Published in final edited form as:
Nat Methods. 2020 December ; 17(12): 1237–1244. doi:10.1038/s41592-020-00986-4.

Avant-garde: An automated data-driven DIA data curation tool.

Alvaro Sebastian Vaca Jacome[1], **Ryan Peckner**[1], **Nicholas Shulman**[2], **Karsten Krug**[1], **Katherine C. DeRuff**[1], **Adam Officer**[1], **Karen E. Christianson**[1], **Brendan MacLean**[2], **Michael J. MacCoss**[2], **Steven A. Carr**[1], **Jacob D. Jaffe**[1,3]

[1]Broad Institute of MIT and Harvard, Cambridge, MA

[2]University of Washington Genome Sciences, Seattle, WA

[3]Inzen Therapeutics, Cambridge, MA

Abstract

Multiple challenges remain in Data-Independent Acquisition (DIA) data analysis, like confidently identifying peptides, defining integration boundaries, removing interferences, and controlling false discovery rates. In practice, a visual inspection of the signals is still required, which is impractical with large datasets. We developed *Avant-garde* as a tool to refine DIA (and PRM) data. *Avant-garde* uses a novel data-driven scoring strategy; signals are refined by learning from the data itself, using all measurements in all samples to achieve the best optimization. We evaluated *Avant-garde's* performance with benchmarking DIA datasets. We showed that it can determine the quantitative suitability of a peptide peak, and reaches the same levels of selectivity, accuracy, and reproducibility as manual validation. *Avant-garde* is complementary to existing DIA analysis engines and aims to establish a strong foundation for subsequent analysis of quantitative MS data.

Editorial summary:

A computational tool, Avant-garde, automates refinement of data-independent acquisition mass spectrometry-based quantitative proteomics data.

Corresponding authors: Jacob D. Jaffe: jjaffe@inzentx.com; Alvaro Sebastian Vaca Jacome: svaca@broadinstitute.org.

Author Contributions
A.S.V.J. conceived of the idea, designed and performed experiments, collected the data, authored software, and wrote the manuscript. R.P. provided help with the data analysis of the benchmarking datasets. N.S. and B.M. adapted Skyline to facilitate the use of *Avant-garde* as an External tool. B.M. performed data processing and validation of the LFQBench and Extended Benchmark datasets. K.K. provided help with the statistical analysis of the benchmarking datasets. K.C.D. and A.O. carried out experiments and collected data for the P100 dataset. K.C.D., A.O. and K.E.C. beta-tested the software and provided help with the data analysis. M.J.M. and S.A.C. provided laboratory resources and guidance on the manuscript. J.D.J. provided laboratory resources, provided input on the software, provided guidance on the experimental design and wrote the manuscript.

Present addresses:
RP: Cogen Therapeutics, Cambridge, MA
JDJ: Inzen Therapeutics, Cambridge, MA

Conflicts of Interest
The MacCoss Lab at the University of Washington (members N.S., B.M., and M.J.M.) has a sponsored research agreement with Thermo Fisher Scientific, the manufacturer of the instrumentation used in this research. Additionally, M.J.M. is a paid consultant for Thermo Fisher Scientific. J.D.J is employed by Inzen Therapeutics and declares that he has no conflict of interest. The remaining authors declare no competing interests.

Introduction

Data-Independent Acquisition (DIA) combines the protein-wide coverage of Data-Dependent Acquisition (DDA) with the reproducibility, sensitivity, and accuracy of targeted methods[1–4]. In DIA, MS instrumentation co-isolates and fragments multiple peptides, either in sequential isolation windows traversing an *m/z* range, or all at once[1–9]. DIA has the potential to comprehensively analyze all peptides in a sample that are above the instrument's limit of detection.

DIA data are quantified with a chromatogram-based approach. For each peptide, several transitions (precursor/fragment ion pairs) are monitored over time, producing a set of chromatographic peak traces. Peak area is integrated and used as a proxy for analyte abundance. Ideally, all transitions of a given analyte should have: 1) the same elution peak shape, 2) relative areas mirroring the relative intensities found in their reference spectrum from a library, 3) a low mass error, and 4) consistency across all LCMS experiments being compared. However, due to the complexity of DIA data, it is difficult to obtain signals that correspond to this archetype, and data analysis remains challenging.

Several tools have been developed to analyze DIA data[10]. Each one can produce a different set of detectable peptides and quantitative results, even with standardized samples and data sets[11]. This variability is introduced by differences at all stages of data analysis (i.e. raw data processing, protein database search, peak detection, transition selection, chromatogram extraction, peak integration, and statistical analysis), each of which can affect detection and quantification.

Most tools focus on statistical validation of peptide detection (using target/decoy approaches[12,13]) but do not address the *quantitative suitability* of the signals extracted. Targeted analyses of DIA data begin with DDA spectral libraries, which may be built from a single, fractionated "master sample". Transition selection from spectral libraries may not be suitable for quantification as these practices mask the complexity found in real DIA data and do not anticipate interferences present in real biological samples. In practice, further curation by visual inspection of the signals by an expert in the field is required for rigorous quantitation to remove transitions subject to interference, and manually corrects peak integration boundaries. However, time-consuming manual curation is impractical with large datasets, and produces subjective user-dependent results. Curation is thus often omitted and the output of these tools is used at face-value for downstream analysis.

Another problem for DIA approaches is missing values, that can have a biological origin (e.g. peptides truly not present), a technical origin (e.g., peptide loss during sample processing), or a computational origin, (e.g., failure to assign the correct signal to the respective peptide, improper retention time prediction of chromatographic peak boundaries in real samples). Non-curated data can also produce missing data as a peptide subject to interference might fail to be identified in a subset of samples (e.g. indistinguishable from a decoy peptide), even though the peptide is present. However, if another set of interference-free transitions had been used, this peptide might be detected in all samples, subsequently providing accurate and reproducible quantification.

The issues discussed above motivated us to create a tool for automated targeted MS data curation. Here we present *Avant-garde* (AvG), a modular tool meant to polish the results of DIA and PRM analysis tools. Building upon earlier work on DIA and PRM data optimization[14], AvG refines DIA signals to reach the highest possible levels of sensitivity, selectivity, and accuracy. AvG refines peak detection, adjusts peak boundaries, removes transitions subject to interference, eliminates noise, and estimates the FDR of analytes for *quantitative suitability*, which we define as the proper attribution of a high-quality abundance signal (e.g., MS intensity) to a peptide believed to be present in a biological sample. Unlike other tools where MS runs are scored independently from each other, AvG uses an ensemble data-driven scoring strategy. DIA signals are refined by learning from the data itself, using all measurements in all samples together to achieve the best optimization.

Results

Principle of Avant-garde:

AvG is a tool designed for automated data curation, meant to complement common DIA analysis tools such as mProphet[12], OpenSWATH[13], DIA-Umpire[15], EncyclopeDIA[16], and Specter[17]. To ease use and adoption of AvG, we have chosen Skyline[18] to extract chromatogram data as a vendor-independent and user-friendly tool. It enables data visualization and provides a common framework to refine results from different upstream tools. Skyline requires only the peptide sequences and peak integration boundaries determined by these tools. *Avant-garde* can be used on any type of data that produces fragment-ion chromatograms at the MS2 level, including PRM and DIA data.

AvG uses extracted ion chromatogram data and employs three independent modules for refinement (Fig. 1). First, a transition refinement module improves the choice of transitions to eliminate interferences and reduce noise. Second, a peak refinement module adjusts integration boundaries without the need for spiked-in retention time peptides. A third module scores peaks using a number of intuitive metrics and estimates the false discovery rate (FDR) for quantitative suitability. The refinement results and scoring metrics are then imported back into Skyline.

Like other tools, AvG assigns and quantifies peptides using composite scores (built from subscores) as the basis for quality filtering of results and estimation of the FDR. Each module of AvG produces its own composite score: the "*AvG fitness score*" for transition selection, the "*AvG chromatographic score*" for peak integration boundaries, and the "*AvG score*" for the final scoring of peaks and FDR estimation. However, its composite scores are calculated as the product rather than the sum of its subscores. This approach avoids allowing any *single* subscore to push the composite score over an arbitrary "significance" threshold employed to control the FDR. Uniquely, AvG calculates its module scores in an ensemble-driven manner, curating transitions and peak boundaries while *considering data from all samples in a set*.

The scoring strategy is designed to produce very conservative results. AvG penalizes peptides with any single metric that indicates poor quality. This scoring mechanism imposes strong penalties on transitions subject to interference. A high final AvG score ensures that minimal interference is present and that the signals are suitable for quantification. A detailed explanation on the automated refinement of the transition selection and peak boundary definition is given in the Supplementary Note.

AvG ensures high accuracy and precision

The precision and accuracy of AvG were evaluated using a dilution series of 95 synthetic phosphopeptides that were spiked into a HEK293T digest. A typical AvG performance on the spiked-in synthetic peptide S[+80]LTAHSLLPLAEK is shown in Fig. 2A. Skyline's initial extraction of signals from this peptide is incorrect in some runs, due to misassignment of peak boundaries. These aberrant signals cause departure from the expected linear relationship between concentration and peak area ($r^2 = 0.44$, Fig. 2A, top). After curation by AvG, the expected linear relationship is recovered ($r^2 = 0.99$, Fig. 2A, bottom) as AvG automatically corrected the peak boundaries.

There was marked improvement for aggregate measurements of all 95 synthetic peptides after application of AvG (as compared to initial, unoptimized Skyline extractions) in several figures of merit (Fig. 2B). The precision, measured by the CVs of triplicates, improved from 43.2% to 5.6%. The correlation coefficient between peptide concentration and peak area improved from r^2=0.85 to r^2=0.99. Improvements in the fraction of measurements with less than 20% absolute error (for our definition of accuracy, see Online Methods) were also evident. Finally, the relative quantification accuracy was also evaluated. We calculated the ratios between the mean area of each calibration point to the mean area of the fourth calibration point (P4 in Fig. 2C). The distribution of measured ratios after AvG refinement is clearly much tighter and closer to the expected values.

AvG results are concordant with expert manual curation

To further evaluate the performance of the automated data curation, we applied AvG to a reduced-representation phosphoproteomics dataset obtained for our LINCS project[19]. This dataset, acquired in DIA mode, had previously been manually curated by an expert in our laboratory, which we consider the gold-standard against which other approaches were compared. We curated data across 96 samples for 95 phosphopeptides for which isotopically-labeled heavy peptide counterparts were present. For the "unoptimized" analysis, the 5 most intense transitions from the spectral library were chosen and the peak boundaries were defined by Skyline. For the optimized version, all possible b- and y-ions above b_4 and y_4 were extracted and subjected to further curation by AvG. AvG was run in two modes: 1) "open" curation, where no subscore or composite score filters were applied, and 2) "filtered" curation, where subscore filters were introduced.

The comparison of light-to-heavy ratios between the manually curated and unoptimized analyses (Fig. 3A) had many points deviating from the ideal x=y line. After "open" curation by AvG (Fig. 3B), many fewer points deviated from this line. The disagreements that remained could be explained by peptides where AvG chose different transitions than the manual curator, producing discrepant light-to-heavy ratios. These differences were enhanced if either the light or the heavy peptide had low intensity. In that case, any small change in the signal would have a large impact on the ratio. The results of filtered curation correlated even better with manual curation (r=0.99, Fig. 3C). Signals creating discrepancies in the open curation analysis were filtered out showing that they were derived from low-quality data.

Analysis of 190 precursors and 18240 individual measurements was theoretically possible (95 peptides x 2 isotopic label states x 96 samples). AvG improved the data completeness over unoptimized analysis and even manual curation (19% and 12%, respectively, at the measurement level, Fig. 3D). Overall, the data curation by AvG enabled the quantification of 92% of all peptides with data completeness of 79%. AvG performed its curation of this dataset in < 1 hr of unsupervised time, while typically it takes a manual curator > 10 hours of “hands-on” work.

Evaluation of AvG with LFQBench

We asked whether AvG could further improve quantification when applied to a DIA benchmarking dataset, LFQBench[11], as compared to the many tools with which it has already been analyzed. This dataset was collected on a time-of-flight mass spectrometer, resulting in different data characteristics (resolution, mass accuracy, baseline noise level) than the Orbitrap-class data on which AvG was developed.

The example shown in Fig. 4 compares the Skyline analysis to the AvG curation of the LFQBench HYE110 dataset, acquired with a SWATH method with 64 variable m/z windows[11]. The first dataset corresponds to the Skyline analysis using its implementation of mProphet for peak picking, FDR estimation (filtered using a q-value<0.01) and data validation. The two samples are each mixtures of three complex proteomes - *E. coli*, human and yeast - formulated as shown in Fig. 4A. Three expected ratios are possible when comparing sample A to B (0.1, 1 and 10 for the *E. coli*, human and yeast peptides respectively). The results extracted using Skyline show that some ratio data points deviate from the expected values (Fig. 4B). Indeed, the mean and median percent errors were 171% and 26.3% for *E. coli* and 42.7% and 28.6% for yeast (Fig. 4E and Supp. Table 1). These deviations also affected the precision of the measurement, calculated for triplicate values (mean and median CV of 14.0% and 8.0% for *E. coli*, 11.3% and 7.0% for human and 13.9 and 7.9%for yeast peptides respectively, Fig. 4F and Supp. Table 1).

After curating and filtering the data with AvG, the ratio distributions were closer to the expected values (Fig. 4C). AvG produced very conservative results. The total number of reported peptides with high-quality and properly attributed signals after curation was lower than initially reported by the upstream tools. Indeed using Skyline and mProphet (with a q-value cutoff of 0.01) and following the metrics established in the LFQBench paper[11], 34,494 peptides could be identified, of which 19,534 provided valid quantified ratios. After AvG, 20% fewer peptides and ratios could be identified (Fig. 4D and Supp. Table 1). However, the median A/B ratio for each proteome improved and was 0.11 ± 0.016 for *E. coli*, 1.03 ± 0.09 for human, and 10.87 ± 1.92 for yeast peptides. The precision improved (mean and median CV of 6.8 and 5.6 for *E. coli,* 6.4 and 5.5% for human, and 6.8 and 5.7 % for yeast peptides respectively; Fig. 4F), as did the accuracy (mean and median percent error 68.5% and 16.5% for *E. coli* and 26.1% and 19.2% for yeast respectively; Fig. 4E and Supp. Table 1). The mProphet results yielded similar results in terms of accuracy and precision as AvG when applying a much more stringent q-value cutoff value of 0.0001. However the difference in the order of magnitude in the threshold required to obtain the same levels of signal quality could be explained because the current implementation of mProphet in Skyline and the FDR estimation in AvG does not yet consider different context-dependent error rate estimation strategies that could correct a potential miscalibration of the q-values calculation[20].

To demonstrate that AvG can improve results generated with multiple upstream DIA tools, we have also examined the results of the LFQBench dataset analyzed with OpenSWATH and the TRIC retention time alignment tool[21]. We have found the use of *Avant-garde* improves the accuracy of the quantification (Fig. S1). AvG identified 25% fewer peptides and ratios and the reproducibility of the measurements was in this case very slightly improved after the curation (mean CV of 6.8%, 6.4%, and 6.8% for *E. coli*, human and yeast peptides respectively, Fig. S1F). However, the most important change was observed for the accuracy of the quantification. The median percent error decreased from 24.8% to 18.8% and38.1% to 16.4% for yeast and *E. coli* respectively after curation with *Avant-garde* (Fig. S1D).

To further evaluate AvG, we created a complex benchmarking set of 4 samples consisting of a mixture of three complex proteomes. The total amount of protein and the proportion of the human proteome was kept constant in all samples, while the proportion of *E. coli* and yeast varied (Fig. S2). Six pairwise combinations of the samples are possible, resulting in 12 “ground truth” ratios ranging from 1.2-fold to 10-fold, plus a constant 1:1 ratio of human peptides for all possible comparisons. This experimental design enabled the estimation of reproducibility across many MS runs having different sample compositions, with some compositions more prone to interferences than others.

The resulting dataset has a large peptide abundance dynamic range and emulates ratios close to typical thresholds of biological significance for the evaluation of DIA analysis tools[11,22]. After data refinement with AvG the quantification accuracy and precision improved. The results of this evaluation can be found in the Supplementary Note.

Performance of AvG under conditions emulating real biological data

Detection of changes in protein levels between two sample classes (e.g., diseased vs. healthy, treated vs. control) is a major paradigm for quantitative proteomics. To evaluate whether AvG curation would help achieve this goal, we simulated biological data to create a realistic scenario in which most peptides in the data set were “unchanged,” while a small minority were up- or down-regulated. This was practically achieved by downsampling the benchmarking data (Supplementary Note) to include 90% human analytes (3000 peptides, unchanged), 5% *E. coli* peptides (positive fold-changes), and 5% yeast peptides (negative fold-changes). The analytes were chosen at random from the larger pool of peptides, allowing us to bootstrap the analysis by selecting different subsets.

We calculated ratios of peptides between the different sample compositions before and after AvG curation, and compared them to the expected ratios (Fig. S2). Significance (p) values were assigned to the ratios using a moderated t-test and corrected for multiple hypothesis testing[23]. Peptides were classified as differentially expressed if their adjusted p-value was lower than 0.05 and their absolute fold change was $> 2\sigma$ of the fold changes for the (unchanged) human peptides present in the downsampled data, and considered “accurate” if their observed ratio was 80-120% of expected. Knowledge of the species-of-origin for each peptide allowed us to classify results as true- or false-positive.

AvG improved the ability to detect changes in protein expression. As an example, we compared sample A to B before and after AvG curation (Fig. 5A). The improved accuracy and precision obtained after AvG resulted in a much higher number of true positive hits (blue and red full disks) and lower number of false positives hits (green full disks not within the grey area). Additionally, the number of accurate measurements (observations between the dashed lines) increased after curation.

To quantify the improvement in performance, we calculated the recall, % of “accurate” measurements, and false positive rates for detection of differentially expressed peptides across the range of fold-changes using 1000 bootstrap iterations as described in Supp. Methods. The results are illustrated in Fig. 5B, with the shaded areas indicating improvement in figures-of-merit achieved after AvG curation. AvG increased the recall and led to a higher number of correct calls of significance for differentially expressed peptides (Fig. 5B, left). After curation, a recall higher than 92% was achieved for any absolute fold change above 2.0. In comparison, the mProphet data had a lower recall by 2 to 12% in the same range. In addition, we observed an improvement in the percentage of true positive hits that were classified as being *accurate*, with a median improvement of 13% (Fig. 5B, middle). Furthermore, the false positive rate decreased by a median value of 7.9%. Additionally, we verified that the similar results to AvG can be obtained by filtering mProphet results using a stringent q-value cutoff of 0.0001 (Fig. S3 and Supp. Table 2).

Discussion

Data curation is an extremely important but often overlooked step in transition-based quantitative proteomics. We have demonstrated that AvG can curate DIA data in an automated manner. AvG tailors the choice of transitions to each dataset to minimize noise and increase the reliability of quantification. A key feature of AvG is that each peptide is reassessed with an independent global scoring module after curation to control a dataset-level FDR for *quantitative suitability*, not just detection. Quantitative suitability is an important new concept for DIA. It is a shorthand for the correct attribution of abundance signals to peptides thought to be present in a dataset, and illustrates why AvG is complementary to and not competitive with other DIA analysis tools. AvG takes the position that a peptide has been properly “identified” in the data set by some other search strategy (i.e., it “exists” in the samples that were analyzed). By scoring the signals after the data curation and peak boundaries refinement, the AvG score becomes a measurable and quantitative metric of the quality of the signals. Counterintuitively, the number of detected peptides may go down after AvG, but the quality of quantitation for those peptides will be higher. We have empirically demonstrated that the AvG score tends to be low for “decoy” analytes, and data curated with AvG consistently produces FDRs < 1.0% at the thresholds we have defined.

Other software tools for DIA analysis typically simply extract the 5-10 most intense transitions from the spectral library. This approach does not guarantee interference-free transitions for analyzing complex biological samples, where curation is paramount. An alternative approach is to select the transitions that are predicted to be unique to their precursor ion using tools like SRMCollider[24]. However, interferences are difficult to predict due to run-to-run chromatographic variability and changes in sample composition. These tools do not consider retention time or fragment ion intensities, hindering accurate curation. *A priori* prediction of interference, reliant on protein databases and other user choices, does not anticipate real-world LCMS data artifacts. Another *a priori* approach to curation, SWATHProphet[25], uses spectral libraries embedded with retention time information to anticipate quantitative interferences. In this case, library completeness (again traceable to experimental and user decisions) governs the success of the approach. In contrast, AvG uses an *a posteriori* approach to curate data that explicitly considers LCMS artifacts *and* utilizes prior knowledge but is not limited by it. AvG starts from a larger initial number of transitions (5-25 transitions) and optimizes the choice to find the most suitable set of transitions for quantification(at least 4 transitions) (Fig. S4). SWATHProphet, based on the mProphet discriminant score, also implements an approach for *a posteriori* flagging of poor transitions. Its application requires iterative cycles of optimization and data re-extraction, and again relies on spectral libraries as the primary source of interference detection. Further, it focuses on improving quantitation for peptides that already have a high mProphet score, rather than potentially improving scores of borderline peptides. This approach is apt to produce false negatives and will fail to rescue suitable data signals. Other tools, such as MSstats[26] and mapDIA[27], also calculate a correlation of transitions to detect interfered transitions during the post-acquisition analysis of the data. These tools do a single-pass refinement of transitions as a data post-processing step, rather than an active iterative optimization to achieve the best signals. MSstats[26,28] also has an automated statistical approach to detect features that contain few missing values and having intensities that are consistent with the majority of the corresponding protein intensity profiles across the MS runs. This approach flags outlier features not meeting these criteria for further investigation, curation or removal. This work illustrates the necessity of performing a transition refinement step in DIA data curation in order to obtain high accuracy and sensitivity[28]. AvG focuses on flagging and automatically curating the “cleanest” transitions that are the best suited for quantification, and can improve signal quality for marginal cases. It does not require iterative cycles and is not bound to any specific DIA or PRM workflow, it also includes a peak boundary refinement algorithm and is fully implemented as an external tool in Skyline.

The objective of quantitative proteomics is to identify differentially expressed proteins or peptides. Therefore, it is important to evaluate new methods with scenarios that mimic conditions found in real biological milieux. For us, that meant creating a dataset where the majority of peptide analytes were unchanged between two sample classes, while a minority were changing with a known ratio (Supplementary Note). We evaluate AvG in ranges of borderline biological significance (1.5 - 2.0-fold) as well as extreme significance (>5-fold) to truly assess method performance. We were pleased to find that AvG could enable discrimination of low fold changes with fairly high sensitivity and accuracy, and that, across the board, it can add value to the work done by other DIA analysis tools by improving *quantitative suitability* of the data. To us, this means that it can help produce more accurate and reproducible quantification results, providing more granularity in the elucidation of the complex dynamics of proteomes.

AvG's ensemble-driven scoring strategy is designed to produce very conservative results by penalizing poor-quality signals. Its combined score is a weighted product of run-specific and dataset-wide subscores that intuitively map to common LCMS data quality metrics. AvG penalizes sets of transitions for peptides with any single poorly-scoring metric, making it very sensitive to interferences. However, its evolutionary optimization approach ultimately selects sets of transitions that produce the lowest levels of noise and the highest level of parsimony for signals across the entire dataset. Application of AvG improves selectivity, accuracy, and reproducibility of quantitative DIA proteomics data. The resulting curated data is comparable to the current gold-standard of expert human curation but obtainable in a fraction of the time. Similar results as AvG can be obtained with mProphet-based tools using a very stringent FDR control (q-value<0.0001; Figure 5, Fig. S3 and Supp. Table 2). Most MS practitioners have come to expect a degree of statistical rigor in interpreting their data sets. The field has almost universally adopted a standard of FDR (or q-value) < 1%. However, there may be competing motivations for how this important statistical parameter is determined. Our analysis here suggests that the q-value reported by one tool is two orders of magnitude overly optimistic compared to another. This illustrates the need for careful inspection of actual results and tools, like AvG, that can be used to facilitate that process. AvG is meant to complement these tools to improve the quantification results. For very large datasets and to optimize the analysis time, Skyline can be used as a first pass for the analysis and AvG can be used to curate a subset of targets of interest (statistically changing between conditions for example). AvG's compatibility with a variety of acquisition modes (DIA or PRM), data sources (e.g. Orbitrap and TOF), upstream DIA identification tools (e.g. EncyclopeDIA, Specter, mProphet, etc.), and Skyline integration should make it attractive for broad utilization in the field.

Online Methods:

HEK293T cell digest

HEK293T cells were cultured in DMEM (Gibco; 11995) supplemented with 10% heat-inactivated FBS (Sigma; F4135). Once cells reached ~95% confluence they were harvested by scraping. Cells were pelleted at 1,000g for 2 min. The supernatant was then removed, and the cell pellet was frozen in liquid nitrogen. HEK293T cells were lysed by 5 min of exposure on ice to a lysis buffer (8 M urea, 75 mM NaCl, 50 mM Tris-HCl, pH 8.0, 1 mM EDTA, 2 μg/mL aprotinin (Sigma; A6103), 10 μg/mL leupeptin (Roche; 11017101001), 1 mM PMSF (Sigma; 78830)). The sample was centrifuged for 10 min at 20,000g. The protein concentration of HEK293T proteins was determined by BCA assay to be 4.3 μg/μl. 10 mg of protein was reduced (5 mM dithiothreitol, 45 min) and alkylated (10 mM iodoacetamide, 45 min). A Tris-HCl solution (50 mM, pH 8) was used to dilute the samples by a factor of 4 to reach a concentration of 2 M urea. A two-step digestion protocol was used to digest the lysate: Lys-C was used in a 1:50 enzyme-to-substrate ratio (Wako Chemicals; 129-02541) for 2 h at 30 °C, then the lysate was digested overnight at room temperature with trypsin in a 1:50 enzyme-to-substrate ratio (Promega; V511X) on a shaker. Formic acid (FA; 0.5% final concentration) was added to stop the digestion. The sample was split into four aliquots and loaded onto four 100-mg-capacity C18 Sep-Pak cartridges (Waters) for desalting. The four aliquots were eluted with 50% acetonitrile (ACN)/0.1% FA, pooled together, and vacuum-concentrated to dryness.

E. coli Digest

DH5α E. coli were grown in Luria broth at 37 °C overnight. Cells were pelleted by centrifugation, washed once with cold PBS, flash-frozen in liquid nitrogen, and stored at −80 °C until processing. For generation of the E. coli lysate digest, the cell pellet was thawed on ice. Lysozyme (Sigma) was added to the thawed pellet, and the mixture was placed on ice with periodic vortexing until viscous. The cells were resuspended in 8 M urea, 50 mM ammonium bicarbonate plus protease inhibitors (Roche), and the solution was sonicated with a probe sonicator for 2 min, 3 s on, 2 s off, until no longer viscous. After centrifugation at 15,000g for 30 min at 4 °C, protein concentration was measured by Bradford assay (Bio-Rad). Disulfide bridges were reduced (10 mM TCEP (tris(2-carboxyethyl)phosphine), Thermo) and alkylated (10 mM iodoacetamide, Thermo; 30 min; room temperature; in the dark). The lysate was diluted to 1.5 M urea with ammonium bicarbonate (50 mM) and digested overnight with a trypsin-to-substrate ratio of 1:100. The digest was desalted on C18 Sep-Pak cartridges (Waters). After vacuum centrifugation, dried peptides were resuspended to 1 mg/mL in 30% ACN/0.1% FA and stored at −80 °C.

Extended benchmarking DIA dataset

Mass spec-compatible Yeast digest was purchased from Promega. The E. coli, yeast and HEK293T digest were resuspended to 1 mg/mL in 30% ACN/0.1% FA. Four samples, with sample composition described in Fig. S2, were generated for the benchmarking dataset. A fifth sample that had the same quantity of protein of each proteome was also generated to obtain the spectral library. To generate the samples, the volume of each proteome digest corresponding to the desired protein quantity were mixed together, vacuum-concentrated to dryness, and resuspended to 500 ng/ul in 5% ACN/0.1%TFA. The 4 samples for the DIA benchmarking dataset were analyzed in DIA mode in 4 replicates. In total 16 runs were analyzed.

Calibration curve in HEK293T cell digest

The HEK293T digest was resuspended using 0.1% TFA, and a mixture containing 95 synthetic peptides at known individual concentrations was spiked into it to generate a five-point calibration curve. Each point was designed to contain 1 μg of HEK293T digest and 6.75, 13.5, 27, 54, or 108 ng of total amount of peptide on the column. The solution of heavy peptides used for this experiment was a mixture of 95 peptides that were combined in different concentrations in order to get a concentration-balanced mixture. Since not all these peptides have the same response factor, we have adjusted the concentration of each peptide to ensure its detection. The 5 samples were analyzed in DIA mode in triplicates. In total 15 runs were analyzed.

We generated a spectral library by first searching the DDA runs with Spectrum Mill v. B.06.01.201 using a FASTA containing the 95 synthetic peptide sequences and the UniProt human protein sequences (version dated 17 October 2014). Results were auto-validated to a false discovery rate of 1%, exported as a PepXML search result file, and loaded into Skyline to generate a spectral library in blib format.

The data analysis was performed using Skyline. Eleven unmodified peptides from the HEK293T background were chosen as standards to calibrate the retention times. The transition selection and peak boundaries was performed by AvG. The coefficients of variation were calculated on the triplicate measurement of the area of each analyte. The percent error was obtained by calculating the concentrations of each calibration point with the linear regression equation found for each peptide. The percent error was calculated as the absolute value of the difference between the measured and the expected concentration over the expected concentration.

P100 dataset:

The P100 samples were prepared exactly as described in Abelin et al. 2016[19]. In short, cells were cultivated, perturbed with 32 drugs each with 3 biological replicates. Cells were lysed for 30 min at 4 °C in lysis urea buffer (8 m urea; 75 mm NaCl, 50 mm Tris HCl pH 8.0, 1 mm EDTA, 2 μg/ml aprotinin (Sigma), 10 μg/ml leupeptin (Roche), 1 mm PMSF (Sigma), 10 mm NaF, Phosphatase Inhibitor Mixture 2 (1:100, Sigma), Phosphatase Inhibitor Mixture 3 (1:100, Sigma). Lysates were centrifuged at 15,000 × g for 15 min. Protein concentrations were measured (660 protein assay, Pierce). Reduction, Alkylation and digestion were performed on a Bravo robotic liquid handling platform (Agilent). Five hundred micrograms of protein were used for reduction in 100 mm DTT, alkylation in 200 mm IAA, dilution to 2 m urea in 50 mm Tris (pH 8.0), and digestion with 0.5 μg/μl (1:50) sequencing-grade modified trypsin (Promega) in 400μl volumes per sample at 37 °C overnight. Digestion was stopped by bringing the samples to a final concentration of 0.5% TFA. Acidified samples were loaded onto a 25 mg capacity C18 SepPak in a 96-well plate format (Waters) for desalting. Samples were eluted using 50% ACN/0.1% trifluoroacetic acid and vacuum concentrated to dryness. Then the samples were phospho-enriched on an AssayMAP Bravo robotic system (Agilent). The desalted samples were reconstituted in 80% ACN/0.1% TFA. Prior to sample loading the Agilent AssayMAP Fe-(III)-NTA cartridges were washed with water, stripped with 100 mm EDTA, and loaded with 100 mm FeCl3. Fe-(III)-NTA cartridges were primed with 1:1:1 ACN/methanol/0.01% acetic acid. Samples were loaded at 20 μl/min and flow-throughs were re-loaded onto cartridges eight additional times. Cartridges were washed with 80% ACN/0.1% TFA, and peptides were eluted with 500 mm K2HPO4 (pH7) at 5 μl/min. Eluates were vacuum concentrated to dryness, and subsequently desalted using AssayMAP RP-S cartridges according to the manufacturer's instructions. The 96 samples were analyzed in DIA mode.

For the data analysis we focused on the 95 phosphopeptides that constitute the P100 assay[2], which had isotopically labelled heavy peptide counterparts spiked into the sample. The dataset was analyzed using Skyline. For the manual validation: 3 to 5 transitions per peptide were extracted. The transitions were the ones chosen for the P100 assay. The data were visually inspected, interfered transitions were removed, and peak boundaries manually corrected by an expert in the field. For the unoptimized dataset: the 5 most intense transitions from the spectral library were chosen and the peak boundaries were defined by Skyline. For the optimized version, all possible transitions (b,y above b4 and y4) were extracted and the transition selection and peak boundaries were performed by AvG.

Evaluation of AvG using the LFQBench dataset

The LFQBench data[11] was downloaded from ProteomeXchange (dataset PXD002952). We focused on the HYE110 dataset analyzed with 64 variable isolation width windows on an AB Sciex Triple-TOF 6600. The raw WIFF files were imported into Skyline Daily (v.4.1.1.18118). We used the spectral library provided by the study's authors (ecolihumanyeast_concat_mayu_IRR_cons_openswath_64w_var_curated.csv) that consisted of precursors with only 6 annotated fragment ions in CSV format compatible with OpenSWATH. To reproduce the results published in the LFQBench paper, we first extracted the signals of all peptides present in the reports present in the same ProteomeXchange dataset, for Skyline we used the file called "Skyline_HYE110_TTOF6600_64var_160305_fix1603.tsv, for OpenSWATH+TRIC we used the report called "E1603141345_feature_alignment.tsv". We applied the Skyline default parameters for the signal extraction (30000 resolving power for MS/MS, chromatograms were extracted 5 minutes around the predicted RT, all 6 six ions were extracted for each precursor). The integration boundaries were changed to match the values listed in the reports. Only features having a q-value lower than 0.01 were integrated. Features having a q-value>0.01 were not integrated and thus were missing values. For *Avant-garde*, decoy peptides were added to the skyline files and all features were integrated. For peptides having a q-value <0.01, the peak boundaries were changed to the values reported in the LFQBench reports. After running *Avant-garde*, the FDR was controlled using Percolator 3.0 at a false discovery rate lower than 1%. Percolator features included the PSS score, the SLS score, the mass error score, the MPRA score and the AvG score. The peak boundaries and the selected transitions were imported back into the corresponding Skyline files. Following the metrics established in the LFQBench paper, an "identified peptide" is defined as a peptide observed in at least one MS run, a "valid quantifiable ratio" is defined as a peptide that is observed at least once in at least two conditions.

Evaluation of AvG using our extended benchmarking DIA dataset

To build the spectral library a sample constituted of an equal protein amount of the three proteomes was analyzed. This sample was analyzed using gas phase fractionation using 12 MS runs[29]. Each injection was analyzed with narrow-window DIA where the instrument cycles through 25 2-m/z DIA windows and focuses only on 50 m/z at the MS1 level per MS run. Twelve injections are necessary to systematically and comprehensively monitor the 400 to 900 m/z range with narrow windows. The data were searched with SpectrumMill against a merged database of the human, E. coli and yeast database. The search was done with large tolerance at the MS1 level (1 m/z) and small tolerance at the MS2 level 10ppm. The search results were validated using Percolator 3.0 at a false discovery rate lower than 1%. The list of validated spectra was imported into Spectrum Mill, a PepXML search result file was generated and loaded into Skyline to generate a spectral library in blib format.

Eighteen thousand peptides (6,000 of each proteome) were randomly chosen from the spectral library and extracted using Skyline. A subset of 15 unmodified human peptides whose retention times were distributed across the chromatographic gradient were chosen as retention time standard peptides.

For the unoptimized dataset the chromatograms were extracted using Skyline and validated using Skyline's implementation of mProphet. Raw files were converted to MzML and demultiplexed using MSConvert. Skyline was set with the following parameters: Precursor charges:2, 3, 4; Ion charges:1,2; Ion types:y,b; Ion match tolerance 0.05 m/z; Product mass analyzer: Centroided; MS/MS mass accuracy: 20ppm: chromatograms were extracted 5 minutes around the predicted RT. The 10 most intense transitions from the spectral library were extracted and a mProphet model was trained using the corresponding 18 thousand decoy peptides. Only the peaks with a q-value lower than 0.01 were reintegrated.

For the optimized dataset the chromatograms were extracted using Skyline. The 10 most intense transitions for each peptide were extracted and the peak boundaries were determined by Skyline, which was using our 15 RT standard peptides for retention time prediction. The data was refined a posteriori by AvG to select at least 4 transitions per peptide and correct the peak boundaries. The data was scored and filtered to less than 1% FDR (spectral library similarity >0.7, mass error score>0.7, peak shape similarity score >0.85, MPRA score> 0.9 and AvG Score>0.1). An "identified peptide" is defined as a peptide observed in at least one MS run, a "valid quantifiable ratio" is defined as a peptide that is observed at least once in at least two conditions.

In order to simulate "real life" biological samples, we downsampled the dataset by randomly selecting a 3,000 human peptides, and then randomly selecting a lower number of peptides for yeast and E. coli that represented 5% percent of the total number of human peptides. For each downsampled dataset, a two-tailed two-sample moderated t-test from the limma R-package were calculated for the quadruplicate log2 transformed areas[23]. The p-values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg method. The peptides were classified as significantly differentially expressed (positive hits) if their adjusted p-value was lower than 0.05 and their absolute fold change was higher than 2 times the standard deviation of the fold changes for the human peptides. Peptides were classified as not differentially expressed (negative hits) if either the adjusted p-value was lower than 0.05 or the absolute fold change was lower than 2 times the standard deviation of the fold change for the human peptides. Peptides were classified as accurate if their observed ratio was in the 80-120% range of the expected value. The recall for the detection for differentially expressed peptides was estimated here by the number of true positive hits over the total number of differentially expressed peptides, i.e. yeast and E. coli peptides (Recall= TP/(Total number of yeast and E. coli peptides in the downsampled dataset)). The false positive rate (FPR) was determined by calculating the number of false positive hits over the total number of proteins found to be differentially expressed (FP/(TP + FP)). The process described above (downsampling, statistical testing and performance evaluation) was iterated a thousand times to reduce the effect of outlier in the evaluation of the performances of the unoptimized and the optimized dataset by AvG.

LC-MS method

Overlap DIA method on Q-Exactive HF+ (used for P100 dataset and calibration curve spiked in HEK 293T digest): The P100 dataset and the calibration curve spiked in HEK 293T digest were analyzed with an Orbitrap Q-Exactive HF Plus (Thermo Fisher Scientific) mass spectrometer coupled to a nanoflow Proxeon EASY-nLC 1000 UHPLC system (Thermo Fisher Scientific). The mass spectrometer was used in positive mode and was equipped with a nanoflow ionization source (James A. Hill Instrument Services, Arlington, MA); the spray voltage was set at 2.00 kV. The LC system, the column, and the electrospray voltage source (platinum wire) were connected via a stainless steel cross (360 μm; IDEX Health & Science; UH-906x). The column was heated to 50 °C. A volume of 3 μl was injected onto an in-house packed 20 cm × 75 μm diameter C18 silica picofrit capillary column (1.9-μm ReproSil-Pur C18-AQ beads, Dr. Maisch GmbH, r119.aq; Picofrit 10-μm tip opening, New Objective, PF360-75-10-N-5). The mobile phase had a flow rate of 250 nL/min and consisted of 3% ACN/0.1% FA (solvent A) and 90% ACN/0.1% FA (solvent B). The column was conditioned before each sample injection. Peptides were separated using the following LC gradient: 0–3% B in 3 min, 5–40% B in 50 min, 40–90% B in 1 min, stay at 90% B for 5.5 min, and 90–50% B in 30 s. DDA and DIA data were acquired on the same instrument. For the MS1 scans, the resolution was set at 60,000 at 200 m/z and the automatic gain control (AGC) target was 3×10^6 with a maximum inject fill time of 20 ms. For DDA, MS2 scans on the top 12 peaks doubly charged and above were acquired at a resolution of 15,000, AGC target of 5×10^4 with maximum inject fill time of 50 ms. Isolation widths were set to 1.5 m/z with a 0.3 m/z offset. The normalized collision energy (NCE) was set to 27 and dynamic exclusion was set to 10 s. For DIA, an overlap DIA method was used with 56 × 22 m/z isolation windows covering the 400–1,000 m/z range. In this method, the isolation windows in two consecutive cycles have an offset of 11 m/z. The default charge state was 4, the resolution was 30,000 at 200 m/z, the AGC target was 1×10^6, the maximum inject fill time was 50 ms, the loop count was 27 and the NCE was set to 27.

Overlap DIA and narrow-window DIA method on Q-Exactive HFX (used for the extended benchmarking DIA dataset): The extended benchmarking DIA dataset was analyzed with an Orbitrap Q-Exactive HFX (Thermo Fisher Scientific) mass spectrometer coupled to a nanoflow Proxeon EASY-nLC 1200 UHPLC system (Thermo Fisher Scientific). The mass spectrometer was used in positive mode and was equipped with a nanoflow ionization source (James A. Hill Instrument Services, Arlington, MA); the spray voltage was set at 2.00 kV. The LC system, the column, and the electrospray voltage source (platinum wire) were connected via a stainless steel cross (360 μm; IDEX Health & Science; UH-906x). The column was heated to 50 °C. A volume equivalent to 500ng of protein on column was injected onto an in-house packed 20 cm × 75 μm diameter C18 silica picofrit capillary column (1.9-μm ReproSil-Pur C18-AQ beads, Dr. Maisch GmbH, r119.aq; Picofrit 10-μm tip opening, New Objective, PF360-75-10-N-5). The mobile phase had a flow rate of 200 nL/min and consisted of 3% ACN/0.1% FA (solvent A) and 90% ACN/0.1% FA (solvent B). The column was conditioned before each sample injection. Peptides were separated using the following LC gradient: 2–6% B in 1 min, 6–30% B in 74.5 min, 30–60% B in 7.5min, 60–90% B in 1 min, stay at 90% B for 5 min, and 90–50% B in 2min.

For the MS1 scans, the resolution was set at 60,000 at 200 m/z and the automatic gain control (AGC) target was 3 × 10^6 with a maximum inject fill time of 20 ms.

For DIA, an overlap DIA method was used with 68× 18 m/z isolation windows covering the 400–1,000 m/z range. In this method, the isolation windows in two consecutive cycles have an offset of 9 m/z. The default charge state was 4, the resolution was 15,000 at 200 m/z, the AGC target was 1 × 10^6, the maximum inject fill time was 18ms, the loop count was 34 and the collision energy was fixed for each window using the following equation: CE=0.0459 m/z - 6×10^{-5}. The m/z value corresponded to the center of each DIA window.

For narrow-window DIA, we used 12 different instrument methods. Each one used 25× 2 m/z non-overlapped isolation windows covering 50 m/z range. Together the 12 MS runs covered the 400-1000 m/z range. The default charge state was 4, the resolution was 15,000 at 200 m/z, the AGC target was 1 × 10^6, the maximum inject fill time was 25ms, the loop count was 25 and the collision energy was fixed for each window using the equation described in the paragraph above.

Data analysis setup

For each dataset, the corresponding spectral library and the list of peptides of interest were loaded into Skyline. The spectral library is used to select the most intense transitions per peptide, ideally selecting a large number (8-15 transitions per peptide) to allow AvG to optimize the transition selection a posteriori based on the signals of the whole dataset. The raw data was then imported into Skyline. A report containing the chromatograms, mass errors, spectral library information and total areas was generated. This report is the input to AvG, and the outputs are a file containing the chosen transitions for each peptide, a file containing the peak boundaries for each peptide in each MS acquisition and a report containing the subscores and AvG score of all peptides in all MS acquisition. These reports were uploaded into Skyline to adjust the peak boundaries and define the transitions to be used for the quantification. A complete tutorial for the use of AvG and a demonstration dataset are provided in *Avant-garde*'s GitHub page (https://github.com/SebVaca/Avant_garde).

Computational parallelization and benchmarking

AvG can be parallelized as the analysis of each analyte is completely independent from the others. *Avant-Garde* is designed to determine the number of available cores in the computer (N) and run using N-1 cores. Each core will analyze one analyte at the time, so in order to increase the performance of AvG it is recommended to increase the number of cores available.

For example, the calibration curve dataset (96 peptides in 15 runs) took 6 minutes to run. The P100 dataset (96 peptides in 96 runs) ran in 20 minutes, on a system with similar configuration as the following: Windows 7, Intel Core i7-3770 CPU @ 3.40Ghz, 8-core, memory (RAM) 24Gb, 64-bit operating system.

The LFQBench and the Benchmarking datasets in 1-2 hours using the Broad Institute's dynamic parallel computing platform (Univa Grid Engine, Linux RedHat6, Memory (Ram) 4Gb per core).

Calculation of AvG subscores

Peak shape similarity (PSS) score: For a given peptide and a given subset of transitions, the similarity score shows the resemblance of the peak shape of a given transition to all the others transitions in the subset. This score is calculated by first normalizing each transition to the maximum intensity value within the integration boundaries. By doing this, the peak shape comparison is not dependent on the intensity of each transition but only on its shape. A mean peak shape profile is then created by calculating the mean of the normalized intensities for all transitions at each time point. The mean was chosen in order to better reflect the presence of interferences would be reflected and avoid smoothing them away. This will amplify the difference of peak shapes when a interference is present.

The similarity is determined by calculating the mean of all the dot products, calculated for each transition, of the normalized intensities of each transition and the mean profile.

To make sure that the similarity score is not influenced by just a single highly-scored transition, a second mean is calculated after removing the transition with the highest dot product value. This penalizes even further the set of transitions where an interference is present, ensuring that only the set of transitions that have similar peak shapes obtain a high score. The mean of these two dot product values corresponds to the peak shape similarity score.

For a given peptide using a given set of transitions in a given MS run:

$$n = total\ number\ of\ transitions\ in\ the\ set$$
$$u_i = vector\ of\ the\ intensities\ of\ a\ transition\ i\ in\ elution\ time\ order$$
$$v = vector\ of\ the\ intensities\ of\ the\ mean\ peak\ shape\ profile\ in\ elution\ time\ order$$
$$k = Index\ of\ the\ transition\ for\ which\ the\ normalized\ dot\ product\ is\ the\ highest$$
$$K = set\ of\ all\ indices\ from 1 to\ n\ except\ k$$
$$Peak\ Shape\ Similarity\ score = \frac{\sum_{i=1}^{n} \frac{u_i \cdot v}{\|u_i \cdot v\|}}{n} \times \frac{\sum_{i \epsilon K} \frac{u_i \cdot v}{\|u_i \cdot v\|}}{n-1}$$

Mass error score:

The mass error score is a mean of the mass error measured at each chromatographic point between the integration boundaries weighted by the intensity. The user defines a tolerance threshold in ppm below which the score is equal to 1 and a cut-off threshold above which the score is equal to 0 (Fig. S5A). The mass error score is defined as follows:

For a given peptide using a given set of transitions in a given MS run:

$$
\begin{aligned}
&m_{measured} = absolute\ value\ of\ the\ mass\ error\ measured\ at\ a\ given\ chromatographic\ point\ (in\ ppm)\\
&m_{tol} = mass\ error\ tolerance\ (in\ ppm)\\
&m_{cutoff} = mass\ error\ cutoff\ (in\ ppm)\\
&mass\ error\ score = 1\ if\, m_{measured} \le m_{tol}
\end{aligned}
$$

$$
\begin{aligned}
&mass\ error\ score = \frac{m_{cutoff}}{(m_{cutoff} - m_{tol})} + \frac{m_{measured}}{(m_{tol} - m_{cutoff})} if\, m_{tol} < m_{measured} \le m_{cutoff}\\
&mass\ error\ score = 0\ if\, m_{measured} > m_{cutoff}
\end{aligned}
$$

Mean profile of relative areas (MPRA) score

For a given peptide and a given subset of transitions, the area under the curve of each transition is normalized to the sum of the areas of all transitions. A mean profile is obtained by calculating the mean for each transition of all normalized areas across all runs. The mean profile is then normalized. For each peptide in each run, the MPRA is the dot product between the vector containing the normalized areas for each transition and the mean profile calculated previously. This score reflects how similar to each other are the relative peak areas across the entire dataset. The objective of the MPRA is to detect interferences. By optimizing it we can maximize the similarity between the data used to quantify a peptide. By doing this the resulting choice of transitions will be tailored to the dataset to be analyzed, producing the lowest number of interferences in the entire dataset, and providing the solution where the remaining noise has the least impact on the quantification.

For a given peptide using a given set of transitions in a given MS run:

$$
\begin{aligned}
&n = total\ number\ of\ transitions\ in\ the\ set\\
&r = total\ number\ of\ MS\ runs\ in\ the\ dataset\\
&a_i^j = area\ under\ the\ curve\ between\ the\ integration\ boundaries\ for\ transition\ i\ in\ run\ j\\
&b_i^j = normalized\ area\ for\ transition\ i\ in\ run\ j\\
&b_i^j = \frac{a_i^j}{\sum_{i=1}^{n} a_i^j}
\end{aligned}
$$

$$
\begin{aligned}
&c_i = mean\ area\ for\ transition\ i\ across\ all\ MS\ runs\ in\ the\ dataset\\
&c_i = \frac{\sum_{j=1}^{r} a_i^j}{r}
\end{aligned}
$$

$$
\begin{aligned}
&c_i' = normalized\ mean\ area\ for\ transition\ i\ across\ all\ MS\ runs\ in\ the\ dataset\\
&c_i' = \frac{c_i}{\sum_{i=1}^{n} c_i}
\end{aligned}
$$

$$u^j = vector\ of\ all\ b_i^j\ for\ MS\ run\ j$$
$$v = vector\ of\ all c_i'$$

$$MPRA^j = MPRA\ score\ for\ MS\ run\ j$$
$$MPRA^j = u^j \cdot v$$

$$MPRA = MPRA\ score\ for\ the\ entire\ dataset$$
$$MPRA = \frac{\sum_{j=1}^{r} MPRA^j}{r}$$

Spectral library similarity (SLS) score

For a given peptide and a given subset of transitions, the spectral library similarity is calculated based on the dot product between the intensity of each transition in the spectral library and the areas integrated from the DIA signals. Then the dot product value undergoes a transformation (Fig. S5B). The aim of this transformation is 1) to set a cut-off value defined by the user under which the SLS score will be low. 2) To examine how similar the measured DIA signals are to the signals present in the spectral library. However, due to instrumental variations over time, changes in instrumental calibration and the use of different acquisition methods (DDA and DIA) the fragmentation patterns observed in the spectral library and the ones obtained when analyzing the samples in DIA might not be exactly the same. In order to be more tolerant to these small changes, another threshold is determined above which the spectral library similarity score will be equal to 1.,e.g. two sets of transitions that have a high dot product will both obtain a spectral library similarity score equal to one. This ensures that both sets of transitions are scored equally instead of biasing one of the two with the dot product that does not take into account the variation of the fragmentation patterns.

Intensity and intensity product chromatographic-scores

For a given peptide and a given run, the intensity chromatographic score is equal to the sum of the intensities of all transitions at a given time point normalized by the maximum value of the summed intensity across the entire chromatogram. To calculate the intensity product chromatographic score, first 1 is added to the areas of each transition to avoid missing values or values equal to zero. For each time point the areas are multiplied together, log2 transformed, and normalized by the maximum value across the entire chromatogram.

Potential peak score

In order to define a signal as a potential peptide chromatographic peak, at least 3 transitions should have an intensity higher than the level of noise for at least N number of consecutive points. N is defined by the user and needs to be adapted for each instrument. The level of noise was estimated in the following way: for a given peptide, the level of noise was set as the median of the all the lowest intensity values among all transition at each time point, plus

2 times the standard deviation of these values. This parameter is necessary when analyzing data from Q-TOF instruments as the background noise is higher in these datasets. For Overlap DIA Orbitrap data the noise if very low and the level of noise was set to zero. If the criteria described here is met the Potential peak score is equal to 1, and 0 if not.

Score combination:

Avant-garde is very conservative and uses a novel ensemble-driven scoring strategy. The combined score from *Avant-garde* is not calculated by adding subscores like other tools. We avoided adding subscores because it can lead to a high number of false positives due to the fact that the combined score can be mainly influenced by one single very good-scoring subscore. *Avant-garde* changes the way peak groups are scored. The main idea behind *Avant-garde* is to reduce noise to the minimum, thus obtaining very high-quality signals. We manage this by penalizing peptides having any low-scoring metric. All *Avant-garde* scores have values between 0 and 1. To combine them each metric is weighted by an exponent and multiplied together. This means that if a peptide does not score well with any given metric the combined score will be severely penalized. Three different combinations are used in AvG for different purposes. The combined scores are called “AvG chromatographic score” for the refinement of peak integration boundaries, “AvG fitness score” for the refinement of the transition selection and “AvG score” for the scoring of peaks.

For transition refinement:

For each step (or generation) the genetic algorithm selects a population of randomly chosen subset of transitions, scores each one with a fitness function and selects the best-scoring solutions as the starting point for the next generation.

The fitness function used by the genetic algorithm was defined as follows:

If a randomly selected set of transitions has a number of transitions below a minimal number defined by the user then Fitness Score =0.

If for a randomly selected set of transitions at least n transitions are not among the top N most intense transitions in the spectral library then Fitness Score =0. Where n and N are user-defined values.

Otherwise:

$$AVG\ Fitness\ Score = (1 - \alpha - \beta) \times (MPRA + PSS) + \alpha \times IntensityScore + \beta \times MassErrorScore + 0.05 \times SLS$$

Where α and β are user-defined values (default to 0.05).

For the fitness function, more weight is given to the PSS and the MPRA score, given that they are the most sensitive metrics to detect interferences. This enables to confidently identify interferences and remove them in the entire dataset. The intensity, mass errors and library score have a smaller weight on the fitness score. They are used to decide between possible solutions with similar peak shape similarity and MPRA scores. When two solutions are possible the fitness function is designed to choose the one providing the highest intensity, lower mass deviations and matching the best the spectral library.

Often interferences can overshadow the signal from the analyte of interest, especially for low abundant peptides. Reducing the influence of interferences with high intensity on the fitness score by giving a small weight to the intensity score is extremely important. Additionally, the algorithm ensures that at least n transitions are among the N most intense fragments in the spectral library (n and N are user-defined values) guaranteeing that the solution will have intense signals and ensures that the transition selection step will not have a negative impact on the sensitivity of the quantification.

For peak boundaries refinement: In order to refine a peptide identification and its peak boundaries, *Avant-garde* uses chromatographic scores that are combined to form the *AvG chromatographic score*. The maximum value of the *AvG chromatographic score* corresponds to the peaks’ retention time. The boundaries correspond to the retention time where the intensity score is at 4% of the maximum value of the summed intensities of all transitions in the subset.

In order to robustly combine the subscores and estimate the trend of the scores without being too susceptible to rapid changes and isolated anomalies, the scores were first transformed using a moving median. A moving median is a function which replaces each data value with the median of neighboring values. The combined score, termed *AvG chromatographic score*, is calculated in the following way for each time point:

$$AVG\ chormatographic\ score = moving.average\left(SLS^3 \times MPRA^3\right) \times intensity.Product.Score^3 \times moving.average(MassErrorScore) \times PotentialPeak$$

1. For peak scoring and FDR calculation:

After the transition refinement and peak boundaries refinement steps, each peak group is scored in order to filter the data and control the FDR. The data being scored here are the chromatogram traces between the new integration boundaries. For each peptide in each run, the AvG score is calculated as:

$$AVG\ score = PSS^{9.5} \times SLS^{4.5} \times \mathrm{MassErrorScore}^{2.5} \times MPRA^{0.5}$$

The exponents were found using the HEK293T dataset where one thousand peptides and their corresponding decoys were extracted and the data was curated by AvG. We then determined the set of exponents (multiples of 0.5 between 1 and 10) that increased the separation between target and decoy peptides. These exponents were empirically determined and fixed to this value so we can compare different datasets to each other. The intensity score was not included in the score at this stage in order to avoid penalizing low-abundant peptides and give a strong influence to high-intensity interfered transitions.

AvG score and FDR estimation

In order to define the AvG score that is used to estimate the FDR we used two datasets that were acquired on two different instruments. The first one was acquired on a Q-Exactive HF (Thermo Fisher Scientific) plus instrument. For this dataset, one thousand peptides, and their corresponding shuffled-sequence decoy peptides, from the HEK293T digest data were randomly selected in 15 MS runs (Fig. S6 and S7). The second dataset was acquired on a Q-TOF instrument (Triple-TOF 6600, Sciex). We used a subset of the LFQBench data that contained 4000 targets and their corresponding 4000 decoy peptides in 6 runs (Fig. S8 and S9).

Avant-garde was used to curate the data and score each peptide in each of the 15 MS runs.

The AvG score was defined as:

$$\text{AvG score} = \text{PSS}^{a1} \times \text{SLS}^{a2} \times \text{MassErrorScore}^{a3} \times \text{MPRA}^{a4}$$

An optimization algorithm was used to find the optimal values of the exponents (a1 to a4) of each subscore that enabled obtaining the largest number of validated measurements. This algorithm chose a random set of 4 numbers that corresponded to the exponents in the equation. The random values were a multiple of 0.5 between 1 and 10. The set of exponents that provided the largest number of validated measurements at an FDR of 1% over 1000 iterations were chosen.

The coefficients found to calculate the AvG score were the following:

$$\text{AvG score} = \text{PSS}^{9.5} \times \text{SLS}^{4.5} \times \text{MassErrorScore}^{2.5} \times \text{MPRA}^{0.5}$$

The results obtained by this heuristic approach were compared to the results obtained with a standard linear discriminant analysis classifier. To obtain comparable results we used the logarithmic transformation of the equation above:

$$\text{K} = \text{a1} \times \log10(\text{PSS}) + \text{a2} \times \log10(\text{SLS}) + \text{a3} \times \log10(\text{MassErrorScore}) + \text{a4} \times \log10(\text{MPRA})$$

The LDA was used to define the set of coefficients (a1 to a4) that enabled obtaining the best separation between targets and decoys. To obtain an equivalent result as the AvG score, the LDA score was then defined as:

$$\text{LDA score} = 10\text{K}$$

However, using the LDA score to reach an FDR of 1 % produced a lower number of validated measurements than the one found using the AvG score. This was expected as the AvG score was designed to optimize the number of validated measurements for an FDR below 1%. The LDA is affected by very low scoring decoy signals that are most likely background noise. In order to improve the results of the LDA, we performed a second round of analysis on a subset of the data. We removed all measurements having a LDA score lower than the 25% percentile of the LDA score for decoy peptides. By removing the lower scoring signals, the 2nd round LDA was better at separating target and decoy signals. This is due to the fact that, in the second round, the characteristics of the decoy population is more similar to the target signals as the background noise was removed (Fig. S10). The LDA coefficients were determined using the reduced dataset but the FDR was determined using the complete dataset. In these two datasets, the AvG score shows similar performance as two rounds of linear discriminant analysis.

In this manuscript, we decided to use the AvG score in all the results presented here. This was done to obtain scores that can be compared from dataset to dataset. Additionally, AvG does not produce any output for signals that have a lower number of transitions than the minimal number allowed or that do not have any potential peak. The latter is often the case for decoy peptides. For data acquired in centroid mode, the background noise level is drastically reduced and consequently the number of reported decoys is lower than for data acquired in profile mode. Only signals that have the characteristics of a peptide signal obtain an AvG score. Hence this can produce datasets with a very small population of decoys having an AvG score. In this case it is not possible to perform an LDA classification on these datasets. This does not mean that the decoys were ignored. It means that the quality of the decoy is so low that they do not obtain an AvG score. By fixing the exponents used to calculate the AvG score, we were able to estimate the FDR in all datasets without the need to perform an LDA for each one.

In this manuscript the data acquired on a Q-Exactive series instrument using an overlapped DIA method was filtered using the AvG score and using the following thresholds: SLS >0.7, mass error score>0.7, PSS >0.85, MPRA > 0.9, AvG score >0.1. This guaranteed that all the reported signals had at least a minimal level of signal quality. We then calculated the FDR to verify that it was below 1%. For Q-TOF stepwise DIA data, the data was filtered using the following thresholds: SLS >0.7, mass error score>0.7, PSS >0.85, MPRA > 0.9, AvG score >0.61.

Statistical analysis

All statistical analyses were carried out using R statistical software[30] (v.3.4.3). DIA data analysis was done using Skyline (v.4.2).

Reporting Summary

Further information on research design is available in the Life Sciences Reporting Summary linked to this article.

Data availability:

The original mass spectra have been deposited in the public proteomics repository MassIVE and are accessible at ftp://MSV000085540@massive.ucsd.edu.

Code availability:

Avant-garde is an open-source software tool available as an R package and as an Skyline External tool at https://github.com/SebVaca/Avant_garde. *Avant-garde* can be directly downloaded from the tool Store interface within Skyline or from the Skyline tool Store at https://skyline.ms/tool-AvG.url.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Nicholas Pythoud, Joanna Bons, Alexandre Burel and Christine Carapito for beta-testing the software. This work was funded by U54 HG008097 to JDJ. This work was also supported in part by grants from the National Cancer Institute (NCI) Clinical Proteomic Tumor Analysis Consortium grants NIH/NCI U24-CA210986 and NIH/NCI U01 CA214125 (to SAC) and NIH/NCI U24-CA210979 to D.R. Mani, principal computational scientist in the Proteomics Platform at the Broad Institute of MIT and Harvard.

References

1. Panchaud A et al. Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. Anal. Chem. 81, 6481–6488 (2009). [PubMed: 19572557]
2. Gillet LC et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol. Cell. Proteomics 11, O111.016717 (2012).
3. Egertson JD, MacLean B, Johnson R, Xuan Y & MacCoss MJ Multiplexed peptide analysis using data-independent acquisition and Skyline. Nat. Protoc. 10, 887–903 (2015). [PubMed: 25996789]
4. Chapman JD, Goodlett DR & Masselon CD Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. Mass Spectrom. Rev. 33, 452–470 (2014). [PubMed: 24281846]
5. Purvine S, Eppel J-T, Yi EC & Goodlett DR Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. Proteomics 3, 847–850 (2003). [PubMed: 12833507]
6. Silva JC et al. Quantitative proteomic analysis by accurate mass retention time pairs. Anal. Chem. 77, 2187–2200 (2005). [PubMed: 15801753]
7. Silva JC et al. Simultaneous qualitative and quantitative analysis of the Escherichia coli proteome: a sweet tale. Mol. Cell. Proteomics 5, 589–607 (2006). [PubMed: 16399765]
8. Prakash A et al. Hybrid data acquisition and processing strategies with increased throughput and selectivity: pSMART analysis for global qualitative and quantitative analysis. J. Proteome Res. 13, 5415–5430 (2014). [PubMed: 25244318]
9. Geiger T, Cox J & Mann M Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. Mol. Cell. Proteomics 9, 2252–2261 (2010). [PubMed: 20610777]
10. Bilbao A et al. Processing strategies and software solutions for data-independent acquisition in mass spectrometry. Proteomics 15, 964–980 (2015). [PubMed: 25430050]
11. Navarro P et al. A multicenter study benchmarks software tools for label-free proteome quantification. Nat. Biotechnol. 34, 1130–1136 (2016). [PubMed: 27701404]
12. Reiter L et al. mProphet: automated data processing and statistical validation for large-scale SRM experiments. Nat. Methods 8, 430–435 (2011). [PubMed: 21423193]
13. Röst HL et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat. Biotechnol. 32, 219–223 (2014). [PubMed: 24727770]
14. Jaffe JD, Feeney CM, Patel J, Lu X & Mani DR Transitioning from Targeted to Comprehensive Mass Spectrometry Using Genetic Algorithms. J. Am. Soc. Mass Spectrom. 27, 1745–1751 (2016). [PubMed: 27562500]
15. Tsou C-C et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat. Methods 12, 258–64, 7 p following 264 (2015). [PubMed: 25599550]
16. Searle BC et al. Comprehensive peptide quantification for data independent acquisition mass spectrometry using chromatogram libraries. bioRxiv 277822 (2018) doi:10.1101/277822.

17. Peckner R et al. Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics. Nat. Methods 15, 371–378 (2018). [PubMed: 29608554]
18. MacLean B et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 26, 966–968 (2010). [PubMed: 20147306]
19. Abelin JG, Patel J, Lu X, Feeney CM & Fagbami L Reduced-representation phosphosignatures measured by quantitative targeted MS capture cellular states and enable large-scale comparison of drug-induced …. Molecular & Cellular (2016).
20. Rosenberger G et al. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. Nat. Methods 14, 921–927 (2017). [PubMed: 28825704]
21. Röst HL et al. TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. Nat. Methods 13, 777–783 (2016). [PubMed: 27479329]
22. Ramus C et al. Benchmarking quantitative label-free LC-MS data processing workflows using a complex spiked proteomic standard dataset. J. Proteomics 132, 51–62 (2016). [PubMed: 26585461]
23. Ritchie ME et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47 (2015). [PubMed: 25605792]
24. Röst H, Malmström L & Aebersold R A computational tool to detect and avoid redundancy in selected reaction monitoring. Mol. Cell. Proteomics 11, 540–549 (2012). [PubMed: 22535207]
25. Keller A, Bader SL, Shteynberg D, Hood L & Moritz RL Automated Validation of Results and Removal of Fragment Ion Interferences in Targeted Analysis of Data-independent Acquisition Mass Spectrometry (MS) using SWATHProphet. Mol. Cell. Proteomics 14, 1411–1418 (2015). [PubMed: 25713123]
26. Choi M et al. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics 30, 2524–2526 (2014). [PubMed: 24794931]
27. Teo G et al. mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. J. Proteomics 129, 108–120 (2015). [PubMed: 26381204]
28. Tsai T-H et al. Selection of features with consistent profiles improves relative protein quantification in mass spectrometry experiments. Mol. Cell. Proteomics 19, 944–959 (2020). [PubMed: 32234965]
29. Ting YS et al. PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. Nat. Methods 14, 903–908 (2017). [PubMed: 28783153]
30. Team, R. C. R: A language and environment for statistical com-puting. R Foundation for Statistical Computing, Vienna, Austria (2017).

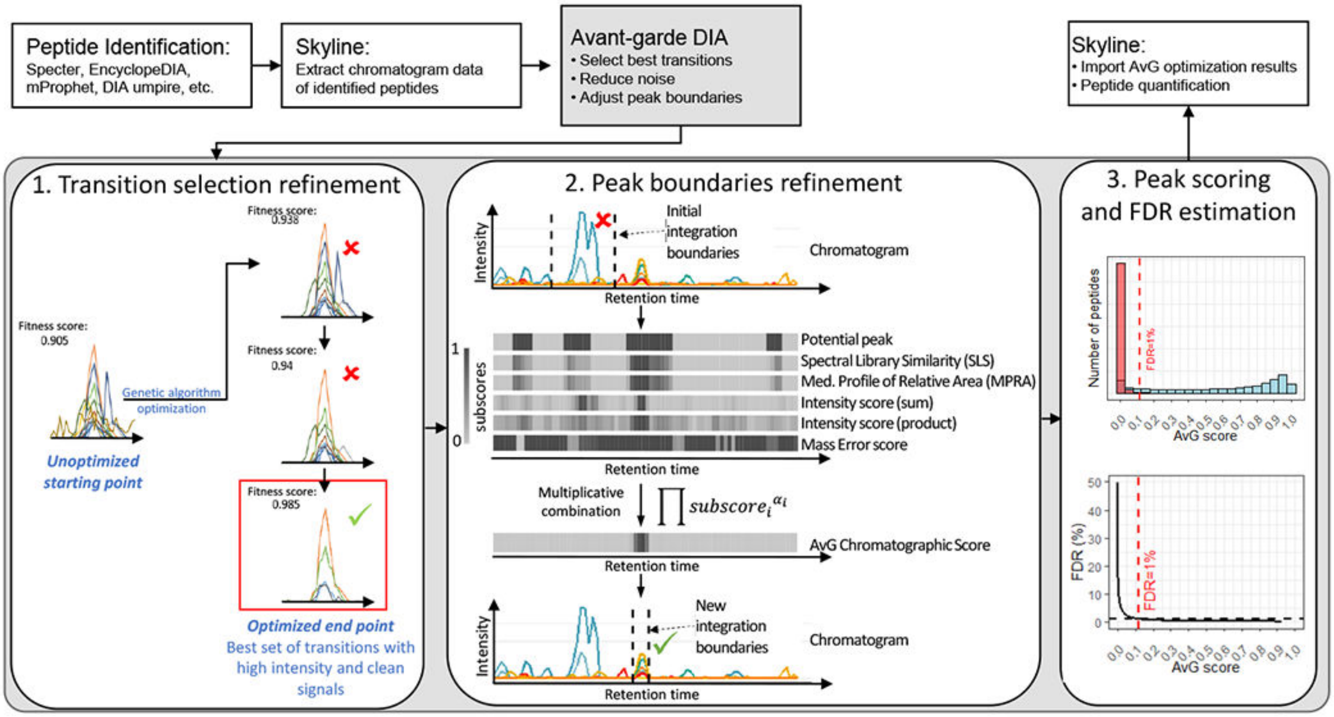
Peptide Identification:
Specter, EncyclopeDIA, mProphet, DIA umpire, etc.
Skyline:
Extract chromatogram data of identified peptides
Avant-garde DIA
• Select best transitions
• Reduce noise
• Adjust peak boundaries
Skyline:
• Import AvG optimization results
• Peptide quantification
1. Transition selection refinement
Fitness score: 0.905
Genetic algorithm optimization
Unoptimized starting point
Fitness score: 0.938
Fitness score: 0.94
Fitness score: 0.985
Optimized end point
Best set of transitions with high intensity and clean signals
2. Peak boundaries refinement
Intensity
Initial integration boundaries
Chromatogram
Retention time
1
0
subscores
Potential peak
Spectral Library Similarity (SLS)
Med. Profile of Relative Area (MPRA)
Intensity score (sum)
Intensity score (product)
Mass Error score
Retention time
Multiplicative combination
$\prod subscore_i^{\alpha_i}$
AvG Chromatographic Score
Retention time
Intensity
New integration boundaries
Chromatogram
Retention time
3. Peak scoring and FDR estimation
Number of peptides
FDR=1%
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
AvG score
FDR (%)
50 40 30 20 10 0
FDR=1%
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
AvG score

Figure 1: *Avant-garde*'s role in data analysis and modular scheme.

Top: AvG is employed downstream of independent DIA identification engines, and depends on Skyline to extract chromatogram data of detected peptides. The output of AvG (suitable transitions, chromatographic boundaries, and scoring metrics) is reimported to Skyline to produce curated quantitative data. *Bottom*: AvG is composed of three modules. Module 1 curates transitions to reduce noise and remove interference using a genetic algorithm, assigning a final quality metric to the selected set (*AvG fitness score*). Module 2 refines peak integration boundaries. AvG calculates chromatographic subscores at each time point in the raw data, and combines them as a weighted product (*AvG chromatographic score*). The maximum value of this score corresponds to the most likely retention time of the analyte. Module 3 scores peaks (*AvG score*), filters the data and estimates the FDR for quantitative suitability.

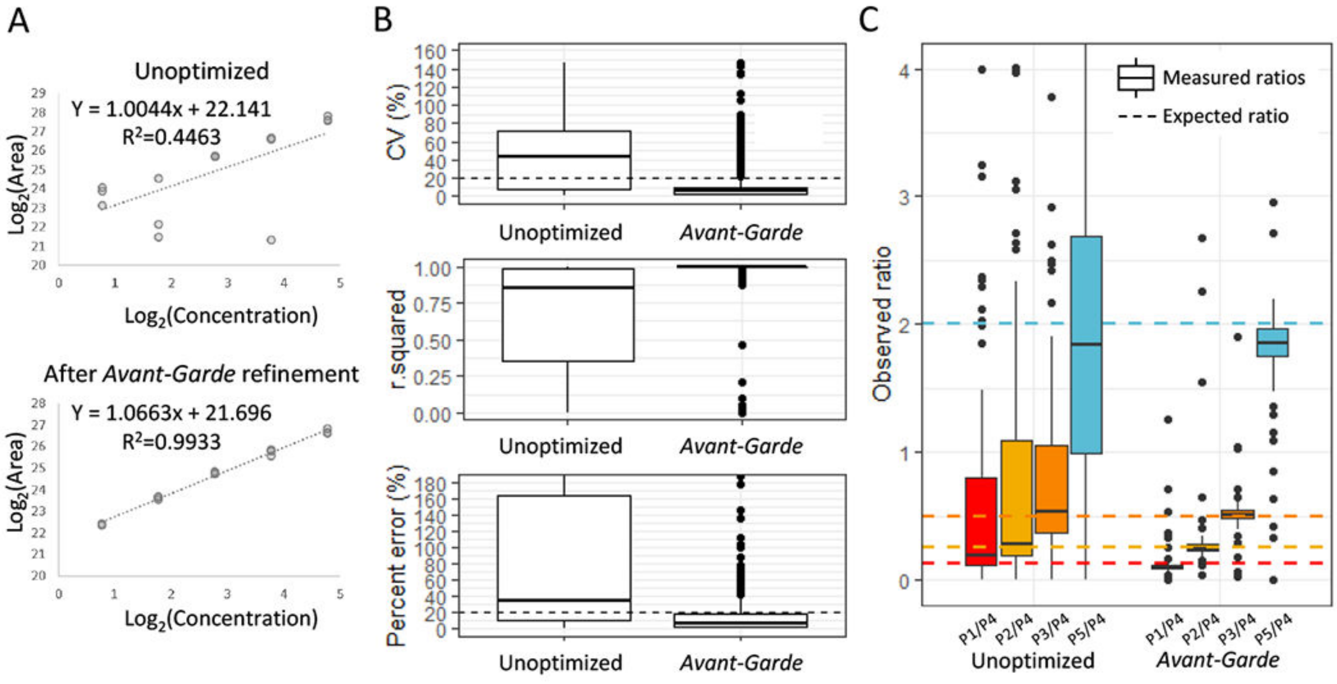

Figure 2: AvG improves quantitative figures-of-merit in a calibration curve.

95 synthetic phosphopeptides were spiked into a HEK293T whole cell digest to create a 5-point calibration curve (5 samples were analyzed in triplicate). (A) Calibration curve before (top) and after (bottom) curation by AvG for peptide S[+80]LTAHSLLPLAEK. For this peptide, the calibration curve spans a range of 1.7 to 27.3 fmol injected on column (n=5, analyzed in triplicate). The r-squared value corresponds to the coefficient of determination. (B) Figures-of-merit summarising the results for all synthetic peptides, pre- and post-optimization (n=5, analyzed in triplicate): % CV of triplicates, r^2 values of linear fits, and absolute percent error of measurements relative to the known concentration. Dashed lines indicate 20% thresholds. The box plot elements are: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers. (C) Expected vs. observed ratiometric quantification, pre- and post-optimization (n=5, analyzed in triplicate). P1 to P5 represent the points of the calibration curve in increasing order of concentration. The ratios between the mean area of each calibration point to the mean area of the fourth calibration point (P4) are shown here. The dashed lines represent the expected ratios (0.125, 0.25, 0.5 and 2) and the boxplots show the distribution of the measured ratios. The boxplot elements are the same as described for panel B.

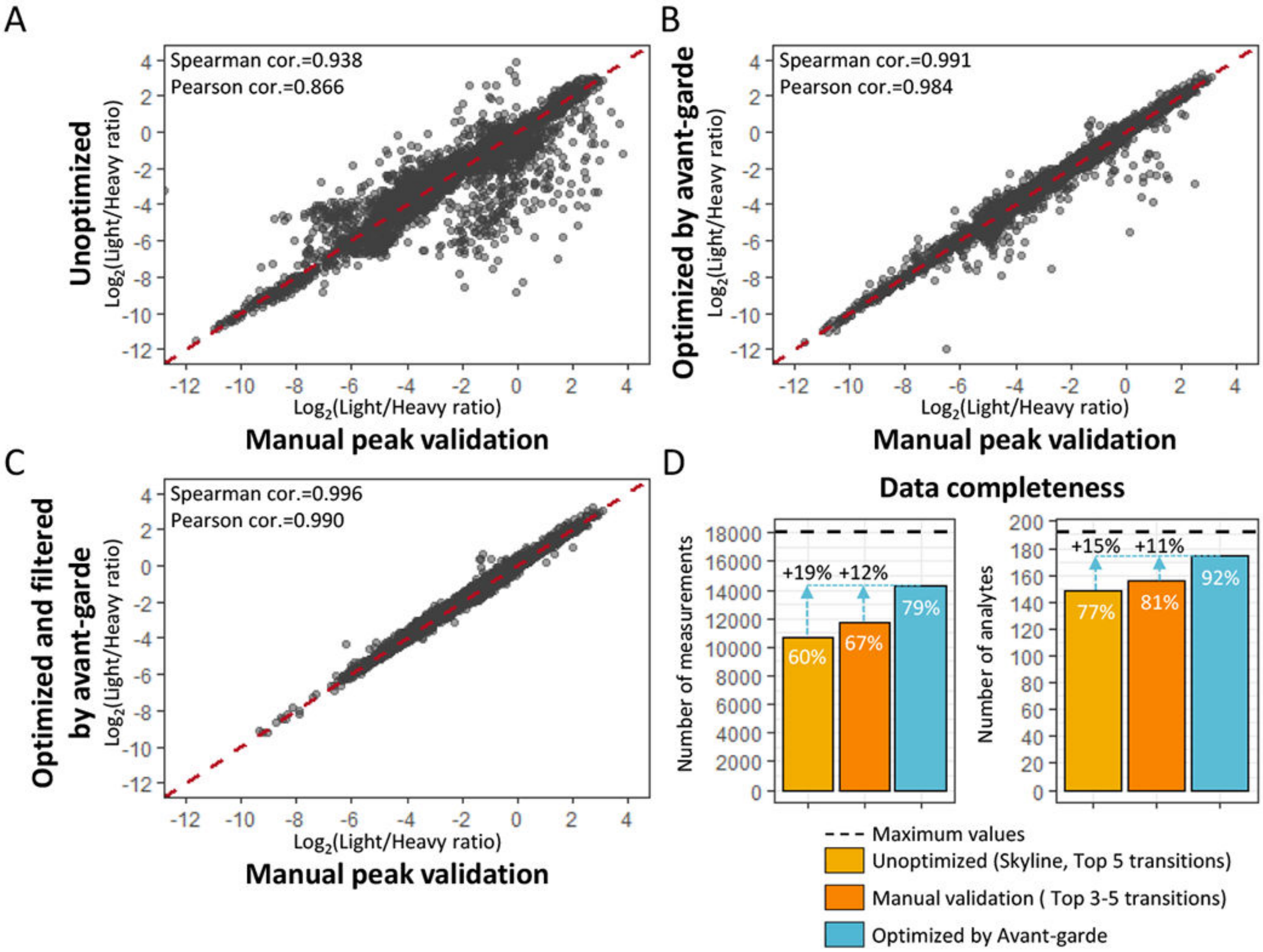

Figure 3: *Avant-garde* equals the performances obtained by expert visual inspection and manual validation.

95 phosphopeptides, and their isotopically labeled heavy peptide counterparts, were analyzed in a cohort of 96 phospho-enriched samples (n=96, 32 drug perturbations in triplicate). The dataset was initially analyzed using Skyline and manually curated by an expert. The scatter plots compare results of light-to-heavy ratios of the (A) unoptimized dataset, (B) the AvG “open” curation dataset, and (C) the AvG filtered curation dataset to the manually curated dataset. (D) Data completeness measured after filtering the data for quantitative suitability at the measurement and at the analyte level.

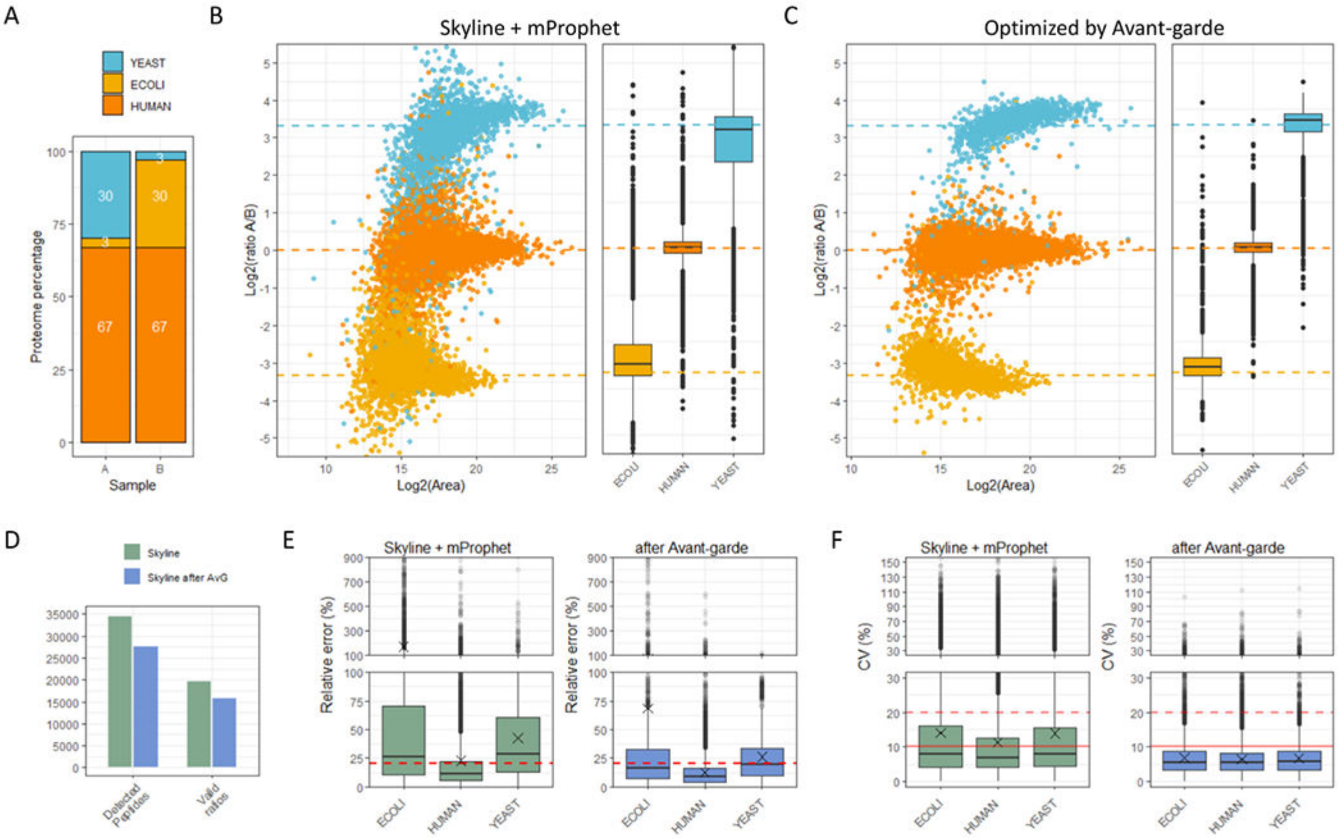

Figure 4: Evaluation of AvG with LFQBench data.

(A) The composition of the LFQBench samples by species proteome (n=2, analyzed in triplicate). (B,C) Results of the relative quantification and distribution of the experimental ratios obtained by Skyline using its implementation of mProphet for the pick peaking(B) and the AvG-curated dataset (C).Each dot represents a ratio calculated for a given peptide in a given run. The dashed lines represent the expected ratios. (D) The number of detected peptides and valid quantifiable ratios are shown for both datasets. The percent relative error (E) and the coefficient of variation (F) for each proteome for the Skyline (top) and AvG curated (bottom) dataset are shown. The horizontal full and dashed lines demarcate the 10% and 20 % threshold. The box plot elements are: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers.

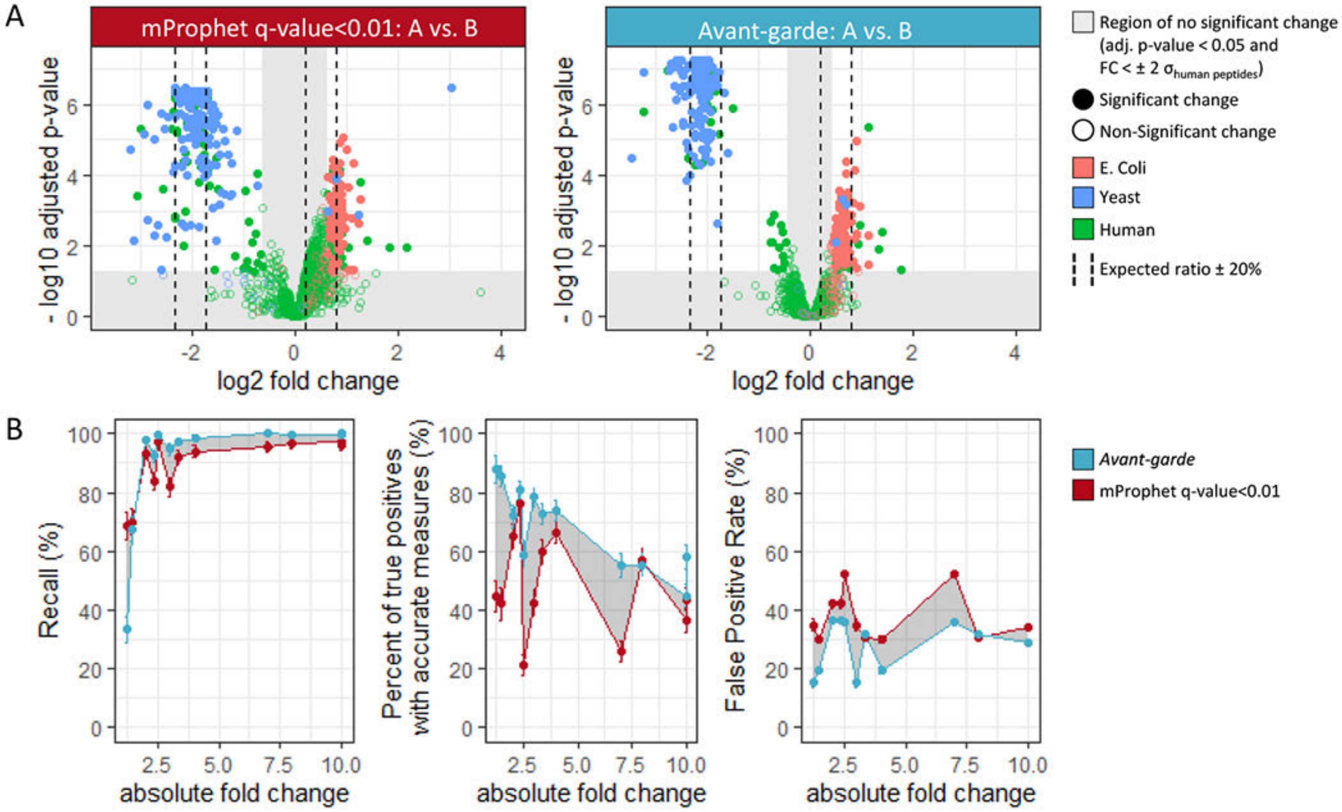

Figure 5: Detection of differentially expressed peptides in unoptimized and curated data.

(A) An example of a pairwise comparison (sample A vs. B, n=2 analyzed in quadruplicate), with volcano plots of unoptimized, Skyline+mProphet filtered with a q-value cutoff of 0.01 (left),and AvG curated (right) data. Each point represents an *E. coli* (red), Yeast (blue) or Human (green) peptide. The shaded regions demarcate ranges where detection of differential expression is not statistically viable (two-tailed two-sample moderated t-test, p-values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg method). The dashed lines represent accuracy boundaries of +/− 20%. (B) Bootstrap (n=1000) analysis of downsampled datasets for recall (sensitivity), accuracy, and false positive rate. Shaded regions indicate improvement in area under the curve after AvG curation. Error bars connote the standard deviation across bootstrap iterations and the center point represents the median value for each metric.