MSstatsPTM: Statistical relative quantification of post-translational modifications in bottom-up mass spectrometry-based proteomics

Devon Kohler[1], Tsung-Heng Tsai[2], Erik Verschueren[4], Ting Huang[1], Trent Hinkle[3], Lilian Phu[3], Meena Choi*[3], and Olga Vitek*[1]

[1]Khoury College of Computer Science, Northeastern University, Boston, MA, USA
[2]Kent State University, Kent, OH, USA
[3]MPL, Genentech, South San Francisco, CA, USA
[4]ULUA BV, Arendstraat 29, 2018 Antwerp, Belgium
*Corresponding Authors

Abstract

Liquid chromatography coupled with bottom up mass spectrometry (LC-MS/MS)-based proteomics is increasingly used to detect changes in post-translational modifications (PTMs) in samples conditions. Analysis of data from such experiments faces numerous statistical challenges. These include the low abundance of modified proteoforms, the small number of observed peptides that span modification sites, and confounding between changes in the abundance of PTM and the overall changes in the protein abundance. Therefore, statistical approaches for detecting differential PTM abundance must integrate all the available information pertaining to a PTM site, and consider all the relevant sources of confounding and variation. In this manuscript we propose such a statistical framework, which is versatile, accurate, and leads to reproducible results. The framework requires an experimental design, which quantifies, for each sample, both peptides with post-translational modifications and peptides from the same proteins with no modification sites. The proposed framework supports both label-free and tandem mass tag (TMT)-based LC-MS/MS acquisitions. The statistical methodology separately summarizes the abundances of peptides with and without the modification sites, by fitting separate linear mixed effects models appropriate for the experimental design. Next, model-based inferences regarding the PTM and the protein-level abundances are combined to account for the confounding between these two sources. Evaluations on computer simulations, a spike-in experiment with known ground truth, and three biological experiments with different organisms, modification types and data acquisition types demonstrate the improved fold change estimation and detection of differential PTM abundance, as compared to currently used approaches. The proposed framework is implemented in the free and open-source R/Bioconductor package $MSstatsPTM$.

Introduction

Signaling mechanisms allow cells to mount a fast and dynamic response to a multitude of biomolecular events. Signaling is facilitated by the modification of proteins at specific residues, acting as molecular on/off switches [1, 2, 3]. Characterizing relative abundance of a modification site's occupancy repertoire across experimental conditions provides important insights [4]. For example, meaningful patterns of changes in post-translational modifications (PTMs) abundance can serve as biomarkers of a disease [5]. Alternatively, distinguishing the quantitative changes in a PTM from the overall changes of the protein abundance helps gain insight into biological and physiological processes operating on a very short timescale [6, 7, 8]. This helps to distinguish between relative site occupancy changes at steady-state protein levels, typical for short time-scale signaling events, and observed relative changes of PTMs as a result of underlying gene expression or protein abundance levels.

Bottom-up liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) is a tool of choice for unbiased and large-scale identification and quantification of proteins and their PTMs [9, 10]. However, LC-MS-based interrogation of the modified proteome is challenging, for a number of reasons. First, the relatively lower abundance of modified proteoforms dictates that a global interrogation can only be achieved through large-scale enrichment protocols with modification-specific antibodies or beads [11]. Variability in the enrichment efficiency inevitably affects the reproducibility of the number of spectral features (e.g., peptide precursor ions or their fragments) and their intensities. Second, contrary to the often large number of identified peptides that can be used to quantify protein abundance, there are relatively few representative peptides that span a modification site, and there may be multiple modified sites on a single peptide [4]. Third, unless early signaling events are interrogated, the interpretation of the relative changes in modification occupancy are inherently confounded with changes in the overall protein abundance, complicating the interpretation of the results [6, 12]. Finally, technological aspects of bottom-up MS experiments, such as presence of labeling by tandem mass tag (TMT), introduce additional sources of uncertainty and variation.

The technological difficulties in PTM identification and quantification increase the uncertainty and the variation in the data, and challenge the downstream statistical analyses. Frequently data from these experiments are analyzed using statistical methods that were not originally designed for this task. Researchers use methods such as *t*-test[13], Analysis of Variance (ANOVA)[14], or Limma[15], by taking as input the intensity ratios of modified and unmodified peptide features, and comparing the mean abundance of different PTM sites. Such approaches do not fully account for all the sources of uncertainty. As the result, these approaches are either not directly applicable to experiments with non-trivial designs (such as experiments with multiple conditions, paired and time course designs, and experiments with labeling), or require the analysts to exercise non-trivial statistical expertise.

This manuscript proposes a versatile statistical analysis framework that accurately detects relative changes in post-translational modifications. The framework requires an experimental design, which quantifies, for each sample, both the peptides with post-translational modifications, and peptides from the same proteins with no modification sites. The framework supports data-dependent acquisitions (DDA) that are label-free or tandem mass tag (TMT)-based. The statistical methodology separately summarizes the abundances of peptides with and without the modification sites, and fits separate linear mixed effects models that reflect the biological and technological aspects of the experimental design. Next, model-based inferences regarding the PTM and the unmodified protein-level abundances are combined to account for the confounding between these two sources.

We evaluated the proposed framework on two datasets from computer simulations, one benchmark controlled mixture, and three biological investigations. The datasets illustrate a diverse set of organisms, modification types, acquisition methods and experimental designs, showing the applicability of the framework to a variety of situations. By appropriately leveraging the information from the unmodified proteins, the proposed approach improved the accuracy of the estimates of PTM fold changes, and produced a better calibrated false positive rate of detecting differentially abundant PTMs as compared to existing methods. In particular, accounting for the confounding from unmodified protein abundance allowed us to characterize the true effect of the modification, avoiding the need for more manual and time intensive follow-up investigation.

The proposed approach is implemented as a freely available open source R package *MSstatsPTM*, as part of the *MSstats* family of packages [16, 17], and is available on Bioconductor.

Experimental procedures

Data overview and availability

Table 5.1 summarizes the experiments. Two computer simulations had known ground truth, and varied in experimental realism. The first simulation produced a perfectly clean dataset, with many replicates and no missing values. The second simulation introduced real-world characteristics, such as limited modified features and missing values. Details of computer simulations are available in **Supplementary Sec. 2.1** and on GitHub (`https://github.com/devonjkohler/MSstatsPTM_simulations`).

One spike-in experiment also had known changes in modified spike-in peptides, but had real-world experimental characteristics. Finally, three biological experiments demonstrated the applicability of the proposed approach across different biological organisms, modifications, experimental designs and acquisition strategies. The experimental data, R scripts with *MSstatsPTM* analysis, and results of the statistical analysis are available in MassIVE.quant (`https://massive.ucsd.edu/ProteoSAFe/static/massive-quant.jsp`) [18].

Dataset 1: Computer simulation 1 - Label-free clean

Simulation design: The simulation represented an idealistic case. 24 synthetic label-free datasets were generated with different experimental designs and different biological variation. In each dataset, 1,000 proteins had 10 unmodified features per protein. Each of the 1,000 proteins had one PTM. Each PTM was represented by 10 modified features. The PTMs of 500 proteins had a differential fold change between conditions, while the other 500 proteins were generated with no changes in abundance between conditions. Furthermore, the fold changes of half of the 500 differential PTMs were fully masked by changes in the unmodified protein. Finally, the fold change of half the 500 non-differential PTMs was entirely due to changes in the unmodified protein. All the differential PTMs were generated with an expected log base 2 fold change of 0.75 between conditions.

Each simulation was generated with random biological variation. The observed peptide abundances were simulated by adding random noise $\mathcal{N}(0, \sigma^2)$ to the deterministic abundances described above. Two values $\sigma^2 = \{.2, .3\}$ were motivated by the experimental datasets in this manuscript.

Evaluation: We evaluated the ability of the statistical methods to correctly detect differentially abundant PTMs. We gauged the methods ability to avoid false positives (i.e. specificity), accurately estimate the fold change between conditions, and analyzed the sensitivity of detecting differentially abundant PTMs. The evaluation was performed both in the presence of confounding with changes in the unmodified protein and after applying adjustment to correct for the confounding.

Dataset 2: Computer simulation 2 - Label-free with few low feature counts and missing values

Simulation design: The data were simulated as above, while providing a more realistic representation of the experiments. The feature counts and the proportion of missing values were as observed on average over all the the experimental datasets in this manuscript. Specifically, PTMs were simulated with 2 modified peptide features, and unmodified proteins were simulated with 10 features. Additionally, 20% of observations for both modified and unmodified peptides were missing completely at random.

Evaluation: The methods were evaluated as above. We evaluated their ability to correctly detect PTM's specificity, fold change estimation, and sensitivity. These statistics were analyzed both in the presence of, and without, confounding with the overall changes in protein abundance.

Dataset 3: Spike-in benchmark - Ubiquitination - Label-free

Experimental design: Figure 1(a) overviews the experimental design. Four mixtures (i.e., conditions) were created with varying amounts of human lysate, background *E. Coli* lysate, and human spike-in ub-peptide mixture. Unmodified peptides from human lysate were viewed as the global proteome. Background *E. coli* lysate were used to equalize total protein levels. 50 heavy-labeled KGG motif peptides from 20

human proteins were spiked into the mixed background of the lysates. Quantitative changes in protein and site abundance of these 20 human proteins were the target of the benchmark. In particular, we distinguished the unadjusted changes (i.e. changes in the abundances of the modified peptides) and the protein-level adjusted changes of (i.e., changes in the abundances of the modified peptides relative to the changes in the abundances of the human lysate). The true log-fold changes between the relevant components of the relevant mixtures are summarized in **Figure 1**(b). Two replicate mixtures were created per condition.

Data acquisition: Each mixture was analyzed with KGG enrichment, and without KGG enrichment (i.e., in a global profiling run), with label-free LC-MS/MS. There was a 90.2% overlap of protein identifications between the identified background modified peptides and proteins quantified in the global profiling run.

Evaluation: We expect the relative abundances of the spike-in peptides to change as in **Figure 1**(b). The changes in peptide abundances in all the comparisons except Mix 4 vs Mix 1 were distinct from changes in the global proteome abundances and distinct from zero, and were viewed as positive controls. In the comparison of Mix4 vs Mix 1 both the modified peptides and the global proteome background changed two-fold, and as the result the peptides in this comparison were viewed as a negative control. The background *E. Coli* lysate peptides were not expected to change in abundance in comparison, after accounting for adjustment, and were viewed as additional negative controls. We evaluated the statistical methods ability to avoid false positives, as well as their sensitivity in detecting the differentially abundant spike-in peptides and accurately estimate their expected fold change.

Dataset 4: Human - Ubiquitination - 1mix-TMT

Experimental Design: Luchetti et al. [19] profiled human epithelial cells engineered to express IpaH7.8 under a dox inducible promoter. Uninfected cells were measured at 0 and 6 hours, while cells infected with *Shigella Flexneri* (*S. Flexneri*) bacteria were measured at 1, 2, 4, and 6 hour increments, resulting in six total conditions. 11 samples were allocated to 1 TMT mixture in an unbalanced repeated measure design. All conditions had two biological replicates except for the Dox1hr condition, which was allocated one replicate.

Data acquisition: The ubiquitinated peptides, and the total proteome (i.e., global profiling) were each conducted in a single LC-MS/MS run. There was a 95% overlap between the identified modified peptides and proteins that were quantified in the global profiling run.

Evaluation: We evaluated the ability of the statistical methods to detect changes in the abundance of modified peptides both before and after adjusting for changes in global protein abundance. The six condition were labeled Dox1hr, Dox2hr, Dox4hr, Dox6hr, NoDox0hr, and NoDox6hr. All conditions were compared with each other, resulting in 15 pairwise comparisons. Since the dataset was a biological investigation, the true positive modifications were unknown. Shigella ubiquitin ligase IpaH7.8 was shown to function as an inhibitor of the protein Gasdermin D (GSDMD). GSDMD was actively degraded when IpaH7.8 expression was induced by dox treatment in human cells. We expect IpaH7.8 to function as an inhibitor of GSDMD in the global profiling run.

Dataset 5: Mouse - Phosphorylation - 2mix-TMT

Experimental Design: Maculins et al. [20] studied primary murine macrophages infected with *S. Flexneri*. The experiment quantified the abundance of total protein and of phosphorylation in wild type (WT), and in ATG16L1-deficient (cKO) samples, uninfected and infected with *S. Flexneri*. The abundance of total protein and post-translation modifications were quantified at three time points, uninfected, early infection (45-60 minutes), and late infection (3-3.5 hours). 22 biological samples were allocated to 2 TMT mixtures in an unbalanced repeated measure design, with 11 samples allocated to each mixture. 16 replicates were spread equally between the early and late WT and cKO conditions, resulting in four replicates per condition. Both the uninfected WT and cKO contained 3 replicates, with mixture one allocating one replicate to uninfected WT and two replicates to uninfected cKO. Conversely, mixture two contained one replicate of uninfected cKO and two uninfected WT.

Data acquisition: This experiment included a total proteome (i.e., a global profiling run) and a phosphopeptide enrichment run. There was a 90% overlap between the identified modified peptides and proteins that were quantified in the global profiling run.

Evaluation: We evaluated the ability of the statistical methods to detect changes in the abundance of modified peptides both before and after adjusting for changes in global protein abundance. The six condition were labeled KO_Uninfect, KO_Early, KO_Late, WT_Uninfect, WT_Early, and WT_Late. 9 total comparisons were made, namely KO_Early-WT_Early, KO_Late-WT_Late, KO_Uninfected-WT_Uninfected, KO_Early-KO_Uninfected, KO_Late-KO_Uninfected, WT_Early-WT_Uninfected, WT_Late-WT_Uninfected, Infected-Uninfected, and KO-WT. Since the dataset was a biological investigation, the true positive modifications were unknown.

Dataset 6: Human - Ubiquitination - Label-free no global profiling run

Experimental Design: Cunningham et al. [21] investigated the relationship between USP30 and protein kinase PINK1, and their association with Parkinson's Disease. The experiment profiled ubiquitination sites, and analyzed changes in the modified site abundance. The experiment had four conditions, CCCP, USP30 over expression (USP30 OE), Combo, and Control. Cell lines were used to create two biological replicates per condition. The abundance of modified peptides was quantified with label-free LC-MS/MS.

Data acquisition: This experiment did not include a separate global profiling run to measure unmodified peptides. In addition to low feature counts for unmodified peptides, this lead to substantially fewer matches between modified and unmodified peptides. There was a 41.9% overlap between the identified background modified peptides and proteins that were quantified in the global profiling run.

Evaluation: We evaluated the ability of the statistical methods to detect changes in the abundance of modified peptides both before and after adjusting for changes in global protein abundance. All the conditions were compared with each other in a full pairwise comparison, resulting in 6 comparisons. Since the dataset is a biological investigation, the true positive modifications were unknown.

Background

Goals of PTM characterization, input to statistical analyses, and notation

Consider a label-free LC-MS/MS experiment in the special case of a balanced design with I conditions and J biological replicates per condition. For simplicity, we assume that the experiment has no technical replicates, such that each biological replicate is represented by a single LC-MS/MS run. **Figure 2** schematically illustrates this data structure for one protein and one PTM site, $I = 2$ and $J = 2$. For one protein, the PTM site is represented by K spectral features (i.e., peptide ions, distinguished by their cleavage residues and charge states). The number of modified and unmodified features typically varies across proteins. Some log-intensities can be outliers, and some spectral features can be missing. The $\log_2$-intensity of Feature k, in Replicate j of Condition i is denoted by y^*_{ijk}. Conversely, the unmodified protein is represented by L spectral features, and the log-intensity of Feature l from the unmodified protein in the same run is denoted by y_{ijl}. The features can be quantified as part of a same mass spectrometry run, or in a separate enrichment and global proteome profiling run.

The population quantity of interest is the difference between the $\log_2$ abundances of a PTM site in Condition i and Condition i', denoted by μ^*_i and $\mu^*_{i'}$ respectively. We are interested in testing the null hypothesis

$$H_0 : \Delta_{PTM} = \mu^*_i - \mu^*_{i'} = 0 \text{ vs } H_a : \Delta_{PTM} = \mu^*_i - \mu^*_{i'} \neq 0 \quad (1)$$

Unfortunately, this population quantity is inherently confounded with the overall changes in protein abundance. To account for this, it is advantageous to consider a different null hypothesis:

$$H_0 : \Delta_{adj} = (\mu^*_i - \mu_i) - (\mu^*_{i'} - \mu_{i'}) = 0 \text{ vs } H_a : \Delta_{adj} = (\mu^*_i - \mu_i) - (\mu^*_{i'} - \mu_{i'}) \neq 0 \quad (2)$$

where μ_i and $\mu_{i'}$ reflect the overall $\log_2$ protein abundances in Condition i and Condition i'. These quantities are estimated using protein features with and without the modification site.

Existing statistical methods for detecting differentially abundant PTMs

ANOVA on summarized modified $\log_2$-intensities

Analysis of Variance (ANOVA) [22] is the simplest statistical model for summarized modified features in each biological replicate. The summarization often consists of averaging (or taking the median or other robust summary) of the $\log_2$ intensities of the modified features in each replicate, e.g. $\hat{y}^*_{ij} = \sum_{k=1}^{K} y^*_{ijk}/K$. Alternatively, summarization sums the intensities of the modified features on the original scale, and then takes the log

$$\hat{y}^*_{ij} = \log_2 \left(\sum_{k=1}^{K} 2^{y^*_{ijk}} \right) \tag{3}$$

The basic ANOVA model is then

$$\hat{y}^*_{ij} = \mu^*_i + \epsilon^*_{ij},\ \epsilon^*_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^{*2}),\ i = 1, \ldots, I,\ j = 1, \ldots, J \tag{4}$$

The model allows us to estimate $\hat{\Delta}$ and its standard error. The estimates are used to test the null hypothesis in Eq. (1) by comparing the model-based test statistic against the Student distribution with $df = I(J-1)$ degrees of freedom in balanced designs. Unfortunately, this approach is fundamentally flawed as it does not account for the confounding between changes in the PTM abundance and the overall changes in the abundance of the unmodified protein.

ANOVA based on ratios of modified and unmodified $\log_2$-intensities

The basic ANOVA can be extended to account for the confounding of changes in PTM abundance and overall changes in protein abundance [23, 24, 25]. Typically this is done by first calculating sums of the intensities of the modified and unmodified features on the original scale, and then considering replicate-wise ratios of the sums and taking the log

$$u_{ij} = \log_2 \frac{\sum_{k=1}^{K} 2^{y^*_{ijk}}}{\sum_{l=1}^{L} 2^{y_{ijl}}} = \log_2 \left(\sum_{k=1}^{K} 2^{y^*_{ijk}} \right) - \log_2 \left(\sum_{l=1}^{L} 2^{y_{ijl}} \right) \tag{5}$$

The approach then models these values with the basic ANOVA, which corresponds to

$$u_{ij} = (\mu^*_i - \mu_i) + \epsilon'_{ij}, \text{where } \epsilon'_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma'^2),\ i = 1, \ldots, I,\ j = 1, \ldots, J \tag{6}$$

The model allows us to estimate $\hat{\Delta}_{adj}$ and its standard error. Based on this model, we can test the more relevant null hypothesis in Eq. (2), by comparing the test statistic against the Student distribution with $df = I(J-1)$ degrees of freedom in balanced designs.

Although effective, the approach is somewhat simplistic. It is not applicable to experimental designs with more complex sources of biological and technological variation, such as experiments with repeated measurements, experiments with multiple batches or experiments with TMT labeling. Since Eq. (5) performs the adjustment on the replicate level, the experiment must contain a matching number of replicates in both the modified and unmodified runs. Technological artifacts such as missing values further undermine the calculation of u_{ij} in Eq. (5). Finally, there is no self contained, straightforward implementation of the method, such as in the form of a coding package, and therefore the approach requires a manual implementation.

Limma

The estimation of nuisance variation of the two ANOVA models above is often further expanded with Empirical Bayes moderation implemented in Limma [15, 23, 26, 27, 28, 29]. A typical application of Limma on summarized modified $\log_2$-intensities takes as input $\hat{y}_{ij}$, and for each PTM fits the linear model in Eq. (4).

A typical application of ratio-based Limma takes as input u_{ij}, and for each PTM fits the linear model in Eq. (6). The Limma versions of the models differ in that they specifies additional prior distributions for the model parameters. The priors are estimated from the same data by combining the information across all the proteins and all the PTM as described in [26]. With this approach, testing the null hypothesis is enhanced by combining the PTM- and protein-specific estimates of variation with a consensus estimate obtained from all the PTM and all the proteins. As the result, in experiments with few biological replicates the standard errors are often smaller, and the degrees of freedom are often larger than without moderation [15]. Thus the approach tends to increase the sensitivity of detecting differential abundance.

Since Limma only improves upon the estimation of variation, its limitations are similar to those of ANOVA. In particular, the method is only directly applicable to experiments with at most two variance components, and cannot account for all the sources of variation in experiments with either isobaric labeling or complex designs. There is no self contained implementation of the methods to PTMs, requiring manual transformation and application by the user.

Isobar-PTM

Isobar-PTM was also proposed for experiments with LC-MS/MS quantitative strategies that employ isobaric labels such as TMT, or isobaric tag for relative and absolute quantification (iTRAQ)[30]. Isobar-PTM expresses MS measurements with a linear model and performs adjustment with respect to protein abundance using the difference between log-ratio of modified peptides in two channels and log-ratio of protein level. Unfortunately, this statistical modeling framework is not applicable to either label-free workflows or experiments with complex designs.

Relative protein quantification in MSstats

MSstats [16] and *MSstatsTMT* [17] are a family of R/Bioconductor packages for statistical relative quantification of proteins and peptides in global, targeted and data-independent proteomics. The packages take as input $\log_2$-intensities y_{ijk}. For each protein, the $\log_2$-intensities are first summarized into a single value per protein per run using Tukey's median polish [31]. The summaries are then used as input to fit a flexible family of linear mixed-effects models [32, 33, 34]. The models are fit separately for each protein. The specific model depends on the design of the experiment, labeling type and data acquisition type as summarized in **Supplementary Figure S1**. For example, the unmodified protein features in the simple design in **Figure 2** are modeled with one-way ANOVA

$$\hat{y}_{ij} = \mu_i + \epsilon_{ij}, \text{where } \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \tag{7}$$

In contrast, a group comparison experiment with multiple TMT mixtures is modeled as

$$\hat{y}_{imj} = \mu_i + Mixture_m + \epsilon_{imj}, \text{where } Mixture_m \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_M), \epsilon_{imj} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \tag{8}$$

Moreover, the model fit for a particular protein depends on the pattern of missing values in that protein. If some of the terms of the model reflecting the experimental design are not estimable, a simpler model is fit for that protein instead.

Parameters of the model are estimated using restricted maximum likelihood (REML) [35]. The parameters allow us to estimate the pairwise comparison $\hat{\Delta}_{Protein} = \hat{\mu}_i - \hat{\mu}_{i'}$ and its standard error. Similarly to *Limma*, *MSstatsTMT* includes an optional Empirical Bayes moderation of the standard error [17], increasing the sensitivity of detecting differential abundance when the number of biological replicates in each condition is small.

MSstats and *MSstatsTMT* can also be used at the feature or at the modification cite level, as opposed to protein level. For example, summarizing the features per PTM cite instead of per protein, the approach allows us to test the null hypothesis in Eq. (1).

The *MSstats* framework has a number advantages over the methods above. First, unlike ANOVA and *Limma*, *MSstats* and *MSstatsTMT* are applicable to arbitrary complex experimental designs, including designs with multiple sources of variation, and unbalanced designs. Second, the approach is applicable to various data acquisition types, including label-free DDA and DIA, and experiments with TMT labeling.

Third, the *MSstats* packages are compatible with various data processing tools such as Skyline, Spectronaut, MaxQuant, Progenesis, Proteome Discoverer, and OpenMS. Finally, the custom *MSstats* and *MSstatsTMT* implementation accounts for potential data artifacts, is numerically scalable and stable, and is available through both command line and a dedicated graphical user interface.

Unfortunately, the *MSstats* framework focuses on overall protein abundance, and as the result tests the null hypothesis in Eq. (1). It does not account for the confounding between the changes in PTM abundance and the overall changes in protein abundance. This manuscript proposes a simple extension to the methodology in *MSstats* and *MSstatsTMT*, to test the null hypothesis in Eq. (2).

Results

Statistical methods in MSstatsPTM

Detecting changes in PTMs, adjusted for global changes in protein abundance

The overall statistical analysis workflow and its implementation are summarized in **Figure 3**. *MSstatsPTM* takes as input the modified spectral features y^*_{ijk}, and the corresponding unmodified features y_{ijk}. Ideally, the modified features are acquired separately, after an enrichment to maximize the information content in the resulting dataset, and the unmodified features are acquired separately as part of a global proteome profiling. However the method can also take as input a combination of modified and unmodified features acquired within a same run.

Each feature type is first analyzed separately using *MSstats* or *MSstatsTMT*. In particular, the modified features are summarized into run-level summaries $\hat{y}^*_{ij}$. The estimated summaries of the modified features are then used as the input to the appropriate model in *MSstats* or *MSstatsTMT*, such as in Eq. (7) or Eq. (8) [Could we show in Fig 3 where *MSstats* or *MSstatsTMT* come into place?]. The resulting model-based estimates include $\hat{\Delta}_{PTM} = \hat{\mu}^*_i - \hat{\mu}^*_{i'}$, and its standard error $\widehat{SE}(\hat{\Delta}_{PTM})$. Similarly, the unmodified features of each protein are summarized for each run into $\hat{y}_{ij}$, and the summaries are used as input to a separate analysis by *MSstats* or *MSstatsTMT* producing $\hat{\Delta}_{protein} = \hat{\mu}_i - \hat{\mu}_{i'}$ and $\widehat{SE}(\hat{\Delta}_{protein})$.

From the summaries above, the proposed approach estimates the adjusted difference $\hat{\Delta}_{adj}$ in Eq. (2)

$$\hat{\Delta}_{adj} = (\hat{\mu}^*_i - \hat{\mu}_i) - (\hat{\mu}^*_{i'} - \hat{\mu}_{i'}) = (\hat{\mu}^*_i - \hat{\mu}^*_{i'}) - (\hat{\mu}_i - \hat{\mu}_{i'}) = \hat{\Delta}_{PTM} - \hat{\Delta}_{protein} \tag{9}$$

Assuming that the sources of variation in the modified features that are unexplained by the model are independent from the corresponding sources of variation in the modified features, the standard error $\widehat{SE}(\hat{\Delta}_{adj})$ is obtained by combining the standard errors from the unmodified and modified model fits

$$\widehat{SE}(\hat{\Delta}_{adj}) = \sqrt{\widehat{SE}(\hat{\Delta}_{PTM})^2 + \widehat{SE}(\hat{\Delta}_{protein})^2} \tag{10}$$

For example, in the simple case of **Figure 2** with $J = 2$ conditions, where $\hat{\sigma}^2_{PTM}$ and $\hat{\sigma}^2_{protein}$ are respectively the estimates of the error variance for the PTM and protein model described in Eq. (7), the standard error is calculated as

$$\widehat{SE}(\hat{\Delta}_{adj}) \quad = \quad \sqrt{\widehat{SE}(\hat{\Delta}_{PTM})^2 + \widehat{SE}(\hat{\Delta}_{protein})^2} = \sqrt{\frac{1}{J}\hat{\sigma}^2_{PTM} + \frac{1}{J}\hat{\sigma}^2_{protein}} \tag{11}$$

The estimated standard error is larger than the standard errors associated with each individual feature type, reflecting the combined uncertainty in the two estimates. Finally, the degrees of freedom associated with Eq. (10) are obtained via the Satterthwaite approximation [22, 36]

$$df\left(\widehat{SE}(\hat{\Delta}_{adj})\right) = \left(\widehat{SE}(\hat{\Delta}_{PTM})^2 + \widehat{SE}(\hat{\Delta}_{protein})^2\right)^2 \Big/ \left(\frac{\widehat{SE}(\hat{\Delta}_{PTM})^4}{\mathrm{df}(\widehat{SE}(\hat{\Delta}_{PTM}))} + \frac{\widehat{SE}(\hat{\Delta}_{protein})^4}{\mathrm{df}(\widehat{SE}(\hat{\Delta}_{protein}))}\right) \tag{12}$$

To test the null hypothesis in Eq. (2), the test statistic $\hat{\Delta}_{adj}/\widehat{SE}(\Delta)_{adj}$ is compared with the Student distribution with the degrees of freedom in Eq. (12). The p-values of the comparison are adjusted for

multiple testing using the approach by Benjamin and Hochberg [37]. [Add 'df' to Step 5 of Fig 3?]

Sample size calculation for future PTM experiments

The proposed statistical framework enables sample size calculation for future PTM experiments. The procedure has been described in general in [22], and for protein significance analysis specifically in [38]. It requires us to specify the desired levels of the following quantities: a) q, the False Discovery Rate of detecting differential abundance, b) β, the average Type II error rate, c) Δ_{adj}, the minimal $\log_2$-fold change in adjusted PTM abundance of interest, d) $m_0/(m_0+m_1)$, the fraction of truly differentially modified PTM sites in the comparison, and e) σ^2_{PTM} and $\sigma^2_{protein}$, the anticipated variances associated with the summaries of the modified and unmodified features, respectively. Typically, the variances are estimated from an existing experiment, conducted with the same biological material and measurement workflow. Given the above quantities, and assuming a balanced design, the minimal number of replicates J across I conditions is chosen to bound the variance of the estimated $\log_2$-fold change $\mathrm{SE}^2(\Delta_{adj})$:

$$SE(\Delta_{adj})^2 = \left[\frac{2}{J}\left(\sigma^2_{PTM} + \sigma^2_{protein}\right)\right] \leq \left(\frac{\Delta_{adj}}{z_{1-\beta} + z_{1-\alpha/2}}\right)^2 \tag{13}$$

where

$$\alpha = (1-\beta)\cdot\frac{q}{1+(1-q)\cdot m_0/m_1} \tag{14}$$

and $z_{1-\beta}$ and $z_{1-\alpha/2}$ are the $100(1-\beta)^{\text{th}}$ and the $100(1-\alpha/2)^{\text{th}}$ percentiles of the Standard Normal distribution. Solving for J, the number of biological replicates per condition is

$$J \geq \frac{(2\sigma^2_{PTM} + 2\sigma^2_{protein})(z_{1-\beta} + z_{1-\alpha/2})^2}{\Delta^2_{adj}} \tag{15}$$

The numerator has two sources of variation, reflecting a larger uncertainty in the adjusted calculation. Therefore, the adjustment typically requires a larger sample size to gain the same sensitivity as the unadjusted estimation. Details of applications of this approach to experiments with complex design are in **Supplementary Sec. 1.1**. Examples of power analyses are in **Supplementary Sec. 3**.

Implementation of MSstatsPTM

The implementation of the open source R package *MSstatsPTM* is also overviewed in **Figure 3**. By leveraging the implementations in *MSstats* and *MSstatsTMT*, the proposed approach is very versatile. It is applicable to a wide variety of experimental designs, including group comparison, paired designs, time course designs and unbalanced designs. It is applicable to label-free data acquisitions such as DDA, DIA, SRM and label-based data acquisitions such as TMT. It can model experiments where the experimental designs for PTM profiling and global proteome profiling vary in properties such as number of biological replicates, data acquisition strategies and runs.

MSstatsPTM takes as input lists of identified and quantified spectral features, produced by spectral processing tools such as MaxQuant, Progenesis, or Spectronaut (Step 3 of **Figure 3**). Conversion is performed separately for the runs enriched in modified peptides, and separately for the global profiling runs. We require the processing tools to identify the modification site (i.e., the amino acid in the protein sequence where the modification occurred). This will generally include the amino acid abbreviation, plus it's number in the protein sequence. For example, a modification on a 70th amino acid in the sequence, serine should be marked as "S70". Occasionally the outputs of data processing tools only include the peptide sequence with the modified amino acid highlighted, without indicating the location in the protein sequence. For these cases *MSstatsPTM* includes functionality for identifying the location, given the modified peptide sequence and a FASTA file with the entire protein sequence. The converters output the modified spectral features y^*_{ijk} and the corresponding unmodified features y_{ijk} in the format required for summarization.

The next step is PTM/protein summarization using the *dataSummarizationPTM*() function (Step 4 of **Figure 3**). Summarization is performed separately for the PTM and the unmodified protein. When summarizing the PTM, modified peptide features that span the same modification site are summarized together.

Peptides that include multiple modifications are not included in the single modification summarization, and are grouped separately. The unmodified protein summarization is performed as discussed above for *MSstats*. When summarizing the unmodified protein features, the package optionally imputes missing values using an Accelerated Failure Time (AFT) model [39]. When summarizing the modified features, missing value imputation is also possible but should be performed with care. PTMs generally exhibit low feature counts and may be missing due to reasons other than low abundance. These issues can violate the assumptions underlying the imputation, and lead to numerically unstable results. The outputs of this step are the run-level summaries for the modified, $\hat{y}^*_{ij}$, and unmodified, $\hat{y}_{ij}$, features.

Finally, separate statistical models are fit to both feature summaries using the *groupComparisonPTM*() function (Step 5 of **Figure 3**). The models are automatically selected to reflect the experimental design and the data acquisition. If the base model is not applicable for a particular PTM or protein, e.g. due to missing data, a simplified model is fit. The output of the models are the estimates $\hat{\Delta}_{PTM}$ and $\hat{\Delta}_{protein}$, and their standard errors $\widehat{SE}(\hat{\Delta}_{PTM})$ and $\widehat{SE}(\hat{\Delta}_{protein})$.

After modeling, the modified model is adjusted for changes in unmodified protein abundance, using the methods described above. Modification sites which lack corresponding global profiling information cannot be adjusted for changes in protein abundance. In this case the implementation reverts to testing the null hypothesis in Eq. (1) using the statistical methods seen in *MSstats*, applied separately to each modified peptide. [This is indicated in the 'issues' column in the output?] The final output is the estimate $\hat{\Delta}_{adj}$ and it's standard error $\widehat{SE}(\hat{\Delta}_{adj})$.

In addition to the above functionalities, the implementation includes visualizations for quality control, *dataProcessPlotsPTM*(), and assessment of the quality of model fit, *groupComparisonPlotsPTM*().

The implementation relies on functionalities from the R packages *MSstats* [16] and *MSstatsTMT* [17], which in turn rely on the R packages *lme4* [40] and *lmerTest* [41]. *MSstatsPTM* is available on Bioconductor, `http://www.bioconductor.org/packages/release/bioc/html/MSstatsPTM.html`, and Github, `https://github.com/Vitek-Lab/MSstatsPTM`.

Evaluation

Evaluation strategy

We compared the performance of *MSstatsPTM* to that of [Need more details on the 'adjusted' vs 'unadjusted' versions ANOVA and Limma. Give each version a name that you'll later use in the figures, and describe what type of summaries each version takes.] ANOVA and of Limma. Since IsobarPTM is only applicable to experiments with TMT labeling, it could not be applied to the datasets with known ground truth in this manuscript, and was excluded from the comparisons. *MSstatsPTM* [Also describe the 'adjusted' version (MSstatsPTM) and the 'unadjusted' version (I guess it is MSstats/MSstatsTMT)] was used without imputing missing values, and without Empirical Bayes moderation. All the evaluations were done at the FDR-adjusted p-value cutoff of $q = .05$. More details are in **Supplementary Sec. 2**.

We evaluated *MSstatsPTM* on simulated and spike-in datasets with known ground truth in terms of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) differentially abundant PTMs. The true positives were defined as PTMs with changes distinct from the overall changes in unmodified protein abundance. The true negatives were defined as PTMs which, after accounting for the changes in the overall protein abundance, were not differentially abundant. Additional summaries were the false discovery rate $FDR = FP/(TP+FP)$ [I am not sure we should call this FDR. By definition FDR is a population-level quantity (it is an expected value of the proportion of false positives. Here it is an observed value). Maybe replace with 1-PPV?], $Recall = TP/(TP + FN)$, and $Accuracy = (TP + TN)/(TP + TN + FP + FN)$. For biological experiments with unknown ground truth, we compared the differentially abundant PTMs with and without adjusting for changes in unmodified protein abundance.

Protein-level adjustment was required to control FDR in differentially abundant PTM

Figure 4a(a) summarizes the FDR [I am not sure we should call this FDR. By definition FDR is a population-level quantity (it is an expected value of the proportion of false positives. Here it is an observed value). Maybe replace with 1-PPV?] reported on Computer Simulation 1 dataset by the unadjusted and adjusted versions of methods [XXX]. All the analyses were performed to control the FDR at at most 5%. The

simulation mimics a "clean" label-free experiment, not compromised by issues such as deviations from model assumptions, missing values and outliers. Yet, even under these favorable circumstances, the models that did not adjust for confounding from changes in overall protein abundance produced an excessive number of false positives. The versions of the models that accounted for the confounding produced error rates that were much better calibrated at the desired level. Since the simulation mimicked the assumptions of [model XXX], the error rates of this model were calibrated best.

In noisy simulations, MSstatsPTM improved the estimation of fold change over the existing methods [Hmm... Fig 4b is not about the fold change... Rephrase?]

[I suggest this order for this section. (Fig 4b has multiple panels, but the current narrative doesn't make it quite clear what is the difference between the panels) Fig 4b summarizes the results on Sim 2 dataset. The dataset is more realistic, and includes XXX. For low noise, methods compared in terms of sample size and number of conditions XXX. As the noise increased, XXX. The difference is due primarily to the robust nature of summarization in MSstats. Supplementary Fig XXX details the estimation of fold change]

In simulations with missing values and few features, $MSstatsPTM$ outperformed ANOVA and Limma. Changes in unmodified protein level still needed to be accounted for to control the FDR. Once controlling for changes in the unmodified protein, the proposed method outperformed the other methods, as seen in **Figure 4b**. The proposed method calibrated model accuracy well, even when the number of replicates were low. $MSstatsPTM$ can handle low feature counts better than the existing methods due to using Tukey's Median Polish for summarization. The existing methods summation approach requires more observations to correctly summarize the intensities in the run, whereas TMP is robust to few measurements.

Supplementary Figure S2 further compares the fold change estimation across all modified peptides. $MSstatsPTM$ showed a tighter distribution of estimated fold changes around the true fold change. Specifically, the inter-quartile range (IQR) of the estimated fold change for the proposed method was on average 32.5% smaller than Limma and ANOVA's IQR. While the mean of the estimated fold changes was generally correct for all methods, the proposed approach correctly estimated the fold change more often across all PTMs.

In the label-free benchmark experiment, MSstatsPTM had a higher sensitivity than the existing methods

In this experiment all models incorrectly estimated the fold change of the modified spike-in peptides before adjusting for changes in unmodified protein abundance. After adjustment, the spike-in peptides' fold change was generally in line with expectation in all methods, however the existing method's distribution of estimated fold changes was visibly wider (**Figure 5a**). Of the three approaches, the proposed method showed the tightest distribution around the true log fold change. Comparing the IQR of the spike-in peptide's log fold change, the proposed method's IQR was 32.86% smaller than Limma and ANOVA's IQR. While Limma and ANOVA overall estimated the correct fold change, they were inconsistent in their estimation, resulting in a higher error rate. The proposed method's estimation was more consistent, with a smaller error. As discussed in the previous section, the summarization methods of the proposed methods excel when the number of observations are low. This was the case here, with each PTM only averaging 1.37 features. The existing approaches used a non-robust summarization method, that was very sensitive to few observations, causing their fold change estimation to be more inconsistent, with a larger error compared to the proposed method.

In **Figure 5b** we can clearly see the fold change of the red labeled spike-in peptides was only correctly estimated when accounting for changes in the unmodified protein abundance. Additionally, the background peptides, serving as the null model, show many false positives before adjustment. After adjustment the number of false positives substantially decreased. Specifically, for the proposed method, the number of false positives went from 20.88% to 1.84% after adjustment was applied. While the proposed method and Limma both correctly estimated the fold change of the spike-in peptides, using Limma resulted in many large adjusted p-values. This was due to Limma estimating a 35.7% larger variance compared to the proposed method when estimating the spike-in peptides. This experiment introduced more variance components, which is challenging for the existing methods and therefore the difference in the results is more pronounced

than in the simulations. Volcano plots for all methods and comparisons can be seen in **Supplementary Section 2.1.3**.

In two group comparison TMT experiments, MSstatsPTM corrected for confounding with the unmodified protein

The results of these experiments are summarized in **Figure 6** and **Figure 7**. **Figure 6** shows the number of differentially abundant modified peptides before and after adjustment for Datasets 4 & 5. Adjusting for changes in the unmodified protein caused more PTMs to change from differentially abundant to non-differentially abundant than vice versa. A question that must be addressed is if the decrease in differentially abundant peptides is due to the increased variance that comes from adjustment. This was tested by looking for modified peptides whose adjusted log fold change was within 10% of the unadjusted log fold change but became statistically insignificant after adjustment. In other words, the fold change was the same between models but variance increased. When the test was applied to Dataset 4, only one PTM became non-differentially abundant due to an increase in variance. When applied to Dataset 5, 548 PTMs became non-differentially abundant due to an increase in variance (only 3.4% of all PTMs that became non-differentially abundant). Thus we can conclude that the drop off in differentially abundant PTMs was due to changes in global protein abundance, not simply increased model variance.

Figure 6 demonstrates that the unmodified protein was more likely to cause the observed changes in PTM abundance, rather than mask changes in the PTM. The later case can occur when the unmodified protein exhibits a change in abundance, while the modification does not. Additionally, this case may come about due to the high variability of PTMs compared to the global protein. An example of this can be seen in **Figure 7b**. Originally the modification at $S178$ of protein TTP was shown to be differentially abundant between KO_Uninfect and KO_Early, with a log fold change of 2.9. However the unmodified protein was shown to contribute 2.014 of this change, meaning that the modification only accounted for .886 after adjustment (seen in **Supplementary Figure S7**). This caused the modification to go from differentially abundant to non-differentially abundant. These cases occur frequently and can be thought of as false positives. The proposed approach corrects for this confounding, allowing us to see the true fold change of PTMs, while excluding those that do not exhibit differential changes.

Conversely, cases where the unmodified protein masks changes in the PTM are more infrequent. In this case, the modification and unmodified protein exhibit contradictory changes in abundance, moving in opposite directions. One such case can be seen in **Figure 7a**. Luchetti et al. [19] showed that $GSDMD$ was actively degraded when IpaH7.8 expression was induced by dox treatment. Our reanalysis confirmed this, with the $GSDMD$ protein being down-regulated when Dox treatments reached the 4 and 6 hour marks. Conversely, ubiquitination of $GSDMD$ at site $K62$ up-regulated abundance between the same conditions. This up-regulation was originally confounded by the down-regulation of unmodified $GSDMD$, and made the modification appear to have little change between no Dox and Dox 4 and 6 hour conditions. The proposed approach accounted for this confounding and the modification was shown to be differentially abundant, with an absolute fold change of 1.415 between the Dox 1 hour and Dox 4 hour conditions. This was an interesting result, as the modification contradicts what previous research showed for the unmodified protein. This effect can be thought of as a false negative and would have been challenging to observe without the proposed approach. There are many non-differentially abundant PTMs, and having to comb through each one individually to determine false negatives is generally an unfeasible task. The proposed approach removes the confounding on an experimental level, revealing modifications of interest that are entirely masked due to changes in the unmodified protein.

In label-free experiment without a separate global profiling run, MSstatsPTM eliminated the confounding due to changes in the unmodified protein, albeit less effectively than in the presence of a global profiling run

As discussed in **Section 2**, there was no unmodified global profiling run performed in this experiment. Once identification and quantification of the Ubiquitination profiling was performed, peptides which were unmodified were extracted and used in place of a global profiling run. This resulted in a lack of overlap between PTMs and unmodified proteins. Any PTM without a corresponding unmodified protein could not be adjusted. Of the 10,799 ubiquitin sites identified, only 4,526 had a corresponding unmodified protein and

could be adjusted. Additionally, the lack of a separate global profiling run resulted in low feature counts for the unmodified protein model compared to other experiments, seen in Table 5.1.

The number of differentially abundant PTMs before and after adjustment are shown in **Figure 8**. After adjusting for changes in the unmodified protein, there were fewer differentially modified peptides. However, this was mainly due to not having a global profiling run. When looking only at PTMs with a corresponding unmodified protein, the number of differentially modified PTMs exhibited a small drop, but was more consistent. As above, we tested if this drop in differentially abundant PTMs was due to an increase in variance. Here only 25 PTMs became insignificant due to increased variance. Without a global profiling run, the method was less effective at removing confounding with the unmodified protein. This was due to not being able to perform adjustment without a corresponding global protein. In this case adjustment caused us to toss out a large number of the PTMs in the original model. In cases without a global profiling run, the proposed method can still be used, but it should be used in conjunction with the unadjusted model so that information is not lost.

Discussion

[When you have a chance, could you do minor clean ups of the references? Remove the doi; check that the same journal has the same abbreviation in all the references; keep the reference format consistent (e.g., remove capitalization in article titles; italic for book/volume titles etc)]

[Somewhere in the discussion it would be good to comment that the proposed approach relies on the commutative property of the null hypothesis in Eq 2. While the ratio-based approaches look at the differences within one sample first, we look at the differences between conditions first. This provides us with a greater flexibility as far as modeling, accounting for outliers and missing values, and handling complex designs and robust implementations.]

We proposed a general statistical modeling framework and implementation for PTM characterization. The framework is designed for bottom-up MS workflows, which are characterized with variations from multiple confounded sources, frequent missing data, and associated uncertainty in the conclusions. The framework is general and is applicable to a variety of experimental designs. It outperforms the ad-hoc methods underlying ANOVA and Limma, and yields accurate results in the broad type of experimental circumstances, including the presence of missing values, changes in protein abundance, few representative peptides, and different acquisition methods. The framework allows us to plan for subsequent experiments, and choose the appropriate number of replicates in consideration of adjustment with respect to protein abundance. The implementation allows for straightforward application of the methods discussed and allows for reproducible experimental analysis.

Our results show that the proposed approach for modeling and summarization leads to more sensitive PTM significance analysis and more accurate and precise quantification. The gain is due to a more efficient use of the data, and to a more accurate understanding of the systematic and random variations. The proposed framework can be extended beyond the experimental designs with variation from multiple sources discussed above. Although demonstrated here on DDA, is also applicable to DIA, SRM and PRM acquisitions. Additionally, the approach can handle experiments with modified peptides processed using label-free methods and unmodified peptides processed using TMT labeling, or vice versa. In this case summarization and modeling is still done separately for both the modified and unmodified data, and then combined after modeling.

A potential limitation of the proposed framework is the assumption that all the peptides are correctly mapped to the underlying proteins and PTM sites, and the features are informative of the abundances of underlying protein and PTM. Also, characterizing PTMs with current data-dependent acquisition workflows is prone to being under sampled, leading to a sparse dataset with a large number of missing values for the analysis. Statistical methods accounting for effects due to experimental units and missing values introduced in this manuscript help interpret the data in a more objective manner. The latest development of targeted acquisition and data-independent acquisition methods are expected to further alleviate these issues.

Additionally, abundance levels of PTM sites can be confounded with each other if there are multiple modification sites per peptide, or confounded with changes in the unmodified peptide (as opposed to the unmodified protein). In the current implementation the effect of a specific modification in a peptide with

multiple modifications cannot be quantified. One potential solution to this is to measure the abundance of peptides with one modification and use this to adjust the peptide with multiple sites to remove the confounding. However, this method would likely run into challenges due to sparsity of features for modified peptides with both a single and multiple modification sites. A more complex approach to addressing this problem is likely necessary.

Overall, the proposed approach balances accuracy and practicality, and enables the analysis of complex experiments in high throughput. Future work is to carry out the inference and testing for not only the relative change of PTM abundance, but also the fraction of the protein that is modified at the particular site (site occupancy, or stoichiometry), and attempt to remove the confounding of individual PTMs in peptides with multiple modifications.

References

[1] Y. L. Deribe, T. Pawson, and I. Dikic (2010). "Post-translational modifications in signal integration". In: *Nature Structural & Molecular Biology* 17, pp. 666–672.

[2] P. Cohen (2000). "The regulation of protein function by multisite phosphorylation–a 25 year update". In: *Trends in Biochemical Sci.* 25.12, pp. 596–601.

[3] I. Bludau et al. (2022). "The structural context of PTMs at a proteome wide scale". In: *bioRxiv Preprint*. URL: `https://doi.org/10.1101/2022.02.23.481596`.

[4] M. Mann and O. Jensen (2003). "Proteomic analysis of post-translational modifications". In: *Nature Biotechnology* 21, pp. 255–261.

[5] N. A. Petushkova et al. (2017). "Post-translational modifications of FDA-approved plasma biomarkers in glioblastoma samples". In: *PLOS ONE* 12.5, e0177427.

[6] R. Wu et al. (2011). "Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes". In: *Mol Cell Proteomics* 10.8, p. M111.009654.

[7] K. Chandramouli and P. Y. Qian (2009). "Proteomics: challenges, techniques and possibilities to overcome biological sample complexity". In: *Human genomics and proteomics : HGP* 1.1. URL: `https://doi.org/10.4061/2009/239204`.

[8] M. S. Kim, J. Zhong, and A. Pandey (2016). "Common errors in mass spectrometry-based analysis of post-translational modifications". In: *Proteomics* 16.5, pp. 700–714.

[9] L. Käll and O. Vitek (2011). "Computational Mass Spectrometry–Based Proteomics". In: *PLoS Computational Biology* 7.12.

[10] P. Roepstorff (1997). "Mass spectrometry in protein studies from genome to function". In: *Current Opinion in Biotechnology* 8.1, pp. 6–13.

[11] J. Huang et al. (2014). "Enrichment and separation techniques for large-scale proteomics analysis of the protein post-translational modifications". In: *Journal of Chromatography A* 1372, pp. 1–17.

[12] J. Olsen and M. Mann (2013). "Status of Large-scale Analysis of Post-translational Modifications by Mass Spectrometry". In: *Molecular & Cellular Proteomics* 12.12, pp. 3444–3452.

[13] D. Kalpić, N. Hlupić, and M Lovrić (2011). "Student's t-Tests". In: *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1559–1563.

[14] Ellen R Girden (1992). *ANOVA: Repeated measures*. 84. Sage.

[15] M. E. Ritchie et al. (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7, e47.

[16] M. Choi et al. (2014). "MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments". In: *Bioinformatics* 30.17, pp. 2524–2526.

[17] T. Huang et al. (2020). "MSstatsTMT: Statistical Detection of Differentially Abundant Proteins in Experiments with Isobaric Labeling and Multiple Mixtures". In: *Molecular & Cellular Proteomics* 19.10, pp. 1706–1723.

[18] M. Choi et al. (2020). "MassIVE.quant: a community resource of quantitative mass spectrometry–based proteomics datasets". In: *Nat Methods* 17, pp. 981–984.

[19] G. Luchetti et al. (2021). "Shigella ubiquitin ligase IpaH7.8 targets gasdermin D for degradation to prevent pyroptosis and enable infection". In: *Cell Host & Microbe* 29.10, pp. 1521–1530.

[20] T. Maculins et al. (2021). "Multiplexed proteomics of autophagy-deficient murine macrophages reveals enhanced antimicrobial immunity via the oxidative stress response". In: *eLife* 10, e62320.

[21] C. Cunningham et al. (2015). "USP30 and Parkin homeostatically regulate atypical ubiquitin chains on mitochondria". In: *Nature Cell Biology* 17.2, pp. 160–169.

[22] M. H. Kutner et al. (2004). *Applied Linear Statistical Models*. 5th ed. McGraw-Hill/Irwin.

[23] V. Schwämmle, T. Verano-Braga, and P. Roepstorff (2015). "Computational and statistical methods for high-throughput analysis of post-translational modifications of proteins". In: *Journal of Proteomics* 129. Special Issue : Computational Proteomics, pp. 3–15.

[24] S. P. Thomas et al. (2020). "A practical guide for analysis of histone post-translational modifications by mass spectrometry: Best practices and pitfalls". In: *Methods* 184, pp. 53–60. ISSN: 1046-2023.

[25] P. Mertins et al. (2013). "Integrated proteomic analysis of post-translational modifications by serial enrichment". In: *Nature Methods* 10, pp. 634–637.

[26] G. K. Smyth (2003). "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments". In: *Statistical Applications in Genetics and Molecular Biology* 3.1, Article 3.

[27] — (2005). "limma: Linear Models for Microarray Data". In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Ed. by R. Gentleman et al. New York, NY: Springer New York, pp. 397–420.

[28] Y. Zhu et al. (2020). "DEqMS: A Method for Accurate Variance Estimation in Differential Protein Expression Analysis". In: *Molecular & Cellular Proteomics* 19.6, pp. 1047–1057.

[29] K. Chappell et al. (2021). "PTMViz: a tool for analyzing and visualizing histone post translational modification data". In: *BMC Bioinformatics* 22.275.

[30] F. Breitwieser and J. Colinge (2013). "IsobarPTM: A software tool for the quantitative analysis of post-translationally modified proteins". In: *Journal of Proteomics* 90, pp. 77–84.

[31] J. W. Tukey (1977). *Exploratory data analysis*. Addison-Wesley.

[32] Robert A. McLean, William L. Sanders, and Walter W. Stroup (1991). "A Unified Approach to Mixed Linear Models". In: *The American Statistician* 45.1, pp. 54–64.

[33] J. J. Faraway (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC Press.

[34] B. Bolker et al. (2009). "Generalized linear mixed models: a practical guide for ecology and evolution". In: *Trends in Ecology and Evolution* 24.3, pp. 127–135.

[35] M. Kenward and J. Roger (1997). "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood". In: *Biometrics* 53.3, pp. 983–997.

[36] F. E. Satterthwaite (1946). "An approximate distribution of estimates of variance components". In: *Biometrics Bulletin* 2.6, pp. 110–114.

[37] Y. Benjamini and Y. Hochberg (1955). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *J.R. Statist. Soc. B* 57.1, pp. 289–300.

[38] A. L. Oberg and O. Vitek (2009). "Statistical design of quantitative mass spectrometry-based proteomic experiments". In: *Journal of Proteome Research* 8.5, pp. 2144–2156.

[39] L.J. Wei (1992). "The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis". In: *Statistics in medicine* 11.14-15, pp. 1871–1879.

[40] D. Bates et al. (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48.

[41] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen (2017). "lmerTest Package: Tests in Linear Mixed Effects Models". In: *Journal of Statistical Software* 82.13, pp. 1–26.

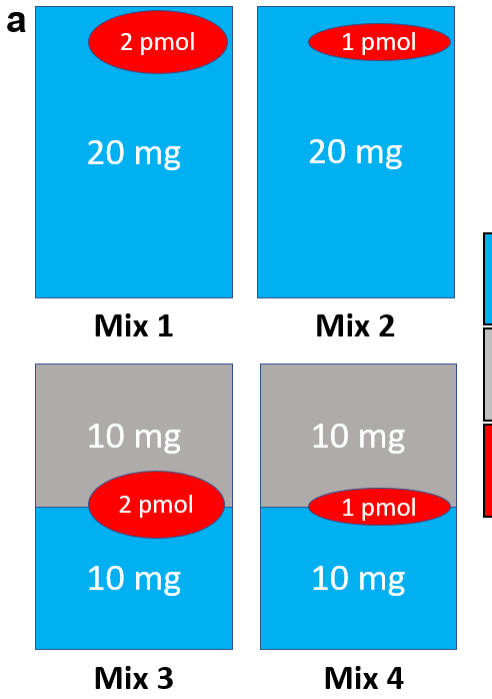
Tables

Experimental Datasets Overview

	Dataset	No. of Conditions	No. of bio. replicates	No. of mod. peptides	No. of mod. features/site	No. of unmod. features/protein	Data availability	Analysis
Known Ground Truth	Computer simulation 1 - Label-free	2/3/4	2/3/5/10	1,000	10	10	Github	
	Computer simulation 2 - Missing and low features	2/3/4	2/3/5/10	1,000	2	10	Github	
	SpikeIn benchmark - Ubiquitination - Label-free	4	2	12,137	1.37	10.17	MSV000088971	TBD
Biological Experiment	Human - Ubiquitination - 1mix-TMT	6	2 or 1	8,848	1.21	11.01	MSV000088966	RMSV000000356
	Mouse - Phosphorylation - 2mix-TMT	6	4 or 3	26,433	1.67	11.61	MSV000085565	RMSV000000357
	Human - Ubiquitination - Label-free	4	2	10,799	1.40	1.65	MSV000078977	RMSV000000358

Table 5.1: **Simulated and experimental datasets.** "Dataset" is the dataset code name. "No. of bio. replicates" shows the number of biological replicates per condition. Simulations were generated with different numbers of replicates. The designs of two biological experiments were unbalanced with unequal replicates per condition. "No. of mod. features/site" is the number of features (i.e., peptide ions) used to estimate the abundance of a single modification. "No. of unmod. peptides/protein" is the number of peptide ions without modifications that were used to estimate the global protein abundance. "Data availability" is the ID of the MassIVE.quant repository or the GitHub repository. "Analysis" is the ID of the MassIVE.quant reanalysis container, containing analysis code and modeling results. All the experiments were conducted in data-dependent acquisition (DDA) mode.

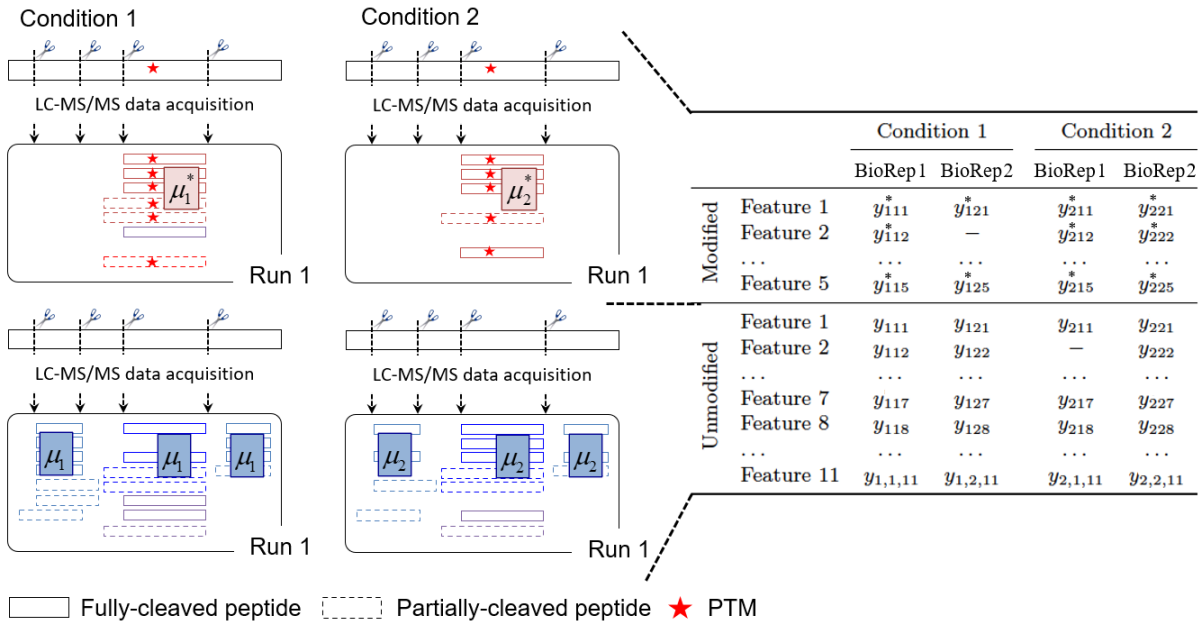
Figures



b

Comparison	Unadj. true $\log_2$FC	Adj. true $\log_2$FC
mix2 vs mix1	-1	-1
mix3 vs mix1	0	1
mix4 vs mix1	-1	0
mix3 vs mix2	1	2
mix4 vs mix2	0	1
mix4 vs mix3	-1	-1

Figure 1: **Dataset 3: Spike-in benchmark - Ubiquitination - Label-free**. (a) Four mixtures (i.e., conditions) were created with varying amounts of human lysate, background *E. Coli* lysate, and human spike-in ub-peptide mixture. Unmodified peptides from human lysate were viewed as the global proteome. Background *E. coli* lysate were used to equalize total protein levels. 50 heavy-labeled KGG motif peptides from 20 human proteins were spiked into the mixed background of the lysates. Quantitative changes in protein and site abundance of these 20 human proteins were the target of the benchmark. (b) We distinguished the unadjusted changes (i.e. changes in the abundances of the modified peptides) and the protein-level adjusted changes of (i.e., changes in the abundances of the modified peptides relative to the changes in the abundances of the human lysate). "Unadj. true $\log_2$FC" are the log-ratios of the abundances of the spiked peptides between each condition. "Adj. true $\log_2$FC" was calculated by determining the ratios of the abundances of the spiked peptides and human lysate between each condition and then adjusting the ratio of the spiked peptides by the human lysate, similarly to Eq. (2).



		Condition 1		Condition 2	
		BioRep1	BioRep2	BioRep1	BioRep2
Modified	Feature 1	y^*_{111}	y^*_{121}	y^*_{211}	y^*_{221}
	Feature 2	y^*_{112}	–	y^*_{212}	y^*_{222}
	...	...	...	...	...
	Feature 5	y^*_{115}	y^*_{125}	y^*_{215}	y^*_{225}
Unmodified	Feature 1	y_{111}	y_{121}	y_{211}	y_{221}
	Feature 2	y_{112}	y_{122}	–	y_{222}
	...	...	...	...	...
	Feature 7	y_{117}	y_{127}	y_{217}	y_{227}
	Feature 8	y_{118}	y_{128}	y_{218}	y_{228}
	...	...	...	...	...
	Feature 11	$y_{1,1,11}$	$y_{1,2,11}$	$y_{2,1,11}$	$y_{2,2,11}$

Figure 2: Schematic representation of one PTM site, in a special case of a label-free experiment with $I = 2$ conditions and $J = 2$ biological replicates per condition. After a $\log_2$ transform, we are interested in estimating the difference between the population-level PTM abundance between Condition 1 and Condition 2 (i.e, $\mu_1^* - \mu_2^*$), relative to the population-level difference of the overall protein abundance (i.e., $\mu_1 - \mu_2$). These quantities are characterized by the observed spectral Features (boxes), i.e. peptides of different charge states. The peptides can be fully cleaved (solid lines), or partially cleaved (dashed lines). The $\log_2$-intensities of the modified peptides in Condition i, Run j, and Feature k are denoted by y^*_{ijk}. The $\log_2$-intensities of Feature k corresponding to the unmodified peptide in Condition i and Run j are denoted by y_{ijk}.

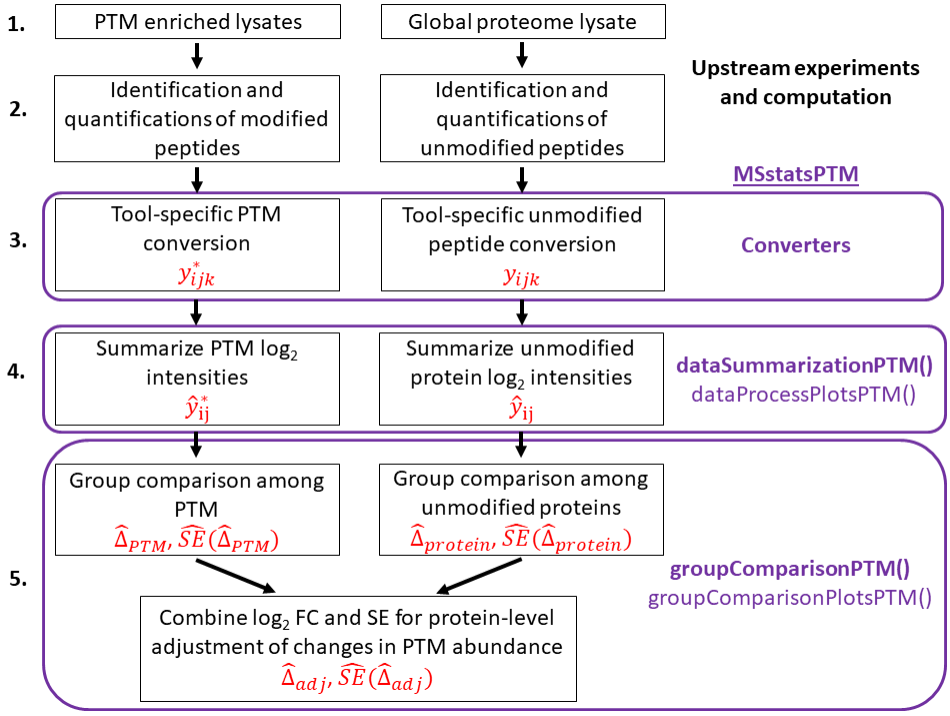
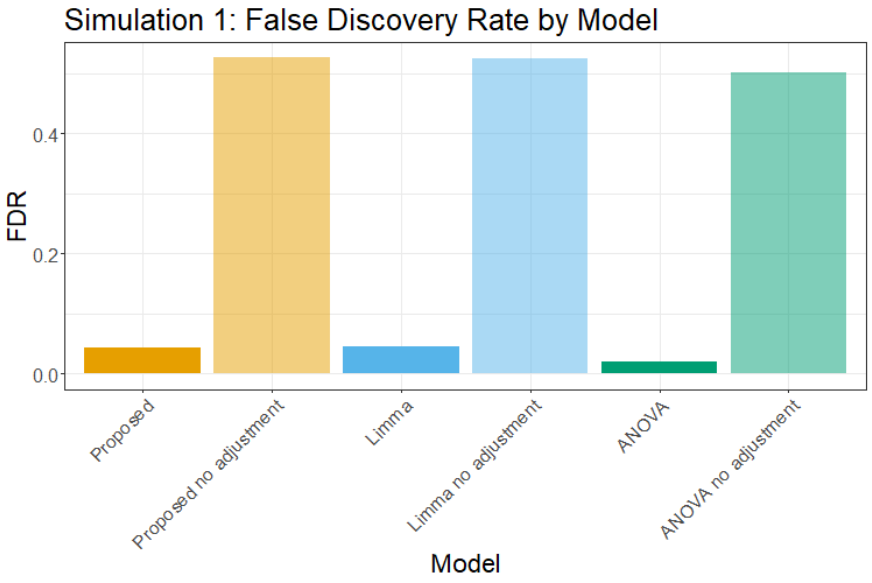
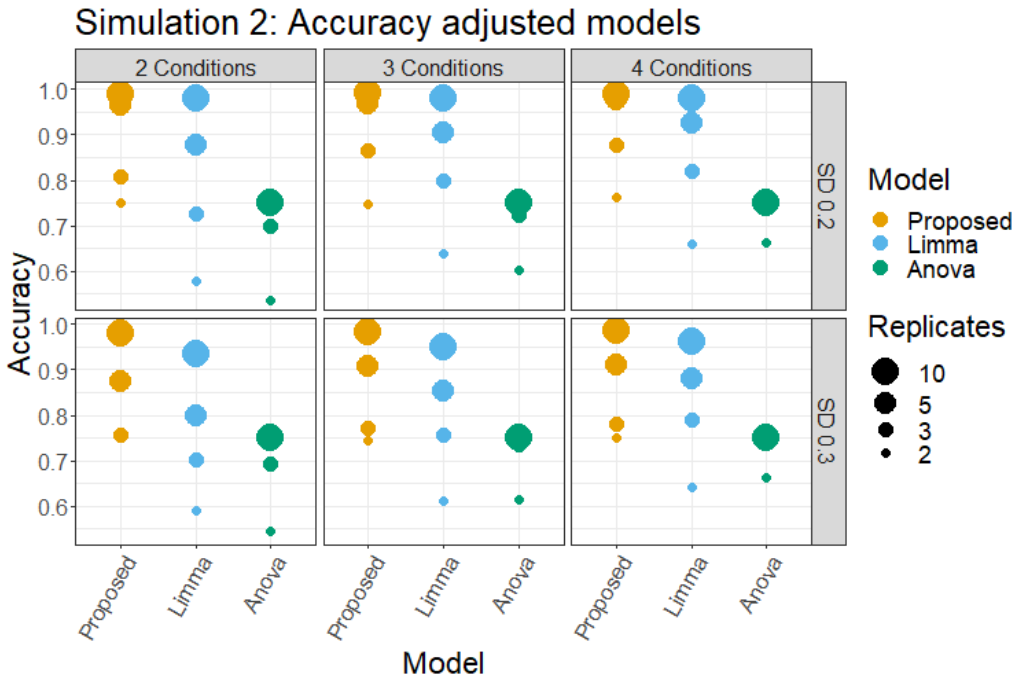


Figure 3: [Where does the sample size calculation fit into that?] [Could you rotate the 'Upstream...' Text and 'MSstatsPTM text and place it vertically to the left of the corresponding steps? It will be a bit more clear] [It is not very clear which steps leverage MSstats/MSstatsTMT. Should we say in relevant places that these are MSstats implementations?] The $MSstatsPTM$ workflow. The names of the R functions used for each step are highlighted in purple and the output notations are highlighted in red. The workflow begins with the acquisition of the enriched and global proteome lysates. The package is applicable to label-free data acquisitions such as DDA, DIA, SRM and label-based data acquisitions such as TMT. It takes as input lists of identified and quantified spectral features for the PTM and unmodified protein, produced by spectral processing tools such as MaxQuant, Progenesis, or Spectronaut. Conversion, summarization and statistical modeling are performed separately for the PTM and for the unmodified proteins. At the end, the model-based summaries are combined to adjust the changes in the PTM abundance for changes in unmodified protein abundance. Data visualizations can optionally be created after steps 4 and 5 of the workflow.

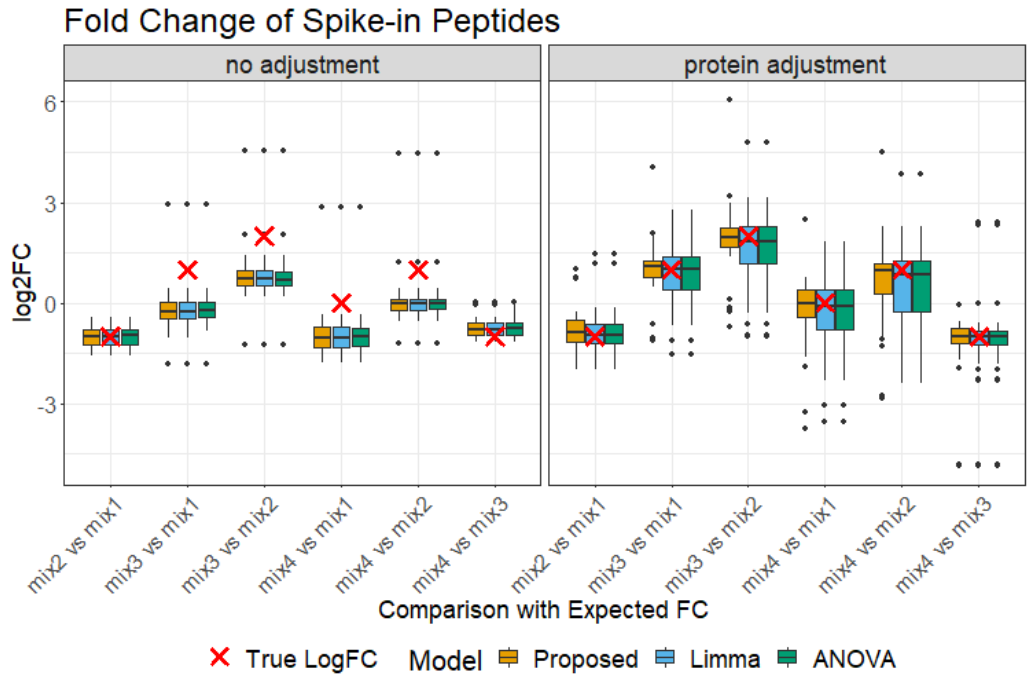


(a)

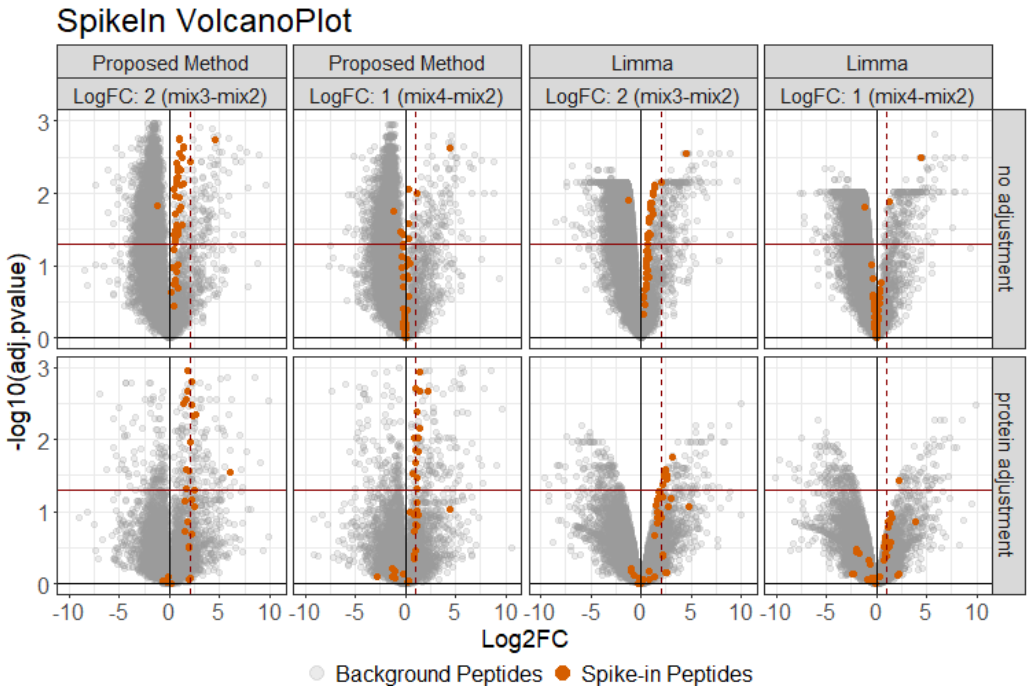


(b)

Figure 4: [Could you clarify what 'with' and 'without' adjustment means for each model? E.g., what is 'proposed with no adjustment'? We did not propose anything without adjustment - did you mean base MSstats? Also, maybe clarify what you mean by limma with and without adjustment, since we technically didn't introduce limma without the adjustment. Maybe explain the inputs to each version of the model?] [In (a), could you add a horizontal line at 5%?] Computer simulations. a) Dataset 1, clean simulation, analyzed to control FDR at at most 5%. The methods not accounting for the confounding between changes in PTM and overall changes in protein abundance produced exceedingly high numbers of false positives. In contrast, the methods accounting for the confounding correctly calibrated the proportion of false positive differentially abundant PTM. b) In a noisy simulation, including limited feature observations and missing values, the advantage of the proposed approach was apparent. Looking at accuracy, the proposed method outperformed ANOVA and Limma in nearly every model.

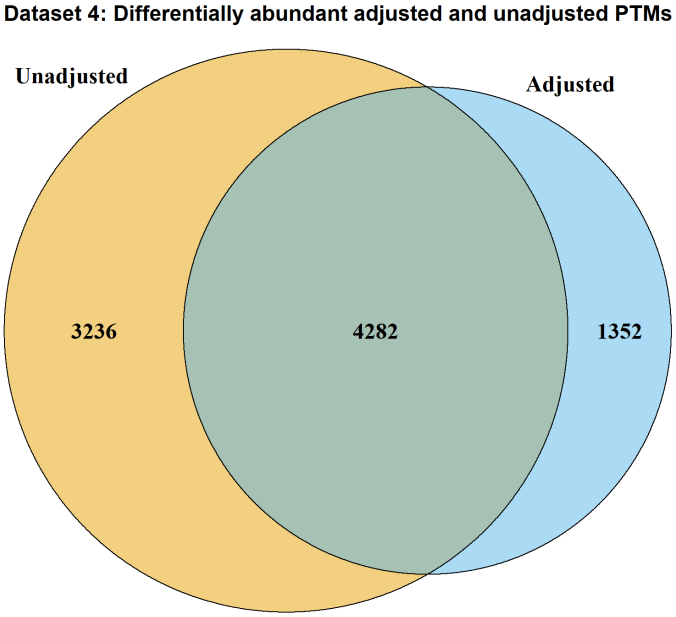


(a)

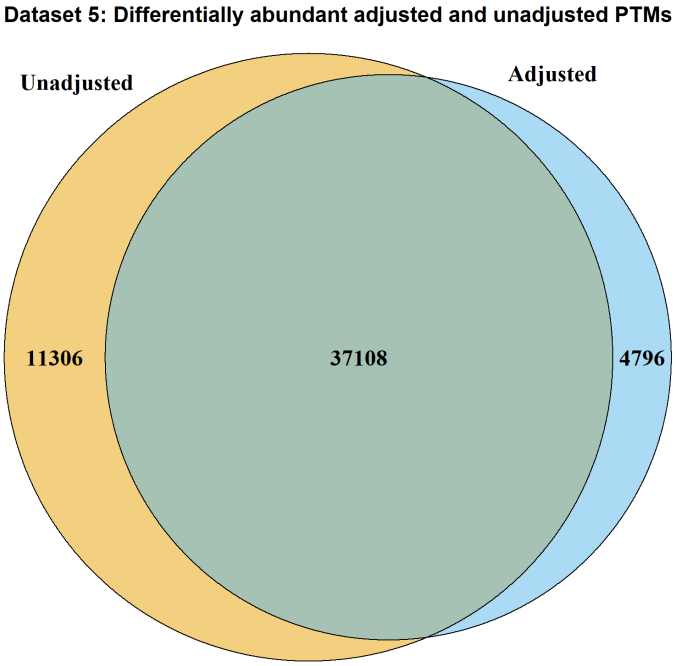


(b)

Figure 5: Dataset 3: Spike-in benchmark - Ubiquitination - Label-free. a) Before adjustment the fold change of the spike-in peptides' were systematically different from the expected fold change in all models. After adjustment, this systemic difference was removed, however the inner quartile range of the Limma and ANOVA models was wider than the proposed method. b) [The title of the figure reads as if it is accuracy-adjusted. Maybe rephrase?] Before adjustment the spike-in peptides (colored red) did not follow the expected log fold change; after adjustment, the spike-in peptides were more in line with expectation. Using Limma, the spike-in peptides followed the expected log fold change after adjustment, however the majority of spike-in peptides did not have a significant adjusted p-value.

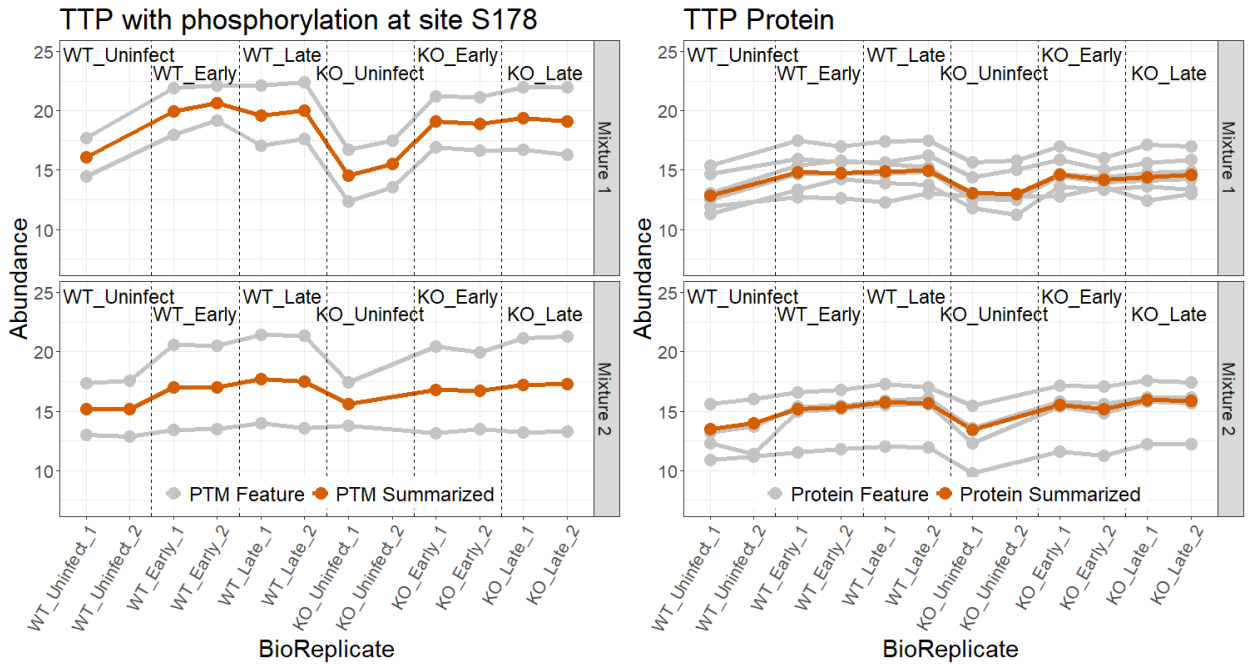


(a)

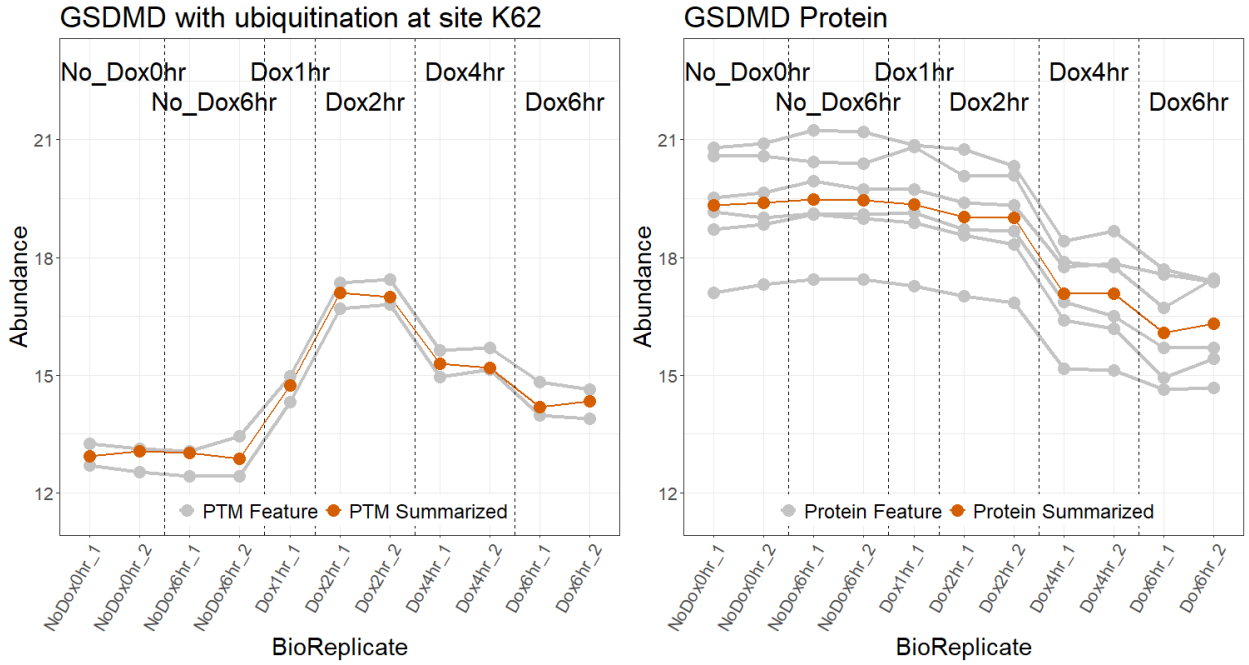


(b)

Figure 6: a) Dataset 4: Human - Ubiquitination - 1mix-TMT. The overlap of differential modified peptides for the PTM model with and without global protein level adjustment across all comparisons. 3,236 modified peptides became insignificant, 1,352 became significant, while 4,282 were significant in both models. b) Dataset 5: Mouse - Phosphorylation - 2mix-TMT. The overlap of differentially modified peptides between the PTM model with and without global protein level adjustment across all comparisons. 19,286 peptides became insignificant, 4,947 became significant, and 41,552 were significant in both models.

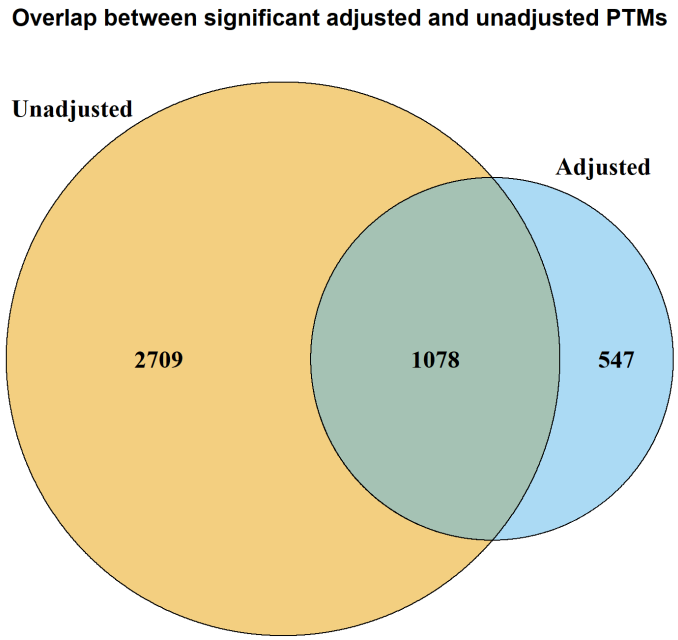


(a)

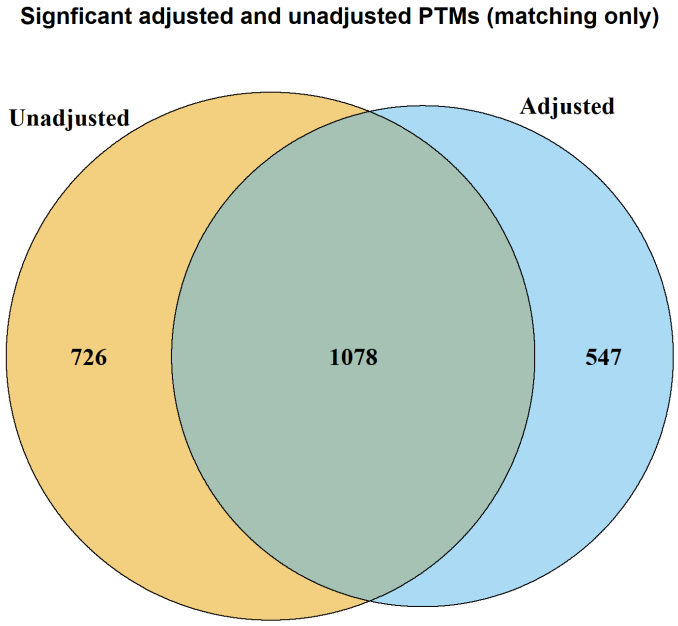


(b)

Figure 7: a) Dataset 5: Mouse - Phosphorylation - 2mix-TMT. Comparing the global profiling of protein TTP with the modification at site $S178$. The modification and unmodified protein abundance followed the same trend across conditions. Specifically, there was a positive adjustment in abundance when comparing WT_Uninfect to WT_Late in both the modification and global profiling run. b) Dataset 4: Human - Ubiquitination - 1mix-TMT. Comparing the global profiling of protein $GSDMD$ with the ubiquitination of the protein at site $K62$. Here the modification and global protein follow different trends. There appeared to be no change in abundance between Dox1hr and Dox4hr in the modified plot, however there was a large negative change when looking at the unmodified plot. This indicated the modification was confounded with changes in the unmodified protein.



(a)



(b)

Figure 8: Dataset 6: Human - Ubiquitination - Label-free no global profiling run. a) The overlap of differential modified peptides for the PTM model with and without global protein level adjustment across all comparisons. More PTMs became insignificant than became significant after adjustment. In total, 2709 modified peptides became insignificant, while only 547 became significant. b) Here we made the same comparison but only looked at modified peptides where adjustment could be performed, ie they had a matching unmodified protein. In this case there were significantly less peptides that became insignificant after adjustment. 726 modified peptides became insignificant, 547 became significant, and 1,078 were significant in both models.