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]. 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 [8, 9]. 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 [10]. 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 [11]. 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[12], Analysis of Variance[13], or Limma[14], 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 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 peptides, 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 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 [15, 16], and is available on Bioconductor. [OV: could you clean up the references a little bit? E.g., remove DOI, ISBN]

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 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. 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`) [17].

Additional details on all experiments are available in Supplementary Section 2.

Dataset 1: Computer simulation 1 - Label-free

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 convolution with changes in the unmodified protein and after applying adjustment to correct for convolution.

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

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, convolution with changes in the unmodified protein.

Dataset 3: SpikeIn 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 prior to enrichment or global protein profiling between Mix 1 and 3, and between Mix 2 and 4. 50 heavy-labeled KGG motif peptides from 20 human proteins were used as spike-in peptides in a mixed background of lysates from *E. coli* and a human cell line. Quantitative changes in protein and site abundance of these 20 human proteins were the target of the benchmark. In particular, we distinguished the changes in the abundances of the modified peptides (i.e., unadjusted changes) and the changes of their abundance relative to the changes in the abundances of the human lysate (i.e., protein-level adjusted changed). 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 that were quantified in the global profiling run.

Evaluation: We expect the spike-in peptides to change corresponding to the values in Figure 1(b). These peptides are treated as positive controls in all comparisons except Mix 1 vs Mix 4. Additionally, the background *E. coli* lysate peptides are treated as negative controls and are not expected to change in any comparison. Therefor 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.

[Here is the place to describe: 'we expect XXX to not change'. Therefore we evaluated the ability of statistical methods to avoid false positives (i.e., specificity). We expect XXX to have fold changes XXX between conditions XXX. We evaluated the accuracy of estimation of fold changes... and the sensitivity of detecting differentially abundant...] [I think some mixtures are not comparable because the total protein amount is not the same. I believe only in Mix 1 vs 2 and Mix 3 vs 4 are comparable. I do not think it is correct to compare Mix 1 vs 3 and 2 vs 4 directly because they differ in the total amount of the proteome. Or was it normalized based on what we know is the same? If yes, we need to describe in the beginning of 'Evaluation'] [The unadjusted proteins... estimation... The adjusted proteins... estimation... Positive controls... Negative controls...]

Dataset 4: Human - Ubiquitination - 1mix-TMT

Experimental Design: Luchetti et al. [18] 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. [19] 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 phospho-peptide 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. [20] 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 corresponds to an LC-MS/MS run. 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 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 peptides in the same run is denoted by y_{ijl}. The features can be quantified as part of a same mass spectrometry run, or in separate enrichment and global proteome profiling runs. Figure 2 schematically illustrates this data structure for one PTM site and the corresponding unmodified features of the protein, in a special case of $I = 2$ conditions and $J = 2$ biological replicates per condition. The number of modified and unmodified features varies across proteins. Some log-intensities may be outliers, and some spectral features can be missing.

The population parameter of interest is the difference between the abundances of a PTM site in condition 1 and condition 2, denoted by μ_1^* and μ_2^* respectively. We are interested in testing the null hypothesis

$$
\begin{aligned}
H_0 &: \Delta = \mu_1^* - \mu_2^* = 0 \\
H_a &: \Delta = \mu_1^* - \mu_2^* \neq 0
\end{aligned}
\tag{1}
$$

To test the hypothesis, a statistical model must represent the values in the table in Figure 2, and provide a model-based estimate of $\mu_1^* - \mu_2^*$, and its standard error.

Unfortunately, the population parameter is inherently convoluted with the overall changes in protein abundance. To account for this, it is advantageous to change the null hypothesis to

$$
\begin{aligned}
H_0 &: \Delta = (\mu_1^* - \mu_1) - (\mu_2^* - \mu_2) = 0 \\
H_a &: \Delta = (\mu_1^* - \mu_1) - (\mu_2^* - \mu_2) \neq 0
\end{aligned}
\tag{2}
$$

where μ_1 and μ_2 reflect the overall protein abundances in condition 1 and condition 2. These quantities can be estimated using protein features without the modification site.

Existing statistical methods for detecting differentially abundant PTMs

Two-sample t-test [DK: I mainly explained the ratio version. I briefly mentioned the feature version, which is used to model ptms w/o adjustment. We could reference the versions back to hypotheses in Equation 1 and 2 but I didnt yet.]

Many investigations perform differential abundance analysis of PTMs using two-sample t-test or its extensions [21]. These methods can be performed without accounting for convolution with global protein abundance by taking as input intensities of individual features from modified peptides, denoted by y_{ijk}^*. Additionally, they can be extended to remove convolution by taking as input intensity ratios of modified and unmodified peptide features and comparing the mean abundance of a PTM site from one condition to another [22][23]. This is done by taking the sum of peak intensities in each run for modified and unmodified features separately and dividing the estimate of PTM abundance by the protein abundance estimate. This is equivalent to taking the difference of their log abundances. For condition i and biological replicate j, the estimated quantity is denoted by $\hat{u}_{ij}^*$ and is given by Equation 3.

$$
\hat{u}_{ij}^* = \log\left(\sum_{k=1}^{K} 2^{y_{ijk}^*}\right) - \log\left(\sum_{l=1}^{L} 2^{y_{ijl}}\right)
\tag{3}
$$

These estimated ratios are then used as input to the two sample t-test.

$$
\hat{u}_{ij}^* = \hat{\mu}_i^* + \epsilon_{ij}^*, \text{where } \epsilon_{ij}^* \overset{iid}{\sim} N(0, \sigma^2), i \in \{1, 2\}
\tag{4}
$$

t-test assumes that the data follows a normal distribution, sampled from a representative portion of the total population, and that the conditions of interest have equal variance. The method is only applicable comparisons between two conditions of interest. The output of this model is the estimate of our quantity of interest $\hat{\Delta}$ and the standard error, $\widehat{SE(\hat{\Delta})}$. The statistical significance is determined by comparing the test statistic against the t distribution, with degrees of freedom $df = 2J - 2$ in balanced designs.

While simple, these approaches do not fully account for all sources of variations, and are not directly applicable to experiments with complex designs. These include experiments with comparisons of multiple conditions, acquisition in multiple batches, and unbalanced designs. Additionally, because the adjustment is done on the replicate level, there must be a matching number of replicates in both the modified and unmodified experimental designs. If the designs are different, this method is no longer directly applicable. Finally, while these approaches can be applied to experiments studying PTMs, there is not a self contained, straightforward implementation of the methods, making application challenging.

Limma Modifications of t-test such as moderated t-test with Limma were also proposed [14][24][25]. The inputs to Limma are the same as the inputs to t-test. The intensities of individual features from modified peptides, y_{ijk}^*, can be used directly (in the presence of confounding with the unmodified protein). To remove

confounding we transform the input by taking the intensity ratios of modified and unmodified peptide features, $\hat{u}_{ij}^*$, calculated in the same way as Equation 3. Limma then fits a linear model using the transformed inputs.

$$\hat{u}_{ij}^* = \beta_0 + c_i\beta_1 + + r_j\beta_2 + \epsilon_{ij}^*, \text{where } \epsilon_{ij}^* \overset{iid}{\sim} N(0, \sigma^2/w_j), i \in \{1,2\}, j \in \{1,2\} \tag{5}$$

Using a linear model allows Limma to incorporate the replicate, r_j, as an additional source of variation. Additionally, the model allows for the sharing of variance information across comparisons of interest. The model is used to estimate our condition of interest, $\hat{\Delta}$. To estimate $\widehat{SE(\hat{\Delta})}$, Limma leverages empirical bayes moderation [26] to adjust the model variance of the individual PTM using the shared variance of all PTMs modeled in the experiment. [DK: I introduced a new subscript here for the PTM p. It feel awkward to add it across the other methods so I only included it in the moderation describtion.] For a single PTM p with an estimated fold change $\hat{\Delta}_p$, the estimated variance is calculated as $var(\hat{u}_{pij}^*) = \sigma_p^2/w_{pj}$ where σ_p^2 is the variance resulting from the linear model of PTM p and w_{pj} is the global variance across all PTMs in replicate j. The statistical significance is then determined using $\hat{\Delta}_p$ and $\widehat{SE(\hat{\Delta}_p)}$ calculated using the adjusted variance $var(\hat{u}_{pij}^*)$.

[DK: @Olga can you expand on the good things/limitations? I think I am missing some points here but not sure what exactly]

Limma provides an alternative that overcomes limitations of t-test. First, utilizing empirical bayes moderation can increase the sensitivity of the analysis and is especially useful in cases of low replicates. Additionally, sharing variance information across comparisons means we can more accurately measure the variance as the number of comparisons increases.

While Limma offers improvements over t-test, the method still has a number of limitations. First, the method is limited in the types of experimental designs it can be applied to. The method is only directly applicable to experiments with at most two variance components, meaning that in more complex experiments it cannot account for all sources of variation. Additionally, it still requires the same experimental design for the modified and unmodified features so that the ratio calculation can be performed. Finally, there is no self contained implementation of the methods to PTMs, requiring manual transformation and application by the user.

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)[27]. 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

[DK: This section could use some rewording]

$MSstats$ [15] is designed for protein significance analysis and statistical relative quantification of proteins and peptides in global, targeted and data-independent proteomics. The package takes as input a list of intensities of individual features, denoted by y_{ijk}, used to characterize the identified protein. For each protein, the feature log-intensities are first summarized into a single value per protein per run in the subplot model. This is done using Tukey's median polish (TMP) [28], a simple and robust procedure, which iteratively fits a two-way additive model with the effects of run and feature, summarizing them into a single value per run. These summarized values are then used as input to fit a linear model based on the family of linear mixed-effects models [29] [30]. For example, the model for a TMT experiment would be as follows

[need to fix the centering]

$$u_{ijm} = \beta_0 + c_i\beta_1 + + r_j\beta_2 + mix_m\beta_3 + \epsilon_{ij}^*, \text{where } \epsilon_{ijm}^* \overset{iid}{\sim} N(0, \sigma^2),$$

$$mix_m \overset{iid}{\sim} N(0, \sigma_m^2), r_j \overset{iid}{\sim} N(0, \sigma_j^2), \sum_{i=0}^{I} c_i = 0 \tag{6}$$

We then use this model to estimate the quantity of interest $\hat{\Delta}$ and the standard error, $\widehat{SE(\hat{\Delta})}$. These methods are implemented in the R packages $MSstats$ and $MSstatsTMT$ [16].

$MSstats$ has a number advantages over simpler methods. First, the approach extends to many types of complex experimental design, including those with multiple sources of variation, such as the TMT model above, and unbalanced designs. Run-level summarization has been shown to be more robust compared to directly using individual features as input to the statistical model [15]. Additionally, missing value imputation can be performed before run-level summarizing using an accelerated failure time (AFT) model [31]. Finally, the methods include optional implementation of empirical bayes moderation, described above, when applied to TMT experiments [16].

While $MSstats$ is a powerful tool for the quantification of the global protein, it does not have built in tools to analyze PTMs. It is designed to estimate the difference in a protein between conditions, answering the Null Hypothesis

$$\begin{aligned} H_0 &: \Delta = \mu_1 - \mu_2 = 0 \\ H_a &: \Delta = \mu_1 - \mu_2 \neq 0 \end{aligned} \tag{7}$$

It is not designed to remove confounding with the global protein meaning that it cannot directly address the Null Hypothesis in Equation 2.

Results

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

We propose a statistical method for detecting changes in PTMs that allows us to remove confounding with changes in the global protein and is applicable to experiments with complex designs. To do this, we leverage the existing methods in $MSstats$ and extend them to remove confounding between changes in the PTM and global protein. The method takes as input the spectral features of the PTM, y_{ijk}^*, and it's corresponding global protein, y_{ijk}. The features are summarized separately using the methods described previously for $MSstats$, resulting in run level values for the PTM, $\hat{y}_{1++}^*$, and unmodified peptide, $\hat{y}_{1++}$. These summarized values are then used as the input to our models. Two separate linear mixed-effects models are fit, one using the summarized PTM values and another using the summarized global protein values. The models give estimates of the difference in log-abundance of the underlying PTM, $\hat{\Delta}_{PTM}$, and global protein, $\hat{\Delta}_{protein}$. Finally, the adjusted difference, $\hat{\Delta}$, is given by the difference in log-abundance of the PTM site, subtracted by the difference in log-abundance of the underlying protein.

$$\hat{\Delta} = \hat{\Delta}_{PTM} - \hat{\Delta}_{protein} \tag{8}$$

In the simple case of a balanced design with one source of variation, this is equivalent to the log of the ratio of PTM abundance difference to protein abundance difference.

$$\hat{\Delta} = \left[\frac{1}{J} \left(\hat{y}_{1++}^* - \hat{y}_{2++}^* \right) \right] - \left[\frac{1}{J} \left(\hat{y}_{1++} - \hat{y}_{2++} \right) \right] \tag{9}$$

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. The equation accounts for the increased uncertainty when incorporating more information from the unmodified protein. This may lead to an increased SE, when compared to modeling the PTM alone.

$$\widehat{SE(\Delta)} = \left[\left(\hat{\sigma}_{\gamma^*}^2 + \hat{\sigma}_{\gamma}^2 \right) \right]^{1/2} \tag{10}$$

Finally the degrees of freedom are calculated as follows

$$\left(\hat{\sigma}_{\gamma^*}^2 + \hat{\sigma}_{\gamma}^2 \right)^2 \Bigg/ \left(\frac{\hat{\sigma}_{\gamma^*}^4}{\mathrm{df}(\gamma^*)} + \frac{\hat{\sigma}_{\gamma}^4}{\mathrm{df}(\gamma)} \right) \tag{11}$$

The results of this model are used to answer the Null Hypothesis stated in Equation 2. 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 [32]. 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 [33].

[DK: I'm not sure where this part should be. could move it to the first paragraph of this section, or in the extensions to complex design] This approach allows us to model experiments with separate designs for the PTM and global profiling runs. This includes differing number of runs, biological replicates, and/or types of acquisition methods. This is achieved through fully modeling the PTM and global protein separately before adjusting. Because we are not combining on the replicate level, the number of replicates for each group do not need to match for adjustment. This makes the proposed method more flexible and applicable to a wider variety of experimental designs.

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) $\sigma_{\gamma^*}^2$ and σ_γ^2, 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 [34]. Given the above quantities, the minimal number of replicates J is determined by the variance of the estimated log-fold change $\text{SE}^2(\hat{\Delta})$ as

$$\text{SE}^2(\hat{\Delta}) = \left[\frac{2}{J} \left(\hat{\sigma}_{\gamma^*}^2 + \hat{\sigma}_\gamma^2 \right) \right] \leq \left(\frac{\Delta}{t_{1-\beta,df} + t_{1-\alpha/2,df}} \right)^2 \tag{12}$$

where

$$\alpha = (1 - \beta) \cdot \frac{q}{1 + (1 - q) \cdot m_0/m_1} \tag{13}$$

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

$$J \geq \frac{(2\hat{\sigma}_{\gamma^*}^2 + 2\hat{\sigma}_\gamma^2)(t_{1-\beta,df} + t_{1-\alpha/2,df})^2}{\Delta^2} \tag{14}$$

More details on sample size calculation can be found in [35].

Detection of changes in PTMs - Extension to complex designs

As discussed above, the proposed approach can be extended to complex designs, including experiments with multiple sources of variation and unbalanced designs. This extension is achieved through utilizing the methods in *MSstats* which, in complex cases, use restricted maximum likelihood (REML) [36] to estimate the parameters of the models. Once the parameters are estimated, we can adjust the PTM model for changes in the global protein using modified versions of Equations 9, 10, and 11.

The log-fold change in the adjusted PTM abundance, $\hat{\Delta}$, is now estimated by REML, and Equation 9 becomes

$$\hat{\Delta} = (\hat{\mu}_{RML_i}^* - \hat{\mu}_{RML_{i'}}^*) - (\hat{\mu}_{RML_i} - \hat{\mu}_{RML_{i'}}) \tag{15}$$

The standard error of the estimate $\text{SE}(\hat{\Delta})$ 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 REML.

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 $MSstatsPTM$. An overview of the steps of the package are illustrated in Figure 3. $MSstatsPTM$ 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 peptide and global profiling runs. If the global profiling information is not available, the package can still analyze the modified run, but will do so without adjusting for changes in unmodified protein abundance. This will reduce our null hypotheses to Equation 1; which amounts to running the methods seen in $MSstats$ on modified peptide level data [15] [16].

After using the converters, the next step is peptide/protein summarization and missing value imputation. For the modified run, the package summarizes features 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 can optionally impute missing values using an AFT model [31]. Missing value imputation should be performed with care when applied to PTM data. PTMs generally exhibit low feature counts and may be missing due to reasons other than low abundance. These issues can cause the missing value imputation to be unstable.

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 above.

Beyond the core functionality, the package also includes functions for plotting the results. These include summarization plots to help with quality assurance and identifying sources of variation (such as quality control plots and profile plots), as well as modeling plots (such as volcano plots and heatmaps), which evaluate the fold change between conditions and peptides.

The package relies on functionalities from the R packages $MSstats$ [15] and $MSstatsTMT$ [16]. The statistical modeling relies on the functionality from the R packages $lme4$ [37] and $lmerTest$ [38]. 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. When 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 PTMs that were expected to be differential after accounting for changes in global protein abundance. True negatives were PTMs that were not expected to be 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. These results were then compared to the ground truth to determine TP and TN. Summary statistics were calculated, including 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.

Further details reviewing the results of each experimental analysis can be seen in Supplementary Sec. 2.

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 4a we see that not accounting for changes in the unmodified protein resulted in a high FDR. 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 information across comparisons, resulting in improved power in simulations with more comparisons.

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

In simulations with missing values and low features 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 4b. 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 (IQR) of the estimated fold change for the proposed method was on average 21.8% smaller than Limma's IQR and 10.4% smaller than t-test's IQR. This fold change comparison can be seen in Supplementary Figure S6.

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 in all methods, however the 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 36.78% smaller than Limma's and 32.98% smaller than t-test's. While Limma and t-test generally estimated the correct fold change, the proposed method's estimation was more consistent across all modifications.

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 decreased significantly. 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 insignificant adjusted p-values. Using Limma would have resulted in missing the majority of differential PTMs.

MSstatsPTM corrected for bias in group comparison TMT experiment

The results of this experiment are summarized in Figure 6. Figure 6a shows the number of significant modified peptides before and after adjustment. More peptides became insignificant after adjustment than became significant. 3,236 modified peptides became insignificant, 1,352 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 were corrected in adjustment, and the unconvoluted 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 6b, the modification of protein $GSDMD$ at site $K62$ 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 two mixture TMT experiment

The results of this experiment are summarized in Figure 7. Figure 7a shows the number of significant modified peptides before and after adjustment. 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 removing convolution with changes in the umodified protein or increased variance after adjustment was applied. 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 7b 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.

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 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 results of this experiment are summarized in Figure 8. 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 became insignificant after adjustment. 726 modified peptides became insignificant, 547 became significant, and 1,078 were significant in both models. Only 25 PTMs became insignificant due to increased variance from adjustment.

Discussion

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. 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 convoluted with each other if there are multiple modification sites per peptide, or convoluted 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 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 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.

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]   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.

[7]   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.

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

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

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

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

[24] 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.

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

[26] H. Robbins (1992). "An Empirical Bayes Approach to Statistics". In: *Breakthroughs in Statistics: Foundations and Basic Theory*. New York, NY: Springer New York, pp. 388–394.

[27] 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.

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

[29] 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.

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

[31] 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.

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

[33] 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.

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

[35] 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.

[36] M. Kenward and J. Roger (1997). "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood". In: *Biometrics* 53.3, pp. 983–997. (Visited on 06/16/2022).

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

[38] 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	2/1	8,848	1.21	11.01	MSV000088966	RMSV000000356
	Mouse - Phosphorylation - 2mix-TMT	6	4/3	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.
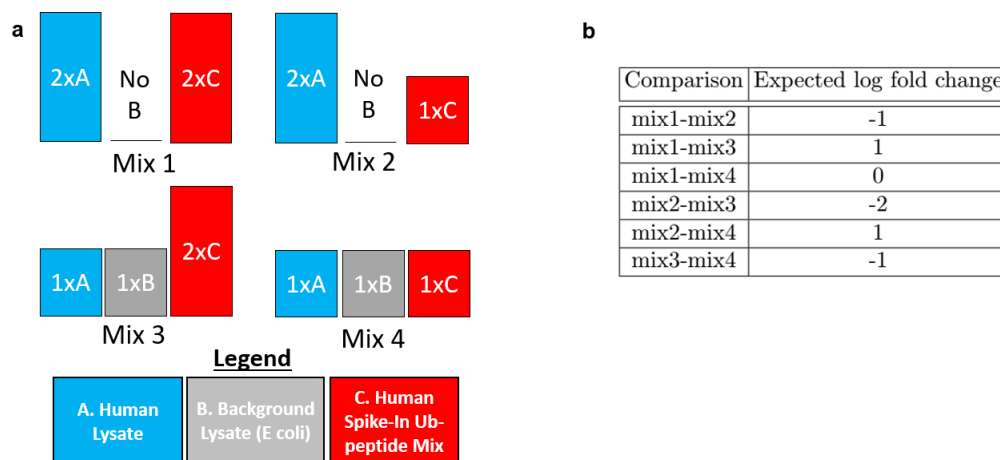
Figures

Figure 1: **Dataset 3: SpikeIn benchmark - Ubiquitination - Label-free**. (a) Relative amounts of human lysate, background *E. Coli* lysate, and spike-in peptides. [Clarify what is changing how] (b) The expected fold change of the spike-in peptides, after adjusting for changes in the unmodified peptide (human lysate). The expected fold change is calculated using the values of the spike-in and human lysate in part (a).[Could you make the columns in the table more complete? True fold-fold changes between the backgrounds; unadjusted true fold changes, adjusted true fold changes - essentially any value for any comparison that we may want to evaluate]
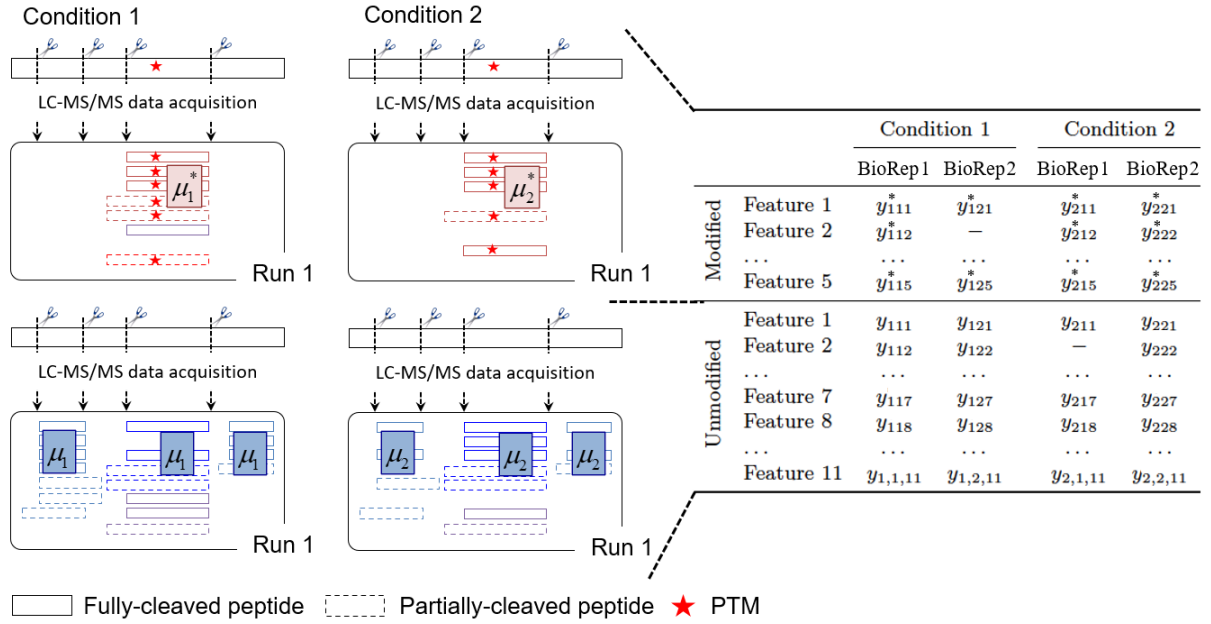
Figure 2: Goals of PTM characterization, input to statistical analyses, and notation. (a) Schematic data representation, in a simplified case of two conditions and two biological replicates. Each PTM site is modeled and characterized separately, where a PTM is quantified with observed spectral features (boxes), distinguished by different charge states of a peptide. The log-intensity of the PTM in Condition i, Run j, and feature k are denoted by y^*_{ijk}. Log-intensity of feature corresponding to unmodified peptide in Condition i, Run j, and feature k are denoted by y_{ijk}. Peptides can be fully cleaved (solid lines) and/or partially cleaved (dashed lines). (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.
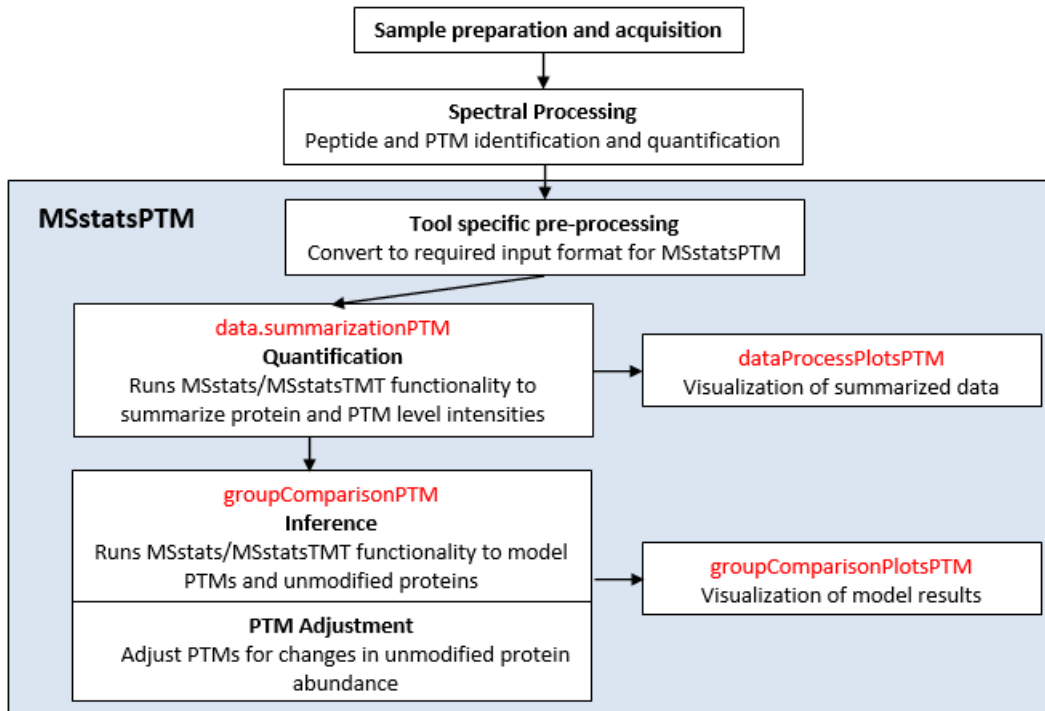
Figure 3: The workflow of *MSstatsPTM* and how it fits into the experimental pipeline. The name of the R functions used for each step are highlighted in red. The package's workflow starts after modified and unmodified peptide quantification. First tool specific pre-processing is performed, this includes 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. After both the summarization and group comparison steps, plots can be created to summarize the results.

(a)

(b)

Figure 4: 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.
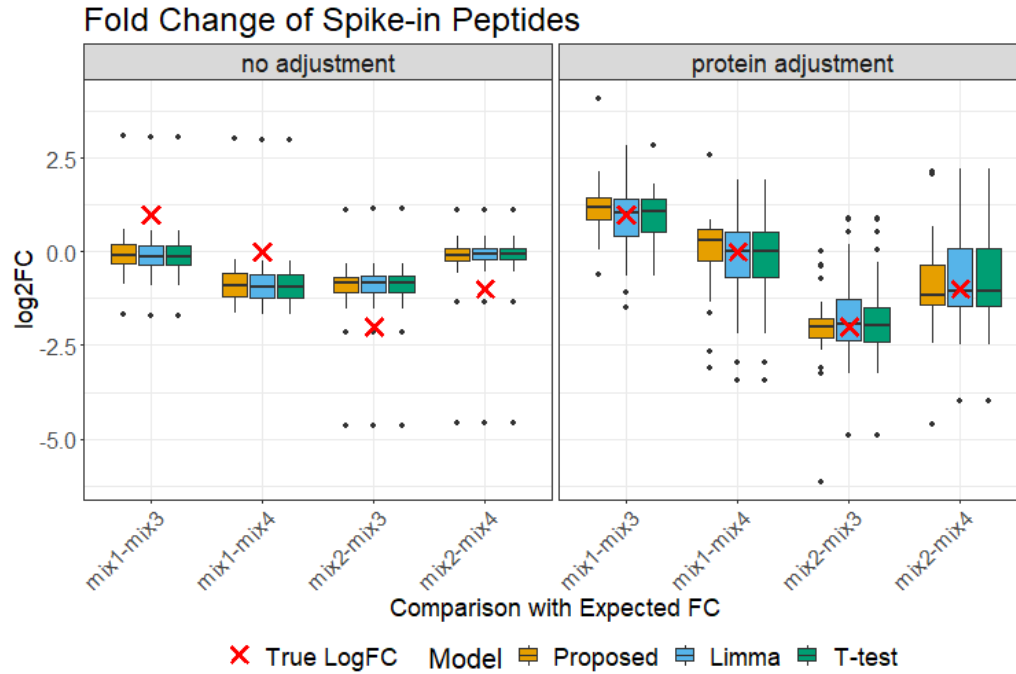
(a)
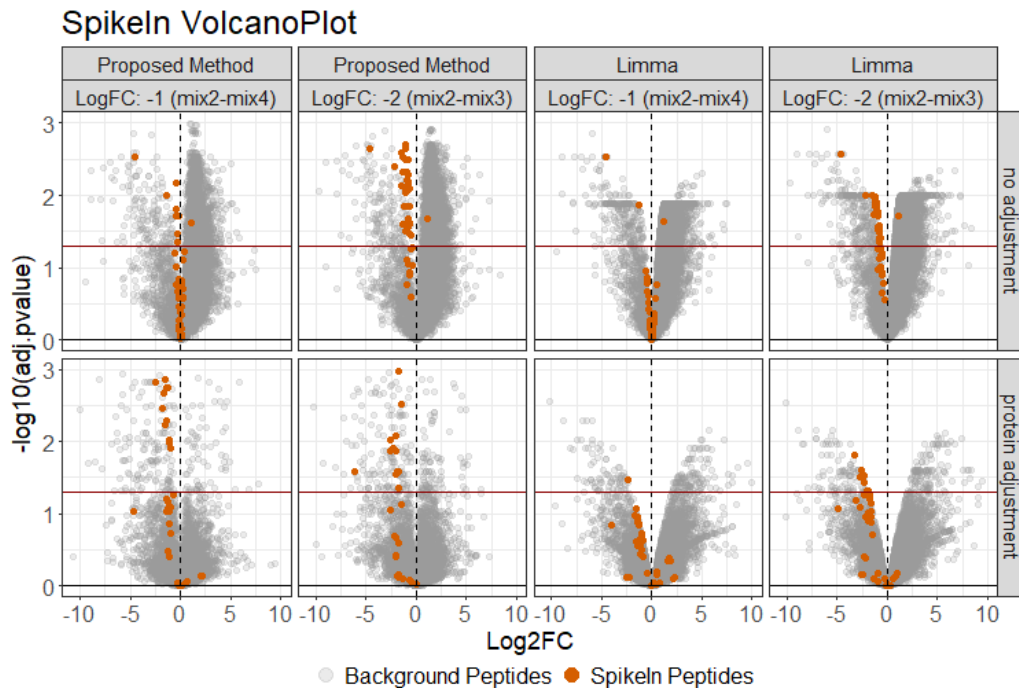
(b)

Figure 5: 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) 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.
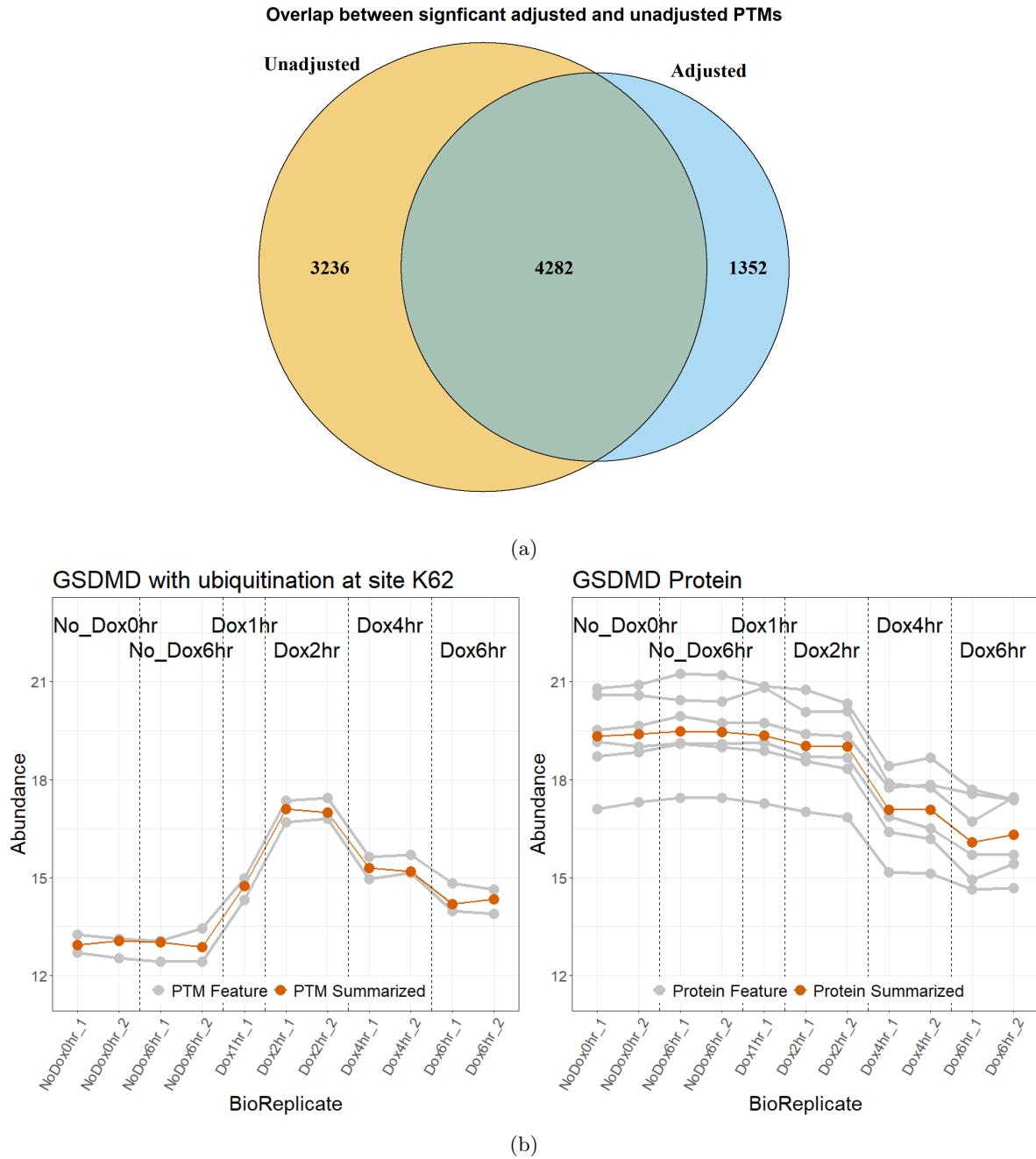
Figure 6: Dataset 4: Human - Ubiquitination - 1mix-TMT. a) The overlap of differential modified peptides for the PTM model with and without global protein level adjustment across all comparisons. More PTMs became insignificant after adjustment than became significant. For the peptides that became insignificant in the adjusted model, their change in abundance was driven by changes in the global protein. In contrast, peptides that became significant after adjustment had 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.
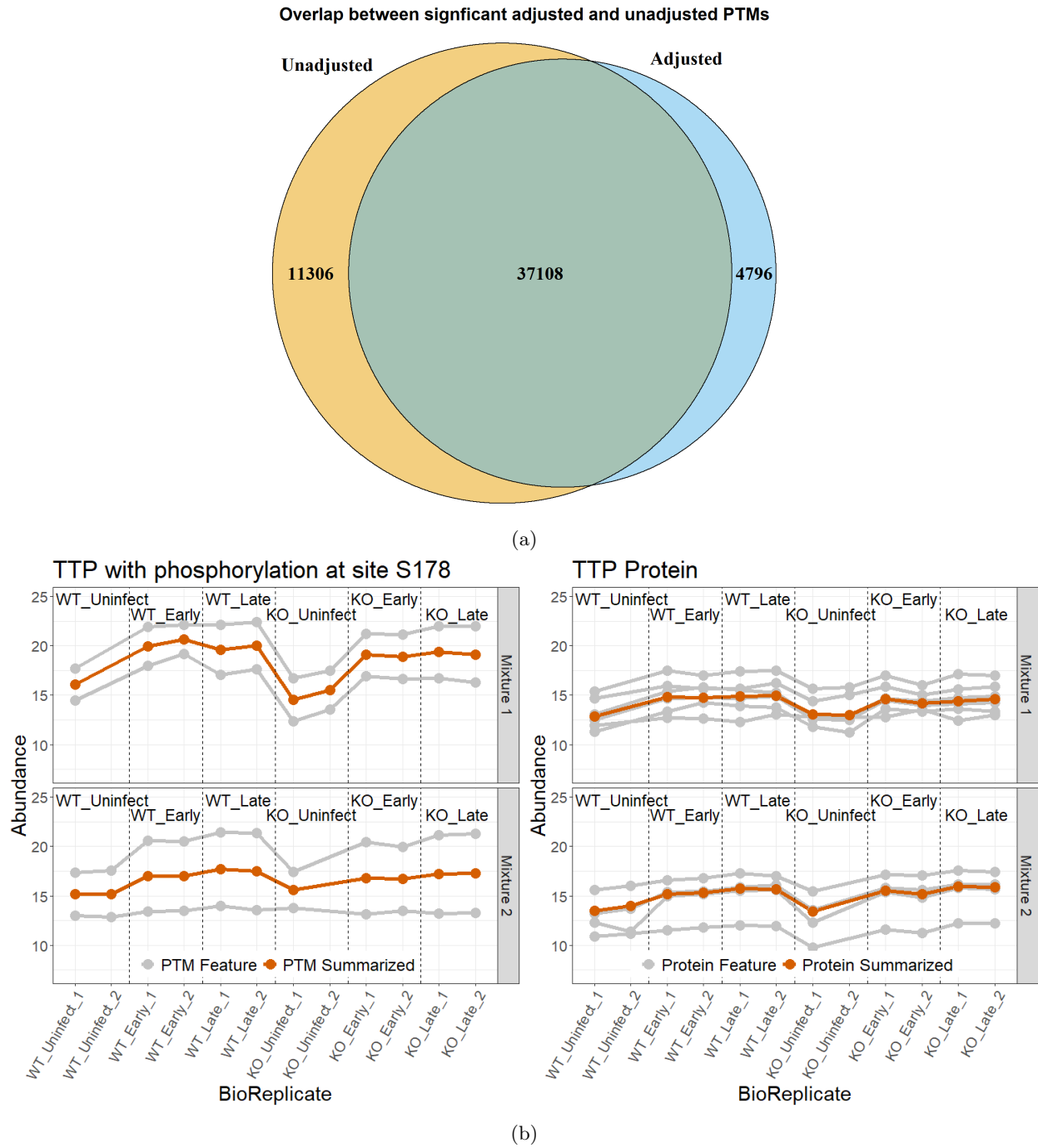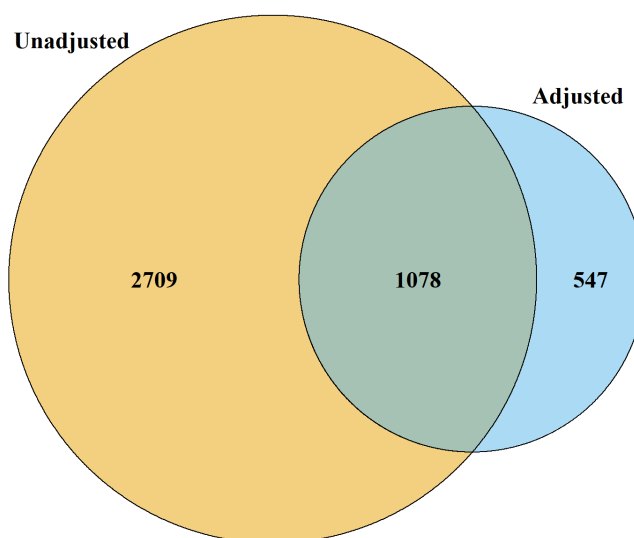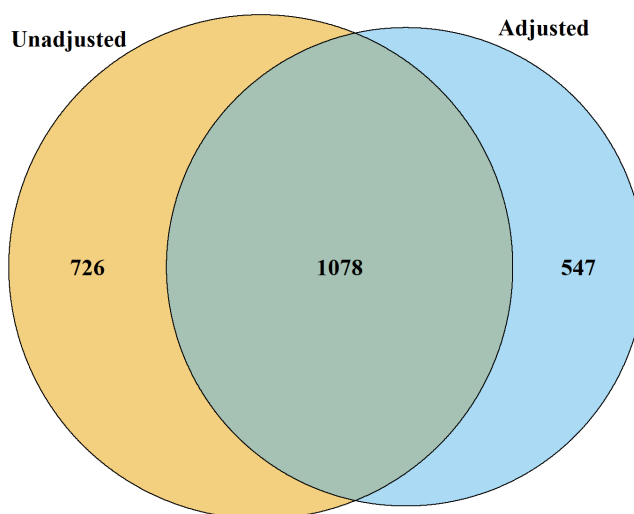
(a)

(b)

Figure 7: Dataset 5: Mouse - Phosphorylation - 2mix-TMT. a) The overlap of differentially modified peptides between the PTM model with and without global protein level adjustment across all comparisons. 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.

Overlap between significant adjusted and unadjusted PTMs

(a)

Signficant adjusted and unadjusted PTMs (matching only)

(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. 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.