MSstatsPTM statistical relative quantification of post-translational modifications in bottom-up proteomic experiments

Devon Kohler[1], Tsung-Heng Tsai[2], Ting Huang[1], Erik Verschueren[4], Trent Hinkle[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 mass spectrometry (LC-MS/MS)-based proteomics is increasingly used to detect changes in post-translational modifications (PTMs) between samples from various conditions. Analysis of data from such experiments faces numerous statistical challenges. These include the low abundance of modified proteoforms, the small number of representative 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 for each sample quantifies both the peptides with post-translational modifications, and peptides from the same proteins with no modification sites. The 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, 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 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 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

[This section would benefit from more biology/technology background, and more extensive references] Signaling mechanisms allow cells to mount a fast and dynamic response to a multitude of bimolecular events. Signaling is facilitated by the modification of proteins at specific residues, acting as molecular on/off switches [1, 2]. Localizing the modification sites, or characterizing relative abundance of a modification site's occupancy repertoire across experimental conditions provides important insights [3]. For example, meaningful patterns of changes in post-translational modifications (PTMs) abundance can serve as biomarkers of a disease [4]. 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 [would be good to have a reference].

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 [5, 6]. 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 [7]. 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 [3]. 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 [8]. 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 challenges in PTM identification and quantification challenge the downstream statistical analyses. Yet frequently data from these experiments are analyzed using statistical methods that were not originally designed for this task. In particular, methods such as t-test, Analysis of Variance, or Limma take as input the intensity ratios of modified and unmodified peptide features, and compare the mean abundance of different PTM sites [9]. Such methods do not fully account for all the sources of uncertainty and variation in the experiment. As the result, they 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 general statistical analysis framework that detects relative changes in post-translational modifications. The framework requires an experimental design, which for each sample quantifies 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 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, acquisition methods, and experimental designs, showing the applicability of the framework to a variety of situations. The evaluations demonstrated that by appropriately leveraging the information from the unmodified peptides, the proposed approach improves the accuracy of the estimates of PTM fold changes and results in a better calibrated false positive rate of detecting differentially abundant PTMs. In particular, accounting for the confounding from protein abundance allows us to characterize the true effect of the modification, avoiding the need for more manual and time intensive methods.

The proposed approach is implemented as a freely available open source R package $MSstatsPTM$, available on Bioconductor, which employs similar input format as $MSstats$ and $MSstatsTMT$ [10, 11].

Experimental Procedures

The datasets in this manuscript benchmarked the proposed approach in situations with known ground truth, and represented a variety of experimental designs and acquisition methods. Two computer simulations 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. The spike-in experiment took the real world characteristics a step further and allowed us to compare the methods in a real-world experimental settings with known changes in modified spike-in peptides. Finally, three biological experiments demonstrated the applicability of the proposed approach across different biological organisms, PTM types, experimental designs and acquisition strategies. Table 2.1 summarizes the experiments. Details of 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`) [12]. Details of computer simulations are available on GitHub (`https://github.com/devonjkohler/MSstatsPTM_simulations`).

	Dataset	Experimental Design	No. of Conditions	No. of Bio. Replicates	Data availability
Known Ground Truth	Computer simulation 1 - Label-free	Group comparison	2/3/4	2/3/5/10	Github
	Computer simulation 2 - missing and low features	Group comparison	2/3/4	2/3/5/10	Github
	SpikeIn benchmark - Ubiquitination - Label-free	Group comparison	4	2	MSV000088971
Biological Experiment	Human - Ubiquitination - 1mix-TMT	Group comparison	6	2	MSV000088966
	Mouse - Phosphorylation - 2mix-TMT	Time course	6	4	MSV000085565
	Human - Ubiquitination - Label-free	Group comparison	4	2	MSV000078977

Table 2.1: **Simulated and experimental datasets in this manuscript** "'Dataset" is the code name of the dataset in this manuscript. "Data availability" shows the ID of the MassIVE.quant repository or the GitHub repository. All the experiments were conducted in data-dependent acquisition (DDA) mode.

Dataset 1 : Computer simulation 1 - Label-free

The simulation represented an idealistic case. 24 synthetic datasets were generated exhibiting scenarios with different experimental designs and biological variation. In each dataset, 1000 proteins had 10 unmodified features per protein. Each of the 1000 proteins had one PTM, each PTM 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 difference between conditions. Furthermore, the fold changes of half of the 500 differential PTMs were fully masked by changes in the unmodified protein. Additionally, the fold change of half the 500 non-differential PTMs was entirely due to changes in the unmodified protein. All differential PTMs were generated with an expected log fold change of 0.75 between conditions.

Each simulation was generated with random biological variation. Specifically, let Y_{ibc} denote the abundance of peptide i in biological replicate b of condition c. The simulated peptide abundance Z_{ibc} was generated as $Z_{ibc} = Y_{ibc} + \epsilon_{ibc}$ where $\epsilon_{ibc} \sim^{iid} N(0, \sigma^2)$. The simulations were generated with one of two values $\sigma^2 = \{.2, .3\}$. The values for σ^2 were motivated [motivated how specifically? corresponded to median variances?] by the biological experiments in this manuscript. Further details are in Supplementary Sec. 2.1.2.

Dataset 2 : Computer simulation 2 - Label-free missing values and low features

The data were simulated as above, while including missing values and a low number of modified features to provide a more realistic representation of the experiments. In this simulation, 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, while unmodified proteins

were simulated with 10 features. Additionally, 20% of observations for both modified and unmodified peptides were missing completely at random. [There will likely be a comment that missing values are more common among low-abundance features. Did the simulation account for that? Devon: The simulation did not include any missing values due to low abundance. Should we mention this or just leave it as is?] Further details are in Supplementary Sec. 2.1.3.

Dataset 3 : SpikeIn benchmark - Ubiquitination - Label-free

In the custom spike-in experiment 8 biological [I am not sure these are bio reps. If this is the same lysate, then these are in fact technical reps (there is no between-subject variation)] replicates were allocated to 4 conditions, resulting in 2 biological replicates per condition, and a balanced design. Additional information can be found in Supplementary Sec. 2.1.4.

Experimental Design: 50 heavy-labeled KGG motif peptides from 20 human proteins were used as spike-in peptides. Quantitative changes in protein and site abundance of these 20 proteins were the target of the benchmark. Unmodified peptides from Human lysate were used to estimate changes in global protein abundance. Background *E. coli* lysate was used to normalize total protein levels prior to enrichment or global protein profiling. The background lysate was treated as control, i.e. it had no expected changes in any comparison. [I am a bit confused. Fig S7 shows that there is a difference in Background between Mix 1,2 and Mix 3,4. Could you explain better? In fact, would it be possible to move the figure here for clarity, and have the figure as panel (a), and the table as panel (b)? And in the text above refer to the figure. This will help the understanding] [Also, why is it a balanced design - as stated in the paragraph above - when the background is not present everywhere?]

The spike-in peptides were mixed with human lysate to create four mixture conditions. The true log fold changes of spike-in modified peptides in these comparisons can be seen in Table 2.2. Two sets of data were acquired for each mixture in label-free LC-MS/MS mode: KGG enriched + LC-MS/MS, and LC-MS/MS only.

Global Profiling Run: This experiment included a separate global profiling run to measure unmodified peptides. There was a 90.2% overlap between the identified background modified peptides and proteins that were quantified in the global profiling run.

Pairwise Comparisons: We evaluated the ability of the statistical methods to detect known changes in abundance between pairs of conditions. The four mixtures were treated as individual conditions and labeled mix1, mix2, mix3, and mix4. The pairwise comparisons were mix1-mix2, mix1-mix3, mix1-mix4, mix2-mix3, mix2-mix4, and mix3-mix4.

Comparison	Expected log fold change
mix1-mix2	-1
mix1-mix3	1
mix1-mix4	0
mix2-mix3	-2
mix2-mix4	-1
mix3-mix4	1

Table 2.2: The expected log fold change of modified spike-in peptides in Dataset 3.

Dataset 4 : Human - Ubiquitination - 1mix-TMT

In this experiment 11 biological replicates were allocated to 1 TMT mixture in an [group comparison or time course?] unbalanced design. Further details are in Supplementary Sec. 2.2.1.

Experimental Design: Luchetti et al. [13] quantified the abundance of total protein and ubiquitination in human epithelial cells. Cells were engineered to express IpaH7.8 under a dox inducible promoter and measurements were taken at different time periods. GSDMD was actively degraded when IpaH7.8 expression was induced by dox treatment. Uninfected cells were measured at 0 and 6 hours, while infected cells were measured at 1, 2, 4, and 6 hour increments, resulting in six total conditions. [If this is a group comparison, explain why, even though the conditions are times. The table above says it is group comparison] The proteome was quantified using isobaric multiplexing via tandem mass tagging (TMT) in combination with LC-MS/MS. [One MS run total? Only one MS run was used. Not sure the best way to say this.]

Global Profiling Run: This experiment included a separate global profiling run, with the same design, to measure unmodified peptides. There was a 95% overlap between the identified modified peptides and proteins that were quantified in the global profiling run.

Pairwise Comparisons: 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.

Dataset 5 : Mouse - Phosphorylation - 2mix-TMT time series

In this experiment 8 biological replicates were allocated to 2 TMT mixtures in an unbalanced [time course] design. More details are in Supplementary Sec. 2.2.2.

Experimental Design: Maculins et al. [14] studied primary murine macrophages infected with *Shigella flexneri* (*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 uninfected 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).

The proteome was quantified in 11-plex isobaric multiplexing via tandem mass tagging (TMT) in combination with LC-MS/MS. The six conditions were split between 11 channels, and as the result the experiment was unbalanced. Each mixture contained two replicates per early and late WT and KO conditions. Mixture one contained one replicate of uninfected WT and two replicates of uninfected KO. Mixture two contained one replicate of uninfected KO and two uninfected WT. [was there a reference channel?]

Global Profiling Run: This experiment included a separate global profiling run with the same design to measure unmodified peptides. There was a 90% overlap between the identified background modified peptides and proteins that were quantified in the global profiling run.

Time series comparison: 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

In this experiment 8 biological replicates were split between 4 conditions in a balanced [group comparison] design. Further details in Suppler Sec 2.2.3.

Experimental Design: Cunningham et al. [15] 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. Each condition had two biological replicates. The abundance of modified peptides was quantified with label-free LC-MS/MS.

Global Profiling Run: This experiment did not include a separate global profiling run to measure unmodified peptides. In this case *MSstatsPTM* could still be used by extracting unmodified peptides from the modified run. However, 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.

Pairwise comparisons: 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 an LC-MS/MS experiment with I conditions and J mass spectrometry runs (technical [of biological?] replicates) per condition in the experiment. For one protein, the PTM site is represented by K spectral features (peptide ions, distinguished by their cleavage residues and charge states). The log-intensity (base 2) of Feature k, in Run 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 peptides in the same run is denoted by y_{ijl}.

Figure 1(a) schematically illustrates this data structure for one protein in a simplified experiment with $I = 2$ conditions. [Describe the population parameters in the figure (quantities of interest]. The data are collected in $J = 2$ [biological?] replicate runs [describe how the data are representative of the population parameters of interest] The number of quantified features vary across replicate LC-MS/MS runs of the same sample, and across conditions. [There may be outliers]. Some spectral features can be missing.

[in Figure 2 b, there are two quantities of interest, stated at the population level. Describe the quantities, what is of interest, why relevant]

Existing statistical methods for experiments targeting PTMs

Two-sample t-test [When describing the existing methods, could you refer to Fig 1? Here is the suggested order: (1) What is the null hypothesis? Do these methods test the same null hypothesis? (If not, it may be a good idea to add this null hypothesis to the figure for contrast) (2) What are the data (how summarized prior to the testing); (3) what is the model assumptions? (4) how inference is performed] Many investigations perform differential abundance analysis of PTMs using two-sample t-test or its extensions.

[These sentences were combined from different places. Could you edit?] The approach takes as input intensities of individual features from modified peptides, or intensity ratios of modified and unmodified peptide features, and compares the mean abundance of a PTM site from one condition to another [16][17]. Two-sample t-test is based on the null hypothesis that there is no difference in mean PTM abundance between Conditions i and i^*. The abundance in each run is taken as input and is often estimated by sum of peak intensities. For example, the log-abundance estimate for the PTM in Run j of Condition i is given by

$$\log\left(\sum_{k=1}^{K} 2^{y^*_{ijk}}\right).$$

For adjustment with respect to unmodified peptides, the estimate of PTM abundance is divided by the protein abundance estimate, and the t-test for the adjusted PTM abundance on log scale takes as input the

difference of their log-estimates. The quantity is denoted by d_{ij} and is given by

$$d_{ij} = \log\left(\sum_{k=1}^{K} 2^{y^*_{ijk}}\right) - \log\left(\sum_{l=1}^{L} 2^{y_{ijl}}\right).$$

Alternatively, run-level summarization can also be used. The difference between the means of PTM abundance in Conditions i and i' are estimated as

$$\hat{\Delta} = \frac{1}{J} d^*_{i+} - \frac{1}{J} d_{i'+},$$

where $d^*_{i+} = \sum_{j=1}^{J} d^*_{ij}$ and the test statistic for the t-test is given by $\hat{\Delta}/\mathrm{SE}(\hat{\Delta})$. The statistical significance of the difference is determined by comparing the test statistic against the t distribution, with degrees of freedom $df = 2J - 2$ in balanced designs.

Limma [Could you fine-tune, using the same order as above for Limma? I.e., null hypothesis, the data, the model, the inference] Modifications of the t-test such as moderated t-test with Limma were also proposed [9][18]. While simple, these approaches do not fully account for the sources of variations, and are not directly applicable to experiments with complex designs, e.g., comparisons of multiple conditions, acquisition in multiple batches, etc. Additionally, while these approaches can be applied to experiments studying PTMs, there is not a self contained, straight forward implementation of the methods, making application challenging.

Limma uses linear models to test the null hypothesis that there is no difference in mean PTM abundances between conditions. It leverages Empirical Bayes moderation to share pooled variance information across individual modification models and moderate the individual residual variances. Using a linear model allows Limma to share variance information across conditions, providing a more accurate estimate [9].

With respect to PTM analysis, feature level summarization and adjustment for changes in the unmodified peptides are performed in the same way as t-test. After fitting a linear model, the final variance is moderated using the global variance across all modified peptides. This is done as $s_i = P(v_i|v_\Theta)$ where v_i is the resulting variance when modeling PTM i and v_Θ is the global variance.

Isobar-PTM was also proposed for experiments with LC-MS/MS quantitative strategies that employ isobaric labels such as tandem mass tags (TMT) or isobaric tag for relative and absolute quantification (iTRAQ)[19]. 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.

Statistical modeling and parameter estimation in MSstats

[Could you also refer to Figure 1? Essentially, MSstats only analyzes a part of the figure. Refer to the null hypothesis. Then describe the summarization (including AFT etc), and the model for he summarized data (the fact that different designs require different models - this is your argument for the limitation of the methods above.)] The proposed approach leverages the large amount of work done to create *MSstats* [10]. *MSstats* takes as input a list of log-transformed intensities of spectral features which are used to characterize the identified protein. For each protein, the feature log-intensities are expressed using a linear mixed model in consideration of the effects of condition, run, feature and interaction between run and feature. The model parameters are estimated using a split-plot approach, where the feature log-intensities are first summarized into a single value per protein per run in the subplot model, and the protein-level summaries are then used for the inference of the protein abundance. This approach allows the method to be extended to cases where the experimental design is unbalanced and where additional sources of variation are present. In the protein-level summarization, Tukey's median polish (TMP), a simple and robust procedure, is applied to iteratively fit a two-way additive model with the effects of run and feature, which in turn summarizes the log-intensities

for each protein [20]. After summarization, the inference of the protein abundance in each condition is carried through fitting a model based on the family of linear mixed-effects models [21] [22]. We leverage this statistical modeling and quantification workflow and apply it to PTMs, in addition to unmodified proteins.

Additionally, *MSstats* provides a method for imputing missing feature intensities if desired. When values are imputed, it is assumed they are missing for reasons of low abundance. Missing features are imputed in each MS run using the Accelerated Time Failure (AFT) model [20]. In order to impute a feature's missing values, the feature must be present in at least one MS Run. If the feature is not present across all MS Runs, the value will be left missing. Missing value imputation is done before summarization with Tukey's Median Polish in order to correct for the impact of outliers. Missing value imputation is done separately for both the PTM and global protein datasets.

Results

Detection of changes in PTMs - Balanced design with one source of variation

We propose a statistical method for detecting changes in PTMs while removing confounding with changes in global protein abundance. For simplicity, we discuss the method in the special case of a balanced design with multiple conditions and one technical replicate. However, in practice the experimental design can be unbalanced and variation can come from multiple sources, as shown in the next section.

[Put it in the context of the figure again, and also in the context of the existing methods in Background - how is it similar to the existing methods, and how different] The null hypothesis states that there is no difference in log-abundance of the PTM site between conditions, adjusted with respect to protein abundance. The expected values of log-abundances of the PTM and protein in Condition i are denoted by μ_i^* and μ_i respectively. The feature and run level summarization [Summarized how?] for condition i is denoted as $\hat{y}_{i+}^*$ for the modified peptide and $\hat{y}_{i+}$ for the unmodified peptide. For two conditions i and i' the hypothesis test is formulated as follows:

$$\begin{aligned} H_0 &: \Delta = (\mu_i^* - \mu_{i'}^*) - (\mu_i - \mu_{i'}) = 0 \\ H_a &: \Delta = (\mu_i^* - \mu_{i'}^*) - (\mu_i - \mu_{i'}) \neq 0 \end{aligned} \tag{1}$$

Specifically, the adjusted difference is given by the difference in log-abundance of the PTM site, subtracted by the difference in log-abundance of the underlying protein, which is equivalent to the log of the ratio of PTM abundance difference to protein abundance difference.

$$\hat{\Delta} = \left[\frac{1}{J}\left(\hat{y}_{i+}^* - \hat{y}_{i'+}^*\right)\right] - \left[\frac{1}{J}\left(\hat{y}_{i+} - \hat{y}_{i'+}\right)\right] \tag{2}$$

The estimate of the adjusted difference and the standard error (SE) of the estimate are obtained by combining the difference estimates and the associated SEs from both counterparts.

$$SE(\hat{\Delta}) = \left[\left(\hat{\sigma}_{\gamma^*}^2 + \hat{\sigma}_{\gamma}^2\right)\right]^{1/2} \tag{3}$$

Finally the degrees of freedom are calculated as follows.

$$\left(\hat{\sigma}_{\gamma^*}^2 + \hat{\sigma}_{\gamma}^2\right)^2 \Bigg/ \left(\frac{\hat{\sigma}_{\gamma^*}^4}{\mathrm{df}(\gamma^*)} + \frac{\hat{\sigma}_{\gamma}^4}{\mathrm{df}(\gamma)}\right) \tag{4}$$

The test statistic for the hypothesis testing is the ratio of the estimate of the adjusted difference to its SE. To determine the statistical significance of the difference in terms of p-value, the test statistic is compared against the t distribution with degrees of freedom approximated by the Satterthwaite method [23]. Adjustment for multiple comparisons is performed using the Benjamini-Hochberg procedure to control the false discovery rate at a desired level, e.g., 0.05 [24].

Design of PTM experiments in the balanced case

The proposed statistical framework allows for design of PTM experiments in terms of sample size calculation and power analysis. Sample size calculation takes as input a) q, the desired false discovery rate, b) β, the average Type II error rate, c) Δ, the minimal log-fold change in adjusted PTM abundance that we would like to detect, d) $m_0/(m_0+m_1)$, the fraction of truly differentially modified PTM sites in the comparison, and e) $\sigma^2_{\gamma^*}$ and σ^2_{γ}, the anticipated variances associated to modified and unmodified peptide features, respectively. The variances can be derived based on the dataset being analyzed, assuming similar quantitative properties and variations. With these values and a user-specified number of conditions, the corresponding number of technical replicates per condition can then be derived, as described in [25]. Given the above quantities, the minimal number of replicates J is determined by the variance of the estimated log-fold change $\mathrm{SE}^2(\hat{\Delta})$ as

$$\mathrm{SE}^2(\hat{\Delta}) = \left[\frac{2}{J}\left(\hat{\sigma}^2_{\gamma^*} + \hat{\sigma}^2_{\gamma}\right)\right] \leq \left(\frac{\Delta}{t_{1-\beta,df} + t_{1-\alpha/2,df}}\right)^2 \tag{5}$$

where

$$\alpha = (1-\beta)\cdot\frac{q}{1+(1-q)\cdot m_0/m_1} \tag{6}$$

and $t_{1-\beta,df}$ and $t_{1-\alpha/2,df}$ are the $100(1-\beta)^{\text{th}}$ and the $100(1-\alpha/2)^{\text{th}}$ percentiles of the t distribution, with $df = I(J-1)$ degrees of freedom in balanced designs.

Solving for J, the number of biological replicates per group is calculated as

$$J \geq \frac{(2\hat{\sigma}^2_{\gamma^*} + 2\hat{\sigma}^2_{\gamma})(t_{1-\beta,df} + t_{1-\alpha/2,df})^2}{\Delta^2} \tag{7}$$

More details on sample size calculation can be found in [26].

Detection of changes in PTMs - Extension to complex designs

The statistical modeling approaches discussed above can be extended to complex designs including experiments with additional sources of variation and unbalanced designs. In this case our null hypothesis and Equation 1 are unchanged.

Model inference is done in the same way to how MSstats [10] and MSstatsTMT [11] target the global protein. In cases where the design is unbalanced, a linear mixed effects model is fit, which takes into account all potential sources of variation, and restricted maximum likelihood (RML) is used to estimate the parameters of each model. Once the parameters are estimated, we can combine the models using modified versions of Equations 2, 3, and 4.

The log-fold change in the adjusted PTM abundance, Δ, is now estimated by RML, and Equation 2 becomes

$$\hat{\Delta} = (\hat{\mu}^*_{RML_i} - \hat{\mu}^*_{RML_{i'}}) - (\hat{\mu}_{RML_i} - \hat{\mu}_{RML_{i'}}) \tag{8}$$

The standard error of the estimate $\mathrm{SE}(\hat{\Delta})$ and degrees of freedom in Equations 3 and 4 are unchanged from the balanced case, although $\hat{\sigma}_{\gamma^*}$ and $\hat{\sigma}_{\gamma}$ are now estimated using RML.

Details on how the proposed method can be used to run sample size calculations and power analysis on experiments with complex design can be seen in Supplementary Sec. 1.1

Implementation

The proposed methods are implemented in the open source R package *MSstatsPTM*, available on Bioconductor. *MSstatsPTM* includes converters for multiple spectral processing tools, including MaxQuant, Progenesis, and Spectronaut. The converters take as input the raw data from the tool, identify the modification site for modified peptide, and put the data into the correct format for analysis in *MSstatsPTM*. Conversion is done separately for the modified and global profiling runs. If the global profiling run is not available, the package can still analyze the modified run, but will do so without adjusting for changes in

unmodified protein abundance. Specifically, if a global profiling run is not available the null and alternative hypothesis in Equation 1 will reduce to:

$$\begin{aligned} H_0 &: \Delta = \mu_i^* - \mu_i = 0 \\ H_a &: \Delta = \mu_i^* - \mu_i \neq 0 \end{aligned} \tag{9}$$

Which amounts to running the methods seen in MSstats and MSstatsTMT on peptide level data [10] [11].

After using the converters, the next step is peptide/protein summarization and missing value imputation. In the methods sections it was assumed that all the input data was summarized; in the implementation we need to run a specific summarization step to make this happen. For the modified run, the package summarizes features, PSMs, which include the same modification together. Features with multiple modification sites are not included with single site features and are summarized separately. For the global profiling run all unmodified features from the same protein are summarized together, up to the protein level. Additionally, the summarization includes global median normalization and normalizes between MS runs. The package uses an AFT model to impute missing values, although this step is optional.

The final step of the package is modeling the summarized dataset. A linear mixed effects model is fit for both the summarized modified and global profiling runs. This model is automatically adjusted depending on the experimental design and acquisition method. The comparisons of interest can either be predefined or a full pairwise comparison will be tested. After fitting a model to both the modified and unmodified data, the modified model is adjusted for changes in unmodified protein abundance, using the methods described in this paper.

Beyond the core functionalities of conversion, summary, and modeling, the package also includes functions for plotting the results. These include plots for the summarized and modeled data to assist in the analysis of the experiment. The summarized plots help with quality assurance analysis and identifying sources of variation. This includes a quality control plot, summarizing the peptide abundance per run in the form of a boxplot, and a profile plot, plotting each feature and the overall feature summarization as a line plot over each run. Additionally, the model plots include a volcano plot, showing all peptides adjusted p-values and fold changes, as well as a heatmap, which evaluates the fold change between conditions and peptides.

The package relies on functionalities from the R packages $MSstats$ [10] and $MSstatsTMT$ [11]. The statistical modeling relies on the functionality from the R packages $lme4$ [27] and $lmerTest$ [28]. An overview of the steps of the package are illustrated in Figure 2.

The code is available on Bioconductor, `http://www.bioconductor.org/packages/release/bioc/html/MSstatsPTM.html`, and Github, `https://github.com/Vitek-Lab/MSstatsPTM`.

Evaluation

The performance of the proposed method was evaluated on simulated and spike in datasets with known ground truth, as well as biological experimental data where the ground truth was not known. For the experiments where ground truth was known, we calculated the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). True positives were modified peptides that were differential after accounting for changes in global protein abundance. True negatives were modified peptides that were not differential after accounting for changes in global protein abundance. To determine positives and negatives we adjusted the p-values for multiple testing and used an $\alpha = .05$ cutoff to determine differential sites. These results were then compared to the ground truth to determine TP and TN. We used these results to calculate the false discovery rate (FDR = FP/(TP+FP)), recall rate (TP/(TP + FN), and accuracy ((TP + TN) / (TP + TN + FP + FN)). Finally, the summary statistics were compared across methods to analyze method performance.

For biological experiments where ground truth was not known, we adjusted the p-values for multiple testing and used an $\alpha = .05$ cutoff to determine differential sites. Comparisons were made before and after adjusting for changes in unmodified protein abundance and differences between the comparisons were noted. Each biological dataset presented a different experimental scenario, showing the applicability of the

proposed method on a variety of experimental designs. All datasets in this paper were evaluated using the $MSstatsPTM$ package.

The proposed approach corrected for high FDR in computer simulations

The proposed approach was evaluated against two-sample t-test and Limma in computer simulations. In Figure 3a we see that not accounting for changes in the unmodified protein resulted in a high false discovery rate. After adjusting for changes in the unmodified protein, all methods performed similarly in terms of FDR. In a clean simulation, recall and accuracy of the proposed approach and Limma performed similarly, although Limma performed slightly better in simulations with fewer biological replicates. In contrast, the performance of the t-test method lagged behind the others. Two-sample t-test only uses data within the groups of interest while ignoring the remainder of the data, leading to similar performance across all simulations. In contrast, the proposed approach and Limma leverage all available information, which resulted in improved power in simulations with more conditions.

The proposed approach outperformed existing methods in simulations with missing values and low features

In this simulation the proposed method outperformed t-test 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 3b. The proposed method calibrated model accuracy and recall rate well, even when the number of replicates were low. Additionally, when comparing the fold change estimation across all modified peptides, the proposed method showed a tighter distribution of estimated fold changes around the true fold change, as compared to t-test and Limma. Specifically, the inner quartile range of the estimated fold change for the proposed method was on average 10.4% smaller than t-test and 21.8% smaller than Limma. This fold change comparison can be seen in Supplementary Figure S6.

Further details reviewing the results of each simulation and method comparison can be seen in Supplementary Sec. 2.1.

In a label-free benchmark experiment the proposed approach corrected for bias and outperformed existing methods

In this experiment all models incorrectly estimated the fold change of the modified spike-in peptides before adjusting for changes in global protein abundance. After adjustment, the spike-in peptides' fold change was generally in line with expectation (Figure 4a), however the distribution of estimated fold changes was visibly wider in all methods. Of the three approaches, the proposed method showed the tightest distribution around the true log fold change. Comparing the inner quartile range (IQR) of the spike-in peptide's log fold change, the proposed method's IQR was 36.78% smaller than Limma's and 32.98% smaller than t-test's. This means that the proposed method's fold change estimate was more consistently correct, when compared to Limma and t-test.

In Figure 4b we can clearly see the red labeled spike-in peptides do not follow the expected fold change before adjusting for changes in global protein abundance. After adjustment, the estimated fold change was more in line with expectation. Additionally, the background peptides serving as the null model, show many false positives before adjustment was made. After adjustment the results improved and the number of false positives decreased significantly. Specifically, for the proposed method, the number of false positives went from 20.88% to 1.84% of all true negatives after adjustment was applied. Using both the proposed method and Limma, the red labeled spike-in peptides were more inline with expected fold change after the adjustment was applied. However, using Limma resulted in the majority of the differential spike-in peptides not showing a significant adjusted p-value. In this case, using Limma would have resulted in missing the majority of differentially modified peptides.

Further results of this experiment are provided in Supplementary Sec. 2.1.4.

MSstatsPTM corrected for bias in group comparison TMT experiment

The results of this experiment are summarized in Figure 5. The experiment was modeled as a group comparison. In Figure 5a the number of significant modified peptides before and after adjustment is shown. We can see that more peptides became insignificant after adjustment than became significant. 3,236 modified peptides became insignificant, 1352 became significant, while 4,282 were significant in both models. For the peptides that became insignificant in the adjusted model, their change in abundance was mainly due to changes in global protein abundance. In contrast, for peptides that became significant after adjustment, their true abundance change was masked by underlying changes in the unmodified protein. Both of these issues are corrected in adjustment, and the true abundance change is shown. An additional question that must be addressed is if the decrease in significant 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 insignificant after adjustment. In other words, the fold change was the same between models but variance increased. When this test was applied on this experiment, only one peptide became insignificant due to an increase in variance. Thus we can conclude that the drop off in significant peptides was due to changes in global protein abundance.

In Figure 5b, the modification of protein $GSDMD$ at $K62$ showed the advantage of the proposed method. The modified peptide originally showed no abundance change between the infected 1 hour, 4 hour, and 6 hour conditions. This was contrasted with a strong negative change in the global profiling run between the same conditions. Looking at the Dox4hr vs Dox1hr conditions and modeling the modified peptide without adjusting for changes in the global profiling, the fold change was $-.501$ and the adjusted p-value was insignificant at .0644. After adjusting for changes in the global protein abundance, the fold change was much higher, 2.79, and the adjusted p-value became significant, $5.25e-8$. In this case the effect of the modified peptide was strongly confounded with changes in the global protein. The proposed method allowed us to remove this confounding and estimate the true effect.

Proposed approach removed confounding in time series TMT experiment

The results of this experiment are summarized in Figure 6. The experiment was modeled as a time series, with the same biological replicate measured at each time point. Modeling the experiment as a time series lowered the model variance as compared to modeling the experiment as a group comparison. In Figure 6a the number of significant modified peptides before and after adjustment is shown. Again more PTMs became insignificant after adjustment than became significant. 19,286 peptides became insignificant, 4,947 became significant, and 41,552 were significant in both models. Again we tested if the decrease in significant peptides was due to increased variance or if it was due to removing convolution with changes in the umodified protein. This was tested by looking for modified peptides whose adjusted log fold change was within 10% of the unadjusted log fold change but became insignificant after adjustment. When this test was applied, 548 peptides became insignificant due to an increase in variance. This is only 3.4% of the total peptides that became insignificant after adjustment. Thus we can conclude that the drop off in significant peptides was mainly due to changes in global protein abundance.

In Figure 6b the profile plot of protein TTP modified at site $S178$ showed the power of the proposed method. Without adjustment, there was a large positive log fold change of 2.9 between the WT_Late and WT_Uninfected conditions. However, the global profiling run showed a similar log fold change of 2.014 between the same conditions. This indicated that the abundance change in the modified peptide is nearly entirely due to changes in the global protein. When adjusting for the global protein the modified peptide's adjusted p-value became insignificant, going from .0009 to .248. Correcting for changes in global protein abundance allowed us to see the true impact of the modification at $S178$ which would have otherwise been challenging to perceive.

Analysis of other modifications in this experiment can be seen in Supplementary Sec. 2.2.2.

In label-free experiment without a separate global profiling run, the proposed approach could eliminate the bias 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 modified and unmodified peptides. Any modified peptide without a corresponding unmodified protein could not be adjusted. Of the 10,799 modified peptides identified, only 4526 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.

The results of this experiment are summarized in Figure 7. After adjusting for changes in the unmodified protein, there were fewer significant modified peptides. In total, 2709 modified peptides became insignificant, while only 547 became significant. However, this was mainly due to not having a global profiling run, resulting in a lack of overlap between modified and unmodified peptides. In the bottom plot, only modified peptides that could be adjusted are shown. Here there were much fewer peptides that become insignificant after adjustment. 726 modified peptides became insignificant, 547 became significant, and 1,078 were significant in both models. As in the previous experiments, we can check if these peptides became insignificant due to an increase in variance. To check this we looked at modified peptides whose adjusted fold change was within 10% of their unadjusted fold change and which became insignificant after adjustment. In this experiment there were only 25 modified peptides that meet this criteria.

Further analysis of modifications before and after protein adjustment can be seen in Supplementary Sec. 2.2.3.

Noisy PTM measurements benefited from additional biological replicates

[I would like to move this section to the supplementary and just referencing it in the main.]

Here we analyzed the sample size needed to achieve a desired statistical power. The proposed approach corrected for confounding between the modified peptide and unmodified protein at the cost of increased variation. This can be seen in the calculation for variance in Section 2. Increased variation required a larger number of replicates to reach the same power. Thus the statistical power was dependent on the variance from both the modified peptide and unmodifed protein.

We compared the statistical power in experiments with differing numbers of replicates, variance, and fold change for both the modified and unmodified runs. In terms of the number of replicates, we tested scenarios with equal replicates in both the modified and unmodified runs, as well as scenarios where the replicates differed between runs. We used the biological experiments to determine what variance values to test. In datasets 4 and 5 the variance of the PTM was higher than the global protein. In dataset 6 the variance of the PTM and Protein was generally the same. We mimicked these scenarios and analyzed the power of experiments when the PTM variance was higher than the protein and when they were equal. When the PTM and protein were the same we chose a variance of .15, whereas when the PTM was higher than the protein we chose a PTM variance of .2 and a protein variance of .1.

The results of the power and sample size analysis can be seen in Supplementary Figure S16. When the variance and replicates were equal, higher replicates predictably lead to higher power. In cases where the replicates were unbalanced, but the variance was still the same, it did not matter if there were more replicates in the modified or unmodified runs. In comparison, with differing variance and equal replicates, higher replicates still lead to higher power. When the replicates were unbalanced and the variance was higher in the PTM, there was more power when there were more replicates in the PTM than the protein. In this case it was clearly more important to have high replicates for the PTM run than the unmodified protein. In cases where the number of replicates has to be limited, it was better to add more to the PTM side.

Discussion

[Although demonstrated here on DDA, is also applicable to DIA, SRP and PRM acquisitions]

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 convoluted 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 t-test 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. For example, it can represent experimental designs with even more complex structures, such as time series or factorial investigations. 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 convoluted with each other if there are two or more modification sites per peptide. 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 convolution. 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 most 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 deconvolute the effect of individual PTMs in peptides with multiple modifications.

Data Availability

[This overlaps with the Datasets section. I would merge all of the below to the experimental data section] All datasets for the four experiments and two simulations, including raw files and analysis results, are available online.

The details of the four experiments, including raw data, and intermediate data processing outputs are located in individual MassIVE repositories, referenced in 2.1.

The R scripts and results of the statistical analysis for the four experiments are available in MassIVE.quant reanalysis containers: SpikeIn benchmark - Ubiquitination - Label-free (MassIVE reanalysis container identifier: TBD), Human - Ubiquitination - 1mix-TMT (MassIVE reanalysis container identifier: RMSV000000356), Mouse - Phosphorylation - 2mix-TMT time series (MassIVE reanalysis container iden-

tifier: RMSV000000357), Human - Ubiquitination - Label-free no global profiling run (MassIVE reanalysis container identifier: RMSV000000358).

The code, data, and analysis results of the two simulations, Computer simulation 1 - Label-free and Computer simulation 2 - Label-free missing values and low features, are available on Github: MSstatsPTM_simulations.

References

[1] Y. L. Deribe, T. Pawson, and I. Dikic. "Post-translational modifications in signal integration". In: *Nature Structural & Molecular Biology* 17 (2010), pp. 666–672. DOI: https://doi.org/10.1038/nsmb.1842.

[2] P. Cohen. "The regulation of protein function by multisite phosphorylation–a 25 year update". In: *Trends in Biochemical Sci.* 25.12 (2000), pp. 596–601. DOI: https://doi.org/10.1016/S0968-0004(00)01712-6.

[3] M. Mann and O. Jensen. "Proteomic analysis of post-translational modifications". In: *Nature Biotechnology* 21 (2003), pp. 255–261. DOI: https://doi.org/10.1038/nbt0303-255.

[4] N. A. Petushkova et al. "Post-translational modifications of FDA-approved plasma biomarkers in glioblastoma samples". In: *PLOS ONE* 12.5 (2017), e0177427. DOI: https://doi.org/10.1371/journal.pone.0177427.

[5] L. Käll and O. Vitek. "Computational Mass Spectrometry–Based Proteomics". In: *PLoS Computational Biology* 7.12 (2011). DOI: https://doi.org/10.1371/journal.pcbi.1002277.

[6] Roepstorff P. "Mass spectrometry in protein studies from genome to function". In: *Current Opinion in Biotechnology* 8.1 (1997), pp. 6–13. DOI: https://doi.org/10.1016/S0958-1669(97)80151-6.

[7] J. Huang et al. "Enrichment and separation techniques for large-scale proteomics analysis of the protein post-translational modifications". In: *Journal of Chromatography A* 1372 (2014), pp. 1–17. DOI: https://doi.org/10.1016/j.chroma.2014.10.107.

[8] J. Olsen and M. Mann. "Status of Large-scale Analysis of Post-translational Modifications by Mass Spectrometry". In: *Molecular & Cellular Proteomics* 12.12 (2013), pp. 3444–3452. DOI: https://dx.doi.org/10.1074%2Fmcp.O113.034181.

[9] M. E. Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7 (2015), e47. DOI: https://doi.org/10.1093/nar/gkv007.

[10] M. Choi et al. "MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments". In: *Bioinformatics* 30.17 (2014), pp. 2524–2526. DOI: https://doi.org/10.1093/bioinformatics/btu305.

[11] T. Huang et al. "MSstatsTMT: Statistical Detection of Differentially Abundant Proteins in Experiments with Isobaric Labeling and Multiple Mixtures". In: *Molecular & Cellular Proteomics* 19.10 (2020), pp. 1706–1723. DOI: https://doi.org/10.1074/mcp.ra120.002105.

[12] M. Choi et al. "MassIVE.quant: a community resource of quantitative mass spectrometry–based proteomics datasets". In: *Nat Methods* 17 (2020), pp. 981–984. DOI: https://doi.org/10.1038/s41592-020-0955-0.

[13] G. Luchetti et al. "Shigella ubiquitin ligase IpaH7.8 targets gasdermin D for degradation to prevent pyroptosis and enable infection". In: *Cell Host & Microbe* 29.10 (2021), pp. 1521–1530. DOI: https://doi.org/10.1016/j.chom.2021.08.010.

[14] T. Maculins et al. "Multiplexed proteomics of autophagy-deficient murine macrophages reveals enhanced antimicrobial immunity via the oxidative stress response". In: *eLife* 10 (2021), e62320. DOI: https://doi.org/10.7554/eLife.62320.

[15] C. Cunningham et al. "USP30 and Parkin homeostatically regulate atypical ubiquitin chains on mitochondria". In: *Nature Cell Biology* 17.2 (2015), pp. 160–169. DOI: https://doi.org/10.1038/ncb3097.

[16] V. Schwämmle, T. Verano-Braga, and P. Roepstorff. "Computational and statistical methods for high-throughput analysis of post-translational modifications of proteins". In: *Journal of Proteomics* 129 (2015). Special Issue : Computational Proteomics, pp. 3–15. ISSN: 1874-3919. DOI: https://doi.org/10.1016/j.jprot.2015.07.016.

[17] S. P. Thomas et al. "A practical guide for analysis of histone post-translational modifications by mass spectrometry: Best practices and pitfalls". In: *Methods* 184 (2020), pp. 53–60. ISSN: 1046-2023. DOI: https://doi.org/10.1016/j.ymeth.2019.12.001.

[18] Y. Zhu et al. "DEqMS: A Method for Accurate Variance Estimation in Differential Protein Expression Analysis". In: *Molecular & Cellular Proteomics* 19.6 (2020), pp. 1047–1057. DOI: https://doi.org/10.1074/mcp.tir119.001646.

[19] F. Breitwieser and J. Colinge. "IsobarPTM: A software tool for the quantitative analysis of post-translationally modified proteins". In: *Journal of Proteomics* 90 (2013), pp. 77–84. DOI: https://doi.org/10.1016/j.jprot.2013.02.022.

[20] J. W. Tukey. *Exploratory data analysis*. Addison-Wesley, 1977.

[21] B. Bolker et al. "Generalized linear mixed models: a practical guide for ecology and evolution". In: *Trends in Ecology and Evolution* 24.3 (2009), pp. 127–135. DOI: https://doi.org/10.1016/j.tree.2008.10.008.

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

[23] F. E. Satterthwaite. "An approximate distribution of estimates of variance components". In: *Biometrics Bulletin* 2.6 (1946), pp. 110–114. DOI: https://doi.org/10.2307/3002019.

[24] Y. Benjamini and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *J.R. Statist. Soc. B* 57.1 (1955), pp. 289–300. URL: http://www.jstor.org/stable/2346101.

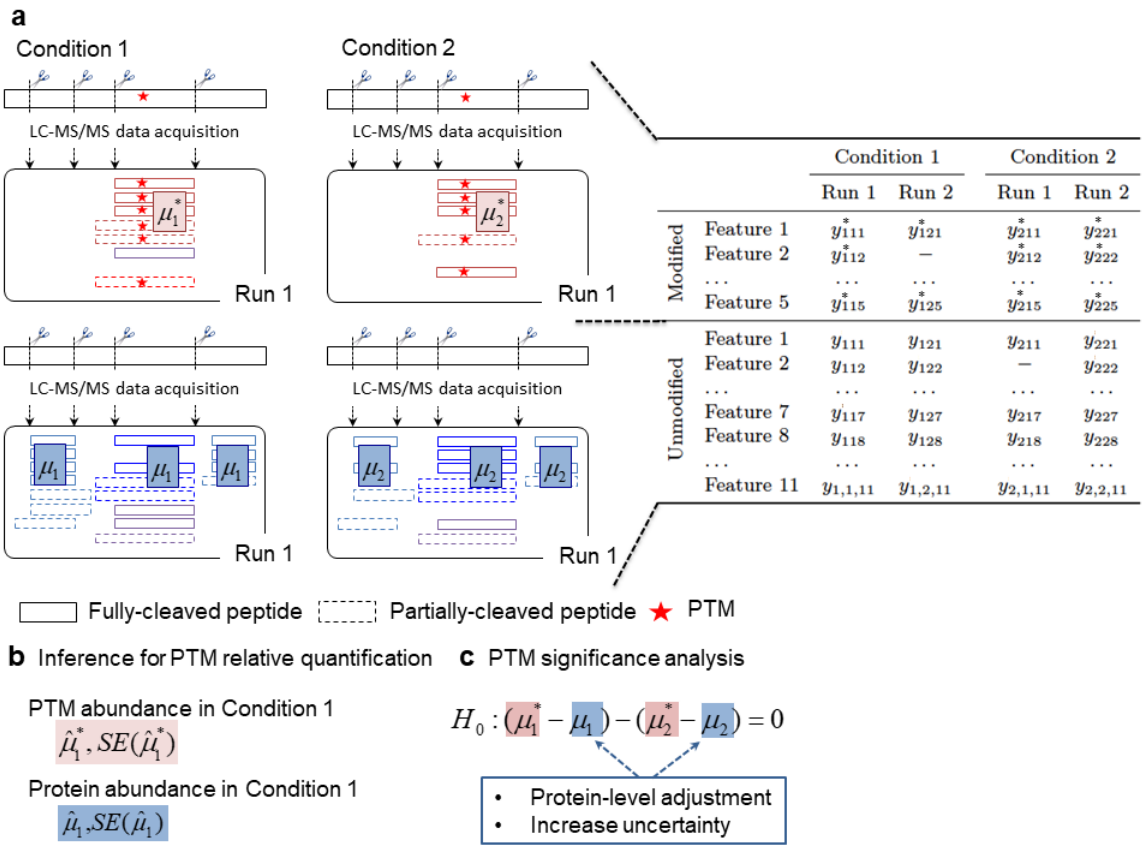
[25] M. H. Kutner et al. *Applied Linear Statistical Models*. 5th ed. McGraw-Hill/Irwin, 2004.

[26] A. L. Oberg and O. Vitek. "Statistical design of quantitative mass spectrometry-based proteomic experiments". In: *Journal of Proteome Research* 8.5 (2009), pp. 2144–2156. DOI: https://doi.org/10.1021/pr8010099.

[27] D. Bates et al. "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1 (2015), pp. 1–48. DOI: https://doi.org/10.18637/jss.v067.i01.

[28] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen. "lmerTest Package: Tests in Linear Mixed Effects Models". In: *Journal of Statistical Software* 82.13 (2017), pp. 1–26. DOI: https://doi.org/10.18637/jss.v082.i13.

Figures



		Condition 1		Condition 2	
		Run 1	Run 2	Run 1	Run 2
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 1: Goals of PTM characterization, input to statistical analyses, and notation. [Could you update the caption? The figure now has only 2 parts. It doesn't have subscript i etc. On the other hand, it distinguishes the population parameters, and the observed data. Could you mention this here briefly, and describe in more detail in the text?] (a) Schematic data representation, in a simplified case of two conditions and two [biological?] replicate runs. Each PTM site is modeled and characterized separately, where a PTM is quantified with multiple spectral features (boxes), distinguished by different charge states of a peptide. The feature intensities the PTM in Condition i are denoted by i*. Features corresponding to unmodified peptides in Condition i is denoted by i. Peptides can be fully cleaved (solid lines) and/or partially cleaved (dashed lines). (b) PTM relative quantification by statistical inference, which makes use of the feature intensities to infer the underlying PTM abundance and protein abundance with an estimate of associated uncertainty. (c) Model-based testing for differential PTM abundance, which corrects for the underlying protein abundance with a cost of increased uncertainty about the estimate of difference between conditions.

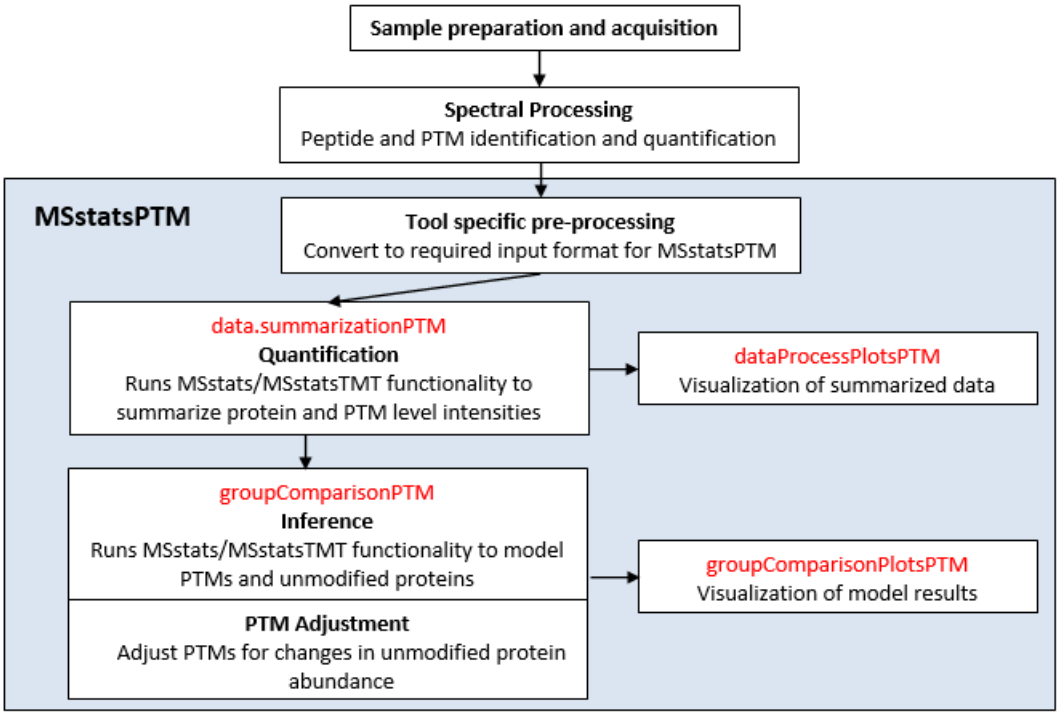
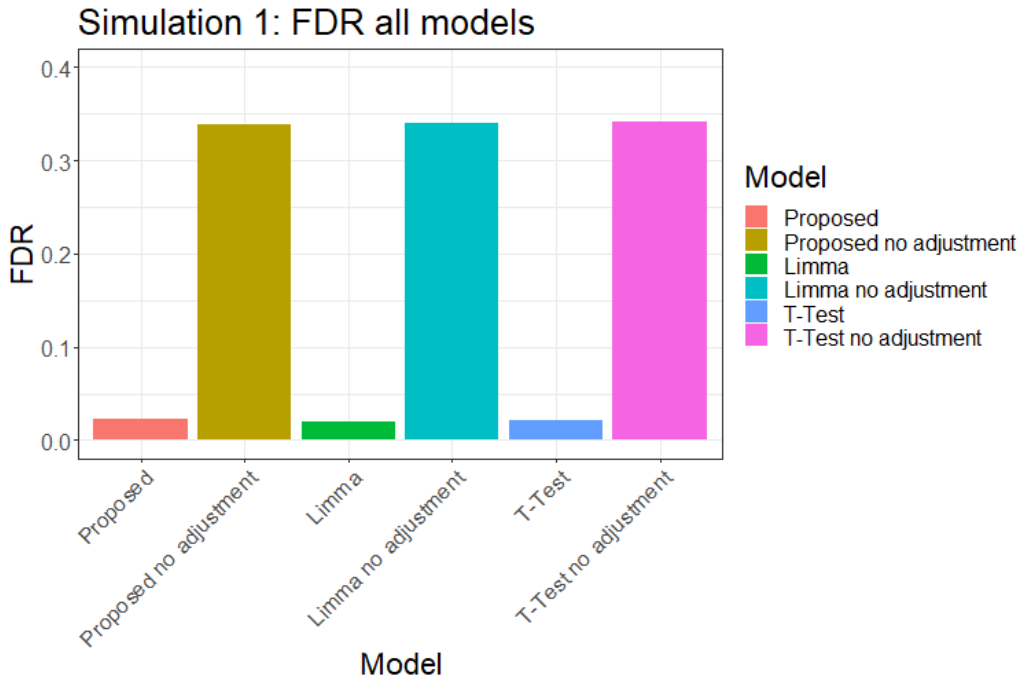
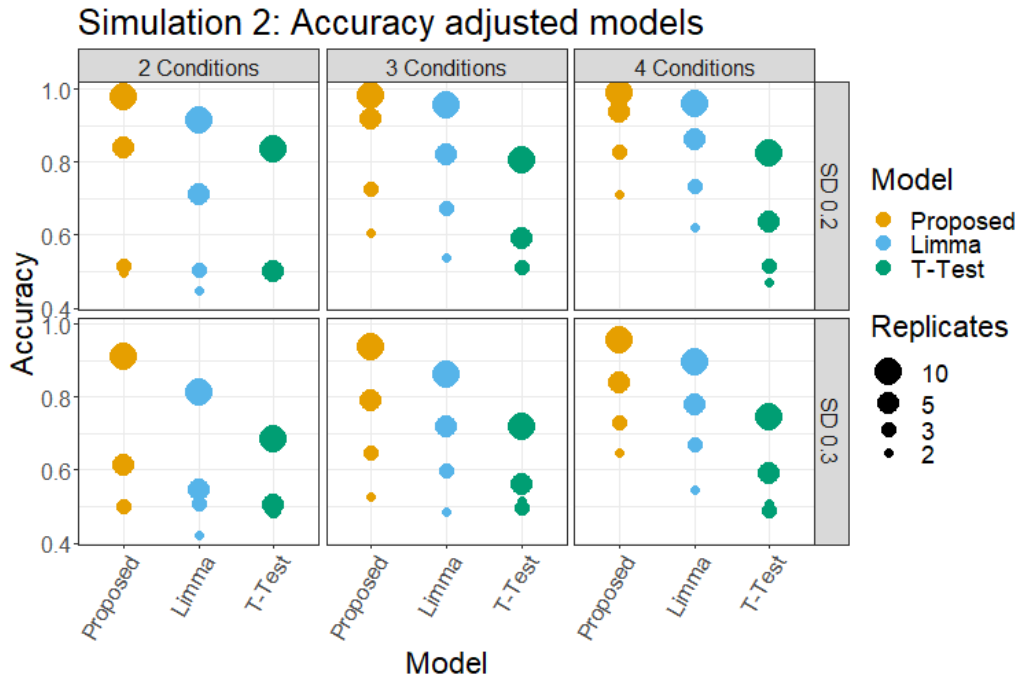


Figure 2: The workflow of MSstatsPTM and how it fits into the experimental pipeline. MSstatsPTM's workflow starts after modified and unmodified peptide quantification. First tool specific pre-processing is done, this includes modification site identification, general data cleaning, and formatting the data into the format needed for the package. The next step is feature level summarization, which summarizes features up to the modification level for the PTM data, and the protein level for the protein data. In the final step a model is fit to identify differential PTMs and unmodified proteins across conditions and the PTM model is adjusted for changes in the unmodified protein.

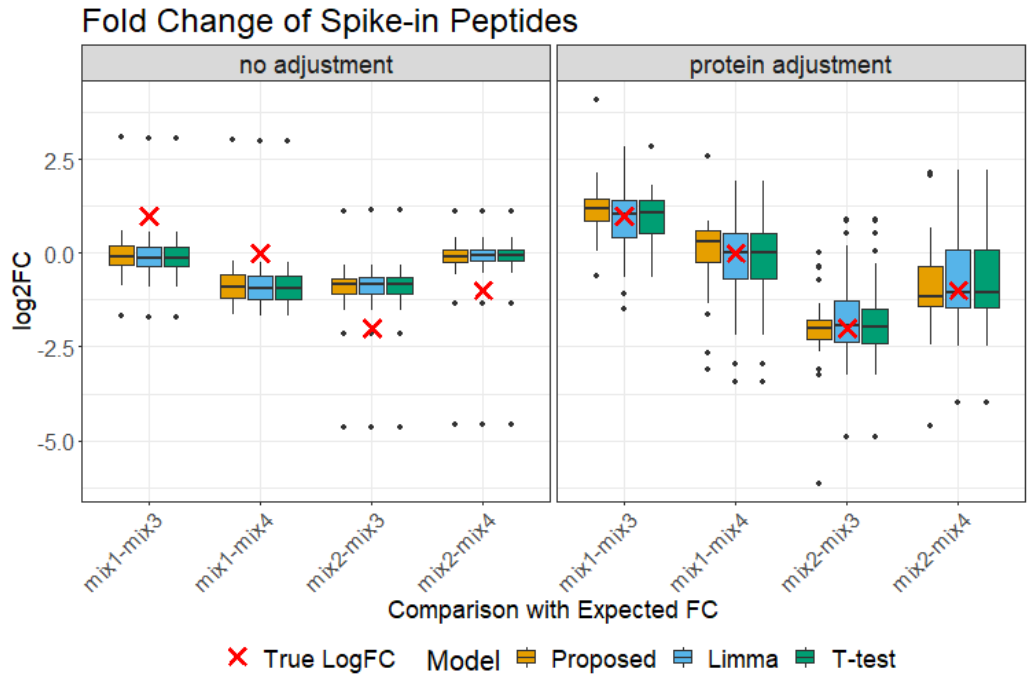


(a)

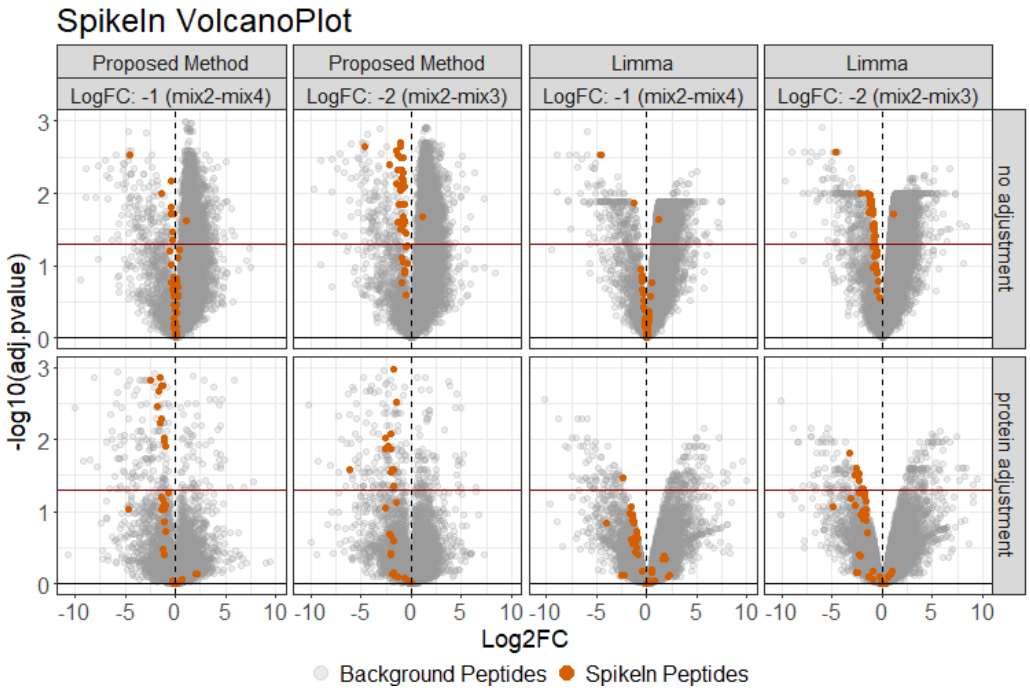


(b)

Figure 3: Dataset 1 & 2: Computer simulation. a) All the considered methods in the first computer simulation correctly calibrated FDR when adjusting for changes in protein abundance. In comparison, the methods without accounting for the protein-level changes resulted in off-target, high false positive rates. b) The advantage of using the proposed approach was apparent when including limited observations and missing values. Looking at accuracy, the proposed method outperformed Limma and *t*-test in nearly every model.

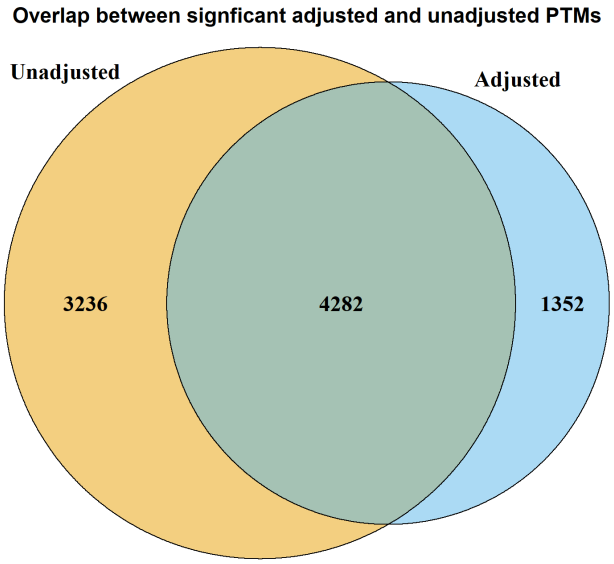


(a)

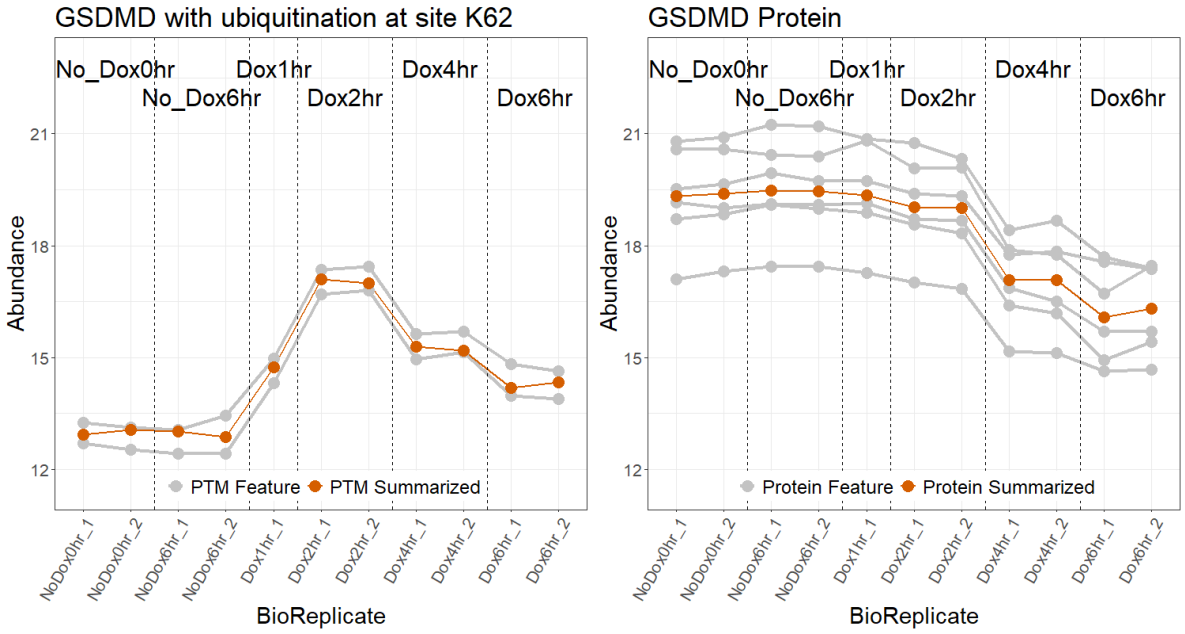


(b)

Figure 4: Dataset 3 : SpikeIn 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 t-test models was wider than the proposed method. b) The spike in peptides (colored red) did not follow the expected log fold change before adjustment. However, 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.

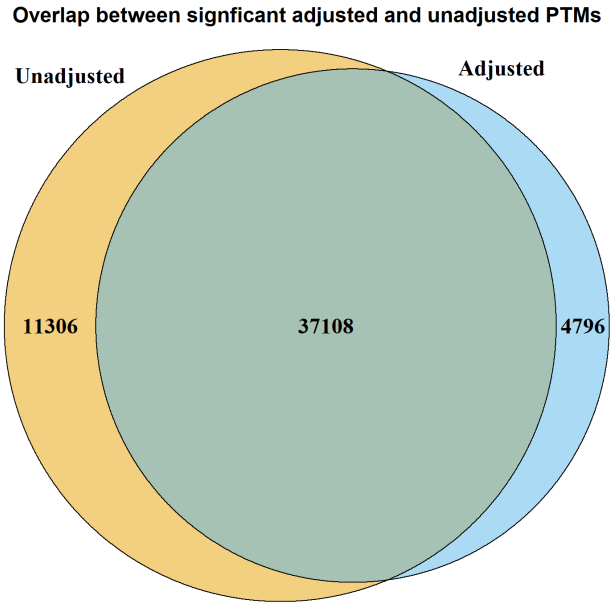


(a)

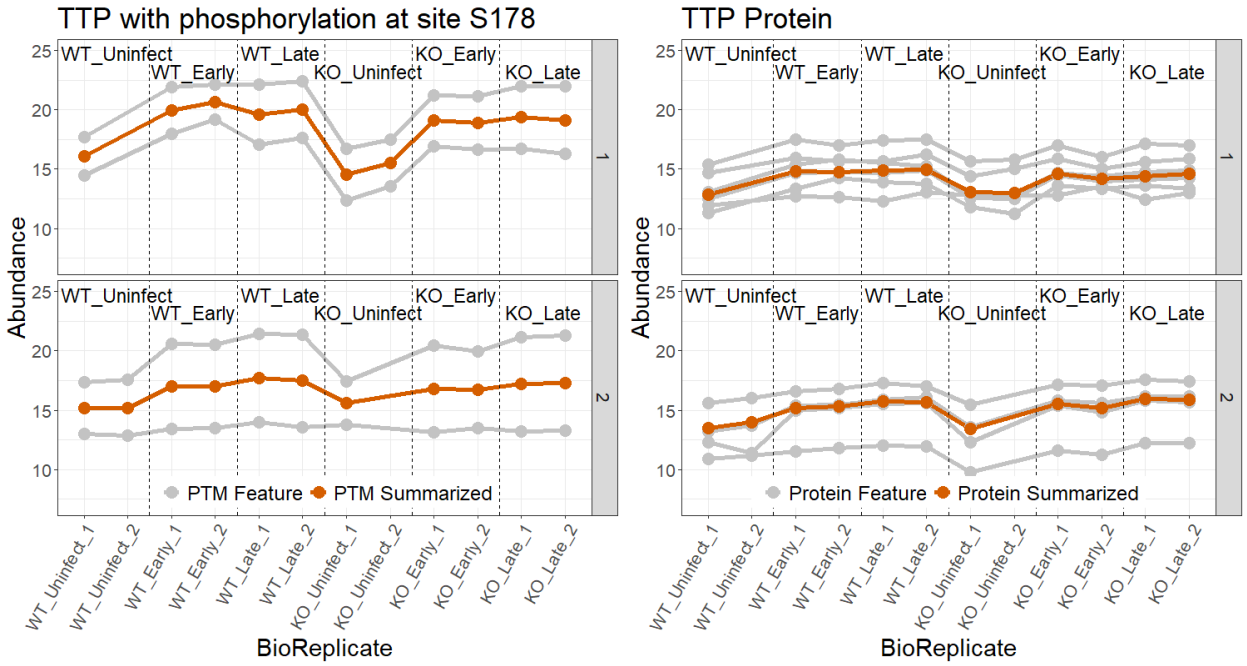


(b)

Figure 5: Dataset 4 : Human - Ubiquitination - 1mix-TMT. a) The overlap of differential modified peptides for the PTM model with and without global protein level adjustment. More PTMs became insignificant after adjustment then became significant. For the peptides that became insignificant in the adjusted model, their change in abundance was driven by changes in the global protein. In contrast, peptides that became significant after adjustment saw their true abundance change masked by underlying changes in the unmodified protein. b) Comparing the global profiling of protein $GSDMD$ with the ubiquitination of the protein at site $K62$. When looking at the summary of the modification and global protein it was clear the conditions follow different trends. Specifically, 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.

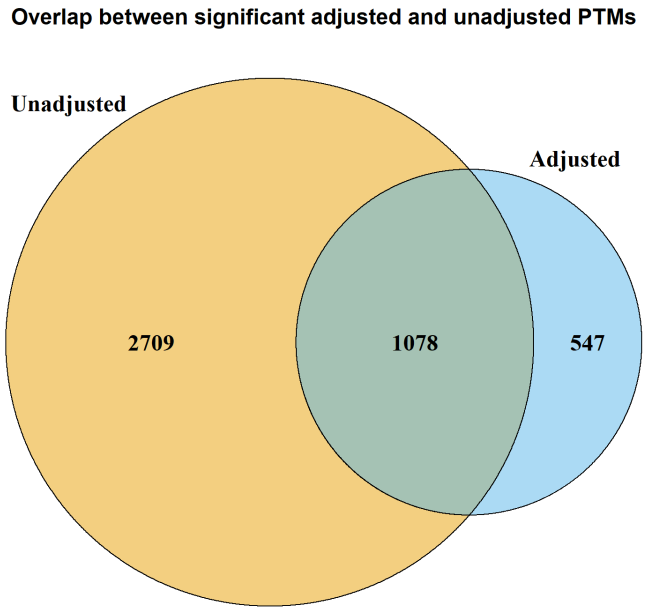


(a)

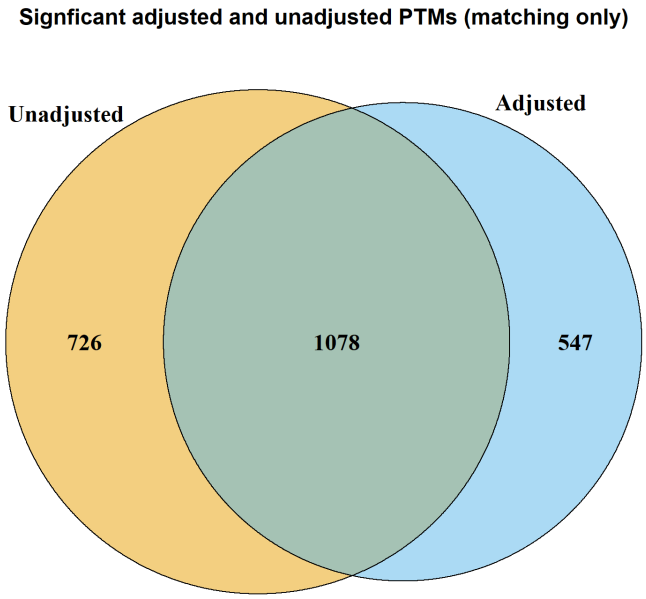


(b)

Figure 6: Dataset 5 : Mouse - Phosphorylation - 2mix-TMT time series. a) The overlap of differentially modified peptides between the PTM model with and without global protein level adjustment. Again more PTMs became insignificant after adjustment then became significant. b) Comparing the global profiling of protein TTP with the modification of the protein at site $S178$. When looking at the summary of the modification and global protein it was clear the difference between conditions followed the same trend. Specifically, there was a positive adjustment in abundance when comparing WT_Uninfect to WT_Late in both the modification and global profiling run. This indicated the movement was driven by changes in global protein that was only accounted for in the model after adjusting for global protein abundance change.



(a)



(b)

Figure 7: 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. More PTMs became insignificant than became significant after adjustment. This was due to not having a global profiling run, resulting in a lack of overlap between modified peptides and unmodified proteins. 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. This highlighted the need for a global profiling run.