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

Supplementary Information

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

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

Contents

1 Proposed approach

1.1 Design of PTM experiments - Extension to complex designs

The proposed approach applied to experiments with complex designs still allows for sample size calculation and power analysis. To make the calculation the standard deviation and degrees of freedom are needed. The estimation of these quantities is dependent on the experimental design. In Figure S1 different experimental designs that the proposed approach can be applied to, along with their estimation of standard error and degrees of freedom is shown.

		Model	Estimated Log-fold change	Theoretical variance	Estimated variance	Degrees of freedom
Label-free (Y_{cs} is $\log_2$ intensity in Condition c and Subject s)	**Group comparison**	$Y_{cs} = \mu + Condition_c + \varepsilon_{cs}$ $\sum_{c=1}^{C} Condition_c = 1$ $\varepsilon_{cs} \sim iid\ N(0,\sigma^2)$	$\bar{Y}_{c.} - \bar{Y}_{c'.}$	$\frac{2\sigma^2}{S}$	$\frac{2S\sum_{c=1}^{C}(y_{cs} - \bar{y}_{c.})^2}{S(CS-C)}$	$CS - C$
	Time course	$Y_{cs} = \mu + Condition_c + Subject_s + \varepsilon_{cs}$ $\sum_{c=1}^{C} Condition_c = 1$ $Subject_s \sim iid\ N(0,\sigma_S^2)$ $\varepsilon_{cs} \sim iid\ N(0,\sigma^2)$	$\bar{Y}_{c.} - \bar{Y}_{c'.}$	$\frac{2\sigma^2}{S}$	$\frac{2\sum_{c=1}^{C}\sum_{s=1}^{S}(y_{cs} - \bar{y}_{s.} - \bar{y}_{c.} + \bar{y}_{..})^2}{S(C-1)(S-1)}$	$(C-1)(S-1)$
TMT (Y_{mcs} is $\log_2$ intensity in Condition c and Subject s from Mixture m)	**Group comparison**	$Y_{mcs} = \mu + Condition_c + Mixture_m + \varepsilon_{mcs}$ $\sum_{c=1}^{C} Condition_c = 1$ $Mixture_m \sim iid\ N(0,\sigma_M^2)$ $\varepsilon_{mcs} \sim iid\ N(0,\sigma^2)$	$\bar{Y}_{.c.} - \bar{Y}_{.c'.}$	$\frac{2\sigma^2}{MS}$	$\frac{2S\sum_{c=1}^{C}\sum_{m=1}^{M}(y_{mcs} - \bar{y}_{m..} - \bar{y}_{.c.} + \bar{y}_{...})^2}{MS(MCS-C-M+1)}$	$MCS - C - M + 1$
	Time course	$Y_{mcs} = \mu + Condition_c + Subject_{sm} + \varepsilon_{mcs}$ $\sum_{c=1}^{C} Condition_c = 1$ $Subject_{sm} \sim iid\ N(0,\sigma_S^2)$ $\varepsilon_{mcs} \sim iid\ N(0,\sigma^2)$	$\bar{Y}_{.c.} - \bar{Y}_{.c'.}$	$\frac{2\sigma^2}{MS}$	$\frac{2S\sum_{c=1}^{C}\sum_{m=1}^{M}(y_{mcs} - \bar{y}_{ms.} - \bar{y}_{.c.} + \bar{y}_{...})^2}{MS(C-1)(MS-1)}$	$(C-1)(MS-1)$

Figure S1: Different models that are fit depending on the experimental design. Models are fit for Label-free and TMT acquisition methods, as well as group comparison and time course experimental designs. The estimation of standard error and degrees of freedom for each model is shown.

The standard error and degrees of freedom are estimated separately for both the PTM and unmodified protein models. Once this is done we can use these these values to calculate the sample size and perform power analysis. Given q, the desired false discovery rate, β, the average Type II error rate, $m_0/(m_0+m_1)$, the fraction of truly differentially modified PTM sites in the comparison. The minimal number of replicates is determined by the variance of the estimated log-fold change

$\mathrm{SE}^2(\hat{\Delta})$ as

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

where

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

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. The degrees of freedom are estimated for the appropriate model using restricted maximum likelihood.

2 Experimental datasets

2.1 Experiments with known ground truth

The proposed statistical approach was evaluated and compared to the *t*-test and Limma methods using computer simulations. Specifically, their statistical properties with respect to protein-level adjustment and real world experimental conditions were evaluated.

2.1.1 Computer simulation

Differential levels of modified peptides may be due to differential modifications and/or changes in protein abundance. Approaches without considering protein-level adjustment lose track of an important aspect in interpreting observed changes in PTM abundance, which may result in misleading conclusions. To highlight the necessity of the adjustment, we compared six different approaches as follows: a) proposed approach, b) proposed approach without adjusting for unmodified peptides c) *t*-test (with adjustment), d) *t*-test (no adjustment), e) Limma (with adjustment), and f) Limma (no adjustment).

In experiments of complex designs, multiple inter-related conditions are often compared together. The proposed approach and Limma leverage measurements in all conditions for the inference of the underlying abundance, whereas *t*-test uses measurements from the two conditions being compared. To highlight this distinction, multiple conditions of data were generated. While multiple conditions were generated, all comparisons were still made with the same fold change between conditions (0.75). This ensured that all peptides were differentially abundant by the same amount and the only difference in calculations was using more conditions in the modeling.

2.1.2 Dataset 1 : Computer simulation 1 - Label-free

In the first simulation an experiment with many features per PTM and unmodified protein was created. Additionally this simulation contained no missing data. These attributes are not representative of a real experiment, but provide a baseline for model performance.

- Mean of log-intensity: 25
- Standard deviations of log-intensities for modified and unmodified peptides: 0.2, 0.3
- Difference in PTM abundance between conditions: 0, 0.75, 1.5, 2.25
- Difference in protein abundance between conditions: 0, 0.75, 1.5, 2.25
- Number of replicates: 2, 3, 5, 10
- Number of conditions: 2, 3, 4
- Number of realizations: 1000
- Number of features per PTM: 10

- Number of features per unmodified protein: 10
- Missing data: no missing value

The results are summarized from Figure S2 to Figure S3. These results include the false discovery rate with and without adjusting for changes in unmodified protein abundance, false positive rate, and overall accuracy. Additionally, the results are compared to simulation 2 described in the next section.

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

In the second simulation we introduced limited feature observations per PTM as well as masking a portion of the observation to simulate missing values. This is more in line with what we would expect in a biological experiment and provides a more realistic expectation of model performance. The number of PTM features and missing data percentage where determined by looking at the features and missing data in the biological experiments in this paper.

- Mean of log-intensity: 25
- Standard deviations of log-intensities for modified and unmodified peptides: 0.2, 0.3
- Difference in PTM abundance between conditions: 0, 0.75, 1.5, 2.25
- Difference in protein abundance between conditions: 0, 0.75, 1.5, 2.25
- Number of replicates: 2, 3, 5, 10
- Number of conditions: 2, 3, 4
- Number of realizations: 1000
- Number of features per PTM: 2
- Number of features per unmodified protein: 10
- Missing data: 20% of the observations for PTMs and Proteins were masked with NA at random

The results are summarized from Figure S5 to Figure S6.

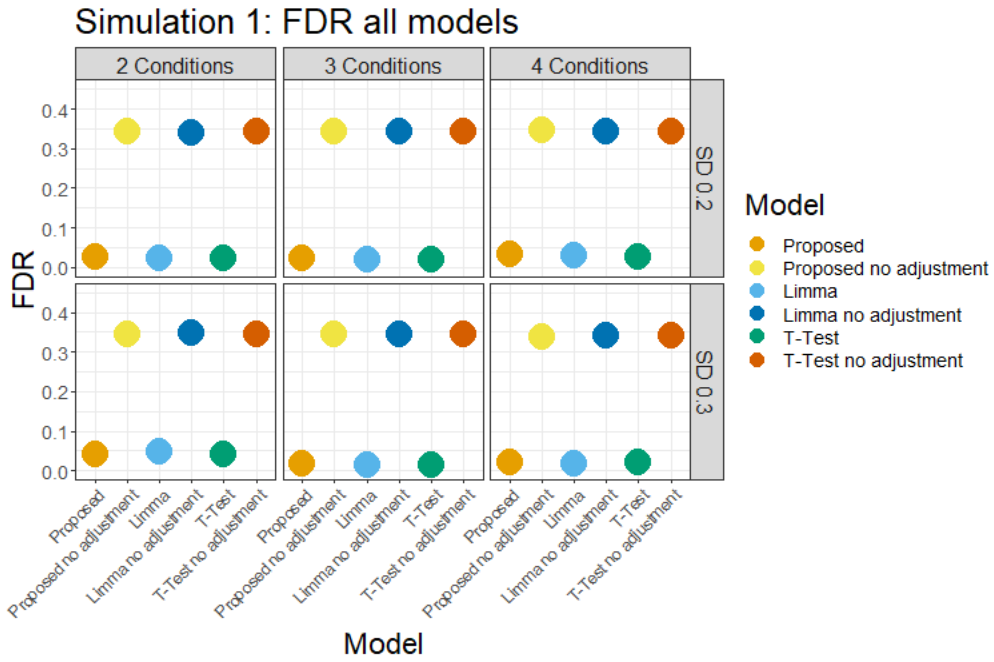

Figure S2.a: All the considered methods in simulation 1 correctly calibrated FDR when adjusting for changes in protein abundance. In comparison, the methods without accounting for the protein-level changes resulted in off-target, high false discovery rates. The performance of the models without adjustment was much lower than those with adjustment, thus only models with adjustment are compared going forward.

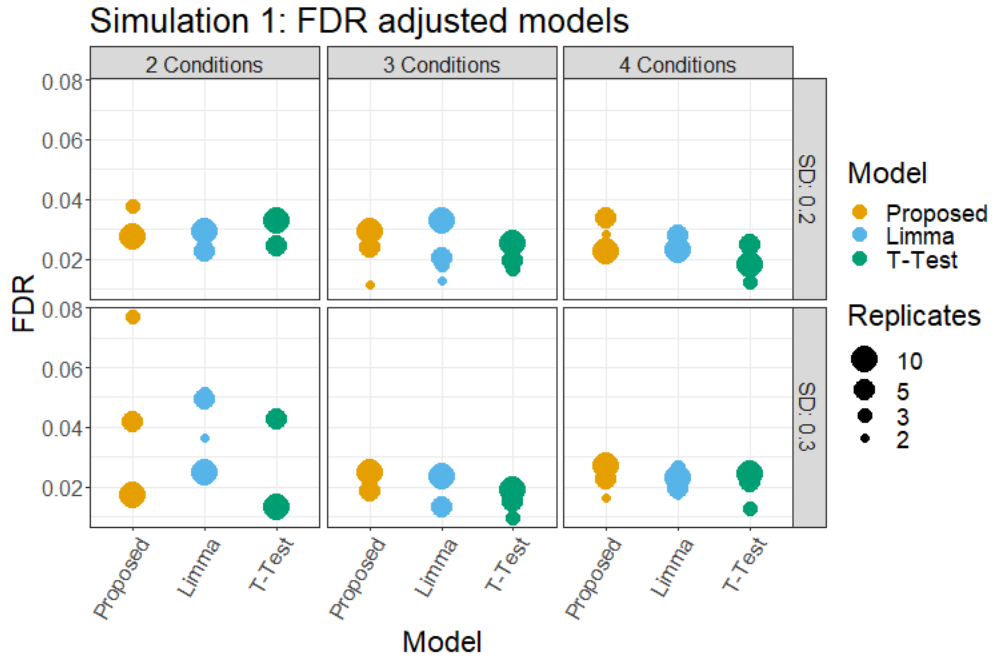

Figure S2.b: The considered methods with protein adjustment were compared in detail. All three methods with adjustment generally performed similarly in terms of FDR.

Figure S2: FDR of Simulation 1.

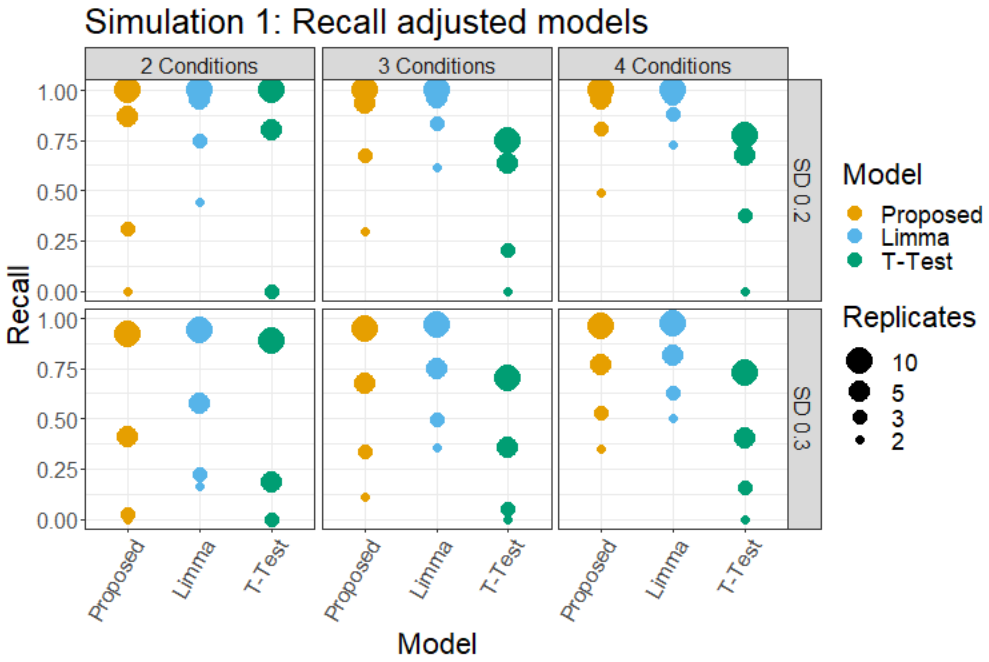

Figure S3.a: All methods with adjustment were compared by comparing recall in simulation 1. Limma performed the strongest here when the number of replicates were low. At higher replicates the performance of the proposed methods and Limma were comparable. *t*-test clearly performed worse across all methods.

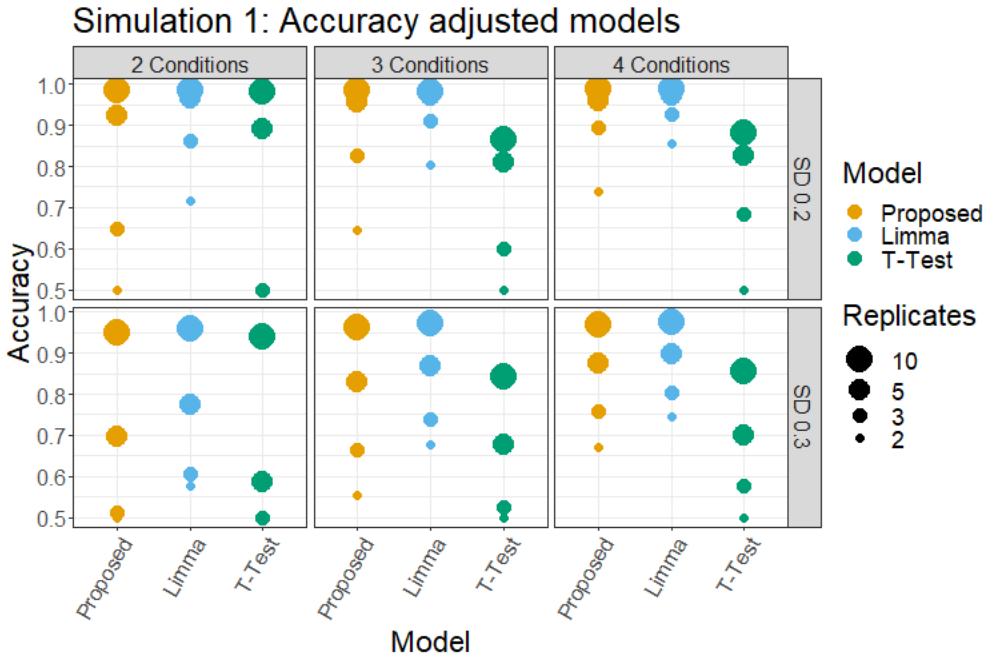

Figure S3.b: The overall accuracy plot mimicked the observations in the recall plot. Limma performed strongest when replicates were low, but was comparable to the proposed method with more replicates.

Figure S3: Recall and Accuracy results of Simulation 1.

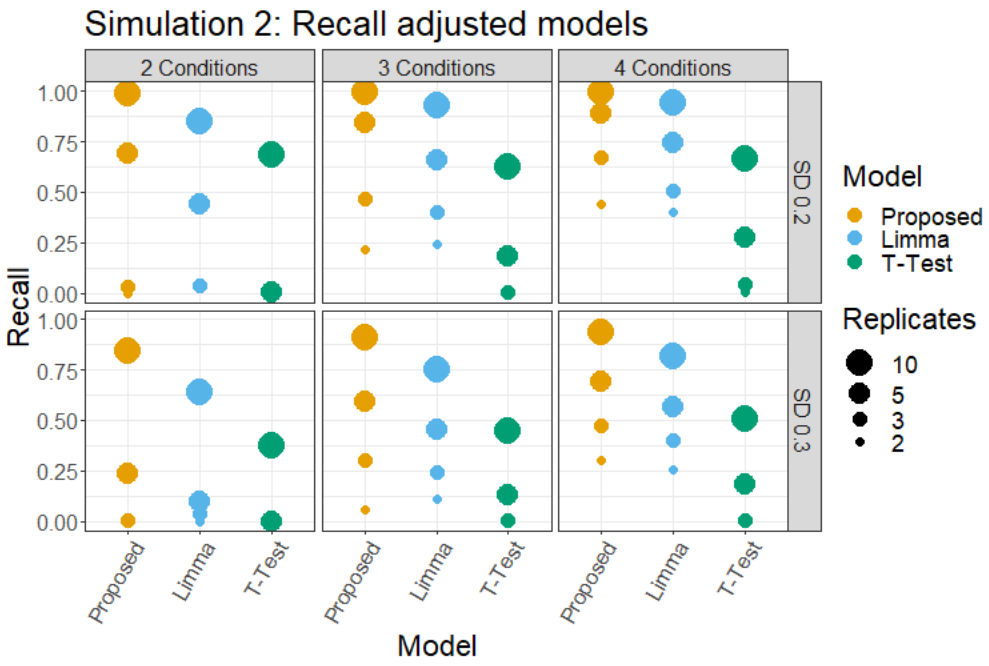

Figure S4

Figure S5: Recall results of Simulation 2. The advantage of using the proposed approach were apparent when looking at simulation 2, which included limited observations and the presence of missing values. In the case of recall the proposed method performed stronger than Limma and t-test in nearly every model. Even at lower replicates the proposed method still outperformed Limma. Again the lowest performing method was *t*-test.

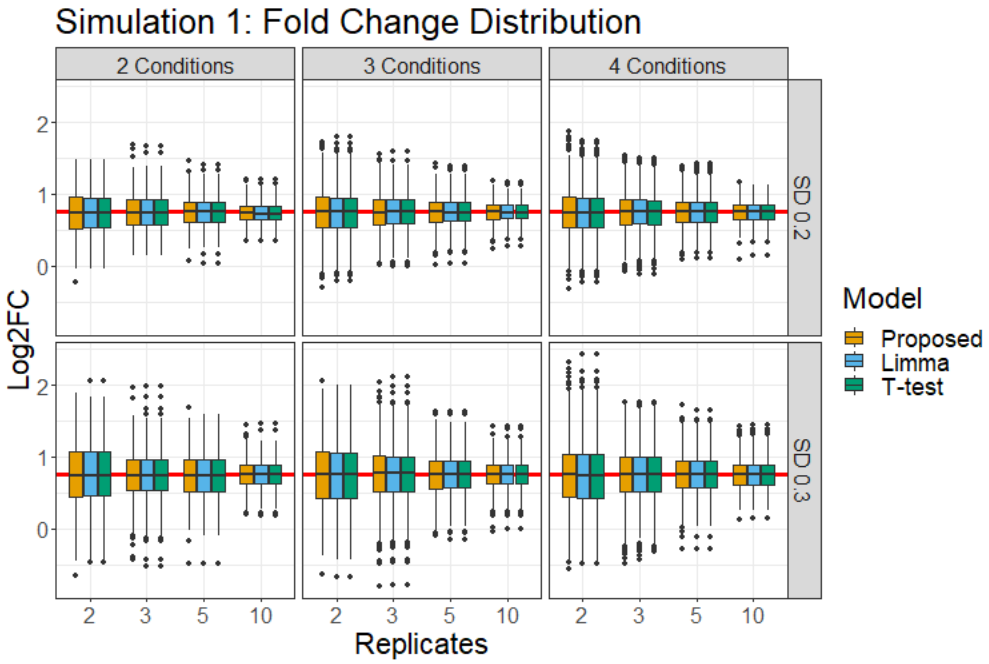

Figure S6.a: In simulation 1 all considered methods correctly estimated the fold change between conditions, with a median fold change estimation of .75. The distributions around the median were also consistent across all methods.

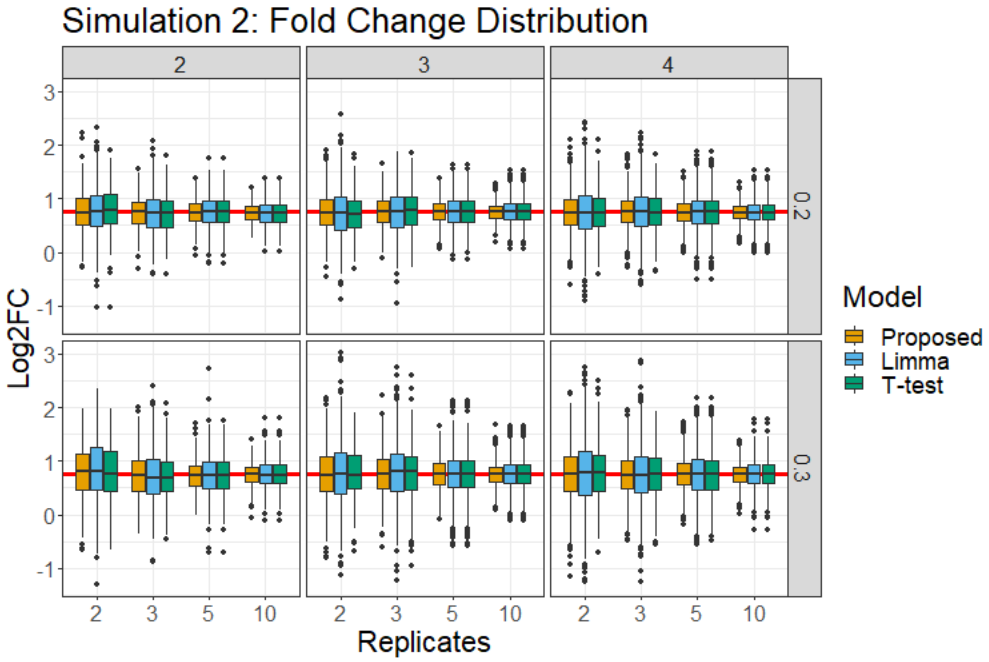

Figure S6.b: In simulation 2 all methods correctly estimated the fold change with a median log change of .75. The proposed method in this simulation had a visibly tighter distribution around the median. Both Limma and *t*-test showed a wider range around the fold change. The inner quartile range of the Proposed method was on average 10.4% smaller than t-test and 21.8% smaller than Limma.

Figure S6: Fold change distribution comparison between Simulations 1 and 2.

2.1.4 Dataset 3 : SpikeIn benchmark - Ubiquitination - Label-free

Again we consider three different methods and assess their performance: the method proposed in this paper, Limma, and two sample t-test. All methods are analyzed after adjusting for changes in overall protein level. The results are summarized from Figure S7 to Figure S10, including volcano plots, model summary statistics, and fold change analysis.

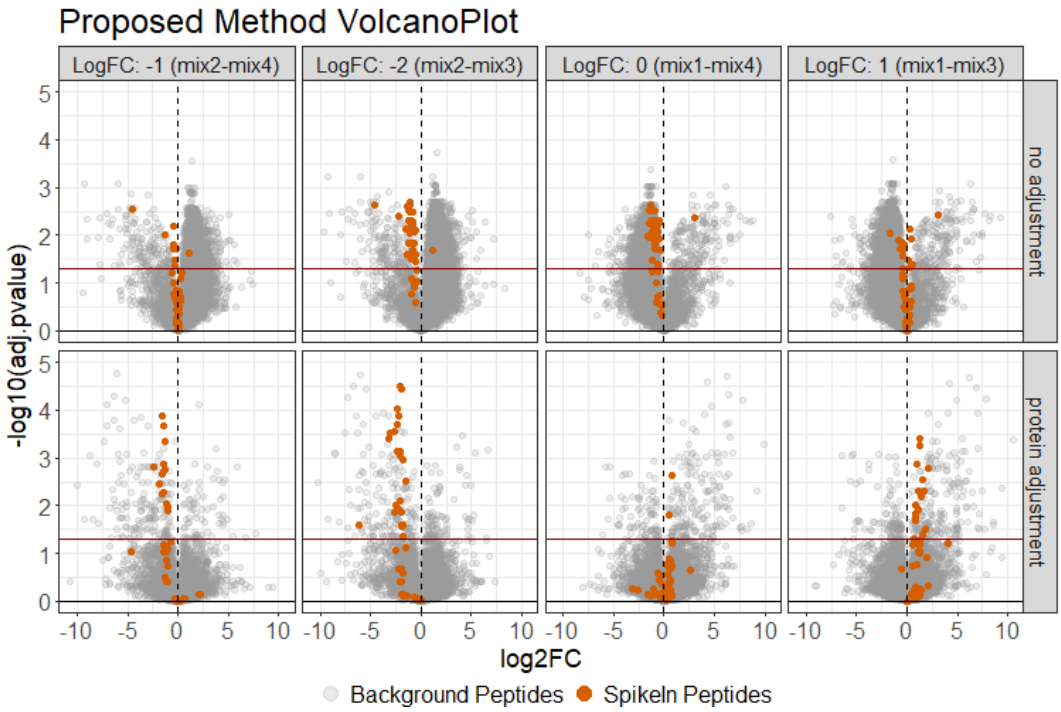

Figure S7: The modeling results of the proposed method both before and after adjustment. The spike-in peptides are colored red and the background peptides are colored grey. All grey peptides are expected to be insignificant. Using the proposed method to model the benchmark experiment, the spike-in peptides (colored red) did not follow the expected log fold change before adjustment. After adjusting for changes in overall protein abundance the spike-in peptides were more in line with expectation. Additionally the background grey colored peptides showed many false positives before adjustment. After adjustment these false positives were decreased considerably.

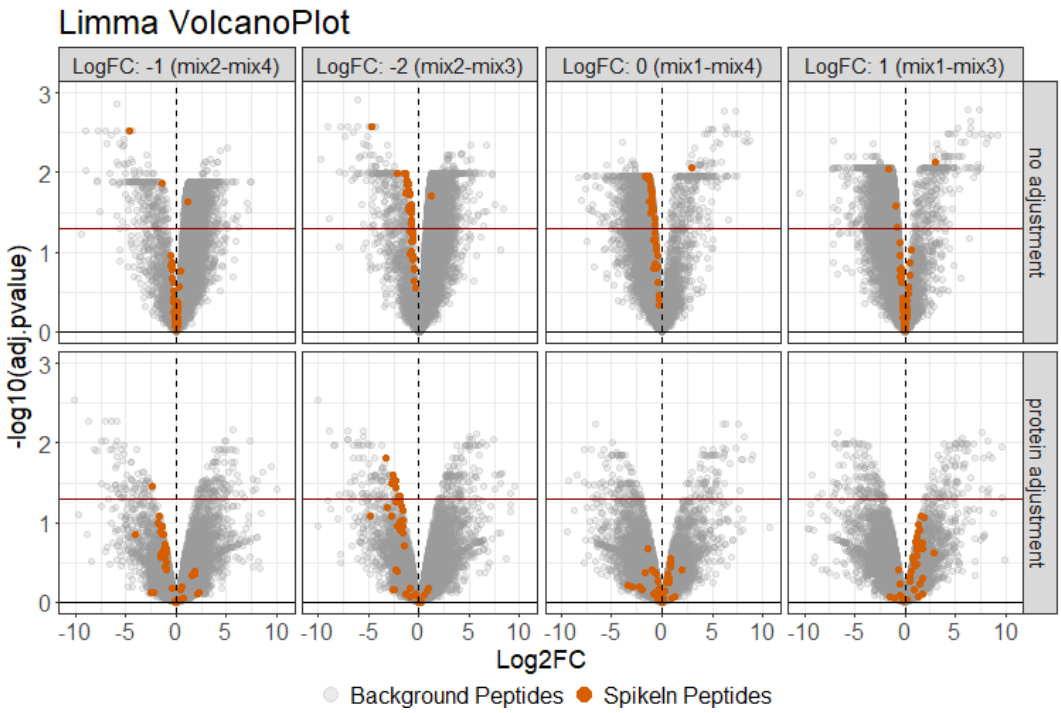

Figure S8: When modeling the experiment with the Limma method, the spike-in peptides again follow the expected log fold change better after adjusting for changes in protein level. However, while the fold change was more accurate, the majority of spike-in peptides did not have a significant adjusted pvalue. In this case, the known differential peptides were missed by the model. In terms of false positives, the results were very similar to the proposed method, with many false positives before adjustment and fewer after.

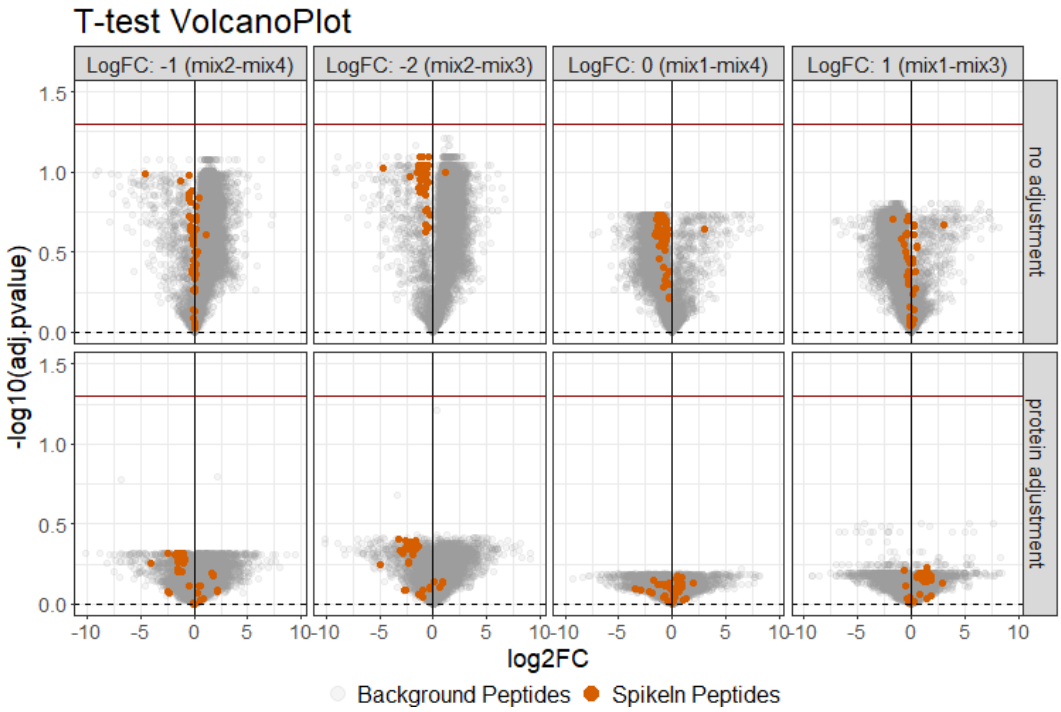

Figure S9: Using the two sample t-test, none of the comparisons either before or after adjustment show any significant peptides. With that being said, the fold change of the spike-in peptides was much closer to expectation after adjusting for global protein abundance.

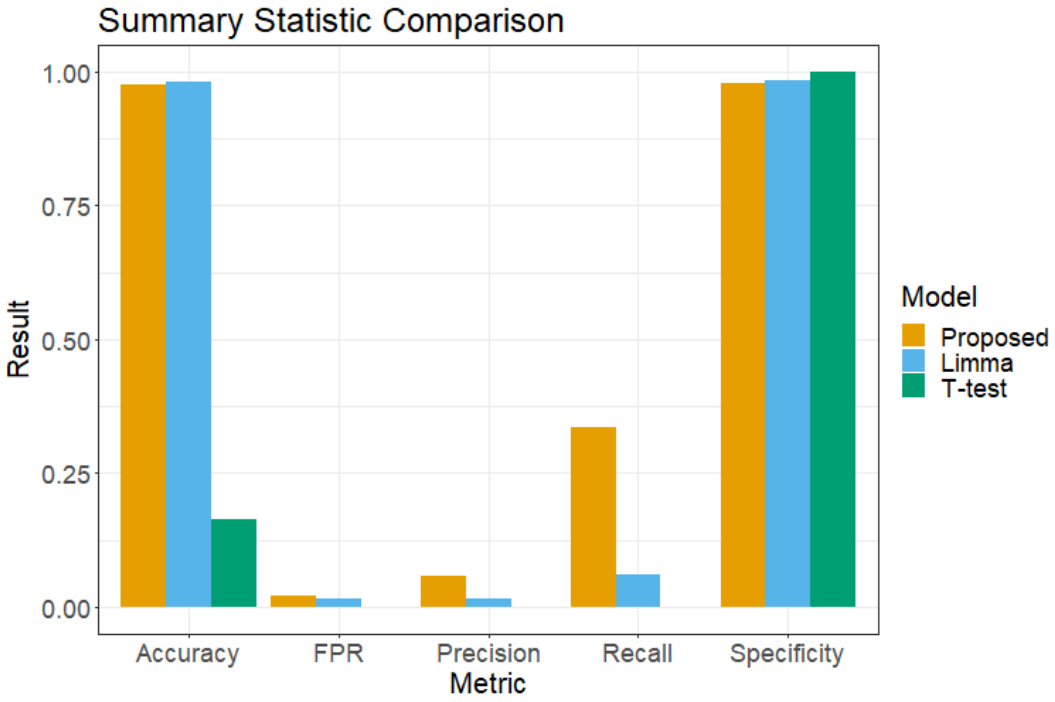

Figure S10: Comparing the summary statistics between methods, the proposed method performed the strongest. In terms of accuracy and specificity the three methods were close, with Limma and *t*-test showing slightly higher values. Accuracy and specificity were dominated by the large number of true negatives (background peptides) compared to the true positives (spike-in peptides). In terms of recall, the proposed approach out performed the other two methods, showing that it correctly labeled the most spike-in peptides.

2.2 Biological investigations

2.2.1 Dataset 4 : Human - Ubiquitination - 1mix-TMT

In this experiment Shigella ubiquitin ligase IpaH7.8 was shown to inhibit the protein gasdermin D (GSDMD) [1]. 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 biological replicates per condition for all conditions except for infected 1 hour. The experiment was modeled as a group comparison, with different replicates measured at each time point. The experimental design can be seen in Table S1.

Condition	BioReplicate	Channel
Dox1hr	Dox1hr_1	127C
Dox2hr	Dox2hr_1	128N
Dox2hr	Dox2hr_2	130C
Dox4hr	Dox4hr_1	128C
Dox4hr	Dox4hr_2	131C
Dox6hr	Dox6hr_1	129N
Dox6hr	Dox6hr_2	131N
NoDox0hr	NoDox0hr_1	126C
NoDox0hr	NoDox0hr_2	129C
NoDox6hr	NoDox6hr_1	127N
NoDox6hr	NoDox6hr_2	130N

Table S1: The experimental design of Dataset 4

A model was fit for the total protein and ubiquitination separately. The model formula can be seen below.

$$Y_{mcb} = \mu + Condition_c + Subject_{mcb} + \epsilon_{mcb}$$

$$\sum_{c=1}^{C} Condition_c = 0,\ Subject_{mcb}\ \sim N(0, \sigma_S^2),\ \epsilon_{mcb}\ \sim N(0, \sigma^2)$$

The results of the proposed method to this experiment can be seen in Figure S11.

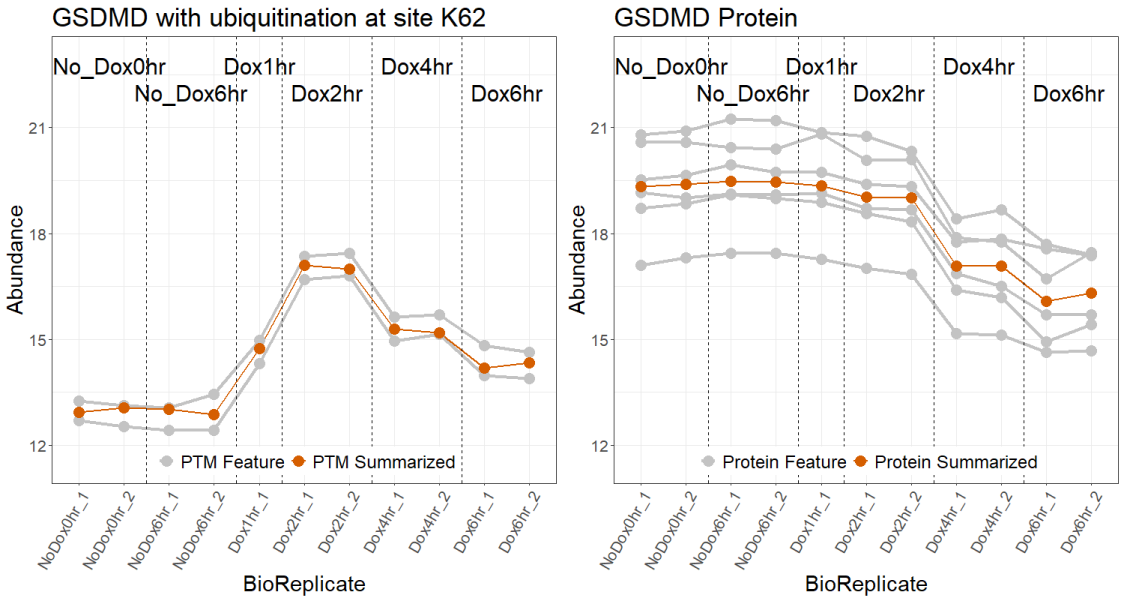

Figure S11.a: Here the global profiling of protein $GSDMD\_HUMAN|P57764$ with the ubiquitination of the protein at site $K62$ was compared. The individual PSM features areshown in grey, while the feature summarization is shown in red. The summary of the modification and global protein showed that the conditions followed different trends. Specifically, there appeared to be no change in abundance between Dox1hr and Dox4hr in the modified plot, however there was a large negative change in the unmodified plot. This indicates the modification was confounded with changes in the unmodified protein.

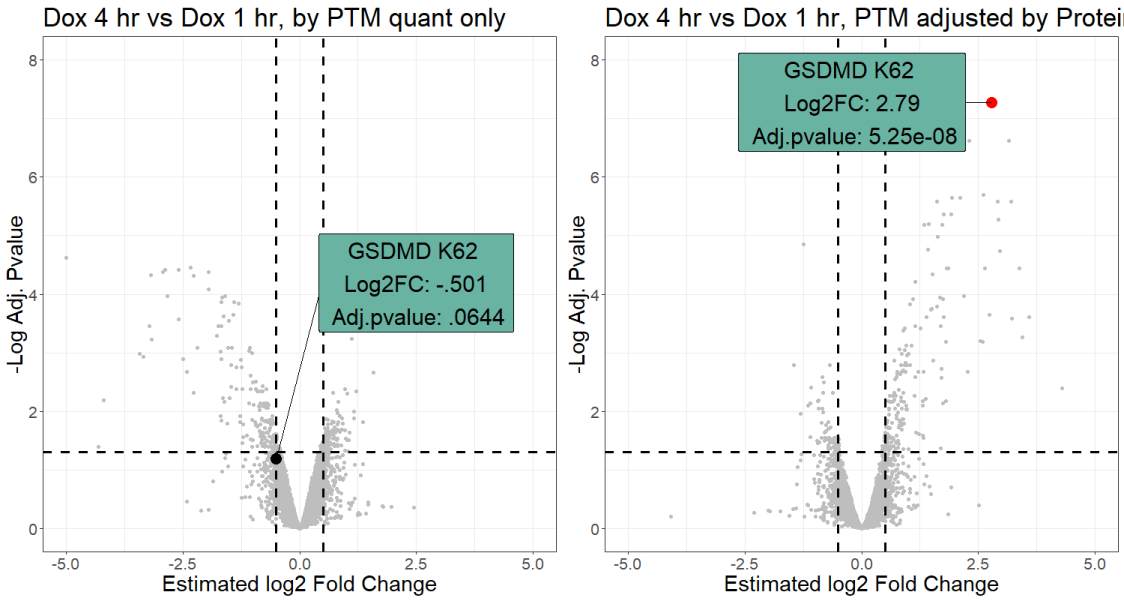

Figure S11.b: Volcano plots of Dox4hr vs Dox1hr both before and after protein adjustment. The $GSDMD\_HUMAN|P57764\_K62$ modification is highlighted. Before adjustment the modification had a small fold change and insignificant adjusted pvalue. After adjustment the fold change was much larger and the adjusted pvalue was significant. In this case the proposed method allowed us to identify a differential modified peptide that could have otherwise been missed.

Figure S11: Summary plots for modification of protein GSDMD at site K62.

2.2.2 Dataset 5 : Mouse - Phosphorylation - 2mix-TMT

In this study, the correlation between the gene Atg16L1 and killing of Shigella flexneri (S.flexneri) was assessed [2]. Multiplex proteomics was used to quantify the abundance of total protein, phosphorylation, and ubiquitination 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 an unbalanced experimental design. 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. The experimental design can be seen in Table S2.

	Mixture 1		Mixture 2		Condition
Uninfected	128C		128C	131C	
Early (1 Hour)	126C	129C	126C	129C	WT
Late (3 Hour)	127C	130C	127C	130C	
Uninfected	129N	131C	129N		
Early (1 Hour)	127N	130N	127N	130N	KO
Late (3 Hour)	128N	131N	128N	131N	

Table S2: The experimental design of Dataset 5

A model was fit for the total protein, phosphorylation, and ubiquitination separately, as described previously for TMT experiments. The model formula can be seen below.

$$Y_{mcb} = \mu + Mixture_m + Condition_c + Subject_{mcb} + \epsilon_{mcb}$$

$$Mixture_m \sim N(0, \sigma_M^2),\ \sum_{c=1}^{C} Condition_c = 0,\ Subject_{mcb} \sim N(0, \sigma_S^2),\ \epsilon_{mcb} \sim N(0, \sigma^2)$$

The results of the proposed method to this experiment can be seen in Figure S12 and Figure S13.

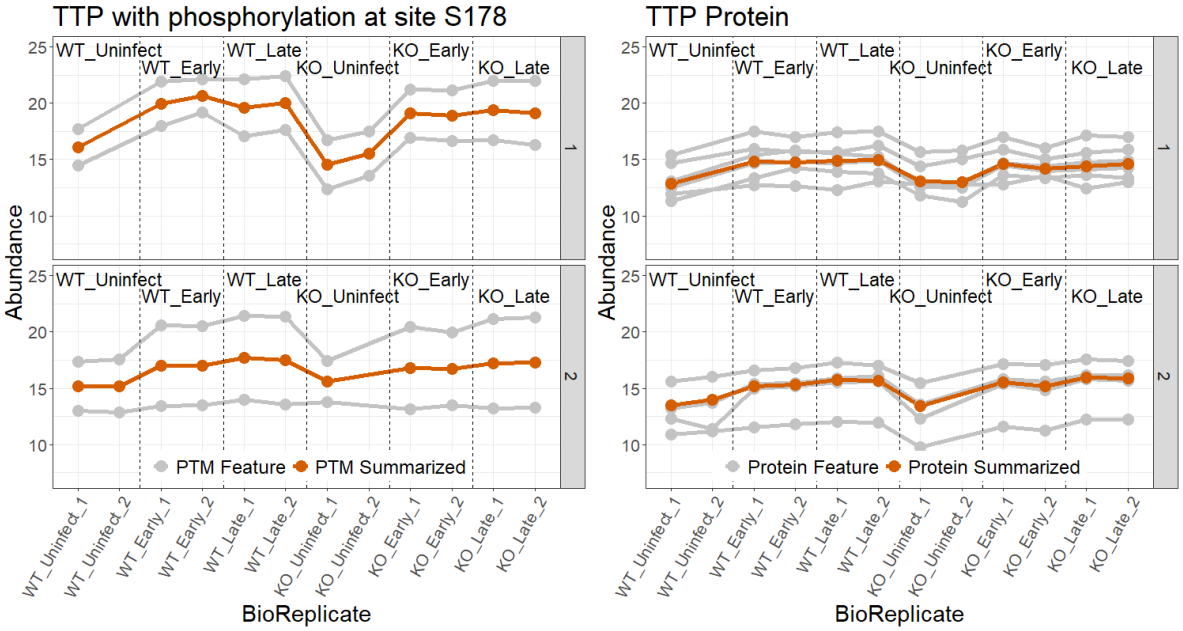

Figure S12.a: The global profiling of protein $TTP\_MOUSE|P22893$ with the modification of the protein at site $S178$ was compared. The individual PSM features are shown in grey, while the feature summarization is shown in red. The summary of the modification and global protein showed that the conditions followed the same trend. Specifically, there was a positive adjustment in abundance when comparing WT_Uninfect to WT_Late in both the modification and global profiling run. This indicated the movement is driven by changes in global protein.

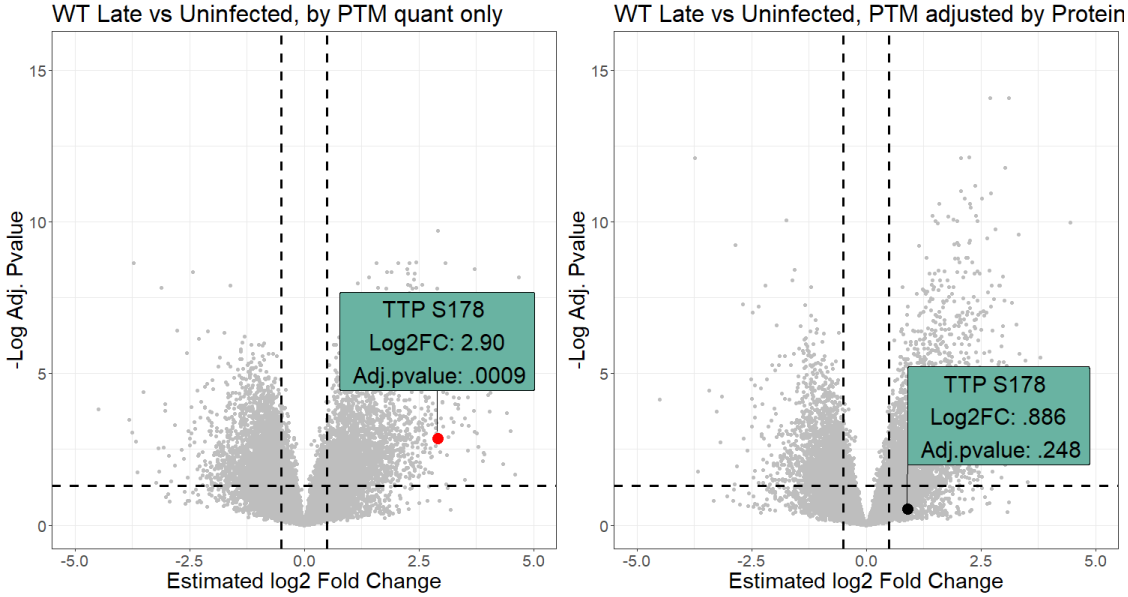

Figure S12.b: Volcano plots of WT_Late vs WT_Uninfect both before and after protein adjustment. The $TTP\_MOUSE|P22893\_S178$ modification is highlighted. Before adjustment the modification had a large fold change and significant adjusted pvalue. After adjustment the fold change was much smaller and the adjusted pvalue was insignificant.

Figure S12: Summary plots for modification of protein TTP at site S178.

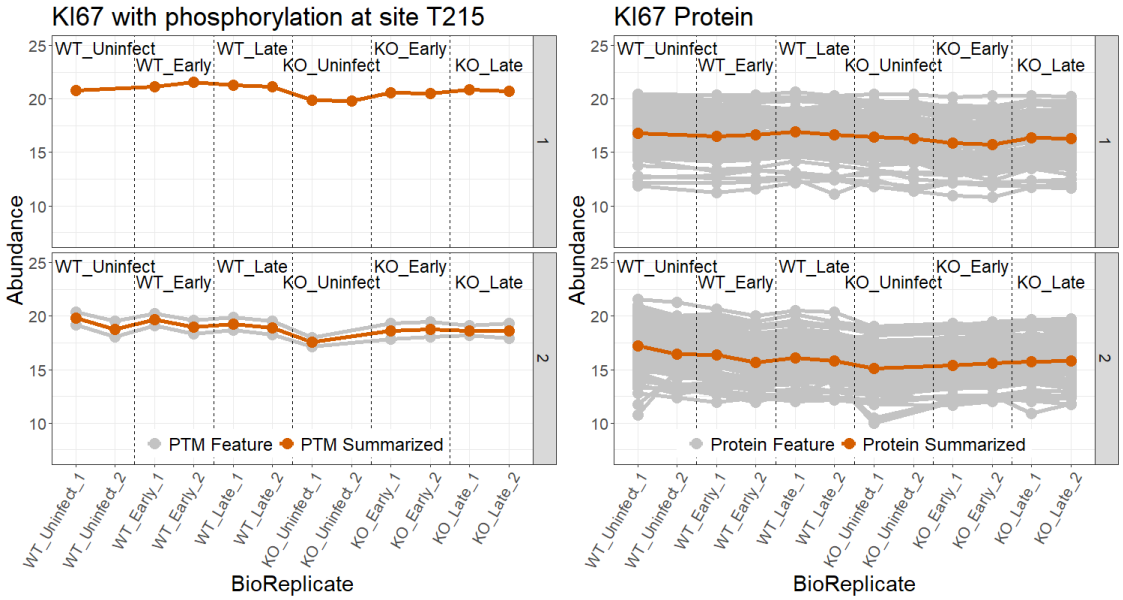

Figure S13.a: The global profiling of protein $KI67\_MOUSE|E9PVX6$ with the modification at site $T215$ was compared. In this case the modification and global protein appeared to show small or no difference between conditions, however after adjusting for change in global protein abundance, the modification was statistically significant. Additionally, this profile plot showed the large difference in available features between modified peptides and global proteins.

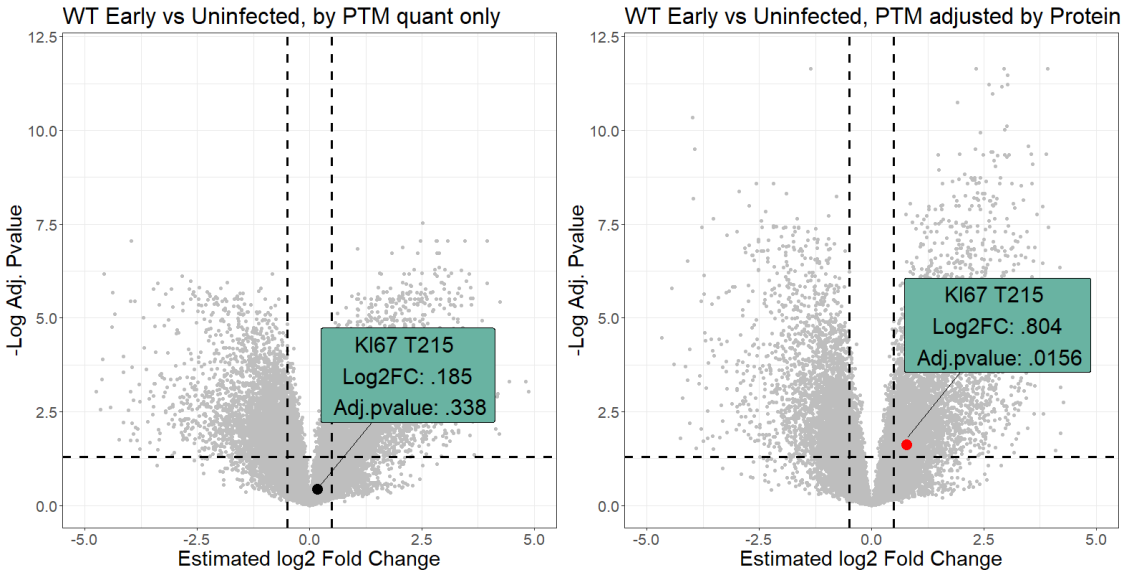

Figure S13.b: The volcano plot of the WT_Uninfect and WT_Early comparison showed the specifics of the adjustment. The modification looked to be flat, with a log fold change of .185, while the global profiling showed a small negative fold change of $-.619$. While both exhibit small changes, when combined we saw a log fold change of .804 and adjusted p-value of .0156.

Figure S13: Summary plots for modification of protein KI67 at site T215.

2.2.3 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 overexpression (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 this supplementary. The experiment was modeled as a group comparison.

In contrast to other experiments analyzed in this paper, there was no unmodified global protein 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 and low feature counts for the unmodified protein model.

An example profile plot for this experiment can be seen in Figure S14.

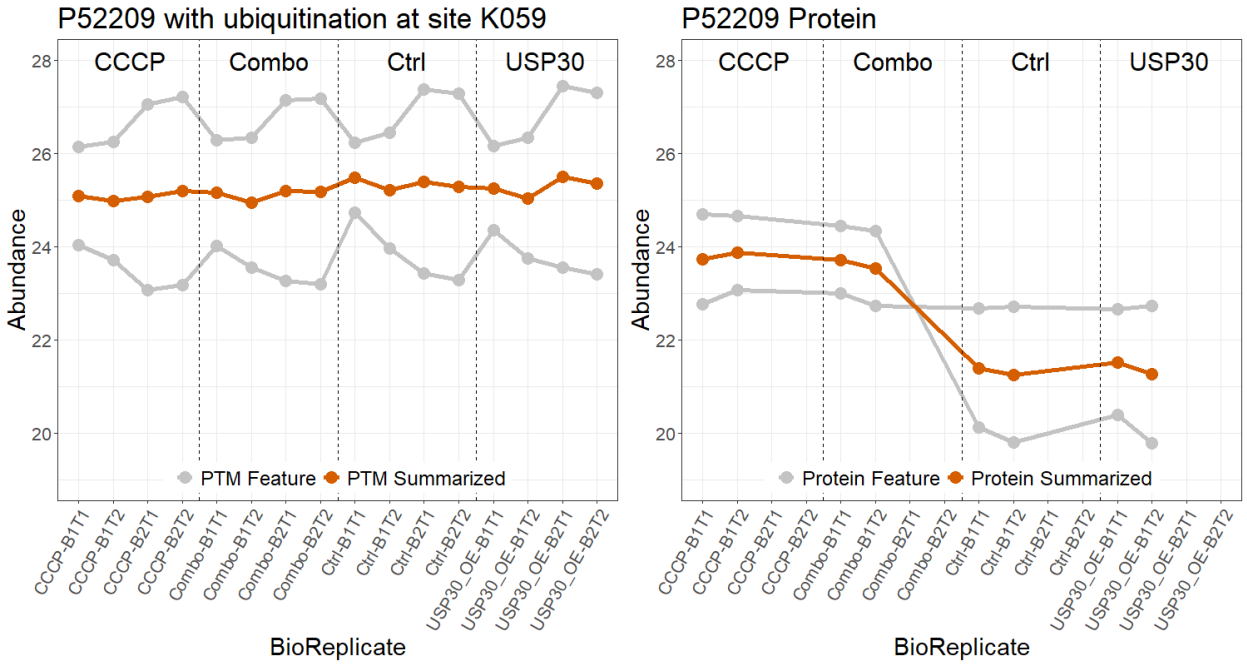

Figure S14: The global profiling of protein $P52209$ with the modification of the protein at site $K059$ was compared. The modification appeared generally unchanged between all conditions, whereas the global profiling run showed the CCCP and Combo conditions had a higher relative abundance compared to the Control and USP30_OE. This indicated that the modification actually had an effect when comparing CCCP and Combo to Control and USP30_OE. However it was not entirely clear as one unmodified peptide feature appeared to be changed, while the other did not. This uncertainty was another result of not running a separate global profiling run. With a global profiling run, many unmodified features are generally quantified, removing the uncertainty that comes with low feature counts.

3 Sample size calculation and power analysis

Noisy PTM measurements benefited from additional biological replicates

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

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

The results of the power and sample size analysis can be seen in Figure S15. When the variance and replicates were equal, higher replicates predictably lead to higher power. In cases where the replicates were unbalanced, but the variance was still the same, it did not matter if there were more replicates in the modified or unmodified runs. In comparison, with differing variance and equal replicates, higher replicates still lead to higher power. When the replicates were unbalanced and the variance was higher in the PTM, there was more power when more replicates were allocated to the PTM run than the unmodified protein run. The results lead to the takeaway that in cases where the number of replicates have to be limited, it is better to allocate more to the PTM run.

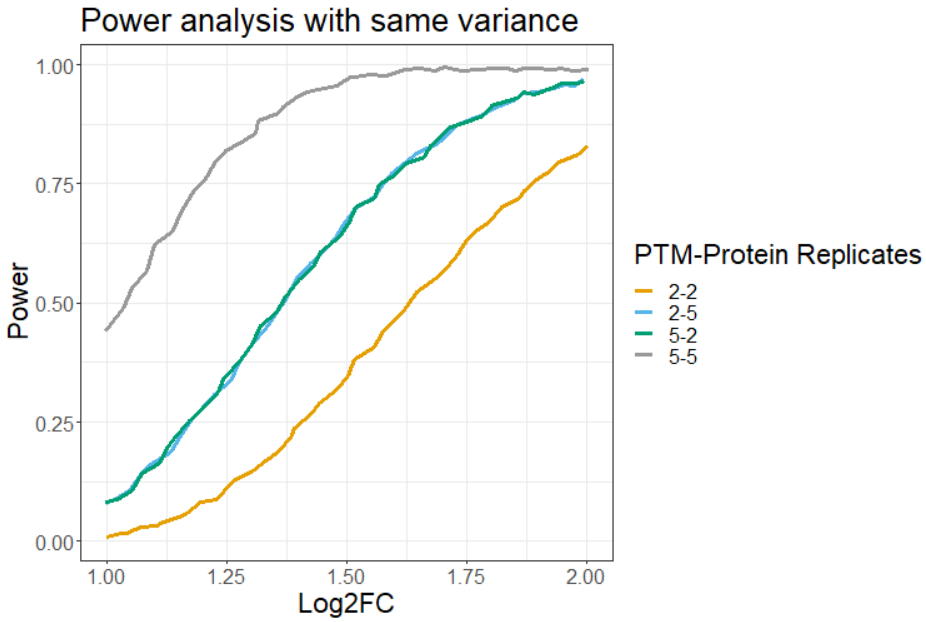

Figure S15.a: The power of an experiment targeting PTMs with the same variance, .15, for the modified and unmodified peptides. Predictably when the replicates were high for both modified and unmodified peptides the power was much higher. Conversely at low replicates for each the power was much lower. With equal variance, it did not matter if the PTM replicates or protein replicates were higher.

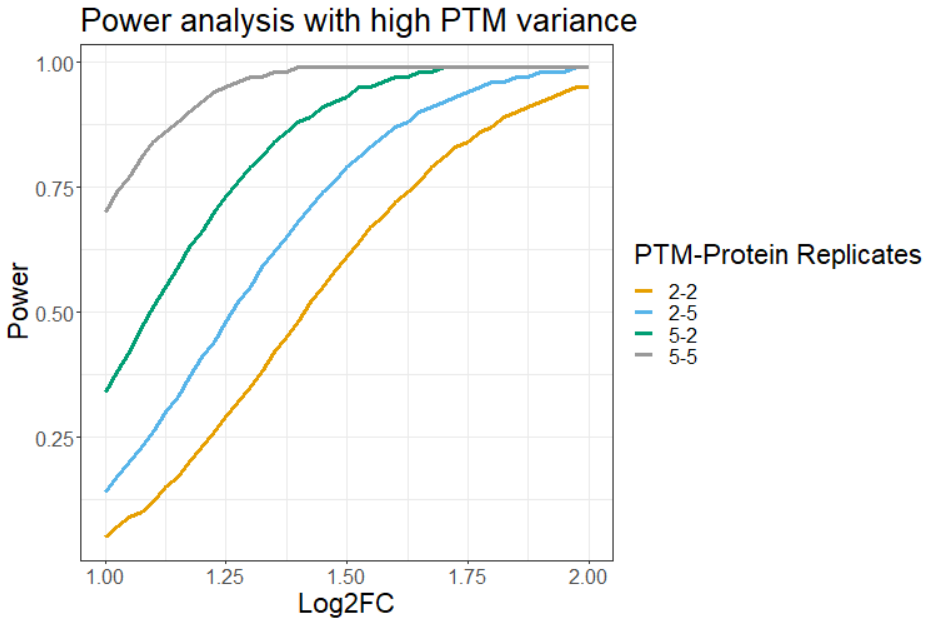

Figure S15.b: In this chart the variance for the PTM was higher than the unmodified protein. The PTM variance was .2, while the unmodified protein variance was .1. With equal replicates the results were the same as above. When the replicates were not equal, more replicates allocated to the PTM runs lead to higher power.

Figure S15: Power analysis of experiments with differing variances.