bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Optimizing data points per protein increases protein identifications while maintaining quantitative precision in short gradient data-independent acquisition proteomics

Joerg Doellinger [1,], Christian Blumenscheit [1], Andy Schneider [1], Peter Lasch [1]*

[1] Robert Koch-Institute, Centre for Biological Threats and Special Pathogens, Proteomics and Spectroscopy (ZBS6), Berlin, Germany

*corresponding author(s): Joerg Doellinger (Doellingerj@rki.de), phone 49-30-18754-2373

KEYWORDS: data-independent acquisition, predicted spectral library, isolation window, data points per peak, quantitative proteomics, SPEED

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ABSTRACT

The combination of short liquid chromatography (LC) gradients and data independent acquisition (DIA) by mass spectrometry (MS) has proven its huge potential for high-throughput proteomics. This methodology benefits from the speed of the latest generation of mass spectrometers, which enable short MS cycle times needed to provide sufficient sampling of sharp LC peaks. However, the optimization of isolation window schemes resulting in a certain number of data points per peak (DPPP) is understudied, although it is one of the most important parameters for the outcome of this methodology. In this study, we show that substantially reducing the number of DPPP for short gradient DIA massively increases protein identifications while maintaining quantitative precision as the number of data points per protein matters. Quantitative precision on protein level is maintained at low DPPP because the number of identified precursors per protein is largely enhanced. This strategy enabled us quantifying 6018 HeLa proteins (> 80,000 precursor identifications) with coefficients of variation below 20% in 30 min using a Q Exactive Orbitrap mass spectrometer, which corresponds to a throughput of 29 samples per day. This indicates that the potential of high-throughput DIA-MS has not been fully exploited and that high-throughput measurements are also possible with rather slow scanning mass spectrometers.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

INTRODUCTION

Technical progress has transformed mass spectrometry (MS)-based proteomics into a high-throughput technology for analysis of large sample cohorts. This contributes to gain new insights in many different areas, including precision medicine, cell biology, biomarker research and single cell analysis. Many impactful approaches rely on the use of short gradients combined with data-independent acquisition (DIA) and AI-supported data analysis [1-3]. This strategy benefits from the increased speed and sensitivity of the latest generation mass spectrometers. However, due to the high costs of these instruments and fast development of new hardware, which requires large investments every few years to keep the equipment of proteomic labs state-of-the-art, the spread of this technology often falls behind topical demands for high-throughput proteomics.

Research in improving DIA-MS has focused on the implementation of the DIA acquisition strategy on new MS hardware [2, 4], the development of novel acquisition schemes [3, 5-8] and the improvement of data analysis [9-10] as well as library generation [10-13]. However, one of the key aspects of setting up of DIA acquisition methods is largely understudied, namely determining the optimal number of data points per peak (DPPP) [14]. Usually, a number of DPPP is chosen based on rule of thumbs or personal experience rather than experimental data, which in turn specifies a certain cycle time from which a corresponding number of isolation windows follows. In general, this strategy favors the use of fast scanning mass spectrometers enabling the use of more windows at a given cycle time and so the identification of more proteins. It is general knowledge, that lowering the number of DPPP results in decreased quantitative accuracy and precision. In this study we show, that the concept of designing DIA acquisition schemes based on a certain number of data points per peak has limitations. The majority of proteomic studies report protein level data based on precursor measurements. Therefore, the number of data points per protein (DPPPr)

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

should be a better predictor of the quantitative performance compared to DPPP. In this study we show, that window optimization based on decreasing the number of DPPP can increase the number of identified precursors so massively, that quantitative precision is maintained at a largely improved proteome depth as the number of DPPPr is almost constant. This strategy enabled us to quantify 6018 HeLa proteins with coefficients of variations below 20% using a Q Exactive HF orbitrap mass spectrometer at rather moderate scanning speed (12 Hz) in 30 min LC gradients. In total, 7318 proteins were identified in this triplicate analysis from an *in silico* predicted human library. These results correspond to a throughput of 29 samples per day and were achieved using a regular nanoLC and a mass spectrometer that came on the market more than 7 years ago. Noteworthy, our results are quite on pair with published data using the latest generations of MS instruments [1-2], which shows, that the potential of high-throughput DIA-MS has not been fully exploited and that slower scanning MS instruments are also well suited for large-cohort proteomic studies.

EXPERIMENTAL PROCEDURES

Cultivation. *Escherichia coli* K-12 (DSM 3871) was cultivated on Tryptic Soy Agar (TSA) ReadyPlates™ (Merck, Darmstadt, Germany) at 37°C overnight. *Saccharomyces cerevisiae* strain S288C (ATCC 204508) was cultivated on MT agar plates supplemented with hemoglobin and charcoal (MTKH) for 48 h at 37°C. Cells were harvested using an inoculating loop and washed in 2 × 1 mL phosphate-buffered saline (PBS) for 5 min at 4,000 × g and 4°C. HeLa cells (ATCC® CCL-2™) were cultivated in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% fetal calf serum (FCS) and 2 mM L–glutamine at 37°C and harvested at 90% confluency by scraping. Cells were washed in 2 × 2 mL PBS for 5 min at 400 × g and 4°C. Cell numbers in

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

aliquots of the samples were determined using Neubauer counting chambers. For each sample cell numbers were determined in triplicates.

Sample preparation. *E. coli, S. cerevisiae* and HeLa cells were prepared for proteomics using *Sample Preparation by Easy Extraction and Digestion (SPEED)* [15]. At first cells were resuspended in trifluoroacetic acid (TFA) (Optima™ LC-MS grade, Thermo Fisher Scientific, Waltham, MA, USA) (sample/TFA 1:10 (v/v)) and incubated at room temperature for 3 min. Samples were neutralized with 2M TrisBase using 10 × volume of TFA and further incubated at 95°C for 5 min after adding tris(2-carboxyethyl)phosphine (TCEP) to a final concentration of 10 mM and 2-chloroacetamide (CAA) to a final concentration of 40 mM. Protein concentrations were determined by turbidity measurements at 360 nm using a NanoPhotometer® NP80 (Implen, Munich, Germany), adjusted to 1 µg/µL using a 10:1 (v/v) mixture of 2M TrisBase and TFA and then diluted 1:5 with water. Digestion was carried out for 20 h at 37°C using Trypsin Gold, mass spectrometry grade (Promega, Fitchburg, WI, USA) at a protein/enzyme ratio of 100:1. Resulting peptides were desalted using Pierce™ peptide desalting spin columns (Thermo Fisher Scientific) according to manufacturer's instructions and concentrated using a vacuum concentrator. Dried peptides were suspended in 20 µL 0.1% TFA and quantified by measuring the absorbance at 280 nm using the NanoPhotometer® NP80 (Implen). Peptide mixtures of different species were prepared from SPEED preparations of *E. coli* and HeLa cells as well as of a commercially available yeast protein digest (Promega), The ratios of the peptide amount of each species within the three mixtures are given in Table S1.

Liquid chromatography and mass spectrometry. Peptides were analyzed on an EASY-nanoLC 1200 (Thermo Fisher Scientific) coupled online to a Q Exactive™ HF mass spectrometer (Thermo Fisher Scientific). 1 µg of peptides were separated on a PepSep column (15 cm length, 75 µm i.d.,

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

1.5 µm C18 beads, PepSep, Marslev, Denmark) using a stepped 30 min gradient of 80% acetonitrile (solvent B) in 0.1% formic acid (solvent A) at 300 nL/min flow rate: 4–9% B in 2:17 min, 9-26% B in 18:28 min, 26–31% B in 3:04 min, 31–38% B in 2:41 min, 39–95% B in 0:10 min, 95% B for 2:20 min, 95–0% B in 0:10 min and 0% B for 0:50 min. Column temperature was kept at 50°C using a butterfly heater (Phoenix S&T, Chester, PA, USA). The Q Exactive™ HF was operated in a data-independent (DIA) manner in the m/z range of 345–1,650. Full scan spectra were recorded with a resolution of 120,000 using an automatic gain control (AGC) target value of 3×10^6 with a maximum injection time of 100 ms. The full scans were followed by various numbers of DIA scans (Tables S2-8). In order to introduce retention time dependent segments of DIA cycles with fixed cycle times but varying window widths, the window centers are deposited in the inclusion list of the QE method editor along with start and end times. Furthermore, the runtime and the window width of each single DIA scan event needs to be harmonized in accordance to the inclusion list (FIGURE S1). DIA spectra were recorded at a resolution of 30,000 using an AGC target value of 3×10^6 with the maximum injection time set to auto and a first fixed mass of 200 Th. Normalized collision energy (NCE) was set to 27% and default charge state was set to 3. Peptides were ionized using electrospray with a stainless-steel emitter, I.D. 30 µm (PepSep) at a spray voltage of 2.1 kV and a heated capillary temperature of 275°C.

Data analysis. Protein sequences of *Homo sapiens* (UP000005640, downloaded 24/11/21), *E. coli* K-12 (UP000000625, downloaded 26/11/21), and *S. cerevisiae* strain ATCC 204508 (UP000002311, downloaded 29/11/21) were obtained from UniProt. Spectral libraries were predicted using the deep-learning algorithm implemented in DIA-NN (version 1.8) [9] with strict trypsin specificity (KR not P) allowing up to one missed cleavage site in the m/z range of 300 – 1,800 with charges states of 1 – 4 for all peptides consisting of 7-30 amino acids with enabled N-

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

terminal methionine excision and cysteine carbamidomethylation. The mass spectra were analyzed in DIA-NN (version 1.8) using default settings including a false discovery rate (FDR) of 1% for precursor identifications with enabled "match between run" (MBR) option for technical triplicates. The resulting precursor.tsv and pg_matrix.tsv (Lib.Q.Value = 1%) files were used for further analysis in Perseus (version 1.6.5.) [16]. Please note, that the numbers of identified protein groups, which were extracted from the pg_matrix.tsv, differs from the numbers reported in the stats.tsv file as DIA-NN uses different q-values for filtering in these two files. Differentially abundant proteins of the species mixture samples were identified using FDR-adjusted p-values from a t-test with a permutation-based FDR of 0.05 and s0 = 0.1 after normalization of the log-2 transformed MaxLFQ intensities using row cluster subtraction of the human proteins.

RESULTS AND DISCUSSION

DIA window optimization. To achieve a balance between sample throughput and effective use of MS scan time, a 30 min LC gradient was used in this study. This corresponds to a throughput of ~ 29 samples per day as the actual duration of one run on this nanoLC setup is ~ 50 min. The peptide elution window is ~ 26 min and so the mass spectrometer is measuring peptides for ~ 50% of the actual run time. A segmented gradient was used to achieve a uniform elution of peptides within the elution window. At first, we aimed to determine the optimal number of data points per peak (DPPP) with respect to protein identification and quantification. HeLa samples were used to determine the retention time (rt) and mass to charge (m/z) distribution of all detectable precursors from measurements using narrow isolation windows (4 m/z widths with 2 m/z overlap) and gas-phase fractionation (GPF) (8 × 100 m/z, 350 – 1,150 m/z). For this purpose, a library of all possible human precursors was predicted using Prosit [13]. The number of isolation windows for selected numbers of DPPP (1, 1.25 and 1.5) were calculated from the average chromatographic peak width

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(full width at half maximum, FWHM) of the precursors identified in the GPF data using DIA-NN and the durations of the MS^1 and DIA scans (TABLE S3). The dynamic widths of the isolation windows were then selected such as the number of precursors identified in the GPF data was kept constant for each window [14]. It should the noted, that the effective numbers of DPPP in the single-run experiments reported by DIA-NN were slightly above the calculated values, because the sensitivity in the GPF experiment exceeds the single-shot experiments. The DIA acquisition strategy with staggered windows and forbidden zones described by Pino et al., was selected as a reference method [17]. Therefore, the number of isolation windows was calculated from the chromatographic peak widths (6σ) of the precursors identified in the GPF data determined by Spectronaut 15 (Biognosys, Schlieren, Switzerland) and the durations of the MS^1 and DIA scans (Table S3) to achieve the recommended average peak sampling of 10 DPPP (6σ), which corresponds to ~ 3.25 DPPP (FWHM) calculated by DIA-NN.

As a further optimization of DIA window placement, we introduced a retention time dependency for the window widths in our methods (FIGURE 1). The retention time distribution of tryptic peptides in reverse-chromatography is not completely independent of the m/z. This means that in general peptides with lower m/z values tend to elute earlier as peptides with higher m/z values. In order to use this information to increase the selectivity of the DIA isolation windows, we split the peptide isolation window into 5 ranges of equal size and calculated the precursor m/z distribution from the GPF data for each of these ranges independently. Afterwards the window widths were selected within each range, while the number of windows and so the number of DPPP was kept constant for all ranges. A schematic representation of this strategy in shown in figure 1. It should be noted at this point that setting up this method in the Q Exactive instrument software is very time-consuming and somewhat cumbersome.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

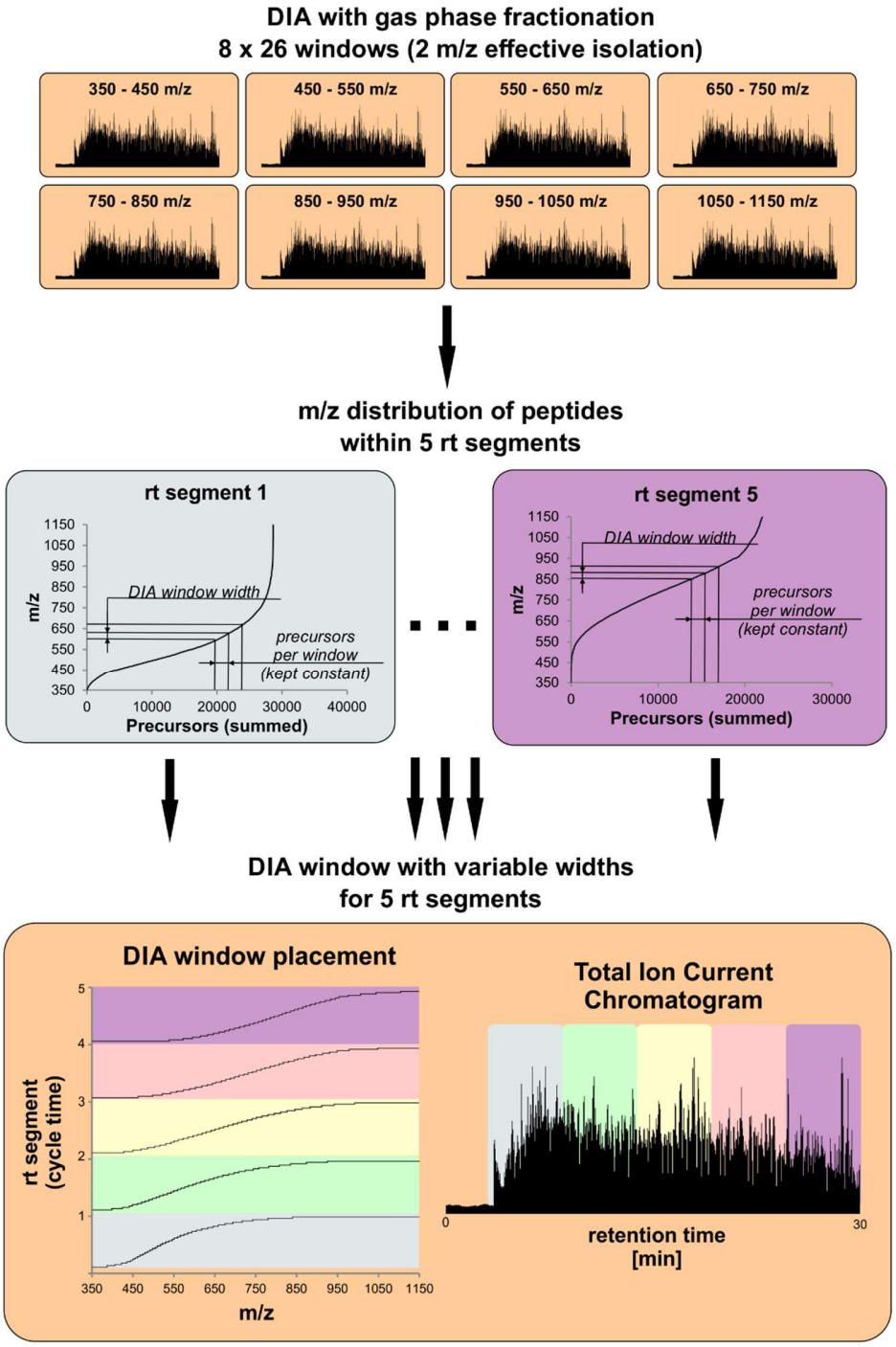


FIGURE 1: Selection of rt dependent DIA isolation window widths.

The m/z distributions of all detectable precursors in HeLa cells within 5 retention time segments were determined from measurements using narrow isolation windows (4 m/z widths with 2 m/z

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

overlap) and gas-phase fractionation (GPF) (8 × 100 m/z, 350 – 1,150 m/z). The resulting m/z distributions illustrate the rt dependency. Isolation window widths were selected separately for each rt segment from the distributions such as the number of precursors identified in the GPF data was kept constant for each window. This information was used to create an instrument method with 5 different rt-dependent isolation schemes but with constant cycle times.

The results of the DIA window optimization experiments are presented in figure 2. All methods were evaluated in triplicate measurements of HeLa cells and independent analysis in DIA-NN using an *in silico* predicted library of the human proteome. Protein identifications increased with decreasing number of DPPP with a maximum of 7,221 proteins using the method with one DPPP ("1 DPPP", 49 variable windows). However, most proteins (5,956) were consistently quantified with coefficients of variation (CV) below 20% in the "1.25 DPPP" method (39 variable windows), which also led to the highest number of precursor identifications (84,175). Therefore the 1.25 DPPP method was used for evaluation of introducing the retention time dependency of the window widths (5 × 39 variable windows). This strategy led to a slight enhancement of the number of precursor and protein identifications and provided the highest identification numbers in all three categories (87,166 precursors, 7,318 identified proteins, 6,018 proteins quantified with CV < 0.2). In total, the 5 × 39 variable windows method represents an increase of 24% in protein identifications and 56% in precursor identifications compared to the reference method, which was set up according to Pino et al [17]. It should be noted that this substantial improvement is solely based on optimizing the window acquisition scheme. The influence of lowering the number of DPPP on the FDR of precursor and protein identifications was analyzed by a subsequent analysis of the 5 × 1.25 DPPP and reference data using a combined *in silico* predicted library of the complete human and *E. coli* proteome. FDRs were calculated based on identified *E. coli* sequences at protein and

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

precursor levels (Table S9) and were with precursor FDR's of ~ 0.1% and protein FDR's of ~ 0.2% proven to be very stable across all methods tested. We wish to note that the number of *E. coli* genes in the database is at the order of 20% of human genes. Although true FDR values are therefore most probably higher, the approach presented nevertheless allows for a relative comparison of the acquisition methods.

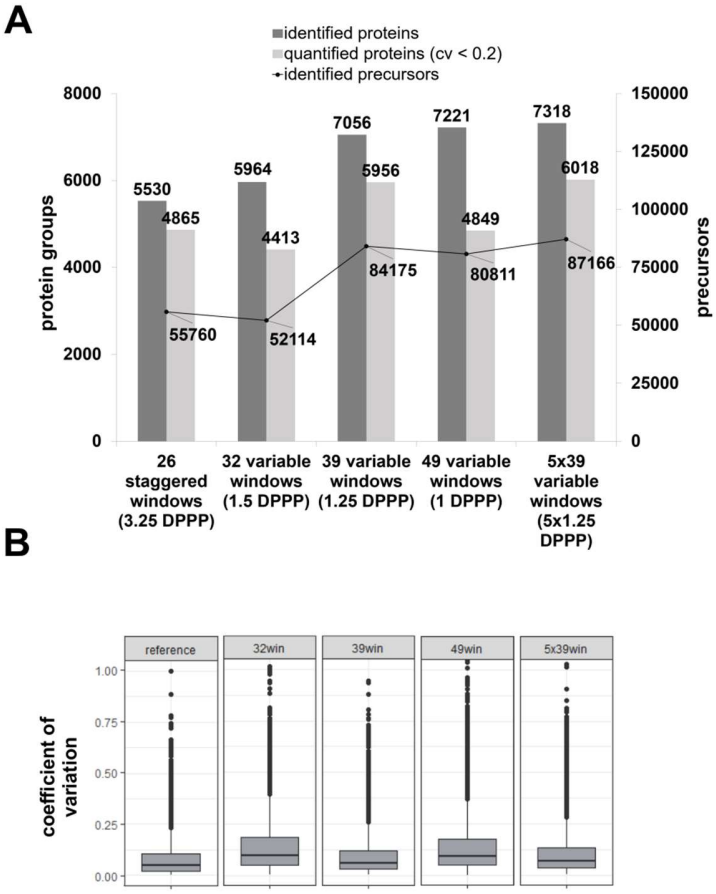


FIGURE 2: Results of DIA isolation window optimization.

HeLa cells were analyzed in triplicates using a 30 min gradient with DIA isolation windows with variable widths. Methods were created for 1, 1.25 and 1.5 data points per peaks (DPPP) on average. Furthermore, a variation of the "1.25 DPPP" method consisting of 5 retention time segments was also included. As a reference, 26 staggered windows of fixed width according to Pino et al. were

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

chosen [17]. Numbers of identified proteins, quantified proteins with coefficients of variation (CV) below 0.2 and identified precursors (mean) are compared in A, while boxplots of the CV's are shown in B.

The protein CV distributions for all acquisition methods are shown in figure 2B. Interestingly, the median protein CV of the 1.25 DPPP method (0.06) exceeds the CV of the reference method (0.05) only slightly, while the 1.25 DPPP method with rt dependent window widths led to another slight increase of the CV (0.07). This is rather unexpected, as the number of DPPP in the reference method is on average 2.6 times larger than in the optimized methods. In order to analyze this observation in more detail, we extracted precursors and proteins, which were detected in the reference and the 5 × 1.25 DPPP methods and compared the CV's on precursor and protein level (FIGURE 3A and B). The CV on precursor level (44,446 shared identifications) is largely affected by the number of DPPP, the median value for the reference method (3.25 DPPP) is 0.078 compared to 0.135 for the 5 × 1.25 DPPP method with rt dependency. However, on protein level (5,312 shared identifications), the difference almost completely disappeared and the median cv is 0.062 for the reference and 0.069 for the 5 × 1.25 DPPP method with rt dependency. The simple reason is, that the distribution of precursors per protein differs largely between the methods (FIGURE 3C). The average number of precursors per shared protein is 12 for the reference method, while it is 27 for the 5 × 1.25 DPPP method with rt dependency. This corresponds to similar average numbers of DPPPr of 34 (5 × 1.25 DPPP) compared to 39 (reference), which explains the aforementioned observations when comparing CV's on precursor and protein level.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

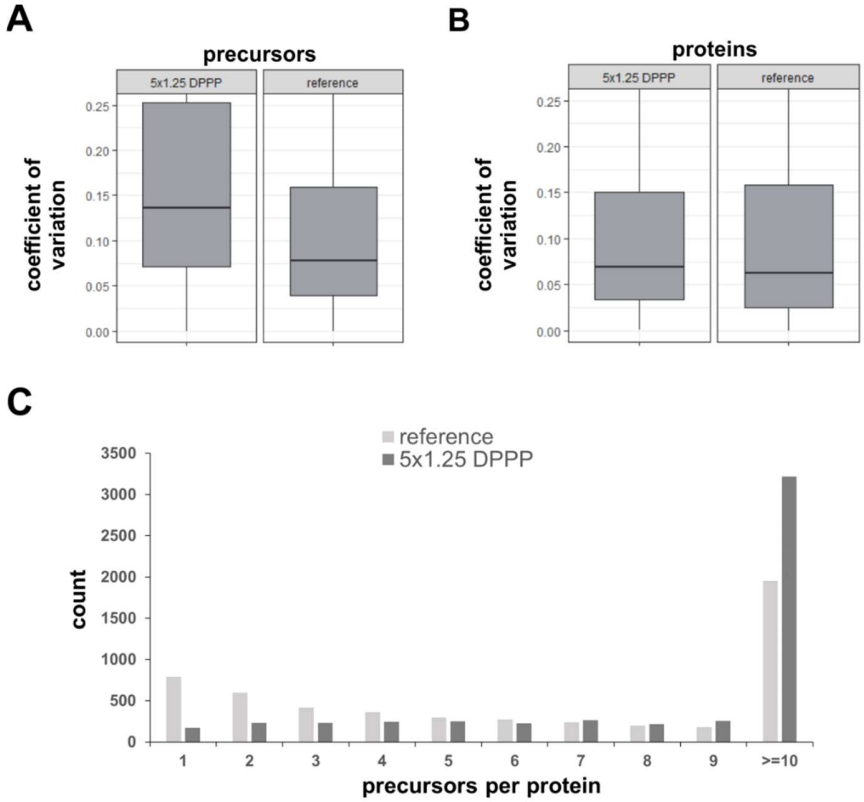


FIGURE 3: Comparison of precursor and protein CV's.

Distributions of precursor (A) and protein (B) CV's of triplicate HeLa cell analyses using a 30 min gradient are compared for the reference acquisition scheme using 26 staggered windows of fixed width according to Pino et al. [17] and the optimized DIA method with variable window widths using 5 rt segments with 1.25 DPPP each. The distribution of precursors per protein is shown in panel C.

Classification of the DIA performance with optimized windows. The performance of the optimized DIA method was compared to published data from the latest generation of mass spectrometers. Therefore, publicly available DIA data of HeLa digests, which were acquired using the 60 samples per day (SPD) method on an EvoSep One LC system coupled either to an Orbitrap Exploris 480 mass spectrometer with and without FAIMS, or a timsTOF Pro instrument were

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

downloaded from PRIDE and compared to the data of this study. All data were analyzed independently for each instrument with enabled MBR using the same *in silico* predicted library of the human proteome in DIA-NN. Of course, this comparison has severe limitations, as the HeLa digests differ between the studies, the gradient of the 60 SPD method is shorter than the one used in this study (21 vs 30 min) and the peptide loading amount varied between 200 to 1000 ng. Nevertheless, this comparison allows us to get a rough overview to assess the performance of each method when using a full organism predicted library. The results are presented in table 1. The optimized method of this study provided the highest numbers in all three categories (87,166 precursors, 7,318 identified proteins, 6,018 proteins quantified with CV < 0.2). The median CV's of all methods are quite similar, except for FAIMS-DIA, which however only represents duplicate measurements. Most strikingly, is the excellent performance of the method optimized in this study with respect to precursor identifications, whose average value e.g. exceeds the data acquired using FAIMS-DIA on an Orbitrap Exploris 480 by 125%. These results show, that the potential of high-throughput DIA-MS is still not fully developed as the strategy for isolation window selection presented in this study could be applied to the latest generation of mass spectrometers as well and that slower scanning MS instruments are also well suited for large-cohort proteomic studies.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

TABLE 1: Comparison of short-gradient DIA data of HeLa cells acquired by different MS instruments

Publicly available datasets of short-gradient DIA HeLa data [1-2] were downloaded from PRIDE [18] and compared to data from this study acquired using the 5 × 1.25 DPPP method. All data were analyzed using the same *in silico* predicted library of the human proteome in DIA-NN with default settings.

	Orbitrap Exploris 480	**Orbitrap Exploris 480**	**Q Exactive HF**	**timsTOF Pro**
method	DIA	FAIMS-DIA	DIA	diaPASEF
PRIDE accession	PXD016662	PXD016662	PXD036451	PXD017703
throughput [samples/day]	60	60	29	60
gradient [min]	21	21	30	21
peptide load [ng]	500	500	1000	200
replicates	3	2	3	3
identified proteins	5588	6542	7318	7079
quantified proteins (CV < 0.2)	4314	5586	6018	5997
identified precursors (mean)	41479	38756	87166	65648
CV (median)	0.077	0.053	0.069	0.062

The proteome depth of the optimized 30 min gradient method was evaluated for cellular proteomes with varying complexity, including *E. coli*, *S. cerevisiae* and *Homo sapiens* (HeLa). The total protein approach [19] was used to calculate the distribution of protein copy numbers in the data (FIGURE 4). The calculations are based on the average protein mass per cell obtained from total protein content measurements and cell counting of the samples. Only proteins, which were identified consistently in all three replicates were considered. The observed limits of detection (95th percentile) for the protein copy numbers per cell are 4 for *E. coli,* 79 for *S. cerevisiae* and

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

5059 for *Homo sapiens* (HeLa). This demonstrates, that almost complete bacterial proteomes are detectable in a 30 min gradient, which offers great opportunities for clinical microbiology including antimicrobial resistance detection and species identification as well as for molecular epidemiology of bacterial infections [20-21].

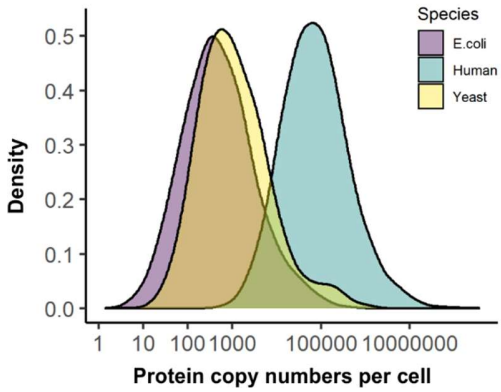


FIGURE 4: Proteome depth of 30 min gradient DIA data for different species.

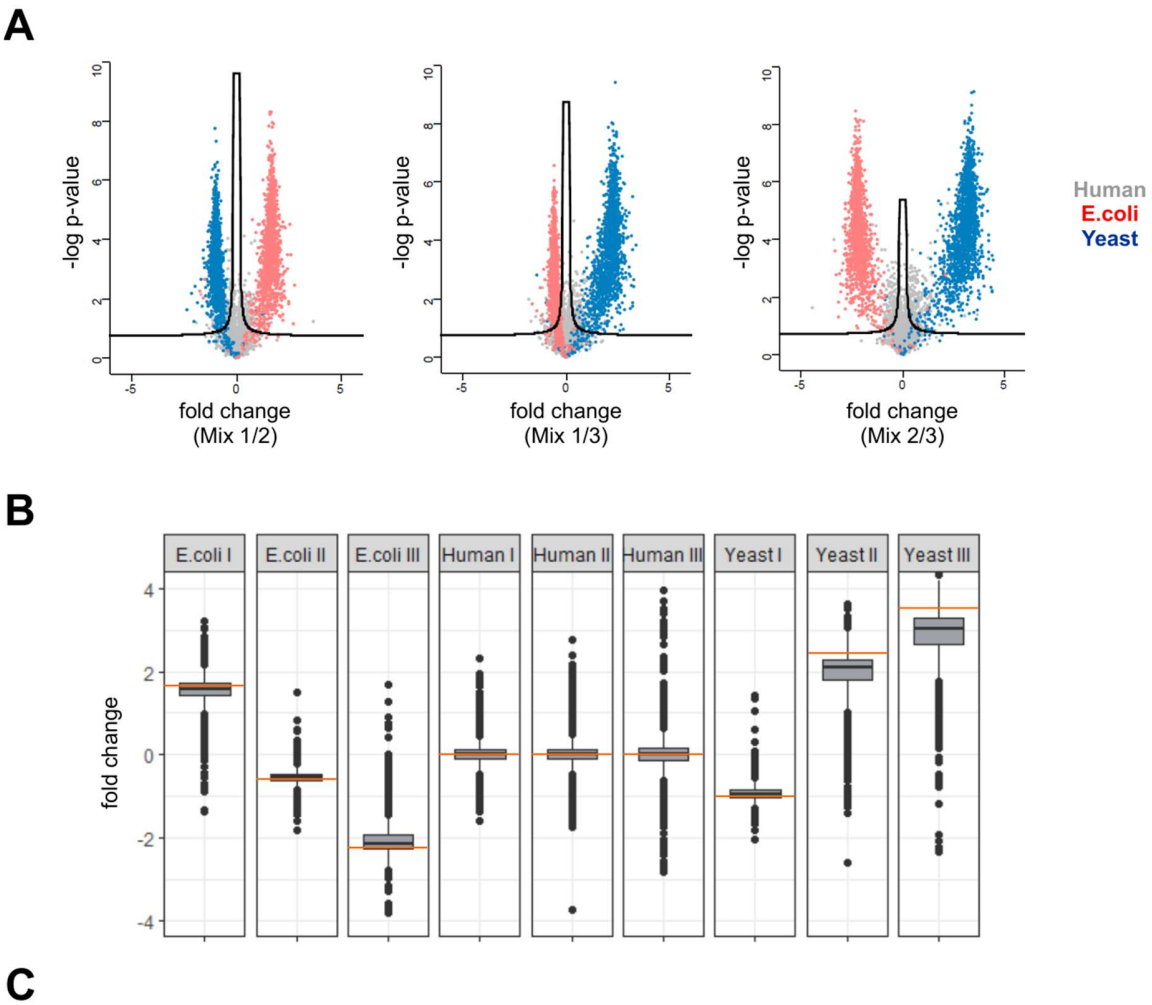
The density distributions of the protein copy numbers per cell were calculated for HeLa, *S. cerevisiae* and *E. coli* cells using the total protein approach [19]. The distributions represent all proteins, which were consistently identified in three replicate measurements.

Evaluation of protein quantification using DIA with optimized windows. The performance of the 1.25 DPPP method with rt dependency was evaluated with respect to protein quantification using mixtures of *Homo sapiens*, *E. coli* and *S. cerevisiae* digests. Purified peptides were mixed at different ratios in order to obtain three different samples (mixtures 1-3) resulting in 9 different fold changes ranging from -5 to 10 (TABLE S1). Differential analysis of protein abundance was performed from triplicate measurements using t-tests with FDR control to account for multiple testing. Percentage of correct protein abundance classification was performed from the statistical results taking the species information of each protein into account. Mixtures were prepared such as the abundance of human proteins was unaltered, while abundance of *E. coli* and *S. cerevisiae*

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

proteins differed between the samples with varying ratios. Volcano plots of the comparisons are shown in figure 5A and the classification results are summarized in 5C. The accuracy of the quantification is visualized as box plots in figure 5B and is summarized as the average percentage deviation from the expected ratio in figure 5C. In general, precision of quantification was high, which resulted in true positive rates > 95% for classification of *E. coli* and yeast proteins in 5 out 6 comparisons. The classification of *E. coli* proteins between the mixtures 1 and 3 was slightly less precise with 91% correct classifications. This comparison has the smallest expected ratio (0.5-fold change), which shows that the precision of classification depends on the abundance ratio. This is quite expectable, as statistical power decreases with lower effect sizes. True negative rates of the human proteins was 94% in two comparisons, which reflects the FDR of 5% used for statistical testing quite well, but decreased to 86% when analyzing mixtures 2 and 3. This comparison contains high expected ratios for *E. coli* (-5-fold change) and *S. cerevisiae* (10-fold change) proteins, which might have introduced some challenges for normalization and statistical testing, which could not be addressed by the quite simple strategy used in this study. The error of abundance ratio estimation was below 10% in 7 out of 9 comparisons and increased up to 25% for yeast proteins when large alterations are expected (-5 and 10-fold change). These results are highly encouraging and show that the short-gradient DIA method with low number of DPPP is well suited for differential protein expression analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



C

Species	Proteins	deviation from expected fold change [%]			correct classification [%]		
		Mix 1/2	Mix 1/3	Mix 2/3	Mix 1/2	Mix 1/3	Mix 2/3
Human	4421	0,0	0,7	0,7	93,8	93,7	85,7
E. coli	1567	-3,7	5,5	9,6	97,3	91	98,9
Yeast	1581	5,7	-22,2	-26,4	97	95,3	98,2

FIGURE 6: Evaluation of quantification precision and accuracy.

Protein quantification accuracy and precision of the 5 × 1.25 DPPP method was analyzed using human, *E. coli* and yeast digests, which were mixed at different ratios in order to obtain three different samples (mix 1-3) corresponding to 9 ratios ranging from -5 to 10 (TABLE S). Differentially abundant proteins were detected using t-tests with permutation-based FDR (5%) control. The results of the statistical analysis are visualized as volcano plots (A). The distributions

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

of the observed fold changes are shown as boxplots along with the expected fold change (orange) in B. The deviation of the observed from the expected fold change is summarized in C in conjunction with the number of identified proteins and the ratios of correct classification, which were calculated independently for each species.

CONCLUSION

High-throughput proteomics benefits greatly from advancements of short gradient DIA methodologies. In this study, we demonstrate that optimization of data acquisition even without any hardware adaptation has a huge untapped potential to increase proteome depth. The strategy presented enabled us to acquire proteomes using a rather slow scanning mass spectrometer up to a depth and precision, achievable to date only by the latest generation of instruments. Our findings should spread the availability of high-throughput proteomics platforms further as it proves, that actually many older instruments can be effectively used for this task. The data also shows, that almost complete bacterial proteomes can be analyzed in just 30 min gradient time. This opens up great opportunities for microbiology applications, a discipline with huge lacks of knowledge on proteome level because of the limited availability of antibodies. The analysis of complete bacterial proteomes in high-throughput mode is thought to be helpful to fully uncover the diagnostic potential of proteomics in clinical microbiology and to provide deeper insights into bacterial evolution when proteomics is used to complement genomics in molecular epidemiology. Furthermore, the presented optimization strategy should in principle be applicable also to faster scanning mass spectrometers and could thus enable further improvements of proteome depth in short gradient DIA proteomics.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ACKNOWLEDGMENTS

Access to proteomics data

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD036451 (Username: reviewer_pxd036451@ebi.ac.uk; Password: 36ZdvZWK) [18].

CONTRIBUTIONS

J.D. and P.L. conceptualized and designed the study. J.D. and A.S. performed the experiments. J.D. analyzed the data, prepared figures and wrote the initial draft of the manuscript. C.B. assisted with data analysis. All co-authors contributed to writing, editing, and reviewing the manuscript.

SUPPORTING INFORMATION

The following supporting information is available free of charge at ACS website http://pubs.acs.org

Supporting tables

Table S1: Species mix preparation

Table S2: Isolation windows for 8 gas-phase fractions with overlapping windows

Table S3: Number of isolation window calculations

Table S4: Isolation windows corresponding to 1.5 data points per peak (FWHM, 32 windows)

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table S5: Isolation windows corresponding to 1.25 data points per peak (FWHM, 39 windows)

Table S6: Isolation windows corresponding to 1.0 data points per peak (FWHM, 49 windows)

Table S7: Isolation windows corresponding to 10 data points per peak (calculated according to Pino et al.)

Table S8: Isolation windows with retention time dependancy corresponding to 1.25 data points per peak (FWHM, 5 × 39 windows)

Table S9: False discovery rate comparison in HeLa samples using a combined library of the human and *E. coli* proteome

Supporting figures

Figure S1: DIA with retention time dependency method setup on a Q Exactive

REFERENCES

1. Bekker-Jensen, D. B.; Martinez-Val, A.; Steigerwald, S.; Ruther, P.; Fort, K. L.; Arrey, T. N.; Harder, A.; Makarov, A.; Olsen, J. V., A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients. *Mol Cell Proteomics* **2020,** *19* (4), 716-729.
2. Meier, F.; Brunner, A. D.; Frank, M.; Ha, A.; Bludau, I.; Voytik, E.; Kaspar-Schoenefeld, S.; Lubeck, M.; Raether, O.; Bache, N.; Aebersold, R.; Collins, B. C.; Rost, H. L.; Mann, M., diaPASEF: parallel accumulation-serial fragmentation combined with data-independent acquisition. *Nat Methods* **2020,** *17* (12), 1229-1236.
3. Messner, C. B.; Demichev, V.; Bloomfield, N.; Yu, J. S. L.; White, M.; Kreidl, M.; Egger, A. S.; Freiwald, A.; Ivosev, G.; Wasim, F.; Zelezniak, A.; Jurgens, L.; Suttorp, N.; Sander, L. E.; Kurth, F.; Lilley, K. S.; Mulleder, M.; Tate, S.; Ralser, M., Ultra-fast proteomics with Scanning SWATH. *Nat Biotechnol* **2021,** *39* (7), 846-854.
4. Wang, Z.; Mülleder, M.; Batruch, I.; Chelur, A.; Textoris-Taube, K.; Schwecke, T.; Hartl, J.; Causon, J.; Castro-Perez, J.; Demichev, V.; Tate, S.; Ralser, M., High-throughput proteomics of nanogram-scale samples with Zeno SWATH DIA. *bioRxiv* **2022**, 2022.04.14.488299.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

5. Amodei, D.; Egertson, J.; MacLean, B. X.; Johnson, R.; Merrihew, G. E.; Keller, A.; Marsh, D.; Vitek, O.; Mallick, P.; MacCoss, M. J., Improving Precursor Selectivity in Data-Independent Acquisition Using Overlapping Windows. *J Am Soc Mass Spectrom* **2019,** *30* (4), 669-684.
6. Mehta, D.; Scandola, S.; Uhrig, R. G., BoxCar and Library-Free Data-Independent Acquisition Substantially Improve the Depth, Range, and Completeness of Label-Free Quantitative Proteomics. *Anal Chem* **2022,** *94* (2), 793-802.
7. Cai, X.; Ge, W.; Yi, X.; Sun, R.; Zhu, J.; Lu, C.; Sun, P.; Zhu, T.; Ruan, G.; Yuan, C.; Liang, S.; Lyu, M.; Huang, S.; Zhu, Y.; Guo, T., PulseDIA: Data-Independent Acquisition Mass Spectrometry Using Multi-Injection Pulsed Gas-Phase Fractionation. *J Proteome Res* **2021,** *20* (1), 279-288.
8. Mun, D. G.; Renuse, S.; Saraswat, M.; Madugundu, A.; Udainiya, S.; Kim, H.; Park, S. R.; Zhao, H.; Nirujogi, R. S.; Na, C. H.; Kannan, N.; Yates, J. R., 3rd; Lee, S. W.; Pandey, A., PASS-DIA: A Data-Independent Acquisition Approach for Discovery Studies. *Anal Chem* **2020,** *92* (21), 14466-14475.
9. Demichev, V.; Messner, C. B.; Vernardis, S. I.; Lilley, K. S.; Ralser, M., DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* **2020,** *17* (1), 41-44.
10. Searle, B. C.; Pino, L. K.; Egertson, J. D.; Ting, Y. S.; Lawrence, R. T.; MacLean, B. X.; Villen, J.; MacCoss, M. J., Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat Commun* **2018,** *9* (1), 5128.
11. Searle, B. C.; Swearingen, K. E.; Barnes, C. A.; Schmidt, T.; Gessulat, S.; Kuster, B.; Wilhelm, M., Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nat Commun* **2020,** *11* (1), 1548.
12. Van Puyvelde, B.; Willems, S.; Gabriels, R.; Daled, S.; De Clerck, L.; Vande Casteele, S.; Staes, A.; Impens, F.; Deforce, D.; Martens, L.; Degroeve, S.; Dhaenens, M., Removing the Hidden Data Dependency of DIA with Predicted Spectral Libraries. *Proteomics* **2020,** *20* (3-4), e1900306.
13. Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; Reimer, U.; Ehrlich, H. C.; Aiche, S.; Kuster, B.; Wilhelm, M., Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat Methods* **2019,** *16* (6), 509-518.
14. Doellinger, J.; Blumenscheit, C.; Schneider, A.; Lasch, P., Isolation Window Optimization of Data-Independent Acquisition Using Predicted Libraries for Deep and Accurate Proteome Profiling. *Anal Chem* **2020,** *92* (18), 12185-12192.
15. Doellinger, J.; Schneider, A.; Hoeller, M.; Lasch, P., Sample Preparation by Easy Extraction and Digestion (SPEED) - A Universal, Rapid, and Detergent-free Protocol for Proteomics Based on Acid Extraction. *Mol Cell Proteomics* **2020,** *19* (1), 209-222.
16. Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M. Y.; Geiger, T.; Mann, M.; Cox, J., The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* **2016,** *13* (9), 731-40.
17. Pino, L. K.; Just, S. C.; MacCoss, M. J.; Searle, B. C., Acquiring and Analyzing Data Independent Acquisition Proteomics Experiments without Spectrum Libraries. *Mol Cell Proteomics* **2020,** *19* (7), 1088-1103.
18. Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Perez, E.; Uszkoreit, J.; Pfeuffer, J.;

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.12.507556; this version posted September 13, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Sachsenberg, T.; Yilmaz, S.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaino, J. A., The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* **2019,** *47* (D1), D442-D450.
19. Wisniewski, J. R.; Rakus, D., Multi-enzyme digestion FASP and the 'Total Protein Approach'-based absolute quantification of the Escherichia coli proteome. *J Proteomics* **2014,** *109*, 322-31.
20. Blumenscheit, C.; Pfeifer, Y.; Werner, G.; John, C.; Schneider, A.; Lasch, P.; Doellinger, J., Unbiased Antimicrobial Resistance Detection from Clinical Bacterial Isolates Using Proteomics. *Anal Chem* **2021,** *93* (44), 14599-14608.
21. Lasch, P.; Schneider, A.; Blumenscheit, C.; Doellinger, J., Identification of Microorganisms by Liquid Chromatography-Mass Spectrometry (LC-MS(1)) and in Silico Peptide Mass Libraries. *Mol Cell Proteomics* **2020,** *19* (12), 2125-2139.