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

Devon Kohler*[1], Tsung-Heng Tsai*[2], Meena Choi*[3], Ting Huang[1], Erik Verschueren[4], Trent Hinkle[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]Galapagos, Mechelen, Antwerp, Belgium
*Corresponding Authors

Abstract

The scientific community widely utilizes mass spectrometry (MS)-based proteomics to quantify the abundance of proteins and their post-translational modifications (PTMs). Experiments targeting PTMs face several specific challenges which are statistical in nature. These include the low abundance of modified proteo-forms, few representative peptides that span modification sites, and convolution with abundance changes in the overall protein expression. Due to these challenges, a robust approach to estimating relative systematic changes in PTMs should combine information pertaining to PTM sites over several peptides, replicates in multiple conditions, and consider sources of confounding and variation present in the experiment. We propose a general statistical model and workflow that is both reproducible and comprehensive. The method measures modified and unmodified peptide abundance by summarizing intensities through Tukey's median polish method. Then a model based on the family of linear mixed-effects models is fit. This model is automatically adjusted to the specific experimental design. Finally, the PTM abundances are adjusted to remove bias from changes in the overall protein. We implemented this model in the free and open-source R package *MSstatsPTM*. The approach is evaluated on computer simulations, a spike-in experiment with known ground truth, and three biological experiments with varying acquisition methods. Finally, the approach is compared against several existing methods and is shown to outperform all other methods.

Introduction

The signaling mechanisms that allow cells to mount a dynamic and fast response to a multitude of events are primarily facilitated by the modification of proteins at specific residues, acting as molecular on/off switches[7] [6]. Mass spectrometry-based label-free proteomics is broadly established as the tool-of-choice for unbiased and large-scale identification and quantification of proteins and their post-translational modifications (PTMs) using liquid chromatography coupled with mass spectrometry (LC-MS)[10] [16]. 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[14]. 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 proteo-forms dictates that a global interrogation can only be achieved through large-scale enrichment protocols with modification-specific antibodies or beads. 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). Third, unless early signaling events are interrogated, the interpretation of the relative changes in modification occupancy are inherently convoluted with changes in the overall protein expression, making the interpretation of the results not straightforward[15]. Therefore, a robust approach to estimate systematic relative changes in post-translational modifications, at scale, should not only combine the quantitative information pertaining to a PTM site over peptides and replicates in multiple conditions, but take into account various sources of variations and confounding factors present in the experiments.

We propose a general statistical approach, which explicitly characterizes the variations and confounding factors present in bottom-up PTM experiments. The proposed approach is aimed at the detection of quantitative changes in PTMs between conditions utilizing procedures developed for summarization of LC-MS data, quantitative characterization of site-specific PTMs, and adjustment with respect to protein abundance. Quantitative analyses of PTMs often involve comparisons between multiple inter-related conditions of the same biological system. The general statistical framework underlying the proposed approach allows for analyzing experiments with complex designs and different acquisition methods.

The proposed approach was implemented using the R coding language and evaluated using datasets from computer simulations, benchmark controlled mixtures, and biological investigations. The approach was compared against the commonly applied t-test and Limma methods[17]. The results demonstrate that by appropriately leveraging the information from the entire dataset, the proposed approach improves the reproducibility and accuracy of the estimates of PTM fold changes and results in a better calibrated type I error rate. Finally, 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[5] [9].

Experimental Procedures

This section describes the datasets and experiments used to evaluate the proposed method. The datasets were chosen to provide a performance benchmark of the proposed approach against existing methods when ground truth is known, as well as represent a variety of experimental designs and acquisition methods where the approach can be applied. To benchmark the dataset, computer simulations and a spike-in experiment were used. The computer simulations vary 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 an real experiment where ground truth is known. Finally the proposed approach was used to model three biological experiments which differed in experimental design, acquisition strategy, and use various organisms. These experiments were chosen to show the proposed method's applicability to many different types of experiments.

Computer simulations

Dataset 1 : Computer simulation 1 - label free Here simulations were generated with a high number of modified and unmodified peptide features and no missing values. Multiple simulations were generated with different numbers of replicates and conditions. In all simulations 1000 modified peptides were generated, half of which were true positives and the other half true negatives. Half of true positives and true negatives were generated with convolution 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 - missing values and low replicates 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 masked 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 the simulations were generated can be found in Supplementary Sec. 4.

Dataset 3 : SpikeIn benchmark - label-free - KGG enriched

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 changes 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. All comparisons with respect to human lysate serve as a null model, there was no change in global protein nor PTM abundance between conditions. Additionally, E coli Lysate was used to normalize total protein levels prior to enrichment or global protein profiling. 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.

Dataset 4 : Human-1mix-TMT - ubiquitination

The proposed approach was evaluated on an experiment where Shigella ubiquitin ligase IpaH7.8 was shown to inhibit 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 either infected or uninfected with IpaH7.8-deficient Shigella flexneri and measurements were taken at different time periods. 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 Section 5.1.

Dataset 5 : Mouse-2mix-TMT - phosphorylation

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. 5.2.

Dataset 6 : Human label free quantification - 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. A corresponding mixed effects model was fit per modification and global protein as described previously. 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 substantial 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 expression 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[20]. 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)[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 2.

Proposed approach

Figure 5.2 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 often acquired. If unmodified features are unavailable for any given modified feature, 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.

Statistical modeling and parameter estimation

The proposed approach 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. In addition mixture may be included for features acquired via tandem mass tag (TMT) methods. 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[5]. 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[19]. 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[3] [8]. Statistical modeling and quantification for global proteomics data are performed by the same procedure as for PTM data.

Detection of changes in PTMs

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 5.2. 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. 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 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[18]. 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[2]. More details are provided in Supplementary Sec 3.2

Missing value imputation

$MSstatsPTM$ gives the ability to impute 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[19]. 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.

Extension to TMT experiments

The statistical modeling approaches discussed above can also be extended to Tandem Mass Tag (TMT) labeling methods. TMT experiments introduce an additional source of variation in the form of different mixtures. To account for mixture variance, first feature intensities are summarized again using Tukey's Median Polish and missing values are imputed, if desired, then a new linear mixed effects model with an added term for the mixture is included. The workflow and statisical model described generally follow the methods used for modeling proteins in $MSstatsTMT$[9]. These methods are repeated to quantify both modified and unmodified peptide abundance. Once both modified and unmodified peptides are modeled, the PTM model is adjusted for changes in global protein abundance using the same methods described previously.

For more information about the extension to Tandem Mass Tag experiments see Supplementary Sec. 3.4.

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 and combined together for input into the rest of the package. 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.

After using the converters, the next step is peptide/protein summarization and missing value imputation. 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, as mentioned in Section 3, 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 automatically 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 Section 3.

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 $MSstats$ [5] and $MSstatsTMT$ [9], depending on the data acquisition type. The statistical modeling relies on the functionality from the R packages $lme4$ [1] and $lmerTest$ [11].

Code availability

Bioconductor: http://www.bioconductor.org/packages/release/bioc/html/MSstatsPTM.html
Github: https://github.com/Vitek-Lab/MSstatsPTM

Evaluation

The performance of the proposed method was evaluated on simulated and spike in datasets with known ground truth, as well as biological experimental data where the ground truth was not known. For the experiments where ground truth is known, we calculated summary statistics using the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). To evaluate the performance of different methods, we considered the false discovery rate (FDR = FP/(TP+FP)), the recall rate (TP/(TP + FN), and total accuracy ((TP + TN) / (TP + TN + FP + FN)). For biological experiments where ground truth was not known, we evaluated significantly differential peptides with an α of .05. Comparisons were made before and after adjusting for changes in unmodified protein abundance and the results are presented in the form of venn diagrams and profile plots. All datasets in this paper were evaluated using the $MSstatsPTM$ package.

Computer simulation 1 - label free

The proposed approach was evaluated against two-sample t-test and Limma. Here we see that not accounting for changes in unmodified peptide level results in a high false discovery rate, seen in Figure 5.3. When adjusting for changes in the unmodified peptide, all methods performed similarly in terms of FDR. Recall and accuracy of the proposed approach and Limma performed similarly, with Limma performing slightly better with low replicates, while t-test lagged behind. Two-sample t-test only uses data within the groups of interest while ignoring the remainder of the data. Consequently, it gave similar performance across cases with different number of conditions. In contrast, the proposed approach and Limma leverage all available information, which resulted in improved power with increased number of conditions.

Computer simulation 2 - missing values and low replicates

In this simulation, changes in unmodified peptide level still needed to be accounted for to control the FDR. Once controlling for changes in unmodified peptide, it is clear that the proposed method performed strongest across all methods, as seen in Figure 5.4. The proposed method well calibrated model accuracy and the recall rate, even when the number of replicates were low. Finally, 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. In comparison, Limma and t-test showed much wider distributions. This fold change comparison can be seen in Figure 5.5.

Further details reviewing the results of each simulation and their comparison with respect to method performance can be seen in Supplementary Section 4.

SpikeIn benchmark - KGG enriched

In Figure 5.6 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, colored grey, show many false positives before adjustment is made. Again after adjustment the results improve and the number of false positives decrease significantly.

In Figure 5.7 we can see the results of the same experiment modeled using Limma with protein adjustment. Again the red labeled spike-in peptides are more inline with expected fold change after adjustment is applied. However, 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 differential modified peptides. In terms of false positives, the results are very similar to the proposed method, with many false positives before adjustment and much fewer afterward.

Further results of this experiment, including a comparison to t-test and fold change distribution of each method before and after adjustment, are provided in Supplementary Section 4.2.

Human-1mix-TMT - ubiquitination

The results of this experiment are summarized in Figures 5.8 and 5.9. In Figure 5.8 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 correct 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 who's 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 5.9, looking at modified peptide $GSDMD\_HUMAN|P57764$ 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 very 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.

Mouse-2mix-TMT - phosphorylation

The results of this experiment are summarized in Figures 5.10 and 5.11. In Figure 5.10 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 39,550 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 who's 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, 776 peptides became insignificant due to an increase in variance. This is still a very small portion (4.02%) 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 5.11 the profile plot of two specific modified peptides, $TTP\_MOUSE|P22893\_S178$ and $KI67\_MOUSE|E9PVX6\_T215$, show the power of the proposed method. The profile plot of modified peptide $TTP\_MOUSE|P22893\_S178$ 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 2.015 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 .275. In contrast $KI67\_MOUSE|E9PVX6\_T215$ originally shows a small positive log fold change of .212 between conditions WT_Early and WT_Uninfected. This is contrast with a negative fold change in the global profiling run of $-.616$. In adjusting the modified peptide for changes in the global protein, the log fold change becomes .828 and the adjusted pvalue goes from .452 to .037.

Human label free quantification - no global profiling run

As discussed previously, there was no unmodified global profiling run performed in this experiment. Once identification and quantification of the Ubiquitinated 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 very low feature counts for the unmodified protein model.

The results of this experiment are summarized in Figures 5.12 and 5.13. In 5.12 the number of significant modified peptides before and after protein adjustment is shown. In the top plot we can see there are a very 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 very small percentage of all modified peptides.

In Figure 5.13 the profile plot of peptide $P52209\_K059$ is shown. The abundance between conditions appears flat in the modified peptide profile plot, however this is contrast with a large negative change in the corresponding global protein. Specifically looking at the comparison between the CCCP and Control conditions, the unadjusted model showed a small log fold change of $-.255$ and insignificant adjusted pvalue of .104. However when adjusting for changes in the undmodified protein, the log fold change increased to -2.73 and adjusted pvalue became significant at $1.067e-04$. Using the proposed method we get increased information that would otherwise have been missed.

Sample size calculation

[Devon: Update this section in main and supp]

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. Additionally, the implementation allows the methods to be applied many different acquisition methods.

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.

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, expression 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 most likely run into major 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). We are also interested in characterizing the interplay of PTMs at multiple sites. The proposed statistical methods are implemented as an R package *MSstatsPTM* available on Bioconductor and Github.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

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

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

[9] 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.

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

[11] 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.

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

[15] 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.

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

[17] 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.

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

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

[20] 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.

Figures

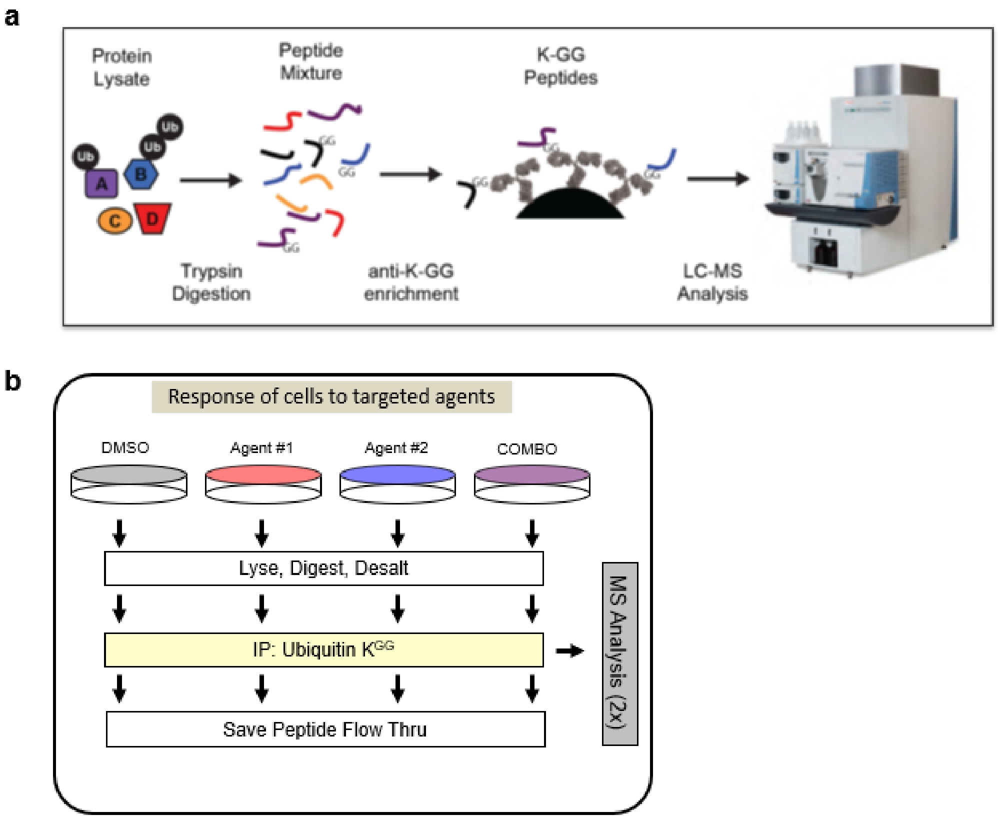
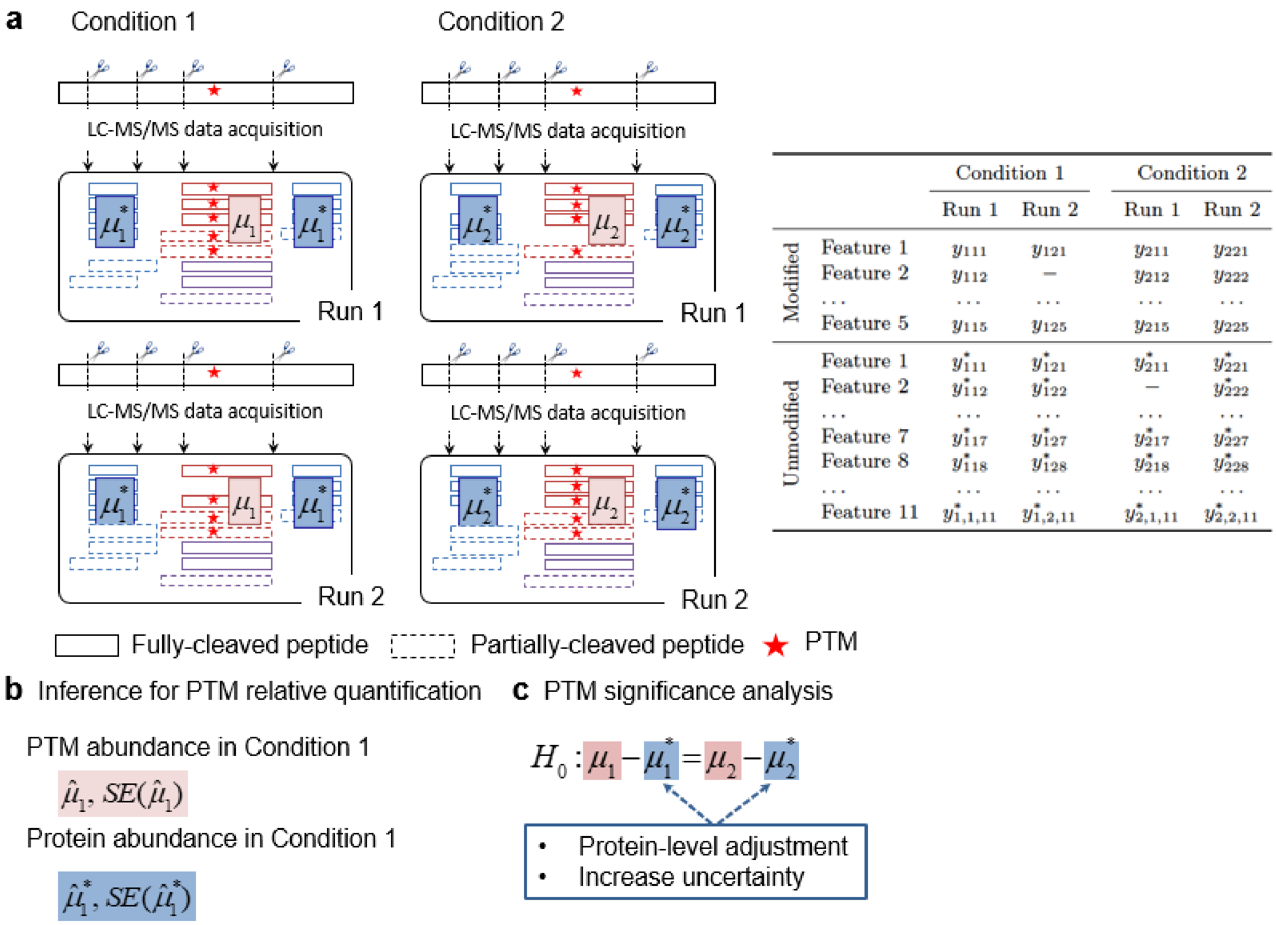


Figure 5.1: [Meena?: Can you write about how MS for PTMs is run?]


a
Condition 1
Condition 2
LC-MS/MS data acquisition
LC-MS/MS data acquisition
μ_1^*
μ_1
μ_2^*
μ_2
Run 1
Run 1
LC-MS/MS data acquisition
LC-MS/MS data acquisition
Run 2
Run 2
Fully-cleaved peptide
Partially-cleaved peptide
PTM
b Inference for PTM relative quantification
c PTM significance analysis
PTM abundance in Condition 1
$\hat{\mu}_1, SE(\hat{\mu}_1)$
Protein abundance in Condition 1
$\hat{\mu}_1^*, SE(\hat{\mu}_1^*)$
$H_0: \mu_1 - \mu_1^* = \mu_2 - \mu_2^*$
Protein-level adjustment
Increase uncertainty

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

Figure 5.2: 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.

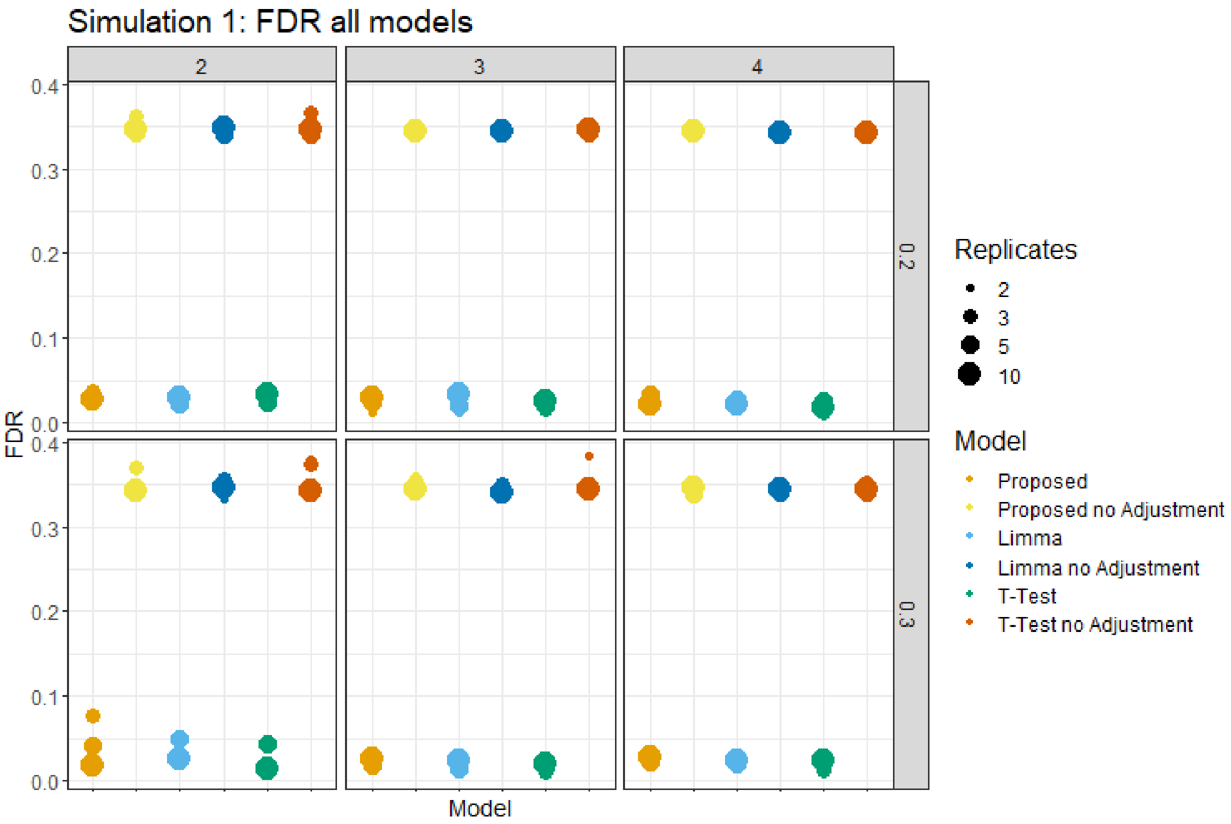


Figure 5.3: Dataset 1 : Computer simulation 1 - label free. All the considered methods in the first computer simulation correctly calibrated FDR when adjusting for changes in protein abundance. In comparison, the methods without accounting for the protein-level changes resulted in off-target, high false positive rates.

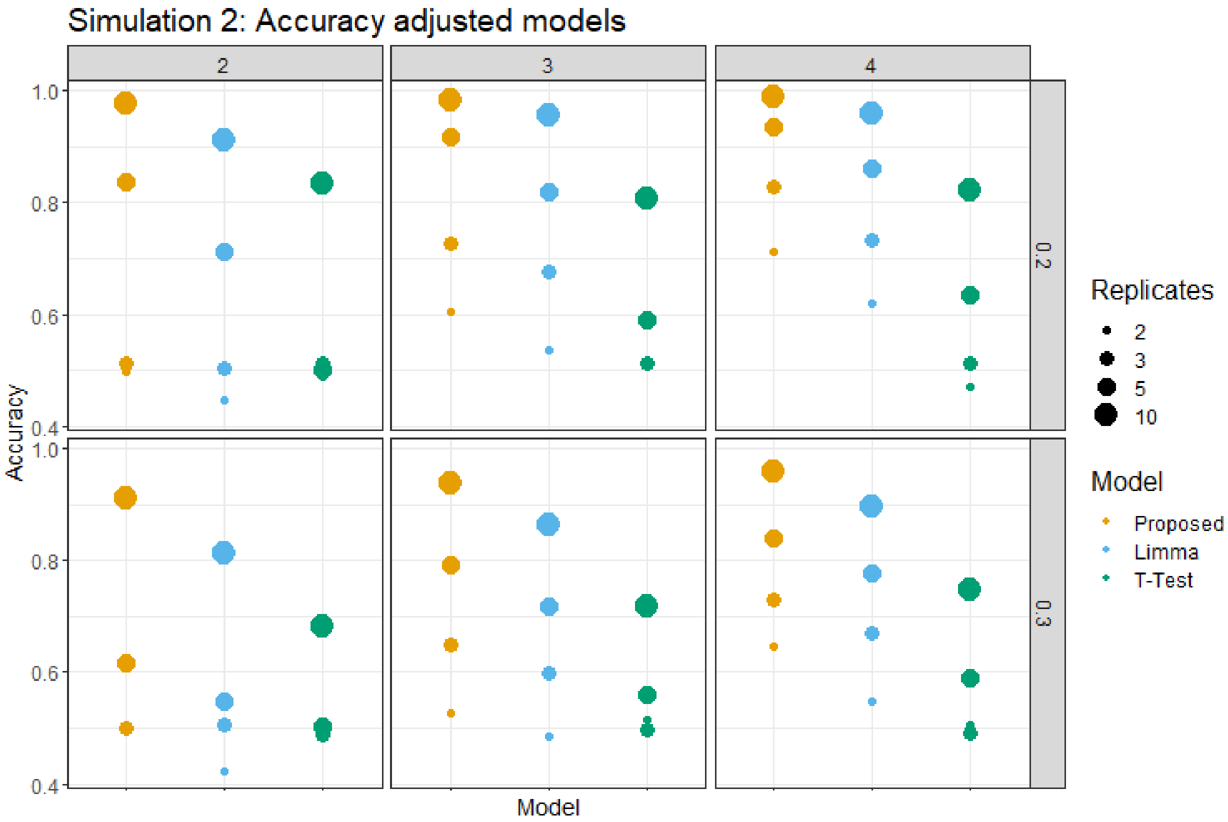


Figure 5.4: Dataset 2 : Computer simulation 2 - missing values and low replicates. The advantage of using the proposed approach is apparent when looking at the second computer simulation, which includes limited observations and the presence of missing values. In overall accuracy, the proposed method performs stronger than Limma and t-test in nearly every model. Even at lower replicates the proposed method still outperformed Limma. The lowest performaning method was *t*-test. Limma shows comparable performance to the proposed method in a clean experiment, however when real world data problems are introduced it is clear the proposed method is more robust.

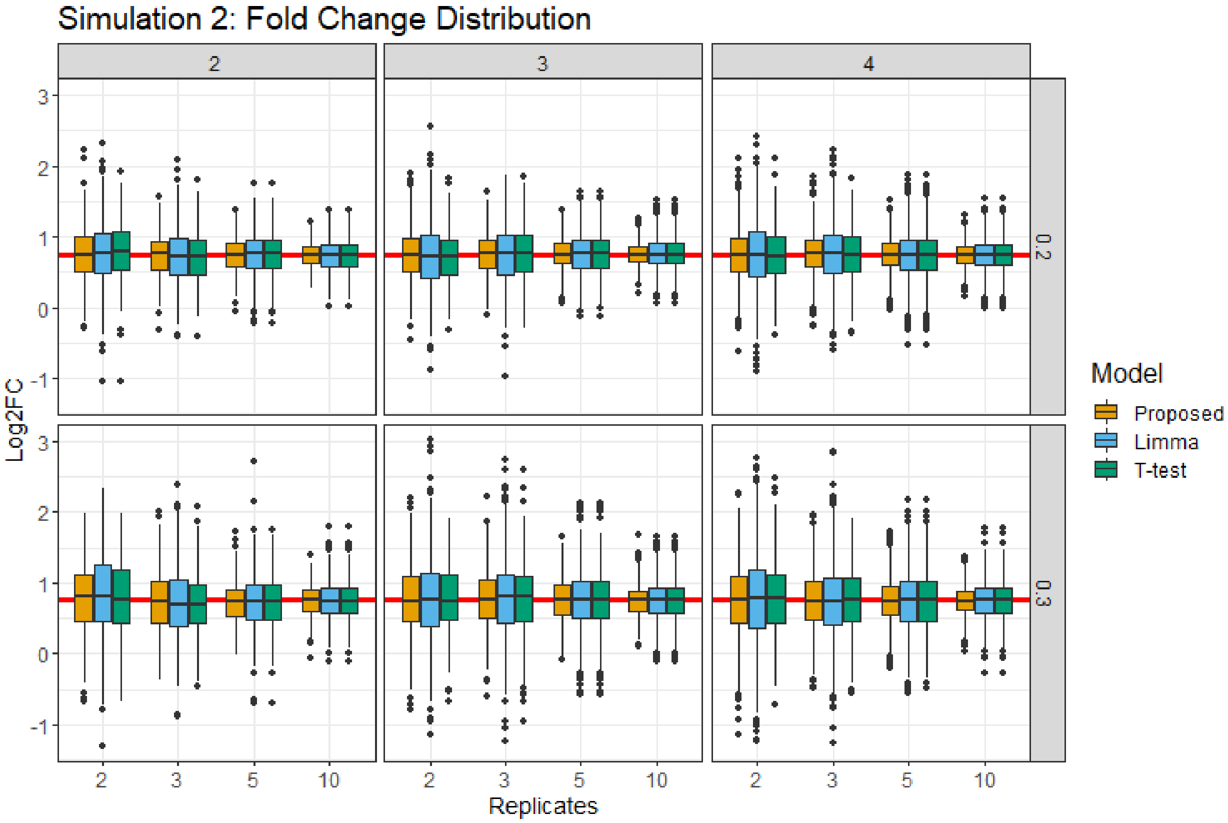


Figure 5.5: Dataset 2 : Computer simulation 2 - missing values and low replicates. In the second computer simulation with missing values and low PTM features all methods correctly estimated the fold change with a median log change of .75. The proposed method in this simulation had a visibly tighter distribution around the median when compared to Limma and *t*-test. While the mean estimated fold changes were similar, the proposed method exhibited a stronger performance in correctly estimating the fold change for all peptides.

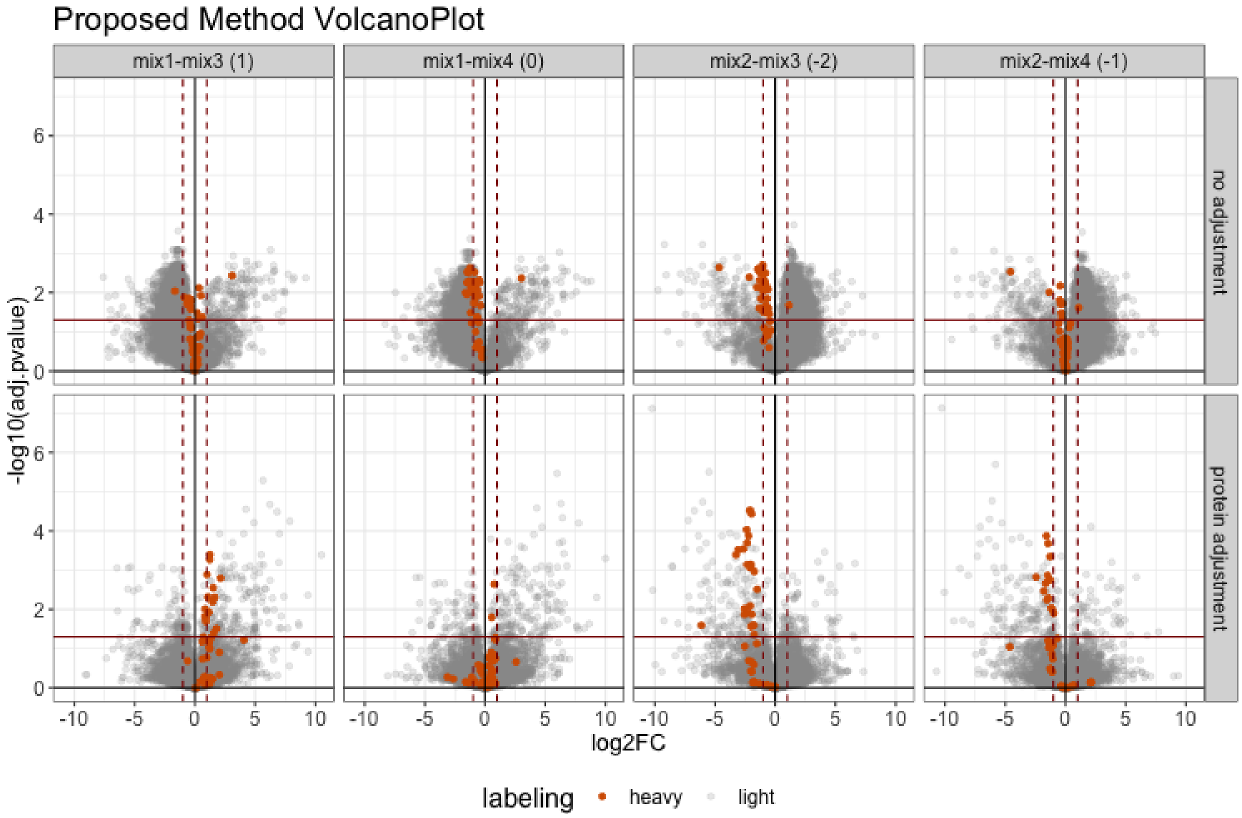


Figure 5.6: Dataset 3 : SpikeIn benchmark - label-free - KGG enriched. Using the proposed method to model the benchmark experiment the spike in peptides (colored red) do not follow the expected log fold change before adjustment. After adjusting for changes in overall protein abundance the spike in peptides are more in line with expectation. Additionally the background grey colored peptides showed many false positives before adjustment. After adjustment the false positives were decreased considerably.

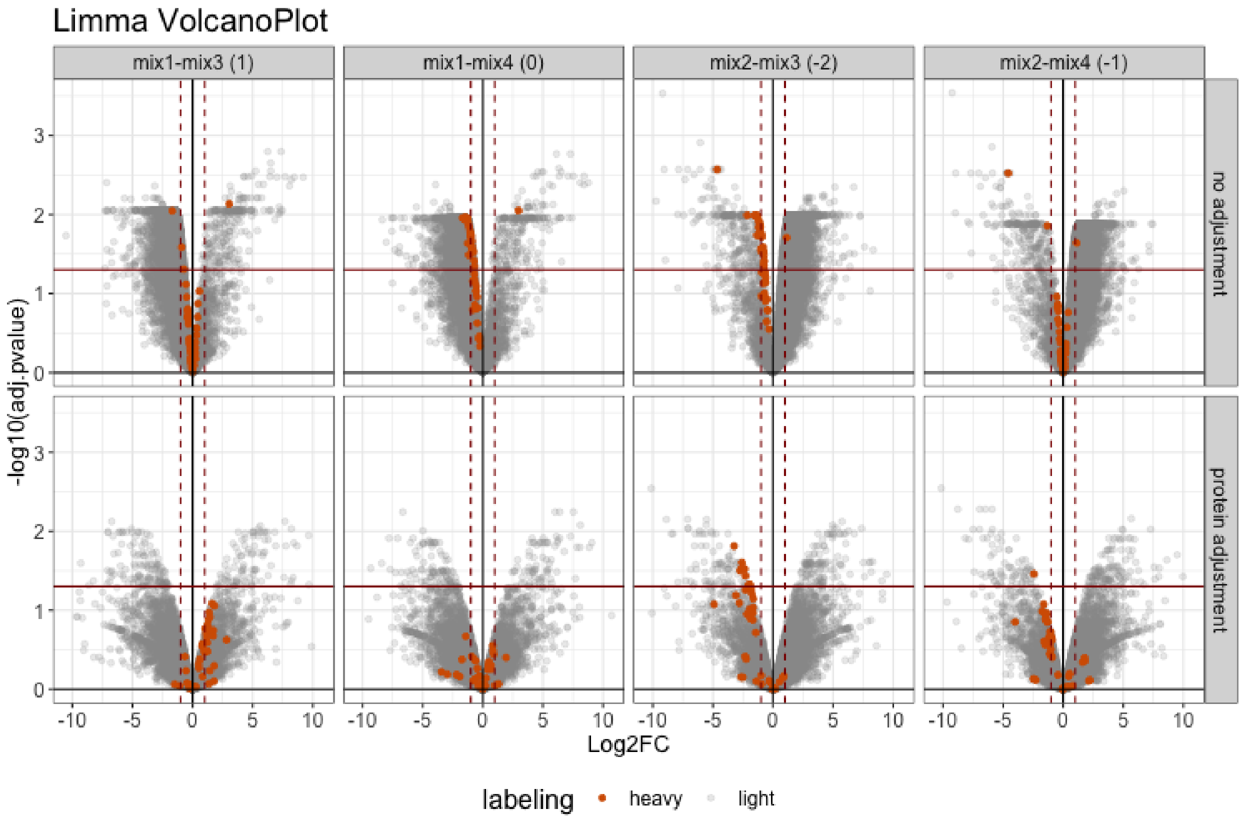


Figure 5.7: Dataset 3 : SpikeIn benchmark - label-free - KGG enriched. Here the Limma method was used to model the spike in experiment. Using Limma the spike in peptides follow the expected log fold change better after adjusting for changes in protein level. However, while the fold change is much more accurate, the majority of spike in peptides do not have a significant adjusted pvalue. In terms of false positives, the results are very similar to the proposed method, with many false positives before adjustment and much fewer after.

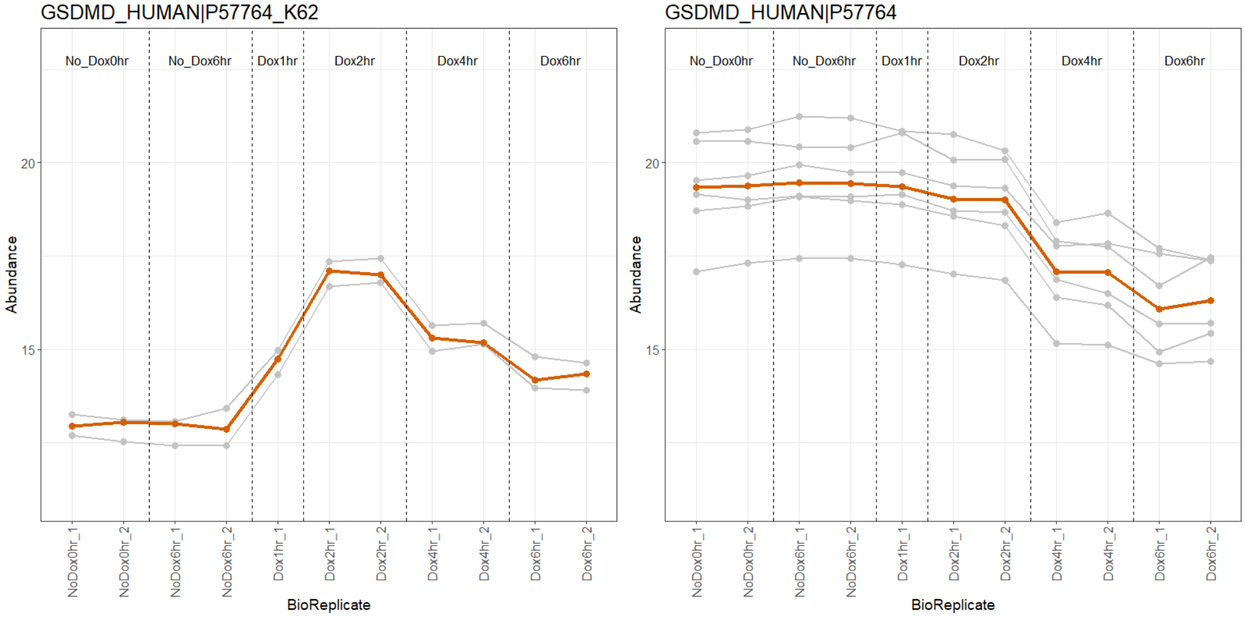


Figure 5.8: Dataset 4 : Human-1mix-TMT - ubiquitination. The overlap of differencial modified peptides for the PTM model with and without global protein level adjustment. More PTMs became insignificant after adjustment then became significant. For the peptides that became insignificant in the adjusted model, their change in abundance was 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.

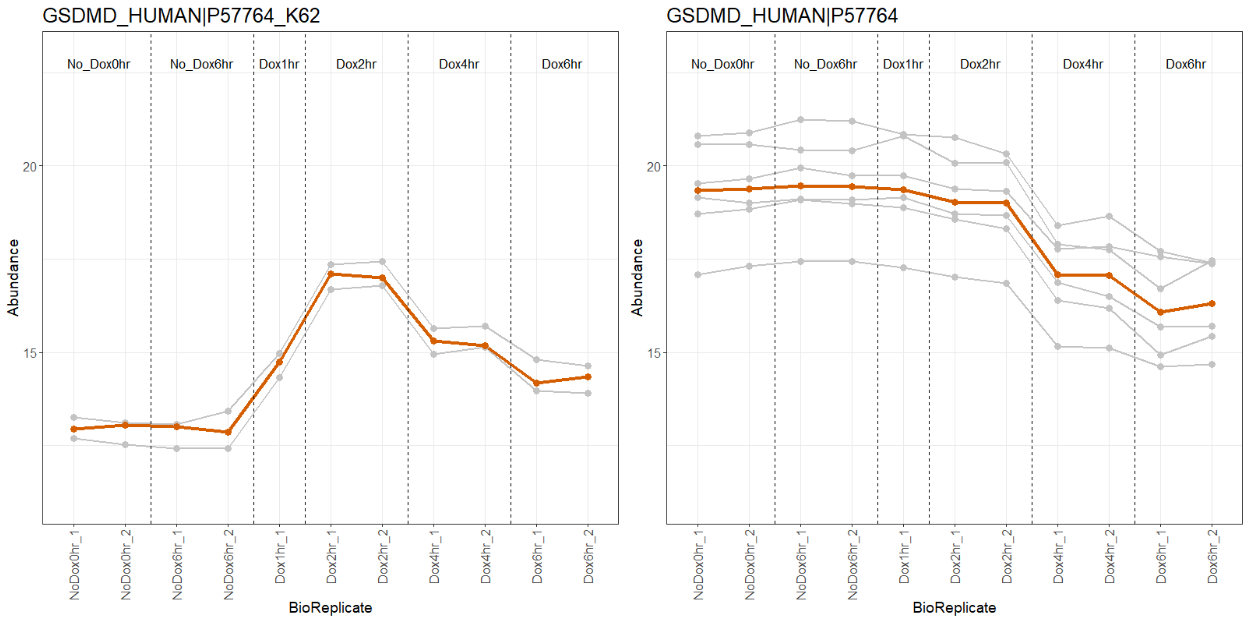


Figure 5.9: Dataset 4 : Human-1mix-TMT - ubiquitination. Comparing the global profiling of protein $GSDMD\_HUMAN|P57764$ with the ubiquitination of the protein at site $K62$. The individual PSM features are shown in grey, while the feature summarization is shown in red. When looking at the summary of 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.

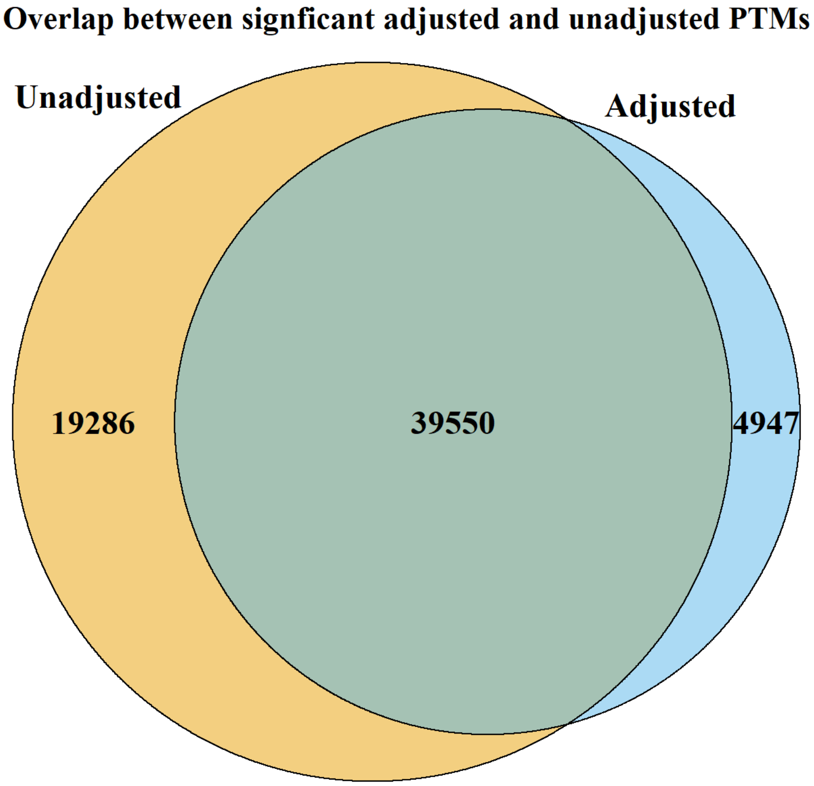
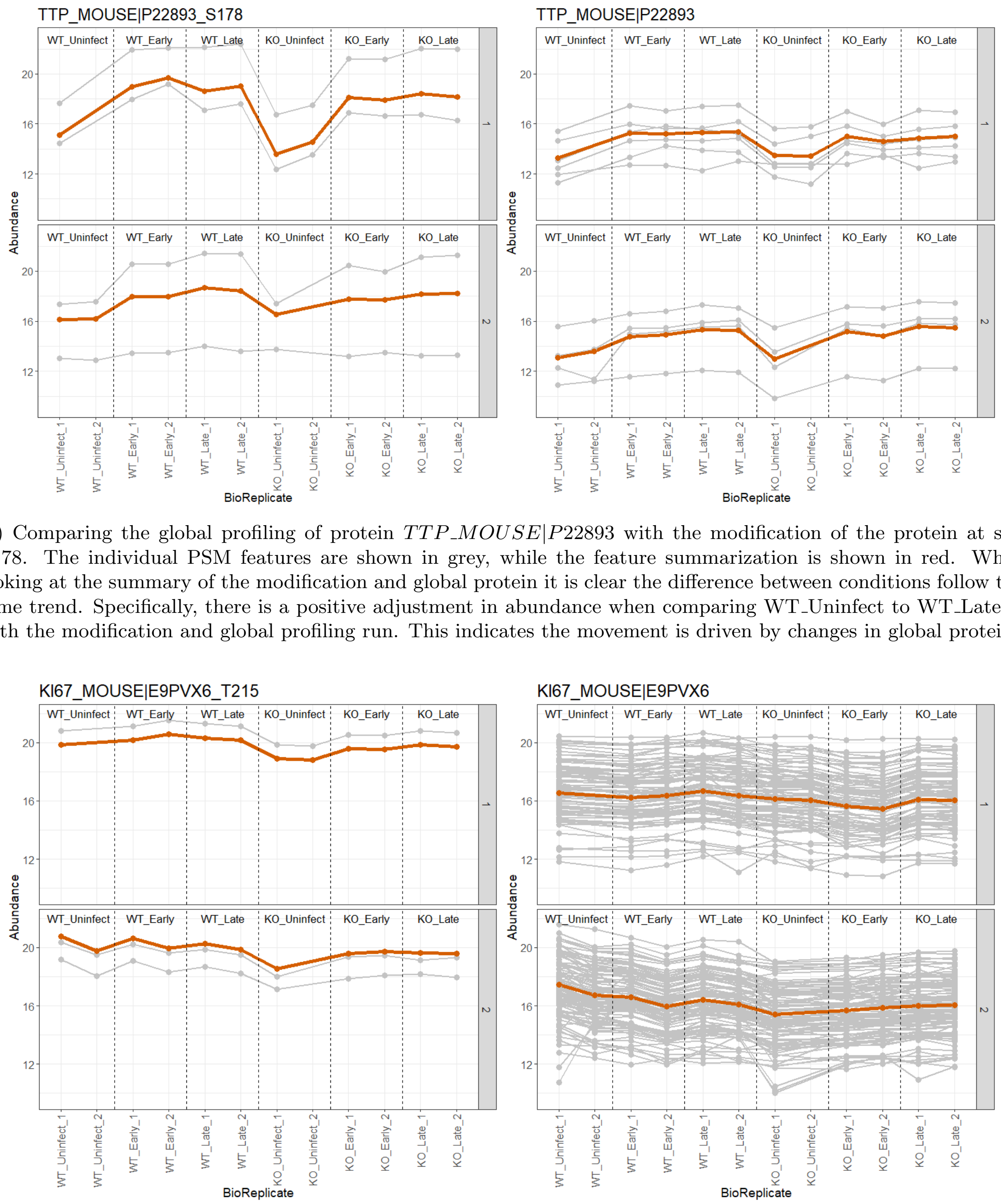


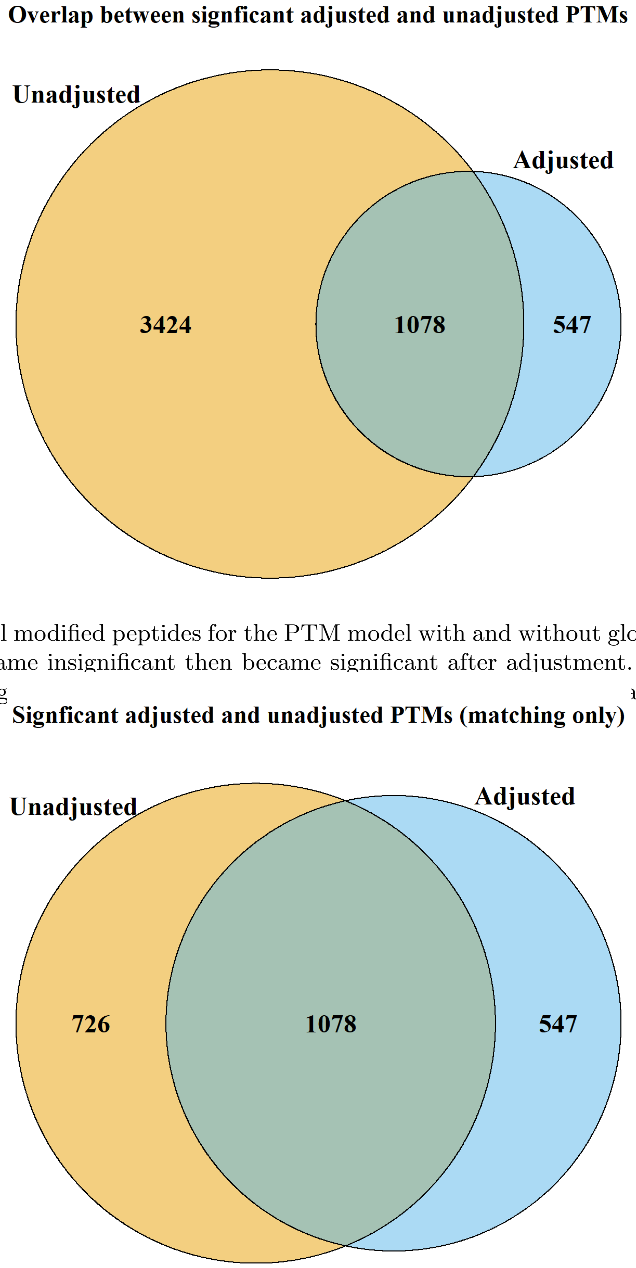
Figure 5.10: Dataset 5 : Mouse-2mix-TMT - phosphorylation. The overlap of differencial modified peptides between the PTM model with and without global protein level adjustment. Again more PTMs became insignificant after adjustment then became significant.



(a) Comparing the global profiling of protein $TTP\_MOUSE|P22893$ with the modification of the protein at site $S178$. The individual PSM features are shown in grey, while the feature summarization is shown in red. 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.

(b) Comparing the global profiling of protein $KI67\_MOUSE|E9PVX6$ with the modification of the protein at site $T215$. In this case the modification and global protein trend in different directions. Specifically, comparing WT_Uninfect and WT_Early there is a slightly positive change in abundance, however in the global profiling there was a negative change. In this case the profile plot indicates the effect of the modification is masked by the change in global protein abundance. Additionally this profile plot shows the large difference in available features between modifications and global protein.

Figure 5.11: Dataset 5 : Mouse-2mix-TMT - phosphorylation. Profile plots for PTMs P22893_S178 and E9PVX6_T215 in the Shigella flexneri experiment.



(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 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.

Figure 5.12: Dataset 6 : Human label free quantification - no global profiling run. Overlap of significant PTMs before and after protein adjustment.

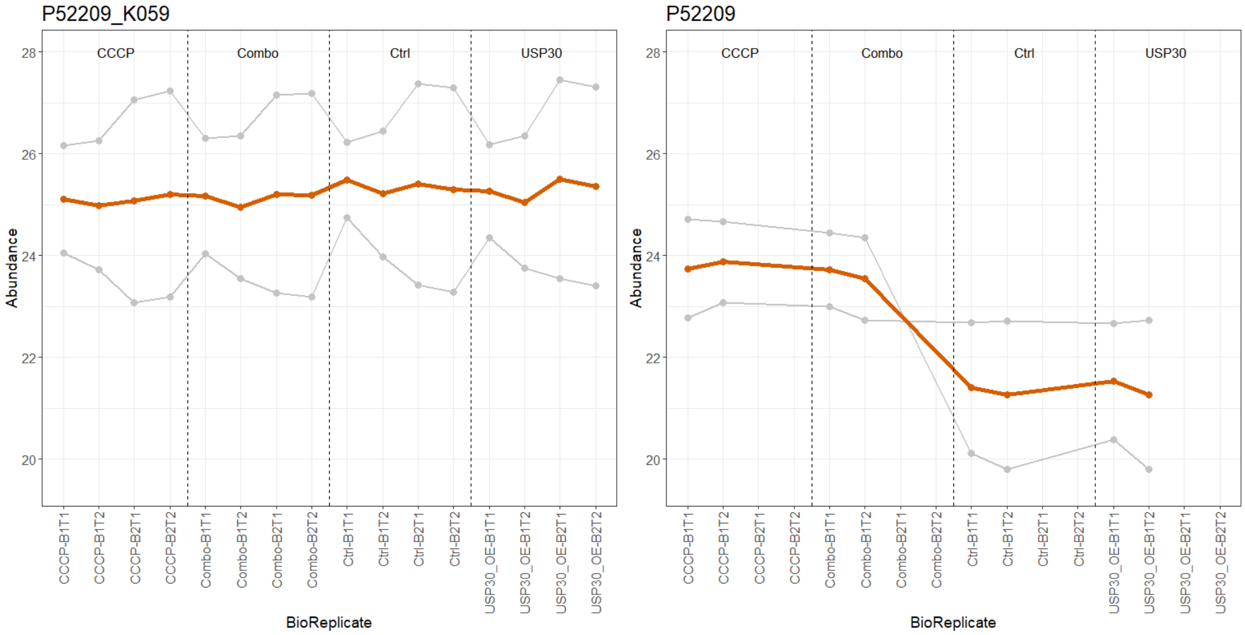


Figure 5.13: Dataset 6 : Human label free quantification - no global profiling run. Comparing the global profiling of protein $P52209$ with the modification of the protein at site $K059$. The modification appears generally unchanged between conditions, whereas the global profiling run shows the CCCP and Combo conditions at a higher relative abundance compared to the Control and USP30_OE. This indicates that the modification actually had a major effect when comparing CCCP and Combo to Control and USP30_OE, which would have been missed without adjusting for global protein changes.