MSstatsPTM statistical relative quantification of post-translational modifications in global proteomics 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

The scientific community increasingly utilizes mass spectrometry (MS)-based proteomics to detect changes in post-translational modifications (PTMs) between samples from various conditions. Experiments targeting PTMs face numerous challenges. These include the low abundance of modified proteoforms, few 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 to estimating relative changes in PTMs, and to detecting differential PTM abundance, should combine information pertaining to PTM sites over several peptides, replicates in multiple conditions, and consider relevant sources of confounding and variation. In this manuscript we propose such a general statistical framework, which is comprehensive, accurate, and leads to reproducible results. The framework requires an experimental design that, for each sample, quantifies the peptides with post-translational modifications, as well as peptides from the same proteins with no modification sites, and is applicable to both label-free and tandem mass tag (TMT)-based acquisitions. The proposed framework summarizes the abundances of peptides with and without the modification sites, and fits a linear mixed effects model that automatically adjusts to the biological and technological aspects of the experimental design. Finally, changes in the PTM and the protein-level abundances are combined to remove bias due to the confounding of these two sources. Evaluations on computer simulations, a spike-in experiment with known ground truth, and three biological experiments which demonstrate the improved fold change estimation as compared to the currently used approaches for relative PTM quantification. The proposed framework is implemented in the free and open-source R package *MSstatsPTM*.

Introduction

The signaling mechanisms that allow cells to mount a dynamic and fast response to a multitude of events are facilitated by the modification of proteins at specific residues, acting as molecular on/off switches[1, 2]. Bottom-up liquid chromatography coupled with mass spectrometry (LC-MS) is broadly established as the tool-of-choice for unbiased and large-scale identification and quantification of proteins and their post-translational modifications (PTMs) [3, 4]. Studies targeting the post-translationally modified proteome focus either on the accurate localization of modification sites on proteins, relative or absolute quantification of a modification site's occupancy repertoire, or relative changes in occupancy across experimental conditions[5]. Some scientific questions, such as discovery of protein biomarkers of a disease, consider meaningful patterns

1

of changes in PTM abundance alone [6]. Other questions, e.g. in biological and physiological studies based on very short timescales, require us to distinguish the confounding of quantitative changes in PTM and of overall changes of the protein abundance.

Regardless of the question at hand, interrogating the modified proteome is challenging due to 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, which imposes challenges in both quantification and statistical modeling. Second, contrary to the often large number of identified peptides that can be used as features to model protein abundance changes, there are relatively few representative peptides that span a modification sites, which often results in sparse, and sometimes, inherently convoluted models (i.e., single versus multiple modified sites on a single peptide) [5]. 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, complication the interpretation of the results [8]. Finally, technological aspects of bottom-up MS experiments, and presence of labeling, e.g. by tandem mass tag (TMT), introduce additional sources of uncertainty and variation.

Currently statistical analyses of these experiments are performed with methods such as t-test and Limma[9]. These approaches generally use the intensity ratios of modified and unmodified peptide features and compare the mean abundance of different PTM sites. These approaches do not fully account for different sources of variations and as such, are not directly applicable to experiments with complex designs, such as multiple conditions and mixtures. Additionally, there is no self contained implementation of these methods, requiring manual application of the methods by the user and increasing complexity.

This manuscript proposes a general statistical analysis framework that detects relative changes in post-translational modifications. The proposed framework combines the quantitative information pertaining to a PTM site over peptides and replicates in multiple conditions, and takes into account various sources of variation and confounding. The framework requires an experimental design that, for each sample, quantifies the peptides with post-translational modifications, as well as peptides from the same proteins with no modification sites. It is applicable to both label-free and tandem mass tag (TMT)-based acquisitions. The proposed framework summarizes the abundances of peptides with and without the modification sites, and fits a linear mixed effects model that automatically adjusts to the biological and technological aspects of the experimental design. [For questions focusing on PTM, can analyze changes in the modified peptides. For eliminating the confounding, combine...] Finally, changes in the PTM and the protein-level abundances are combined to remove bias due to the confounding of these two sources.

We evaluated the proposed framework on datasets from computer simulations, benchmark controlled mixtures, and biological investigations [expand on the relevance and the comprehensive nature of the datasets - diverse organisms, acquisitions]. The evaluations demonstrated that by appropriately leveraging the information from [which sources], the proposed approach improves the accuracy of the estimates of PTM fold changes and results in a better calibrated proportion of false positive [what?]. [In particular, accounting for the confounding from protein abundance...]

The proposed approach is implemented as a freely available open source R package $MSstatsPTM$, available on Bioconductor, which employs similar input format as in $MSstats$ and $MSstatsTMT$, and is compatible with many acquisition methods, such as label free, DIA, DDA, and TMT[10, 11]. [we claim that the results are reproducible - what does this mean?]

Experimental Procedures

The datasets were chosen to provide a performance benchmark of the proposed approach in situations with known ground truth, and to represent a variety of experimental designs and acquisition methods. The computer simulations varied in experimental realism. The first simulation was perfectly clean, with many replicates and no missing values, while the second simulation introduced real world qualities, such as limited modified features and missing values. The spike-in experiment took the real world qualities a step further

and allowed us to compare the methods in a real experiment with known [changes in what?]. Finally, three biological experiments demonstrated the applicability of the proposed approach across different biological organisms, experimental designs, and data acquisition strategies.

Dataset 1 : Computer simulation 1 - Label-free

The simulation mimicked a label-free experiment with a high number of modified and unmodified peptide features and no missing values. Multiple datasets were generated with different numbers of replicates and conditions. In all simulations 1000 modified peptides were generated, half of which were true positives and half true negatives. Half of the true positives and true negatives were generated convoluted with global protein abundance. Before adjustment, the convoluted true positives should appear to be negatives, while the convoluted true negatives would appear differential. Modified peptides and unmodified proteins were both simulated with 10 features.

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

In this simulation real world experimental conditions were introduced, including missing values and low number of modified features. The percentage of missing values and feature counts were determined by looking at the averages from the biological experiments in this paper. Modified peptides were simulated with 2 features, while unmodified proteins were still simulated with 10 features. Additionally, 20% of observations for both modified and unmodified peptides were missing completely at random. Adding missing values and few representative modified features provides a more realistic expectation of model performance. Otherwise this simulation used the same parameters as in simulation 1, with 1000 modified peptides half of which were positives and the other half negative.

    Further details of how both simulations were generated, including exact parameters used, can be found in Supplementary Sec. 4.1Experiments with known ground truthsubsection.4.1.

Dataset 3 : SpikeIn benchmark - Ubiquitination - Label-free

We evaluated our approach using a custom designed spike-in benchmark experiment with known ground truth, where 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 as the estimate of global protein abundance changes. Background E coli Lysate was used to normalize total protein levels prior to enrichment or global protein profiling. The background lysate was treated as controls and are not expected to be differential in any comparison. The spike-in peptides were mixed with human lysate to create four mixture conditions. Two sets of data were acquired for each mixture: KGG enriched + LC-MS, and LC-MS only. The four mixtures were compared with known fold changes for the spike-in peptides.

    More information on the experimental design of the benchmark experiment can be found in Supplementary Sec. 4.1.4Dataset 3 : SpikeIn benchmark - Ubiquitination - Label-freesubsection.4.1.4.

Dataset 4 : Human - Ubiquitination - 1mix-TMT

The proposed approach was evaluated on an experiment where Shigella ubiquitin ligase IpaH7.8 was shown to degrade the protein gasdermin D (GSDMD)[12]. Multiplex proteomics was used to quantify 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. The experiment was unbalanced with two bioreplicates per condition for all conditions except for infected 1 hour. About 95% of the identified modified peptides derived from proteins that were quantified in the global profiling run.

Further details on the design of this experiment are provided in Supplementary Sec. 4.2.1Dataset 4 : Human - Ubiquitination - 1mix-TMTsubsubsection.4.2.1.

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

Here an experiment targeting primary murine macrophages infected with Shigella flexneri (S.flexneri)[13]. Multiplex proteomics was used to quantify the abundance of total protein, and phosphorylation in wild type (WT) and 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). Quantifying the total protein along with the post-translational modifications allowed us to adjust for changes in total protein and see the true impact of the site specific modifications. Two mixtures using 11-plex were ran over the six conditions. The six conditions were split between 11 channels leading to the experimental design being 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. About 90% of the identified modified peptides derived from proteins that were also quantified in the global profiling run.

Further details on the design of this experiment are provided in Supplementary Sec. 4.2.2Dataset 5 : Mouse - Phosphorylation - 2mix-TMT time seriessubsubsection.4.2.2.

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

This experiment looked into the relationship between USP30 and protein kinase PINK1, and their association with Parkinson's Disease. Ubiquitination site profiling was performed and the modified site abundance was analyzed. Four conditions were tested with two biological replicates per condition. The conditions were as follows: CCCP, USP30 over expression (USP30 OE), Combo, and Control. Label-free mass spectrometry quantification was used to quantify the abundance of modified peptides. In contrast to the other biological experiments assessed in this paper, this experiment did not not have a separate global profiling run for the unmodified protein. In this case $MSstatsPTM$ can still be used by extracting unmodified peptides from the modified run, however this generally leads to substantially less matches between modified and unmodified peptides, in addition to low feature counts for unmodified peptides. In this case there only 41.9% of the modified peptides had a corresponding unmodified protein to perform adjustment. In cases where there is no estimate for the matching unmodified protein, adjustment cannot be performed.

Results

Existing statistical methods for experiments targeting PTMs

Despite the important implications of PTMs in biological functions, there is a lack of general framework to summarize the available quantitative information from LC-MS data, to perform statistical inference, and to draw conclusions to characterize the quantitative properties of PTM in a statistically rigorous manner. Many investigations performed differential abundance analysis of PTMs using two-sample t-test or its extensions. 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. Modifications of the t-test such as moderated t-test with Limma were also proposed[14]. While simple, the approach does not fully account for the sources of variations, and it is not directly applicable to experiments with complex designs, e.g., comparisons of multiple conditions, acquisition in multiple batches, etc. Additionally, while this approach can be applied to experiments targeting PTMs, there is not a self contained, straight forward implementation of the methods, making application challenging.

Isobar-PTM was developed for experiments with MS/MS quantitative strategies that employ isobaric labels such as tandem mass tags (TMT) and isobaric tag for relative and absolute quantification (iTRAQ)[15].

4

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. The modeling framework, however, is not applicable for either label-free workflows or experiments with complex designs.

Further details on the specific equations and workflow of existing methods can be seen in Supplementary Sec. 2Existing methodssection.2.

Proposed approach

Figure 1 schematically illustrates a simplified version of the data structure resulting from a typical bottom-up experiment for quantitative analysis of PTMs, in which there are multiple layers of variation present. A PTM site is quantified with multiple spectral features, which vary in sequence (e.g., fully or partially cleaved peptides), ionization efficiency, charge states, etc. The number quantified features vary across replicate LC-MS/MS runs of the same sample, and across conditions. To perform adjustment with respect to protein abundance, features of unmodified peptides are used for the inference of underlying protein abundance. Typically, because of the enrichment step for PTMs, very few of those features are present in original LC-MS runs. For more accurate estimation of protein abundance, separate global proteomics data of unenriched samples are recommended. If unmodified peptides are unavailable for any given modification, unmodified intensity adjustment cannot be performed. As different levels of variability are present in the data, the log-intensities of the features for modified and unmodified peptides are modeled separately using two linear mixed models.

Existing work in statistical modeling and parameter estimation

The proposed approach leverages the large amount of work done to create $MSstats$ [10]. It takes as input a list of log-transformed intensities of spectral features, identified and quantified across LC-MS runs. The features, which are precursor ions of modified or unmodified peptides, are used to characterize the identified PTM sites and proteins. For each PTM site, the feature log-intensities of the modified peptides spanning the site 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 the split-plot approach as in $MSstats$, where the feature log-intensities are first summarized into a single value per site per run in the subplot model, and the site-level summaries are then used for the inference of the PTM site abundance[10]. 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 site-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 site[16]. After summarization, the inference of the PTM site abundance in each condition is carried through fitting a model based on the family of linear mixed-effects models[17] [18]. Statistical modeling and quantification for global proteomics data are performed by the same procedure as for PTM data.

Additionally, the proposed method can include 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[16]. 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.

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

Here we propose a statistical method for detecting changes in PTMs and removing confounding with changes in global protein abundance. The method is shown when applied to a balanced design with multiple conditions and one technical replicate. In practice the experimental design can be unbalanced and variation

can come from multiple sources. We show how the methods can be applied to more complex experimental designs in the next section.

Detection of differentially modified PTM sites is performed through testing the null hypothesis of 'no change' against the alternative. The null hypothesis states that there is no difference in log-abundance of the PTM site between conditions, adjusted with respect to protein abundance. There are I conditions and J mass spectrometry runs (technical replicates) per condition in the experiment. 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 log-intensity of Feature l from the unmodified peptides in the same run is denoted by y_{ijl}. The expected values of log-abundances of the PTM and protein in Condition i are denoted by μ_i^* and μ_i. The feature and run level summarization 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 using the following formula.

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

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[19]. 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[20].

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 [21]. 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

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

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

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

More details on sample size calculation can be found in [22].

Detection of changes in PTMs - Extension to complex designs

The statistical modeling approaches discussed above can also be extended to complex designs including experiments with additional sources of variation and unbalanced designs. Equation (1) is the same in this case and the null hypothesis is still that there is no difference in mean PTM abundance between Conditions i and i' after adjusting for changes in mean unmodified peptide abundance between Conditions i and i'.

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 the formulas in Section 3.

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 $\text{SE}(\hat{\Delta})$ and degrees of freedom in Equations (3) and (4) are unchanged from the balanced case, although $\hat{\sigma}_{\gamma^z*}$ 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. 3.1Design of PTM experimentssubsection.3.1

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 pvalues 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], depending on the data acquisition type. The statistical modeling relies on the functionality from the R packages $lme4$ [23] and $lmerTest$ [24]. 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

[Check that this section uses past tense; look out for informal expressions]

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 is known, we calculated the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). True positives are modified peptides that are differential after accounting for changes in global protein abundance. True negatives were modified peptides that were not differentially abundant after accounting for 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)), the 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 presents 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 corrects for high FDR in computer simulations

The proposed approach was evaluated against two-sample t-test and Limma in computer simulations. In Figure 3 we see that not accounting for changes in unmodified peptide level results in a high false discovery rate. When adjusting for changes in the unmodified peptide, all methods performed similarly in terms of FDR. In a clean simulation, recall and accuracy of the proposed approach and Limma performed similarly, with Limma performing slightly better with low 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 outperforms existing methods in simulations with missing values and low features

In this simulation the proposed method outperformed t-test and limma when real world experimental conditions were introduced. Changes in unmodified peptide level still needed to be accounted for to control the FDR. Once controlling for changes in the unmodified peptide, the proposed method out performed the other methods, as seen in Figure 3. 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 peptides, the proposed method showed a tighter distribution of estimated fold changes around the true fold change than 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 S7Fold change distribution comparison between Simulations 1 and 2.figure.caption.8.

Further details reviewing the results of each simulation and their comparison with respect to method performance can be seen in Supplementary Sec. 4.1Experiments with known ground truthsubsection.4.1.

In a label-free benchmark experiment the proposed approach corrects for bias and outperforms 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 in line with expectation. This can be seen in Figure 4a. While the median fold change for the spike-in peptides was generally in line with the true fold change after adjustment, we can see that the distribution of fold changes for the peptides was visibly wider after adjustment. However, the proposed method a tighter 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 level. After adjustment the estimated fold change is more in line with expectation. Additionally, the background peptides serving as the null model, show many false positives before adjustment is made. After adjustment the results improve and the number of false positives decrease significantly. Specifically, for the proposed method, the number of false positives goes from 20.88% to 1.84% of all true negatives after adjustment is applied. Using both the proposed method and Limma, the red labeled spike-in peptides are more inline with expected fold change after the adjustment is applied. However, when using Limma we can see that the majority of the differential spike-in peptides are not showing a significant adjusted pvalue. 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. 4.1.4Dataset 3 : SpikeIn benchmark - Ubiquitination - Label-freesubsection.4.1.4.

MSstatsPTM corrects 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 become insignificant after adjustment than become significant. 3,236 peptides become insignificant, 1352 peptides become significant, while 4,282 peptides are 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 truly due to changes in global protein abundance.

In Figure 5b, looking at the modification of protein $GSDMD$ at $K62$ we can see the advantage of the proposed method. The modified peptide shows a flat abundance change between the infected 1 hour, 4 hour, and 6 hour conditions. This is 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 pvalue was insignificant at .0644. After adjusting for changes in the global protein abundance, the fold change is much higher, 2.79, and the adjusted pvalue became significant, $5.25e - 8$. In this case the effect of the modified peptide is strongly confounded with changes in the global protein. The proposed method allows us to remove this confounding and estimate the true effect.

Proposed approach removes 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 become insignificant, 4,947 became significant, while 41,552 peptides are significant in both models. Again we tested if the decrease in significant peptides was due to the increased variance that comes from adjustment, or if it was mainly due to removing convolution with change 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 on this experiment, 548 peptides became insignificant due to an increase in variance. This is still a small portion (3.4%) of the 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 $S178$ shows the power of the proposed method. The profile plot shows a large positive log fold change of 2.9 between the WT_Late and WT_Uninfected conditions. However, the global profiling run shows a similar log fold change of 2.014 between the same conditions. This indicates 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 pvalue becomes 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. 4.2.2Dataset 5 : Mouse - Phosphorylation - 2mix-TMT time seriessubsubsection.4.2.2.

In label-free experiment without a separate global profiling run, the proposed approach could eliminate the bias [due to what?], albeit less effectively than [in presence of a global profiling run?]

As discussed previously, 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 significant 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, not having 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. In 7a the number of significant modified peptides before and after protein adjustment is shown. In the top plot we can see there are a large number of peptides that become insignificant after adjustment. This is mainly due to the lack of overlap between modified and unmodified peptides, stemming from not having a global profiling run. In the bottom plot only modified peptides that could be adjusted are shown. Here there are much less peptides that become insignificant after adjustment. In this plot we see that 726 peptides became insignificant, 547 became significant, and 1,078 are significant in both models. As in the previous experiments, we can check if these peptides are becoming insignificant due to an increase in variance. To check this we look at modified peptides whose adjusted fold change is within 10% of their unadjusted fold change and they become insignificant after adjustment. In this experiment there are only 25 peptides that meet this criteria, a small percentage of all modified peptides.

Analysis of modifications before and after protein adjustment can be seen in Supplementary Sec. 4.2.3Dataset 6 : Human - Ubiquitination - Label-free no global profiling runsubsubsection.4.2.3.

Noisy PTM measurements benefit from additional biological replicates

Here we analyze the sample size needed to achieve a desired statistical power. The proposed approach corrects 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 3. When the variation is increased, the experiment requires a larger number of replicates to reach the same power. Thus the statistical power is dependent on the variance from both the modified peptide and unmodifed protein as well as the number of replicates in each run.

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. For variance we used the biological experiments to determine what scenarios and 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 due to not using a separate global profiling run. 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 S17Power analysis of experiments with differing variances.figure.caption.20. 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. As long as there were the same overall number of replicates the results were the same. 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. It is clearly more important to have high replicates for the PTM run than the unmodified protein, when the PTM variance is higher. In cases where the number of replicates has to be limited, it is better to weigh them on the PTM side.

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. 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

Shigella: https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=4878d777c6b34cf8aaf8477e93140c4d

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.

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

[3] L. Käll and O. Vitek. "Computational mass spectrometry–based proteomics". In: *PLoS Comput. Biol.* 7 (12 Dec. 2011), e1002277.

[4] Roepstorff P. "Mass spectrometry in protein studies from genome to function." In: *Curr Opin Biotechnol.* 8.1 (1997), pp. 6–13.

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

[6] Natalia A. Petushkova et al. "Post-translational modifications of FDA-approved plasma biomarkers in glioblastoma samples". In: *PLoS ONE* 12 (2017).

[7] Junfeng 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. ISSN: 0021-9673. DOI: `https://doi.org/10.1016/j.chroma.2014.10.107`. URL: `https://www.sciencedirect.com/science/article/pii/S0021967314017208`.

[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.

[9] Matthew E. Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7 (Jan. 2015), e47–e47. ISSN: 0305-1048. DOI: `10.1093/nar/gkv007`. eprint: `https://academic.oup.com/nar/article-pdf/43/7/e47/7207289/gkv007.pdf`. URL: `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 (2014), pp. 2524–2536.

[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 Oct. 2020), pp. 1706–1723.

[12] Giovanni Luchetti et al. "Shigella ubiquitin ligase IpaH7.8 targets gasdermin D for degradation to prevent pyroptosis and enable infection". In: *Cell Host & Microbe* (2021). ISSN: 1931-3128. DOI: `https://doi.org/10.1016/j.chom.2021.08.010`. URL: `https://www.sciencedirect.com/science/article/pii/S1931312821003863`.

[13] Timurs Maculins et al. "Proteomics of autophagy deficient macrophages reveals enhanced antimicrobial immunity via the oxidative stress response". In: *bioRxiv* (2020). DOI: `10.1101/2020.09.10.291344`. eprint: `https://www.biorxiv.org/content/early/2020/09/12/2020.09.10.291344.full.pdf`. URL: `https://www.biorxiv.org/content/early/2020/09/12/2020.09.10.291344`.

[14] Y. Zhu et al. "DEqMS: A Method for Accurate Variance Estimation in Differential Protein Expression Analysis". In: *Molecular & cellular proteomics : MCP* 19 (2020), pp. 1047–1057.

[15] Florian P. Breitwieser and Jacques 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`. URL: `https://www.sciencedirect.com/science/article/pii/S1874391913000973`.

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

[17] Benjamin M. 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. ISSN: 01695347. DOI: `10.1016/j.tree.2008.10.008`.

[18] J. J. Faraway. *Extending the linear model with R*. 1st. Boca Raton, FL: Taylor & Francis Group, LLC, 2006.

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

[20] 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.

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

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

[23] Douglas Bates et al. "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1 (2015), pp. 1–48. DOI: 10.18637/jss.v067.i01. URL: https://www.jstatsoft.org/index.php/jss/article/view/v067i01.

[24] Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. "lmerTest Package: Tests in Linear Mixed Effects Models". In: *Journal of Statistical Software* 82.13 (2017), pp. 1–26. DOI: 10.18637/jss.v082.i13. URL: https://www.jstatsoft.org/index.php/jss/article/view/v082i13.
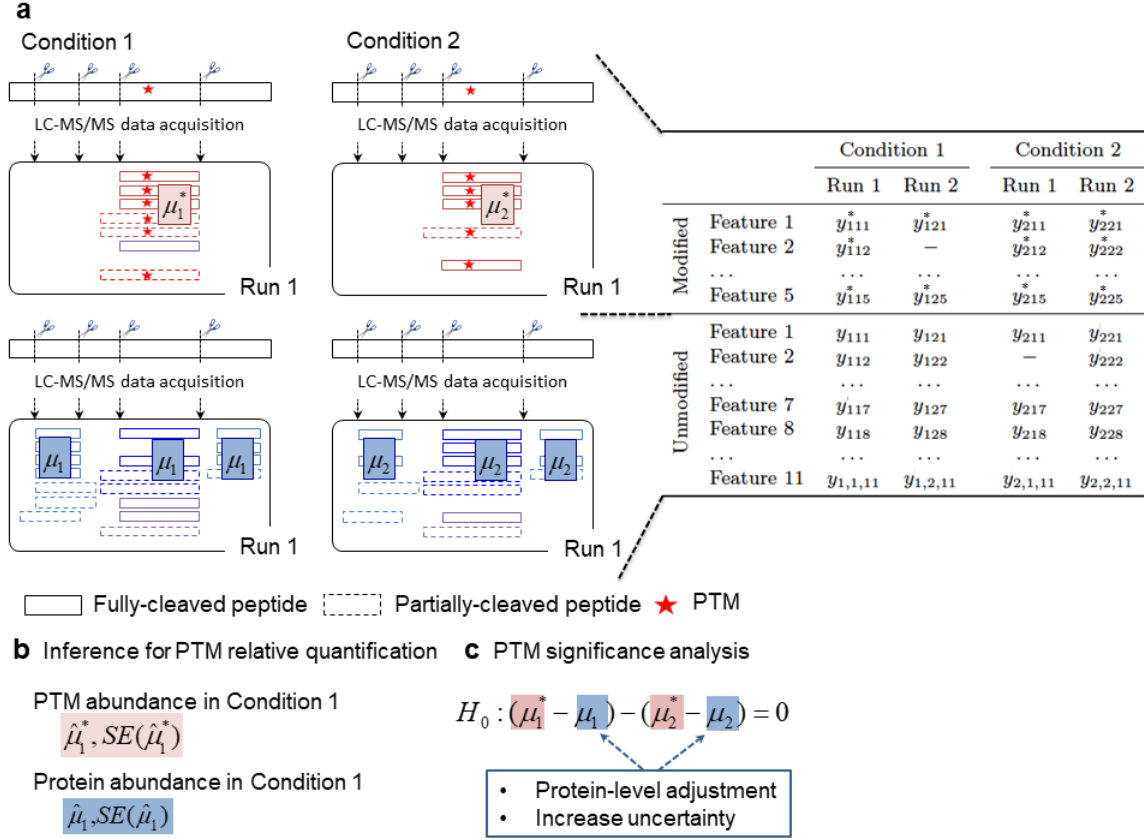
Figures

Figure 1: Data structure of a typical PTM experiment and goals of PTM characterization. (a) Schematic data representation, in a simplified case of two conditions and two 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 are viewed as repeated measurements of the underlying abundance of the PTM, where the abundance in Condition i is denoted by i*. Features corresponding to unmodified peptides are considered together to perform adjustment with respect to protein abundance, where the protein abundance in Condition i is denoted by i. Peptides can be fully cleaved (solid lines) and/or partially cleaved (dashed lines). Some spectral features can be missing. (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.
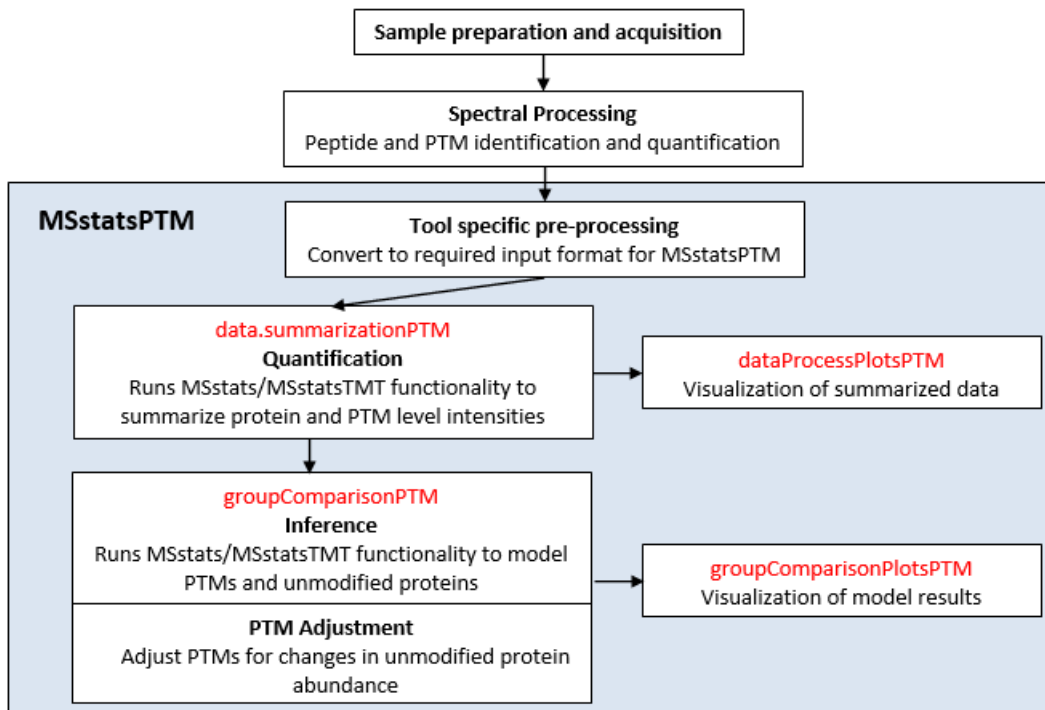
Figure 2: The workflow of MSstatsPTM and how it fits into the experimental pipeline. MSstatsPTM's workflow starts after modified and unmodified peptide quantification. First tool specific pre-processing is done, this includes modification site identification, general data cleaning, and formatting the 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.
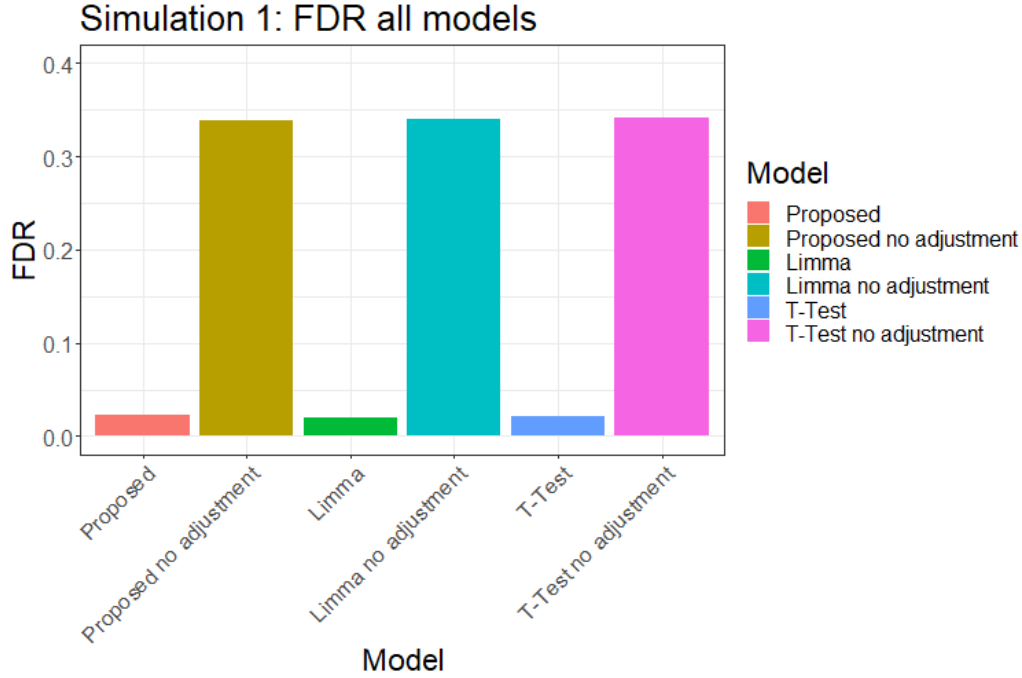
(a)
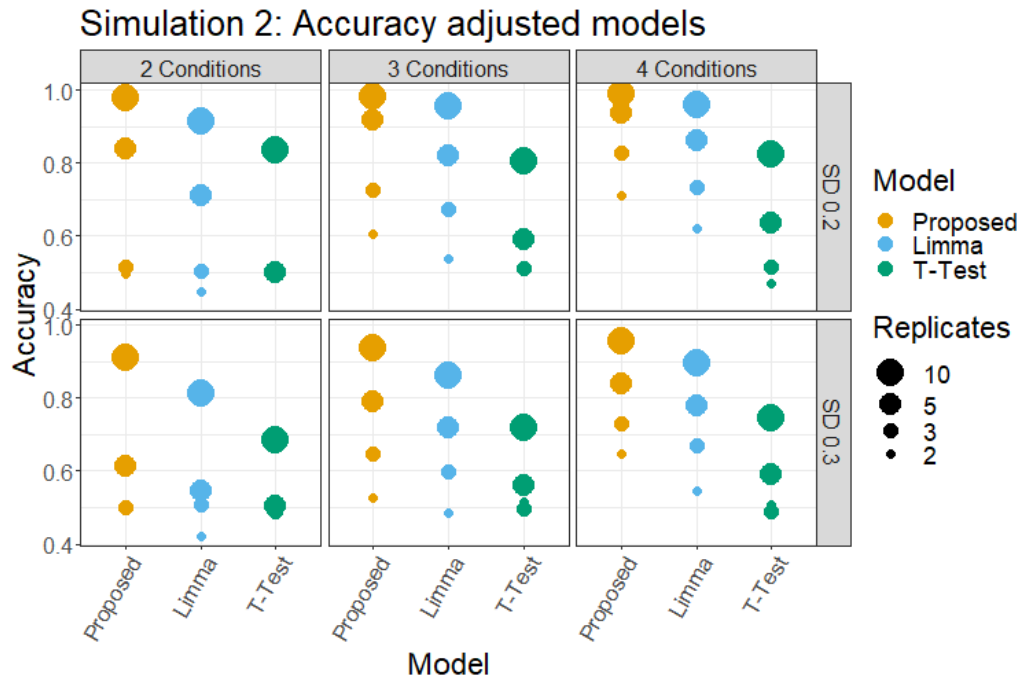
(b)

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 is apparent when including limited observations and missing values. Looking at accuracy, the proposed method outperforms Limma and t-test in nearly every model.
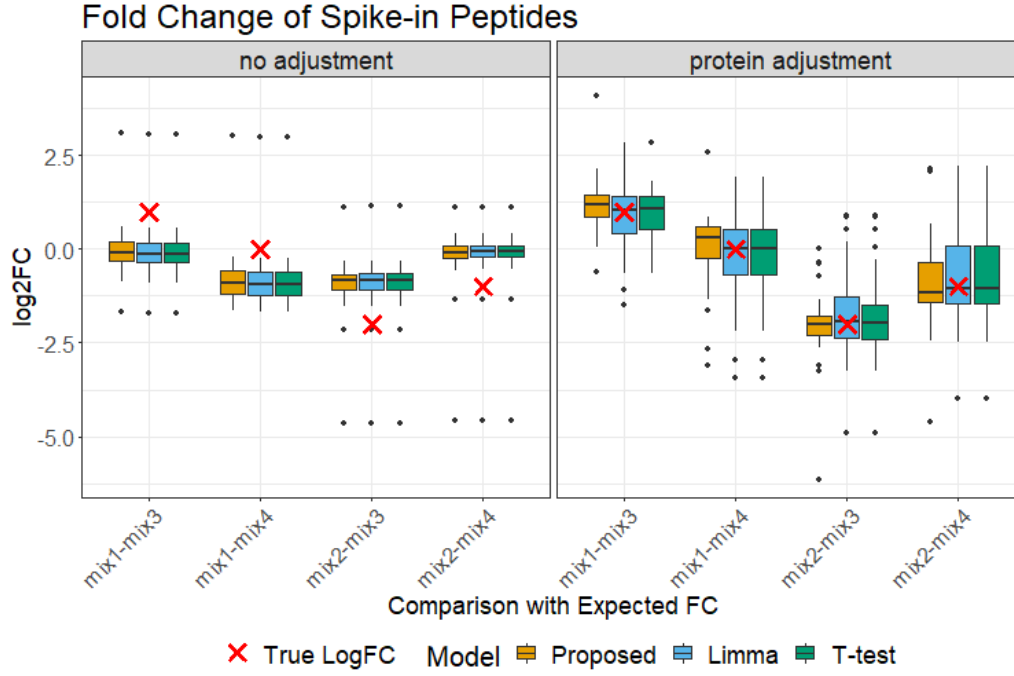
(a)
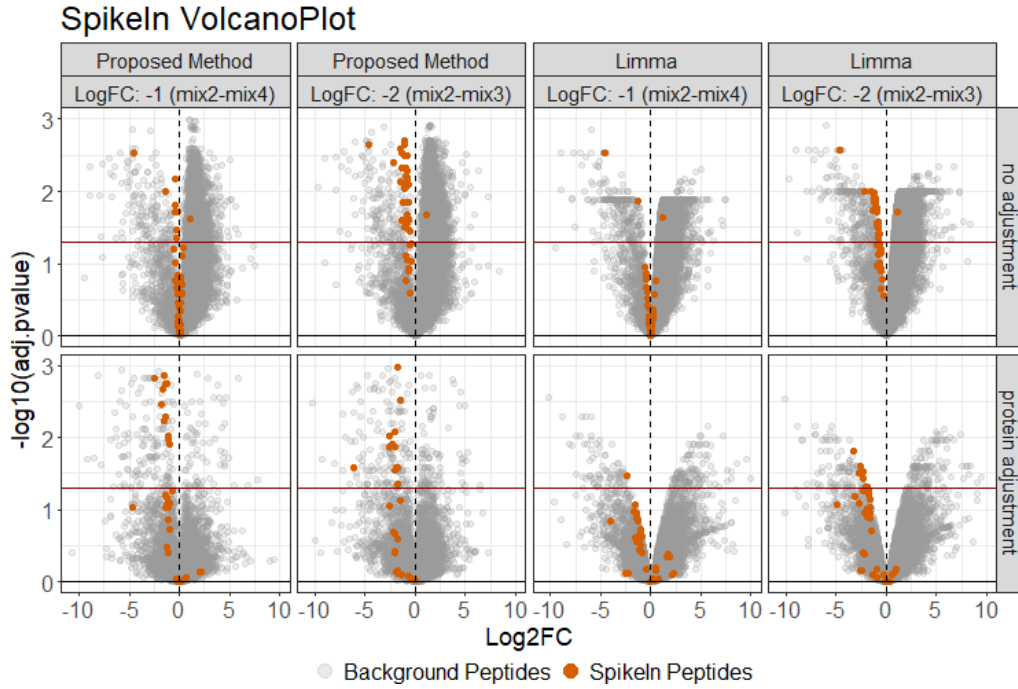
(b)

Figure 4: Dataset 3 : SpikeIn benchmark - Ubiquitination - Label-free. a) Before adjustment all models show the fold change of the spike-in peptides is systematically different from the expected fold change. After adjustment, this systemic difference is removed, however the inner quartile range of the Limma and t-test models is wider than the proposed method. b) The spike in peptides (colored red) do not follow the expected log fold change before adjustment. However, after adjustment, the spike in peptides are more in line with expectation. Using Limma the spike in peptides follow the expected log fold change after adjustment, however the majority of spike in peptides do not have a significant adjusted pvalue.
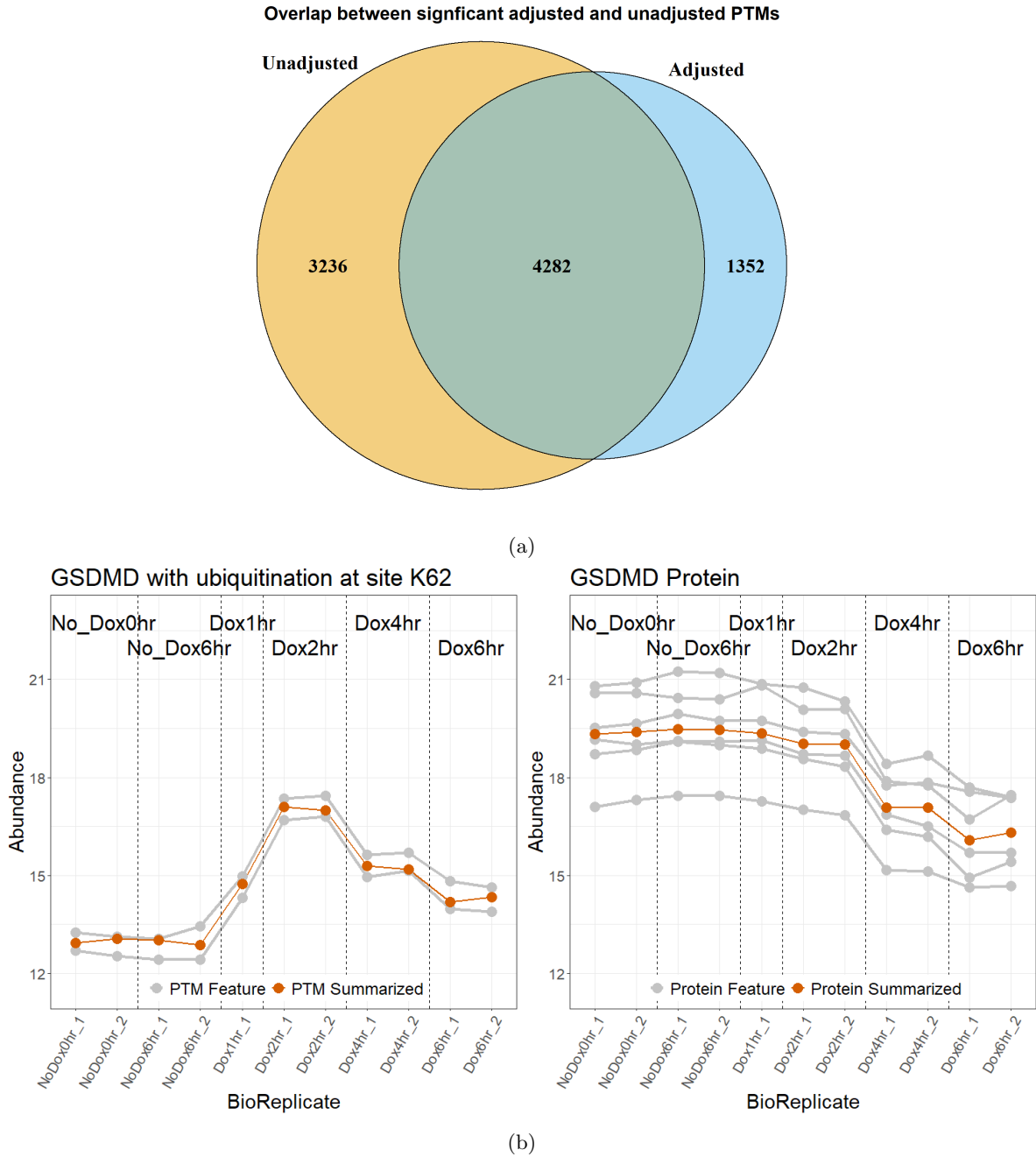
(a)

(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 is clear the conditions follow different trends. Specifically, there appears to be no change in abundance between Dox1hr and Dox4hr in the modified plot, however there is a large negative change when looking at the unmodified plot. This indicates the modification is confounded with changes in the unmodified protein.
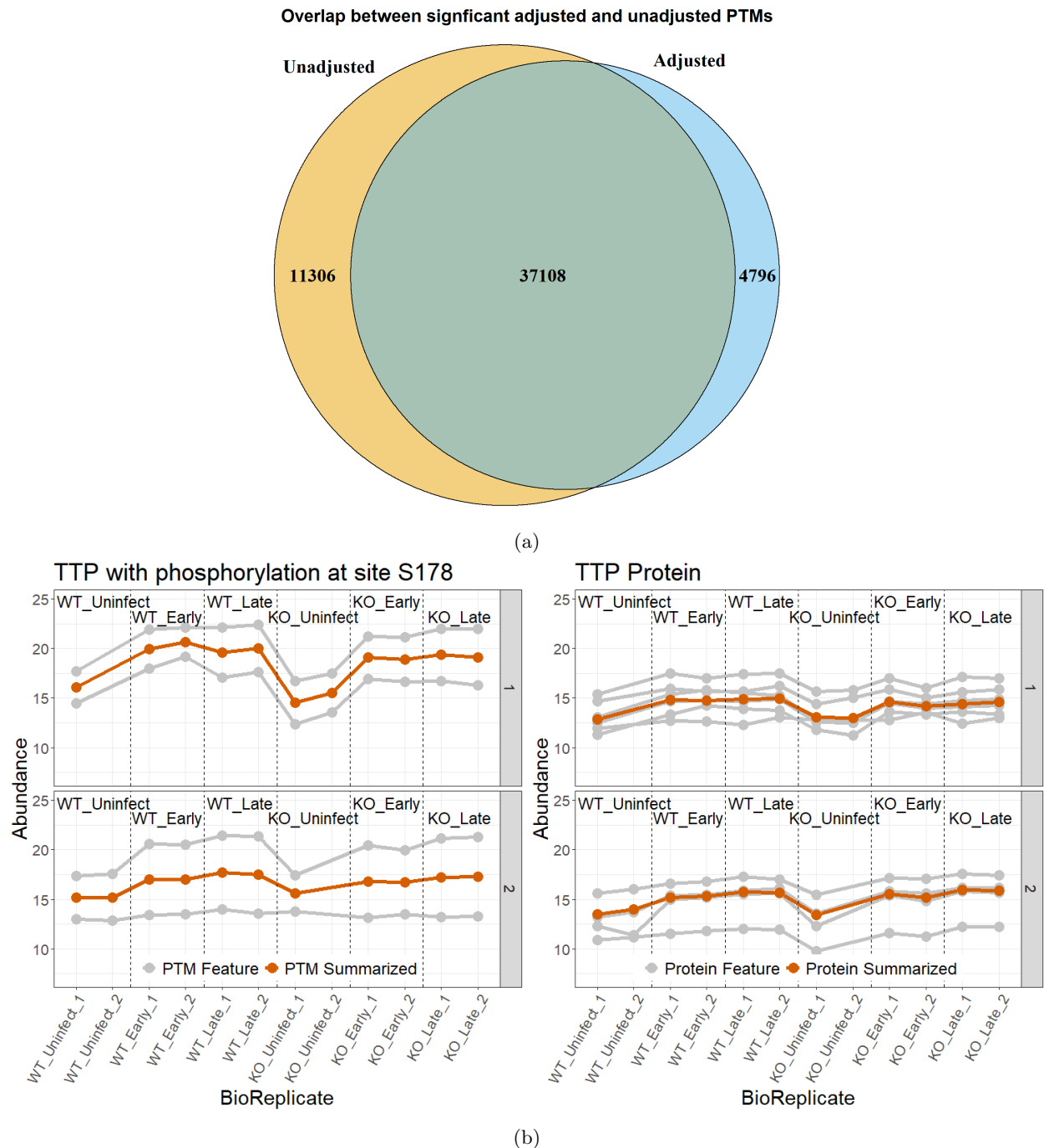
19

(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. 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 is clear the difference between conditions follow the same trend. Specifically, there is a positive adjustment in abundance when comparing WT_Uninfect to WT_Late in both the modification and global profiling run. This indicates the movement is driven by changes in global protein that is only accounted for in the model when adjusting for global protein abundance change.

Overlap between significant adjusted and unadjusted PTMs

(a)
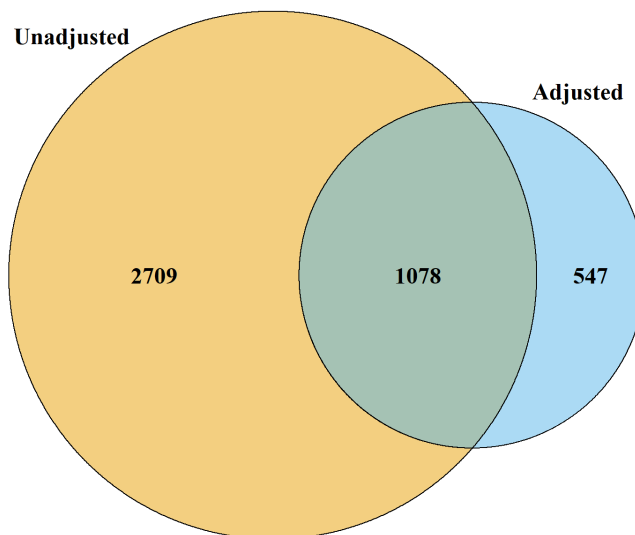
Signficant adjusted and unadjusted PTMs (matching only)
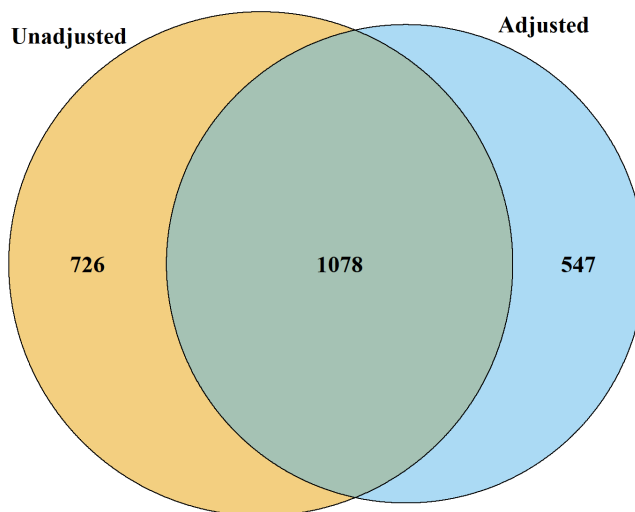
(b)

Figure 7: Dataset 6 : Human - Ubiquitination - Label-free no global profiling run. a) The overlap of differencial modified peptides for the PTM model with and without global protein level adjustment. Here many more PTMs became insignificant then became significant after adjustment. This is due to not having a global profiling run, resulting in a large lack of overlap between the modified peptides and unmodified proteins. b) Here we make the same comparison but only for modified peptides with a matching unmodified protein, so adjustment can be performed. In this case we see significantly less peptides become insignificant after adjustment. This highlights the need for a global profiling run if protein adjustment is going to performed.