Proteoform Suite: Software for Constructing, Quantifying, and Visualizing Proteoform Families

Supporting Information

Anthony J. Cesnik[1], Michael R. Shortreed[1], Leah V. Schaffer[1], Rachel A. Knoener[1], Brian L. Frey[1], Mark Scalf[1], Stefan K. Solntsev[1], Yunxiang Dai[1], Audrey P. Gasch[2,3], and Lloyd M. Smith[1,3,*]

[1]Department of Chemistry, University of Wisconsin-Madison, Madison, WI, USA
[2]Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI, USA
[3]Genome Center of Wisconsin, University of Wisconsin-Madison, Madison, WI, USA
[*]Corresponding author. Tel: +1 608 263 2594. E-mail: smith@chem.wisc.edu.

November 21, 2017

Contents

Supporting Figures

Supporting Tables

All results listed here are tabulated in a supporting Excel file, unless otherwise noted.

- Table S1. Identified experimental proteoforms.
- Table S2. Quantification results.
- Table S3. Raw NeuCode pairs.
- Table S4. Experimental proteoforms.
- Table S5. Theoretical proteoform database.
- Table S6. Exp-Theo peak mass differences.
- Table S7. Exp-Exp peak mass differences.
- Table S8. Proteoform families and orphans.
- Table S9. Quantification method validation using known 2-to-1 mixture.
- Table S10. Balanced permutations used in the original SAM analysis (located on page S-39).
- Table S11. GO analysis results - quantified proteoform background.
- Table S12. GO analysis results - bottom-up detected background.

1 Additional Details on Yeast Proteoform Families

Visualizing proteoform families is an elegant way to represent these results, simplifying many columns of data into a single beautiful graphic. In addition, it allows us to represent what we know about unidentified proteoforms, including mass differences representing known sets of PTMs or amino acid losses in relation to other unidentified proteoforms.

This section covers the information contained in the supporting Cytoscape file, which may be downloaded from `https://github.com/smith-chem-wisc/ProteoformSuite/releases/download/0.2.7/ProteoformFamilies.cys`, and then provides additional details on proteoform families with significant changes in yeast salt stress response.

1.1 Cytoscape Diagrams

The proteoform families with significant changes are shown in the remainder of Section 1, while the many remaining families can be found in the supporting Cytoscape file, which contains 3 network collections: Identification Results, Quantification Results, and Decoy Proteoform Communities.

1.1.1 Identification Results

The first contains the results of constructing proteoform families and identifying proteoforms; it includes unidentified families in the display, but does not include orphans (experimental proteoforms without connections to other proteoforms). The size of the circles in this collection represents the integrated intensity of NeuCode pairs used to construct proteoform families. This type of diagram can be generated without quantitative data.

1.1.2 Quantification Results

The second network collection contains the results of quantification. First, all families are presented again, but this time with quantitative data; quantified proteoforms are represented as pie charts, where the blue portion represents the proteoform abundance under normal growth conditions and the yellow portion represents the abundance under salt stress conditions. The size of the node represents the sum of intensities (abundances) observed for both normal and stress conditions. Orange annuli indicate significant changes. The families with significant changes, which are presented below with individual comments, are also shown in this collection.

Quantification results were used to perform GO analysis to offer descriptions of the biological changes observed in the data. The families related to each significant GO term (corrected p-values < 0.05) are shown in separate diagrams.

Figure S1: Color scheme used for quantitative proteoform families. Experimental proteoforms are represented as pie charts with blue and yellow representing the relative proteoform abundance under normal and salt stress growth conditions, respectively. The orange annulus represents a significant change (by permutation analysis or > 4 fold change in 2 of 3 biological replicates). Grey nodes represent experimental proteoforms that may have been identified, but did not pass selection criteria for quantification (observed in 5 of 6 technical replicates in at least 1 condition).

1.1.3 Decoy Proteoform Communities

Finally, the results of constructing decoy families (discussed in "Identifying proteoforms and constructing proteoform families" in the Online Methods section) are shown in the last collection. These provide a picture of the types of connections that lead to false positive identifications. Notably, we observed that Exp-Theo connections lead to the majority of false positives. This is especially true when combinatorial PTM sets of three or larger are allowed on theoretical proteoforms, and so we limited the theoretical database in this work to have combinatorial sets of two PTMs, or sets with three or four PTMs of one kind. This limited false connections, leading to a $\sim$ 2-fold improvement in the proteoform FDR.

1.2 Proteoform Families with Significant Changes in Salt Stress Response

1.2.1 Sec61 beta homolog 1 (*SBH1*)

The proteoform family for SBH1 was highlighted in the main text. The family is shown here including proteoforms involving adducts and an unassigned mass difference. In addition, one additional interesting proteoform that belongs in this family was found when manually checking for other results related to this family. Experimental proteoform E_1473 appears to be an acetylated form of Sbh1 with the loss of the C-terminal phenylalanine, albeit with a monoisotopic error (an artifact from deconvolution). This indicates that there may be clipping of this proteoform, in addition to the loss of the N-terminal serine shown in the proteoform family diagrams.

Other experimental proteoforms may be related to this family. E_2197 may represent observations of acetylated proteoform that were missed while aggregating NeuCode pairs during identification. E_580 also appears to represent the acetylated form, with a sodium tetradecyl sulfate adduct. These experimental proteoforms did not receive any components during quantification, and so they did not bias the quantification of acetyl-Sbh1.

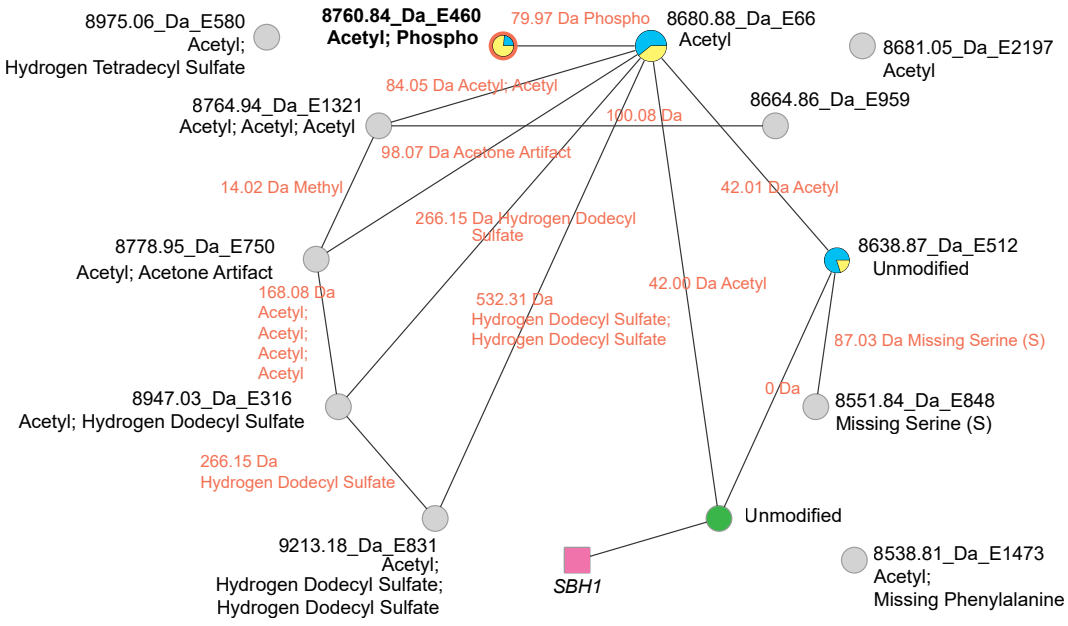

Figure S2: *SBH1* proteoform family with a significant change in the acetylphospho proteoform. Note that adducts and manually identified orphans are included. An unidentified member of the family is also shown, connected by a currently unassigned mass difference. This illustrates what remains to be understood about this family from the results of the experiment, knowledge that can shape future experiments.

1.2.2 Non-histone chromosomal protein 6A (*NHP6A*)

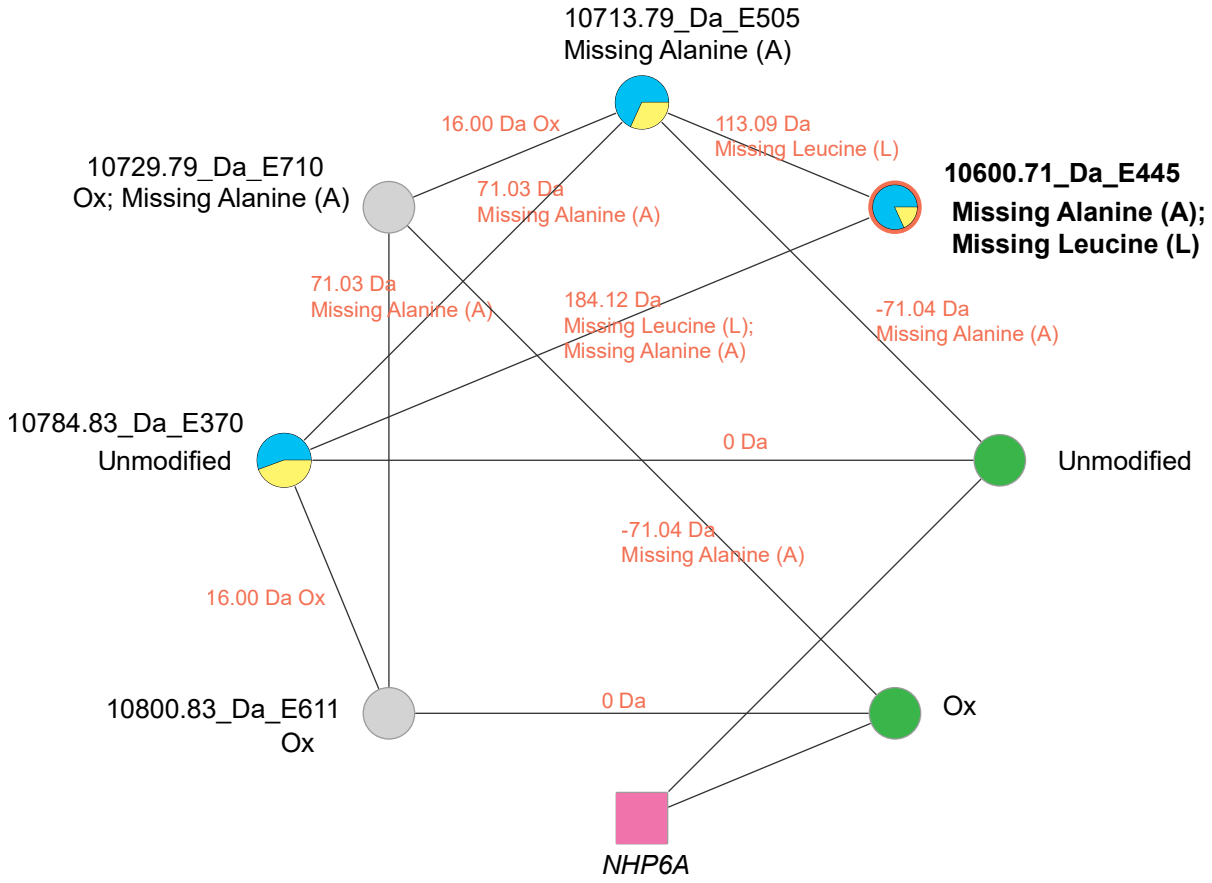

Figure S3: *NHP6A* proteoform family with a significant decrease in C-terminal clipping in response to salt stress. (The amino acid sequence of this protein is MVT...TLA.)

1.2.3 H/ACA ribonucleoprotein complex subunit 3 (*NOP10*)

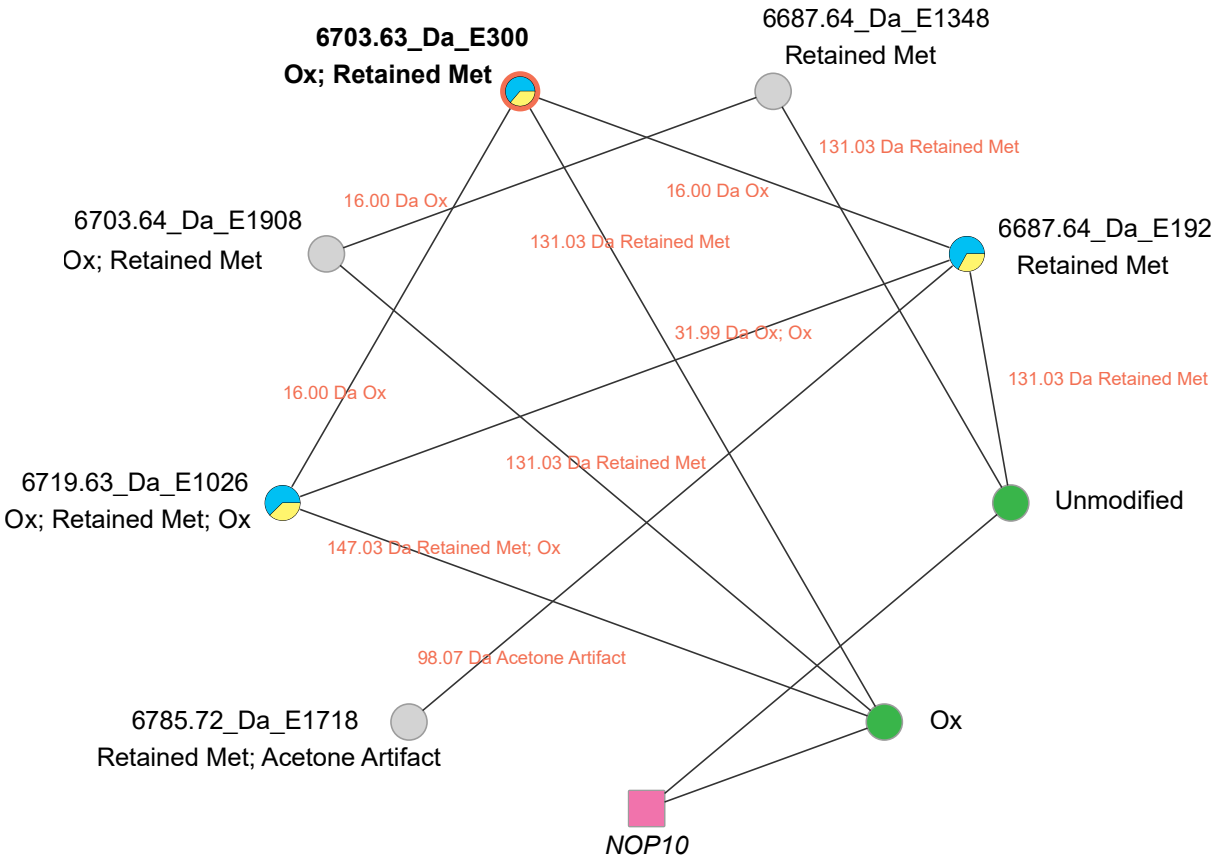

Figure S4: *NOP10* proteoform family with a significant decrease in the proteoform with methionine retention in response to salt stress. This may be the result of an overall decrease in the abundance of this protein, as indicated by the similar decreases in abundance for the other quantified proteoforms.

1.2.4 Restriction of telomere capping protein 3 (*RTC3*)

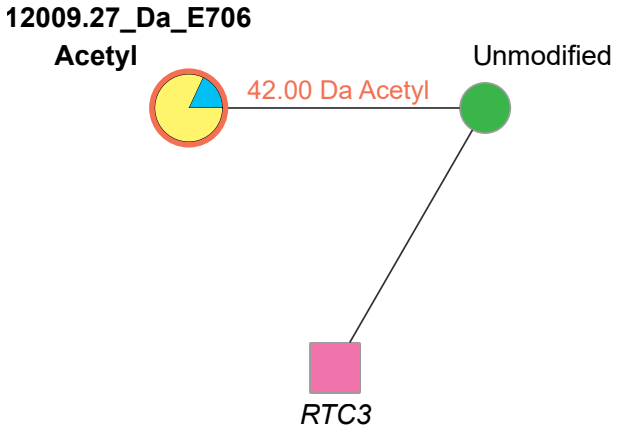

Figure S5: *RTC3* proteoform family with a significant increase of the acetylated proteoform in response to salt stress.

1.2.5 Zeocin resistance protein 1 (*ZEO1*)

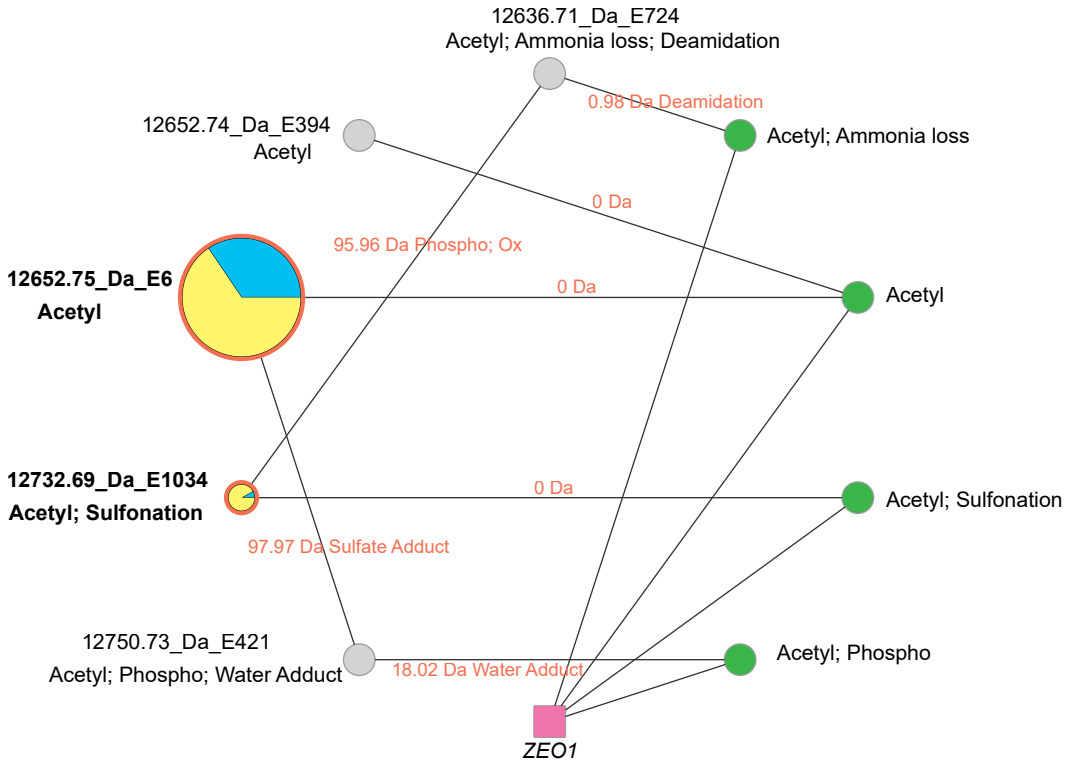

Figure S6: *ZEO1* proteoform family with significant increases in a couple acetylated proteoforms, or perhaps an overall increase in the abundance of the protein.

1.3 Histone Proteoform Families with Significant Changes in Yeast Salt Stress Response

1.3.1 Histone H2A (*HTA2*)

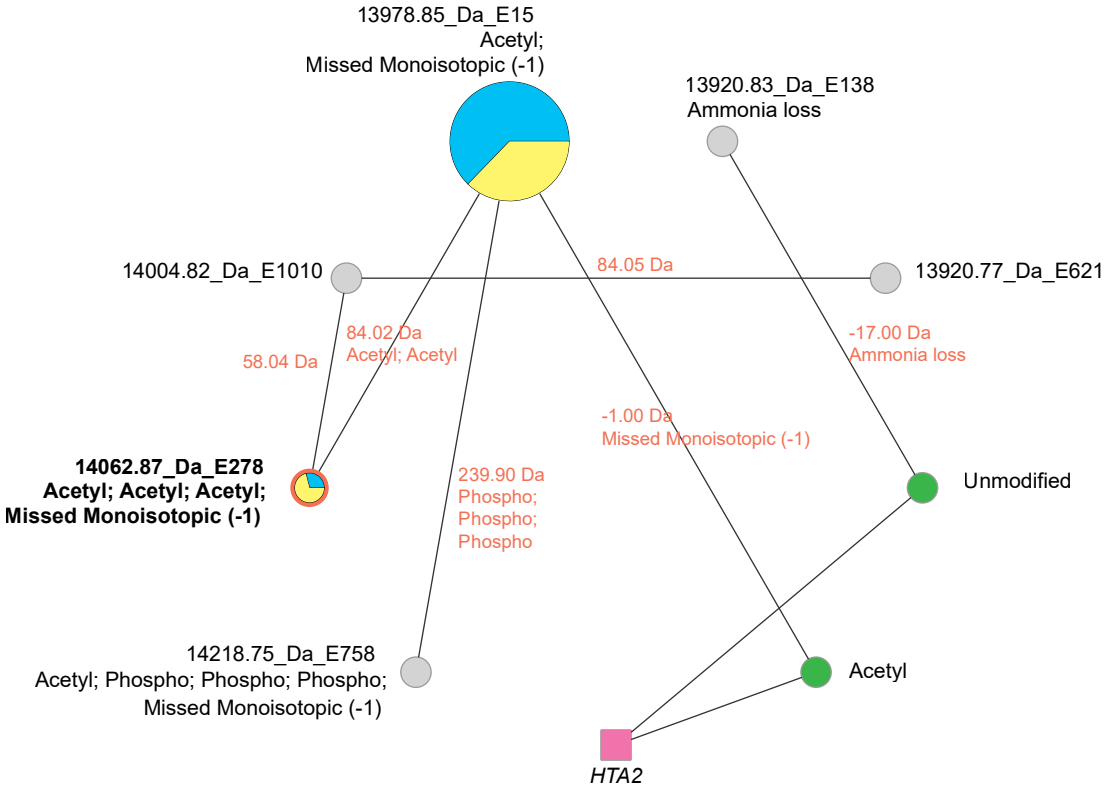

Figure S7: *HTA2* proteoform family with a significant change in the triply acetylated proteoform in response to salt stress. Changes in histone acetylation play a key part in regulating transcription[1]. (Note that "Missed Monoisotopic (-1)" is an artifact of deconvolution.)

1.3.2 Histone H2B (*HTB1*)

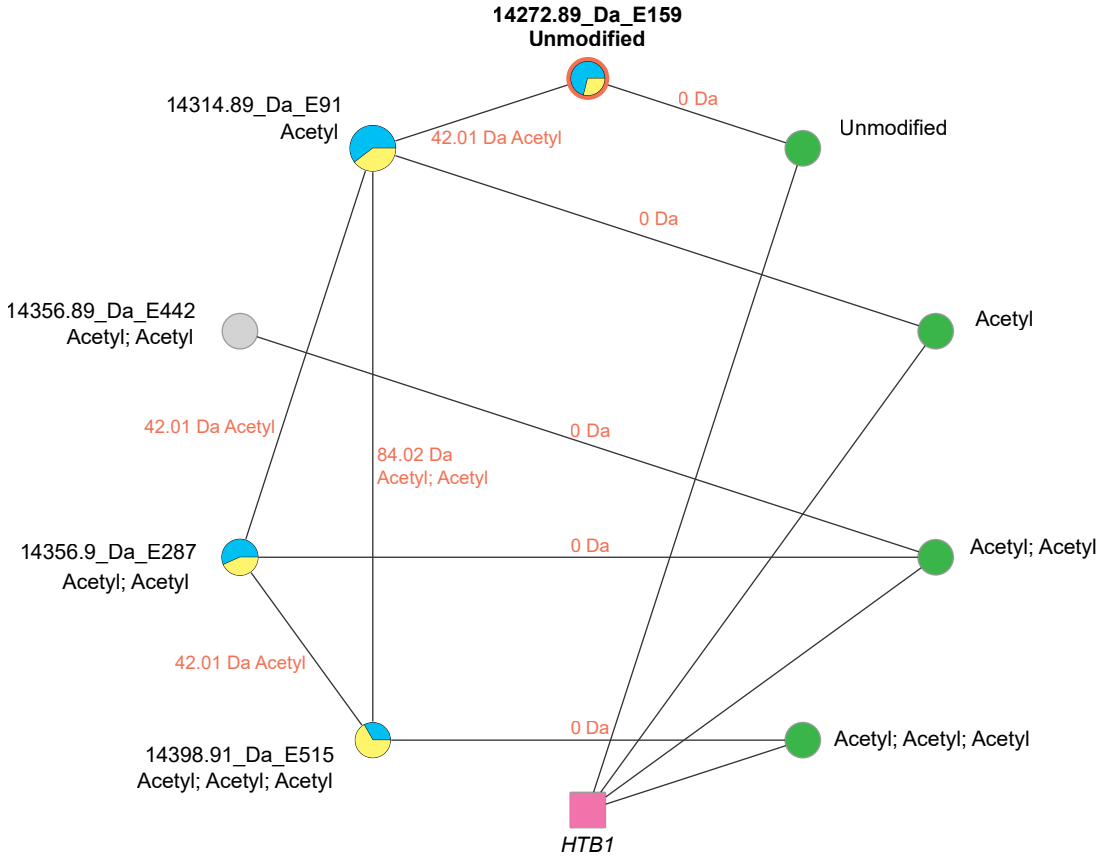

Figure S8: *HTB1* proteoform family with a significant decrease in the unmodified proteoform, which in context of moderate increases in acetylation indicates an increase in acetylation in response to salt stress.

1.3.3 Histone H2B (*HTB2*)

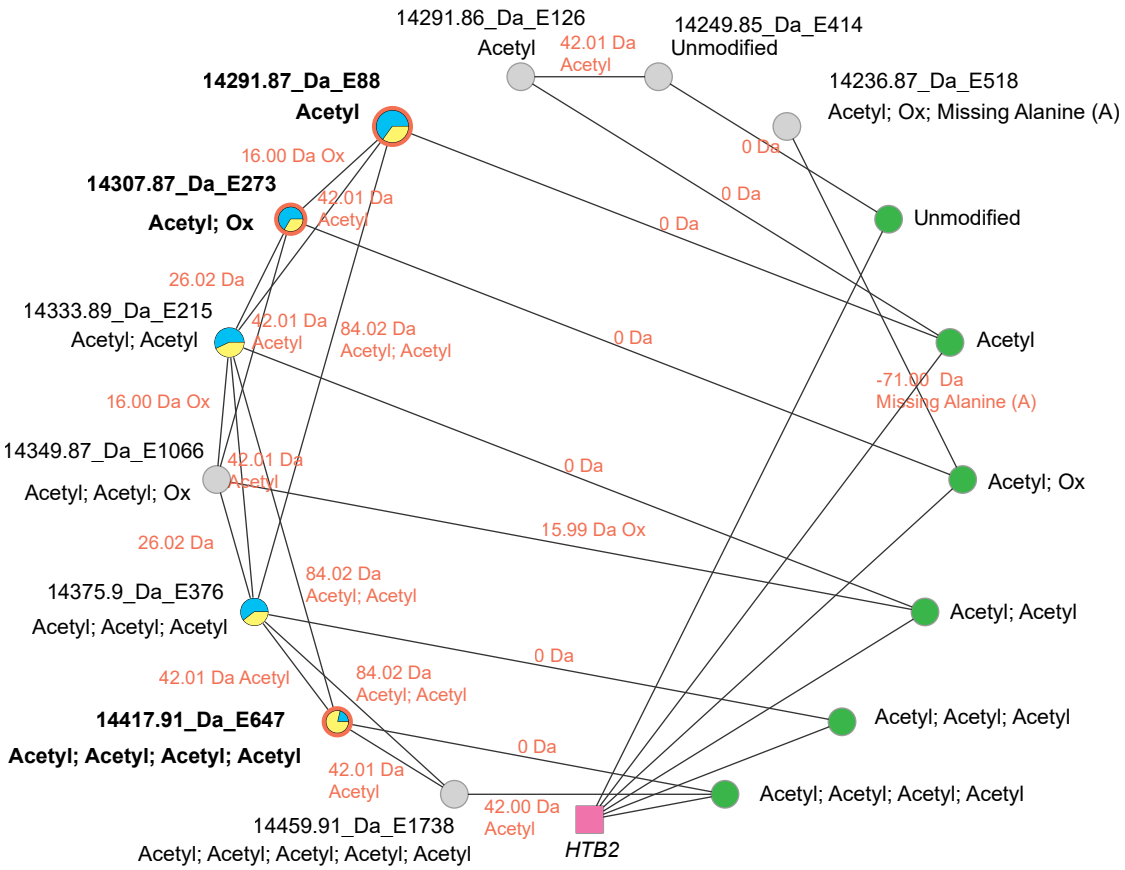

Figure S9: *HTB2* proteoform family with a significant increase in the highly acetylated form and significant decreases for the mono-acetylated form, both in response to salt stress.

1.3.4 Histone H3 (*HHT1*)

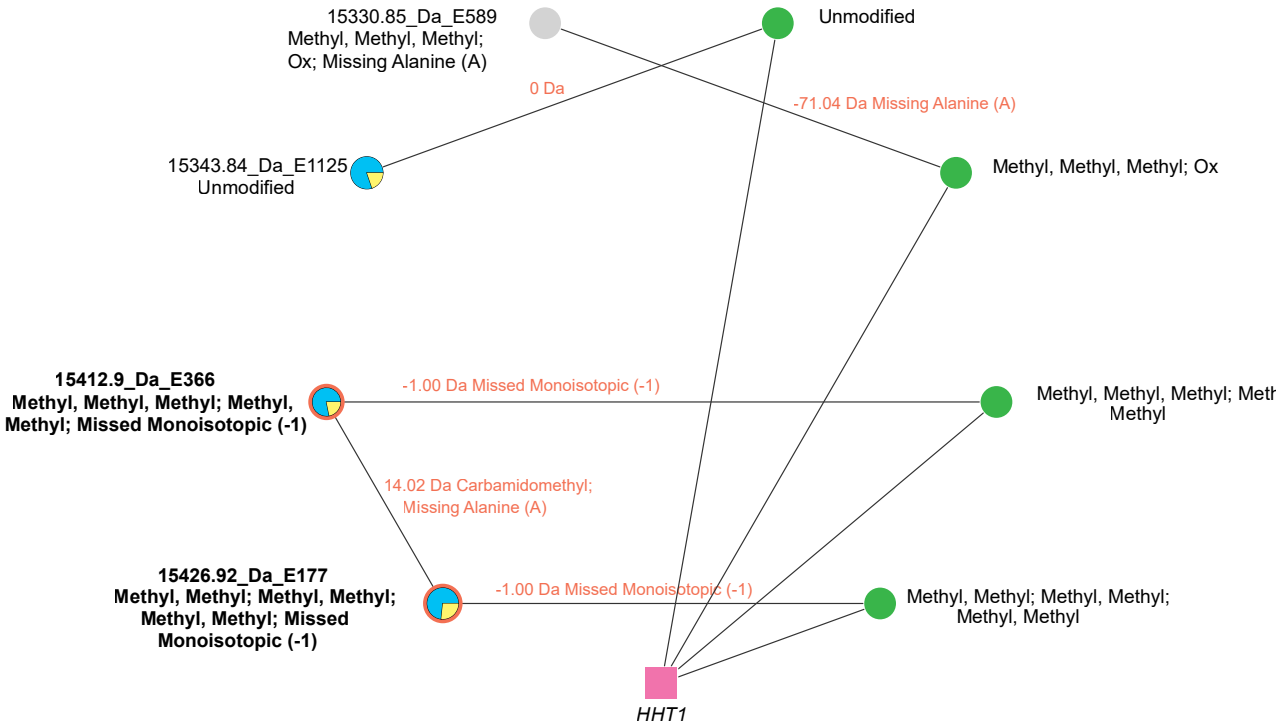

Figure S10: *HHT1* proteoform family with a significant decrease in the abundance of methylated proteoforms in response to stress. It appears these changes may be due to an overall decrease in the abundance of the protein, since the unmodified form has a similar decrease in response to stress. (Note that "Missed Monoisotopic (-1)" is an artifact of deconvolution.)

1.3.5 Histone H4 (*HHF1*)

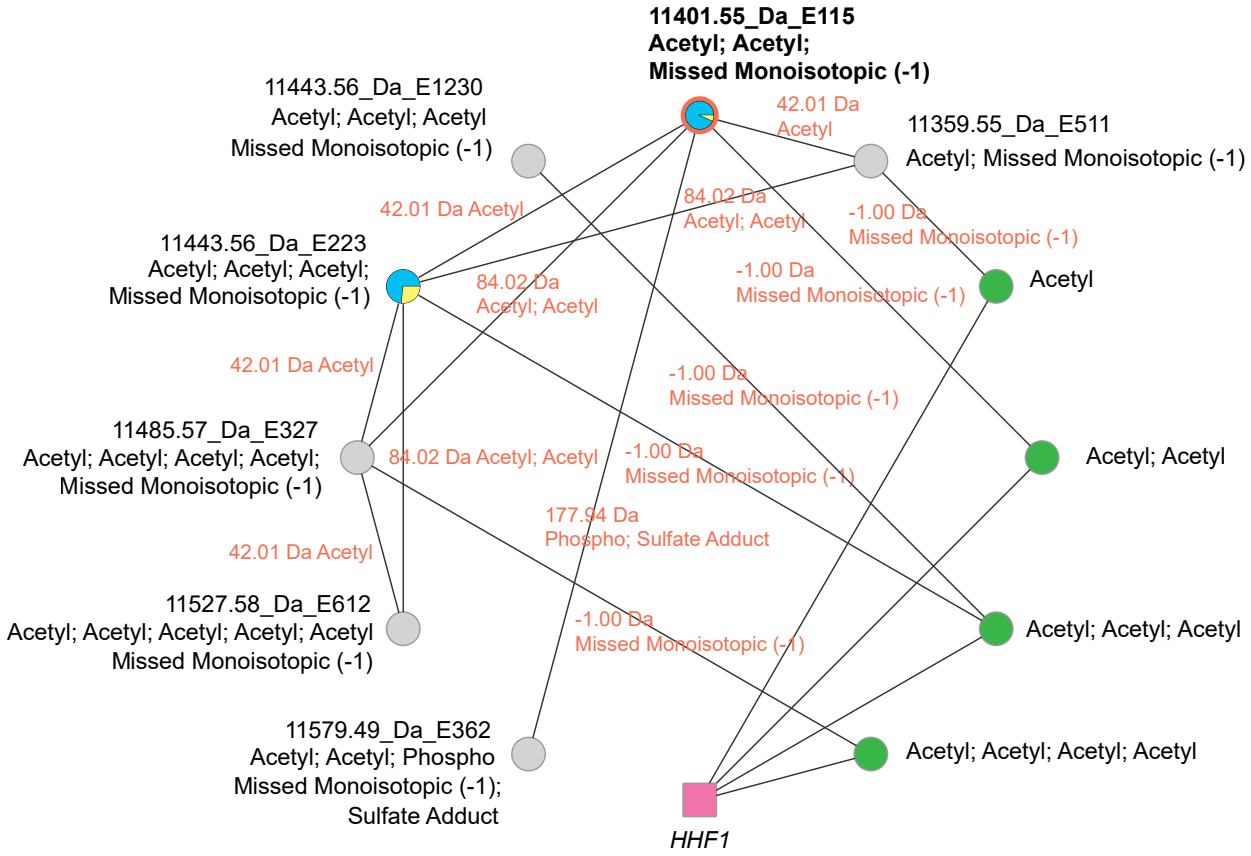

Figure S11: *HHF1* proteoform family with a significant decrease in the abundance of the diacetylated proteoform in response to salt stress. There is extensive acetylation of this proteoform, indicated by the identification of each form with 1 to 5 acetylations, although several of these did not pass the criteria for quantification (in grey). (Note that "Missed Monoisotopic (-1)" is an artifact of deconvolution.)

1.4 Ribosomal Proteoform Families with Significant Changes in Yeast Salt Stress Response

1.4.1 40S ribosomal protein S23-B (*RPS23B*)

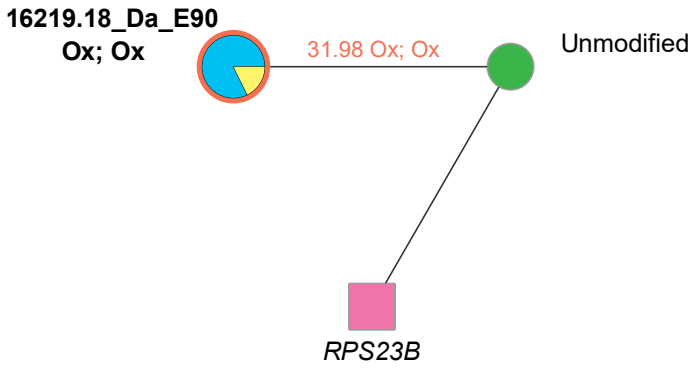

Figure S12: *RPS23B* proteoform family with a significant decrease in the dioxidized proteoforms in response to stress.

1.4.2 40S ribosomal protein S26-A/B (*RPS26A & RPS26B*)

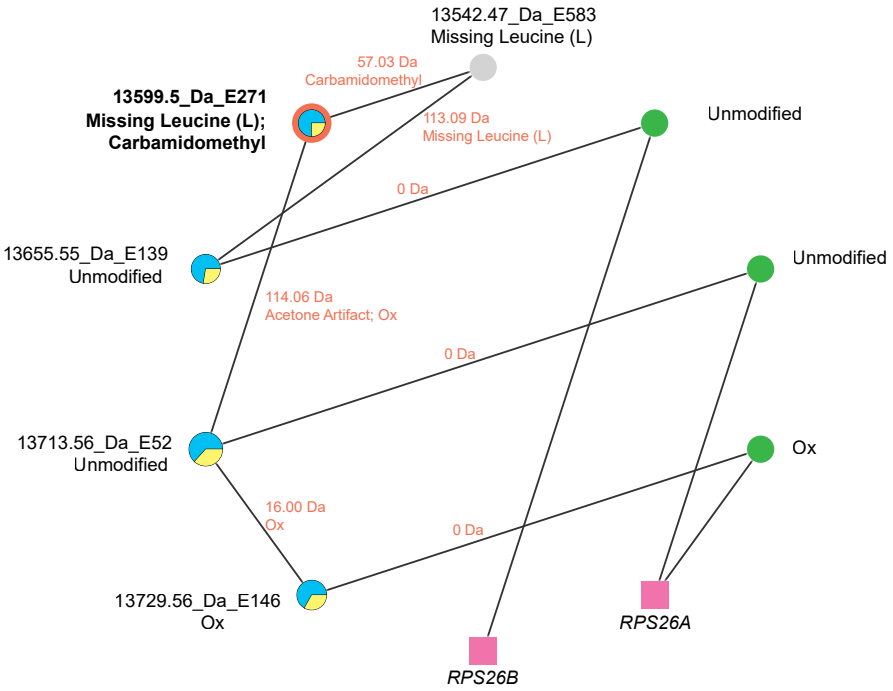

Figure S13: Ambiguous family containing *RPS26A* and *RPS26B* proteoform families. A significant decrease in the abundance of an Rsp26a proteoform missing the C-terminal leucine was observed in response to salt stress. Note that this experimental proteoform was incorrectly assigned the identity of a missing leucine and carbamidomethylation from Rsp26b. The mass difference between unmodified Rps26a and Rps26b proteoforms is 58 Da. Due to a subtle mass error of +1 Da, this appeared more likely to be a carbamidomethylation (+57 Da) to the identification algorithm. The mass error was most likely due to a missed monoisotopic mass during deconvolution; it may also be due to deamidation, but this seems less likely, since it was not observed on the unmodified proteoform.

1.4.3 60S ribosomal protein L9A (*RPL9-A*)

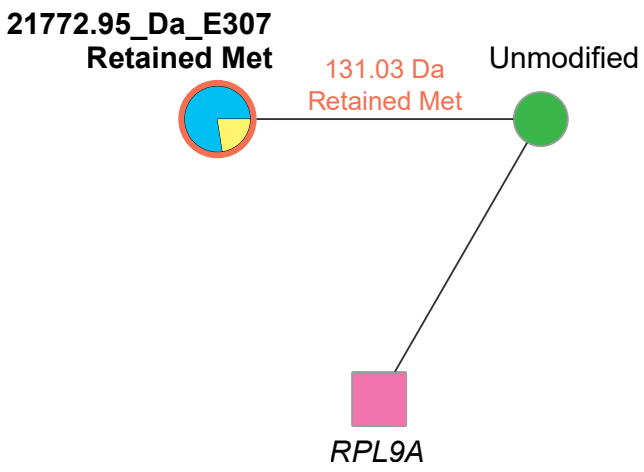

Figure S14: *RPL9A* proteoform family with a significant decrease in the abundance of the proteoform with retained N-terminal methionine in response to salt stress. (See this reference[2] for more on N-terminal methionine cleavage.)

1.4.4 60S ribosomal protein L13-B (*RPL13B*)

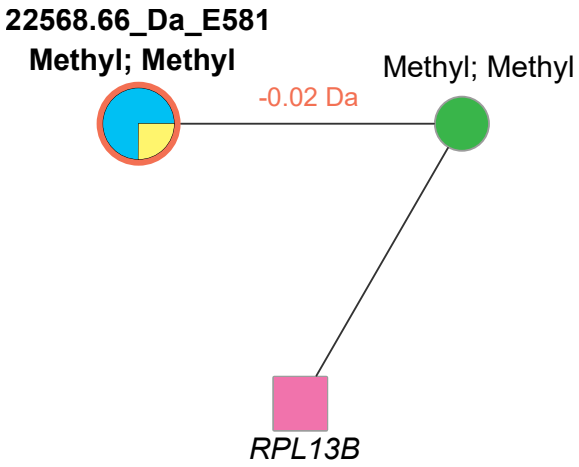

Figure S15: *RPL13B* proteoform family with a significant decrease in the dimethylated proteoform in response to stress.

1.4.5 60S ribosomal protein L14-A (*RPL14A*)

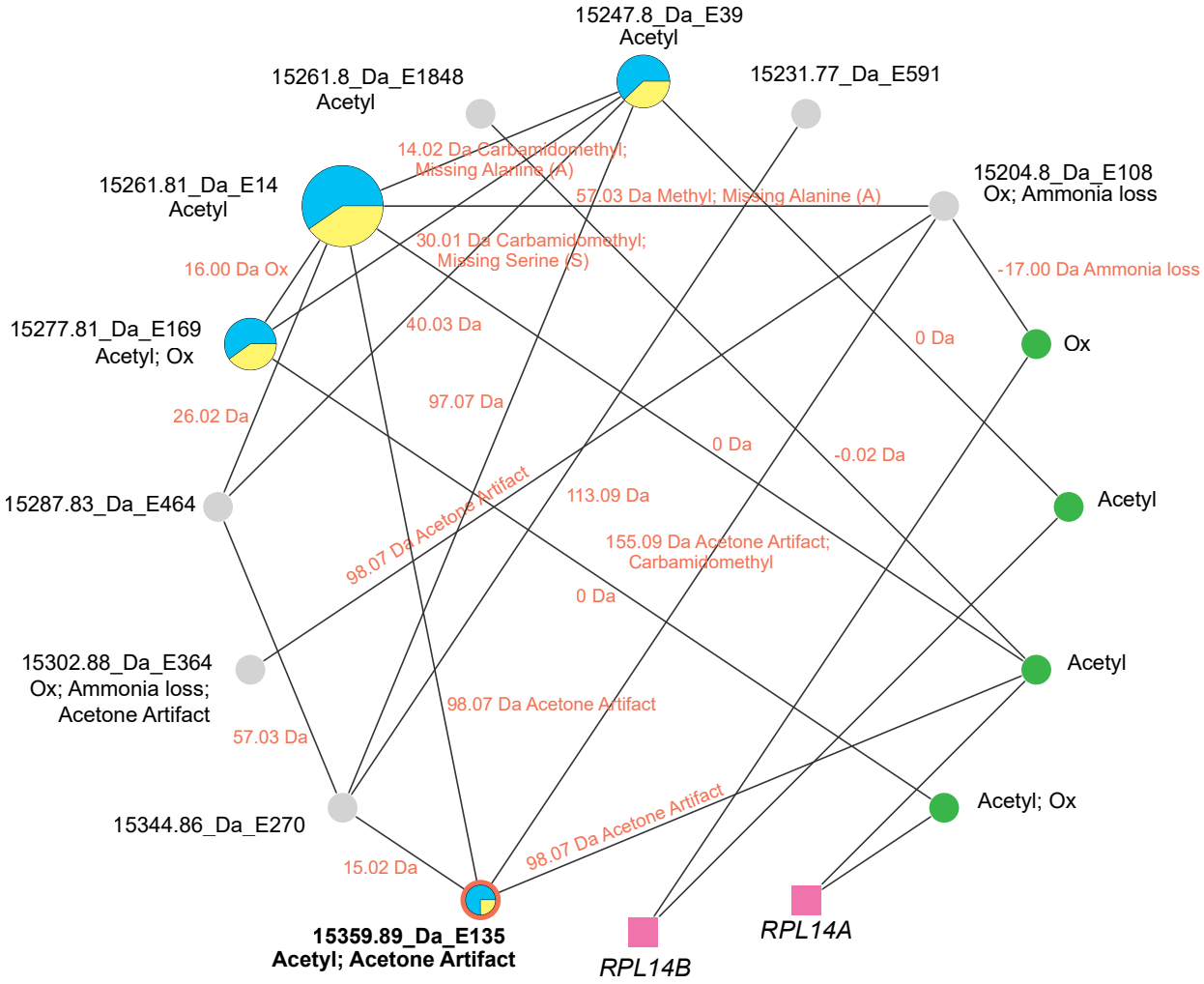

Figure S16: *RPL14A* proteoform family with a significant decrease in the acetylated proteoform with one +98 Da acetone adduct[3]; overall, there appears to be a slight decrease in the acetylated form, although it may not pass significance. Note that this is an ambiguous family; the significantly changing proteoform is one step away from a theoretical proteoform derived from *RPL14A*. Mass differences of 14.02 Da, 30.01 Da, and 155.09 Da form connections between proteoforms in the *RPL14A* and *RPL14B* families, but these connections did not lead to identifications, effectively separating the two families. In the future, we may separate ambiguous families by breaking these weak connections.

1.4.6 60S ribosomal protein L25 (*RPL25*)

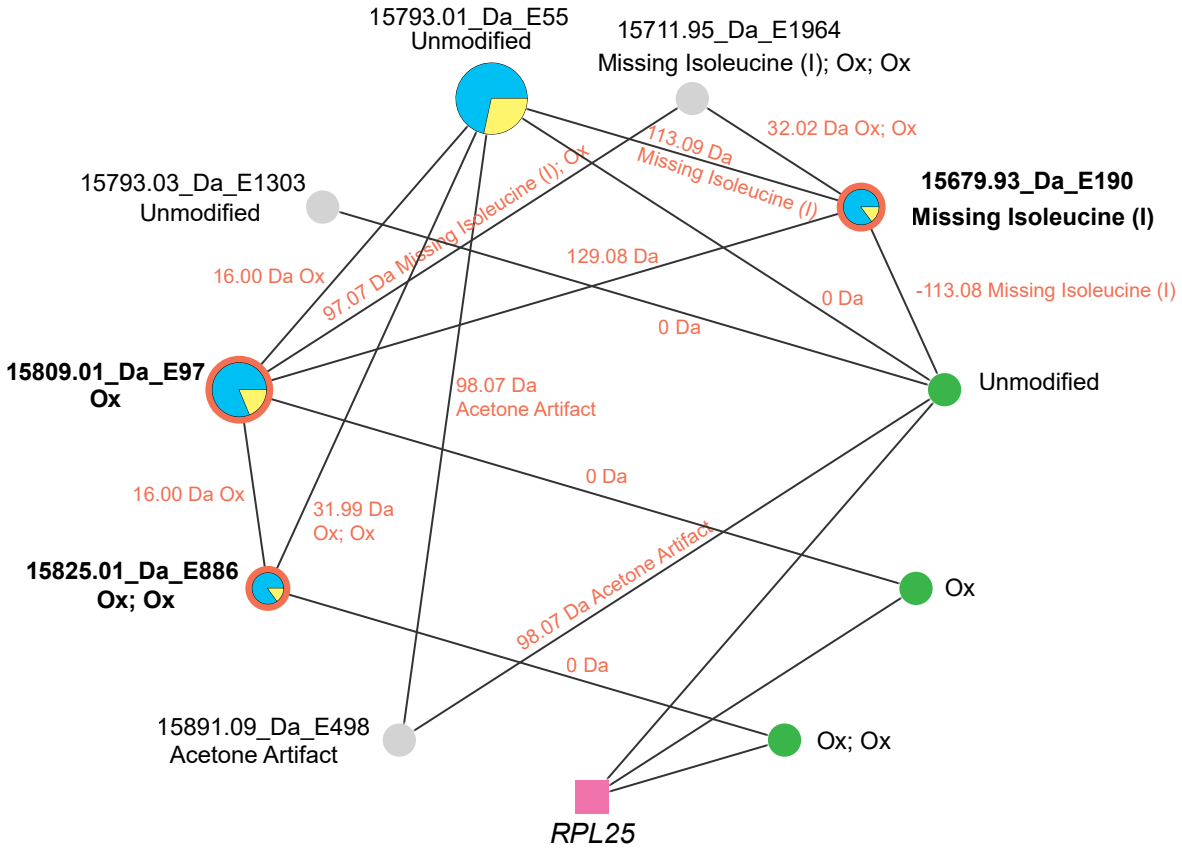

Figure S17: *RPL25* proteoform family with a significant decrease in the proteoform missing the C-terminal isoleucine and two oxidized forms. These changes appear likely to be the result of an overall decrease in the abundance of this protein.

1.4.7 60S ribosomal protein L29 (*RPL29*)

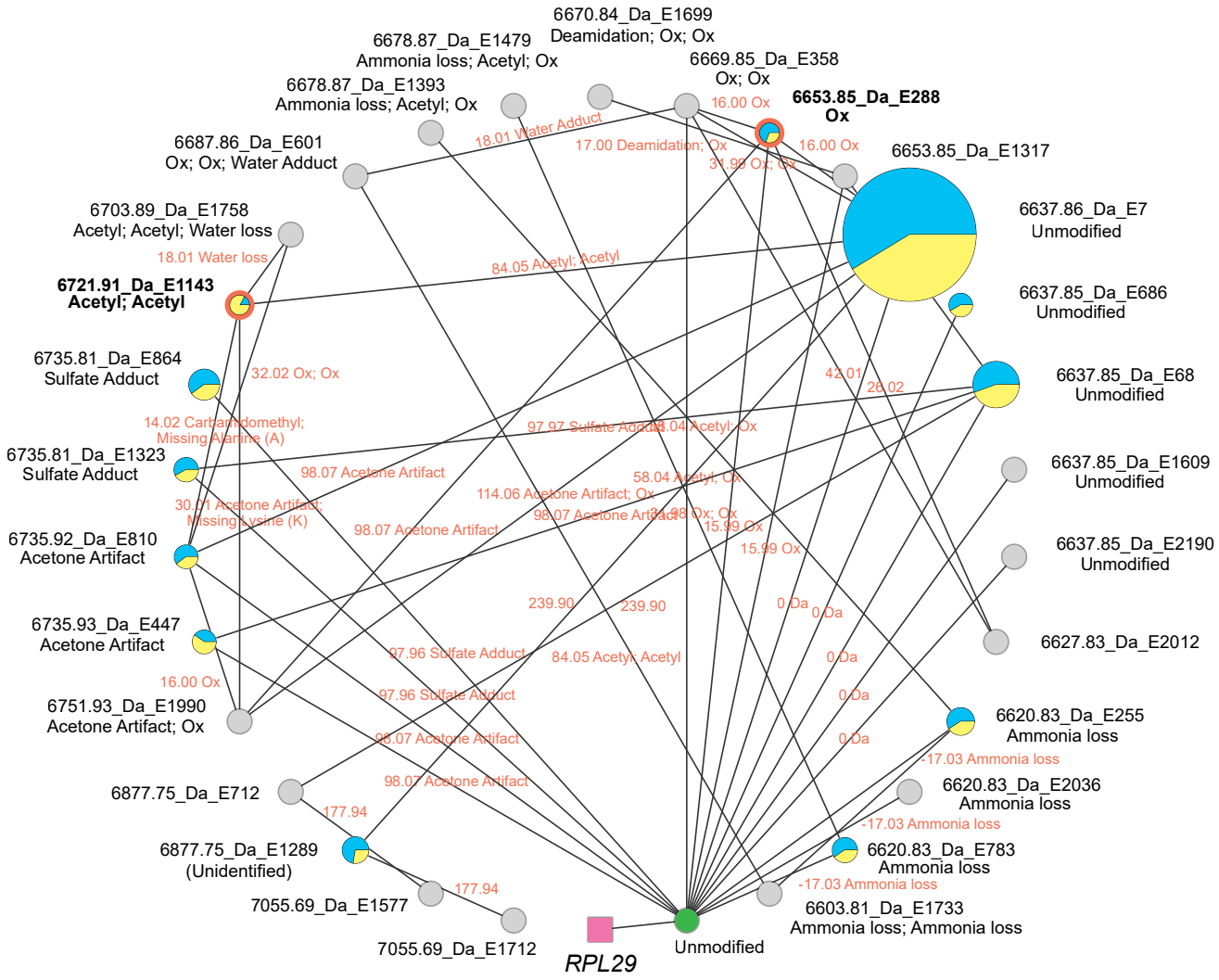

Figure S18: *RPL29* proteoform family with a significant increase in the diacetyl form and a significant decrease in the singly oxidized form in response to salt stress. Notice that a quantified proteoform remains unidentified near the bottom of this diagram (E_1289), connected by a mass difference of 239.90 Da from the unmodified form, which should have been assigned to three phospho groups. Another example can be found in E_172. Continued improvement of the identification algorithm may lead to fewer false negatives of this type. For now, while the automatic identifications have been left for the purpose of this work, this could be easily corrected upon manual inspection.

1.4.8 60S ribosomal protein L32 (*RPL32*)

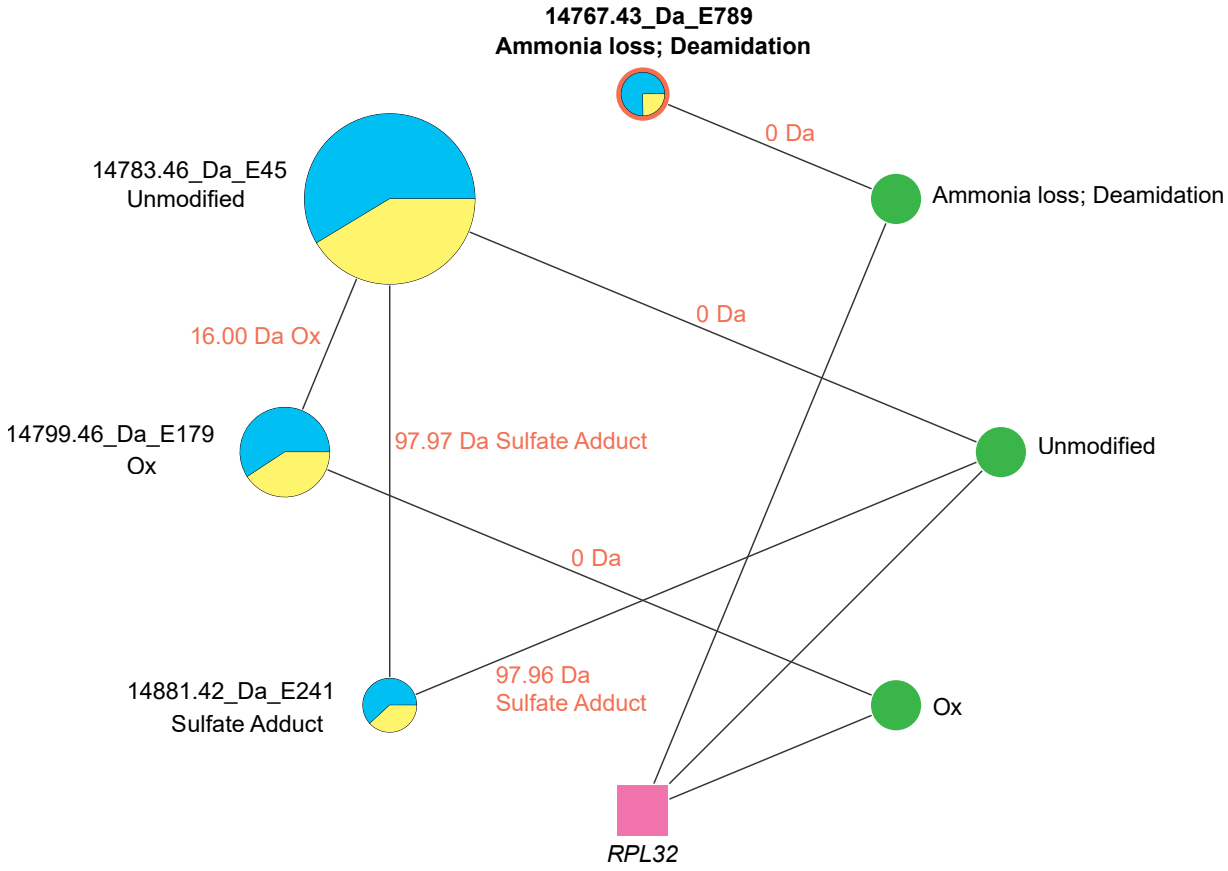

Figure S19: *RPL32* proteoform family with a significant decrease in a proteoform with deamidation in response to salt stress.

1.4.9 60S ribosomal protein L35 (*RPL35B*)

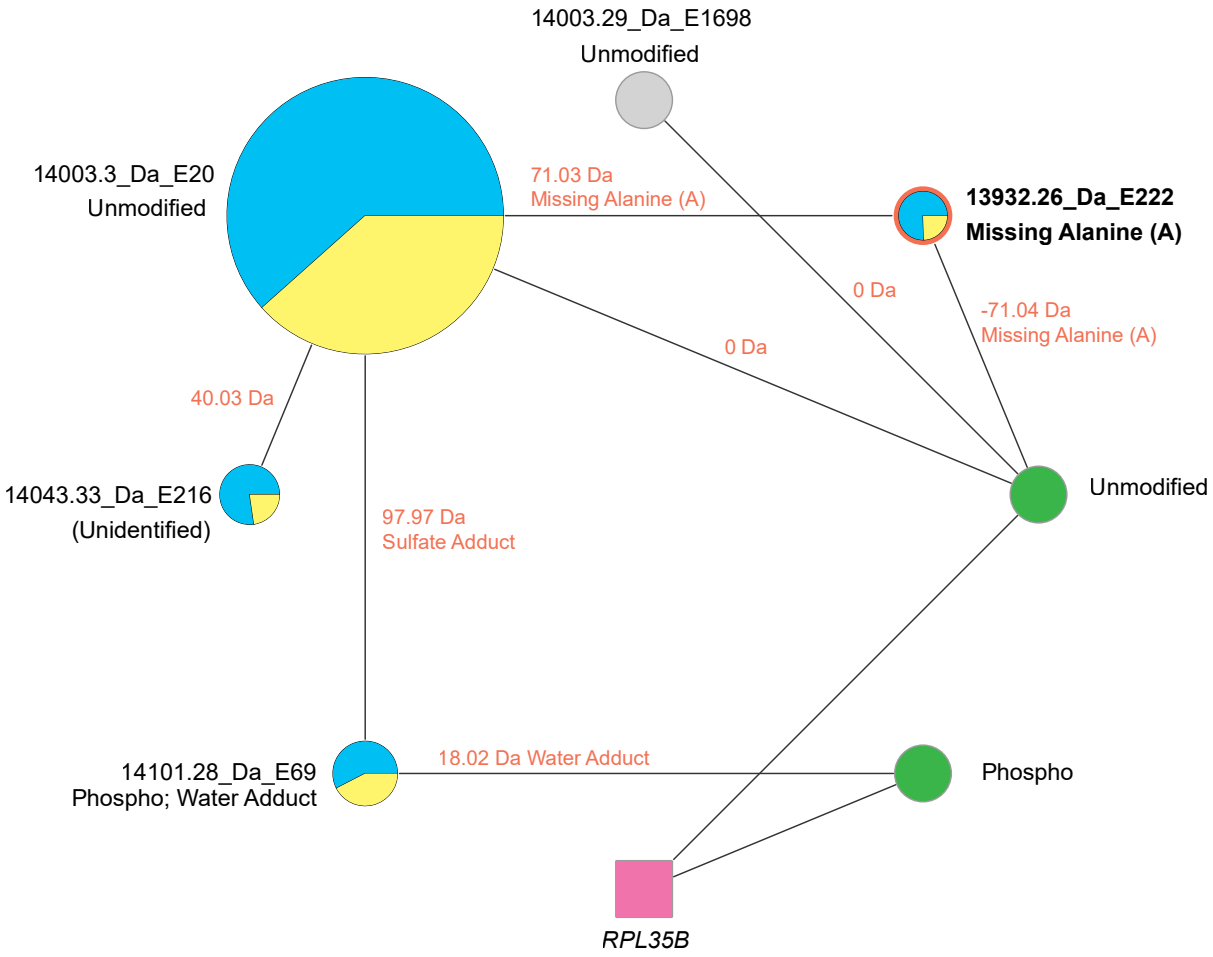

Figure S20: *RPL35B* proteoform family with a significant decrease in the abundance of a proteoform with a cleaved alanine (either at the N- or C-terminus, since the amino acid sequence is MAGV...IKA). One quantified proteoform remains unidentified, connected by a currently unassigned +40.03 Da mass difference from the unmodified form.

1.4.10 60S ribosomal protein L39 (*RPL39*)

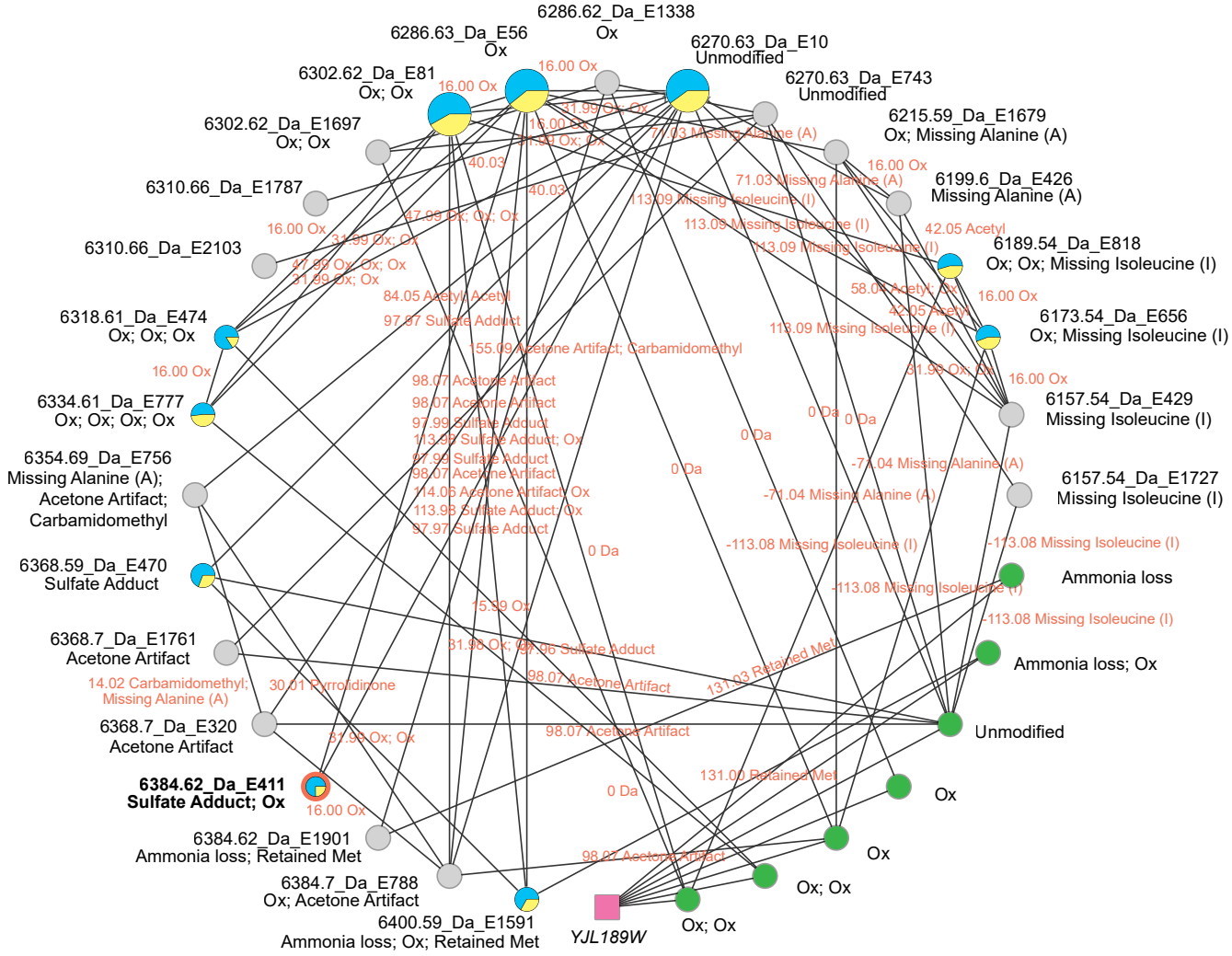

Figure S21: *RPL39* proteoform family containing many adducts. The sulfate adduct of the singly oxidized proteoform was observed at significantly lower abundance in salt stress response. In the future, we may merge quantitative information regarding adducts of the same proteoform.

1.4.11 60S ribosomal protein L41-B (*RPL41B*)

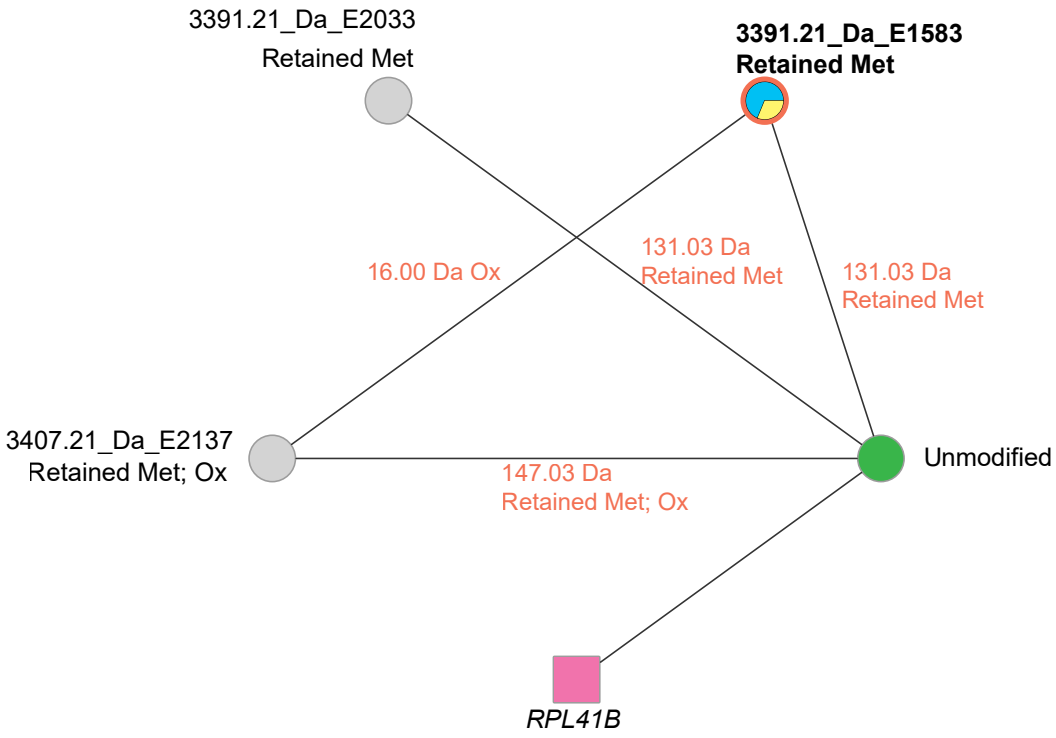

Figure S22: *RPL41B* proteoform family with a significant decrease in a proteoform with a retained N-terminal methionine. (See this reference[2] for more on N-terminal methionine cleavage.)

1.4.12 60S ribosomal protein L42-A (*RPL42A*)

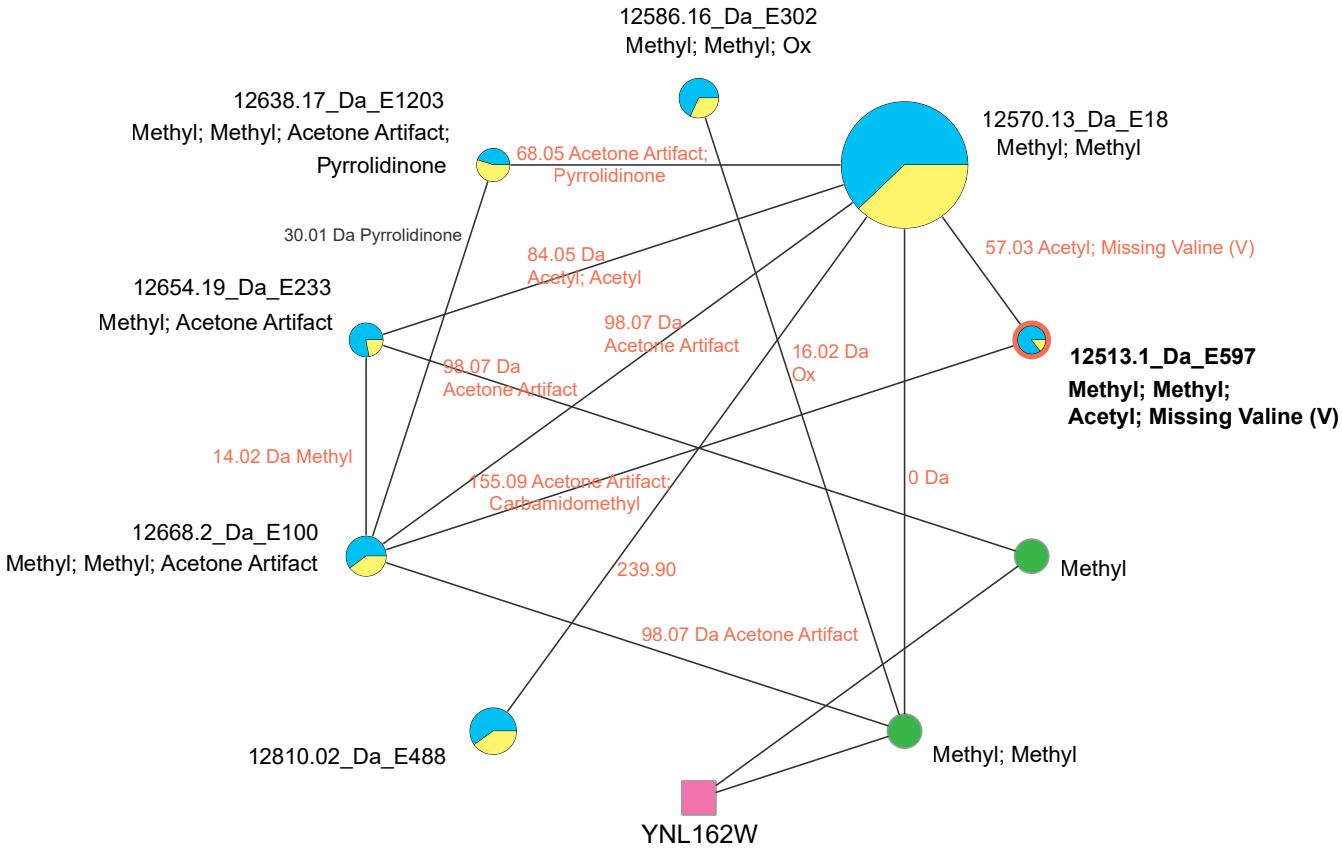

Figure S23: *RPL42A* proteoform family with a significant decrease in the proteoform adorned with two methyl groups and one acetyl group that is also missing the N-terminal valine (the amino acid sequence is MVN...LQF). Note that a quantified proteoform is connected to the dimethylated proteoform by a mass difference of 239.90 Da, which should have been assigned to three phospho groups. Continued improvement of the identification algorithm may lead to fewer false negatives of this type. For now, while the automatic identifications have been left for the purpose of this work, this could be easily corrected upon manual inspection.

2 Quantification Validation

2.1 Quantification Results Using 2-to-1 NeuCode Mixtures

Proteoform Suite uses several processing steps to transform intensity ratios from isotopic labeling into proteoform abundance changes. We confirmed that the intensity ratios measured in individual scans were in fact maintained through those processing steps to give accurate measurements of relative abundance for each quantified proteoform.

The Online Methods section describes how two different isotopically labeled samples were used for identification and quantification. First, proteoform identification was performed using a 2-to-1 mixture of yeast lysates, 2 parts labeled with NeuCode light lysines and 1 part labeled with NeuCode heavy lysines. The known ratio of light to heavy lysines helps identify NeuCode pairs, which in turn helps identify proteoforms. Second, we quantified proteoforms using a 1-to-1 mixture of yeast lysates from two growth conditions, normal growth labeled NeuCode light lysines and salt-stress growth labeled with NeuCode heavy lysines. Measurements in the quantitative experiment are assigned to identified proteoforms as either NeuCode light or NeuCode heavy components by comparing the accurate mass measurements to the expected NeuCode light and heavy masses of the proteoform. Then, by comparing the intensity sums of these measurements, we can draw conclusions about differences in proteoform abundances between the two biological conditions.

To verify that intensity ratios are maintained through these processing steps, we analyzed the data for 2-to-1 mixtures as a quantitative experiment in Proteoform Suite, expecting to observe the known ratio both as the intensity ratio in individual scans and again as the intensity ratio after combining observations across the experiment. Each proteoform in this experiment was observed as a NeuCode pair, with light and heavy NeuCode lysines; the intensity ratios of these NeuCode pairs peaked at 2.0, as shown in the left panel of Figure S24. Quantified proteoforms had intensity ratios that peaked at 2.2, as shown in the right panel of Figure S24, confirming that the data processing involved in quantification maintains the intensities observed in isotopic labeling. The ragged appearance of the second histogram is due to the smaller sample size: the left histogram was generated with 37,431 NeuCode pairs, and the right histogram was generated with 647 quantified experimental proteoforms (only those observed in both technical replicates of both normal and stress 2-to-1 mixtures).

While most proteoforms in this experiment had an intensity ratio near the expected value of 2.0, about forty proteoforms had high intensity ratios (8.0 up to 54.7). As seen in Figure S25, these are predominantly low-abundance outliers. This bias towards high intensity ratios (light NeuCode / heavy NeuCode) at low intensities is due to the loss of observations for heavy-labeled proteoforms, which fall below the limit of detection before their light counterparts that have twice the abundance in this experiment.

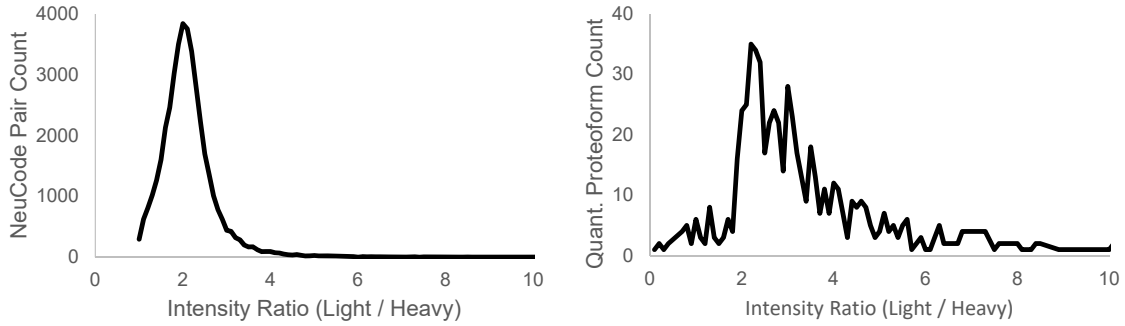

Figure S24: Histograms of intensity ratios for individual NeuCode pairs and for quantified proteoforms are similar, matching the expected 2-to-1 ratio for this experiment. This confirms that intensity ratios from isotopic labeling are maintained through the processing steps in Proteoform Suite.

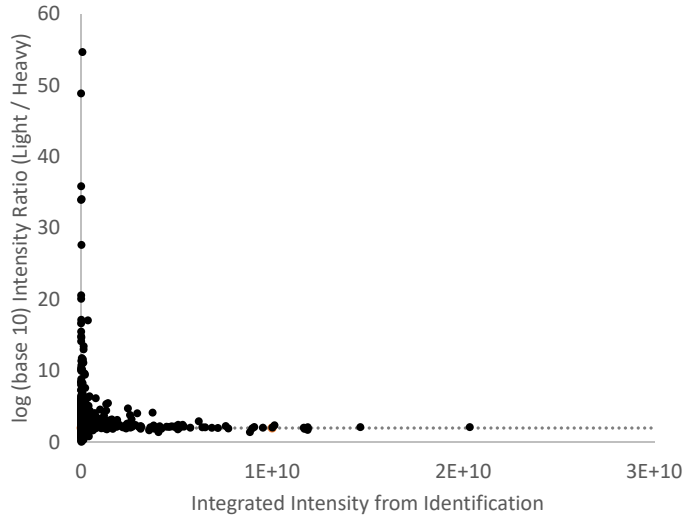

Figure S25: Quantified proteoforms with an expected 2-to-1 intensity ratio (dotted line) have outliers with high intensity ratios that lie predominantly at lower abundances. This is shown in a scatter plot, where the integrated intensity during identification stands in as a proxy for abundance.

3 Supporting Methods

3.1 Protein Database Annotation Using G-PTM-D Strategy

Proteoform Suite takes advantage of modifications annotated in the UniProt XML format. Many modifications are already annotated in UniProt, but we recently reported a strategy that extends the database to include many sample-specific modifications discovered using a bottom-up proteomics strategy[4]. This strategy first uses a wide-mass search to identify and annotate candidate PTM sites, and then uses a second-pass search to identify these modified peptides with a $\sim$ 1% FDR. We analyzed bottom-up proteomics data described below, kept all original UniProt modifications, and added only new modifications that passed a 1% FDR cutoff by target-decoy analysis. This new database was used for the intact mass analysis in this work. MetaMorpheus version 0.0.128 was used for this analysis (available at `https://github.com/smith-chem-wisc/MetaMorpheus`).

3.2 Bottom-Up Proteomics for G-PTM-D Strategy

3.2.1 Cell Culture and Lysate Preparation

Yeast cells were grown in YPD media to an OD_{600} of about 2.0. For stressed conditions, YPD containing NaCl sterile solution was added to give a salt concentration of 0.7 M. Salt-stressed culture continued to shake for 30 min, allowing biological changes to take place. Cells were pelleted and washed with PBS. Pellets were resuspended in lysis buffer (200 mM NaCl, 20 mM EDTA, 50 mM Tris pH 7) containing 200X diluted protease inhibitors DMSO cocktail solution, and then were lysed with a Constant Systems TS series cell disruptor at 30 kpsi. SDS was added to the cell lysate to a 1% final concentration. Lysate was centrifuged at 4°C at 8,000 g for 12 min to clear cell debris. The supernatant was diluted five-fold in lysis buffer (containing protease inhibitors) to lower the SDS concentration to 0.2%.

3.2.2 TCA Precipitation

The protein from the cell lysate samples was precipitated by first adding 320 μL 100% TCA to the 4.8 mL lysate (in four tubes). The sample was then gently mixed, incubated on ice for 10 min, and centrifuged at 20,000 g at 4°C for 20 min. The resulting supernatants were decanted. During centrifugation, the second tube for each sample was precipitated with TCA and added to the pellet in the first tube. The serial addition of TCA-precipitated protein was repeated for all four tubes. The TCA pellet was washed twice with 750 μL chilled acetone and centrifuged at 20,000 g at 4°C for 5 min. The pellet was heated at 95°C for 2 min to remove the residual acetone.

3.2.3 eFASP Procedure

A filter unit and an eFASP collection tube were passivated by soaking them in 5% Tween 20 overnight. The filter unit and collection tube were rinsed at least three times using nanopure water. The lysate protein pellet was resuspended in 810 μL eFASP exchange buffer (8 M urea, 0.10% deoxycholic acid). Then, 90 μL DNaseI reaction buffer and 1 μL DNaseI were added, and the sample was incubated at 37 °C for 10 min. 450 μL of the DNaseI-treated sample was transferred into a passivated filter unit placed in a unpassivated tube and centrifuged at 14,000 g at 15 °C for 10 min. The flowthrough was discarded, and the rest of the sample was added into the filter units and centrifuged again. 200 μL exchange buffer was added in the filter unit, and the sample was centrifuged at 14,000 g at 15 °C for 10 min. The flowthrough was discarded. This washing step was repeated three times. Then, 200 μL eFASP reducing buffer (8 M Urea, 20 mM DTT) was added to the filter unit, and the sample was incubated at room temperature for 30 min, then centrifuged at 14,000 g at 15 °C for 10 min. Flowthrough was discarded again. 200 μL eFASP alkylation buffer (8 M urea, 50 mM iodoacetamide, 50 mM ammonium bicarbonate) was added in each filter unit, and the sample was incubated at room temperature in the dark for 30 min. 15 μL DTT was added in sample and incubated for 10 min. The sample was centrifuged again at 14,000 g at 15 °C for 10 min, and the flowthrough was discarded. 200 μL eFASP digestion buffer (1 M urea, 50mM ammonium bicarbonate, 0.1% DCA) was added in the filter unit, and the sample was centrifuged at 14,000 g at 15 °C for 10 min. The flowthrough was discarded. This washing step was repeated three times. The filter was transferred to passivated collection tubes. 100 μL digestion buffer and 1 μg trypsin was added to the filter unit. The tube and filter unit were wrapped in Parafilm® and incubated at 37 °C overnight with no rotation. When the digestion was complete, Parafilm® was removed, and the tube was centrifuged at 14,000 g at 15 °C for 10 min. 50 μL of 50 mM ammonium bicarbonate was added into the filter unit and centrifuged at 14,000 g at 15 °C for 10 min. This step was repeated twice. The flowthrough was transferred to new a 1.7 mL low-retention tube. 200 μL ethyl acetate and 200 μL 1% TFA was added to the sample, and then shaken for 1 minute. The sample was centrifuged at 15,800 g at 15 °C for 2 min. The top layer was removed from the tube, and another 200 μL ethyl acetate was added, and then shaken for 1 min. The sample was centrifuged again at 15,800 g at 15 °C for 2 min, and then the top layer was removed. The sample was dried in the Savant SpeedVac™ Concentrator to dryness for about 150 min. The dry tube contents were dissolved in 180 μL 0.1% TFA and vortexed to mix.

3.2.4 C18 Solid-Phase Extraction

An extraction pipette tip was activated by pipetting 180 μL 70% ACN up and down three times, and then was washed in 0.1 TFA three time by pipetting and discarding 180 μL of 0.1% TFA. Lysate sample was pipetted up and down 3 times using the washed tip. Extraction tips were than washed in 0.1% TFA as previously described. Peptides were eluted into 150 μL 70% ACN and 0.1% TFA

by pipetting up and down 5 times in 600 μL low retention tubes. Eluted sample was SpeedVacked to dryness for about 50 min. Tube contents were reconstituted in 200 μL 95:5 H2O:ACN and 0.1% formic acid solutions. On the mass spectrometer, one technical replicate injects 2 μL of control. Two technical replicates were performed.

3.2.5 HPLC-ESI-MS/MS Analyses

Samples were analyzed by HPLC-ESI-MS/MS using a system consisting of a high performance liquid chromatograph (nanoAcquity, Waters) connected to an electrospray ionization (ESI) Orbitrap mass spectrometer (LTQ Velos, ThermoFisher Scientific). HPLC separation employed a 100 x 365 μm fused silica capillary micro-column packed with 20 cm of 1.7 μm-diameter, 130 Å pore size, C18 beads (Waters BEH), with an emitter tip pulled to approximately 1 μm using a laser puller (Sutter instruments). Peptides were loaded on-column at a flow-rate of 400 nL/min for 30 min, and then eluted over 120 min at a flow-rate of 300 nl/min with a gradient of 2% to 30% acetonitrile, in 0.1% formic acid. Full-mass profile scans were performed in the FT orbitrap between 300-1500 m/z at a resolution of 60,000, followed by ten MS/MS HCD scans of the ten highest intensity parent ions at 42% relative collision energy and 7,500 resolution, with a mass range starting at 100 m/z. Dynamic exclusion was enabled with a repeat count of two over the duration of 30 seconds and an exclusion window of 120 sec.

3.3 Removing Missed Monoisotopic Masses and Charge State Harmonic Errors

As noted in the Methods section, each component in intact proteoform analysis is comprised of a deconvoluted monoisotopic mass and an intensity value. These results contain some errors resulting from (1) "missed" monoisotopic masses, where a nearby isotopic peak is reported as the monoisotopic mass and (2) charge state harmonics (yielding a multiple of the monoisotopic mass). The charge state harmonic artifact appears to be more severe for our NeuCode data than it is for data obtained from unlabeled samples, likely because of the presence of two overlapping sets of isotopic peaks rather than just a single set.

To correct for these errors, we first join components with others in the same scan that correspond to missed monoisotopic masses. In descending order of intensity, a selected component is merged with all other components that fall within $\pm$10 ppm tolerance of each of a set of missed monoisotopic masses: -3, -2, -1, 1, 2, or 3 Da away from the component. Merging components consists of recomputing the mass for each charge state as an intensity-based weighted average across the same charge states of components in the merger; the mass of the component is then recomputed as an intensity-based weighted average across all of the charge states.

Then, the selected component is merged with all others in the same scan having a charge state harmonic of the selected mass and of several missed monoisotopic masses. Components that were already merged are excluded from this second step and from all subsequent iterations.

Components are ordered by mass in descending order, so that only smaller harmonic masses need to be considered upon each iteration. Specifically, we look for the second harmonic of a component mass that yields a mass value equal to half of the fundamental mass, and the third harmonic that yields a mass value equal to a third of the fundamental mass. In the present analysis, we looked for the second harmonics of up to 4 missed monoisotopics above or below the selected component, as well as the third harmonics of up to 6 missed monoisotopics above or below the mass of the selected component; components within ±10 ppm of these mass values were collected for the next step. With one important exception, the component with more detected charge states (generally also the higher intensity component) serves as the base for merging charge state harmonics, and the smaller intensity component is removed from future consideration. The exception occurs when a proteoform contains 14 lysine residues, which leads to a mass difference of $0.036 * 14 = 0.504$ Da; this is a special case because the NeuCode pair splits the $\sim$ 1 Da isotopic spacing down the middle, producing many deconvolution artifacts with double the mass and double the intensity of the fundamental mass. To account for these cases, we pair all components (excluding those already merged) using the NeuCode pairing algorithm described in the Methods section. In cases where the smaller mass component of a charge state harmonic falls in a NeuCode pair with a lysine count of 14, this smaller mass component serves as the base for the merger, and the higher mass component is removed for future consideration.

3.4 FDRs for Mass Difference Peaks

The FDR for each mass difference can be estimated, as described in our previous work[5]. This assessment accurately represents the FDR for identifications via Exp-Theo connections, since all of these connections are considered direct identifications. However, identifications via Exp-Exp connections now benefit from the additional information taken into consideration about the protein sequence, the PTMs known to be presented on the protein, and the frequency of each PTM in the protein database; not all Exp-Exp mass differences lead to identifications in light of this information. Therefore, the Exp-Exp peak FDRs represent the maximum possible false positive rate for those connections. The new FDR assessment discussed in the Online Methods section using decoy proteoform families gives an accurate, improved proteoform FDR by applying these new identification criteria to decoy families, greatly limiting the number of Exp-Exp connections that result in false identifications.

3.5 Making Permutations and Tusher Plots for Permutation Analysis

The permutation analysis used to evaluate the false positive rate for calling significant changes between normal and salt-stressed yeast was based on the original work[6] by Tusher *et al.* In that paper, microarray hybridization results were analyzed with a method using permutations of the data to assess the significance of the results, named statistical analysis of microarrays (SAM). Applying

this method to determine which proteoforms changed significantly in response to perturbation was not trivial, and so we aim in this section to provide details on how permutations were performed and used for the statistical analysis.

The original experiment had three dimensions: two cell lines (1 and 2) that produced the most variation in results, two conditions (irradiated or induced, I, and unirradiated or uninduced, U) that were the comparison of biological interest, and two hybridizations that represent technical replicates of the array hybridization (A and B). To create a distribution of expected values given the data, 36 balanced permutations of the two cell lines were made from the 8 original labels:

U1A, U1B, U2A, U2B & I1A, I1B, I2A, I2B.

These permutations are listed in Table S10. Note that each set of permuted labels has two samples from cell line 1 and two samples from cell line 2 on each side; this makes these permutations balanced for the cell lines. Also note that the set of labels used for evaluating biological significance are included in the permutations.

The experimental design of the present work differs from this microarray experiment. The first difference is the use of 3 biological replicates, where 2 were used in the original work. Most strikingly, they differ in the fractionation of cell lysates from these 3 biological replicates: each was fractionated into 12 molecular weight fractions before analysis by mass spectrometry, a step that has no similarity to the original work. Each fraction was then injected twice, producing 2 technical replicates. An important similarity of the two studies is that both use induced and uninduced conditions to evaluate the effect of the perturbation.

To mimic the permutations used in the original work, we treated the technical replicates like the hybridizations in the original experiment and summed the intensities observed for each proteoform across all fractions. This produced 6 labels for both normal and stress conditions (N and S), comprised of 3 biological replicates (A, B, and C) and 2 technical replicate labels (1 and 2):

NA1, NA2, NB1, NB2, NC1, NC2 & SA1, SA2, SB1, SB2, SC1, SC2.

Producing balanced permutations on the largest source of variation, biological replicates (analogous to permuting on the cell lines in the original experiment), gave 216 balanced permutations. The main principle to generating these permutations is similar to before: each balanced permutation contains the same number of each biological replicate (2 A's, 2 B's, and 2 C's) on each side of the comparison as in the original labels. Two examples, in addition to the original set, are as follows:

SA1, NA2, NB1, NB2, NC1, NC2 & NA1, SA2, SB1, SB2, SC1, SC2

NA1, SA1, NB1, SB1, NC1, SC1 & NA2, SA2, NB2, SB2, NC2, SC2.

Each meets the criteria for the number of biological replicates on both sides, and all sample labels are represented.

In preparation for calculating the statistics based on these permutations, the original paper normalized and zero-centered the data. Zero-centering after normalization is an important precursor to the statistical analysis that follows. In this study, it involves subtracting the average quantified

proteoform intensity from the individual intensity values for each proteoform. This helps make the statistics calculated below symmetrical around the origin, which is important for drawing the cutoffs discussed below and in the original paper.

The SAM analysis relies on a statistic called the relative difference for each gene (i) (or each quantified proteoform (i) in the present study), $d(i) = (\bar{x}_{induced} - \bar{x}_{uninduced})/(s_p + s_0)$, where s_p is the pooled standard deviation (a formula given in the original work), and where s_0 as a constant. The relative difference statistic was calculated for each gene (each quantified proteoform) using the original set of labels to give the observed relative differences.

The constant s_0 was important in the microarray study for eliminating the dependence of the relative difference on the magnitude of the observed intensities. Namely, very small (*i.e.*, < 1) intensities had very small pooled standard deviations, which could make artificially high relative differences. Because the intensities observed in mass spectrometry experiments are all very high, *i.e.* $> 10^4$, this value was unimportant for the current work and was arbitrarily set to 1.

To establish significance, the observed relative differences are compared to expected ones calculated based on the balanced permutations. The first step in calculating these expected relative differences is calculating the relative difference for each gene (i) (or each quantified proteoform (i) in the present study), taking the difference between the average of the values on the right side ("induced") and the left side ("uninduced"). Note that the pooled standard deviation is calculated independently for each (i) in each permutation. Then, the relative differences are sorted independently for each of the 216 permuted sample labels. This produces 216 ranked lists. The expected relative difference at each rank is the average relative difference at that rank from each permutation.

The observed relative differences are ranked and plotted against these ranked expected relative differences to give the "Tusher plot," which is shown in the Online Methods section of the main text. In this plot, the line $y = x$ represents when the relative difference is the same as would be expected. This should be the case for most proteoforms. Zero-centering ensures that the relative differences are not biased, thereby ensuring that the thresholds applied vertically from this line (*e.g.*, the thresholds $y = x - 0.7$ and $y = x + 0.7$ were used in this work) are not biased towards high or low relative differences. Finally, a significantly changing gene or proteoform is one that has an observed relative difference that falls far from the expected relative difference, *i.e.* outside those thresholds. (We allowed additional criteria to qualify as significance in analyzing proteoform abundance changes; see the Online Methods section in the main text for those specific criteria and the FDR calculation for this analysis.)

References

(1) Krebs, J. E. Moving marks: Dynamic histone modifications in yeast. *Mol Biosyst* **2007**, *3*, 590–597.

(2) Frottin, F.; Martinez, A.; Peynot, P.; Mitra, S.; Holz, R. C.; Giglione, C.; Meinnel, T. The Proteomics of N-terminal Methionine Cleavage. *Mol Cell Proteomics* **2006**, *5*, 2336–2349.

(3) Güray, M. Z.; Zheng, S.; Doucette, A. A. Mass Spectrometry of Intact Proteins Reveals +98 u Chemical Artifacts Following Precipitation in Acetone. *J Proteome Res* **2017**, *16*, 889–897.

(4) Li, Q.; Shortreed, M. R.; Wenger, C. D.; Frey, B. L.; Schaffer, L. V.; Scalf, M.; Smith, L. M. Global Post-Translational Modification Discovery. *J Proteome Res* **2017**, *16*, 1383–1390.

(5) Shortreed, M. R.; Frey, B. L.; Scalf, M.; Knoener, R. A.; Cesnik, A. J.; Smith, L. M. Elucidating Proteoform Families from Proteoform Intact Mass and Lysine Count Measurements. *J Proteome Res* **2016**, *15*, 1213–1221.

(6) Tusher, V. G.; Tibshirani, R.; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **2001**, *98*, 5116–21.

Table S10: The 36 balanced permutations used in the original SAM analysis. The bold-face group is the biologically relevant comparison; it is included as one of the permutations.

U1A	**U1B**	**U2A**	**U2B**	**&**	**I1A**	**I1B**	**I2A**	**I2B**
U1A	I1A	U2A	U2B	&	U1B	I1B	I2A	I2B
U1A	I1B	U2A	U2B	&	U1B	I1A	I2A	I2B
U1B	I1A	U2A	U2B	&	U1A	I1B	I2A	I2B
U1B	I1B	U2A	U2B	&	U1A	I1A	I2A	I2B
I1A	I1B	U2A	U2B	&	U1A	U1B	I2A	I2B
U1A	U1B	U2A	I2A	&	I1A	I1B	U2B	I2B
U1A	I1A	U2A	I2A	&	U1B	I1B	U2B	I2B
U1A	I1B	U2A	I2A	&	U1B	I1A	U2B	I2B
U1B	I1A	U2A	I2A	&	U1A	I1B	U2B	I2B
U1B	I1B	U2A	I2A	&	U1A	I1A	U2B	I2B
I1A	I1B	U2A	I2A	&	U1A	U1B	U2B	I2B
U1A	U1B	U2A	I2B	&	I1A	I1B	U2B	I2A
U1A	I1A	U2A	I2B	&	U1B	I1B	U2B	I2A
U1A	I1B	U2A	I2B	&	U1B	I1A	U2B	I2A
U1B	I1A	U2A	I2B	&	U1A	I1B	U2B	I2A
U1B	I1B	U2A	I2B	&	U1A	I1A	U2B	I2A
I1A	I1B	U2A	I2B	&	U1A	U1B	U2B	I2A
U1A	U1B	U2B	I2A	&	I1A	I1B	U2A	I2B
U1A	I1A	U2B	I2A	&	U1B	I1B	U2A	I2B
U1A	I1B	U2B	I2A	&	U1B	I1A	U2A	I2B
U1B	I1A	U2B	I2A	&	U1A	I1B	U2A	I2B
U1B	I1B	U2B	I2A	&	U1A	I1A	U2A	I2B
I1A	I1B	U2B	I2A	&	U1A	U1B	U2A	I2B
U1A	U1B	U2B	I2B	&	I1A	I1B	U2A	I2A
U1A	I1A	U2B	I2B	&	U1B	I1B	U2A	I2A
U1A	I1B	U2B	I2B	&	U1B	I1A	U2A	I2A
U1B	I1A	U2B	I2B	&	U1A	I1B	U2A	I2A
U1B	I1B	U2B	I2B	&	U1A	I1A	U2A	I2A
I1A	I1B	U2B	I2B	&	U1A	U1B	U2A	I2A
U1A	U1B	I2A	I2B	&	I1A	I1B	U2A	U2B
U1A	I1A	I2A	I2B	&	U1B	I1B	U2A	U2B
U1A	I1B	I2A	I2B	&	U1B	I1A	U2A	U2B
U1B	I1A	I2A	I2B	&	U1A	I1B	U2A	U2B
U1B	I1B	I2A	I2B	&	U1A	I1A	U2A	U2B
I1A	I1B	I2A	I2B	&	U1A	U1B	U2A	U2B