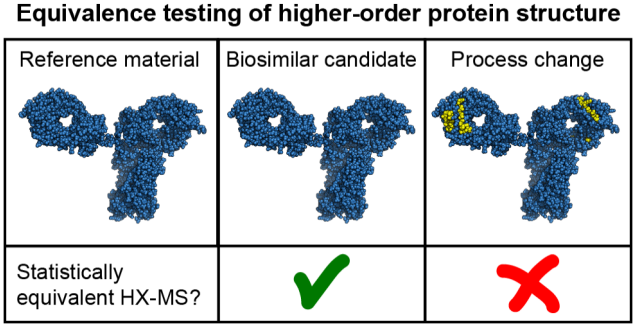
Statistical equivalence testing of higher-order protein structures with differential hydrogen exchange-mass spectrometry (HX-MS)

Tyler S. Hageman[†], Michael S. Wrigley, David D. Weis[*]
Department of Chemistry
The University of Kansas,
1567 Irving Hill Road, Lawrence KS, 66045 USA
[†]Present address Abbvie Bioresearch Center,
100 Research Drive, Worcester, MA, USA
*Correspondence to dweis@ku.edu, +1-785-864-1377

Accepted for publication in *Analytical Chemistry*, 7 May 2021



ABSTRACT

Hydrogen exchange-mass spectrometry (HX-MS) is widely recognized for its potential utility for establishing the equivalence of the higher-order structures of proteins, particularly in comparability and similarity contexts. However, recent progress in the statistical analysis of HX-MS data has instead placed an emphasis on significance testing to identify regions of proteins where there are significant differences in HX between two or more protein states. In the cases involving assessment of similarity or equivalence of the higher-order structure of different protein samples (e.g., biosimilars), significance testing of HX-MS data is unsuitable. To meet this need, we have adapted the univariate two-one sided tests (TOST) equivalence testing method to HX-MS data. Equivalence acceptance criteria were determined using maximum deviations from randomized resampling of truly equivalent samples to define hybrid equivalence criteria (maximum deviation of true equivalents, MDTE). Application of the TOST-MDTE test on differential HX-MS measurements of wild-type and mutated maltose binding protein demonstrate that the equivalence testing method was fit-for-purpose. Three infliximab biosimilars (Remsima, Renflexis, and Inflectra) were found to be equivalent to their Remicade reference product based on differential HX-MS measurements while 5% deglycosylated NIST mAb was not statistically equivalent to the unmodified NIST mAb reference.

INTRODUCTION

Establishing comparability or similarity of the higher-order structure of protein therapeutics to their reference products is an essential procedure in the pharmaceutical industry.[1, 2] Changes or differences in production or manufacturing process could impact higher-order structure of the drug substance.[3] Thus, comparability studies are needed for ongoing surveillance of product quality and to establish comparability to a reference product throughout product life-cycle and also following production-related changes (e.g., scale-up, fill and finishing procedures, cell line changes, facility changes). In this context, the objective is to establish equivalence of the same product from different preparations. For biosimilar proteins, drug manufacturers must reverse-engineer the innovator's proprietary process using only the innovator's reference product. As a result, there could be differences in the manufacturing process that are related to cell-line, growth conditions, and purification procedures.[4] These differences could potentially alter the higher-order structure of the protein product which could in turn lead to undesirable changes in stability and activity profiles.[5] In a similarity context, the objective is to demonstrate that a candidate biotherapeutic protein, developed by an independent process, is highly similar to the reference product. Evaluating comparability or demonstrating similarity of higher-order structure by analytical methods is used widely.

Hydrogen exchange-mass spectrometry (HX-MS) has proven to be a valuable method for analysis of higher-order structure of protein therapeutics. Recently, analysis of protein therapeutics by HX-MS has been included in US FDA biologics licensing applications.[6-10] The versatility and structural resolution of HX-MS, compared to other structural analysis methods, makes it a well-suited method for comparability and similarity studies of protein therapeutics. Previously, Houde et al. outlined the utility of differential HX-MS measurements for higher-order structural comparability studies and performed HX-MS comparability studies with different preparations of interferon-β-1a.[11] Since then, a number of comparability studies of protein therapeutics with HX-MS have been

reported.[12-14] Similarity studies of biosimilars and reference products with HX-MS have also been reported.[7, 15-18] Recently, Fang et al. reported a similarity study of an infliximab biosimilar (Inflectra) and the reference infliximab (Remicade) in which no statistically significant differences were observed in differential HX-MS measurements.[15] However, consistent but statistically insignificant differences in two segments of the infliximab Fc domain that are proximal to the N-linked glycans were apparent. Additionally, differences in the infliximab glycosylation profiles were observed in N-glycan release analysis suggesting that the glycosylation differences might be correlated with the subtle differences observed in HX-MS. Interestingly, one of the segments, heavy-chain residues 244-255 (FLFPPKPKDTLM), identified by HX-MS, has previously been shown to be sensitive to differences in glycosylation by HX-MS.[19-22] Furthermore, other experimental evidence suggests that this particular segment is an aggregation hotspot motif in IgG1 mAbs.[23] Thus, the subtle differences that Fang et al.[15] observed are likely to be consequential even if by HX-MS the differences appeared to be statistically insignificant. Such a situation might indicate that the criteria used to determine statistical significance were overestimated.

Importantly however, even with an appropriately sensitive HX-MS significance testing approach, the absence of significant differences does not establish statistical equivalence. In traditional statistical testing (e.g., Student's *t*-test for comparison of means) one seeks evidence to falsify a null hypothesis of equivalence (H_0). Usually the reliability of this difference is expressed in terms of a probability that the null hypothesis is true. A difference is deemed to be statistically significant if the estimated probability of H_0 is sufficiently low (e.g., $p < 0.05$, or another limit). The nature of the test however is such that the null hypothesis of equivalence cannot be proven.[24] In the case of HX-MS data, significance testing focuses on the differences in HX labeling between two samples, in this case a test sample and a reference sample:

$$\Delta D = m - m_{\text{ref}} \tag{1}$$

where m denotes the centroid mass of an HX-labeled peptide and the subscript "ref" specifies the reference sample. Hence, the null hypothesis would be $H_0 : \Delta D = 0$. Within the context of comparability or similarity, at best one can declare that "there is insufficient evidence to reject H_0." Establishment of comparability or similarity by statistical methods requires a test of statistical equivalence. In such a test, the goal is to falsify a null hypothesis of *non-equivalence*, $H_0 : \Delta D \neq 0$.[25]

Equivalence testing of HX-MS data is at the confluence of several challenging areas of statistical research. HX-MS data sets have high dimensionality in the number of individual HX measurements, but low dimensionality in the number of replicates (*i.e.,* the so-called "large p, small n problem").[26] An additional challenge arises because HX progressively increases with time and overlapping peptides report HX in overlapping regions, leading to complex covariance patterns in the data. Finally, it is widely recognized that traditional univariate approaches to data of this kind are problematic because of the multiple comparisons problem.[27] In the context of pharmaceutical product quality, controlling for multiple comparisons with false discovery methods[28] appears to be untenable from a regulatory perspective. To circumvent all of these challenges, we have developed an empirical hybrid equivalence testing approach that adapts the univariate two one-sided *t*-tests (TOST) method to multivariate data and that also applies a

maximum limit to the difference in means. The equivalence acceptance criteria were determined from the maximum deviations observed in randomized, replicate measurements drawn from a pool of measurements made on a single sample (maximum deviation of true equivalents, MDTE). Since all measurements were made using the same sample, these measurements represent true equivalence—the alternative hypothesis of equivalence testing (H_1) is true. We demonstrate that three approved infliximab biosimilars have HX-MS data that are equivalent to their reference product, Remicade. Furthermore, we establish, using partially deglycosylated NIST mAb, that our acceptance criteria, defined from true equivalence measurements, are sufficiently stringent to reject equivalence between truly non-equivalent samples with subtle differences that approach the HX-MS detection limit.

EQUIVALENCE TESTING HX-MS DATA WITH TOST

The TOST can be used to establish if samples are statistically equivalent.[29] Rejection of the null hypothesis in the TOST requires that ΔD must be significantly greater than a lower limit and simultaneously significantly less than an upper limit. Each of these represents a one-tailed *t*-test. The limit is known as the equivalence acceptance criterion (EAC). To test the alternate hypothesis, the TOST entails two hypothesis tests:

$$H_1: \begin{cases} \Delta D - t(\alpha, df)\sqrt{\frac{s^2}{n} + \frac{s_{\text{ref}}^2}{n_{\text{ref}}}} > -\text{EAC} \\ \Delta D + t(\alpha, df)\sqrt{\frac{s^2}{n} + \frac{s_{\text{ref}}^2}{n_{\text{ref}}}} < +\text{EAC} \end{cases} \tag{2}$$

where n denotes the number of replicate HX measurements, s is the standard deviation, and df denotes the degrees of freedom. The probability (i.e., $1-2\alpha$) that the true difference value is bounded by ±EAC is defined by the significance level, α, at which the critical *t*-value is selected. It is obvious by inspection that this statement of the alternative hypothesis is identical to requiring that the entire 1–2α confidence interval (CI) is bounded by ±EAC.[29, 30]

$$H_1: \begin{cases} \Delta D - \text{CI} > -\text{EAC} \\ \Delta D + \text{CI} < +\text{EAC} \end{cases} \tag{3}$$

Samples having a confidence interval that is bounded by the equivalence limits would be classified as statistically equivalent (i.e., reject H_0 in favor of H_1) while samples having a confidence interval that exceeds either or both the –EAC or +EAC would not be classified as statistically equivalent (i.e., fail to reject H_0). In the latter case, the samples have not been proven to be statistically different, instead the result would be interpreted as insufficient evidence to establish equivalence. This emphasizes the point that the burden of proof for equivalence is placed upon the analytical data: data with wide variance cannot be bounded by stringent EAC limits.

To highlight the differing outcomes from statistical difference and statistical equivalence testing, five scenarios (A-E) are shown in Figure 1.[29] Scenario A is a straightforward outcome: the samples are not equivalent by equivalence testing and are different by significance testing. Scenario B is the alternative straightforward outcome: equivalent and not significantly different from zero. There are three seemingly contradictory results, scenarios C-E. In C and D, neither equivalence nor

difference are established, while in E *both* equivalence and difference are established. There is an important distinction between scenarios B and D, which both have confidence intervals that span zero. While the samples in B are statistically equivalent, the poor precision of the measurements in D prevents establishment of equivalence. These examples highlight the inherent asymmetry in these tests.

A pre-defined EAC is required to test for statistical equivalence by TOST. Various recommendations have been made for establishing the EAC limits including using prior knowledge of the analysis and using limits established by subject matter experts.[29, 31] By design, the TOST for statistical equivalence considers the magnitude of ΔD and its uncertainty (i.e., the confidence interval). Therefore, prior knowledge of the variability in measurements is necessary to determine an appropriate EAC. As the EAC limits are expanded, the probability of committing a type I error (i.e., declaring a false positive or false equivalence of truly non-equivalent samples[32]) will increase. As the EAC limits are narrowed, the probability of committing a type II error, a false negative or a failure to establish equivalence in truly equivalent samples, will increase. In this work we illustrate how HX-MS data from truly equivalent samples can be used to establish EAC.

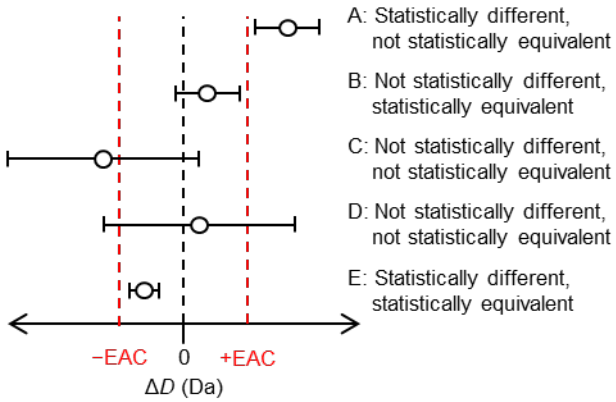


Figure 1. Examples of statistical difference and statistical equivalence testing outcomes (adjacent to each plotted value labeled as scenarios A-E) for ΔD values and TOST EAC limits. The symbols denote mean values and the bars are the confidence intervals. Adapted with permission from reference 29 (American Chemical Society, 2005).

EXPERIMENTAL

Details of sample preparation and HX-MS measurements are provided in the Supporting Information. To allow access to the HX-MS data of this study, HX-MS data summary tables (Tables S1-S2) and the HX-MS data (supporting files) are included in the Supporting Information according to established consensus guidelines.[33]

Randomized resampling of truly equivalent data

Previously published differential HX-MS data from measurements of maltose binding protein (MBP)[34] were used as a truly equivalent dataset to evaluate the criteria for statistical equivalence. Two sets of triplicate measurements were randomly resampled five times without replacement from a pool of seven intra-day HX-MS replicate measurements at each of four HX labeling times.

Independent mean peptide centroid masses were calculated and compared for HX differences. The extent of HX was measured as peptide centroid masses for 115 peptic peptides. In total, the resampled data set was comprised of 2,300 ΔD values with 4,600 standard deviations. A pool of truly equivalent HX-MS data was obtained from eight HX-MS replicate measurements of Remicade measured at six HX labeling times. The extent of HX was measured in 187 peptic peptides, 112 from the heavy chain and 75 from the light chain. A total of five comparisons were performed using two different randomly selected unique triplicate data sets in each. In total, 5,610 ΔD values with 11,220 standard deviations were pooled.

Statistical analysis

The extent of deuteration, 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 $(\overline{m})$ were calculated for sample sizes (n) corresponding to triplicate measurements. Sample standard deviations were calculated as

$$s = \sqrt{\frac{\sum_{i=1}^{n}(m_i - \overline{m})^2}{n-1}} \tag{4}$$

Confidence intervals (CI) were calculated as

$$\mathrm{CI} = \Delta D \pm t\left(\alpha, df\right)\sqrt{\frac{s^2}{n} + \frac{s_{\mathrm{ref}}^2}{n_{\mathrm{ref}}}} \tag{5}$$

Student's *t*-distribution (t) values were determined at a significance level (α) of 0.10, for a two-tailed distribution, with degrees of freedom (df) that were determined using the Satterthwaite approximation:[35]

$$df = \frac{\left(\frac{s^2}{n} + \frac{s_{\mathrm{ref}}^2}{n_{\mathrm{ref}}}\right)^2}{\frac{1}{n-1}\left(\frac{s^2}{n}\right)^2 + \frac{1}{n_{\mathrm{ref}}-1}\left(\frac{s_{\mathrm{ref}}^2}{n_{\mathrm{ref}}}\right)^2} \tag{6}$$

RESULTS

Determination of EAC with resampled replicate measurements of equivalent samples

Our approach for determination of EAC for TOST testing uses truly equivalent samples, replicate HX-MS measurements of a single MBP sample,[34] to empirically define a minimum EAC for which all HX differences are statistically equivalent. HX measurements from a pool of seven intraday replicate measurements of HX were randomly selected in groups of three without replacement into the sample and the reference; the data were resampled five times ($\frac{1}{2}\mathrm{C}_3^7\mathrm{C}_3^4 = 70$ unique re-

samplings are possible). The resulting ΔD_{eq} values and 90% confidence intervals for all truly equivalent MBP data are shown in Figure 2 (the subscript eq denotes measurements of truly equivalent samples). The false non-equivalence rate (type II error) can be driven to zero by empirically defining the EAC using the extrema of the confidence intervals (solid gray lines in Figure 2) that represent the maximum deviations in true equivalents (MDTE):

$$\mathrm{EAC} = \max\left(\left|\min\left(\mathbf{\Delta D}_{\mathrm{eq}} - \mathbf{CI}\right)\right|, \left|\max\left(\mathbf{\Delta D}_{\mathrm{eq}} + \mathbf{CI}\right)\right|\right) \tag{7}$$

where the bold Roman type denotes vectors (i.e., columns) of all differential HX-MS values (equation (1)) and their corresponding confidence intervals (equation (5)). For resampled truly equivalent MBP, EAC = 0.160 Da (see Table 1).

Table 1. Equivalence criteria established with HX-MS measurements of truly equivalent MBP and Remicade samples.

	MBP	Remicade
EAC (Da)	0.160 Da	0.842 Da
ΔD_{max} (Da)	0.091 Da	0.367 Da
Number of HX measurements	2,300	5,610
Pooled standard deviation	0.030 Da	0.096 Da

Reducing false equivalence errors by including a difference in means threshold

If the EAC were overestimated (i.e., overly wide limits), then false equivalence might be established in truly non-equivalent samples (i.e., false positive or type I error). To determine if the EAC is overestimated, truly non-equivalent samples are required. For this purpose, HX-MS measurements of samples containing 5% and 10% of a structural variant of MBP, W169G, were used.[36] Our previous work established that there are subtle, but statistically significant HX differences between the spiked samples and wild-type MBP, taken here as the reference sample. The 5% and 10% structural variant results are ideal because the truly non-equivalent differences present in the HX-MS data are close to the HX detection limit. In this way, the empirically defined EAC limit is challenged. Each structural variant HX-MS data set contains 460 confidence intervals for the 115 MBP peptic peptides at four HX labeling times. Confidence intervals are shown in Figure 2 for samples spiked with 5% and 10% mutant. For the 5% and 10% spiked samples, there are individual confidence intervals that exceed +EAC or –EAC. Thus, these samples, with only subtle differences in HX-MS, were found to be non-equivalent based on the empirically-defined EAC limits.

Beyond the EAC limits, however, we have previously identified all HX differences that exceeded statistical significance limits for differences.[34, 36] However, some of these statistically significant differences here appear to be statistically equivalent because their confidence intervals are bounded by the EAC limits (red symbols in Figure 2D). This suggests that the EAC limit alone is not sufficiently stringent because these results are false equivalents (type I error). Figure 3

compares the structural elements of MBP where HX in the 10% spiked sample was significantly different than wild-type (Figure 3A) and where HX in the 10% spiked

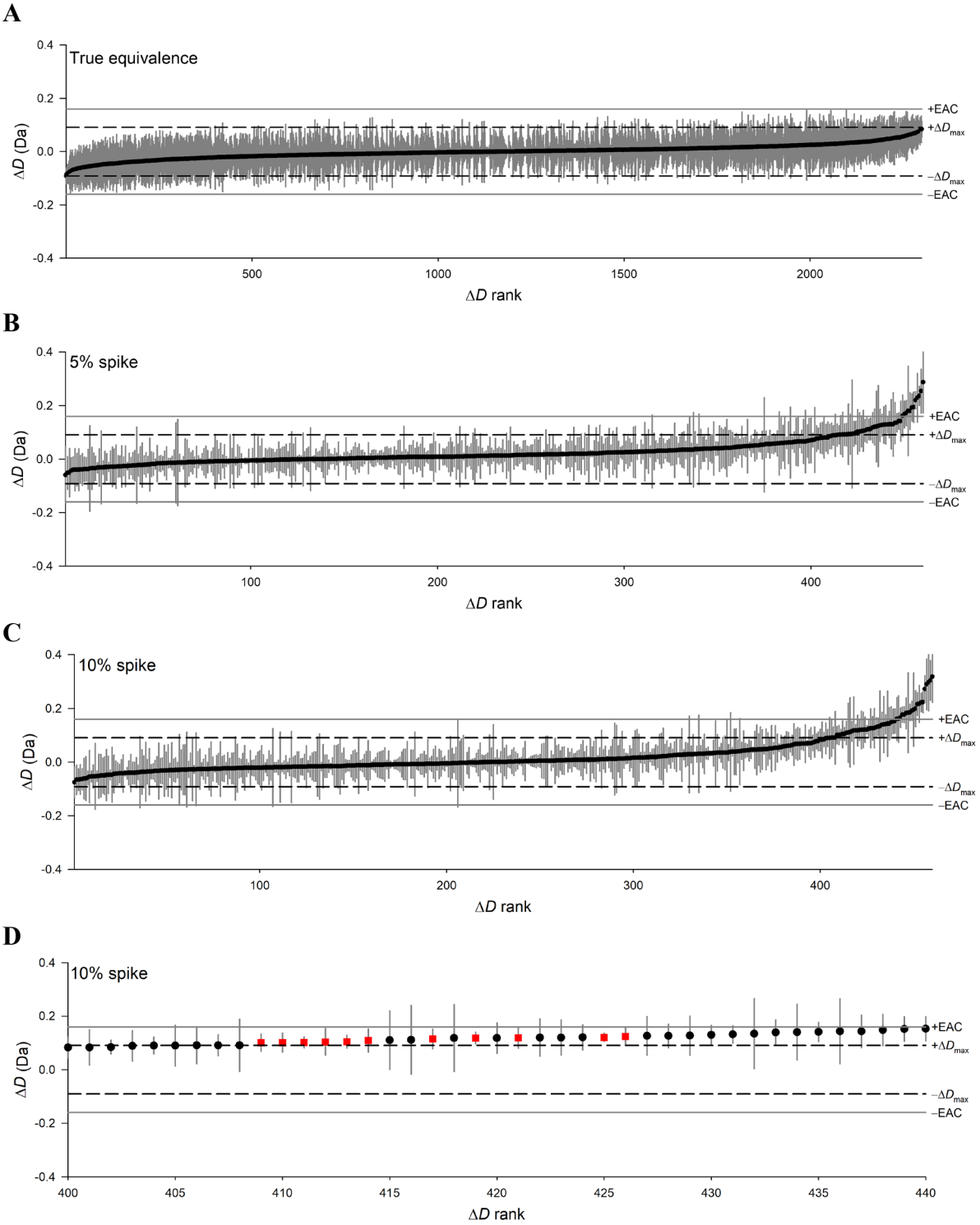


Figure 2. Δ*D* values ranked from smallest to largest with 90% confidence intervals for (A) truly equivalent MBP and MBP spiked with 5% mutant (B) and 10% mutant (C and D). In the zoomed-in view (D) falsely equivalent data based on the EAC limit are shown as red filled squares.

sample was falsely equivalent to wild-type (Figure 3B) based on the EAC. In total, there were 8 false equivalents in the 5% spiked sample and 11 in the 10% spiked sample, yielding type I error rates of 2.4% and 1.7%, respectively.

To increase the stringency of the equivalence test, we impose an additional criterion on the HX data, also defined by the resampled truly equivalent samples, that imposes a limit on the maximum difference in means:

$$\left|\Delta D\right| < \max\left(\left|\mathbf{\Delta D}_{\mathbf{eq}}\right|\right) \equiv \Delta D_{\max} \tag{8}$$

where $\mathbf{\Delta D}_{\mathrm{eq}}$ denotes the vector of all differential HX measurements obtained from the analysis of truly equivalent samples. Therefore, the hybrid equivalence criteria using maximum deviations in true equivalents (MDTE) are:

$$H_1 : \begin{cases} \mathbf{\Delta D} - \mathbf{CI} > -\mathrm{EAC} \\ \mathbf{\Delta D} + \mathbf{CI} < \mathrm{EAC} \\ \left|\mathbf{\Delta D}\right| < D_{\max} \end{cases} \tag{9}$$

here the vector notation indicates that equivalence (alternative hypothesis, H_1) is only established when *all* data simultaneously satisfy all of the equivalence criteria. As evident in Figure 2D, all of the HX measurements that were falsely equivalent using only the EAC limits (red symbols) become non-equivalent when the additional difference in means criterion is also imposed.

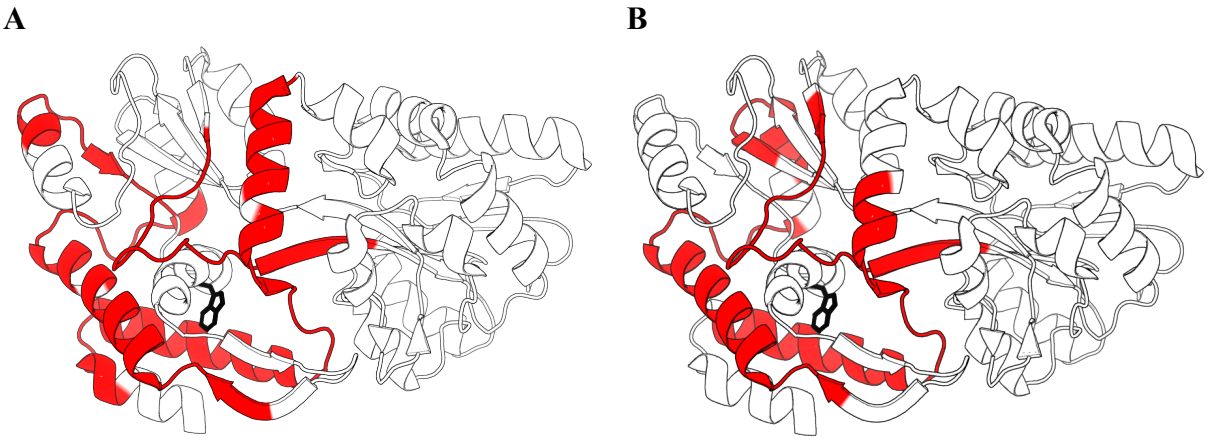


Figure 3. Structure of MBP (PDB 1OMP).[37] Black sticks indicate the site of the W169G mutation. (A) Statistically significant differences (red colored) identified in 10% W169G spiked samples.[36] (B) False equivalence (type I errors) by TOST. False equivalents had confidence intervals bounded by the ±EAC, but $\left|\Delta D\right| > D_{\max}$ (shown in Figure 2D).

Testing equivalence of differential HX-MS of an innovator mAb with its biosimilars using TOST-MDTE

Here, we evaluate the suitability of HX-MS TOST-MDTE equivalence testing within the context of a therapeutic mAb, Remicade (infliximab). A pool of truly equivalent HX-MS data was obtained from eight HX-MS replicate measurements of Remicade measured at six HX labeling times. The extent of HX was measured in 187 peptic peptides, 112 from the heavy chain and 75 from the light chain. A total of five comparisons of were performed using two different randomly selected unique triplicate data sets in each. In total, 5,610 ΔD values with 11,220 standard deviations were pooled. Using the approach outlined for MBP, the data from truly equivalent samples were used to define EAC limits of ±0.842 Da and maximum acceptable HX differences of ±0.367 Da (see Table 1).

Next, we applied these limits to HX-MS differences between each of three approved infliximab biosimilars, Remsima, Renflexis, and Inflectra, and the reference product, Remicade. Here, the null hypothesis of the equivalence test is that the HX-MS data of the biosimilar are not equivalent to the reference product. Each infliximab HX-MS data set contains 1,122 ΔD confidence intervals for the 187 infliximab peptic peptides at six HX labeling times. The biosimilar ΔD confidence intervals are shown in Figure 4. All HX results meet the hybrid equivalence criteria with the reference product, Remicade. This indicates that the higher-order structure of each biosimilar, as measured by HX-MS, is equivalent to the reference product.

Challenging the stringency of the MDTE equivalence acceptance criteria

A potential weakness with equivalence testing is setting overly permissive limits for the HX differences. In such a scenario, even samples that are truly non-equivalent may satisfy the equivalence test limits because the limits are too permissive, leading to false equivalence. To explore the potential of false equivalence in mAb differential HX-MS we used partially deglycosylated NIST mAb as a mock candidate biosimilar and tested for its equivalence to unmodified NIST mAb. Previous HX-MS analysis of NIST mAb and Fc with altered glycosylation have shown that glycan modification has substantial and widespread effects on HX in the Fc domain.[22] However, such HX effects are too large to provide an adequate challenge of the equivalence criteria developed here. To generate a more realistic and challenging sample for equivalence testing, NIST mAb was deglycosylated with PNGaseF and then spiked into unmodified NIST mAb at 5% mole fraction to yield a 5% deglycosylated NIST mAb sample. While a quantitative analysis of the glycan profile would readily detect this change, the goal here was to introduce a subtle perturbation in the ensemble-averaged higher-order structure of the NIST mAb. The 5% deglycosylated NIST mAb samples were then tested for HX-MS equivalence to the unmodified NIST mAb using the equivalence criteria established from Remicade. The data set contained a total of 888 ΔD values and 1,776 standard deviations. The results, shown in Figure 5, indicate that 5% deglycosylated NIST mAb is not equivalent to NIST mAb by HX-MS, with seven measurements failing to satisfy the equivalence criteria (red symbols). Thus, results with the partially deglycosylated NIST mAb demonstrate that HX-MS equivalence criteria are sufficiently stringent that subtle changes in higher-order structure lead to acceptance of the non-equivalence null hypothesis.

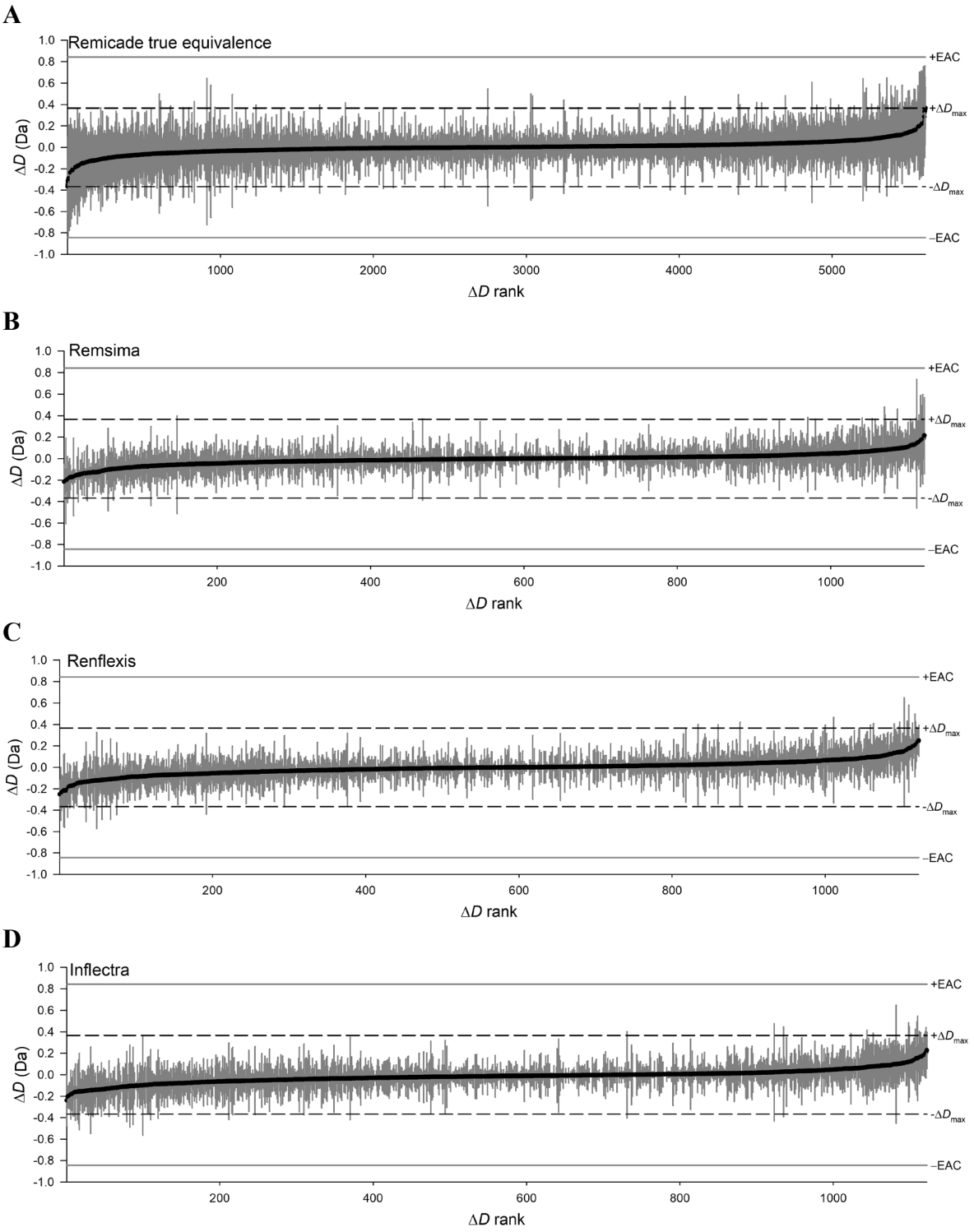


Figure 4. HX differences between truly equivalent Remicade samples and between three approved infliximab biosimilars, Remsima, Renflexis, and Inflectra and the infliximab reference product, Remicade evaluated using TOST-MDTE equivalence testing criteria. ΔD values are ranked from smallest to largest. 90% confidence intervals are shown by the gray bars. All differential HX-MS data from the biosimilars satisfied the hybrid equivalence criteria defined by equation (9).

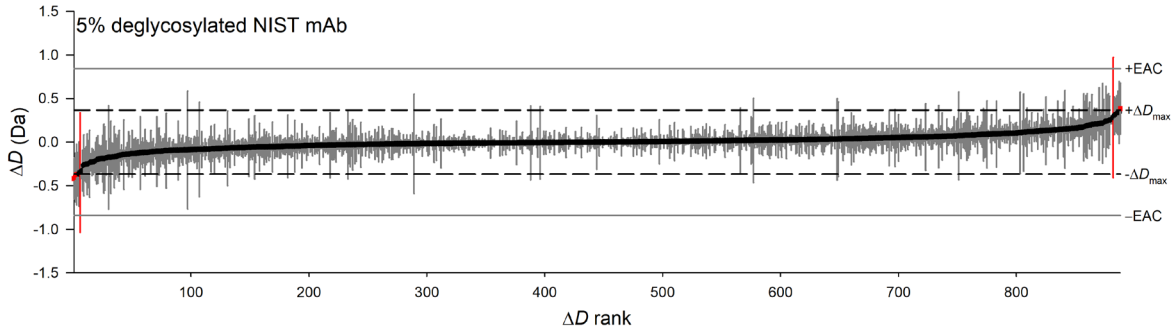


Figure 5. Equivalence testing of 5% deglycosylated NIST mAb against unmodified NIST mAb using TOST-MDTE equivalence testing. ΔD values are ranked from smallest to largest. 90% confidence intervals are shown by the gray bars. The equivalence criteria were determined from truly equivalent Remicade measurements (Figure 4 and Table 1). Data shown in red failed to meet the equivalence criteria defined by equation (9).

DISCUSSION

HX-MS measurements could be used to monitor quality attributes related to the higher order structure of a protein therapeutic. If the HX-MS measurements have high criticality, then using HX-MS data in comparability contexts (e.g., post-approval process changes) or in biosimilarity contexts requires robust statistical methods to demonstrate statistical equivalence. The two one-sided tails (TOST) method is well-established for equivalence testing for univariate data, but methods for defining equivalence acceptance criteria and extending equivalence testing to multivariate, highly correlated data have not been developed. The present work establishes (1) a method to establish empirical equivalence limits by using maximum deviations found upon resampling measurements of truly equivalent samples and (2) a general framework, based on spike-in of structurally perturbed molecules, to demonstrate that the equivalence testing is fit-for-purpose.

The fit-for-purpose evaluation is essential in order to establish that the equivalence testing criteria are not overly permissive. Permissive limits would lead to unacceptable false equivalence determinations. With respect to the spiking approach, we have illustrated here two separate model systems: MBP was based on an aggressive point mutation and the NIST mAb was based on a change in a post-translational modification, glycosylation. This latter change, at about the 5% level reflects the kinds of changes that might be encountered following a cell line change or other changes in upstream conditions. Other kinds of perturbations, such as physically or chemically stressed molecules or slight changes in formulation, could also be used. The specific application and the nature of the criticality will need to be considered when developing a model for a specific application.

The distribution of ΔD values for infliximabs (see Figure 4) is noticeably wider than the distribution for MBP samples (see Figure 2). Thus, the thresholds for equivalence are wider (see Table 1). The larger variability in infliximab than in MBP measurements could be a result of increased protein size and structural complexity. MBP is a 41.5 kDa single chain protein without disulfide bonds while infliximab is a 150 kDa multi-domain and multi-chain protein with 16

disulfide bonds. Furthermore, as a glycosylated protein, infliximab is heterogeneous, with two major glycoforms (G0F, G1F) and various other minor glycoforms present.[38, 39] The increased heterogeneity and complexity of the IgG1 mAb could explain the increased variability. The number of observations is also much larger for the mAbs, thus the probability of results with poor precision (i.e., rare outliers) increases. Variability differences between the two protein systems clearly indicate that acceptance criteria must be determined for the protein of interest.

Figure 4 also shows that the differential HX measurements of the biosimilars have a much narrower range than the HX differences for Remicade. Because the pool of Remicade data was randomly resampled five times, there is a greater probability of obtaining extreme outliers. One way to narrow the estimate of both equivalence criteria, EAC and $\Delta D_{\max}$ equivalence threshold, is to use more replicates. Ideally, increasing the number of replicates will produce a narrower distribution of ΔD values, thus decreasing both EAC and $\Delta D_{\max}$. In addition, as recently suggested by Moroco and Engen, so-called biological replicates would be valuable to capture lot-to-lot variability and aid in interpreting biologically meaningful results.[40]

By defining the EAC based on the maximum deviations in truly equivalent samples observed in the resampled data, the EAC might be overestimated. With an overestimated EAC, the probability of false equivalence increases. To mitigate this, we introduced a $\Delta D_{\max}$ equivalence threshold. However, based on the wider distribution of $\Delta D_{\max}$ values in the Remicade data, the $\Delta D_{\max}$ equivalence threshold might also be overestimated. In fact, the infliximab biosimilar HX-MS results suggest that the limits are overestimated slightly because the largest observed $|\mathbf{\Delta D}|$ values and confidence intervals are less than the limits established with resampled Remicade. However, the biosimilars are true equivalents (i.e., already granted regulatory approval as highly similar) so an outcome of equivalence was expected. While more stringent equivalence limits could be obtained from a smaller Remicade data set, the probability of false non-equivalence would also increase. Non-equivalence between NIST mAb and 5% deglycosylated NIST mAb demonstrates that the equivalence criteria are stringent enough that minor perturbations in higher-order structure result in a finding of non-equivalence.

Understanding the clinical significance of findings from analytical assessments of higher-order protein structural comparability and similarity is important. If significant differences are found, it would be up to subject matter experts to determine if the differences in HX might be clinically meaningful. With well-defined equivalence acceptance criteria, demonstrating similarity of higher-order structure could contribute to the totality of evidence that a biological product is highly similar to its reference product. Ideally, an EAC range should be less than clinically meaningful differences in HX. Establishing that such clinically meaningful differences resulted in non-equivalence would further demonstrate that the equivalence testing was fit-for-purpose. At the present time, this issue is not well-understood. There is a continued unmet need in the comparability and similarity applications of HX-MS to understand how differences in HX correlate to function and stability. An HX-MS study containing a panel of structural variants with differing degrees of structural perturbation and well-defined function and stability profiles would be invaluable for addressing this unmet need and in determining acceptable equivalence criteria.

ACKNOWLEDGEMENTS

This research was supported by the National Institutes of Health NIGMS Biotechnology Predoctoral Training Program (T32-GM008359) to TSH and the University of Kansas Madison and Lila Self Graduate Fellowship to MSW. This material is based upon work supported by the National Science Foundation under grant CHE-1709176.

SUPPORTING INFORMATION

The supporting information consists of a detailed description of the experimental methods; Figure S1 showing peptic peptide coverage map for infliximab; Figure S2 showing peptic peptide coverage map for NIST mAb; Table S1 summarizing the HX-MS measurements of infliximabs; Table S2 summarizing HX-MS measurements of NIST mAb; supporting file remicade_equivalents_hxms_data.txt containing all individual replicates of HX-MS measurements of Remicade in comma separated values format; supporting file infliximab_hxms_data.txt containing all HX-MS data for infliximab in comma separated values format; supporting file nist_mab_hxms_data.txt containing all HX-MS data for NIST mAb in comma separated values format.

REFERENCES

1. Azevedo, V.; Hassett, B.; Fonseca, J. E.; Atsumi, T.; Coindreau, J.; Jacobs, I.; Mahgoub, E.; O'Brien, J.; Singh, E.; Vicik, S.; Fitzpatrick, B., Differentiating biosimilarity and comparability in biotherapeutics. *Clinical Rheumatology* **2016,** *35* (12), 2877-2886.

2. Federici, M.; Lubiniecki, A.; Manikwar, P.; Volkin, D. B., Analytical lessons learned from selected therapeutic protein drug comparability studies. *Biologicals* **2013,** *41* (3), 131-147.

3. McGillicuddy, N.; Floris, P.; Albrecht, S.; Bones, J., Examining the sources of variability in cell culture media used for biopharmaceutical production. *Biotechnol. Lett.* **2018,** *40* (1), 5-21.

4. Vulto, A. G.; Jaquez, O. A., The process defines the product: what really matters in biosimilar design and production? *Rheumatology* **2017,** *56* (suppl_4), iv14-iv29.

5. Blandizzi, C.; Meroni, P. L.; Lapadula, G., Comparing Originator Biologics and Biosimilars: A Review of the Relevant Issues. *Clin Theapeutics* **2017,** *39* (5), 1026-1039.

6. Brown, K. A.; Rajendran, S.; Dowd, J.; Wilson, D. J., Rapid Characterization of Structural and Functional Similarity for a Candidate Bevacizumab (Avastin) Biosimilar using a Multipronged Mass Spectrometry-Based Approach. *Drug Test Anal* **2019,** *11* (ja), 1207-1217.

7. 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., Physicochemical and biological characterization of SB2, a biosimilar of Remicade® (infliximab). *mAbs* **2017,** *9* (2), 365-383.

8. Brokx, S.; Scrocchi, L.; Shah, N.; Dowd, J., A demonstration of analytical similarity comparing a proposed biosimilar pegfilgrastim and reference pegfilgrastim. *Biologicals* **2017,** *48*, 28-38.

9. Cho, I. H.; Lee, N.; Song, D.; Jung, S. Y.; Bou-Assaf, G.; Sosic, Z.; Zhang, W.; Lyubarskaya, Y., Evaluation of the structural, physicochemical, and biological characteristics of SB4, a biosimilar of etanercept. *mAbs* **2016,** *8* (6), 1136-1155.

10. Rogstad, S.; Faustino, A.; Ruth, A.; Keire, D.; Boyne, M.; Park, J., A Retrospective Evaluation of the Use of Mass Spectrometry in FDA Biologics License Applications. *J. Am. Soc. Mass. Spectrom.* **2017,** *28* (5), 786-794.

11. Houde, D.; Berkowitz, S. A.; Engen, J. R., The utility of hydrogen/deuterium exchange mass spectrometry in biopharmaceutical comparability studies. *J. Pharm. Sci.* **2011,** *100*, 2071-2086.

12. Bonnington, L.; Lindner, I.; Gilles, U.; Kailich, T.; Reusch, D.; Bulau, P., Application of HDX-MS to Biopharmaceutical Development Requirements: Improved sensitivity to detection of conformational changes. *Anal. Chem.* **2017,** *89* (16), 8233-8237.

13. Houde, D.; Berkowitz, S. A., Conformational comparability of factor IX-Fc fusion protein, factor IX, and purified Fc fragment in the absence and presence of calcium. *J. Pharm. Sci.* **2012,** *101* (5), 1688-700.

14. Wei, H.; Ahn, J.; Yu, Y. Q.; Tymiak, A.; Engen, J. R.; Chen, G., Using hydrogen/deuterium exchange mass spectrometry to study conformational changes in granulocyte colony stimulating factor upon PEGylation. *J. Am. Soc. Mass. Spectrom.* **2012,** *23* (3), 498-504.

15. Fang, J.; Doneanu, C.; Alley, W. R.; Yu, Y. Q.; Beck, A.; Chen, W., Advanced assessment of the physicochemical characteristics of Remicade® and Inflectra® by sensitive LC/MS techniques. *mAbs* **2016,** *8* (6), 1021-34.

16. Lee, J.; Kang, H. A.; Bae, J. S.; Kim, K. D.; Lee, K. H.; Lim, K. J.; Choo, M. J.; Chang, S. J., Evaluation of analytical similarity between trastuzumab biosimilar CT-P6 and reference product using statistical analyses. *mAbs* **2018**, 1-25.

17. Pan, J.; Zhang, S.; Borchers, C. H., Comparative higher-order structure analysis of antibody biosimilars using combined bottom-up and top-down hydrogen-deuterium exchange mass spectrometry. *Biochim. Biophys. Acta* **2016,** *1864* (12), 1801-1808.

18. Visser, J.; Feuerstein, I.; Stangler, T.; Schmiederer, T.; Fritsch, C.; Schiestl, M., Physicochemical and Functional Comparability Between the Proposed Biosimilar Rituximab GP2013 and Originator Rituximab. *BioDrugs* **2013,** *27* (5), 495-507.

19. Houde, D.; Arndt, J.; Domeier, W.; Berkowitz, S.; Engen, J. R., Characterization of IgG1 conformation and conformational dynamics by hydrogen/deuterium exchange mass spectrometry. *Anal. Chem.* **2009,** *81*, 2644-2651.

20. Houde, D.; Peng, Y.; Berkowitz, S. A.; Engen, J. R., Post-translational modifications differentially affect IgG1 conformation and receptor binding. *Mol. Cell. Proteomics* **2010,** *9* (8), 1716-28.

21. More, A. S.; Toth, R. T., IV; Okbazghi, S. Z.; Middaugh, C. R.; Joshi, S. B.; Tolbert, T. J.; Volkin, D. B.; Weis, D. D., Impact of Glycosylation on the Local Backbone Flexibility of Well-Defined IgG1-Fc Glycoforms Using Hydrogen Exchange-Mass Spectrometry. *J. Pharm. Sci.* **2018,** *107* (9), 2315-2324.

22. Groves, K.; Cryar, A.; Cowen, S.; Ashcroft, A. E.; Quaglia, M., Mass Spectrometry Characterization of Higher Order Structural Changes Associated with the Fc-glycan Structure of the NISTmAb Reference Material, RM 8761. *J. Am. Soc. Mass. Spectrom.* **2020,** *31* (3), 553-564.

23. Majumdar, R.; Middaugh, C. R.; Weis, D. D.; Volkin, D. B., Hydrogen-Deuterium Exchange Mass Spectrometry as an Emerging Analytical Tool for Stabilization and Formulation Development of Therapeutic Monoclonal Antibodies. *J. Pharm. Sci.* **2015,** *104* (2), 327-45.

24. Barchard, K. A., Null Hypothesis Significance Testing Does Not Show Equivalence. *Analyses Soc Issues Pub Policy* **2015,** *15* (1), 418-421.

25. Burdick, R.; Coffey, T.; Gutka, H.; Gratzl, G.; Conlon, H. D.; Huang, C.-T.; Boyne, M.; Kuehne, H., Statistical Approaches to Assess Biosimilarity from Analytical Data. *AAPS J* **2017,** *19* (1), 4-14.

26. Johnstone, I. M.; Titterington, D. M., Statistical challenges of high-dimensional data. *Phil Trans Roy Soc A* **2009,** *367* (1906), 4237-4253.

27. Dudoit, S.; van der Laan, M. J., *Multiple Testing Procedures with Applications to Genomics*. Springer Science+Business Media: New York, 2008.

28. Benjamini, Y.; Hochberg, Y., Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. Roy. Stat. Soc. Ser. B.* **1995,** *57* (1), 289-300.

29. Limentani, G. B.; Ringo, M. C.; Ye, F.; Bergquist, M. L.; McSorley, E. O., Beyond the t-Test: Statistical Equivalence Testing. *Anal. Chem.* **2005,** *77* (11), 221 A-226 A.

30. Lakens, D., Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Soc. Psychol. Pers. Sci.* **2017,** *8* (4), 355-362.

31. Burdick, R. K.; LeBlond, D. J.; Pfahler, L. B.; Quiroz, J.; Sidor, L.; Vukovinsky, K.; Zhang, L., Analytical Proceedures. In *Statistical Applications for Chemistry, Manufacturing and Controls (CMC) in the Pharmaceutical Industry*, Springer International Publishing: Cham, 2017; pp 193-225.

32. Munroe, R. Error Types. https://xkcd.com/2303/.

33. Masson, G. R.; Burke, J. E.; Ahn, N. G.; Anand, G. S.; Borchers, C.; Brier, S.; Bou-Assaf, G. M.; Engen, J. R.; Englander, S. W.; Faber, J.; Garlish, R.; Griffin, P. R.; Gross, M.

L.; Guttman, M.; Hamuro, Y.; Heck, A. J. R.; Houde, D.; Iacob, R. E.; Jørgensen, T. J. D.; Kaltashov, I. A.; Klinman, J. P.; Konermann, L.; Man, P.; Mayne, L.; Pascal, B. D.; Reichmann, D.; Skehel, M.; Snijder, J.; Strutzenberg, T. S.; Underbakke, E. S.; Wagner, C.; Wales, T. E.; Walters, B. T.; Weis, D. D.; Wilson, D. J.; Wintrode, P. L.; Zhang, Z.; Zheng, J.; Schriemer, D. C.; Rand, K. D., Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry (HDX-MS) experiments. *Nat. Methods* **2019,** *16* (7), 595-602.

34. Hageman, T. S.; Weis, D. D., Reliable Identification of Significant Differences in Differential Hydrogen Exchange-Mass Spectrometry Measurements Using a Hybrid Significance Testing Approach. *Anal. Chem.* **2019,** *91* (13), 8008-8016.

35. Satterthwaite, F. E., An Approximate Distribution of Estimates of Variance Components. *Biometrics Bull.* **1946,** *2* (6), 110-114.

36. Hageman, T. S.; Weis, D. D., A Structural Variant Approach for Establishing a Detection Limit in Differential Hydrogen Exchange-Mass Spectrometry Measurements. *Anal. Chem.* **2019,** *91* (13), 8017-8024.

37. Sharff, A. J.; Rodseth, L. E.; Spurlino, J. C.; Quiocho, F. A., Crystallographic evidence of a large ligand-induced hinge-twist motion between the two domains of the maltodextrin binding protein involved in active transport and chemotaxis. *Biochemistry* **1992,** *31* (44), 10657-10663.

38. Pisupati, K.; Tian, Y.; Okbazghi, S.; Benet, A.; Ackermann, R.; Ford, M.; Saveliev, S.; Hosfield, C. M.; Urh, M.; Carlson, E.; Becker, C.; Tolbert, T. J.; Schwendeman, S. P.; Ruotolo, B. T.; Schwendeman, A., A Multidimensional Analytical Comparison of Remicade and the Biosimilar Remsima. *Anal. Chem.* **2017,** *89* (9), 4838-4846.

39. Lee, C.; Jeong, M.; Lee, J. J.; Seo, S.; Cho, S. C.; Zhang, W.; Jaquez, O., Glycosylation profile and biological activity of Remicade® compared with Flixabi® and Remsima®. *mAbs* **2017,** *9* (6), 968-977.

40. Moroco, J. A.; Engen, J. R., Replication in bioanalytical studies with HDX MS: aim as high as possible. *Bioanalysis* **2015,** *7* (9), 1065-1067.