Selection of features with consistent profiles improves relative protein quantification in mass spectrometry experiments

Tsung-Heng Tsai\textsuperscript{1,1}, Meena Choi\textsuperscript{1}, Balazs Banfai\textsuperscript{2}, Yansheng Liu\textsuperscript{3}, Brendan X. MacLean\textsuperscript{4}, Tom Dunkley\textsuperscript{2}, Olga Vitek\textsuperscript{1,*}

\textsuperscript{1}Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
\textsuperscript{†}Current affiliation: Department of Mathematical Sciences, Kent State University, Kent, OH, USA
\textsuperscript{2}Roche Pharmaceutical Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, Basel, Switzerland
\textsuperscript{3}Department of Pharmacology, Yale University School of Medicine, New Haven, CT, USA
\textsuperscript{4}Department of Genome Sciences, University of Washington, Seattle, WA, USA

**Corresponding author:** marked as *
Olga Vitek (o.vitek@northeastern.edu)

**Running Title:**
Feature selection for relative protein quantification

**Abbreviations:**
LC-MS/MS, liquid chromatography-tandem mass spectrometry; MS, mass spectrometry; DDA, data dependent acquisition; SRM, selected reaction monitoring; DIA, data independent acquisition; FDR, false discovery rate
Abstract

In bottom-up mass spectrometry-based proteomics, relative protein quantification is often achieved with data-dependent acquisition (DDA), data-independent acquisition (DIA), or selected reaction monitoring (SRM). These workflows quantify proteins by summarizing the abundances of all the spectral features of the protein (e.g., precursor ions, transitions or fragments) in a single value per protein per run. When abundances of some features are inconsistent with the overall protein profile (for technological reasons such as interferences, or for biological reasons such as post-translational modifications), the protein-level summaries and the downstream conclusions are undermined. We propose a statistical approach that automatically detects spectral features with such inconsistent patterns. The detected features can be separately investigated, and if necessary removed from the dataset. We evaluated the proposed approach on a series of benchmark controlled mixtures and biological investigations with DDA, DIA and SRM data acquisitions. The results demonstrated that it can facilitate and complement manual curation of the data. Moreover, it can improve the estimation accuracy, sensitivity and specificity of detecting differentially abundant proteins, and reproducibility of conclusions across different data processing tools. The approach is implemented as an option in the open-source R-based software MSstats.
1 Introduction

Bottom-up proteomics is a method to analyze peptides from enzymatic digestion of proteins with liquid chromatography coupled with mass spectrometry (LC-MS). It is the most widespread proteomic workflow [2, 3] compatible with data-dependent acquisition (DDA), data-independent acquisition (DIA) [28, 11], and selected reaction monitoring (SRM) [18].

In DDA, the mass spectrometer alternates between acquiring a full MS1 spectrum of the co-eluting precursor ions for analyte quantification, and acquiring the fragment MS/MS spectra of the peptide ions for identification. SRM, an important method for targeted analysis, requires an a priori knowledge of the unique peptide sequences of the protein targets. The peptides are selectively profiled based on their precursor ions and representative fragment ions, and quantified by the intensity of each pair of precursor/fragment, also known as transition. Label-based SRM workflows further enhance the accuracy of the quantification by spiking an isotopically labeled reference counterpart of each targeted peptide into the samples. In DIA, the mass spectrometer covers the entire relevant mass range with a set of wide windows. All the precursor ions in a window are co-isolated and subjected to fragmentation, to produce their combined fragment spectrum.

A common use of DDA, SRM and DIA is relative protein quantification. Protein quantification summarizes all the quantitative information pertaining to a protein across the LC-MS runs, and performs statistical analyses that detect differentially abundant proteins between conditions. Since mass spectrometers quantify peptide ions or their fragments instead of proteins, deriving protein-level conclusions can be challenging. First, the quantitative profile of a peptide ion or a fragment can deviate from the representative protein profile for biological reasons (e.g., in presence of post-translational modifications or unknown protein isoforms), or for technological reasons (e.g., due to interferences of co-eluting precursor ions such as ion suppression). Second, some of the precursor ions or fragments can be incorrectly identified or quantified across the LC-MS runs. This is particularly common for low abundant analytes. Third, in DDA workflows, the stochastic nature of the acquisition of fragmentation spectra makes it challenging to consistently identify and measure the same species in multiple LC-MS runs, and produces missing values. Finally, computational processing of the spectra adds uncertainty to protein quantification. A variety of computational tools, e.g., MaxQuant [9, 8], Progenesis (Nonlinear Dynamics), Proteome Discoverer (Thermo Fisher Scientific), Skyline [15], DIA Umpire [26], OpenSWATH [20] and Spectronaut [4], identify, quantify and align the peaks across the LC-MS runs. Each tool optimizes different computational tasks, and the resulting list of differentially abundant proteins may be, at least to some extent, tool-dependent. Overall, these issues undermine the accuracy and reproducibility of protein-level summarization and relative quantification.

This manuscript aims to alleviate these challenges. We propose an automated statistical approach, which takes as input a set of identified and quantified spectral features reported by a data processing tool, characterizes the information content in the individual features, and performs relative protein quantification using an informative subset of the features. We consider DDA, SRM and DIA acquisition types, and use the generic term feature to refer to precursor ions in DDA, transitions in SRM, or fragments in DIA, identified and quantified across multiple LC-MS runs. We define informative features features that contain few missing values, and for which the intensities are consistent with the majority of the protein intensity profiles across the runs. We define uninformative features features that do not satisfy these properties. Among those, low-coverage features contain
a large number of missing values. Outliers are inconsistent with the overall protein profile. Finally, noisy features exhibit substantial variation between the LC-MS runs, beyond the variation of the majority of the features of the protein.

2 Background

Detecting uninformative features in LC-MS data often involves manual curation. A feature of a peptide is evaluated by plotting the proportion of its intensity in the sum of the intensities of all the features of the peptide. For the informative features, this proportion is roughly constant across all the LC-MS runs in the investigation. Deviations from the constant pattern indicate that the feature may be compromised by noise or artifacts, and should be further investigated. Such manual curation is limited in throughput. Also, when performed separately for each peptide as in common practice, it does not account for potentially distinct peptide profiles of the same protein.

Automatic approaches for detecting uninformative features have been introduced as alternatives to manual curation. They can be loosely classified into two groups: those detecting interferences or errors in identification and quantification within a single LC-MS run, and those detecting these patterns across multiple LC-MS runs. Many methods consider properties of spectral features in a single LC-MS run. These methods evaluate feature-specific patterns such as relative intensities, chromatographic traces, or patterns of isotopic distributions. The patterns are compared to their expected patterns for the analyte (or, in the case of SRM, to the observed patterns of their isotopically labeled reference counterparts). For example, the scoring model in mProphet [19] combines multiple quality metrics, and filters out peaks with likely incorrect identification, presence of interfering ions, etc. Originally developed for SRM, the mProphet algorithm has since been implemented in Spectronaut, OpenSWATH and Skyline for DIA workflows. Similar steps are implemented in other data processing tools, e.g., SWATHProphet [14] and EncyclopeDIA [21] for DIA.

To select informative features in multiple LC-MS runs, a common heuristic, called “top-n”, selects n features with the highest average intensities across the runs. The heuristic is based on the insight that higher intensity features are less impacted by interference, and therefore have a higher information content. More recent approaches make more sophisticated considerations. For SRM, AuDIT [1] calculates the intensity ratio between each pair of features (i.e., transitions) generated from a same precursor in all runs, compares them to the ratios of the isotopically labeled reference counterparts, and flags significantly deviating ratios using the t-test. Although the method is simple and performs well, it requires high-quality information from the standards. For DIA, mapDIA [25] implements a suite of data processing steps, including intensity normalization, fragment filtering, and detection of differentially abundant proteins. The fragment filtering step includes outlier detection, and uses pairwise correlation to select features with consistent profiles across the runs. The step is tightly integrated with the rest of the mapDIA workflow, and requires tuning multiple parameters that control the cross-fragment correlation and the number of peptides and fragments. Using several quality metrics, Avant-garde [27] automatically curates features with a genetic algorithm, and adjusts their peak boundaries for quantification. The method is integrated as an external tool of Skyline. Diffacto [29] performs protein summarization through robust factor analysis of the features, views the features corresponding to the majority factor as representative of the protein, but does not select a subset of informative features.
This manuscript aims to improve upon the existing methods above. We propose a statistical approach that takes as input the identified and quantified features that pass single-run filtering criteria (e.g., by mProphet), and complements the filtering using the information available from multiple LC-MS runs. It selects a subset of informative spectral features for detecting differentially abundant proteins. The uninformative features can be separately investigated, curated, or removed. The approach is compatible with multiple data acquisition workflows and data processing tools.

3 Experimental procedures

The input datasets are briefly described below, and summarized in Tables 1 and 2. The datasets are available on either MassIVE.quant [5] or Panorama [23]. Details about data processing, R script for statistical analysis, and the differential abundance analysis results are available on MassIVE.quant. Dataset identifiers and reanalysis container identifiers on MassIVE.quant are listed in Tables 1 and 2. Additionally, the R script for summarizing all the analysis results in this manuscript is available at https://github.com/tsunghengtsai/MSstats-feature-selection.

Benchmark controlled mixtures We evaluated the proposed approach on four benchmark controlled mixtures of known composition. Raw spectra from a same investigation were processed with up to four data processing tools. Data from two controlled mixtures were acquired with DDA: i) the iPRG benchmark [6], where the data were processed by MaxQuant, Progenesis, and Skyline; and ii) the Choi benchmark [7], where the data were processed by MaxQuant, Progenesis, Proteome Discoverer, and Skyline. Data from two controlled mixtures were acquired with DIA: i) the Bruderer benchmark [4], where the data were processed by DIA Umpire, Skyline, and Spectronaut; and ii) the Navarro benchmark [16], where the data were processed by DIA Umpire, OpenSWATH, Skyline, and Spectronaut.

This work did not aim to compare the data processing tools. Rather, we used these datasets to evaluate the impact of the proposed approach on accuracy of fold change estimation, sensitivity and specificity of detecting differentially abundant proteins, and reproducibility of conclusions across the data processing tools. We compared the performance to i) all-features protein quantification, as currently implemented in MSstats, and ii) protein quantification based on the n features with the highest average intensities across runs (i.e., top-n features) using MSstats.

Biological investigations We evaluated the proposed approach using two biological investigations. The first is a label-based SRM experiment by Dunkley et al. [10], investigating in vivo neuro-development. In addition to the biological samples, the experiment included 35 replicate runs from a pooled quality control (QC) sample to assess the technical performance. Also, spike-in experiments with pooled isotope-labeled peptides spiked into the QC sample were performed to characterize the linear portion of the dynamic range for each analyte. The data were processed with Skyline, and the quality of each feature was manually curated. The manual curation consisted of two steps: i) removing peptides with no signal, and transitions with interference, and ii) removing peptides that failed to meet a set of thresholds based on coefficient of variation (CV) of the QC runs, and the range of intensity values compared to the linear portion of the dynamic range. Whenever either an endogenous or a reference feature did not meet the quality thresholds,
the entire pair was removed. Further details regarding the manual curation can be found in [10]. We used this dataset to evaluate the capability of the proposed approach to automatically detect uninformative features, as compared to the manual curation.

The second biological investigation is a DIA-based experiment by Selevsek et al. [22]. The main investigation quantified the proteome of *Saccharomyces cerevisiae* in response to osmotic stress at six time points. In addition, a separate refinement experiment performed eight technical DIA runs on a same biological sample. Two analyses of the resulting data were performed in the study, both using Skyline. The first analysis, which we call Full, considered all the peptide precursors in the main investigation, and quantified each precursor with the top-6 fragments. The second analysis, which we call LowCV, selected peptides detected in at least 50% of the refinement runs, and with CV < 20% in those runs. The selected peptides were then quantified in the biological samples in the main investigation. After these steps, the features selected in both analyses were further subjected to mProphet to characterize the quality of the peaks in individual LC-MS runs in terms of $q$-values. Compared to the Full analysis, the LowCV analysis queried against a reduced spectral library. Therefore, a peak in the two analyses could have different $q$-values. We considered two types of $q$-filters: i) the Sparse $q$-filter eliminated peaks with $q$-value > 0.01, and ii) the 50% $q$-filter eliminated the entire LC-MS features where the $q$-values of over 50% of the peaks exceeded 0.01. Combinations of the two data analyses with the two $q$-filter strategies produced four datasets: Full_Sparse, Full_50%, LowCV_Sparse, and LowCV_50%. We used these datasets to evaluate the impact of the proposed approach on relative protein quantification under different choices of data processing. All the downstream statistical analyses were performed with MSstats v3.16.2.

**Simulated datasets** In addition to the experimental data, computer simulations evaluated the effect of outliers, and of different number of features and runs, on the performance of the proposed approach. The details are in Supplementary Section S1.

4 Proposed statistical approach for selecting informative features

4.1 Overview and notation

The input to the approach is a list of log-transformed intensities (peak areas, heights at apex, etc.), identified, quantified and aligned across LC-MS runs by a data processing tool. We denote $y_{ijk}$ the log-intensity from protein $i$ ($i = 1, \ldots, I$), feature $j$ ($j = 1, \ldots, J$) and run $k$ ($k = 1, \ldots, K$). Some values $y_{ijk}$ can be missing due to the stochastic nature of data acquisition (e.g., in DDA), or due to filtering steps (e.g., with respect to $q$-values). We denote $n_{ij}$ the number of observed log-intensities in protein $i$ and feature $j$. We order the features of a protein by descending order of the number such that $n_{ij} \geq n_{i,j+1}$ for $j = 1, \ldots, J - 1$. The proposed approach consists of: i) developing a probability model for a representative pattern of missing log-intensities of a protein, and detecting features that deviate from this pattern, and ii) developing a probability model for a representative pattern of observed log-intensities of a protein, and detecting features that deviate from this pattern in terms of outlying log-intensities and between-run variation. We describe these aspects below, and illustrate each step in Fig. 1a, using as an example protein P32915 from the DDA iPRG benchmark, quantified with MaxQuant.
4.2 Detection of low-coverage features

Some amount of missing log-intensities is expected in any experiment, in particular when protein abundance is close to the limit of detection, or when a protein is absent in the samples in some conditions. Our first step is to detect low-coverage features, which contain fewer log-intensities than what is representative for the protein.

Statistical model We assume that the number \( n_{ij} \) of observed log-intensities in protein \( i \) and feature \( j \) follows a Binomial distribution:

\[
P(n_{ij}|K, \pi_i) = \binom{K}{n_{ij}} \pi_i^{n_{ij}} (1 - \pi_i)^{K-n_{ij}},
\]

where \( K \) is the number of runs and \( \pi_i \) is the representative probability of observing log-intensities in protein \( i \). The Binomial model characterizes both the expected number and the variability of the observed log-intensities in every feature of protein \( i \).

Estimation of representative probability \( \pi_i \) The value of parameter \( \pi_i \) in Eq. 1 is unknown and must be estimated from the data. We estimate \( \pi_i \) as the average proportion of observed log-intensities among all the features of protein \( i \):

\[
\hat{\pi}_i = \frac{1}{J} \sum_{j=1}^{J} \frac{n_{ij}}{K}.
\]

For example, protein P32915 (Fig. 1a) has five features quantified across 12 runs. One feature (TWIEISGTSPR, highlighted in dark gray) has six observed log-intensities, and each of the other four features has 12 observed log-intensities. The representative probability of observing log-intensities is estimated as \( \hat{\pi}_i = ((6 + 12 \cdot 4)/12)/5 = 0.9 \).

Detection of low-coverage features According to the Binomial distribution, the cumulative probability of observing \( n_{ij} \) or fewer log-intensities in protein \( i \) and feature \( j \) is \( \sum_{n=0}^{n_{ij}} P(n|K, \hat{\pi}_i) \). We label a feature as low-coverage if this probability is below a pre-defined, typically very low confidence level \( \alpha_c \) (e.g., \( \alpha_c = 0.01 \)):

\[
\sum_{n=0}^{n_{ij}} P(n|K, \hat{\pi}_i) \leq \alpha_c.
\]

The low-coverage features are excluded from the subsequent analyses. After this step, the protein has \( J' \) remaining features (\( J' \leq J \)), and for each of the remaining features \( \sum_{n=0}^{n_{ij}} P(n|K, \hat{\pi}_i) > \alpha_c \). For example, the first panel of Fig. 1a illustrates that the cumulative probability for feature TWIEISGTSPR, \( \sum_{n=0}^{6} P(n|12, 0.9) \) is below the confidence level (\( \leq 0.01 \)). Therefore, the feature is labeled as low-coverage, and is excluded. Note that, in general, the dataset may still contain missing values after this filtering.
4.3 Detection of outliers and noisy features

**Statistical model** The model describes the patterns of systematic and random variations across LC-MS runs, expected in situations where all the remaining features of protein $i$ are informative. We represent the log-intensities of all the features of a protein with a relatively simple linear model

$$y_{ijk} = \mu_i + \alpha_{ij} + \beta_{ik} + \epsilon_{ijk},$$

$$\sum_{j=1}^{J'} \alpha_{ij} = 0, \quad \sum_{k=1}^{K} \beta_{ik} = 0, \quad E[\epsilon_{ijk}] = 0 \text{ and } \text{Var}[\epsilon_{ijk}] = \sigma_i^2,$$  \hspace{1cm} (3)

Here $\mu_i$ is the mean log-intensity of the protein $i$ across all the features and all the runs, $\alpha_{ij}$ is the deviation of the mean log-intensity of the feature $j$ from the overall mean (expressing differences in ionization efficiency, fragmentation patterns, etc.), and $\beta_{ik}$ is the deviation of the log-intensity of the run $k$ from the overall mean (expressing differences induced by conditions, biological replicates, technical replicates, etc.). The term $\epsilon_{ijk}$ expresses the stochastic variation unexplained by the terms above. Note that the stochastic variation needs not to be Normally distributed. We only assume that it is non-systematic (i.e., $E[\epsilon_{ijk}] = 0$) and has protein-specific variance $\sigma_i^2$.

**Estimation of representative protein profiles** The values of the parameters in Eq. 3 are unknown, and must be estimated from the data. Since we anticipate the presence of missing values, outliers and features with noise, it is important to develop an estimation procedure that is robust to such artifacts, especially in experiments with few features and runs.

Multiple robust methods for estimating the parameters in Eq. 3 exist, e.g., Tukey’s median polish (TMP) implemented for run-level summarization in MSstats [7]. However, when a feature systematically deviates from the representative protein profile, the robustness of TMP is quickly reduced [12]. Here we propose to estimate the parameters by robust regression with M-estimation [13], a generalization of the maximum-likelihood estimation. The estimator defines the residuals $r_{ijk} = y_{ijk} - (\mu_i + \alpha_{ij} + \beta_{ik})$ and minimizes the objective function

$$\hat{\mu}_i, \hat{\alpha}_{ij}, \hat{\beta}_{ik} = \arg\min_{\mu_i, \alpha_{ij}, \beta_{ik}} \sum_{j=1}^{J'} \sum_{k=1}^{K} \rho(r_{ijk}) = \sum_{j=1}^{J'} \sum_{k=1}^{K} \rho(y_{ijk} - (\mu_i + \alpha_{ij} + \beta_{ik})), \hspace{1cm} (4)$$

where $\rho(r)$ is the Huber function

$$\rho(r) = \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \leq c \\ c|r| - \frac{1}{2}c^2 & \text{if } |r| > c \end{cases} \hspace{1cm} (5)$$

The constant $c$ is set to $1.345 \times \sigma_i$, such that the estimator is 95% as efficient as the least squares estimator when the stochastic variation is Normally distributed [13]. The standard deviation $\sigma_i$ is unknown, and is estimated with either Huber’s proposal 2 [13] or with the median absolute deviation (MAD, $\hat{\sigma}_i = \text{median}(|r_{ijk}|) / 0.6745$), where the constant 0.6745 scales the MAD to approximately match the standard deviation of a Normal distribution. Note that the Huber function $\rho(r)$ requires an estimate of standard deviation $\hat{\sigma}_i$, which depends on the residuals with the estimated parameters in Eq. 4. An iteratively reweighted least squares method is therefore required to estimate $\sigma_i$, $\mu_i$, $\alpha_{ij}$, and $\beta_{ik}$. For estimating $\hat{\sigma}_i$, we chose Huber’s proposal 2, based on the simulations in Supplementary Section S1.
With the estimated parameters, the protein abundance in run $k$ is given by $(\hat{\mu}_i + \hat{\beta}_{ik})$, as highlighted in yellow in the second panel of Fig. 1a, and the residuals are

$$\hat{r}_{ijk} = y_{ijk} - (\hat{\mu}_i + \hat{\alpha}_{ij} + \hat{\beta}_{ik}).$$

(6)

**Estimation of representative variation** Since the estimates $\hat{\sigma}_i^2$ can be unstable in experiments with a small number of features per protein or a few runs, we enhance the estimation with the empirical Bayes (EB) shrinkage estimator, originally implemented in the limma package [24]. The EB estimator leverages the information across all the proteins in the dataset to stabilize the variance estimates. Briefly, it assumes that the estimate $\hat{\sigma}_i^2$ follows a scaled chi-square distribution with $d_i$ degrees of freedom

$$\hat{\sigma}_i^2 \mid \sigma_i^2 \sim \sigma_i^2 \chi_{d_i}^2,$$

and the true protein-specific variance $\sigma_i^2$ is in turn sampled from a scaled inverse chi-square distribution with degrees of freedom $d_0$ and location $s_0^2$:

$$\sigma_i^2 \sim \frac{s_0^2 d_0}{\chi_{d_0}}.$$

(7)

Under these assumptions, $\hat{\sigma}_i^2$ follows a scaled $F$-distribution on $d_i$ and $d_0$ degrees of freedom

$$\hat{\sigma}_i^2 \sim s_0^2 F_{d_i,d_0},$$

and $\log \hat{\sigma}_i^2$ follows Fisher’s $z$-distribution [24], which is roughly symmetric and has finite moments of all orders.

The degree of freedom $d_i$ in Eq. 7 is determined by the experimental design, as in two-way fixed effects models. For example, in a balanced dataset (where each feature has fully observed log-intensities across runs), $d_i = (J' - 1)(K - 1)$. The hyperparameters $s_0^2$ and $d_0$ in Eq. 8 are obtained by matching the first two moments (i.e., mean and variance) of the empirical distribution of $\log \hat{\sigma}_i^2$ to that of the $z$-distribution [24]:

$$\text{E}(\log \hat{\sigma}_i^2) = \log s_0^2 + \psi \left( \frac{d_i}{2} \right) - \psi \left( \frac{d_0}{2} \right) + \log \left( \frac{d_0}{d_i} \right)$$

and

$$\text{var}(\log \hat{\sigma}_i^2) = \psi' \left( \frac{d_i}{2} \right) + \psi' \left( \frac{d_0}{2} \right)$$

where $\psi()$ and $\psi'()$ are the digamma and trigamma functions, respectively. As the moment matching estimation is sensitive to the extreme values of $\hat{\sigma}_i^2$, we employ a robust solution originally proposed in limma [17], and only consider the 5–90% percentiles of the empirical distribution of $\log \hat{\sigma}_i^2$ for the match.

After obtaining the hyperparameters $s_0^2$ and $d_0$, the posterior mean of $1/\sigma_i^2$ given $\hat{\sigma}_i^2$ is $1/\tilde{\sigma}_i^2$ with

$$\sigma_i^2 = \frac{d_0 s_0^2 + d_i \hat{\sigma}_i^2}{d_0 + d_i}.$$

(9)

The $\tilde{\sigma}_i^2$ is the EB shrinkage estimator of the representative protein variation. As shown in the second panel of Fig. 1a, the EB variance estimates ‘shrink’ towards the consensus value $s_0^2$. 
Detection of outlying log-intensities  Residuals $\hat{r}_{ijk}$ of outlying log-intensities exceed the representative protein variation $\hat{\sigma}_i^2$. Therefore, if the residual

$$|\hat{r}_{ijk}| > 3\hat{\sigma}_i,$$  \tag{10}$$
we label the corresponding log-intensity as an outlier, and replace the log-intensity and the residual with NA. For example, in protein P32915 with $\hat{\sigma}_i = 0.32$ (Fig. 1a), the log-intensity from feature GQIGIYPIK_2 and run 12 has a residual greater than the cutoff $3\hat{\sigma}_i = 0.96$, and is labeled as an outlier (highlighted in dark gray in the third panel). With this cutoff, we expect a 0.3\% false positive rate of outliers when the residuals are Normally distributed and have a known standard deviation.

Detection of noisy features  Residual variation of noisy features exceeds the representative protein variation $\hat{\sigma}_i^2$. We evaluate the variation in feature $j$ of protein $i$ according to the second moment of its residuals, scaled by the variance estimate of the protein, i.e.,

$$\hat{\tau}_{ij} = \frac{1}{n'_{ij} - 1} \sum_{k=1}^{K} \frac{\hat{r}_{ijk}^2}{\hat{\sigma}_i^2},$$  \tag{11}$$
where larger values indicate noisier features. Note that the summation does not include the outlying log-intensities replaced with NA in the previous step, and $n'_{ij}$ is the number of non-outlying log-intensities in protein $i$ and feature $j$. The scaling factor $\hat{\sigma}_i^2$ allows us to compare all the features of all the proteins simultaneously. To determine a significance threshold, we generate a reference summary for each feature

$$\hat{\tau}_{ij}^{ref} = \frac{1}{n'_{ij} - 1} \sum_{k=1}^{K} \frac{(y_{ijk} - \bar{y}_i)^2}{\hat{\sigma}_i^2},$$  \tag{12}$$
where $\bar{y}_i$ is the average log-intensity over runs and features in protein $i$, and it can be viewed as a less accurate prediction of $y_{ijk}$ ignoring the contributions from both the runs and the features in Eq. 3. Therefore, in most cases the squared deviations $(y_{ijk} - \bar{y}_i)^2$ are larger than the squared residuals $\hat{r}_{ijk}^2$ (Eq. 6), and the values of the reference summary $\hat{\tau}_{ij}^{ref}$ are generally larger than the targets $\hat{\tau}_{ij}$. We use the empirical distribution of the reference summary $\hat{\tau}_{ij}^{ref}$ to determine the largest acceptable value of $\hat{\tau}_{ij}$, where a threshold corresponding to a 5\% false positive rate on the lower tail of the distribution is selected. For example, in the fourth panel of Fig. 1a, the empirical distributions of the feature summary $\hat{\tau}_{ij}$ and the reference $\hat{\tau}_{ij}^{ref}$ are shown in log scale. With the reference distribution, a threshold for log $\hat{\tau}_{ij}$ is set to $-0.23$. The dark gray feature GQIGIYPIK_2 exceeds the threshold, and is labeled as a noisy feature.

Detection of outliers and noisy features in SRM experiments with labeled reference peptides  In these experiments, the proposed approach is performed separately for features from endogenous and reference peptides. For reference peptides, the procedure is modified to i) skip the outlier detection/removal step (Eq. 10), and ii) consider all the log-intensities when calculating the feature summary and the reference (Eqs. 11 and 12). The rest of the procedure is applied identically to the endogenous and reference peptides.
4.4 Implementation in the open-source R-based software package MSstats

Supplementary Fig. S3 presents the pseudocode of the proposed approach for label-free DDA, DIA and SRM workflows. The approach is implemented as an option within the open-source R-based software MSstats v3.16.2 (or later), which can be downloaded through Bioconductor (https://bioconductor.org/packages/release/bioc/html/MSstats.html). It is applied after the data processing steps such as normalization, and before the protein-level summarization and statistical inference such as class comparison, sample summarization, or sample size calculation, as illustrated in Fig. 1b.

The approach is integrated in the dataProcess function as the option featureSubset=’highQuality’. The option highlights the uninformative features, which can be separately investigated, curated, or removed. The option remove_uninformative_feature_outlier=TRUE removes the detected uninformative features and outliers. The implementation of the robust regression is based on the rlm function in the R package MASS, and the implementation of the empirical Bayes shrinkage estimator is based on the squeezeVar function in limma. Further details about the usage of the feature selection approach can be found in the vignette and user manual of MSstats, available at http://msstats.org/.

5 Results

We evaluated the proposed approach on a series of benchmark controlled mixtures and biological investigations, analyzed by multiple data processing tools. All the downstream statistical analyses were performed with MSstats. Using the benchmark datasets, we evaluated the impact of the proposed selection of informative features on accuracy of fold change estimation, sensitivity and specificity of detecting differentially abundant proteins, and reproducibility of conclusions across the data processing tools. We compared the performance to that with all the features and with the top-n features. Moreover, we used the biological investigations to compare the proposed approach with the manual curation, and to evaluate the impact of the proposed approach on relative protein quantification under different choices of data processing.

In this section, we assume that all the inconsistencies in the features are due to noise, and should be removed. However, in many investigations the inconsistencies may be due to genuine biological reasons such as presence of post-translational modifications. Such features should be considered separately, as discussed in Section 6. When removing the uninformative features, some proteins may be eliminated from the dataset. As a result, the number of comparisons between conditions can change before and after the feature selection procedure.

5.1 Evaluation with the DIA benchmarks

For DIA experiments, we used the Bruderer benchmark and the Navarro benchmark to evaluate the proposed feature selection approach. The two benchmarks differ in the number of differentially abundant proteins and their fold changes. In the Bruderer benchmark, three mixes of 12 non-human proteins were spiked into a constant background proteome (HEK-293), covering a range of fold
changes from 1.1 to 50. On the other hand, the Navarro benchmark had tryptic digests of human, yeast and *Escherichia coli* proteins mixed in defined proportions (1:1 for human, 2:1 for yeast, and 1:4 for *Escherichia coli* proteins), resulting in a large number of true differentially abundant proteins (around 3,000) with fold changes in the range of 2 to 4. Data from both benchmarks were analyzed with at least three data processing tools.

### 5.1.1 Improved estimation accuracy and detection of differential abundance

Fig. 2a illustrates the number of features per protein across the DIA benchmarks, before and after the proposed selection of informative features. As shown in the figure, the number of features varied substantially, and depended on both the experimental design and the data processing tool. With the proposed approach, the number of selected informative features became more consistent, with a median around 10-20.

Fig. 2b and Fig. 2c detail the proposed approach in the case of protein C8ZIG9 from the Navarro benchmark. Skyline and Spectronaut quantified the protein with different numbers of features, and with somewhat different patterns of between-run variation and outliers. In both cases, protein-level summarization with top-3 features was compromised by outliers in the high-intensity features, and the estimated log-fold changes were in the opposite direction of the truth. In the case of Skyline, the change in the opposite direction was large enough to warrant statistical significance (adjusted \( p \)-value 0.002). The all-features summarization was in turn compromised by the noise in low-intensity features. Although the estimates of the log-fold changes were in the right direction, the variation inflated the standard errors, and undermined the adjusted \( p \)-values. In contrast, the proposed approach effectively eliminated the uninformative features, and produced the most accurate estimation of log-fold changes, as well as adjusted \( p \)-values consistent with differential abundance. Additional examples illustrating the impact of missing values are shown in Supplementary Fig. S4.

Beyond considering individual proteins, Fig. 3a and Fig. 3b summarize the absolute errors of the fold change estimates and associated standard errors, considering multiple benchmarks and data processing tools, and focusing on background proteins with no differential abundance. The figures compare protein quantification with the selected features to that with top-\( n \) for a range of \( n \). The top-10 quantification approximately matches the median number of features per protein selected by the proposed approach. With a small \( n \), the top-\( n \) quantification was vulnerable to uninformative features and outliers, and increasing \( n \) increased its robustness to data artifacts. Selecting informative features performed consistently better than the top-\( n \) in terms of estimates of log-fold changes and their standard errors.

Table 3 summarizes the results of detecting differentially abundant proteins across conditions at the false discovery rate (FDR) of 0.05, across the benchmarks and the data processing tools, using all the features and using the informative features. Selecting informative features consistently improved the sensitivity, while maintaining the specificity of detecting the changes. Supplementary Figs. S5 and S6 show the receiver operating characteristic (ROC) curves for all the considered methods for the Bruderer and Navarro benchmarks. The performance of the top-\( n \) quantifications varied across the datasets. For example, the top-3 quantification had the highest sensitivity in the Bruderer benchmark analyzed by Skyline (Supplementary Fig. S5), but performed the worst in the Navarro benchmark datasets (Supplementary Fig. S6). In contrast, selecting informative features consistently optimized the detection of differential abundance in these comparisons.
5.1.2 Reproducibility of detecting true differentially abundant proteins across tools

Since data processing may introduce discrepancy in protein quantification, we evaluated the ability of the proposed selection of informative features to improve the between-tool reproducibility of relative protein quantification. Fig. 4 shows the agreement of detected true changes in protein abundance between the data processing tools for the two DIA benchmarks. DIA Umpire identified substantially fewer proteins in the Navarro benchmark than the other tools (the fact also noted in the original publication [16], and shown in Table 1). We therefore excluded it from this evaluation. Selecting informative features maximized the agreement in detected true changes in protein abundance across the tools. Similarly to the discussion in Section 5.1.1, performance of the top-3 and top-5 methods in detecting differentially abundant proteins was inconsistent across the datasets.

5.2 Evaluation with the DDA benchmarks

For DDA experiments, we used the Choi benchmark and the iPRG benchmark to evaluate the proposed selection of informative features. In the Choi benchmark, three mixes of 30 proteins were spiked into a constant background of *Escherichia coli* proteome in different amounts, with fold changes of 1, 2 and 4. The iPRG benchmark considered six individual proteins spiked into a constant *Saccharomyces cerevisiae* proteome, covering a wide range of true fold changes (from 1.1 to 833). Data from both benchmarks were analyzed with at least three data processing tools.

5.2.1 Effective detection of uninformative features and outliers

DDA experiments typically quantify fewer features than DIA, and have more missing values and noise. Fig. 5a illustrates the number of quantified features per protein across the DDA benchmarks, before and after the selection of informative features. The number of features was smaller than in DIA (Fig. 2a). Similarly to the DIA, the original number of features per protein depended on the experimental design and on the data processing. For example, while the iPRG benchmark analyzed by Progenesis produced fewer features per protein (the median of 3 features) than Skyline, the Choi benchmark analyzed by Progenesis produced more features than Skyline. Selecting informative features mitigated this discrepancy, and resulted in a more comparable number of features per protein (median of 5) in most datasets.

Fig. 5b and Fig. 5c illustrate the proposed approach in the case of protein P32898 from the iPRG benchmark. Quantified with Progenesis, the protein had only four features, including two with inconsistent, high-intensity profile in runs 4-6. The noisy features undermined the protein-level summaries and resulted in a false positive detection of differential abundance (adjusted *p*-value close to 0 with both all-features and top-3 quantifications). Selection of informative features excluded one noisy feature and successfully corrected this artifact. Skyline reported more features for this protein. While some of the low-intensity features contained outliers, the majority of feature profiles formed a consistent pattern, and all the methods led to the correct conclusion of no change in abundance.
5.2.2 Performance in terms of relative protein quantification

Using the same evaluation measures as in Section 5.1, we evaluated the impact of the proposed approach on relative protein quantification with the DDA benchmarks. Similarly to Fig. 3, Supplementary Fig. S7 illustrates the effect of selecting informative features on the estimates of log-fold changes and their standard errors for the DDA benchmarks processed with different tools. The proposed approach more accurately estimated log-fold changes and had smaller standard errors than the top-3 and top-5 quantifications, but performed similarly to the all-features quantification.

Table 4 summarizes the results for detecting differentially abundant proteins between conditions at FDR=0.05. Overall, selecting informative features improved the sensitivity over the all-features method while maintaining a similar specificity. The improvement, however, was not as pronounced as in the DIA datasets. The ROC curves for the DDA benchmark datasets are shown in Supplementary Figs. S8 and S9. In the Choi benchmark analyzed with Proteome Discoverer, the proposed approach had a reduced sensitivity compared to the all-features quantification (Table 4 and Supplementary Fig. S8), with two fewer detected true changes at the FDR cutoff of 0.05. The proteins with missed true changes by the proposed approach typically had few or no features with a consistent profile across all the conditions. Removing uninformative features alone did not fully eliminate the inconsistency and the relative protein quantification was compromised, as illustrated in Supplementary Fig. S10. These cases would require further process to curate or re-quantify the features.

In terms of reproducibility of detecting true positives across data processing tools, there was no clear change between the proposed approach and the all-features quantification. The top-n methods, on the other hand, resulted in worse reproducibility across tools (Supplementary Fig. S11). Overall, compared to the all-features quantification, the proposed approach effectively removed uninformative features in the DDA benchmarks, but did not affect the relative protein quantification.

5.3 The selected features overlapped with manually curated features in the SRM investigation

As a targeted method, SRM quantifies a protein with fewer features than DIA and DDA, and most of the quantified features are of high quality. Thus, the evaluation of the proposed approach for SRM focused on individual uninformative features. As described in Section 3, the SRM investigation by Dunkley et al. manually curated the quality of each feature. We used this dataset to compare the proposed approach with the manual curation.

For the 246 proteins in the dataset, four proteins labeled as iRT standard and three quantified with only one feature were excluded from the evaluation. Among the remaining 239 proteins, 89 with inconsistent features were detected by the proposed approach as well as the manual curation. Supplementary Fig. S12 illustrates an example of protein P63104, where the features of peptide TAFDEAIAELETLSSEESYK had significant run-to-run variation, compared to those of peptides DIC[+57]NDVSLLEK and FLIPNASQAESK. Qualities of these features were also reflected on their chromatographic profiles and intensity proportions. In the case of peptide FLIPNASQAESK, the features were quantified in the expected range of retention time and had high consistency across runs (Supplementary Fig. S13). In contrast, the uninformative features of peptide TAFDEAIAELETLSSEESYK were inconsistent in terms of their chromatographic profiles, retention times and intensity propor-
The proposed approach detected uninformative features that were not detected by manual curation in 64 more proteins. In the case of protein P51532 (Supplementary Fig. S15a), the uninformative features of peptide AIEEGTLEEIEEEVR were only detected by the proposed approach. While these features were consistent among themselves, their profiles deviated from the majority of other features of the protein. It is therefore difficult to detect such an unwanted deviation with the per-peptide manual curation.

Moreover, 33 proteins were characterized as having uninformative features by the manual curation alone. This was particularly common in situations where a protein was targeted with two peptides with distinct quantitative profiles. When a consensus profile of the features was unclear or even nonexistent, the proposed approach had difficulties establishing a consensus protein profile. Among the 33 proteins with bad-quality features flagged by the manual curation alone, 32 were targeted with one or two peptides. In the case of protein Q9ULK0 with two peptides (Supplementary Fig. S15b), all features were manually labeled as bad-quality, because the endogenous peptides had inconsistent chromatographic profiles (e.g., GFSIDVLDALAK in Supplementary Fig. S16). Based on the log-intensities, however, the proposed approach was undermined by the lack of consensus protein profile, and did not detect any of these features as uninformative. For proteins with few quantified features, manual curation may still be necessary.

5.4 Appropriate data processing facilitated the selection of informative features

We evaluated the impact of data processing on the proposed selection of informative features and protein-level conclusions, using the DIA experiment in Selevsek et al. Two Skyline analyses were performed: the Full analysis considered all the peptide precursors and the LowCV analysis selected peptides that were consistently detectable and with a low CV. We also considered two types of q-filter strategies, on the basis of individual peaks (the Sparse q-filter) and entire LC-MS features (the 50% q-filter). Combinations of the two analyses and two q-filter strategies produced four datasets: Full_Sparse, Full_50%, LowCV_Sparse, and LowCV_50%. Full_Sparse was processed with less stringent criteria and had more proteins than the other three datasets. Fig. 6a illustrates the fact that the choice of data processing affected the detection of differentially abundant proteins.

In the case of protein YPL117C, Full_Sparse quantification had several noisy features and many missing values (Fig. 6b). Features of better consistency were reported with either the 50% q-filter (Full_50%, Fig. 6c) or low-CV analysis (LowCV_Sparse, Fig. 6d). This difference affected the reproducibility of detecting differential abundance. In the T2-T0 comparison, the all-features quantification gave inconsistent conclusions across these analyses, in terms of estimated log-fold change and adjusted p-value. Selecting informative features reduced this discrepancy. Moreover, for all the detected changes in T2-T0, the proposed approach enhanced the reproducibility across data processing options (Fig. 6e and Fig. 6f), especially for Full_50%, LowCV_Sparse and LowCV_50%. The remaining discrepancy with Full_Sparse was likely due to more missing values and noise in the dataset, which undermined the selection of informative features.
5.5 Relationship between the initial number of features and the performance of the proposed approach

Accurate estimation of representative patterns for each protein is key to the proposed selection of informative features, with different levels of challenges across proteins and across datasets. In the previous sections we saw that selecting informative features improved the detection of differentially abundant proteins (as compared to the all-features quantification) in the DIA benchmarks (Section 5.1), but had a limited benefit in the DDA benchmarks (Section 5.2). Using the benchmark datasets, we investigated whether the initial number of features impacted the performance of the proposed approach.

Supplementary Fig. S17 illustrates that the errors of the estimated log-fold changes decreased with an increase in the number of features. Compared to the all-features quantification, the proposed selection of informative features had a reduced or similar estimation error across proteins with different number of features in the DIA benchmarks (Supplementary Fig. S17a). The reduction in the estimation error was the most pronounced for proteins with 7-20 features. There was only a moderate improvement for proteins with 1-6 features, likely due to less reliable estimates of the protein patterns with fewer features. On the other hand, with a sufficient number of features (e.g., > 20), the TMP summarization in MSstats was less affected by outliers and uninformative features, and the proposed selection of informative features only resulted in marginal improvement.

For the DDA benchmarks, the impact of the proposed approach was moderate and appeared to be tool-dependent (Supplementary Fig. S17b). It did not result in substantial reduction in estimation error across proteins with different number of features and across the evaluated datasets, except for proteins with 7-20 features in the iPRG benchmark quantified with Progenesis. DDA experiments typically have larger technical variation than DIA, which makes it more challenging to obtain reliable estimates of protein patterns. Some proteins even lacked a consensus profile among the quantified features across all the conditions (Supplementary Fig. S10), and selecting informative features did not improve the performance. For proteins with > 20 features, the discrepancy across data processing and quantification options was minimized. Overall, the efficacy of the proposed approach depended on both the number of features and the data processing, as well as other factors such as technical variation.

6 Discussion

The proposed approach automatically detects uninformative features and outliers to be separately investigated, curated, or removed. For estimation and detection of differential abundance, the uninformative features can be confidently removed in a dataset with a large number of good-quality features (e.g., DIA), where a detailed and efficient curation may not be possible. In such cases the proposed approach was robust to the uninformative features and outliers (Fig. 2), increased the estimation accuracy (Fig. 3) and the sensitivity of detecting true changes (Supplementary Fig. S5 and Fig. S6), and enhanced the reproducibility of the results across data processing tools (Fig. 4).

The proposed selection of informative features relies on accurate estimation of representative patterns for each protein. DDA workflows typically quantify much fewer features, and do this in a less consistent manner, than DIA. This makes it challenging to obtain a reliable estimate of protein
abundance. While the proposed approach is useful in detecting uninformative features and outliers (Fig. 5), and helps direct the search for poor-quality data, further evaluation of the detected features is recommended. The evaluation may be based on manual curation, which is more manageable compared to examining the entire dataset. It may also be useful to obtain additional evidence using other quality metrics. For SRM datasets, special attention is needed as a consensus profile may not be existent from the data if the targeted peptides have distinct quantitative profiles.

It should be noted that the proposed approach cannot completely eliminate the impact of poor experimental design (e.g., lack of randomization) or flawed data processing (e.g., inappropriate normalization).

The proposed selection of informative features assumes that features of the same protein share a representative pattern. This assumption may be violated due to biological reasons such as polymorphisms, alternative splice variants, or PTMs. For quantitative analysis of PTMs, it is more suitable to proceed with site-level quantification and inference. While the proposed approach is applicable for the site-level analysis, it would require a sufficient number of features for each PTM site, in order to accurately estimate the representative patterns.

The proposed approach evaluates information content for individual features. For design of targeted assay, instead of focusing on individual features, the approach can be extended to detect peptides whose quantitative profiles deviate from the patterns of the targeted protein. In label-based experiments, the proposed approach analyzes endogenous and reference peptides separately. When both endogenous and reference features are required, it is possible to integrate both counterparts, prior to applying the proposed approach as in label-free analysis.

Combining information from multiple sources is a common strategy in many statistical and machine learning practices, which may as well improve the selection of informative features. The goal to distinguish between informative features and uninformative features without a ground-truth is an ill-posed classification problem, and supervised methods for classification are not directly applicable. However, methods for unsupervised learning may be well suited for facilitating the estimation of protein profile, which can be framed as a clustering problem. When manual curation is performed on a subset of the data, semi-supervised methods may further improve the detection of uninformative features, by using the curated features to train predictors with learned representations.

Acknowledgments

This work was supported in part by NSF DBI-1759736 and NIH NLM 1R01LM013115-01 to O.V. We thank Erik Verschueren (University of California, San Francisco), Oliver M. Bernhardt (Biognosys), Eduard Sabidó (Centre for Genomic Regulation, The Barcelona Institute of Science and Technology), and the group of Alexey I. Nesvizhskii (University of Michigan) for processing the controlled mixture datasets. We also thank the maintainers of MassIVE.quant (the group of Nuno Bandeira at University of California, San Diego), in which the experimental datasets are deposited.
References

[1] Abbatiello, S. E., Mani, D. R., Keshishian, H., and Carr, S. A. Automated detection of inaccurate and imprecise transitions in peptide quantification by multiple reaction monitoring mass spectrometry. *Clinical Chemistry* 56, 2 (2010), 291–305.

[2] Aebersold, R., and Mann, M. Mass spectrometry-based proteomics. *Nature* 422, 6928 (2003), 198–207.

[3] Aebersold, R., and Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* 537, 7620 (2016), 347–355.

[4] Bruderer, R., Bernhardt, O. M., Gandhi, T., Miladinović, S. M., Cheng, L.-Y., Messner, S., Ehrenberger, T., Zanotelli, V., Butscheid, Y., Escher, C., Vitek, O., Rinner, O., and Reiter, L. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Molecular & Cellular Proteomics* 14, 5 (2015), 1400–1410.

[5] Choi, M., Carver, J., Chiva, C., Tzouros, M., Huang, T., Tsai, T.-H., Pullman, B., Bernhardt, O. M., Hüttenhain, R., Teo, G. C., Pavlou, M., Verschueren, E., Wollscheid, B., Nesvizhskii, A. I., Reiter, L., Dunkley, T., Sabidó, E., Bandeira, N., and Vitek, O. MassIVE.quant: a community resource of curated quantitative mass spectrometry-based proteomics datasets. Submitted.

[6] Choi, M., Eren-Dogu, Z. F., Colangelo, C., Cottrell, J., Hoppmann, M. R., Kapp, E. A., Kim, S., Lam, H., Neubert, T. A., Palmblad, M., Phinney, B. S., Weintraub, S. T., MacLean, B., and Vitek, O. ABRF Proteome Informatics Research Group (iPRG) 2015 study: detection of differentially abundant proteins in label-free quantitative LC-MS/MS experiments. *Journal of Proteome Research* 16, 2 (2017), 945–957.

[7] Choi, M., Tsai, T.-H., Huang, T., Shulman, N., Bernhardt, O. M., Nesvizhskii, A. I., Reiter, L., Sabidó, E., MacLean, B. X., and Vitek, O. MSstats maximizes the reproducibility of detecting differentially abundant proteins across tools that process raw mass spectra. Submitted.

[8] Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & Cellular Proteomics* 13, 9 (2014), 2513–2526.

[9] Cox, J., and Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* 26, 12 (2008), 1367–1372.

[10] Dunkley, T., Costa, V., Friedlein, A., Lugert, S., Aigner, S., Ebeling, M., Miller, M. T., Patsch, C., Piraino, P., Cutler, P., and Jagasia, R. Characterization of a human pluripotent stem cell-derived model of neuronal development using multiplexed targeted proteomics. *Proteomics Clinical Applications* 9, 7–8 (2015), 684–694.

[11] Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & Cellular Proteomics* 11, 6 (2012), O111.016717.
[12] Hoaglin, D. C., Mosteller, F., and Tukey, J. W. Understanding Robust and Exploratory Data Analysis. Wiley, 1983.

[13] Huber, P. J. Robust estimation of a location parameter. *Annals of Mathematical Statistics 35*, 1 (1964), 73–101.

[14] Keller, A., Bader, S. L., Shteynberg, D., Hood, L., and Moritz, R. L. Automated validation of results and removal of fragment ion interferences in targeted analysis of data-independent acquisition mass spectrometry (MS) using SWATHProphet. *Molecular & Cellular Proteomics 14*, 5 (2015), 1411–1418.

[15] MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., Kern, R., Tabb, D. L., Liebler, D. C., and MacCoss, M. J. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics 26*, 7 (2010), 966–968.

[16] Navarro, P., Kuharev, J., Gillet, L. C., Bernhardt, O. M., MacLean, B., Röst, H. L., Tate, S. A., Tsou, C.-C., Reiter, L., Distler, U., Rosenberger, G., Perez-Riverol, Y., Nesvizhskii, A. I., Aebersold, R., and Tenzer, S. A multicenter study benchmarks software tools for label-free proteome quantification. *Nature Biotechnology 34*, 11 (2016), 1130–1136.

[17] Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S., and Smyth, G. K. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Annals of Applied Statistics 10*, 2 (2016), 946–963.

[18] Picotti, P., and Aebersold, R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nature Methods 9*, 6 (2012), 555–566.

[19] Reiter, L., Rinner, O., Picotti, P., Hüttenhain, R., Beck, M., Brusniak, M.-Y., Hengartner, M. O., and Aebersold, R. mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nature Methods 8*, 5 (2011), 430–435.

[20] Röst, H. L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S. M., Schubert, O. T., Wolski, W., Collins, B. C., Malmström, J., Malmström, L., and Aebersold, R. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology 32*, 3 (2014), 219–223.

[21] Searle, B. C., Pino, L. K., Egertson, J. D., Ting, Y. S., Lawrence, R. T., MacLean, B. X., Villén, J., and MacCoss, M. J. Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature Communications 9*, 1 (2018), 5128.

[22] Selevsek, N., Chang, C.-Y., Gillet, L. C., Navarro, P., Bernhardt, O. M., Reiter, L., Cheng, L.-Y., Vitek, O., and Aebersold, R. Reproducible and consistent quantification of the Saccharomyces cerevisiae proteome by SWATH-mass spectrometry. *Molecular & Cellular Proteomics 14*, 3 (2015), 739–749.

[23] Sharma, V., Eckels, J., Schilling, B., Ludwig, C., Jaffe, J. D., MacCoss, M. J., and MacLean, B. Panorama Public: A public repository for quantitative data sets processed in Skyline. *Molecular & Cellular Proteomics 17*, 6 (2018), 1239–1244.
[24] Smyth, G. K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3 (2004), Article3.

[25] Teo, G., Kim, S., Tsou, C.-C., Collins, B., Gingras, A.-C., Nesvizhskii, A. I., and Choi, H. mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. *Journal of Proteomics* 129 (2015), 108–120.

[26] Tsou, C.-C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A.-C., and Nesvizhskii, A. I. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature Methods* 12, 3 (2015), 258–264.

[27] Vaca Jacome, A. S., Peckner, R., Shulman, N., Krug, K., DeRuff, K. C., Officer, A., MacLean, B., MacCoss, M. J., Carr, S. A., and Jaffe, J. D. Avant-garde: An automated data-driven DIA data curation tool. *bioRxiv* (2019).

[28] Venable, J. D., Dong, M.-Q., Wohlschlegel, J., Dillin, A., and Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods* 1, 1 (2004), 39–45.

[29] Zhang, B., Pirmoradian, M., Zubarev, R., and Käll, L. Covariation of peptide abundances accurately reflects protein concentration differences. *Molecular & Cellular Proteomics* 16, 5 (2017), 936–948.
Figures and tables

![Feature Selection](image1)

![Protein-level Summarization](image2)

**Figure 1:** Overview of the proposed feature selection approach, illustrated with the case of protein P32915 from the DDA iPRG experiment, analyzed by MaxQuant. (a) Four major steps, and their intermediate results: detection of low-coverage features, estimation of representative patterns of the protein, detection of outlying log-intensities, and detection of noisy features. The detection of low-coverage features is detailed in Section 4.2. The remaining steps are detailed in Section 4.3. The example protein has five features quantified across 12 runs, where feature TWIEISGTPR_2 (highlighted in dark gray in the first panel) has six observed log-intensities, which are significantly fewer than expected. The feature is labeled as low-coverage, and is excluded. The remaining four features are used to estimate the representative protein profile and variation, as highlighted in yellow in the second panel. The log-intensity from feature GQIGIYPK_2 and run 12 significantly deviates from the estimated patterns, and is labeled as an outlier (highlighted in dark gray in the third panel). Finally, feature GQIGIYPK_2 (highlighted in dark gray in the fourth panel) exhibits substantial variation between the runs, and is labeled as a noisy feature. (b) Application of the proposed feature selection approach for protein-level summarization and statistical inference. The profile plots of the protein with all the features and with the selected informative features are shown in the first two panels, where the detected outlier and uninformative features (TWIEISGTPR_2 and GQIGIYPK_2) are removed in the second panel. The informative features are used as input to perform subsequent statistical analyses. The third panel shows the results of protein-level summarization, using the informative features (shown in light gray, solid lines), or using all the features that also include the uninformative features (shown in dashed lines). The protein-level summary with the proposed approach and that with all the features are shown in yellow and dark gray, respectively.
Figure 2: Performance of the proposed feature selection approach in the DIA benchmarks. (a) Number of features per protein, before and after the proposed selection of informative features. (b)-(c) Example protein C8ZIG9 from the Navarro benchmark, quantified with (b) Skyline and (c) Spectronaut, highlights representative patterns of uninformative features, and their impact on the relative protein quantification. The plots show the features before and after the selection of informative features. They contrast the protein-level summaries by the proposed approach to those with all the features and with the top-3 features. The table below each plot summarizes the relative protein quantification in terms of the estimate of log-fold change (FC), its standard error, and FDR-adjusted \( p \)-value, as determined by MSstats. Some low-intensity outliers are due to missing values, which are not considered by the proposed approach.
Figure 3: Performance evaluation for the background proteins (i.e., true log₂-fold change between conditions is zero) in the DIA benchmarks. (a) The absolute errors, i.e., deviation of the estimated log-fold change from the truth, for all background proteins in the Bruderer benchmark (left panel) and the Navarro benchmark (right panel), quantified with DIA Umpire, OpenSWATH, Skyline or Spectronaut. The proposed selection of informative features was compared to the all-features and top-n quantifications. Smaller values indicate better performance. (b) The standard errors associated with the log-fold change estimates. Smaller values indicate better performance.
Figure 4: Euler diagrams for detected true changes across data processing tools in (a) the Bruderer benchmark and (b) the Navarro benchmark, using the proposed selection of informative features, and the all-features and top-\(n\) quantifications.
Figure 5: Performance of the proposed feature selection approach in the DDA benchmarks. (a) Number of features per protein, before and after the proposed selection of informative features. (b)-(c) Example protein P32898 from the iPRG benchmark, quantified with (b) Progenesis and (c) Skyline, highlights representative patterns of uninformative features, and their impact on the relative protein quantification. Progenesis quantified the protein with only four features, including two with inconsistent, high-intensity profile in runs 4-6. On the other hand, the Skyline analysis included more features, and majority of them formed a consistent pattern.
Figure 6: Impact of data processing options on the proposed selection of informative features, and the relative protein quantification in the four Selevsek datasets. (a) Number of total proteins and number of detected changes over time with the all-features quantification. (b) Profile plot for protein YPL117C in Full_Sparse, where uninformative features are shown in dashed lines and outliers are depicted with open circles. Protein-level summaries with the proposed approach (yellow) and the all-features quantification (dark gray) are shown on top of quantified features. The table summarizes the relative protein quantification of T2-T0, with both approaches. (c)-(d) Same as in (b), but for the datasets Full_50% and LowCV_Sparse. (e) Euler diagrams for detected changes in the T2-T0 comparison across the four datasets, using all the features. (f) Same as in (e), but using the selected informative features.
Table 1: Summary of the experimental datasets of benchmark controlled mixtures.

| Benchmark (MassIVE ID) | Evaluation | No. of conditions | No. of runs | Data processing          | No. of proteins | Median no. of features per protein | MassIVE reanalysis ID |
|------------------------|------------|-------------------|-------------|--------------------------|-----------------|------------------------------------|----------------------|
| DDA Choi [7] (MSV000084181) | True FC    | 5                 | 15          | MaxQuant, Progenesis, Proteome Discoverer, Skyline | 1,549           | 10                                 | A: RMSV000000261      |
|                         |            |                   |             |                          | 1,330           |                                    | S: RMSV000000263      |
|                         |            |                   |             |                          | 1,271           |                                    |                      |
|                         |            |                   |             |                          | 1,244           |                                    |                      |
| DDA iPRG [6] (MSV000079843) | True FC    | 4                 | 12          | MaxQuant, Progenesis, Skyline | 2,875           | 7                                  | A: RMSV000000249      |
|                         |            |                   |             |                          | 3,015           |                                    | S: RMSV000000259      |
|                         |            |                   |             |                          | 3,027           |                                    |                      |
| DIA Bruderer [4] (MSV000081828) | True FC    | 7                 | 21          | DIA Umpire, Skyline, Spectronaut | 3,611           | 25                                 | A: RMSV000000252      |
|                         |            |                   |             |                          | 3,076           |                                    | S: RMSV000000258      |
|                         |            |                   |             |                          | 3,134           |                                    |                      |
| DIA Navarro [16] (MSV000081024) | True FC    | 2                 | 6           | DIA Umpire, OpenSWATH, Skyline, Spectronaut | 4,383           | 18                                 | A: RMSV000000250      |
|                         |            |                   |             |                          | 5,376           |                                    | S: RMSV000000257      |
|                         |            |                   |             |                          | 5,387           |                                    |                      |
|                         |            |                   |             |                          | 5,486           |                                    |                      |

Table 2: Summary of the experimental datasets from biological investigations. Data from the investigation by Dunkley et al. are available on Panorama (https://panoramaweb.org/neuroSRM.url).

| Investigation (MassIVE ID) | Evaluation        | No. of conditions | No. of runs | Data processing | Feature_q-filter | No. of proteins | Median no. of features per protein | MassIVE reanalysis ID |
|---------------------------|-------------------|-------------------|-------------|-----------------|-----------------|-----------------|------------------------------------|----------------------|
| DIA Selevsek [22] (MSV000081677) | Low-CV analysis   | 6                 | 18          | Skyline         | Full_Sparse     | 2,189           | 24                                 | A: RMSV000000251      |
|                           |                   |                   |             |                 | Full_50%        | 1,821           |                                    | S: RMSV000000262      |
|                           |                   |                   |             |                 | LowCV_Sparse    | 2,212           |                                    |                      |
|                           |                   |                   |             |                 | LowCV_50%       | 2,204           |                                    |                      |
| SRM Dunkley [10]          | Manual curation   | 10                | 95          | Skyline         | Full            | 246             | 6                                  |                      |
Table 3: Results of protein significance analysis at the FDR cutoff of 0.05. Performances of the all-features and proposed methods in terms of sensitivity and specificity were evaluated on the DIA benchmark controlled mixtures. Sensitivity is calculated as TP/(TP + FN), where TP is the number of true positives and FN is the number of false negatives. Specificity is calculated as TN/(FP + TN), where FP is the number of false positives and TN is the number of true negatives.

| Benchmark and data processing | Quantification method | No. of comparisons | No. of changes | TP  | FP  | Sensitivity | Specificity |
|------------------------------|-----------------------|--------------------|----------------|-----|-----|-------------|-------------|
| DIA Broeder                 | All-features          | 72,261             | 241            | 170 | 53  | 0.705       | 0.999       |
|                             | Proposed              | 72,176             | 241            | 177 | 79  | 0.734       | 0.999       |
| Skyline                     | All-features          | 63,946             | 252            | 171 | 42  | 0.679       | 0.999       |
|                             | Proposed              | 63,935             | 252            | 177 | 23  | 0.702       | 1.000       |
| Spectronaut                 | All-features          | 65,803             | 252            | 191 | 122 | 0.758       | 0.998       |
|                             | Proposed              | 65,803             | 252            | 207 | 79  | 0.821       | 0.999       |

| Benchmark and data processing | Quantification method | No. of comparisons | No. of changes | TP  | FP  | Sensitivity | Specificity |
|------------------------------|-----------------------|--------------------|----------------|-----|-----|-------------|-------------|
| MaxQuant                     | All-features          | 15,486             | 224            | 188 | 76  | 0.839       | 0.995       |
|                             | Proposed              | 15,486             | 224            | 192 | 94  | 0.857       | 0.994       |
| Progenesis                   | All-features          | 13,300             | 216            | 190 | 53  | 0.880       | 0.996       |
|                             | Proposed              | 13,300             | 216            | 201 | 52  | 0.931       | 0.996       |
| Proteome                     | All-features          | 12,542             | 216            | 187 | 104 | 0.866       | 0.992       |
| Discoverer                   | Proposed              | 12,538             | 216            | 185 | 120 | 0.856       | 0.990       |
| Skyline                      | All-features          | 12,440             | 192            | 169 | 96  | 0.880       | 0.992       |
|                             | Proposed              | 12,440             | 192            | 172 | 98  | 0.896       | 0.992       |

Table 4: Results of protein significance analysis at the FDR cutoff of 0.05. Performances of the all-features and proposed methods in terms of sensitivity and specificity were evaluated on the DDA benchmark controlled mixtures.