Reliable identification of significant differences in differential HX-MS measurements using a hybrid significance testing approach

Tyler S. Hageman[1†] and David D. Weis[1,2,*]

[1]Department of Chemistry and the Ralph N. Adams Institute for Bioanalytical Chemistry,

[2]Department of Pharmaceutical Chemistry, University of Kansas

1567 Irving Hill Road, Lawrence, KS 66045

[†]Present address: AbbVie Bioresearch Center, 100 Research Drive, Worcester, MA 01605

[*]Corresponding Author: dweis@ku.edu

ABSTRACT

Differential hydrogen exchange-mass spectrometry (HX-MS) measurements are valuable for identification of differences in the higher order structures of proteins. Typically, the data sets are large with many differential HX values corresponding to many peptides monitored at several labeling times. To eliminate subjectivity and reliably identify significant differences in HX-MS measurements, a statistical analysis approach is needed. In this work, we performed null HX-MS measurements (i.e., no meaningful differences) on maltose binding protein and infliximab, a monoclonal antibody, to evaluate the reliability of different statistical analysis approaches. Null measurements are useful for directly evaluating the risk (i.e., falsely classifying a difference as significant) and power (i.e., failing to classify a true difference as significant) associated with different statistical analysis approaches. With null measurements, we identified weaknesses in the approaches commonly used. Individual tests of significance were prone to false positives due to the problem of multiple comparisons. Incorporation of Bonferroni correction led to unacceptably large limits of detection, severely decreasing the power. Analysis methods using a globally estimated significance limit also led to an over-estimation of the limit of detection, leading to a loss of power. Here, we demonstrate a hybrid statistical analysis, based on volcano plots, that combines individual significance testing with an estimated global significance limit, simultaneously decreased the risk of false positives and retained superior power. Furthermore, we highlight the utility of null HX-MS measurements to explicitly evaluate the criteria used to classify a difference in HX as significant.

INTRODUCTION

Hydrogen exchange-mass spectrometry (HX-MS) has emerged as a routine method for evaluating protein higher order structure.[1] The success of HX-MS is largely derived from its ability to overcome limitations of protein size, sample conditions, experiment time, or resolution that limit alternative methods.[2,3] HX-MS measurements have revealed localized structural changes associated with protein-protein interactions,[2,4-6] protein-ligand interactions,[7-11] environmental stresses,[12-14] and primary sequence modifications.[15-18] In addition, HX-MS has been used for comparability studies of protein therapeutics.[19-22]

Traditional HX-MS experiments are performed in a differential manner using a bottom-up workflow in which peptide-level HX measurements of two samples are compared.[23] The objective of a differential HX-MS experiment is to determine if meaningful HX differences between the two samples are present. In many cases, meaningful differences are identified by subjective classification of peptide deuterium uptake plots. Depending on the application and protein studied, there could be few or many differences observed and those differences could be small or large. For example, HX differences observed in an epitope mapping study might be large and localized to the epitope.[24] In contrast, HX differences observed in a conformational study of varying protein drug product formulations might be small and globally distributed.[25] To probe for structural differences between two samples with HX-MS, multiple HX labeling time measurements, with technical replicates, are obtained to sample a range of protein dynamics. After proteolytic digestion, mean peptide HX values at each HX labeling time from each sample are compared to identify HX differences. With many peptides and multiple HX labeling times, differential HX-MS data sets are large, resulting in many HX comparisons. To reveal meaningful differences in large HX-MS data sets and to eliminate subjectivity, statistical analysis is necessary.

Significance testing is a suitable approach for comparing two independent sample means to determine if a significant difference is present between two populations. For significance testing, a null hypothesis (H_0) is defined as the population means (μ_a, μ_b) do not differ (i.e., $H_0: \mu_a - \mu_b = 0$) and an alternative hypothesis (H_1) as the population means differ (i.e., $H_1: \mu_a - \mu_b \neq 0$). Hypothesis testing is subject to type I error and type II error. In the context of HX differences, a type I error is any differential HX measurement in which the null hypothesis is falsely rejected (i.e., false positive), in other words, classifying an observed HX difference as significant when there is not a true difference. Type II error is any event in which the null hypothesis is false, but not rejected (i.e., false negative), which would be failing to classify a true HX difference as significant. The probability of committing type I error (i.e., risk) and the probability of not committing type II error (i.e., power) will vary depending on the test and criteria used to determine significance. A simple hypothesis test for differential HX-MS data can be performed, under the assumption that the HX measurement errors follow a normal distribution, by calculating a confidence interval for the difference between observed mean peptide masses ($\Delta\overline{HX} = \overline{m}_a - \overline{m}_b$) for each peptide at each label time where the subscripts, a and b, denote each protein sample state measured. The standard deviations ($s_a\ ,s_b$) and number of replicates ($n_a\ ,n_b$) for each mean peptide mass can be used with a critical value (t) from the Student's t-distribution to calculate a confidence interval for hypothesis testing:

$$H_1 : \left|\Delta\overline{HX}\right| > t\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}} \qquad (1)$$

Rearranging equation (1) yields a confidence interval:

$$\Delta\overline{HX} \pm t\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$$

If zero is contained in the confidence interval, then the null hypothesis cannot be rejected (i.e., the difference is not considered statistically significant). Conversely, if the confidence interval does not include zero, then the null hypothesis is rejected (i.e., the difference is considered statistically significant). The statistical confidence for rejecting the null hypothesis is defined by the significance level (α) at which the critical t-value is selected.

For significance testing of differential HX-MS data, both individual and global confidence intervals have been used.[26-31] With individual confidence intervals, significance is tested by calculating a confidence interval for each $\Delta\overline{HX}$ value by using associated sample standard deviations ($s_a\ ,s_b$) for each mean difference. Any $\Delta\overline{HX}$ confidence interval not including zero is classified as significant. Alternatively, with global confidence intervals, significance is tested by calculating a global confidence interval for all $\Delta\overline{HX}$ values. Following the global confidence interval approach, significance testing results are easily illustrated in a difference plot of $\Delta\overline{HX}$ values, generally with significance limit lines representing the global confidence intervals. Any $\Delta\overline{HX}$ value exceeding the significance limit lines, in either direction, is classified as significant (i.e., reject the null hypothesis), while any value that does not exceed limit is classified as insignificant (i.e., fail to reject the null hypothesis). Unlike the individual confidence interval approach, where equal sample standard deviations are not assumed, the global confidence interval approach assumes equal sample standard deviations. Although most of the work involving significance testing of HX-MS data has focused on differences at individual HX times, significance testing has also been employed when HX data is averaged or integrated across the entire HX time range.[32-34] Both global and individual confidence interval approaches have been applied for classifying significant results in differential HX-MS measurements.[27-31,35] However, the rates of type I and type II error when applying these approaches to determine significance in differential HX-MS measurements have not been evaluated. A null experiment, with an expected outcome, would be useful to evaluate the error rates from these significance-determining approaches for differential HX-MS data. In this work, we designed such null experiments to explicitly evaluate the relationship of risk and power associated with individual and global significance testing. We demonstrate that a hybrid approach for significance testing overcomes weaknesses we identified for individual and global significance testing. We validate the hybrid approach with our experimental null results. Furthermore, in a companion paper, we demonstrate an application of this hybrid significance testing approach to reliably identify subtle differences in differential HX-MS measurements.

EXPERIMENTAL

Details for preparation of maltose-binding protein (MBP) and IgG1 monoclonal antibody (mAb), infliximab, are described in the Supporting Information.

Hydrogen exchange-mass spectrometry

Labeling was performed on a LEAP Technologies HDX PAL robot. MBP deuterated samples were prepared by diluting 3 µL of 8 µM MBP in 57 µL of labeling buffer (20 mM sodium phosphate, 100 mM sodium chloride, pD 7.0 in D_2O, pH was corrected for isotope effect[35]).

Samples were labeled seven times for each label time (30, 240, 1800, and 14400 s) at 25 °C. After labeling, 50 µL of each labeled sample was quenched with 50 µL of precooled quench buffer (200 mM sodium phosphate, pH 2.5 in water) at 1 °C. To minimize the influence of inter-day variations, all replicates of each individual HX labeling time were completed within a single day. Non-deuterated controls were prepared in a similar manner except with MBP buffer (20 mM sodium phosphate, 100 mM sodium chloride, pH 7.0 in water) substituted for labeling buffer. Following quenching, 75 µL of sample was injected into a temperature-controlled chromatography cabinet connected to an Agilent 1260 Infinity series LC. Cabinet temperature and equipped LC solvent pre-cooler were maintained at 0 °C for all experiments. Injected sample was passed over an immobilized pepsin column, prepared in house,[36] at 200 µL min^{-1} for 120 s with 0.1% formic acid in water. The resulting peptic peptides were captured on a C8 trap (Poroshell 120EC-C8 trap 2.1 x 5 mm, 2.7 µm particles) and washed at 200 µL/min for 60 s with 0.1% formic acid in water. Desalted peptic peptides were separated on a C18 column (ZORBAX RRHD 300SB-C18, 2.1 x 50 mm, 1.8 µm particles) with an 8-minute linear gradient of 0.1% formic acid in acetonitrile increasing from 15% to 45% acetonitrile. Peptide masses were subsequently measured with an Agilent 6530 Q-TOF mass spectrometer running in ESI-positive mode.

Deuterated mAb samples were prepared in a manner similar to MBP samples except for the following differences. MAb samples were labeled six times with labeling buffer (150 mM sucrose, 5 mM sodium phosphate, pD 7.2 in D_2O (pH corrected for isotope effect)[37]) for each label time (20, 100, 500, 2500, 12500, and 62500 s) at 20 °C. Samples were quenched with quench buffer (200 mM sodium phosphate, 3 M guanidine hydrochloride, 500 mM tris(2-carboxyethyl)phosphine hydrochloride, pH 2.4 in water). 80 µL of sample was injected and passed over the immobilized pepsin column for 180 s. Peptic peptides were captured on the C8 trap and washed for 90 s, then separated on the C18 column with a 12-minute linear gradient of 0.1% formic acid in acetonitrile increasing from 15% to 35% acetonitrile.

Hydrogen exchange-mass spectrometry data analysis

Peptic peptide databases were generated prior to HX-MS experiments with separate tandem mass spectrometry measurements of non-deuterated MBP and non-deuterated mAb samples. 115 peptic peptide (94% sequence coverage) assignments were confirmed for MBP. 112 heavy chain (88% sequence coverage) and 75 light chain (97% sequence coverage) peptic peptide assignments were confirmed for mAb. After automated HX-MS analysis in Sierra Analytics HDExaminer (versions 2.3 and 2.4), a single charge state that contained high quality spectra for all replicates across all HX labeling times was selected to represent HX for each peptide. Replicates with abnormally large deviations in HX values were individually evaluated; chromatographic peak integration limits were applied where necessary. The extent of HX based on the peptide centroid mass (m) for each peptide at each HX label time was exported to Microsoft Excel and Systat SigmaPlot for post-processing. For each peptide at each HX label time, mean centroid masses $(\bar{m})$ were calculated for sample sizes (n) corresponding to triplicate measurements. All sample standard deviations (s_m) presented within this work were calculated using equation (2).

$$s_m = \sqrt{\frac{\sum_{i=1}^{n}(m_i - \bar{m})^2}{n-1}} \qquad (2)$$

RESULTS AND DISCUSSION

Null measurements to evaluate significance testing approaches

During the HX-MS workflow, many sources can contribute to error including the sample, sample handling, mass measurement, and data processing.[36,38] Careful sample preparation complimented with comprehensive orthogonal characterization, automated sample handling, reliable mass spectrometers, and robust data processing software minimizes these sources of error. However, error still exists as a sum of all sources, some sources contributing more than others. Accurately estimated measurement error is essential for determining significance of observed HX differences by statistical testing. Inaccurate estimates of measurement error can impact type I or type II error rates. In this work, we designed null experiments to explicitly evaluate type I error rates from significance testing approaches used for determining significance in differential HX-MS data.

To generate a pool of experimental null data, seven intra-day HX-MS replicate measurements of MBP were acquired at four HX labeling times (see Supporting **Figure S1** for an outline of the workflow used to generate differential HX-MS data from our null measurements). The extent of HX (m), in Daltons (Da), was monitored for 115 MBP peptic peptides. Two unique triplicate data sets (a, b) were drawn at random from the pool of experimental null data, independent means were calculated $(\bar{m}_a, \bar{m}_b)$, and compared for HX differences $(\Delta\overline{HX} = \bar{m}_a - \bar{m}_b)$. In total, five comparisons were performed following this workflow. Each null experiment produced 460 $\Delta\overline{HX}$ values with 920 standard deviations (s_a, s_b) for the peptic peptides monitored at four HX labeling times. All experiments combined produced 2,300 $\Delta\overline{HX}$ values with 4,600 standard deviations. By design of the experiment, the theoretical value of $\Delta\overline{HX}$ is zero (i.e., the null hypothesis is true); thus, the distribution of $\Delta\overline{HX}$ values, from all comparisons, ranging from –0.091 to +0.084 Da, represents measurement error between the expected value and observed values (**Figure 1A**). Meanwhile the distribution of standard deviations, ranging from 0.000 to

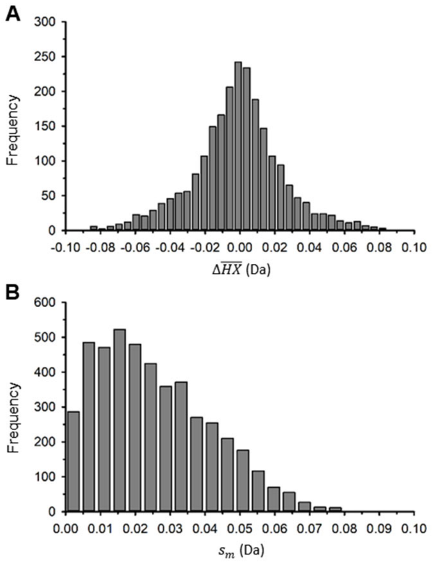


Figure 1. MBP null comparison distribution of 2,300 $\Delta\overline{HX}$ values (**A**) and 4,600 standard deviations (**B**).

0.080 Da, illustrates the variance in $\bar{m}$ values used to calculate $\Delta\overline{HX}$ values (**Figure 1B**). With our null comparison data, we scrutinized commonly used approaches for identifying meaningful differences in differential HX-MS data.

False positives using individual significance tests

One approach to identify meaningful differences in differential HX-MS data is significance testing of individual differences. Individual confidence intervals can be calculated for each $\Delta\overline{\mathrm{HX}}$ value by using corresponding standard deviation values and a critical t-value, from a Student's *t*-distribution, at a defined significance level (α) (see equation (1)). However, for large HX-MS data sets it is more convenient to interpret individual significance testing results with a probability value (p) that is related to the spread and location of the confidence interval. The probability can be determined using a two-sample Student's *t*-test. Welch's *t*-test[39] is a variation that does not require equal standard deviations in the two samples. It is suitable in cases where the standard deviations might differ substantially. To examine the rate of type I error for individual significance testing, we applied Welch's *t*-tests (two-tail) to our null comparison data. Any *p*-value, from the *t*-tests, less than the defined *α* is a type I error because the alternative hypothesis is false: there are no true differences between the samples. With *α* = 0.05, we observed many type I errors, shown in **Table 1**, with a mean false positive rate of 5.6% (126 type I errors from 2,300 *t*-tests). Upon increasing the stringency to *α* = 0.01, fewer type I errors were observed, as expected, with a mean false positive rate of 1.1% (25 type I errors from 2,300 *t*-tests). The Welch's *t*-test performed as expected based on the similarity between observed false positive rates and the probability of false

Table 1. Type I errors from MBP null comparisons with various approaches for determining significance at *α* = 0.05 and *α* = 0.01.

	Welch's *t*-test		**Bonferroni corrected**		**Global $\Delta\overline{HX}$ threshold**		**Hybrid significance test**	
	α = 0.05	*α* = 0.01	*α* = 0.05	*α* = 0.01	±0.067 Da	±0.110 Da	±0.067 Da *α* = 0.05	±0.110 Da *α* = 0.01
Null 1	24	4	0	0	5	0	1	0
Null 2	40	9	0	0	17	0	9	0
Null 3	6	0	0	0	1	0	0	0
Null 4	10	0	0	0	2	0	1	0
Null 5	49	12	1	0	22	0	7	0
Combined	129 (5.6%)	25 (1.1%)	1 (0.04%)	0 (0.0%)	47 (2.0%)	0 (0.0%)	18 (0.8%)	0 (0.0%)

positives from the α applied. However, in an actual differential HX-MS experiment, where many true differences might be present, it is doubtful that the probability of false positives, α, could be used to differentiate between a false positive from a true positive.

Upon detailed inspection of the type I errors observed, we found that many result from abnormally tight clustering of the data arising from the random nature of the measurements. For example, one specific type I error at $\alpha = 0.01$ from our null experiments has a seemingly negligible HX difference, $\Delta\overline{HX}$ = –0.009 Da but also even smaller standard deviations, ±0.002 Da, for both $\bar{m}$ values, that results in the classification of this miniscule HX difference as significant, a type I error. This event highlights a vulnerability of significance testing of individual differences: a small difference in means can be classified as significant if the estimate of the experimental error from replicate uncertainties is unreliable due to sampling statistics.

The overall frequency of type I errors depends on the number of significance tests performed, a problem known as the multiple comparisons problem.[40] For $w = 460$ (i.e., the number of t-tests performed in a single null experiment), the probability of having zero type I errors (i.e., one minus the family-wise error rate, $(1–\alpha)^w$) is 6E–11 when using $\alpha = 0.05$ and 0.0098 when using $\alpha = 0.01$. The median number of false positives can be estimated from the binomial distribution as αw (see Supporting **Figure S2**). On this basis, we would expect 23 type I errors at $\alpha = 0.05$ and 4 type I errors at $\alpha = 0.01$ for 460 comparisons, which is consistent with the mean number of type I errors observed across our five null experiments (see **Table 1**). Correction approaches, such as the Bonferroni correction,[41] are commonly used in statistical analyses to compensate for multiple comparisons and reduce probability of type I error. Implementing the Bonferroni correction (αw^{-1}) for a single null experiment, $w = 460$, results in corrected values for significance of 1.1E–4 and 2.2E–5 for desired $\alpha = 0.05$ and $\alpha = 0.01$, respectively. When Bonferroni-corrected significance limits are applied to all null comparisons, the number of observed type I errors is 1 for corrected $\alpha = 0.05$ and 0 for corrected $\alpha = 0.01$ (**Table 1**). To explore the impact of these corrections, we calculated theoretical $|\Delta\overline{HX}|$ values that would be required for a significant discovery when using Bonferroni-corrected significance values (see detailed calculation in caption of Supporting **Figure S3**). The resulting $|\Delta\overline{HX}|$ values required for a significant discovery are 10- to 100-fold larger than the distribution of actual $|\Delta\overline{HX}|$ values from null comparisons (see Supporting **Figure S3**). With a Bonferroni-corrected significance for α = 0.05, the mean of $|\Delta\overline{HX}|$ values required for a significant HX difference is 1.2 Da and it becomes 2.4 Da for a corrected significance for α = 0.01. In some cases, the differences would only be considered significant if they exceeded 4.0 Da for α = 0.05 or 10.0 Da for α = 0.01. The $\Delta\overline{HX}$ values required to achieve significance after applying a multiple comparisons correction are outrageous for individual HX times in differential HX-MS data, and in some cases might actually exceed the theoretical maximum deuteration. In a non-null experiment, applying Bonferroni correction for multiple comparisons will drastically increase the probability of type II error, i.e., false negatives. Thus, Bonferroni correction to reduce the probability of type I error at the expense of type II error is not suitable for typical differential HX-MS data where a large number of comparisons (>10^3) are often analyzed.

With our null experiment, we have demonstrated a major limitation for using individual significance tests to identify meaningful differences in differential HX-MS data: there will be false positives at a rate of approximately α. Although significance testing is a valuable approach to identify significant differences when comparing individual experimental means, the limitations for application to differential HX-MS data outweigh the value. The problem of multiple comparisons is inevitable for applications of individual significance tests to differential HX-MS data. Multiple comparison corrections can reduce the probability of type I error, but at the expense of losing power. Loss of power decreases the probability of finding small but significant differences.

A delicate balance of risk and power with a global $\Delta\overline{HX}$ significance threshold

A widely accepted approach to identify meaningful differences in differential HX-MS data is to apply a global $\Delta\overline{HX}$ threshold for significance that is based on an estimate of experimental error. With this threshold approach, the $\Delta\overline{HX}$ values from differential HX-MS data are compared to a defined $\Delta\overline{HX}$ threshold to determine significance. Defined thresholds are typically derived by taking the product of a probability distribution value (e.g., Student's t-distribution) and a globally estimated value to represent all experimental error. This approach relies on a global estimate of measurement error to determine significance that is intended to be representative of all of the data. A widely-reported, but perhaps not well-substantiated, figure-of-merit for a global $\Delta\overline{HX}$ threshold to determine significance is ±0.5 Da for analyses of single time point differential HX-MS data from triplicate measurements.[42] Applying the threshold H_1: $|\Delta\overline{HX}|$ > 0.5 Da to our distribution of $\Delta\overline{HX}$ values from null comparisons, we would not observe any type I errors, recalling that the range is –0.091 to +0.084 Da in **Figure 1A**. However, this $\Delta\overline{HX}$ threshold range is clearly a large overestimation of measurement error of our null data. Overestimating will certainly decrease the probability of type I error but at the expense of losing power, similar to a multiple comparisons correction. Alternatively, a $\Delta\overline{HX}$ threshold could be defined empirically from the largest difference observed in a null distribution, here ±0.091 Da, however, running a null experiment is not desirable for routine HX-MS experiments.

An approach to derive a global significance threshold, without the use of a null experiment, based on standard deviations from experimental data was proposed by Houde et al.[28] Following this approach and implementing some recently-proposed corrections,[28] we used our experimental standard deviations to calculate a $\Delta\overline{HX}$ significance threshold as detailed in the Supporting Information. We used a pooled standard deviation from our 4,600 standard deviations to estimate a global uncertainty of ±0.030 Da (s_p) in any $\bar{m}$ value. Using the pooled standard deviation for triplicate measurements (n_a, n_b) a propagated standard error of the mean (SEM) of ±0.024 Da for any $\Delta\overline{HX}$ value was calculated by equation (3).

$$SEM_{\Delta\overline{HX}} = \sqrt{\frac{s_p^2}{n_a} + \frac{s_p^2}{n_b}} \tag{3}$$

By selecting a suitable value of *k*, a confidence interval for any $\Delta\overline{HX}$ value can be calculated by equation (4).

$$H_1 : \left|\Delta\overline{HX}\right| > k \times SEM_{\Delta\overline{HX}} \tag{4}$$

Here we define *k* by a Student's *t*-distribution value ($t_{\alpha/2}$), for one-tail, with four degrees of freedom ($n_a + n_b - 2$). Using Student's *t*-distribution values of 2.7764 and 4.6041, corresponding to α = 0.05 and α = 0.01, resulted in $\Delta\overline{HX}$ threshold values for significance of ±0.067 Da and ±0.110 Da, respectively. These values are much smaller than the commonly reported ±0.5 Da threshold. This difference can be explained both by the narrow distribution of standard deviations in our null experiments and by recently-proposed corrections[43] to the method used to estimate the ±0.5 Da threshold.

The $\Delta\overline{HX}$ threshold of ±0.067 Da (corresponding to α = 0.05) underestimates the error in our null comparison data as is evident in a dot density plot of $\Delta\overline{HX}$ values from all null comparisons (**Figure 2** red dashed lines**)**. Using α = 0.05, 47 $\Delta\overline{HX}$ values exceed the $\Delta\overline{HX}$ significance threshold of ±0.067 Da, resulting in a false positive rate of 2.0% (**Table 1**). In contrast, with a significance threshold of ±0.110 Da, corresponding to α = 0.01, the significance limits closely bracket our distribution of $\Delta\overline{HX}$ values from null comparisons (**Figure 2** grey wide dashed lines). No $\Delta\overline{HX}$ values exceed the $\Delta\overline{HX}$ significance threshold value of ±0.110 Da, indicating a reasonable estimate for significance (**Table 1**). The false positive rates for α = 0.05 and α = 0.01 are much lower using the global $\Delta\overline{HX}$ thresholds for significance compared to the respective Welch's *t*-test false positive rates of 5.6% (α = 0.05) and 1.1% (α = 0.01). Furthermore, our estimated error at α = 0.01 yields an estimate that is reasonably close to the ±0.091 Da significance value we would obtain if we empirically determined a $\Delta\overline{HX}$ significance threshold from our maximum observed $|\Delta\overline{HX}|$ value in null comparisons. Importantly, the threshold estimated from equation (4) can be estimated by pooling the standard deviations for any differential HX-MS experiment, without using a null experiment.

Pooling the standard deviations to define a global significance threshold carries the assumption of equal experimental error for all measurements. To justify this assumption, experimental error must be independent of peptide and other experimental parameters. To determine if there was any correlation of experimental error with peptide and experimental parameters in our data, we determined the standard deviation using all seven replicates at each HX labeling time for each peptide, which produced 460 standard deviations. We did not observe strong evidence that supports the proposition that standard deviation is correlated to HX labeling time, retention time, and mass-to-charge ratio (see Supporting **Figure S4**). However, there is a weak positive correlation between standard deviation and charge, mass, theoretical maximum HX, percent HX, and magnitude HX (see Supporting **Figure S4**), which are all parameters related to the measured mass increase. From this, we can conclude that peptide size and isotopic distribution width are slightly correlated to standard deviation. However, the observed correlation is mostly driven by a tight cluster of data from small peptides. The deuterated isotopic distributions of the small peptides does not shift along the m/z axis. Instead, the distribution of peak intensities changes with increasing isotopic peak intensity and decreasing monoisotopic peak intensity. This observation of decreased experimental error in this scenario is consistent with previously reported

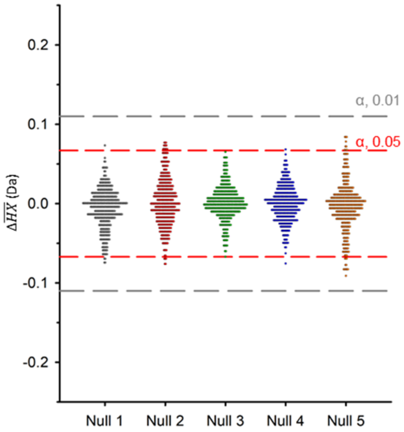


Figure 2. Dot density plot of $\Delta\overline{HX}$ values for each MBP null comparison (460 $\Delta\overline{HX}$ values each), with significance limits at α = 0.05 (red dashed lines) and α = 0.01 (grey wide dashed lines), calculated from the pooled standard deviation (s_p) of all 4,600 null standard deviations based on equation (4) with k = 2.7764 and 4.6041, as described in the text.

correlations by Houde et al.[43] and Schriemer et al.[28] Considering minimal correlation, our results indicate that the assumption of equal experimental error for all measurements is justified. Thus, our calculation to derive a global threshold based on pooled standard deviation is reasonable. A null experiment is not necessary to determine significance by reducing probability of type I error. Meanwhile, the estimated $\Delta\overline{HX}$ threshold potentially conserves power by not substantially overestimating the significance limits.

From our results, it is clear that a reliable $\Delta\overline{HX}$ threshold can be estimated from experimental data without using a null experiment and that this threshold, by itself, was sufficient to eliminate type I errors in our null MBP data. However, it is important to consider the good precision observed in our MBP null experiment. A more complex protein may show a broader distribution of precision that will challenge reliably estimating a $\Delta\overline{HX}$ threshold to determine significance. To investigate this possibility, we also performed a null experiment with a well-characterized, marketed IgG1 mAb therapeutic to evaluate the rate of type I error for a more complex protein sample. Following a workflow similar to the MBP null experiments, we compared two unique triplicate data sets consisting of 187 mAb peptic peptides in which HX was monitored at six HX labeling times. Unlike MBP, only a single null comparison was performed rather than five comparisons used with MBP. The null comparison produced 1122 $\Delta\overline{HX}$ values with 2244 standard deviations. The distribution of mAb standard deviations ranges from 0.000 to 0.307 Da (**Figure 3A**). The distribution of mAb standard deviations is noticeably wider than the distribution of null MBP standard deviations, that was 0.000 to 0.080 Da, in **Figure 1B**. The range of mAb $\Delta\overline{HX}$ values, –0.315 to +0.333 Da, in **Figure 3B** is also wider than the range of MBP $\Delta\overline{HX}$ values, –0.091 to +0.084 Da, in **Figure 1A**.

The pooled standard deviation, 0.063 Da, for the 2244 mAb standard deviations is greater than the MBP pooled standard deviation (0.030 Da). Therefore, the resulting mAb $\Delta\overline{HX}$ significance thresholds were greater as well with ±0.145 Da for α = 0.05 and ±0.240 Da for α =

0.01. We observe a slightly underestimated $\Delta\overline{HX}$ threshold when applying the ±0.145 Da

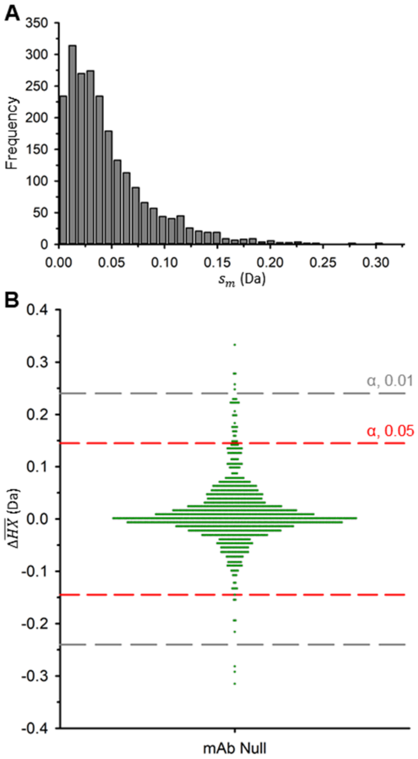


Figure 3. Distribution of 2244 standard deviations (**A**) and dot density plot of 1122 $\Delta\overline{HX}$ values (**B**) from all mAb null comparisons. Significance limits at *α* = 0.01 (grey wide dashed lines) and *α* = 0.05 (red dashed lines) calculated from the pooled standard deviation (s_p) of all mAb null standard deviations.

significance value (corresponding to *α* = 0.05) to our mAb null comparison data the dot density plot of $\Delta\overline{HX}$ values (**Figure 3B** red dashed lines**)**. At *α* = 0.05, 42 $\Delta\overline{HX}$ values exceed the $\Delta\overline{HX}$ significance threshold value of ±0.145 Da resulting in a false positive rate of 3.7% (see Supporting **Table S1**). Increasing stringency by applying the ±0.240 Da significance value (corresponding to *α* = 0.01), shown as grey wide dashed lines in **Figure 3B,** we observe 8 $\Delta\overline{HX}$ values that exceed the threshold resulting in a false positive rate of 0.7% (see Supporting **Table S1**). Unlike the MBP null results with no false positives for a $\Delta\overline{HX}$ significance threshold value corresponding to *α* = 0.01 (**Table 1**), the mAb null results do return false positives. False positives in the mAb data are not surprising since the calculated $\Delta\overline{HX}$ significance threshold value was less than the largest

$|\Delta\overline{HX}|$ values observed in mAb results. Thus, our mAb results highlight the challenge of estimating a global significance threshold for data with a wider distribution of uncertainty. In the case of the mAb results, with a wider distribution of standard deviations, the assumption of equal experimental error can increase the risk if the significance threshold is underestimated. Conversely, the assumption will potentially decrease power if the significance threshold is overestimated.

Hybrid significance testing minimizes type I and type II errors

To decrease probability of type I errors, an overestimated $\Delta\overline{HX}$ threshold for significance could be applied but at the cost of losing power by increasing probability of type II errors (i.e., false negatives). Alternatively, Welch's *t*-tests can be applied to gain power by reducing probability of type II error, but as we have shown, this increases risk of type I errors (i.e., false positives). A solution to manage both type I and type II error is significance testing using a hybrid approach. The significance testing results are easily visualized with a volcano plot, a scatter-plot of unstandardized signal (magnitude of change, $\Delta\overline{HX}$) versus noise-adjusted standardized signal (represented using the *p*-value from Welch's *t*-test), as shown in **Figure 4**. The fields of differential gene expression and metabolomics have widely adopted volcano plots to determine significance and visualize results.[44] Significance is tested using the relationship of the signals which determines significance based on $\Delta\overline{HX}$ versus individual replicate variance and an estimate of measurement error. Thus for the hybrid significance testing approach of HX-MS data, the first-pass significance test is whether the difference observed is greater than the globally-estimated measurement error (i.e., equation (4)). If the difference is greater than measurement error, then a Welch's *t*-test is performed to confirm significance based on individual standard deviations. By using these double-filtering criteria, large differences that are greater than measurement error but that have large standard deviations, are penalized by the Welch's *t*-test for the large error. Negligibly small HX differences that appear significant based on the *t*-test are penalized by the magnitude of the HX difference. Although volcano plot representations of HX-MS data have not been widely reported, Mass Spec Studio software[45] implements similar statistical significance testing using a Woods plot. Recently, a volcano plot feature, using calculations similar to those presented within this work, was implemented in HDExaminer (Sierra Analytics, Modesto, CA). However, to the best of our knowledge no reports in the HX-MS field have validated this type of approach against a true null data set.

A volcano plot representation of the MBP null comparison hybrid significance testing results is shown in **Figure 4A**. Observed $\Delta\overline{HX}$ values are plotted on the horizontal axis and *p*-values from Welch's *t*-tests are plotted on the vertical axis. The horizontal significance limit line at 10^{-2} represents α = 0.01 for the *p*-values from the Welch's *t*-tests. The $\Delta\overline{HX}$ significance limit, the vertical lines, represent an estimate of global uncertainty at α = 0.01. In **Figure 4A,** we have

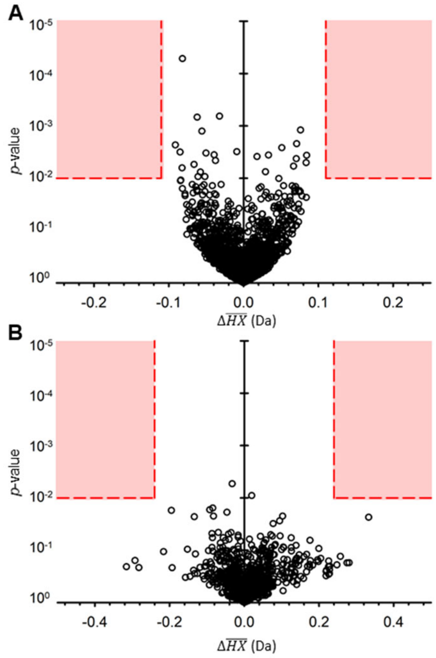


Figure 4. Volcano plot of MBP (**A**) and mAb (**B**) null comparisons. Observed $\Delta\overline{HX}$ values (horizontal axis) and Welch's *t*-test *p*-values (vertical axis). Horizontal *p*-value significance limits are defined at α = 0.01 and vertical significance limits are defined at ±0.110 Da (MBP) and ±0.240 (mAb) from s_p calculated $\Delta\overline{HX}$ significance threshold representative of α = 0.01.

defined this limit as ±0.110 Da which is the global $\Delta\overline{HX}$ significance threshold we calculated corresponding to $\alpha = 0.01$ (vide supra). Any $\Delta\overline{HX}$ value and *p*-value exceeding significance limits in both dimensions (within the red shaded regions) is classified as significant. Any significant result from our null data would be a type I error, but we do not observe any type I errors for MBP results with the hybrid significance criteria defined at $\alpha = 0.01$ (**Table 1**). This was expected because all null comparison results are less than the defined ±0.110 Da significance threshold. To compare the type I error rate of the hybrid significance test versus individual testing approaches, we decreased the stringency by applying significance limits corresponding to $\alpha =$ 0.05 for MBP null comparisons. By lowering the stringency, we decrease the $\Delta\overline{HX}$ significance limit to ±0.067 Da in which 47 type I errors were observed (**Table 1**). By applying this $\Delta\overline{HX}$ threshold and a *p*-value significance limit of $\alpha = 0.05$ for hybrid significance testing, we observe 18 type I errors (**Table 1**). The type I error rate for $\alpha = 0.05$ hybrid significance testing (18 type I errors) is less than the type I error rates observed for both the approach using individual Welch's *t*-tests (129 type I errors) and the approach using a global $\Delta\overline{HX}$ threshold (47 type I errors). These results indicate that for our MBP null results, the probability of type I error with the hybrid significance test is lower than with either individual significance tests and global significance tests.

We observe a similar outcome, reduced type I error, when applying hybrid significance testing to mAb null data. MAb null comparison hybrid significance testing results are illustrated in **Figure 4B** with a volcano plot. The horizontal significance limit line at 10^{-2} represents $\alpha = 0.01$ for the *p*-values from the Welch's *t*-tests. The vertical $\Delta\overline{HX}$ significance limit lines are defined at ±0.240 Da corresponding to $\alpha = 0.01$ (vide supra). There are not any type I errors for mAb results with the double-filtering criteria defined at $\alpha = 0.01$ (see Supporting **Table S1**). In contrast to MBP, there are 8 $|\Delta\overline{HX}|$ values from mAb null comparisons that are greater than the ±0.240 Da ($\alpha =$ 0.01) significance threshold. The type I error rate for $\alpha = 0.01$ hybrid significance testing (0 type I errors) is less than the type I error rates observed for global $\Delta\overline{HX}$ threshold (8 type I errors) (see

Supporting **Table S1**). In addition, the type I error rate for α = 0.05 hybrid significance testing (2 type I errors) is less than the type I error rates observed for global $\Delta\overline{HX}$ threshold (42 type I errors). The larger $|\Delta\overline{HX}|$ values that appear to be significant based on the global $\Delta\overline{HX}$ significance threshold are penalized by the Welch's *t*-test because of their large standard deviations. For both MBP and mAb null experiments, the hybrid significance test returned lower type I error rates than individual significance tests.

The advantage of this hybrid significance testing approach, compared to individual and global significance testing approaches, is that individual replicate error is considered for measurement *reliability* while simultaneously considering the magnitude of HX difference for measurement *plausibility*. In addition, statistical testing results are easily visualized in the volcano plot to determine if statistically significant differences are present. In a companion paper we demonstrate that this hybrid approach can reliably identify subtle differences that challenge the detection limit of HX-MS. It is important to stress that significant HX differences identified using this approach should be scrutinized by the HX-MS analyst. Validation of significant differences should consist of confirming similar trends in overlapping peptides, identifying differences at multiple HX labeling times, reviewing the quality of raw spectra, and inspecting chromatographic peak integration limits for discrepancies. (We have found that peak integration errors are a large source of variance, results not shown.) Expert review is also needed for results with missing technical replicate data or abnormally large standard deviations. The reliability of peptides with missing replicate HX data or missing data at specific HX labeling times should be carefully considered. Missing replicate data (e.g., *n* = 2) will substantially increase the critical *t*-values in the Welch's *t*-test making it more difficult for large differences to be classified as significant in the *p*-value dimension. Also, a key component of the $|\Delta\overline{HX}|$ threshold calculation for the hybrid significance testing approach, detailed within this work, is the number of technical replicates is equal for all sample means. It is also important to emphasize that the $|\Delta\overline{HX}|$ threshold for this approach should be calculated from experimental standard deviations for each experiment. As our mAb results compared to MBP showed, HX-MS measurements of more complex proteins can result in larger measurement error that will alter the calculated $|\Delta\overline{HX}|$ threshold. Another important consideration is the experimental timeframe. In inter-day experiments slight deviations in experimental conditions could affect the observed measurement error.[46] Although a pooled standard deviation can be calculated to estimate a global $|\Delta\overline{HX}|$ threshold for any experiment without the use of a null experiment, a null experiment is valuable to evaluate measurement error and criteria used for classifying significant differences. The null experiment approach, demonstrated here, to evaluate statistical significance should be extended to the use of differential HX-MS measurements for higher order structural comparability applications. The present work has focused exclusively on treating HX labeling times as discrete experiments. Significance testing using HX-MS data that has been integrated[32] or averaged[33,34] across all HX times might also be useful, though we note that those differences could become diluted by averaging them into a set of data where other HX times exhibit no measurable differences.

ACKNOLEDGEMENTS

We thank Agilent Technologies for an instrument loan, Jeff Morrow for HDExaminer software support, Anna Schwendeman for gifting the mAb protein, and the National Institutes of Health NIGMS Biotechnology Predoctoral Training Program (T32-GM008359) and the National Science Foundation (CHE-1709176) for financial support.

REFERENCES

(1) Pirrone, G. F.; Iacob, R. E.; Engen, J. R. *Anal Chem* **2015**, *87*, 99-118. DOI: 10.1021/ac5040242.

(2) Houde, D.; Arndt, J.; Domeier, W.; Berkowitz, S.; Engen, J. R. *Anal Chem* **2009**, *81*, 2644-2651. DOI: 10.1021/ac802575y.

(3) Engen, J. R. *Anal Chem* **2009**, *81*, 7870-7875.

(4) D'Arcy, S.; Martin, K. W.; Panchenko, T.; Chen, X.; Bergeron, S.; Stargell, L. A.; Black, B. E.; Luger, K. *Mol Cell* **2013**, *51*, 662-677. DOI: 10.1016/j.molcel.2013.07.015.

(5) Smith, B. C.; Underbakke, E. S.; Kulp, D. W.; Schief, W. R.; Marletta, M. A. *Proc Nat Acad Sci USA* **2013**, *110*, E3577-E3586. DOI: 10.1073/pnas.1313331110.

(6) Burke, J. E.; Williams, R. L. *Advances in Biological Regulation* **2013**, *53*, 97-110. DOI: 10.1016/j.jbior.2012.09.005.

(7) Iacob, R. E.; Krystek, S. R.; Huang, R. Y. C.; Wei, H.; Tao, L.; Lin, Z.; Morin, P. E.; Doyle, M. L.; Tymiak, A. A.; Engen, J. R.; Chen, G. *Expert Rev Proteom* **2015**, *12*, 159-169. DOI: 10.1586/14789450.2015.1018897.

(8) Hamuro, Y.; Coales, S. J.; Morrow, J. A.; Molnar, K. S.; Tuske, S. J.; Southern, M. R.; Griffin, P. R. *Protein Sci* **2006**, *15*, 1883-1892. DOI: 10.1110/ps.062103006.

(9) Dai, S. Y.; Burris, T. P.; Dodge, J. A.; Montrose-Rafizadeh, C.; Wang, Y.; Pascal, B. D.; Chalmers, M. J.; Griffin, P. R. *Biochemistry* **2009**, *48*, 9668-9676. DOI: 10.1021/bi901149t.

(10) Bennett, M. J.; Barakat, K.; Huzil, J. T.; Tuszynski, J.; Schriemer, D. C. *Chem Biol* **2010**, *17*, 725-734. DOI: 10.1016/j.chembiol.2010.05.019.

(11) Chalmers, M. J.; Wang, Y.; Novick, S.; Sato, M.; Bryant, H. U.; Montrose-Rafizdeh, C.; Griffin, P. R.; Dodge, J. A. *ACS Med Chem Lett* **2012**, *3*, 207-210. DOI: 10.1021/ml2002532.

(12) Landgraf, R. R.; Goswami, D.; Rajamohan, F.; Harris, M. S.; Calabrese, M. F.; Hoth, L. R.; Magyar, R.; Pascal, B. D.; Chalmers, M. J.; Busby, S. A.; Kurumbail, R. G.; Griffin, P. R. *Structure* **2013**, *21*, 1942-1953. DOI: 10.1016/j.str.2013.08.023.

(13) Zhang, A.; Singh, S. K.; Shirts, M. R.; Kumar, S.; Fernandez, E. J. *Pharm Res* **2012**, *29*, 236-250. DOI: 10.1007/s11095-011-0538-y.

(14) Xiao, Y.; Konermann, L. *Protein Sci* **2015**, *24*, 1247-1256. DOI: 10.1002/pro.2680.

(15) Bommana, R.; Chai, Q.; Schoneich, C.; Weiss, W. F., IV; Majumdar, R. *J Pharm Sci* **2018**, *107*, 1498-1511. DOI: 10.1016/j.xphs.2018.01.017.

(16) Houde, D.; Peng, Y.; Berkowitz, S. A.; Engen, J. R. *Mol Cell Proteomics* **2010**, *9*, 1716-1728. DOI: 10.1074/mcp.M900540-MCP200.

(17) Zhang, A.; Hu, P.; MacGregor, P.; Xue, Y.; Fan, H.; Suchecki, P.; Olszewski, L.; Liu, A. *Anal Chem* **2014**, *86*, 3468-3475. DOI: 10.1021/ac404130a.

(18) Yan, Y.; Wei, H.; Fu, Y.; Jusuf, S.; Zeng, M.; Ludwig, R.; Krystek, S. R., Jr.; Chen, G.; Tao, L.; Das, T. K. *Anal Chem* **2016**, *88*, 2041-2050. DOI: 10.1021/acs.analchem.5b02800.

(19) More, A. S.; Toth, R. T.; Okbazghi, S. Z.; Middaugh, C. R.; Joshi, S. B.; Tolbert, T. J.; Volkin, D. B.; Weis, D. D. *J Pharm Sci* **2018**, *107*, 2315-2324. DOI: 10.1016/j.xphs.2018.04.026.

(20) Houde, D.; Berkowitz, S. A. *J Pharm Sci* **2012**, *101*, 1688-1700. DOI: 10.1002/jps.23064.

(21) Visser, J.; Feuerstein, I.; Stangler, T.; Schmiederer, T.; Fritsch, C.; Schiestl, M. *Biodrugs* **2013**, *27*, 495-507. DOI: 10.1007/s40259-013-0036-3.

(22) Fang, J.; Doneanu, C.; Alley, W. R., Jr.; Yu, Y. Q.; Beck, A.; Chen, W. *MAbs* **2016**, *8*, 1021-1034. DOI: 10.1080/19420862.2016.1193661.

(23) Hong, J.; Lee, Y.; Lee, C.; Eo, S.; Kim, S.; Lee, N.; Park, J.; Park, S.; Seo, D.; Jeong, M.; Lee, Y.; Yeon, S.; Bou-Assaf, G.; Sosic, Z.; Zhang, W.; Jaquez, O. *MAbs* **2017**, *9*, 364-382. DOI: 10.1080/19420862.2016.1264550.

(24) Wales, T. E.; Engen, J. R. *Mass Spectrom Rev* **2005**, *25*, 158-170. DOI: 10.1002/mas.20064.

(25) Malito, E.; Faleri, A.; Lo Surdo, P.; Veggi, D.; Maruggi, G.; Grassi, E.; Cartocci, E.; Bertoldi, I.; Genovese, A.; Santini, L.; Romagnoli, G.; Borgogni, E.; Brier, S.; Lo Passo, C.; Domina, M.; Castellino, F.; Felici, F.; van der Veen, S.; Johnson, S.; Lea, S. M., et al. *Proc Nat Acad Sci USA* **2013**, *110*, 3304-3309. DOI: 10.1073/pnas.1222845110.

(26) Majumdar, R.; Manikwar, P.; Hickey, J. M.; Samra, H. S.; Sathish, H. A.; Bishop, S. M.; Middaugh, C. R.; Volkin, D. B.; Weis, D. D. *Biochemistry* **2013**, *52*, 3376-3389. DOI: 10.1021/bi400232p.

(27) Chalmers, M. J.; Busby, S. A.; Pascal, B. D.; Southern, M. R.; Griffin, P. R. *J Biomolec Techniques* **2007**, *18*, 194-204.

(28) Houde, D.; Berkowitz, S. A.; Engen, J. R. *J Pharm Sci* **2011**, *100*, 2071-2086. DOI: 10.1002/jps.22432.

(29) Wei, H.; Ahn, J.; Yu, Y. Q.; Tymiak, A.; Engen, J. R.; Chen, G. *J Am Soc Mass Spectrom* **2012**, *23*, 498-504. DOI: 10.1007/s13361-011-0310-x.

(30) Iacob, R. E.; Bou-Assaf, G. M.; Makowski, L.; Engen, J. R.; Berkowitz, S. A.; Houde, D. *J Pharm Sci* **2013**, *102*, 4315-4329. DOI: 10.1002/jps.23754.

(31) Leurs, U.; Lohse, B.; Ming, S.; Cole, P. A.; Clausen, R. P.; Kristensen, J. L.; Rand, K. D. *Anal Chem* **2014**, *86*, 11734-11741. DOI: 10.1021/ac503137u.

(32) Mazur, S. J.; Weber, D. P. *J Am Soc Mass Spectrom* **2017**, *28*, 978-981. DOI: 10.1007/s13361-017-1615-1.

(33) Chalmers, M.; Pascal, B.; Willis, S.; Zhang, J.; Iturria, S.; Dodge, J.; Griffin, P. *Int J Mass spectrom* **2011**, *302*, 59-68.

(34) Chalmers, M. J.; Busby, S. A.; Pascal, B. D.; West, G. M.; Griffin, P. R. *Expert Review of Proteomics* **2011**, *8*, 43-59. DOI: 10.1586/epr.10.109.

(35) Eisinger, M. L.; Dörrbaum, A. R.; Michel, H.; Padan, E.; Langer, J. D. *Proc Nat Acad Sci USA* **2017**, *114*, 11691.

(36) Glasoe, P. K.; Long, F. A. *J Phys Chem* **1960**, *64*, 188-190. DOI: 10.1021/j100830a521.

(37) Busby, S. A.; Chalmers, M. J.; Griffin, P. R. *Int J Mass spectrom* **2007**, *259*, 130-139. DOI: 10.1016/j.ijms.2006.08.006.

(38) Iacob, R. E.; Engen, J. R. *J Am Soc Mass Spectrom* **2012**, *23*, 1003-1010. DOI: 10.1007/s13361-012-0377-z.

(39) Engen, J. R.; Wales, T. E. *Annu Rev Anal Chem* **2015**, *8*, 127-148. DOI: 10.1146/annurev-anchem-062011-143113.

(40) Welch, B. L. *Biometrika* **1947**, *34*, 28-35. DOI: 10.1093/biomet/34.1-2.28.

(41) Tamhane, A. C. In *Handbook of statistics*; Elsevier, 1996, pp 587-630.

(42) Dudoit, S.; Shaffer, J. P.; Boldrick, J. C. *Stat Sci* **2003**, *18*, 71-103. DOI: 10.1214/ss/1056397487.

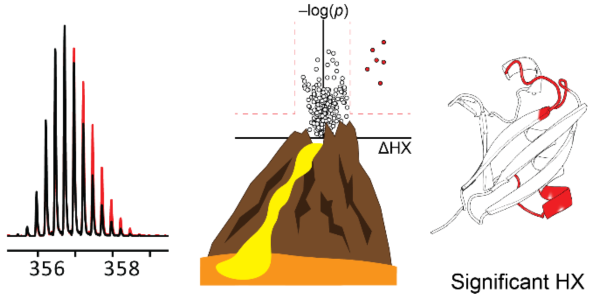
(43) Weis, D. D. *J Pharm Sci* **2019**, *108*, 807-810. DOI: 10.1016/j.xphs.2018.10.010.

(44) Slysz, G. W.; Percy, A. J.; Schriemer, D. C. *Anal Chem* **2008**, *80*, 7004-7011. DOI: 10.1021/ac800897q.

(45) Li, W. *J Bioinf Comp Biol* **2012**, *10*, 1231003. DOI: 10.1142/S0219720012310038.

(46) Rey, M.; Sarpe, V.; Burns, K. M.; Buse, J.; Baker, C. A. H.; van Dijk, M.; Wordeman, L.; Bonvin, A. M. J. J.; Schriemer, D. C. *Structure* **2014**, *22*, 1538-1548. DOI: 10.1016/j.str.2014.08.013.

TOC FIGURE



Differential HX-MS

Volcano Plot Analysis

Significant HX Differences