ACS Partner Journal

Journal of the American Society for Mass Spectrometry

pubs.acs.org/jasms

Research Article

High-Resolution Hydrogen–Deuterium Protection Factors from Sparse Mass Spectrometry Data Validated by Nuclear Magnetic Resonance Measurements

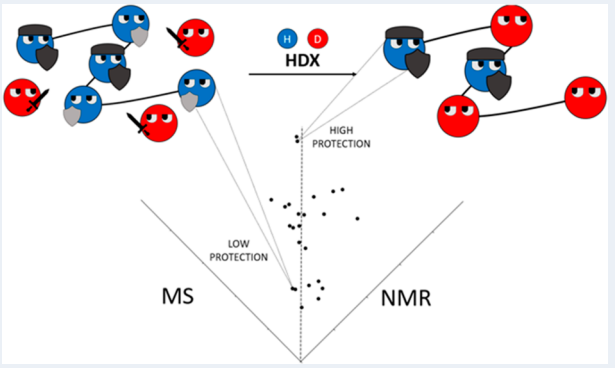
Michele Stofella, Simon P. Skinner, Frank Sobott, Jeanine Houwing-Duistermaat, and Emanuele Paci*

Cite This: https://doi.org/10.1021/jasms.2c00005

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Experimental measurement of time-dependent spontaneous exchange of amide protons with deuterium of the solvent provides information on the structure and dynamical structural variation in proteins. Two experimental techniques are used to probe the exchange: NMR, which relies on different magnetic properties of hydrogen and deuterium, and MS, which exploits the change in mass due to deuteration. NMR provides residue-specific information, that is, the rate of exchange or, analogously, the protection factor (i.e., the unitless ratio between the rate of exchange for a completely unstructured state and the observed rate). MS provides information that is specific to peptides obtained by proteolytic digestion. The spatial resolution of HDX-MS measurements depends on the proteolytic pattern of the protein, the fragmentation method used, and the overlap between peptides. Different computational approaches have been proposed to extract residue-specific information from peptide-level HDX-MS measurements. Here, we demonstrate the advantages of a method recently proposed that exploits self-consistency and classifies the possible sets of protection factors into a finite number of alternative solutions compatible with experimental data. The degeneracy of the solutions can be reduced (or completely removed) by exploiting the additional information encoded in the shape of the isotopic envelopes. We show how sparse and noisy MS data can provide high-resolution protection factors that correlate with NMR measurements probing the same protein under the same conditions.



INTRODUCTION

Hydrogen–deuterium exchange (HDX) is the spontaneous exchange of covalently bonded hydrogens of a protein with deuterium in solution.[1] In his pioneering work, Lindestrøm-Lang probed the phenomenon through density gradient tubes.[2] Since then, nuclear magnetic resonance (NMR) has been the leading technique used to probe HDX until the early 2000s,[3] when mass spectrometry (MS) began to emerge as an alternative with many advantages (no sample size limitations, no labeling required, low protein concentration, low costs, and highly automated processing), counting an increasing number of applications in fundamental biophysics and applied biotechnology.[4−6] With both NMR and MS, only the exchange of amide hydrogens can be observed because other hydrogens exchange either too fast (side chain acidic and basic hydrogens and polar groups) or too slowly (carbon-bonded hydrogens as well as side chain aliphatic and aromatic hydrogens) to be detected. Hence, in principle, both techniques probe the properties of single amino acids.[7]

NMR exploits the different magnetic properties of hydrogen and deuterium to determine the rate of exchange of individual residues. Their measurement is limited by the resolution of the amide signals themselves, or of cross peaks in homo- or heteronuclear multidimensional NMR spectra.[3] MS measures directly the mass variation as a function of exchange time of peptides obtained by proteolytic digestion.[4] The spatial resolution of HDX-MS measurements depends on the digestion pattern of the protein, the overlap between peptides, and the MS/MS fragmentation methods used.[8] Most current approaches use collision-induced dissociation (CID) for MS/MS fragmentation of peptides, but because of H/D scrambling during collisional activation, no information is gained on the exact location of deuterium labels within the peptides. Instead, such MS/MS data merely serve to unambiguously identify peptides by their sequence tags. Different approaches have been proposed to increase spatial resolution, including the use

Received: January 11, 2022
Revised: March 1, 2022
Accepted: March 21, 2022

of alternative MS/MS methods, which minimize H/D scrambling during fragmentation.[9−11]

Several computational strategies have been proposed to extract single-residue protection factors from peptide-level HDX-MS data.[12−19] Here, we demonstrate the advantages of a method recently proposed[20] that exploits self-consistency (i.e., data consistency among overlapping peptides) and finds alternative sets of protection factors equally consistent with experimental data. These solutions can be classified into a finite number of clusters whose degeneracy can be further reduced by exploiting the additional information contained in the shape of the isotopic envelope. We show how sparse and noisy MS data can provide high-resolution protection factors that correlate with NMR measurements probing the same protein at the same conditions.

The exchange kinetics of an amide proton is highly dependent on the environment, hence, a unique probe of the structure and dynamics of proteins. Since the seminal work from Linderstrøm-Lang,[2] HDX has been modeled as a two-step process. The deuteration of a residue in a D_2O solution is possible if a local *opening* of the structure occurs:

$$\mathrm{NH_{cl}} \underset{k_o}{\overset{k_c}{\rightleftharpoons}} \mathrm{NH_{op}} \xrightarrow[D_2O]{k_{int}} \mathrm{ND} \tag{1}$$

Here, k_o and k_c refer to *opening* and *closing* rates, respectively, which allow the residue to switch from an exchange-incompetent state (i.e., a closed or folded state NH_{cl}) to an exchange-competent state (i.e., an open or unfolded state NH_{op}). The intrinsic exchange rate k_{int} is the exchange rate of the residue in a completely unstructured protein and depends on the pH, temperature of the solution, and side chains of the two adjacent amino acids.[21−25]

Since, for a folded protein $k_c \gg k_o$ (native state approximation), the observed exchange rate can be written as

$$k_{obs} = \frac{k_{int}k_o}{k_{int} + k_c} \tag{2}$$

This expression suggests two limiting cases depending on the relative size of k_{int} and k_c. If $k_{int} \ll k_c$ (EX2 regime), the deuteration of a single residue is

$$d(t, P) = 1 - e^{-\frac{k_{int}}{P}t} \tag{3}$$

where the opening equilibrium constant $P \equiv k_c/k_o$, known as the protection factor, is linked to the dynamic properties of the residue by definition; moreover, several studies have shown a correlation between the protection factors of a protein and its structure.[26,27] If instead $k_c \ll k_{int}$ (EX1 regime), $k_{obs} = k_o$. Under physiological conditions, the EX2 regime dominates the exchange kinetics in natively folded proteins.[28]

In HDX-NMR experiments, the proton signal decays exponentially as deuteration occurs because deuterium is ^1H NMR silent, and the experimental curves can be fitted with eq 3 to obtain P.[29,30]

On the other hand, HDX-MS measures the exchange of proteolytic peptides, with experimental curves resulting in a sum of exponentials. The fractional deuterium uptake at time t of a peptide of N exchangeable residues (i.e., excluding prolines and the N terminus) is

$$D(t, \{P_i\}) = \frac{1}{N}\sum_{i=1}^{N}(1 - e^{-k_{int,i}/P_i t}) \tag{4}$$

where P_i and $k_{int,i}$ are the protection factor and the intrinsic exchange rate of the residue i. If exchange rates (or, equivalently, protection factors) are known for each residue, the exchange kinetics of peptides is uniquely defined, but not vice versa.[20]

The possibility of estimating individual protection factors from HDX-MS data depends on four factors:[8] (i) peptide overlap, (ii) time point resolution, (iii) time window coverage, and (iv) experimental error. (i) The protection factor of an individual amino acid can, in principle, be extracted only if two proteolytic peptides differ by exactly one amino acid. When multiple peptides partly overlapping are available, protection factors are ambiguous, with the ambiguity decreasing with an increase in the number of overlapping peptides.[12,20] In the case of "exact" measurements (i.e., not affected by experimental error), the problem is combinatorial: for an isolated peptide formed by N residues, there are $N!$ possible solutions (Supplementary Figure 1A); for two overlapping peptides formed by N1 and N2 residues, respectively, and with N_c residues in common, there are $(N1 - Nc)!(N2 - Nc)!Nc!$ alternative solutions (Supplementary Figure 1B). Reporting a solution in terms of observed rates ($k_{obs} = k_{int}/P$) or protection factors yields equivalent results with different numerical values arising from the different intrinsic exchange rates between residues (Supplementary Figure 1C). Although the observed rates span several orders of magnitude depending on the experimental conditions (pH, temperature), the protection factor can be restricted to the boundaries $0 < \ln(P) \leq 20$, facilitating the convergence of fitting algorithms. (ii) The fractional uptake of a peptide (eq 4) is measured for a discrete set of times (N_{times}); if these are fewer than the exchangeable amino acids in the peptide (N_{res}), the individual residues' protection factor is underdetermined: multiple solutions are equally consistent with experimental data. Even for small peptides, though, where in principle the number of time points is sufficient to extract all the exchange rates ($N_{res} \ll N_{times}$), the solutions are degenerate because eq 4 does not contain information on the relative contribution of the fitting parameters (protection factors). A necessary condition, albeit not sufficient, is that the number of experimental points should be no less than some multiple Q (quality factor) of the number of adaptive parameters in the model:[31] $N_{exp} \approx QN_{res}$, where $N_{exp} = N_{times}$ for an isolated peptide. (iii) To properly sample the multiexponential uptake of a peptide (eq 4), these exchange times should follow a log-uniform distribution between the *beginning* and the *end* of the exchange process, which can be deduced from the exchange of the whole protein (i.e., without digestion). Typical HDX-MS measurements report time-resolved exchange between tens of seconds and hours. The detection of exchange at shorter times (e.g., subsecond) is now possible, with recent developments giving access to millisecond time scales.[32−34] A simultaneous fitting of the information encoded in multiple overlapping peptides reduces the degeneracy on the rate-to-residue assignment by adding local information. Moreover, it increases the number of experimental points N_{exp}: for a region formed by N_{res} residues and covered by N_{pep} peptides, $N_{exp} = N_{times}N_{pep}$. Experimentally, the overlap of peptides depends on the choice of the protease, which is limited because of the acidic conditions needed to quench the exchange. (iv) The presence of experimental uncertainties affects the accuracy on the final predictions. The law of large numbers ensures that the average value among independent measurements (replicates) tends to

https://doi.org/10.1021/jasms.2c00005
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

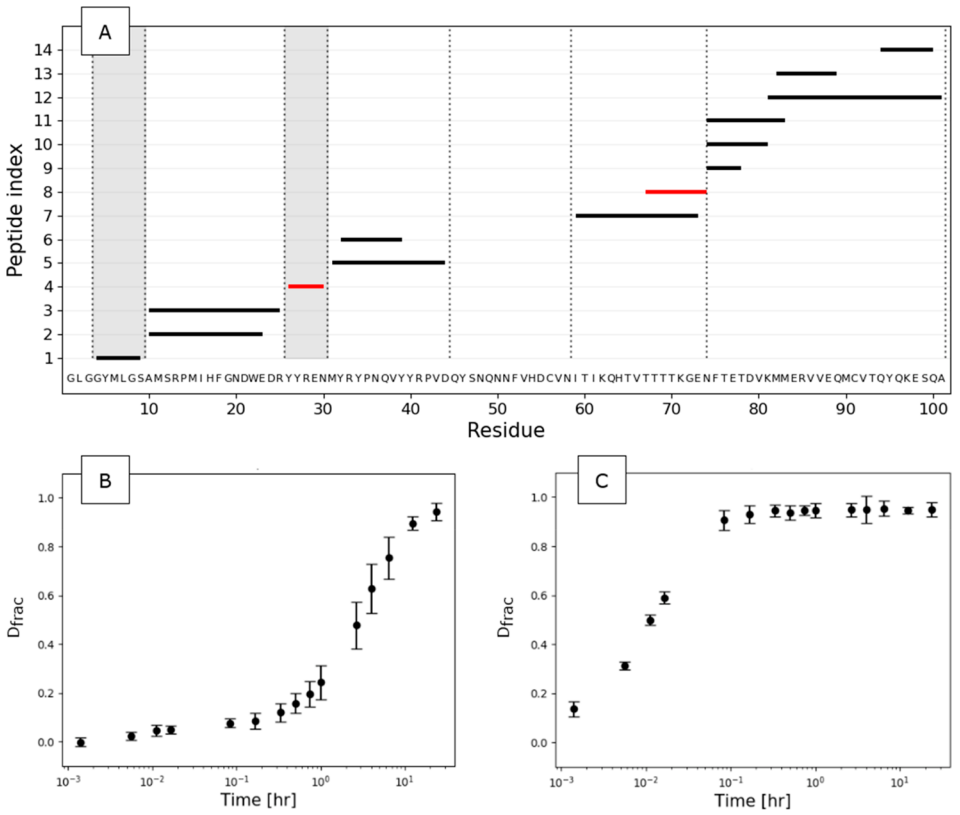


Figure 1. HDX-MS data set previously published in ref 36. (A) The coverage map localizes the 14 peptides identified after pepsin digestion. The six regions covered by isolated (gray) or contiguous overlapping peptides are separated by vertical dotted lines. (B, C) The fractional uptake of peptides 4 (B) and 8 (C)—highlighted in red in the peptide map—is shown at 15 time points.

the mean of the measurements, that is, the true value of the estimated quantity in the limit of an infinite number or replicates. The number of replicates provided in HDX-MS experiments (generally three) limits the accuracy of the measured quantity (i.e., the fractional uptake), and consequently of the estimated protection factor.

Two computational approaches aim to extract protection factors at the highest resolution possible from HDX-MS data sets. HDSite[12,35] uses the isotopic envelopes to derive the extent of deuteration of each residue of the peptide at different exchange times ($0 \leq d \leq 1$), and the obtained curve can be further fitted with a single exponential (eq 3) to obtain the protection factor. An initial guess on the deuteration of each residue is refined to reproduce the isotopic pattern. The probability of exchange for a residue follows a binomial distribution where the "success probability" is given by the deuteration of the residue and is therefore a function of time. Hence, the isotopic pattern can be calculated as the product of binomial probability distributions (one per amino acid) further convoluted with the natural abundance of elements. In practice, HDSite derives single residue protection factors only when the uptake of a residue can be calculated as the difference in uptake of two peptides, otherwise an averaged value is returned. Therefore, the method strongly depends on the data set, and the prediction is limited by the number of peptides available and their overlap. Analogously to HDSite, other methods aim to extract single residue information from HDX-MS data by fitting the isotopic envelopes of peptides.[15−17] A method more recently proposed (ExPfact)[20] simultaneously fits the uptake curves of contiguous overlapping peptides with multiexponential curves (eq 4), determining all alternative patterns of protection factors compatible with experimental data. This method can be applied to any data set, and the ambiguity on the predicted protection factors provides a measurement of the degree of underdetermination of single residue properties. A similar approach has been implemented by pyHDX,[13] HDXModeller,[14] HR-HDXMS,[18] and HDX Workbench.[19]

In this article, we analyze a data set previously published,[36] containing sparse HDX data from MS and NMR measurements under the same experimental conditions for the small monomeric mouse prion protein (103 amino acids). Using ExPfact,[20] we show that a discrete number of sets of protection factors can be extracted from sparse HDX-MS data, that the ambiguity on the estimate can be reduced when a proper temporal sampling is coupled with minimal overlap, and completely removed by exploiting the additional information contained in the isotopic envelopes *a posteriori*. The extracted protection factors correlate with NMR measurements, with discrepancies providing insights on the compatibility between the two techniques as well as strengths and limitations of the statistical approach implemented.

METHODS AND MATERIALS

Data Set. The measurements analyzed here were previously published[36] and probed the mouse prion protein (103 residues) at pH 4 and a temperature of 25 °C at different urea concentrations. To ensure the validity of the EX2 approximation (and thus of eq 3 and eq 4), we focused on the exchange of the protein in its native state (i.e., in the absence of urea). In the HDX-MS experiment, the exchange was quenched at pH 2.4 and a temperature of 0 °C, and the

https://doi.org/10.1021/jasms.2c00005
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

protein was digested by pepsin, providing a data set (Figure 1) that includes 14 peptides covering most of the sequence (75/103 residues were covered) but with marginal overlap. Six regions covered by contiguous overlapping peptides were identified. The exchange was monitored at 15 exchange times ranging from 5 s to 24 h, and the experiment was conducted in triplicate. Data are available in the Supporting Information.

The fractional deuterium uptake of a peptide D was calculated as the intensity-weighted average (centroid) of the isotopic envelope at a specific time D_t and was normalized using the centroid of the experimentally fully deuterated sample D_{FD} (which was lower than the theoretical, fully deuterated centroid because of back exchange) and the centroid of the fully protonated sample $D_{0\%}$:

$$D = \frac{D_t - D_{0\%}}{D_{FD} - D_{0\%}} \tag{5}$$

The fractional uptake was then averaged over the three replicates.

HDX was also measured by NMR under the same experimental conditions, and exchange rates were derived for 34 amino acids. A subset of 27 residues was covered by both data sets. Because NMR experiments were performed only once, we assumed that the protection factors provided by NMR represent their true values.

Prediction of Protection Factors from HDX-MS Data. ExPfact is a computational method aiming to extract protection factors at the resolution of the single amide.[20] Considering regions covered by contiguous overlapping peptides one by one, the method finds multiple solutions of a system of equations (the size of which depends on the number of overlapping peptides in each region, and each equation has the functional form in eq 4), and then clusters these, reducing the degeneracy and providing a discrete number of alternative averaged solutions.

To find one possible solution, we performed a best fit on the experimental data. The experimental fractional uptake D_j^{exp} was simultaneously fitted for every peptide j at every time point t_k with eq 4 (D_j^{pred}), and the set of protection factors $\{P_i\}$ was adjusted to minimize the cost function

$$C(\lambda, \{P_i\}) = \underbrace{\sum_j \sum_k w_{jk}[D_j^{pred}(t_k,\{P_i\}) - D_j^{exp}(t_k)]^2}_{\text{SSR}} + \underbrace{\lambda \sum_i (\ln(P)_{i-1} - 2\ln(P)_i + \ln(P)_{i+1})^2}_{\text{penalty term}} \tag{6}$$

The cost function in eq 6 consists of a regular term, the sum of squared residuals (SSR), which depends on the experimental data, and a penalty term, which was introduced to avoid overfitting and, given the correlation between exchange rates and structure of the protein,[26,27] to disfavor large variations in the protection factors of adjacent residues. Gaps between peptides and prolines do not influence the penalty term, which is set to 0 unless ln P_{i-1}, ln P_i, and ln P_{i+1} are simultaneously greater than 0 ($\ln(P)$ is set to -1 for prolines and for any residue not covered by peptides). The penalty constant was set to $\lambda = 10^{-8}$ after cross validation (Supplementary Figure 2). Following the recommendations for the propagation of error in HDX-MS data,[37] a pooled standard deviation can be associated with each measure; therefore, the weights w_{jk} are all equal. When reliable error estimates are available—which is unlikely the case when the number of replicates is limited to three—then it is more accurate to consider the weights as the inverse of the standard deviation. The cost function in eq 6 represents a rough fitting landscape, and depending on the initial guess for the set $\{P_i\}$, the minimization algorithm converges to different local minima. When not specified, the initial guess is chosen through a random search: 10 000 sets of protection factors are randomly initialized with the constraint $0 < \ln(P) \leq 20$, and the set with the best agreement with experimental data (i.e., with the lowest cost function) is selected as the initial guess for a least-squares minimization. To explore alternative local minima in the fitting landscape, and thus to calculate several possible solutions, this minimization procedure is repeated 5000 times. To reduce the degeneracy of the sets of protection factors, we applied a clustering algorithm based on Gaussian mixture models (GMM), implemented in the R package mclust.[38] The histograms of the predicted protection factors, which are often multimodal (Figure 2), are combined into an

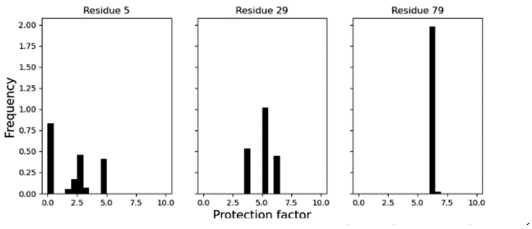


Figure 2. Histograms of protection factors predicted for residues 5 (left), 29 (center), and 79 (right) from 5000 minimizations. In most cases, histograms are multimodal distributions: Three modes can be identified for residue 5; 4 modes for residue 29.

M-dimensional probability distribution (M being the length of the region covered by overlapping peptides), which is fitted with a mixture of Gaussians with variable means and covariances (Supplementary Figure 3). The clustering algorithm returns a finite number of clusters of sets of $\{P_i\}$, each one in agreement with HDX-MS experimental data. The final number of identified clusters is determined by BIC (Bayesian information criterion). The minimization procedure is repeated until the addition of new solutions does not alter the outcomes of the clustering algorithm.

Performances. One minimization procedure requires on average 12.7 s on the data set analyzed here (processor: Intel Xeon W-1290P 3.7 GHz) using the default tolerance parameter (--tol), which controls the convergence of the algorithm. To speed up the process, the code was parallelized to run on multiple cores (parameter --ncores). Splitting the calculations over four cores is sufficient to complete 5000 minimizations in less than 5 h. We recommend running ExPfact overnight and setting up the number of minimizations and the tolerance parameter according to the computational power available.

Prediction of Isotopic Envelopes. For a peptide, the fractional deuterium uptake at time t (Figure 1B,C) is the mean of the centroids of the isotopic envelopes of different replicates. However, the same centroid value corresponds to different isotopic envelopes depending on the deuterium uptake of individual amino acids. Isotopic envelopes estimated from a predicted set of protection factors provide additional information to select the correct solution among all those that fit the time evolution of the centroid of each isotopic envelope.

https://doi.org/10.1021/jasms.2c00005
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

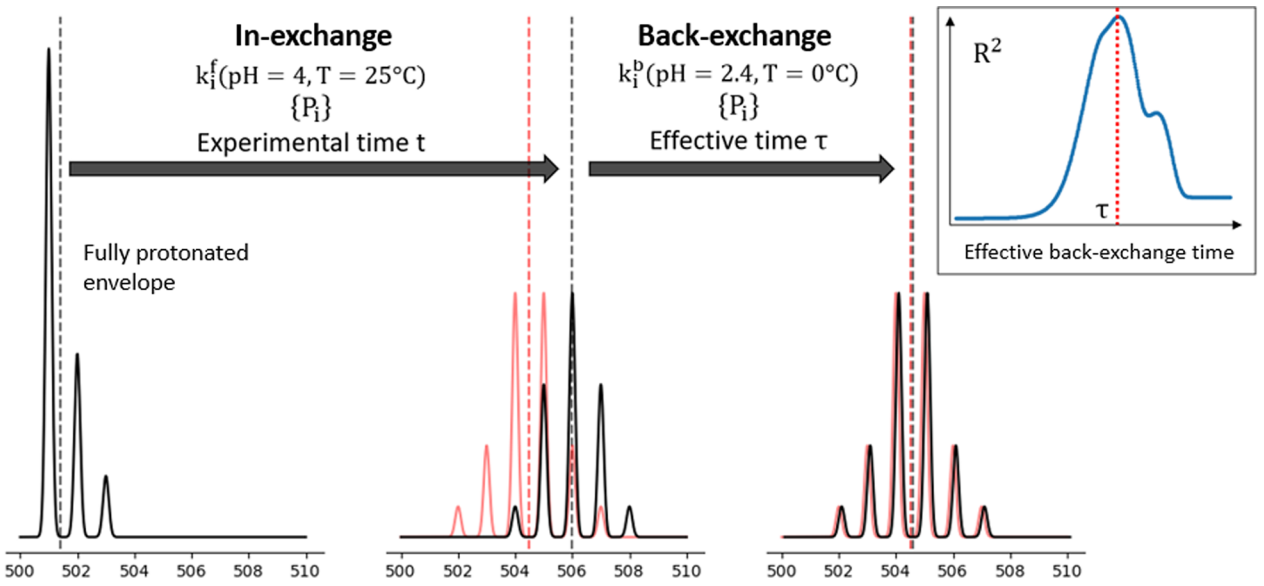


Figure 3. Schematic representation of the calculations for the reproduction of the experimental isotopic envelope. The fully protonated envelope can be calculated from the knowledge of the peptide sequence. The isotopic envelope at time t is evaluated by applying the evolution in eq 7 to the fully protonated envelope, using a specific pattern of protection factors $\{P_i\}$ and the intrinsic "forward" exchange rates k_i^f calculated at pH 4 and a temperature of 25 °C for a protonated protein in a deuterated buffer. The in-exchange predicts an envelope (black) that lies at higher m/z values with respect to the experimental spectrum (red); vertical dashed lines indicate the centroid of the envelopes. To correct for back exchange, the evolution in eq 7 is applied toward protonation, using the same $\{P_i\}$ and the intrinsic "back"-exchange rates k_i^b calculated under quenching conditions (pH 2.4 and temperature 0 °C) for a deuterated protein in water. The back-exchange evolution is applied for an effective back-exchange time τ, which maximizes the agreement between the predicted and the experimental envelope (insert).

To simulate the time evolution of the isotopic envelope of a peptide formed by n exchangeable residues, we need to calculate the probability that k residues have exchanged at time t:

$$\Pi(k, t) = \sum_{\substack{A\subset\{1,\dots,n\} \\ |A|=k}} \prod_{i\in A} d_i(t) \prod_{j\in\{1,\dots,n\}\setminus A} (1 - d_j(t)) \tag{7}$$

Equation 7 can be built following these considerations: (i) The probability of a residue to exchange is a function of time and is given by eq 3; (ii) the probability of k residues to have exchanged is the product of their individual probabilities (assuming they are independent events); (iii) the probability that only k residues of an n-residue peptide have exchanged is given by the probability that k residues have exchanged times the probability that $n - k$ residues have not exchanged; (iv) the calculations in points (i)−(iii) must be summed over all possible combinations of k residues in the n-residue peptide.

The isotopic envelope of a peptide can be calculated by applying the evolution in eq 7 to the fully protonated envelope of the peptide (calculated using the python library pyOpenMS[39]). Given the intensity of the fully protonated envelope π_i for a species with isotope number i, the simulated intensity of the isotopic envelope at time t is given by $\pi_i\Pi(k{=}0,t) + \pi_{i-1}\Pi(k = 1,t) + \dots + \pi_{i-N}\Pi(k{=}N,t) = \sum_{j=0}^{N}\pi_{i-j}\Pi(j,t)$, where the species $i - N$ corresponds to the monoisotopic mass of the peptide.

To calculate the shape of the isotopic envelope at time t from a set of $\{P_i\}$, the evolution in eq 7 was applied until deuteration time t, that is, toward higher m/z values, using the intrinsic exchange rates calculated at a temperature of 25 °C and pH 4, the conditions at which the experiment was performed, for a protonated protein in a deuterated buffer. However, the predicted envelope always appeared at higher m/z values with respect to the experimental ones because the deuteration in eq 7 does not account for back exchange. Back exchange occurs at the protein level in the labeling buffer, which is never 100% deuterated (generally 90−95% D_2O), and in the quench buffer before injection into the pepsin column, after which back exchange also occurs at the peptide level. To reproduce the shape of the isotopic envelope, we applied the evolution in eq 7 toward protonation, that is, toward lower m/z values, using the same set of $\{P_i\}$ and the intrinsic exchange rates calculated at a temperature of 0 °C and pH 2.4 for a deuterated protein in a protonated buffer. This back-exchange correction was applied for the "effective back-exchange time" τ that minimizes the difference between the predicted and the experimental shape. The underlying assumption is that back exchange can be modeled analogously to "in exchange" (i.e., using the multiexponential in eq 4). We used the predicted envelopes to discriminate whether some pattern of protection factors was able to better reproduce the shape of the isotopic envelope; the agreement was evaluated with R^2. The procedure for the prediction of isotopic envelopes is summarized in Figure 3.

RESULTS

The protection factors derived from HDX-MS measurements probing the mouse prion protein in its native state are shown in Figure 4A with their associated error. The value(s) and the error(s) associated with protection factors derived from MS measurements are the mean(s) and standard deviation(s) of the Gaussian cluster(s). For most of the sequence, a single cluster is found (i.e., all possible solutions correspond to a single cluster, see Methods and Materials), whereas multiple clusters are found in the two regions (residues 5−9 and 27−30) in which proteolytic peptides do not overlap (Figure 4B,C). In both regions, the NMR protection factors fall within

https://doi.org/10.1021/jasms.2c00005
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

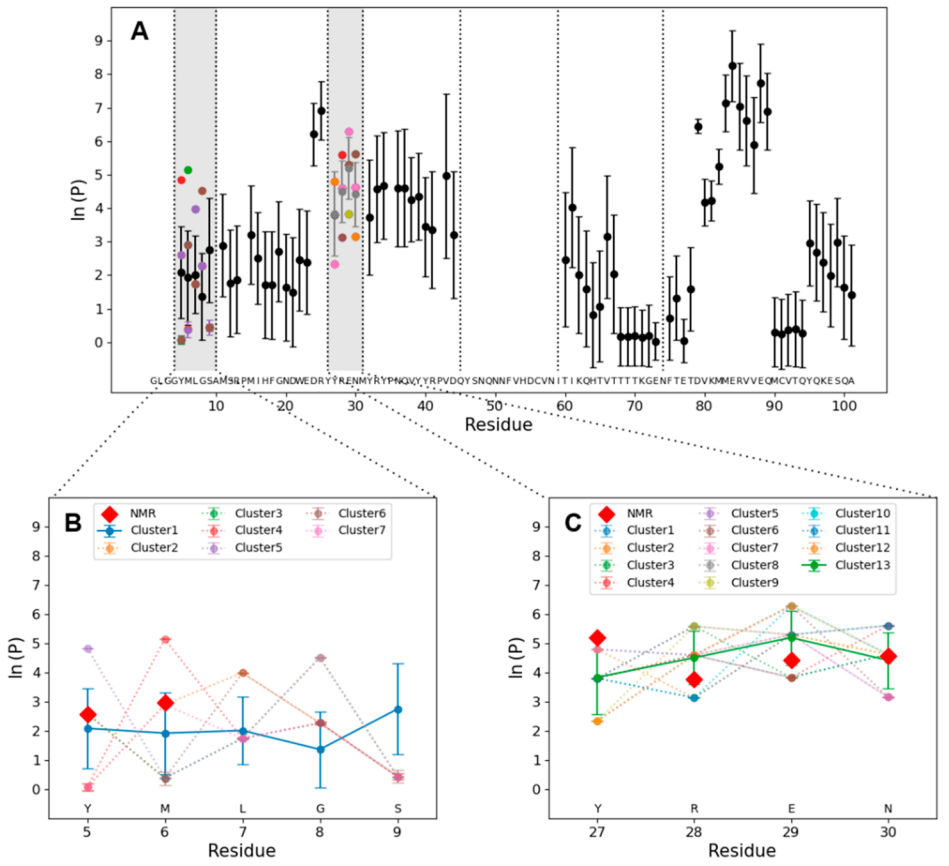


Figure 4. (A) Protection factors corresponding to cluster means of 5000 least-squares solutions obtained by minimizing the cost function in eq 6. Vertical dashed lines show regions covered by isolated (gray) or overlapping peptides (compare to Figure 1). Dots and error bars represent the mean and standard deviation of the estimated clusters. In regions where multiple clusters are identified, different clusters are shown with different colors. (B, C) Comparison of the estimated clusters with protection factors from NMR (red diamonds) in the regions where multiple clusters are identified; clusters compatible with NMR measurements, namely, clusters 1 and 13 in the regions covered by residues 5−9 and 27−30, respectively, are highlighted.

1σ (1 standard deviation) of the MS estimation. The predicted profiles of protection factors reflect the known structural properties of the protein. Indeed, higher protection against exchange is observed at helices $\alpha 1$ (residues 21−30) and $\alpha 3$ (residues 77−101), with completely unprotected residues surrounding Cys91, which forms a disulfide bond with Cys56. Lower protection is also observed in the loop between $\alpha 2$ and $\alpha 3$ (residues 72−76).

The shape of the experimental isotopic envelope can be exploited to define the quality of each cluster of solutions. We show the results for peptide 1 (residues 4−9, Figure 1), where the clustering algorithm identified seven clusters (Figure 4). We randomly select a set of $\{P_i\}$ from each cluster and predict the isotopic envelope as discussed in the Methods and Materials section. The outcomes (Figure 5) show that the solutions belonging to cluster 1, which was the only cluster compatible with NMR measurements, can reproduce the shape of the experimental isotopic envelope better than any other cluster. This proves that the isotopic envelopes encode a greater amount of information relative to centroided data, and that this information can be used *a posteriori* to reduce the ambiguity on the estimated value of protection factors.

Protection factors for 27 residues covered by the MS data set are also available from NMR measurements. We were able to extract single residue protection factors from MS centroided data for all but two regions (Figure 4). Moreover, we were able to assess the quality of different solutions in one of these regions, therefore deriving one "top-scoring" pattern of protection factors (Figure 5). The region covered by peptide 4 (Figure 1) remains underdetermined (Figure 4C) because the experimental isotopic envelopes for this peptide were not available. In this region, we selected cluster 13 as the final pattern of protection factors because it showed compatibility with NMR measurements. The comparison between protection factors extracted by MS and NMR (Figure 6A,B) showed a high degree of compatibility between the protection factors extracted by the two techniques, with 23/27 values compatible with at most 2σ, and a correlation coefficient $\rho = 0.71$ (Figure 6D) when the outlier residues 25, 91, and 94 (which are not compatible within 3σ) are not considered.

DISCUSSION

Despite the only partial coverage provided by the MS data set (Figure 1), we showed how alternative patterns of protection factors with similar agreement to experimental data can be accurately derived at the resolution of a single amino acid (Figure 4A). Moreover, the solutions can be clustered, providing a discrete number of alternative solutions for $\{P_i\}$. In most regions, one unique cluster was identified. Two regions still present ambiguity in the final estimate of the protection factors, but at least one of the clusters identified in these regions is compatible with protection factors derived

https://doi.org/10.1021/jasms.2c00005
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

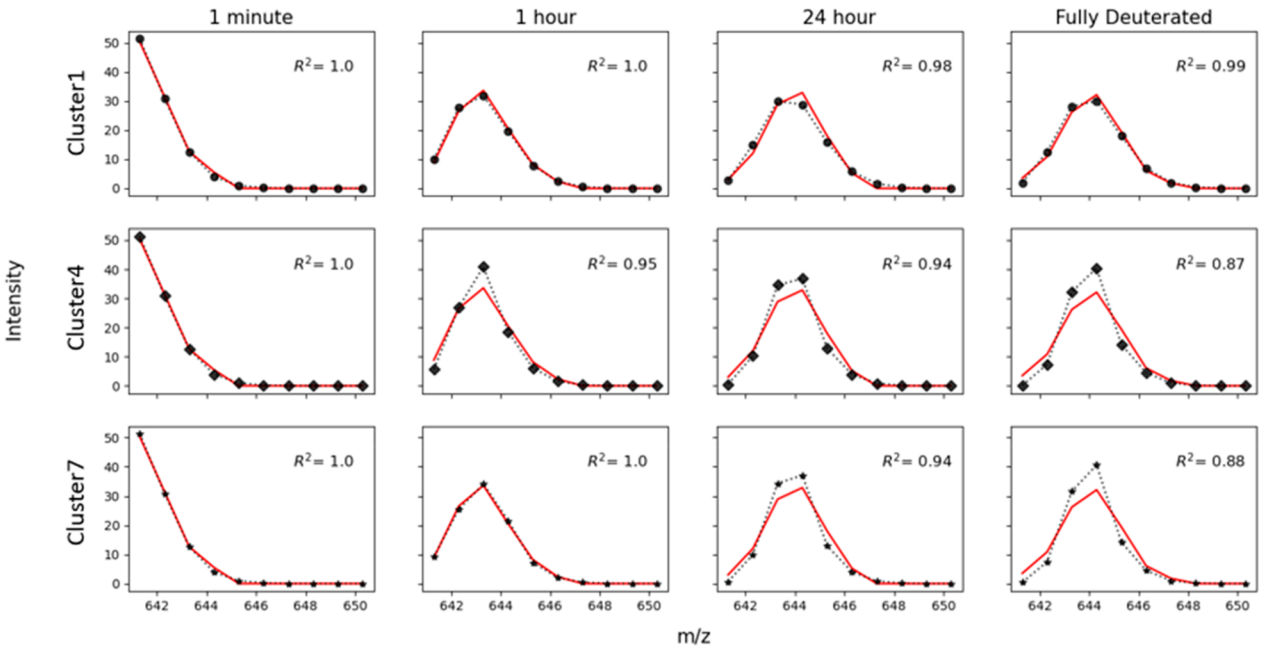


Figure 5. Prediction of isotopic envelopes. Starting from the fully protonated envelope of peptide 1 (sequence YMLGSA), the evolution in eq 5 is applied toward deuteration at times 1 min (column 1), 1 h (column 2), 24 h (column 3), and infinite time (column 4) using intrinsic exchange rates calculated at pH 4 and a temperature of 25 °C and a set of protection factors belonging to cluster C1 (row 1), C4 (row 2), and C7 (row 3). A back-exchange correction is performed by applying eq 5 toward protonation, using intrinsic exchange rates calculated at pH 2.4 and a temperature of 25 °C and the same set of protection factors. The isotopic envelope predicted using protection factors from Cluster1 (black dots), Cluster4 (black diamonds), and Cluster7 (black stars) is compared to the experimental envelopes (red lines). The agreement is evaluated using R^2.

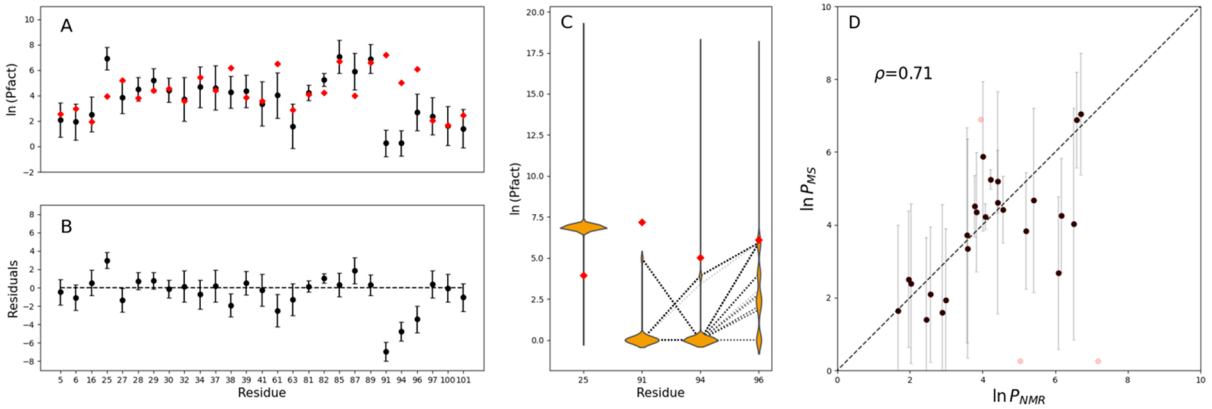


Figure 6. Comparison of protection factors from HDX-MS and HDX-NMR experiments. (A) Clusters of protection factors extracted from HDX-MS data (black dots with error bars) are compared to NMR measurements (red diamonds) for every amino acid covered by both data sets. (B) Residuals of protection factors from HDX-MS and HDX-NMR experiments. (C) Marginal probability distribution of protection factors derived from 5000 minimization procedures for residues 25, 91, 94, and 96. (D) Correlation between protection factors extracted by NMR and MS; Pearson's correlation coefficient ρ = 0.71 excluding outliers (red).

from NMR measurements (Figure 4B,C). Nonetheless, the ambiguity could be completely removed for one of these two regions by exploiting the supplementary information contained in the shape of the experimental isotopic envelope (Figure 5). Therefore, the method used here estimates protection factors from MS data alone (with the exception of the region covered by residues 27−30, where the experimental isotopic envelope was not available).

A comparison of the protection factors estimated from MS with measures from NMR showed a high degree of compatibility (Figure 6), validating the method. The four discrepancies shown by residues 25, 91, 94, and 96 provide insight into the limitations of the data sets and the computational approach. The protection factor of residue 25 is compatible within 3σ with the NMR measurement. Interestingly, the marginal probability distribution of the protection factors estimated for residues 91 and 94 is bimodal, with one of the modes similar to the NMR measurements (Figure 6C). The GMM clustering algorithm selects the final number of components based on the minimum BIC = $k \ln(n) - 2 \ln(\hat{L})$, where n is the number of data points, $\hat{L}$ the maximized value of the likelihood function of the model, and k the number of parameters estimated by the model. Therefore, the BIC tends to favor models with fewer parameters. Considering the low-intensity peaks as outliers of the main distribution leads to a lower BIC than considering them as separate modes of a multimodal distribution. This artifact is even more evident when we look at protection factors of

https://doi.org/10.1021/jasms.2c00005
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

residue 94, which has a multimodal probability distribution with four modes; moreover, one of the modes is compatible with the NMR measurement. Even in this case, the BIC is lower when the projected multimodal distribution is merged into one component.

A univariate clustering approach (i.e., a clustering algorithm considering one residue at a time instead of regions covered by contiguous overlapping peptides, Figure S3) would find the low-intensity peaks approaching NMR measurements shown in Figure 6C for residues 91, 94, and 96. However, a multivariate approach is statistically and physically more rigorous because the protection factor of a residue depends on its neighbors. Indeed, there is not a single pattern of $\{P_i\}$ found by the minimization procedure containing simultaneously all those three values (black dotted lines in Figure 6C show the subset of solution with $4 < \ln(P_{91}) < 6$ or $2.5 < \ln(P_{94}) < 5$). Moreover, a set of $\{P_i\}$ with protection factors of residues 91, 94, and 96 equal to NMR measurements did not fit the uptake curves of the HDX-MS data set. To prove this, we constrained protection factors in the region 75−101 to their NMR value (when available) during the least-squares minimization, whereas the remaining protection factors are adjusted to minimize the cost function in eq 6. For peptide 12, which contains residues 91, 94, and 96, a best fit provides a prediction in deuterium uptake, which is not compatible with MS measurements (Figure 7). The analysis of these discrepancies

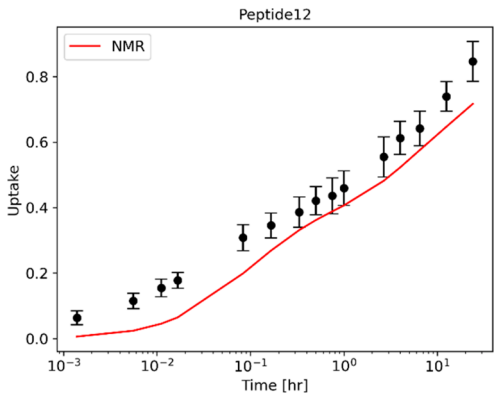


Figure 7. Deuterium uptake prediction for peptide 12 using an optimized set of protection factors with constrained NMR values. In the region covering residues 76−101, the protection factor of 11 residues was measured by NMR. These values are fixed, whereas the remaining protection factors are optimized to minimize the cost function in eq 6. The resulting prediction (red line) is not compatible with MS data.

suggests that an estimation of the same quantity (i.e., the protection factor) from two different techniques is not possible here because the error is either unknown (in the NMR data set) or too large (in the MS data set). The disagreement is however localized in a specific region of the protein and could therefore be caused by artifacts in either the NMR or MS experiment. In the absence of additional measurements, these results cannot be interpreted further.

CONCLUSIONS

In this article, we applied ExPfact[20] to a previously published data set by probing the HDX of the same protein under the same experimental conditions by both MS and NMR.[36] The novelties introduced with respect to the previous publication are (i) the validation of the method via a comparison to NMR data, which is often neglected in related papers;[12−19] (ii) the prediction of the experimental isotopic envelope of peptides (via the back-exchange correction) as another tool to assess the quality of alternative solutions; (iii) several upgrades to the code (introduction of the penalty term, parallelization of the code, additional scripts and tests, and extended documentation).

The approach demonstrated here enables the quantitative analysis of any HDX-MS data set (in the EX2 regime), providing protection factors at the resolution of the single amino acid. We note that the protection factor is a well-defined quantity only when both the native and EX2 approximations are valid. When EX1 kinetics (or mixed EX1/EX2 kinetics) emerge from the isotopic distribution of peptides, single residue information can be extracted via other methods.[17,40] The information extracted in the two regimes is different. For the exchange in EX2 conditions, a protection factor can be extracted: this is a unitless quantity that can be expressed with Gibb's free energy of opening: $\Delta G_{op} = RT \ln P$ (where R is the universal gas constant and T is the temperature).[1] In the case of EX1 kinetics, the exchange of a single residue is $d_{EX1}(t) = 1 - e^{-k_o t}$; therefore, it is possible (in principle) to extract the opening rate k_o of a residue, which has the units of $[\text{time}]^{-1}$ and can be expressed through the Eyring equation[41] as proportional to Gibb's free energy of activation: $k_o = (k_B T/\hbar)\exp\left(-\frac{\Delta G_o^{\ddagger}}{RT}\right)$ (where k_B is the Boltzmann constant and $\hbar$ is Planck's constant).[7] ExPfact aims to extract protection factors from HDX-MS because they encode structural information on the protein, and is consequently limited to the study of data sets with peptides showing EX2 behavior.

HDX-MS is a promising technique for high-throughput and low-cost characterization of proteins' structural and dynamic properties. The principal drawback of the technique is its spatial resolution, providing data at the peptide level, which so far are mostly interpreted qualitatively. The implementation of alternative MS/MS fragmentation methods not affected by H/D scrambling—such as electron capture/transfer dissociation (ExD) and UV photodissociation (UVPD)[9−11] —would be a valuable addition to experimentally increase spatial resolution. However, we believe that single residue resolution will be hardly achieved for the whole sequence of the protein. Therefore, computational methods aiming to extract information at higher resolution will remain essential. The efforts made by the HDX-MS community to acquire higher quality data[32−34] combined with a unified computational approach encompassing the knowledge acquired in the past decade[12−20] will enable HDX-MS data analysis to overcome the obstacle of limited spatial resolution, providing a unique "quick and cheap" experimental validation to assess models from *ab initio* structure determination methods such as AlphaFold.[42]

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jasms.2c00005.

Additional methodological details, including figures and examples (PDF)

https://doi.org/10.1021/jasms.2c00005
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

AUTHOR INFORMATION

Corresponding Author

Emanuele Paci − *School of Molecular and Cellular Biology, University of Leeds, LS2 9JT Leeds, United Kingdom; Dipartimento di Fisica e Astronomia, Università di Bologna, 40127 Bologna, Italy;* orcid.org/0000-0002-4891-2768; Email: e.paci@unibo.it

Authors

Michele Stofella − *School of Molecular and Cellular Biology, University of Leeds, LS2 9JT Leeds, United Kingdom; Dipartimento di Fisica e Astronomia, Università di Bologna, 40127 Bologna, Italy*

Simon P. Skinner − *School of Molecular and Cellular Biology, University of Leeds, LS2 9JT Leeds, United Kingdom*

Frank Sobott − *School of Molecular and Cellular Biology, University of Leeds, LS2 9JT Leeds, United Kingdom;* orcid.org/0000-0001-9029-1865

Jeanine Houwing-Duistermaat − *Dipartimento di Scienze Statistiche, Università di Bologna, 40127 Bologna, Italy;* orcid.org/0000-0002-4505-7137

Complete contact information is available at: https://pubs.acs.org/10.1021/jasms.2c00005

Author Contributions

M.S., E.P., and J.H. designed the analysis; M.S., E.P., and F.S. performed the analysis; M.S. and S.P.S. wrote and maintained the code; M.S., E.P., F.S., and J.H. wrote the manuscript.

Notes

The authors declare no competing financial interest.

The scripts and data needed to reproduce the results shown in this article are available on GitHub at the following link: https://github.com/pacilab/exPfact.

ACKNOWLEDGMENTS

We thank R. Moulick and J. Udgaonkar for providing the raw data corresponding to results published in ref 36.

REFERENCES

(1) Englander, S. W.; Mayne, L.; Kan, Z.-Y.; Hu, W. Protein Folding—How and Why: By Hydrogen Exchange, Fragment Separation, and Mass Spectrometry. *Annu. Rev. Biophys.* **2016**, *45*, 135−152.

(2) Linderstrøm-Lang, K. The pH-dependence of the deuterium exchange of insulin. *Biochim. Biophys. Acta* **1955**, *18*, 308.

(3) Dempsey, C. E. Hydrogen exchange in peptides and proteins using NMR spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.* **2001**, *39* (2), 135−170.

(4) 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.

(5) Deng, B.; Lento, C.; Wilson, D. J. Hydrogen deuterium exchange mass spectrometry in biopharmaceutical discovery and development − A review. *Anal. Chim. Acta* **2016**, *940*, 8−20.

(6) James, E. I.; Murphree, T. A.; Vorauer, C.; Engen, J. R.; Guttman, M. Advances in Hydrogen/Deuterium Exchange Mass Spectrometry and the Pursuit of Challenging Biological Systems. *Chem. Rev. (Washington, DC, U. S.)* **2021**, DOI: 10.1021/acs.chemrev.1c00279.

(7) Hamuro, Y. Tutorial: Chemistry of Hydrogen/Deuterium Exchange Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2021**, *32* (1), 133−151.

(8) Hamuro, Y. Quantitative Hydrogen/Deuterium Exchange Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2021**, *32*, 2711.

(9) Pan, J.; Han, J.; Borchers, C. H. Top-down hydrogen/deuterium exchange and ECD-stitched FTICR-MS for probing structural dynamics of a 29-kDa enzyme. *Int. J. Mass Spectrom.* **2012**, *325−327*, 130−138.

(10) Mistarz, U. H.; Bellina, B.; Jensen, P. F.; Brown, J. M.; Barran, P. E.; Rand, K. D. UV Photodissociation Mass Spectrometry Accurately Localize Sites of Backbone Deuteration in Peptides. *Anal. Chem.* **2018**, *90* (2), 1077−1080.

(11) Pan, J.; Han, J.; Borchers, C. H.; Konermann, L. Hydrogen/Deuterium Exchange Mass Spectrometry with Top-Down Electron Capture Dissociation for Characterizing Structural Transitions of a 17 kDa Protein. *J. Am. Chem. Soc.* **2009**, *131* (35), 12801−12808.

(12) Kan, Z.-Y.; Walters, B. T.; Mayne, L.; Englander, S. W. Protein hydrogen exchange at residue resolution by proteolytic fragmentation mass spectrometry analysis. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (41), 16438−16443.

(13) Smit, J. H.; Krishnamurthy, S.; Srinivasu, B. Y.; Parakra, R.; Karamanou, S.; Economou, A. Probing universal protein dynamics using hydrogen-deuterium exchange mass spectrometry-derived residue-level Gibbs free energy. *Anal. Chem.* **2021**, *93* (38), 12840−12847.

(14) Salmas, R. E.; Borysik, A. J. HDXmodeller: an online webserver for high-resolution HDX-MS with auto-validation. *Commun. Biol.* **2021**, *4* (1), 199.

(15) Babic, D.; Kazazic, S.; Smith, D. M. Resolution of protein hydrogen/deuterium exchange by fitting amide exchange probabilities to the peptide isotopic envelopes. *Rapid Commun. Mass Spectrom.* **2019**, *33* (15), 1248−1257.

(16) Zhang, Z.; Zhang, A.; Xiao, G. Improved protein hydrogen/deuterium exchange mass spectrometry platform with fully automated data processing. *Anal. Chem.* **2012**, *84* (11), 4942−4949.

(17) Zhang, Z. Complete extraction of protein dynamics information in hydrogen/deuterium exchange mass spectrometry data. *Anal. Chem.* **2020**, *92* (9), 6486−6494.

(18) Gessner, C.; Steinchen, W.; Bedard, S.; Skinner, J. J.; Woods, V. L.; Walsh, T. J.; Bange, G.; Pantazatos, D. P. Computational method allowing hydrogen-deuterium exchange mass spectrometry at single amide resolution. *Sci. Rep.* **2017**, *7* (1), 3789.

(19) Saltzberg, D. J.; Broughton, H. B.; Pellarin, R.; Chalmers, M. J.; Espada, A.; Dodge, J. A.; Pascal, B. D.; Griffin, P. R.; Humblet, C.; Sali, A. A residue-resolved Bayesian approach to quantitative interpretation of hydrogen-deuterium exchange from mass spectrometry: application to characterizing protein-ligand interactions. *J. Phys. Chem. B* **2017**, *121* (15), 3493−3501.

(20) Skinner, S. P.; Radou, G.; Tuma, R.; Houwing-Duistermaat, J. J.; Paci, E. Estimating Constraints for Protection Factors from HDX-MS Data. *Biophys. J.* **2019**, *116* (7), 1194−1203.

(21) Molday, R. S.; Englander, S. W.; Kallen, R. G. Primary structure effects on peptide group hydrogen exchange. *Biochemistry* **1972**, *11* (2), 150−158.

(22) Bai, Y.; Milne, J. S.; Mayne, L.; Englander, S. W. Primary structure effects on peptide group hydrogen exchange. *Proteins: Struct., Funct., Genet.* **1993**, *17* (1), 75−86.

(23) Connelly, G. P.; Bai, Y.; Jeng, M.-F.; Englander, S. W. Isotope effects in peptide group hydrogen exchange. *Proteins: Struct., Funct., Bioinf.* **1993**, *17* (1), 87−92.

(24) Englander, S. W.; Sosnick, T. R.; Englander, J. J.; Mayne, L. Mechanisms and uses of hydrogen exchange. *Curr. Opin. Struct. Biol.* **1996**, *6* (1), 18−23.

(25) Nguyen, D.; Mayne, L.; Phillips, M. C.; Englander, S. W. Reference parameters for protein hydrogen exchange rates. *J. Am. Soc. Mass Spectrom.* **2018**, *29* (9), 1936−1939.
(26) Vendruscolo, M.; Paci, E.; Dobson, C. M.; Karplus, M. Rare Fluctuations of Native Proteins Sampled by Equilibrium Hydrogen Exchange. *J. Am. Chem. Soc.* **2003**, *125* (51), 15686−15687.
(27) Best, R. B.; Vendruscolo, M. Structural interpretation of hydrogen exchange protection factors in proteins: characterization of the native state fluctuations of CI2. *Structure (Oxford, U. K.)* **2006**, *14* (1), 97−106.
(28) Ferraro, D. M.; Lazo, N. D.; Robertson, A. D. EX1 Hydrogen Exchange and Protein Folding. *Biochemistry* **2004**, *43* (3), 587−594.
(29) Fitzkee, N. C.; Torchia, D. A.; Bax, A. Measuring rapid hydrogen exchange in the homodimeric 36 kDa HIV-1 integrase catalytic core domain. *Protein Sci.* **2011**, *20* (3), 500−512.
(30) Barnes, C. A.; Shen, Y.; Ying, J.; Takagi, Y.; Torchia, D. A.; Sellers, J.; Bax, A. Remarkable Rigidity of the Single α-Helical Domain of Myosin-VI As Revealed by NMR Spectroscopy. *J. Am. Chem. Soc.* **2019**, *141* (22), 9004−9017.
(31) Bishop, C. M.; Nasser, N. M. *Pattern Recognition and Machine Learning*; Springer: New York, 2006; Vol. *4*, issue 4.
(32) Rob, T.; Liuni, P.; Gill, P. K.; Zhu, S.; Balachandran, N.; Berti, P. J.; Wilson, D. J. Measuring Dynamics in Weakly Structured Regions of Proteins Using Microfluidics-Enabled Subsecond H/D Exchange Mass Spectrometry. *Anal. Chem.* **2012**, *84* (8), 3771−3779.
(33) Kish, M.; Smith, V.; Subramanian, S.; Vollmer, F.; Lethbridge, N.; Cole, L.; Bond, N. J.; Phillips, J. J. Allosteric regulation of glycogen phosphorylase solution phase structural dynamics at high spatial resolution. *bioRxiv* **2019**, 654665.
(34) Svejdal, R. R.; Dickinson, E. R.; Sticker, D.; Kutter, J. P.; Rand, K. D. Thiol-ene Microfluidic Chip for Performing Hydrogen/Deuterium Exchange of Proteins at Subsecond Time Scales. *Anal. Chem.* **2019**, *91* (2), 1309−1317.
(35) Kan, Z.; Ye, X.; Skinner, J. J.; Mayne, L.; Englander, S. W. ExMS2: An Integrated Solution for Hydrogen−Deuterium Exchange Mass Spectrometry Data Analysis. *Anal. Chem.* **2019**, *91* (11), 7474−7481.
(36) Moulick, R.; Das, R.; Udgaonkar, J. B. Partially Unfolded Forms of the Prion Protein Populated under Misfolding-promoting Conditions. *J. Biol. Chem.* **2015**, *290* (42), 25227−25240.
(37) Weis, D. D. Recommendations for the Propagation of Uncertainty in Hydrogen Exchange-Mass Spectrometric Measurements. *J. Am. Soc. Mass Spectrom.* **2021**, *32*, 1610.
(38) Scrucca, L.; Fop, M.; Murphy, T. B.; Raftery, A. E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J.* **2016**, *8* (1), 289−317.
(39) Röst, H. L.; Schmitt, U.; Aebersold, R.; Malmström, L. pyOpenMS: A Python-based interface to the OpenMS mass-spectrometry algorithm library. *Proteomics* **2014**, *14* (1), 74−77.
(40) Na, S.; Lee, J.; Joo, J. W. J.; Lee, K. J.; Paek, E. deMix: decoding deuterated distributions from heterogeneous protein states via HDX-MS. *Sci. Rep.* **2019**, *9* (1), 3176.
(41) Eyring, J. The activated complex in chemical reactions. *J. Chem. Phys.* **1935**, *3* (2), 107−115.
(42) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.

https://doi.org/10.1021/jasms.2c00005
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX