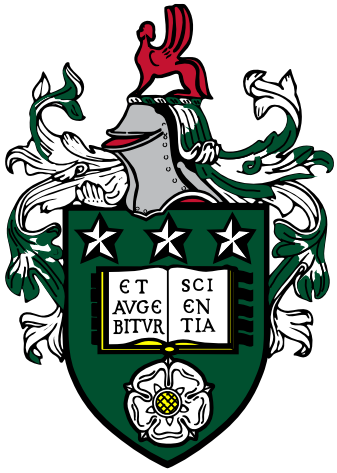
Michele Stofella

Pushing the boundaries: towards single-residue resolution hydrogen/deuterium exchange mass spectrometry



University of Leeds

Astbury Centre for Structural and Molecular Biology

School of Molecular and Cellular Biology

Submitted in accordance with the requirements of the degree of

Doctor of Philosophy

September 2024

To my aunt Carla

Intellectual Property and Publication Statements

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly authored publications has been included. The candidate's contribution and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

List of jointly-author publications included in the thesis

Chapter 2

Michele Stofella, Antonio Grimaldi, Jochem H. Smit, Jurgen Claesen, Emanuele Paci and Frank Sobott, "Computational tools for hydrogen/deuterium exchange mass spectrometry data analysis", accepted with minor corrections at Chemical Reviews (2024)

Author contributions statement. MS, EP and FS designed the review; MS prepared the draft for the manuscript; all authors reviewed the paper and approved the final version of the manuscript.

Chapter 3

Michele Stofella, Simon P. Skinner, Frank Sobott, Jeanine Houwing-Duistermaat and Emanuele Paci, "High-resolution hydrogen–deuterium protection factors from sparse mass spectrometry data validated by nuclear magnetic resonance measurements", Journal of the American Society for Mass Spectrometry (2022), doi: https://doi.org/10.1021/jasms.2c00005

Author contributions statement. MS, EP, and JH designed the analysis; MS, EP, and FS performed the analysis; MS and SPS wrote and maintained the code; MS, EP, FS, and JH wrote the manuscript.

Chapter 4

Michele Stofella, Neeleema Seetaloo, Alexander N. St John, Emanuele Paci, Jonathan J. Phillips, and Frank Sobott, "Recalibrating protection factors using millisecond hydrogen/deuterium exchange mass spectrometry", under review at Analytical Chemistry (2024)

Author contributions statement. MS, FS and JJP. designed the study; MS and NS performed and analysed HDX-MS experiments; ANSJ, MS and EP performed and analysed MD simulations; MS prepared the draft for the manuscript; all authors reviewed the results and approved the final version of the manuscript.

Chapter 5

Michele Stofella, Antonio Grimaldi, Matthew Batchelor, Bob Schiffrin, Anastasia Zhuravleva, Emanuele Paci and Frank Sobott, "Protection factor analysis for hydrogen/deuterium exchange mass spectrometry", paper draft (2024)

Author contributions statement. MS, EP and FS designed the project. MB and RB expressed and purified the sample. MS performed the HDX-MS experiments. MS, AG, EP and FS analysed the HDX-MS data. MS, BS, and AZ performed and analysed the HDX-NMR data. MS prepared the draft for the manuscript. All authors reviewed the results and approved the final version of the manuscript.

Acknowledgements

In December 2021, a year after moving to the UK, I began to understand that my time in Leeds held a purpose far greater than the pursuit of a PhD. God had led me into the quiet solitude of the Yorkshire moors to reach the deepest recesses of my heart. I found resonance in the words of the prophet Hosea: "Look, I am going to seduce her and lead her into the desert and speak to her heart". The desert: a place of immense trials, but also of marvellous wonders!

As this journey comes to an end, I would like to extend my deepest thanks.

To **Prof Frank Sobott** – thank you for your understanding when, overwhelmed by a challenging time, I came to you on the brink of quitting my PhD. Your support, our hikes during the pandemic, and the freedom you gave me to design our research have been invaluable. Thank you for your guidance with conferences and external collaborations, and for the countless hours spent planning and discussing our science. To **Prof Emanuele Paci –** thank you for believing in me, in my ideas and in my capabilities. Your mentorship helped me navigate the intricated labyrinth of academia. I appreciated your philosophical insights into science, your unwavering commitment to rigor (always pushing for that one final proof), and the countless discussions, both in person and online, about the project and beyond. To **Dr Anton Calabrese** – thank you for your technical support in the laboratory and the moral support over coffee.

To the people of the Catacombs and the MS group, especially to **Tiago Moreira** – thank you for tireless work and constant willingness to sacrifice your own time to assist the group. To the (virtual) Pacilab team – I am grateful to **Dr James Ross**, **Dr Matthew Batchelor** and **Antonio Grimaldi** for their invaluable contributions. I also acknowledge the University of Leeds for funding me during this research project, and all internal and external collaborators – thank you for the precious contributions that made this work possible: **Dr Jonathan Phillips**, **Dr Neeleema Seetaloo**, **Dr Anastasia**

Zhuravleva, **Dr Bob Schiffrin**, **Prof Jeanine Houwing-Duistermaat**, **Dr Jochem H. Smit**, **Dr Jurgen Claesen**, **Simon P. Skinner**; and yes, even ChatGPT 3.5 (OpenAI), for assisting with proof-reading parts of this work.

To my housemates **Dr Alex St John**, **Miles Deans** and **Dr Kyle Le Huray** – thank you for your daily companionship during this journey, for our fancy dinners, our board game nights and our adventures, from San Sebastian to Helm's Deep (Gordale Scar).

To all my brothers and sisters in faith of the communities of Wythenshawe and the North-West of England, especially to **David** and **Clara**, **Oscar, Noemi**, **Michaela**, **Daniel**, **Mattia**, **Simone** and **Fr Jim.** I also want to thank the community of Boncellino – thank you for your support and prayers, with a special thanks to my catechists for their spiritual guidance during these years: **Sandro** and **Marina**, **don Walter**, **Nadia**, **don Giuseppe** and **Antonio**. May you receive a hundredfold of what you have given me.

To my family. To my mum **Daniela** and my dad **Sandro**, thank you for encouraging me to pursue this PhD while constantly reminding me to take care of myself because there is more to life than just studying or working. To my brother **Lorenzo** and my sisters **Chiara** and **Anna**, for your unwavering support, especially during the lockdowns we shared. To my grandparents **Enzo** and **Barbara,** for your simplicity in asking me to explain my research in plain terms. To my grandparents **Ezio** and **Luciana,** for sparking in me a deep curiosity about the world. And to my aunt **Carla**, to whom this thesis is dedicated, for her incredible support from Heaven.

After destroying the One Ring, Frodo wraps his arm around Sam's shoulder and says: "I am glad to be with you, Samwise Gange, here at the end of all things". In much the same way, after submitting this One Thesis, I am profoundly grateful to all of you who have stood by my side throughout this pilgrimage.

Abstract

Hydrogen/deuterium exchange (HDX) is a spontaneous process observed in proteins exposed to deuterated solvent. The observed rate of exchange of a residue is the ratio between its 'intrinsic exchange rate', reflecting the chemical properties of its environment, and the 'protection factor', accounting for structural and dynamic properties of the protein. HDX kinetics provide useful insights into protein conformational dynamics, complementing static structures from X-ray crystallography, NMR or cryo-EM.

HDX is primarily studied using NMR and mass spectrometry (MS). HDX-NMR offers single-residue resolution but is limited to smaller proteins, while HDX-MS can study larger systems, though it captures deuteration at peptide-level resolution. While spatial resolution in HDX-MS can be enhanced (e.g. with alternative fragmentation techniques), achieving NMR-like detail remains challenging. As a result, HDX-MS is typically used for differential studies, comparing protein states to qualitatively localize perturbations (e.g. binding sites), rather than quantify absolute biophysical properties such as protection factors.

This thesis explores whether single-residue resolution can be inferred from peptide-level HDX-MS data, despite the inherent underdetermination of the data.

We critically reviewed the strengths and limitations of various strategies used by software tools for analysing HDX-MS data. Building on this, we demonstrated the advantages of a method that exploits self-consistency to identify those alternative protection factor sets which are compatible with experimental data, cluster them into a finite number of solutions, and reduce the degeneracy by incorporating the information encoded in isotopic envelopes. Since our method assumed intrinsic exchange rates as known constants (under fixed environmental conditions), we tested their accuracy by performing millisecond HDX-MS experiments on a mixture of

unstructured peptides. Finally, we proposed and tested against NMR data a protection factor analysis that addresses the ambiguity of peptide-level data by either performing a random search to broadly explore the rugged cost function landscape or incorporating an informed initial guess to focus on physically meaningful solutions.

Overall, our findings show that protection factors can be inferred from peptide-level HDX-MS data with sufficient redundancy and temporal sampling, providing insights comparable to NMR. These results push the boundaries of HDX-MS toward becoming an absolute quantitative experiment rather than a differential qualitative tool.

Table of Contents

List of Figures

Supplementary Figures

List of Tables

Supplementary Tables

List of Abbreviations

3-Ala	Three-alanine peptide
AFM	Atomic force microscopy
AI	Artificial intelligence
AMBER	Assisted model building with energy refinement
ANP	Atrial natriuretic peptide
BIC	Bayesian information criterion
CASP	Critical assessment of protein structure prediction
CD	Circular dichroism
CEM	Chain ejection model
CID	Collision induced dissociation
CRM	Charged residue model
Cryo-EM	Cryogenic electron microscopy
CV	Cross validation
DIA	Data independent analysis
DSSP	Define secondary structure of proteins
ECD	Electron capture dissociation
EM	Electron microscopy
ESI	Electrospray ionization
ETD	Electron transfer dissociation
ExD	Electron capture or transfer dissociation
FDR	False discovery rate
FORTRAN	Formula translation
FTICR	Fourier transform ion cyclotron resonance
GMM	Gaussian mixture models
GUI	Graphical user interface
HCD	Higer-energy C-trap dissociation
HDX	Hydrogen/deuterium exchange
HDX-MS	Hydrogen/deuterium exchange probed by mass spectrometry
HDX-NMR	Hydrogen/deuterium exchange probed by nuclear magnetic resonance
HOHAHA	Homonuclear Hartmann-Hahn spectroscopy
HPLC	High performance liquid chromatography
HSQC	Heteronuclear single quantum coherence spectroscopy

HX	Hydrogen exchange
IDP	Intrinsically disordered proteins
IEM	Ion evaporation model
IM-MS	Ion mobility mass spectrometry
IMP	Integral membrane protein
LC	Liquid chromatography
LC-MS	Liquid chromatography coupled with mass spectrometry
LC-MS/MS	Liquid chromatography coupled with tandem mass spectrometry
LFQ	Label free quantification
MCMC	Markov chain Monte Carlo
MD	Molecular dynamics
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
NMR	Nuclear magnetic resonance
NOESY	Nuclear Overhauser effect spectroscopy
OMSSA	Open mass spectrometry search algorithm
OPS	Overlapping peptide segmentation
PCA	Principal component analysis
PDB	Protein data bank
PDLA	Poly-DL-alanine
PLGS	Protein Lynx global server
PRIDE	Proteomics identification database
RJMCMC	Reverse jump Markov chain Monte Carlo
RNC	Ribosomal nascent chain
RT	Retention time
SEC	Size exclusion chromatography
SMA	Spectral mixture analysis
SSR	Sum of squared residuals
ToF	Time of flight mass spectrometer
UV	Ultraviolet
UVPD	Ultraviolet photodissociation

Chapter 1. General Introduction

1.1 Hydrogen/deuterium exchange: a language to decode protein structure, dynamics and function

Proteins are one of the most important gears in the engine of life. The importance of understanding their structure and dynamics became evident in 1962, when Max Perutz and John Kendrew were awarded the Nobel prize in Chemistry for their pioneering work in determining the structure of globular proteins (Kendrew et al. 1960; Perutz et al. 1960). Since then, proteins have led different branches of sciences to wonder how their linear sequence folds into a three-dimensional structure (or does not, like intrinsically disordered proteins), how they change upon binding, how they maintain health and cause disease (Dill and MacCallum 2012). After seeing the structure of myoglobin at 6 Å resolution, Kendrew commented: "Perhaps the most remarkable features of the molecule are its complexity and its lack of symmetry" (Kendrew et al. 1960).

The concept of 'structure' should be interpreted broadly. Proteins are not rigid entities: they can adapt to their ligands, they can be highly flexible and may exist in random coil conformations under physiological conditions; intrinsically disordered proteins, for example, can sometimes adopt a permanent structure upon binding, or remain biologically active in a disordered state (Serdyuk, Zaccai, and Zaccai 2007).

In 1972, the Nobel prize in Chemistry was given to Anfinsen for his work on ribonuclease. He proposed a ground-breaking postulate for protein folding, known as the thermodynamic hypothesis or Anfinsen's dogma, according to which the native state of a protein is determined by its amino acid sequence, at least for globular proteins (Anfinsen 1973). Today, his hypothesis can be tested using the Protein Data Bank

(Berman et al. 2000), which counts almost 225,000 [a] experimental structures of proteins. Anfinsen's dogma counts several exceptions: some proteins require other proteins (known as chaperones) to fold properly (Ellis 1993); prions are stable conformation of proteins which are different from the native state (Prusiner 1998); some proteins have multiple native structures, and their folding process changes according to external factors (Porter and Looger 2018). These exceptions suggest that the sequence of a protein does not uniquely define its structure, and that the dynamics of the chain plays a crucial role in determining what conformation the protein assumes. The exploration of these limitations led to the discovery and (partial) understanding of a whole new class of misfolding diseases, such as Alzheimer's, Parkinson's and type II Diabetes (Dill and MacCallum 2012).

The protein folding problem is central to protein biophysics and critical for much of cell biology, but despite extensive experimental and theoretical efforts, researchers have yet to reach a consensus on a universally accepted folding model (Sosnick and Barrick 2011). In 1969, Levinthal noted that a polypeptide chain cannot find its native state within any reasonable amount of time by a random search through the vast conformational space available to them (Levinthal 1968). He estimated that if a protein of 100 amino acids were to fold by randomly sampling all possible conformations, it would take an astronomically long time (10^{27} years [b]), longer than the age of the universe. Yet proteins fold in a time range spanning from microseconds to milliseconds. While giving a talk at Stanford, Levinthal exclaimed: "So there must be folding intermediates!" (Baldwin 2017). He imagined that proteins must fold through some programmed structure formation pathway. These intermediates, known as foldons, form in a reproducible sequence, building the native protein step by step. If the initial

[a] 224,004 structures in September 2024

[b] The Levinthal's estimation for the protein folding time is performed with the assumption that each amino acid has about three possible conformations (hence a protein formed by N amino acids has 3^N possible conformations) and that the protein samples a new conformation every 10^{-13} seconds (about the time for a single molecular vibration).

foldon is sufficiently small, the Levinthal paradox is solved. For example, if the foldon consists of 20 amino acids, the random search can be completed in less than a millisecond.

The "energy landscape theory" offers an alternative model for describing protein folding, rooted in hypothetical energy landscapes and statistical mechanical principles (Dill and Chan 1997; Plotkin and Onuchic 2002). According to this theory, a folding protein navigates a rugged energy landscape, with the global minimum representing its native state. Unlike the foldon model, which suggests a more linear pathway, the energy landscape theory allows for multiple trajectories, with the protein progressively moving through various local minima as it folds into its final structure.

Today, the protein folding problem must be discussed in the light of the artificial intelligence (AI) revolution. In 1997, Karplus (Nobel prize in Chemistry in 2013) highlighted the existence of two separate 'protein folding problems', the first consisting in the prediction of the three-dimensional structure of a protein from its primary sequence, the second in the kinetics and dynamics of the folding process. "A complete solution of the second problem would, of course, simultaneously solve the first. However, it is more likely that the prediction of the native structure will be achieved by other methods that are based on the analysis of known structures, rather than by directly folding a polypeptide chain" (Karplus 1997). He was right: almost 25 years later, the unprecedented results of the AlphaFold2 algorithm from DeepMind (Jumper et al. 2021; Senior et al. 2020) at the CASP14 [c] allowed protein science to take a huge step towards solving the first protein folding problem. However, the second protein folding problem remains an open question: how do proteins fold? As the American philosopher Emerson used to say: "It's not the destination, it's the journey!".

[c] CASP (Critical Assessment of protein Structure Prediction) is a community-wide, worldwide experiment for protein structure prediction, taking place every two years since 1994.

In a paper published in 2017, Englander stated that "experimental results provide clear evidence for the existence and ubiquity of protein foldons and their determining role in constructing well-defined protein folding pathways" (Englander and Mayne 2017). The technology that led Englander to this conclusion exploits a natural biophysical process that is universally experienced by all proteins when exposed to deuterated solvent: hydrogen/deuterium exchange (HDX). The main-chain amide hydrogens of proteins, one in every amino acid (except proline) in every protein molecule, spontaneously exchange with the deuterium atoms in the solvent. Their rate of exchange depends on a variety of environmental (pH, temperature, ionic strength) and protein (primary sequence, local- and long-range interactions, conformational changes, dynamics, energetics) parameters. In a real sense, "proteins continually emit signals in the language of hydrogen exchange, in a nonperturbing way resolved to the level of individual residues. We need only receive those signals and understand how to interpret them in structural language" (Englander et al. 2016).

A definitive solution to the protein folding problem remains far beyond the scope of this Thesis. The focus here is to highlight the potential of HDX in addressing fundamental questions that are central to molecular biology. Today, HDX is primarily used to map protein conformational changes under different experimental conditions, but its potential applications extend beyond these current uses.

1.1.1 The exchange of unstructured polypeptides…

When a peptide is diluted in a solution containing D_2O, its hydrogen atoms exchange with deuterium atoms from the solvent (Linderstrøm-Lang 1955). The exchangeable atoms of the peptides are those bound to heteroatoms such as sulfur, oxygen and nitrogen. The heteroatom-bound hydrogens undergo isotopic exchange due to the electronegativity of the heteroatom, while carbon-bonded hydrogens do not exchange to any measurable extent (Weis 2016). The exchangeable atoms in the side chains exchange rapidly and this property makes them invisible to the classic experimental techniques used to probe HDX, namely nuclear magnetic resonance (NMR, **section**

1.2.2) and mass spectrometry (MS, **section 1.2.3**). Note that the exchange of exchangeable hydrogens in the side chains can be probed by gas-phase MS (Suckau et al. 1993). The backbone amide hydrogens are sensitive and very useful probes of protein conformation, as they are distributed along the polypeptide backbone (except in prolines) and form the fundamental hydrogen-bonding network of basic secondary structures (e.g. alpha helices and beta-sheets). For these reasons and the difficulty of measuring the fast exchange of labile hydrogens in the side chains, HDX is modelled as a phenomenon occurring at the level of the individual residue, where the only exchangeable hydrogen that can be measured is the amide hydrogen of the amino acid.

The chemical nature of the hydrogen exchange (HX) or hydrogen/deuterium exchange (HDX) reactions can be described with the proton transfer model (Hamuro 2021b; Weis 2016). The initial event in the proton transfer reaction is the diffusion-limited collision of a proton donor (AH) with a proton acceptor (B) leading to the formation of a hydrogen-bonded complex which is followed by a rapid proton redistribution across the hydrogen bond and finally dissociation of the proton donor and acceptor complex:

Eq. 1.1
$$AH + B \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} A-H\cdots B \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} A\cdots H-B \underset{k_{-1}}{\overset{k_a}{\rightleftharpoons}} A + HB$$

The proton transfer is successful when the proton is transferred to the proton acceptor (dissociation to the right) and unsuccessful if the proton remains with the proton donor (dissociation to the left). The overall forward rate constant for the proton transfer reaction k_{tr} is given by:

Eq. 1.2
$$k_{tr} = k_1\left(\frac{10^{\Delta pK}}{10^{\Delta pK}+1}\right) = k_{int,cat}$$

where k_1 is the collisional rate and ΔpK is the difference in pK between the acceptor and the donor. Thus, the proton transfer rate k_{tr} is the rate of collisions multiplied by the fraction of successful collisions (Englander and Kallenbach 1983). Amide HX can be acid (H_3O^+), base (OH^-) or water (H_2O) catalyzed. Thus, for each specific proton

transfer reaction, k_{tr} is equal to an intrinsic exchange rate constant $k_{int,cat}$ defined for the given catalyst of the reaction, that is the proton acceptor (OH^-, H_3O^+ or H_2O) and the proton donor (amide NH).

The effect of pH on the rate of exchange in an unstructured polypeptide can be written as a linear combination of the proton transfer rates for acid-, base- and water-reactions, regulated by rates $k_{tr,acid}$, $k_{tr,base}$ and $k_{tr,water}$, respectively:

Eq. 1.3 $$k_{int} = k_{tr,acid}[H^+] + k_{tr,base}[OH^-] + k_{tr,water}[H_2O]$$

The transfer rates were determined by Bai et al (Bai et al. 1993) for a poly-DL-alanine (PDLA) peptide at temperature 20°C and at low salt concentrations. In **Chapter 4**, we show that PDLA peptides are not an unstructured reference as a small number of alanine amino acids is sufficient to form double (or even triple) helical bundles. We also suggest that a three-alanine (3-Ala) peptide should be used as unstructured reference. The proton transfer rates for 3-Ala (still at temperature 20°C and low salt concentration) are $k_{tr,acid} = 1.09 \times 10^2\ M^{-1}min^{-1}$, $k_{tr,base} = 2.29 \times 10^{10}\ M^{-1}min^{-1}$ and $k_{tr,water} = 3.16 \times 10^{-2}\ M^{-1}min^{-1}$. The dependence of the intrinsic exchange rate k_{int} on pH (**Eq. 1.3**) has a V-shaped curve with a characteristic minimum at pH 2.5-3.0 (**Figure 1.1A**).

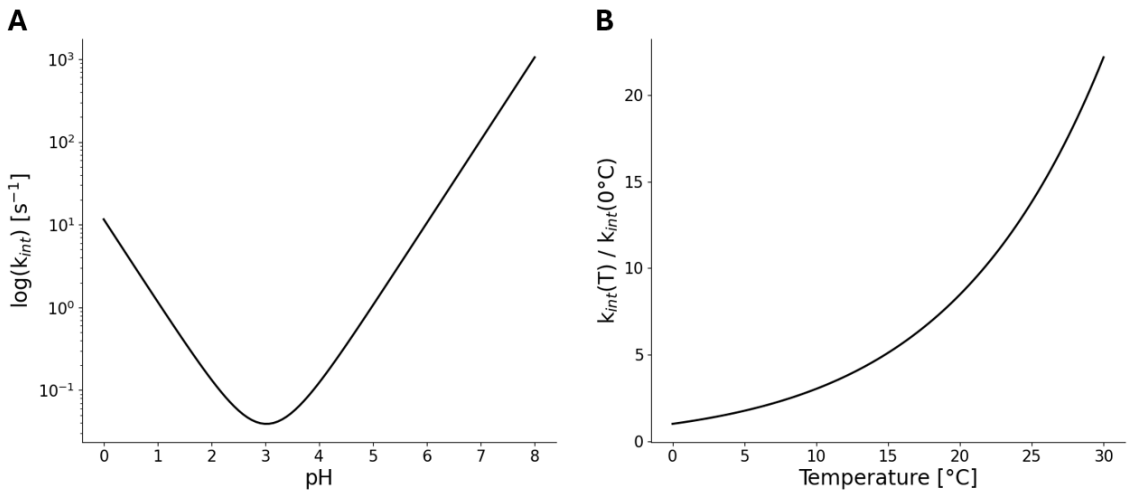


Figure 1.1. Profile of the intrinsic exchange rate dependence on pH and temperature.
(A) Intrinsic exchange rate dependence on pH as dictated by **Eq. 1.3** for an unfolded peptide with sequence AAA at temperature 25°C and pH varying from 0 to 8. (B) Intrinsic exchange rate dependence on temperature as shown in **Eq. 1.4** for an unfolded peptide with sequence AAA at pH 7 and varying temperature from 0 to 30°C. Calculations performed using a python script available at https://github.com/pacilab/exPfact.

The exchange rate is also dependent on the temperature at which the labelling is performed. This is primarily due to the fact that an increase in temperature alters the water ionization constant, consequently increasing the concentration of OH^-. The dependence of the intrinsic exchange rate k_{int} on temperature can be expressed with an Arrhenius equation, which has an exponential behavior (**Figure 1.1B**):

Eq. 1.4 $$k_{int}(T) = k_{int}(293)\exp\left(-\frac{E_a}{R}\left[\frac{1}{T} - \frac{1}{293}\right]\right)$$

Here, $k_{int}(293)$ is the reference rate constant ($k_{tr,acid}$, $k_{tr,base}$ and $k_{tr,water}$) at temperature 293 K; E_a is the activation energy for acid-, base- or water-catalyzed exchange ($E_{a,acid} = 14$ kcal mol^{-1}, $E_{a,base} = 17$ kcal mol^{-1} and $E_{a,water} = 19$ kcal mol^{-1}); R is the gas constant ($R = 8.134$ J mol^{-1} K^{-1}). The Arrhenius equation (**Eq. 1.4**) is valid for relatively low temperatures (0-60 °C), provided that the protein

remains stable, while it needs to be adjusted at higher temperatures (Tajoddin and Konermann 2020).

Molday (Molday, Englander, and Kallen 1972) noted that the exchange rate of a residue surrounded by two isoleucines was much slower (up to 20 times depending on pH) than the rate of the same residue when surrounded by two alanines. This suggested that the intrinsic exchange rate of a residue in an unfolded peptide depends on the side chains of the neighboring amino acids. The effect of the different side chains on the exchange of a residue was systematically studied and tabulated by Bai (Bai et al. 1993). Assuming that the effect is significant only for the sidechains of the first neighbors, these empirical side-chain specific factors allow to rewrite the intrinsic exchange rate (**Eq. 1.3**) as

Eq. 1.5
$$k_{int} = k_{tr,acid}(A_l \cdot A_r)[H^+] + k_{tr,base}(B_l \cdot B_r)[OH^-] + k_{tr,water}(B_l \cdot B_r)[H_2O]$$

where A_l, A_r, B_l and B_r are the acid (A) or base (B) factor for the residue on the left (l) or right (r) of the central residue.

These considerations on the effect of pH, temperature and neighboring side chains on the exchange rate of a residue in a completely unfolded structure allow the prediction of the intrinsic exchange of a residue. We show in **Chapter 4** that these predictions are accurate (when a 3-Ala peptide is used as unstructured reference instead of extended PDLA peptides, which retain some residual structure), and that some additional dependences, such as the ionic strength of the buffer, can affect the intrinsic exchange rate of a residue. An accurate prediction of the intrinsic exchange rate allows the deconvolution of the effect of the chemical environment on the experimental rate of a folded protein from the (more significant) effect of structural and dynamic properties of the protein.

1.1.2 ... and the exchange of folded proteins

While the exchange of unstructured polypeptides (**section 1.1.1**) occurs in milliseconds, the exchange of folded proteins occurs in minutes, hours or even days (Weis 2016). Amide hydrogens are exchange competent when they are surface exposed and not engaged in secondary structure (when they are not hydrogen-bonding other than to water). Some residues are structurally protected against exchange, but local fluctuations, global and sub-global folding events can momentarily disrupt internal amide hydrogen bonding, exposing them to the solvent and enabling deuteration (Englander et al. 2016). Referring to these collective dynamic fluctuations as "breathing motions", which allow the protein to "open" (with rate k_{op}) or "close" (with rate k_{cl}) to the exchange, Linderstrøm-Lang modelled hydrogen/deuterium exchange of each residue of a protein as a two-step process (Linderstrøm-Lang 1955):

Eq. 1.6
$$NH_{cl} \underset{k_{cl}}{\overset{k_{op}}{\rightleftharpoons}} NH_{op} \xrightarrow{k_{int}} ND$$

where NH_{cl}, NH_{op} and ND represent the amide hydrogen in a closed state (or exchange incompetent), opened state (exchange competent) and the deuterated state.

The exact analytical solution for the model in **Eq. 1.6** is double exponential (Hvidt and Nielsen 1966). However, under the so-called native approximation ($k_{op} \ll k_{cl}$, i.e. mostly folded) and the EX2 regime ($k_{int} \ll k_{cl}$, i.e. k_{int} is the rate-limiting step), the deuteration of a residue can be approximated as a single exponential

Eq. 1.7
$$d(t) = 1 - e^{-\frac{k_{int}}{P}t}$$

where the pseudo-(pre)equilibrium constant $P \equiv k_{cl}/k_{op}$ is known as protection factor. A protection factor $P = 1$ corresponds to a fully unstructured polypeptide, while higher values of P describe an increasing protection of the amide against exchange. The

protection factor can be written in terms of Gibb's free energy of opening ΔG_{op} (**Figure 1.2A**) (Hamuro 2021b):

Eq. 1.8 $$\Delta G_{op} = RT\ln(P)$$

where R is the universal gas constant and T is the temperature.

While the EX2 regime described above is the most common, in rare cases the residue may follow the EX1 regime, where the opposite relation between the intrinsic and closing rate holds (i.e., $k_{int} \gg k_{cl}$). In this case, the deuteration of a single residue can be written as

Eq. 1.9 $$d_{EX1}(t) = 1 - e^{-k_{op}t}$$

In this second case, the energetic diagram for the reaction is different (**Figure 1.2B**), and the opening rate can be expressed as proportional to the Gibb's free energy of activation $\Delta G_0^{\ddagger}$ (rather than the Gibb's free energy of opening in the EX2 regime) via the Eyring equation (Eyring 1935):

Eq. 1.10 $$k_{op} = \frac{k_B T}{h} \exp\left(-\frac{\Delta G_0^{\ddagger}}{RT}\right)$$

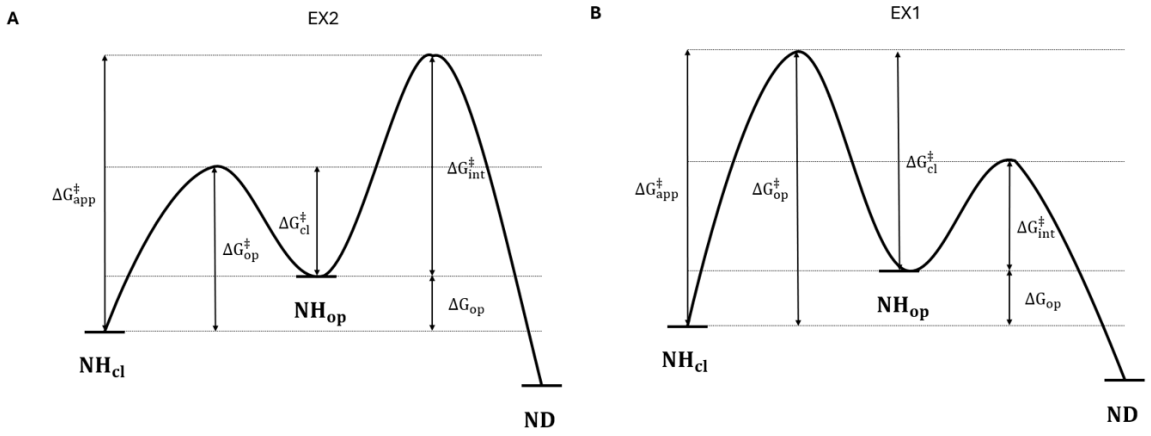


Figure 1.2. Energy diagrams for hydrogen/deuterium exchange reactions.

Energy reaction diagrams for hydrogen/deuterium exchange reactions in the EX2 (A) and EX1 (B) regime. The rate constants $\boldsymbol{k_{op}}$, $\boldsymbol{k_{cl}}$ and $\boldsymbol{k_{int}}$ correspond to the activation free energies $\Delta \boldsymbol{G^{\ddagger}_{op}}$, $\Delta \boldsymbol{G^{\ddagger}_{cl}}$ and $\Delta \boldsymbol{G^{\ddagger}_{int}}$ via the Eyring equation (**Eq. 1.10**). Adapted with permission from Hamuro, "Tutorial: Chemistry of Hydrogen/Deuterium Exchange Mass Spectrometry", 2021, Journal of the American Society for Mass Spectrometry, Copyright © 2021 American Chemical Society.

1.1.3 What is a protection factor?

The Linderstrøm-Lang model provides an energetic interpretation of the protection factor as the Gibb's free energy of opening (**Eq. 1.8**). This interpretation is exclusive to residues following the EX2 regime. A complete characterization of the HDX kinetics of a residue requires the knowledge of the opening and closing rates. However, theoretical models trying to determine these rates are missing, and the experimental techniques used today to probe HDX (**section 1.2**) are far from providing direct measurements of such parameters. Hence, the protection factor remains the *finest* parameter (i.e. at the highest resolution) that can be extracted from experimental data. Several microscopic models have tried to explain the protection factors of a protein in terms of its structural properties (Devaurs, Antunes, and Borysik 2022). The importance of the protection factors in describing the exchange kinetics of proteins is supported by the fact that the vast majority of proteins follows the EX2 regime. The EX1 kinetics, where the protection factor is not a well-defined parameter, is limited to rare cases (Fang et al. 2011).

The first (and still most used) model to describe the protection factors of a protein starting from its structure was developed in the early 2000s by Best, Paci and Vendruscolo (Best and Vendruscolo 2006; Vendruscolo et al. 2003). The model describes the protection factor of a residue for a protein in conformation X as the linear combination of the number of heavy contacts N_c (i.e. the number of atoms in the proximity of the amide hydrogen not belonging to neighbouring residues in the primary sequence) and hydrogen bonds N_h in which the amino acid is involved:

Eq. 1.11 $$\ln(\mathrm{P}) = \beta_{\mathrm{h}}\mathrm{N_h}(\mathrm{X}) + \beta_{\mathrm{c}}\mathrm{N_c}(\mathrm{X})$$

The model was initially tested against a dataset containing seven proteins with known protection factors (measured with NMR, see **section 1.2.2**), providing coefficients $\beta_h = 2.00$ and $\beta_c = 0.35$ (Best and Vendruscolo 2006). The value of the parameters suggest a strong influence of dynamic parameters (hydrogen-bonding) on the protection against exchange rather than structural parameters (heavy contact) (Hamuro 2024).

Alternative models have been developed to connect the structure of a protein to its protection factors, without remarkably improving the correlation with experimental data found by the Best-Vendruscolo model (**Eq. 1.11**). These alternatives can be classified in five groups depending on the structural property exploited (Devaurs et al. 2022):

i) Models based on solvent accessibility only (Petruk et al. 2013; Shan et al. 2013; Truhlar et al. 2006): they provide an intuitive structural interpretation of the protection factors as dependent only on the number of heavy contact but show poor correlation with experimental data.

ii) Models based on hydrogen-bonding only (McAllister and Konermann 2015; Skinner et al. 2012, 2014): they describe the protection factor of a residue using only the number of hydrogen bonds formed by the residue, they provide a good correlation with experimental data, suggesting that hydrogen-bonding is the main factor driving protection against exchange, but this is not enough to completely characterize the exchange.

iii) Electrostatic calculations (Abdolvahabi et al. 2014; Hernández, Anderson, and LeMaster 2009; LeMaster, Anderson, and Hernández 2009): they can explain the protection against exchange for residues that are not involved in hydrogen-bonding, but they are force-field dependent and computationally intensive.

iv) Combination of molecular features (Brand et al. 2007; Sljoka and Wilson 2013; Sowole et al. 2013): they are linear combinations of structural and nonstructural properties of the residue within the protein (such as bond

order or hydrophobicity), but they tend to increase the model complexity without significantly improving the correlation with experimental data.

v) Knowledge-based predictions (Claesen and Politis 2019; Tartaglia, Cavalli, and Vendruscolo 2007; Wang et al. 2018): they attempt to estimate the protection factors directly from the sequence of the protein but show poor correlation with experimental data.

Estimating the protection factors of a protein is the ultimate goal of any HDX experiment, as they allow to quantitatively connect the Linderstrøm-Lang theory with structural and dynamic properties of the protein.

1.2 Detecting hydrogen/deuterium exchange: a technical challenge

1.2.1 The dawn of HDX

The history of hydrogen/deuterium exchange begins in 1932 with the discovery of deuterium (Urey, Brickwedde, and Murphy 1932), which awarded Urey the Nobel prize in Chemistry and made the production of D_2O possible. The first studies probing proteins' secondary structure exploiting hydrogen/deuterium exchange were performed in the 1950s by Linderstrøm-Lang and coworkers at the Carlsberg Laboratory in Denmark (Baldwin 2011). In a world where the first protein structure had yet to be discovered, Linderstrøm-Lang was motivated to discover hydrogen-bonded structures in proteins. He knew that amide hydrogens of proteins undergo exchange with solvent water and that a density gradient column could measure density differences of 1×10^{-6} g/cm^3. He hypothesized that internally bonded amide hydrogens would exchange more slowly than free amide hydrogens, and that he could count the exact number of free amides. He applied his method for free peptides and insulin. He was indeed able to count the number of free amides, but their HDX kinetics resulted in a more complex picture than expected (Hvidt and Linderstrøm-Lang 1954). The question on how to disentangle the contribution of individual amides to the overlapping kinetic curves discouraged further studies, also because in the meantime

X-ray crystallography made it possible to obtain protein structures experimentally (Perutz et al. 1960).

However, on the opposite side of the world, in New Hampshire, Englander wanted to answer the question whether hydrogen bonding could explain the stable helical conformation of polyproline II. He knew about the technique developed by Linderstrøm-Lang, but he had one problem: polyproline II was exchanging too fast. Therefore, he developed a ^{3}H-gel filtration method, which was faster, more sensible and more accurate than the density gradient method (Englander 1963).

This gave a new boost to research around HDX, but the method was still not able to provide information about the exchange of individual amides. It was not until the advent of two-dimensional NMR that significant attention was drawn to the phenomenon. The coupling with mass spectrometry further propelled its impact, sparking widespread interest as testified by the exponential increase of scientific papers published on the topic (**Figure 1.3**).

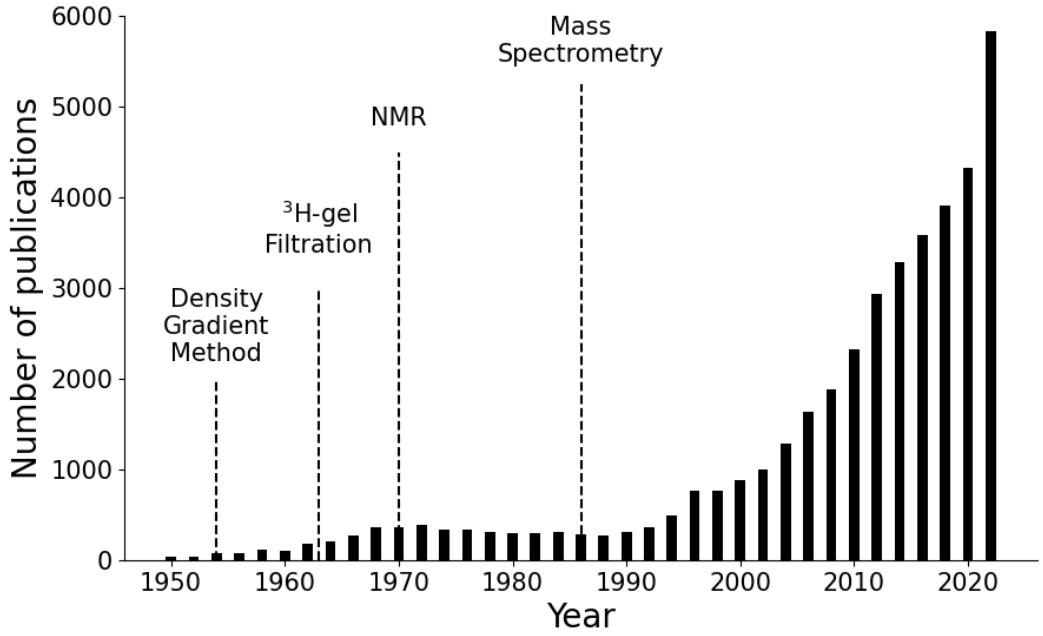


Figure 1.3. Number of publications on hydrogen/deuterium exchange in time.

Number of scientific papers about hydrogen/deuterium exchange published every two years. Vertical lines indicate key years in technological development. Source: google scholar; search terms: "hydrogen deuterium exchange".

1.2.2 Nuclear Magnetic Resonance

The advent of two-dimensional NMR has sparked renewed interest in studying the HDX of proteins, as it addressed a key limitation of previous technologies: the ability to monitor the exchange at the level of individual residues (Dempsey 2001). When a peptide chain or a protein is diluted in D_2O, the amide signals in the NMR spectrum decrease in intensity as the exchange occurs. This decrease in intensity is due to the distinct magnetic properties of hydrogen and deuterium: hydrogen has a spin of ½, while deuterium has a spin of 1. Additionally, their resonating frequencies differ, with hydrogen resonating at 400 MHz and deuterium at 61.4 MHz. Since NMR spectrometers operate over a narrow frequency range, they can only detect one isotope at a time. If the spectrometer is calibrated on the ^{1}H resonating frequency, ^{2}H (D) is invisible, and the deuteration of a residue results in a decreasing signal over time.

The ability to measure exchange rates for individual residues depends primarily on the resolution and assignment of amide signals or cross peaks in multidimensional spectra, as well as the time resolution of NMR acquisition (Dempsey 2001). The optimal approach for monitoring HDX kinetics in proteins involves acquiring HSQC spectra from uniformly ^{15}N-labelled proteins. These spectra offer good resolution in the ^{15}N dimension, can be obtained relatively quickly (within 15-30 minutes or faster depending on the protein concentration) (Andrec, Hill, and Prestegard 1995), allowing to study the slow exchange of folded proteins at physiological pH as well as of unstructured peptides at lower pH. In cases where a ^{15}N-labelled protein is unavailable, alternative methods (such as NOESY or HOHAHA spectra) may be employed (Dempsey 2001), though with reduced signal resolution, sensitivity, and time resolution. Other techniques may be used to track fast exchange processes (with exchange rates of $0.1 - 50\ s^{-1}$) instead of HSQC spectra, e.g. exploiting magnetization transfer (Wójcik et al. 1999).

NMR experiments were the first to enable the measurement of exchange rates at the level of individual residues. The pioneering work of the Englander group, which examined the dependence of exchange rates for unstructured polypeptides on factors such as pH, temperature, and neighboring side chains (**section 1.1.1**), was conducted using NMR (Bai et al. 1993; Connelly et al. 1993; Molday et al. 1972). HDX-NMR experiments allowed for the first time the estimation of protection factors of residues in a folded structure. Indeed, today Start2Fold (Pancsa et al. 2016), a database containing protection factors of different proteins [d], is available. The microscopic models discussed in section **1.1.3** were also optimized on NMR experimental data. However, NMR is primarily suited to studying relatively small protein systems, with a size limit of around 50 kDa (Yu 1999), As molecular weight increases, NMR spectra become more complex, leading to overlapping amide peaks that are difficult to distinguish and assign.

1.2.3 Mass Spectrometry

In 2002, Fenn was awarded the Nobel prize for the development of electrospray ionization (ESI), a breakthrough that revolutionized the analysis of large biomolecules through mass spectrometry. As Fenn famously remarked during his Nobel lecture, "We made molecular elephants fly" (Fenn 2003). The integration of ESI with mass spectrometry expanded the scope of analyzable biomolecular systems, ranging from small from small molecules to big complexes (up to the megadalton scale (Heck and van den Heuvel 2004)), thereby addressing the limitations inherent in NMR spectroscopy.

A fundamental requirement for mass spectrometric analysis is the conversion of analytes into gaseous charged ions (Serdyuk et al. 2007). ESI achieves this by producing multiply charged species, typically in the form of $[M + zH]^{z+}$ ions, where the charge state is considerably greater than one ($z \gg 1$). These high charge states facilitate ion dissociation in tandem MS experiments, therefore providing an optimal ground to

[d] Start2Fold contains protection factors of 57 proteins (September 2024)

couple ESI with liquid chromatography (LC). In ESI, the analyte solution is introduced into a metal capillary, to which a high voltage (of several kV) is applied (**Figure 1.4A**) (Konermann et al. 2013). The tip of the capillary is distorted to create a Taylor cone that emits a fine spray of droplets (with radii in the micrometer range). In positive ion mode, the droplets acquire a positive charge due to the presence of excess cations (such as H^+, NH_4^+, Na^+, K^+) and the generation of protons at the metal/solution interface. As solvent evaporation causes the droplets to shrink, their charge density increases until the Rayleigh limit is reached, i.e. the point at which the surface tension of the droplet is balanced by the Coulombic repulsion (Rayleigh 1882). At this critical limit, the droplets fragment into smaller, highly charged particles through one of three distinct mechanisms **Figure 1.4B**:

i) Ion evaporation model (IEM) (Iribarne and Thomson 1976): low molecular weight species are emitted when the electric field at the surface of a Rayleigh-charged droplet becomes sufficiently strong to eject solvated ions directly from the droplet.

ii) Charged residue model (CRM) (Iavarone and Williams 2003; Kebarle and Verkerk 2009): for larger, globular species, the droplet continues to evaporate until only the analyte remains. When the last solvent layer disappears, the remaining charge on the droplet is transferred to the analyte molecule.

iii) Chain ejection model (CEM) (Ahadi and Konermann 2011; Konermann, Rodriguez, and Liu 2012): unfolded proteins or peptides, characterized by disordered conformations with exposed nonpolar residues, migrate toward the droplet surface due to their hydrophobicity. At the surface, one chain terminus is ejected, followed by a stepwise sequential ejection of the whole chain.

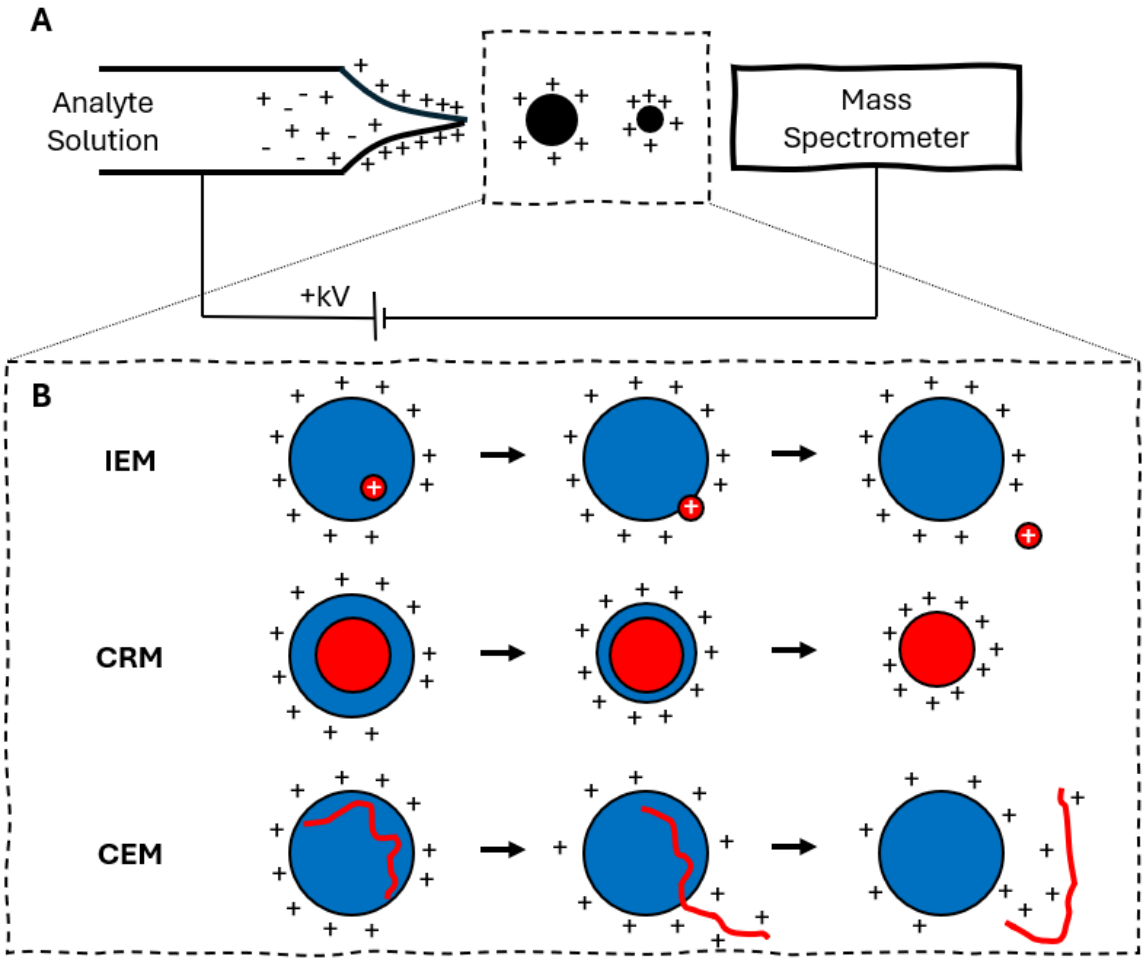


Figure 1.4. Schematic representation of the working principles of electrospray ionisation (ESI). A) Schematic representation of ESI source in positive mode. B) Alternative ESI mechanisms: ion evaporation model (IEM) for small analytes, charged residue model (CRM) for globular proteins, chain ejection model (CEM) for unfolded proteins. Adapted with permission from Konermann et al., "Unraveling the Mechanism of Electrospray Ionization", Analytical Chemistry, Copyright © 2013 American Chemical Society.

These highly charged ions (with final radii in the nanometer range) enter the mass spectrometer, which functions as a molecular scale by measuring their mass-to-charge ratio (m/z) (Serdyuk et al. 2007). The ion, with mass m and charge state z, is accelerated by an applied voltage V, acquiring kinetic energy E_k according to the following equation:

Eq. 1.12 $$E_k = \frac{1}{2}mv^2 = zeV$$

where v represents the velocity of the ion and e is the elementary charge ($e = 1.6 \times 10^{-19}$ C). In a "time of flight" (ToF) mass spectrometer, the principle of operation is relatively simple (**Figure 1.5**) (Boesl 2017). If a detector is placed at a distance d from the exit point of the accelerating source, the time t it takes for the ion to reach the detector depends on its m/z ratio:

Eq. 1.13 $$t = \sqrt{\frac{d^2}{2Ve}}\sqrt{\frac{m}{z}}$$

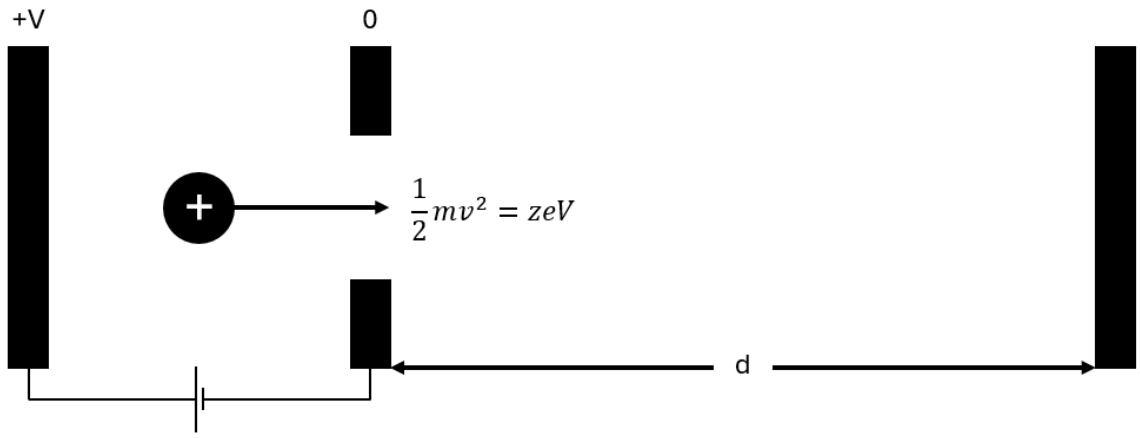


Figure 1.5. Schematic working principle of a "time of flight" (ToF) mass spectrometer.

The mass resolution of a basic ToF instrument is limited by the fact that ions are not at rest prior to acceleration and they are also spatially spread out and might not all start at the same time. Instead, they possess an initial velocity which follows a Boltzmann distribution. The horizontal component of this initial velocity adds to the velocity imparted by the accelerating voltage, meaning that ions of the same mass, but with different initial velocities, will arrive at the detector at slightly different times, reducing resolution. To address this limitation, ToF instruments are commonly equipped with a "reflectron", a series of metal plates with increasing voltage potential designed to reflect the ion trajectories. This correction compensates for variation in kinetic energy: ions

with higher kinetic energy travel longer paths, while those with lower energy follow shorter trajectories. As a result, ions of the same mass, regardless of their initial energy, are timed to arrive at the detector simultaneously, thereby improving the m/z resolution.

Since the invention of the first mass spectrometer by Thompson in 1912, technological advancements have led mass spectrometry to become one of the most used technique in molecular biology (Griffiths 2008). Over time, a variety of mass analyzers have been developed, including sector mass spectrometer, quadrupole mass spectrometer, quadrupole ion trap mass spectrometer, Fourier transform ion cyclotron resonance (FTICR) mass spectrometer, and Orbitrap mass spectrometer (Hoffmann and Stroobant 2007). The versatility of mass spectrometry in analyzing proteins of various sizes, combined with its lower associated costs compared to NMR spectroscopy, has made it an attractive method for investigating protein HDX (Verma et al. 1986). HDX-MS experiments leverage the mass difference between hydrogen and deuterium atoms. When a polypeptide chain is introduced into D_2O, its overall mass increases due to the exchange of backbone hydrogen atoms and solvent deuterium atoms. Although HDX-MS provides valuable insights into protein dynamics, its resolution is generally at the peptide level (typically covering tens of amino acids), rather than at the individual residue level as in NMR. A detailed explanation of the experimental workflow of HDX-MS experiments is provided in **section 1.3.1**.

1.3 Hydrogen/deuterium exchange mass spectrometry

Today, MS and NMR are the main alternative techniques to study the HDX of proteins or unfolded peptides. Of these, MS is generally preferred due to its broader applicability to proteins of varying sizes, simpler sample preparation (no need of ^{15}N-labelled protein) and lower costs. While NMR provides site-specific data, allowing the direct study of individual amino acids, MS offers insights at the peptide level, typically for sequences composed of tens of residues. This section outlines the HDX-MS workflow, the advantages and drawbacks of the experiment with respect to the alternative

labelling techniques in the context of structural mass spectrometry, and introduces the central research question of this Thesis: can single-residue information (protection factors) be inferred from peptide-level HDX-MS data? A detailed discussion on data analysis and interpretation follows in **Chapter 2**. The section ends with a discussion about the versatility of HDX-MS experiments and listing two examples where obtaining single-residue information could be beneficial.

1.3.1 Labelling techniques in structural mass spectrometry

Structural biology has been dominated by X-ray crystallography, NMR spectroscopy and, more recently, cryo-EM. These techniques provide high-resolution protein structures, but they are limited in their use because of the relatively small range of molecular sizes they can analyse. NMR and X-ray crystallography are limited to small sizes (< 50 kDa), while cryo-EM to bigger systems (> 120-150 kDa) (**Figure 1.6**). Only in the last two decades, mass spectrometry (MS) has emerged as a powerful alternative (Kaur et al. 2019). Beyond the wider range of molecular sizes, MS has the great advantage of requiring small amounts of protein (µg). The advent of MS in the field of structural biology was favoured by the advent of electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI), which demonstrated that noncovalent interactions are preserved in the gas phase (Liko et al. 2016). Today, MS methods in structural biology can be grouped into two main families: "native" and "labelling" methods (Lermyte 2020).

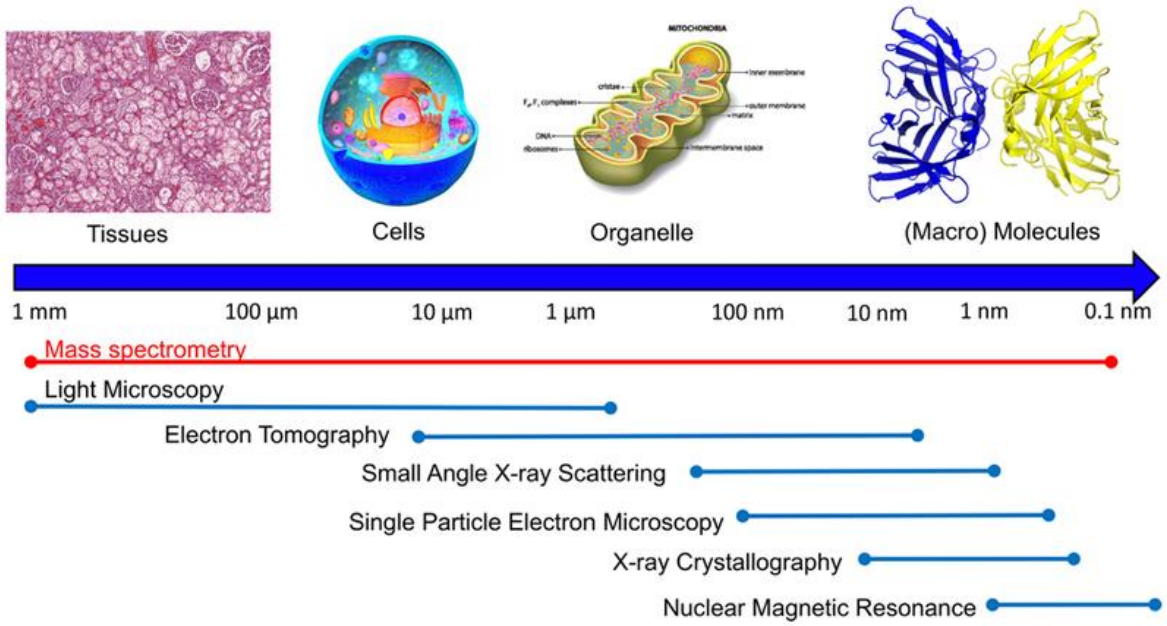


Figure 1.6. Resolution of different biophysical techniques.
NMR and cryo-EM are mainly applicable to macromolecules. MS has a flexible dynamic range providing structures informations from tissues to macromolecules. Reprinted with permission from U. Kaur, D. T. Johnson, E. E. Cheat et al. (2019), "Evolution of structural biology through the lens of mass spectrometry", Analytical Chemistry, © 2019 American Chemical Society.

In native MS methods, the sample is kept under near-native conditions, both during ESI and inside the mass spectrometer, and higher-order structural properties are preserved (Lermyte 2020). A mass spectrum shows the intensity of molecules present in solution with a specific mass to charge ratio m/z. The information contained in the raw data is convoluted: proteins and/or peptides are present with different charge states, with increasing charge values for unfolded proteins (Kafader et al. 2020); moreover, each charge state is represented by multiple peaks depending on the natural distribution of elements. The mass distribution of a peptide is called isotopic envelope.

The two dimensions of the mass spectrum (m/z and intensity) can be coupled to a third dimension, namely the drift time, thanks to ion mobility spectrometry (IMS) (Konijnenberg, Butterer, and Sobott 2013). Traditional drift time measures the time it takes for an ion to migrate through a buffer gas in presence of a low intensity electric

field. Under low-field conditions, the velocity of the ion is directly proportional to the electric field, and the proportionality constant can be written as a function of the collisional cross-section (CCS) of the ion (Kanu et al. 2008). Therefore, IMS provides additional information on the size of the analyte.

In labelling MS methods, mass labels are attached to the native structure in solution, which can be then denatured and/or digested to read out labelling sites and therefore obtain structural information (Lermyte 2020). We here discuss two labelling methods, namely limited proteolysis (LiP) and fast photo-oxidation of proteins (FPOP) before delving into hydrogen/deuterium exchange (HDX).

Limited proteolysis (LiP) has been recently coupled to mass spectrometry to measure protein conformational changes on a proteome-wide scale (Schopper et al. 2017). However, the idea of probing structural properties of proteins exploiting enzymatic digestion has been developed in the last two decades (Fontana et al. 2004). In a typical LiP-MS experiment, proteins are directly extracted from cells or tissues under native conditions, and each proteome extract is split into a control sample and a LiP sample. The LiP sample is first subjected to proteolysis under native conditions by the action of a broad-specificity enzyme like proteinase K; second, the sample is denatured and fully digested with a specific enzyme (such as LysC or trypsin). The control sample is only subjected to the second step of (specific) digestion. The peptide mixture, which contains fully tryptic peptides and half-tryptic peptides (i.e. deriving from the LiP step), is then studied with MS. While specific digestion is site-dependent, the usage of enzymes with broader specificity in the LiP step aims to identify structure-specific peptide fragments (Schopper et al. 2017). LiP-MS counts a vast number of applications, from the analysis of protein structural changes upon specific environmental perturbations to the identification of protein functional alterations from whole proteomes (Cappelletti et al. 2021).

Fast photochemical oxidation of proteins (FPOP) relies on the irreversible labelling of solvent-exposed amino acid side chains by hydroxyl radicals (Johnson, Di Stefano, and

Jones 2019), which can modify 19 of 20 amino acids with a dominant shift in mass of +16 Da (many other modifications can occur). The advantage of FPOP over HDX and LiP is that the data obtained provide residue-level resolution of protein structures and interactions on the microsecond scale. Moreover, labels are permanent, giving more scope for working with complex samples and in vivo. Applications range from epitope mapping to the identification of lipid-interacting regions in membrane proteins, and its potential for *de novo* modelling of protein structures has already been demonstrated (Aprahamian et al. 2018). In a typical FPOP experimental setup, a solution of protein and H_2O_2 is irradiated by an excimer laser at 248 nm to generate hydroxyl radicals. A system of lenses focuses the laser beam on the flow tubing, which is made by silica and has an inner diameter of 150-450 μm. The sample is irradiated through a transparent window exposed on the coated silica tubing. The flow rate and laser frequency are coordinated so that each protein molecule is irradiated only once. After irradiation, the sample is collected into a tube containing catalase and free methionine in buffer to quench H_2O_2 and OH radicals, respectively, thus preventing post-footprinting oxidation artifacts from any remaining reactive species. To correct for the background oxidation, protein control samples are introduced into the flow system without laser irradiation. Performing FPOP under constant flow limits over-oxidation, which could lead to protein unfolding. A radical scavenger, most commonly glutamine, is also added to the sample as another experimental control to prevent over-oxidation. Based on the reactivity of glutamine with OH, FPOP labels proteins on the microsecond timescale and ensures the labelling of the native state of proteins.

Another labelling technique that is worth mentioning is chemical cross-linking (Back et al. 2003; Leitner et al. 2010). The basic principle of a cross-linking MS experiment is that neighbouring residues can be cross-linked and after digestion those cross-linked protein residues can be identified, revealing proximity in the original sample (Lee and O'Reilly 2023).

1.3.2 HDX-MS experimental workflow

When a polypeptide chain is diluted in D_2O, the backbone amide hydrogens spontaneously exchange with deuterium in solution, and the consequent increase in mass can be detected with mass spectrometry. The stock protein solution is typically diluted in D_2O generating a final 80-95% deuterated buffer. The deuterium percentage depends mostly on the desired intensity of the analyte (in other words, on its concentration): if the signal of the analyte is too low in the mass spectrum, the final deuterium percentage should be decreased. Note that the Linderstrøm-Lang model (**Eq. 1.6**) is valid under the assumption of a 100% deuterated buffer, hence a lower deuterium concentration will affect the validity of the model, giving rise to a phenomenon that is referred to as back-exchange (i.e. the exchange of deuterium incorporated in the protein with residual hydrogens in solution).

The workflow of HDX-MS experiments can be highly automated thanks to the implementation of liquid handling robots. Commercial instruments allow to monitor the exchange of proteins for labelling times in a range that goes from 20-30 seconds to hours (provided that the protein is stable in the autosampler kept at 4°C) (Burkitt and O'Connor 2008). This time range is generally optimal to monitor the exchange of folded proteins under physiological conditions (Weis 2016). However, when the system of interest is an unstructured peptide or an unfolded or disordered protein, the exchange happens in the millisecond regime (Al-Naqshabandi and Weis 2017). To monitor the exchange of such molecules, the pH at which the labelling is performed can be decreased to slow down the exchange (Goswami et al. 2013); however, this might not be sufficient to rescale the exchange time to the time window detectable by a traditional instrument. Moreover, a protein is not guaranteed to assume the same conformation at lower pH values (Li et al. 2014). Using a microfluidic system that adjusts the length of capillary tubes and the flow rate of the protein solution, Phillips et al. were able to develop a prototype that enables the measurement of fast (millisecond) HDX (Kish et

al. 2023; Seetaloo and Phillips 2022). Alternative setups to acquire millisecond HDX data were developed by different groups (Lento and Wilson 2022).

If the solution is injected into the mass spectrometer right after deuterium labelling, i.e. when the protein (or the peptide) is still intact, we speak about 'global' analysis (Möller et al. 2020). Despite not being the traditional approach to an HDX-MS experiment, this type of analysis is useful to study the exchange of small systems, such as unfolded peptides (Al-Naqshabandi and Weis 2017) or oligonucleotides (Largy and Gabelica 2020), as it avoids several drawbacks that characterize the traditional workflow, such as quenching optimization and back-exchange (see below). However, this global analysis has a low spatial resolution as it detects information only on the exchange of the whole system.

To increase the resolution and obtain information about the exchange of more localized areas of a protein, the labelling reaction needs to be stopped (quenching). This can be done exploiting the dependence of the exchange on the pH and temperature of the solution (as described in **section 1.1.1**). Indeed, the exchange is generally quenched by lowering the pH near the minimum pH 2.5-3.0 and the temperature to approximately 0°C (**Figure 1.1**) (Masson et al. 2019). After quenching the exchange reaction, the protein is digested (either offline or with an online column) by an enzyme, generally pepsin because it is still active at such acidic conditions. Other enzymes have been proposed as alternatives to pepsin, such as fungal proteases XIII and XVIII (Cravello, Lascoux, and Forest 2003), nepenthesin (Kenji et al. 2005), *Aspergillus niger* prolyn endoprotease (Tsiatsiani et al. 2017), rice field eel pepsin and aspergillopepsin (Ahn et al. 2013). Note that the quenching step often contains denaturants besides being acid. This facilitates the unfolding of the protein, and therefore its digestion. These enzymes are non-specific; hence they cleave the protein in a non-predictable yet reproducible pattern of overlapping peptides.

The combination of quenching and digestion conditions generates a number of proteolytic peptides that are then separated by liquid chromatography and eluted into

the mass spectrum. These peptides can be visualized alongside the primary sequence of a protein in a coverage map (see for instance **Figure 2.2**, **Figure 3.1** or **Figure 5.3**) and the data quality achieved can be defined mainly by two parameters: the sequence coverage (i.e. the percentage of amino acids covered by at least one proteolytic peptide) and the redundancy (i.e. the average number of proteolytic peptides covering a residue, in other words the degree of overlap between peptides) (Masson et al. 2019). Optimizing the quenching and digestion conditions to maximize the sequence coverage and redundancy can be extremely challenging and highly protein dependent, especially for complex systems such as integral membrane proteins (Martens and Politis 2020). For example, the shift towards low pH is sometimes not sufficient to unfold a stable protein, and the addition of denaturant in the quench buffer (typically urea or GuHCl) can yield a significantly higher number of peptides (Hamuro and Coales 2018).

We have already mentioned the possibility for deuterium incorporated in the protein backbone to back-exchange into hydrogen. There are several steps in the experimental workflow where back-exchange can occur:

i) During labelling: we call this form of back-exchange "reverse-exchange" to differentiate it from back-exchange occurring for technical reasons. Reverse-exchange occurs at the level of the native protein.

ii) From quench onwards (Sheff, Rey, and Schriemer 2013): the quench buffer is water-based, and the labelling buffer is typically mixed with the quench buffer at a 1:1 ratio. The low pH and temperature in the quench can slow down, yet not completely stop, forward-exchange (H to D) and back-exchange (D to H), which are competing mechanisms occurring at the protein level from quench to digestion and at the peptide level afterwards (these steps take typically 10-20 minutes).

iii) During ionization (Guttman et al. 2016): additional back-exchange can occur at the peptide level during electrospray ionization.

Back-exchange can be minimized but cannot be completely removed from the experimental workflow. More details on how to treat back-exchange (experimentally or during data analysis) are provided in **section 2.4.2**.

HDX-MS experiments monitor the exchange of proteolytic peptides of the protein, which are generally 5-25 residue long. While this resolution can be sufficient to localize binding sites in a differential experiment (e.g. when the apo and holo conformers of the protein are compared), it does not provide direct information at the level of the single residue and is therefore challenging to connect with the Linderstrøm-Lang theory and, consequently, to dynamic properties of the protein and molecular modelling. The resolution of HDX-MS data can be further increased experimentally by coupling the workflow described so far with fragmentation techniques (Sobott 2020). The proteolytic peptides can be further fragmented using suitable gas-phase dissociation techniques, such as collision-induced dissociation (CID), electro-capture dissociation (ECD), electron-transfer dissociation (ETD) or ultraviolet photodissociation (UVPD) (Mistarz et al. 2018). An issue that may arise while using these fragmentation techniques is that of H/D scrambling (Boyd and Somogyi 2010), i.e. gas-phase randomization of the solution deuterium label. H/D scrambling can be measured for model polypeptides for which the deuterium incorporation pattern is known (Zehl et al. 2008). For example, in a HHHHHHIIKIIK sequence (**Figure 1.7**), the first (histidine-based) part of the sequence exchanges much faster than the second (formed by isoleucines and lysines). At a labelling time of around 1 minute, the first half of the model peptide is completely exchanged, while the second half has not exchanged yet. H/D scrambling can be measured by looking at the deuterium incorporation of the fragments of the peptide model: if there is no scrambling, only fragments coming from the first half of the peptide will have exchanged. Extensive research has shown that ECD, ETD and UVPD induce minimal H/D scrambling (if properly tuned), and can therefore be coupled to HDX-MS experiments to experimentally increase the spatial resolution (Rand, Zehl, and Jørgensen 2014). With this strategy, a 2-amino acid resolution has been achieved by Pan (Pan et al. 2009).

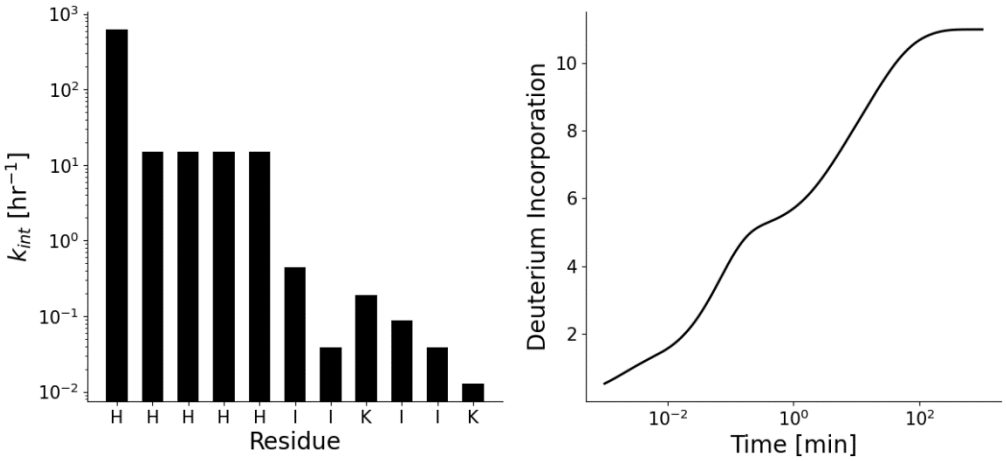


Figure 1.7. Model polypeptide with sequence HHHHHHIIKIIK used for measuring H/D scrambling. A) Intrinsic exchange rates for the model peptide, calculated at pH 4 and temperature 20°C. B) Deuterium uptake curve for the model polypeptide. Intrinsic exchange rate calculations were performed using a python script available at: https://github.com/pacilab/exPfact/blob/master/python/kint.py.

Alternative labelling techniques in structural mass spectrometry provide complementary information. Indeed, while HDX-MS informs on the solvent accessibility, dynamics and hydrogen bonding of backbone amides (Devaurs et al. 2022), FPOP-MS informs on solvent accessible amino acid side chains (Cornwell et al. 2018) and LiP-MS on the overall flexibility of the protein (Fontana et al. 2004). Compared to LiP and FPOP (described in **section 1.3.1**), HDX offers several advantages (Mitra 2021). While FPOP can achieve single-residue resolution, this is generally limited to a small number of residues, resulting in lower overall sequence coverage compared to HDX-MS. LiP, on the other hand, can provide broader sequence coverage and higher redundancy by employing a cocktail of unspecific proteases, although it remains limited to peptide-level resolution. However, LiP has the advantage of being applicable to complex systems *in vivo* (Cappelletti et al. 2021), whereas HDX is only beginning to extend to these environments (Lin et al. 2022). Additionally, state-of-the-art HDX-MS experiments have the unique feature to provide a continuous kinetic measurement, allowing deuteration to be monitored over time rather than observing a static before/after labelling snapshot (Hamuro 2021a). In principle, FPOP and LiP can

also be adapted for kinetic measurements by adjusting the laser intensity or exposure (in FPOP) or the digestion time (in LiP). However, a significant limitation of HDX relative to LiP and FPOP is the reversibility of hydrogen/deuterium exchange, which omplicates data analysis due to challenging back-exchange corrections (see **section 2.4.2**).

1.3.3 Single-residue resolution from peptide-level data: an underdetermined problem

HDX-MS experiments monitor the deuterium uptake of a protein via its proteolytic peptides (or their fragments if a fragmentation technique is used) (Masson et al. 2019). The Linderstrøm-Lang model (**Eq. 1.6**) can be used to write the deuterium uptake of a peptide D(t) as the sum of the contributions of its N exchangeable residues:

Eq. 1.14 $$D(t,\{P\}) = \frac{1}{N}\sum_{i=1}^{N}(1 - e^{-\frac{k_{int,i}}{P_i}t})$$

where $k_{int,i}$ and P_i represent the intrinsic exchange rate and the protection factor of residue i, respectively. The ultimate goal of any HDX experiment is to estimate the protection factors of the protein, which allow to quantitatively connect the experimental data with the Linderstrøm-Lang theory as well as to structural and dynamic properties of the protein (Englander et al. 2016). The search for single-residue resolution has first motivated the HDX field to move from density gradient ^{3}H-gel filtration methods to NMR (**section 1.2.1**). Then, the need to study a broader range of biological systems dictated the move from NMR to MS, at the cost of a lower (peptide level) resolution. Is it possible to use computational strategies and advance statistical methods to retrieve single-residue information from peptide-level (or fragment-level) data?

One of the problems of extracting protection factors from peptide-level (or fragment-level) HDX-MS data is 'underdetermination' (Skinner et al. 2019). A problem is underdetermined when the number of parameters that must be estimated is higher

than the number of experimental points. A trivial example is given by the following problem: suppose that in a closed basket there are N apples and M oranges, for a total of 5 fruits; how many apples and oranges are there? The problem is underdetermined as the number of the required parameters (two: N apples and M oranges) is higher than the number of experimental points (one: 5 fruits). The consequence of underdetermination is that there exist alternative solutions that are equally in agreement with experimental data: in the example, $(N = 1, M = 4)$, $(N = 2, M = 3)$, $(N = 3, M = 2)$ and $(N = 4, M = 1)$ are all possible solutions. Of course, when we open the basket, we realize that only one of these solutions is the true one.

Analogously, HDX-MS data are affected by underdetermination (Stofella et al. 2022). In the case of the isolated deuterium uptake curve of a peptide formed by N residues, the number of parameters to be extracted (N) can be higher or lower than the number of experimental time points available (M). If $N < M$, the problem is underdetermined by definition, hence there exist multiple patterns of protection factors that are equally in agreement with the experimental data. Even when $N > M$ (when the peptide is small or when the number of labelling time points is high), multiple solutions are equally consistent with experimental data because **Eq. 1.14** does not contain information on the relative contribution of the exchange rates. For example, a peptide formed by three amino acids with three different exchange rates yields the same deuterium uptake curve independently of the assignment of a specific rate to a specific residue.

The degree of underdetermination can be decreased by looking simultaneously at the information contained in the deuterium uptake curve of overlapping peptides and is solved in the case of an ideal dataset where all peptides differ by exactly one amino acid (Kan et al. 2013).

Overcoming underdetermination can be achieved by splitting the question into two separate problems: i) calculating all solutions in agreement with experimental data and ii) finding a way to select the *true* one. Throughout this work, we assume that the *true*

value of a protection factor (or the closer we can get to it) is given by the experimental value extracted by NMR.

1.3.4 Applications of HDX-MS: two examples

The versatility of HDX-MS, with its ability to study a wide range of protein systems and provide structural and dynamic information complementary to other techniques like X-ray crystallography, NMR spectroscopy and cryo-EM, has driven its rapid growth in both industry and academia. In its most common applications, HDX-MS has been used to (Narang, Lento, and J. Wilson 2020): i) compare different protein conformers; ii) localize protein-small molecule binding sites while monitoring potential orthosteric and allosteric changes in the protein (Masson, Jenkins, and Burke 2017); iii) map antibody epitopes (Deng, Lento, and Wilson 2016; Wei et al. 2014); iv) characterize large or multi-protein complexes, which are a challenging target for X-ray crystallography and NMR spectroscopy; v) study intrinsically disordered proteins (IDPs) that adopt multiple conformations in their native state and are difficult to crystallize; vi) analyzing the conformational dynamics of membrane proteins, a key focus in drug discovery (Martens and Politis 2020).

In most published studies, HDX-MS experiments are performed at the peptide resolution, revealing differences in deuterium uptake between two conditions (e.g. different conformers or ligand-bound/unbound states). In this Thesis, we propose methods that allow to infer quantitative single-residue information from peptide-level data. This increase in resolution can be powerful to draw more precise structural insights that can drive deeper conclusions about protein systems than peptide-level data alone. We provide two examples.

HDX-MS has been applied to understand prion diseases, where single-residue changes can have a profound impact on protein behavior. A known mutation, A116V, in the prion protein (PrP) enhances the formation of a neurotoxic transmembrane form, which induces neural death (Hegde et al. 1998). By comparing deuterium uptake curves

between wild-type and mutant, Sabareesan and Udgaonkar identified that the peptide 109-132, which is responsible for membrane insertion, becomes more solvent-protected in the mutant (Narang et al. 2020; Sabareesan et al. 2016). Single-residue resolution in this context provides a clear structural advantage by pinpointing precise mutation-induced changes.

HDX-MS is also valuable for studying proteins in different buffer conditions, such as assessing temperature- or pH-dependent conformational changes (Tajoddin and Konermann 2020, 2022). In such cases, differentiating structural changes from buffer-driven effects is essential for accurate conclusions. A traditional differential analysis might confound these factors, but retrieving single-residue information can help discriminating the role played by the buffer (**section 1.1.1**) from that of structure (**section 1.1.2**). For example, α-synuclein, an IDP, shows conformation changes in different *in vivo* environments; IM-MS studies have shown it adopts a more compact form in the presence of metal ions (Byrd et al. 2023). HDX-MS could ideally localize such changes, although the risk of a traditional peptide-level analysis is the misclassification of buffer effects for structural shifts.

1.4 Aims and structure of the thesis

The research question leading to the development of this Thesis is whether it is possible to infer single-residue resolution (protection factors) from peptide-level HDX-MS data. In a broader sense, this work aims to show that HDX-MS experiments have the potential to draw quantitative conclusions on the dynamics of a protein system, rather than only qualitatively localize difference in the uptake of different experimental conditions.

The importance of the Linderstrøm-Lang model in describing the HDX of proteins, and the key role played by the protection factors in connecting experimental data with structural modelling is the focus of **Chapter 1**. We also introduced the different techniques developed during the years to study HDX, with a major focus on the working principles of mass spectrometry. We have described in detail the experimental

workflow of traditional and alternative HDX-MS experiments, introducing underdetermination, i.e. the statistical problem that characterizes HDX-MS data and does not allow a straightforward extraction of single-residue information.

In **Chapter 2**, "Computational tools for hydrogen/deuterium exchange mass spectrometry data analysis", we present a software review that describes and critically evaluates the state-of-the-art tools to analyse HDX-MS data. The section describes the importance of choosing an appropriate statistical test when performing a differential experiment (i.e. when comparing two or more protein states), as well as alternative methods to deal with multimodal analysis of protein spectra and to computationally increase the spatial resolution of peptide-level data.

In **Chapter 3**, "High resolution hydrogen-deuterium protection factors from sparse mass spectrometry data validated by nuclear magnetic resonance", we propose a computational method to estimate the alternative patterns of protection factors in agreement with experimental HDX-MS data. We show how the additional information encoded in the isotopic envelope of the peptides can be exploited to rank the alternative solutions from best to worst. The outcomes are compared with HDX-NMR data probing the same protein under the same experimental conditions, showing a satisfying correlation. The promising results reported by this method, ExPfact, represented the foundation for further studies.

In **Chapter 4**, "Recalibrating protection factors using millisecond hydrogen/deuterium exchange mass spectrometry", we studied the fast HDX of a mixture of peptides to test the validity of the predictions provided by the Englander group (described in section 1.1.1), on which ExPfact relies. Our results show that these predictions are more accurate when a three-alanine peptide is used as unstructured reference instead of poly-DL-alanine (PDLA), and that this is due to the fact that PDLA has high structural propensity as the chain length increases.

In **Chapter 5**, "Inferring residue-level resolution from peptide-level hydrogen/deuterium exchange mass spectrometry data", we expand the method presented in **Chapter 3** to exploit structural information of the protein (when available). We acquired both HDX-MS and HDX-NMR data for two test proteins to show that the protection factors extracted by HDX-MS data only provide reliable estimates. The quality of the predictions depends on the quality of the dataset, namely on the redundancy provided by the peptides and by the number, type and spread of time points available. The strengths and limitations of our approach are discussed.

In **Chapter 6**, we present a general discussion linking the different chapters in the light of the main research question: can single-residue resolution be inferred from peptide-level HDX-MS data? We provide a chronological perspective on how we tried to overcome the problem of underdetermination and discuss future directions of the field.

Chapter 2. Computational Tools For Hydrogen/Deuterium Exchange Mass Spectrometry Data Analysis

Submitted and accepted with minor corrections as: M Stofella, A Grimaldi, JH Smit, J Claesen, E Paci and F Sobott, “Computational tools for hydrogen/deuterium exchange mass spectrometry data analysis”, Chem Rev, 2024

2.1 Abstract

Hydrogen/deuterium exchange (HDX) has become a pivotal method for investigating the structural and dynamic properties of proteins. The versatility and sensitivity of mass spectrometry (MS) made the technique the ideal companion for HDX, and today HDX-MS is addressing a growing number of applications in both academic research and industrial settings. The prolific generation of experimental data has spurred the concurrent development of numerous computational tools, designed to automate parts of the workflow while employing different strategies to achieve common objectives. Various computational methods are available to perform automated peptide searches and identification; different statistical tests have been implemented to quantify differences in the exchange pattern between two or more experimental conditions; alternative strategies have been developed to deconvolve and analyse peptides showing multimodal behaviour; and different algorithms have been proposed to computationally increase the resolution of HDX-MS data, with the ultimate aim to provide information at the level of the single residue. This review delves into a comprehensive examination of the merits and drawbacks associated with the diverse strategies implemented by software tools for the analysis of HDX-MS data.

2.2 Introduction

Proteins are the most important gears in the engine of life. Since the seminal work by Anfinsen in 1960, scientists have wondered how their linear sequence of amino acids folds into a defined three-dimensional structure, how these structures change upon binding, and how they maintain health and cause disease. High-resolution snapshots of protein structures can be captured by X-ray crystallography, NMR spectroscopy or electron microscopy (EM), while their dynamic behaviour in solution is harder to probe. Hydrogen bonding is one of the defining aspects of a protein's structure (or lack thereof), but equally important for how it interacts with the surrounding solvent. One unique feature of proteins is the exchange of their amide hydrogens with hydrogens in solution (Linderstrøm-Lang 1955). "Proteins continuously emit signals in the language of hydrogen exchange" (Englander et al. 2016), and understanding how to detect and interpret these signals is a unique opportunity to harness protein design.

When diluted into a deuterated buffer, the amide hydrogens of the protein spontaneously exchange with deuterium in solution (Linderstrøm-Lang 1955). The phenomenon is referred to as hydrogen-deuterium exchange (HDX). In the case of fully unstructured proteins, the rate of exchange depends on chemical properties of the buffer (pH, temperature, ionic strength) on one side, and on the amino acid's effective pKa (determined by its side chain and its direct neighbours). When a protein acquires its native structure, hydrogen bonding and solvent accessibility lower the rate of exchange by means of 'protection', and HDX measures this perturbed rate of exchange, thereby informing on the protein's structural and dynamic properties (Englander et al. 2016). Measuring the isotopic exchange in proteins posed a technical challenge. In its early years, HDX was measured using an ultracentrifugation procedure (Hvidt and Linderstrøm-Lang 1954); later, by infrared (Osborne and Nabedryk-Viala 1978) or UV (Englander, Calhoun, and Englander 1979) spectroscopy. These techniques have low "spatial resolution": they cannot monitor the exchange at a residue-level, but only the global exchange of the protein (i.e. the summed exchange of labile sites); yet they cannot

determine the overall extent of deuterium incorporation very accurately either. The popularity of HDX increased with the advent of two-dimensional NMR (Englander and Mayne 1992). Hydrogen and deuterium have different spins (hydrogen has spin ½, while deuterium has spin 1); leveraging the decrease of ^{1}H-NMR signal upon deuteration in an HSQC spectrum (Kleckner and Foster 2011), HDX-NMR can monitor the exchange of individual labelled residues (high spatial resolution) but is limited to the study of small proteins (< 50 kDa) (Dempsey 2001) and requires larger amounts of sample as well as ^{15}N labelling. In the last 30 years, HDX coupled with mass spectrometry (MS) has been established as a viable alternative (Masson et al. 2019). The versatility of the technique (Trabjerg, Nazari, and Rand 2018) and recent technological advancements (Engen et al. 2021) led to the generation of large amounts of data, and today the technique needs computational tools for an automated analysis and for retrieving more detailed and statistically accurate information from the raw data (Claesen and Burzykowski 2017).

HDX-MS measures the mass increase of a protein caused by deuteration (**Figure 2.1**) (Masson et al. 2019; Vinciauskaite and Masson 2023). The protein (or complex) is first equilibrated in a suitable biochemical buffer at desired pH, ionic strength and temperature. Continuous H/D exchange starts with dilution into deuterated buffer, with a final deuteration percentage that typically ranges from 80 to 95%, and labelling occurs for a variable amount of time. Labelling times generally range from 10s of seconds to hours, but recent technological developments gave access to the millisecond scale (Kish et al. 2023; Seetaloo and Phillips 2022; Svejdal et al. 2019), which is crucial to probe the fast exchange of highly dynamic regions and intrinsically disordered proteins (Seetaloo and Phillips 2022), as well as unstructured peptides (Al-Naqshabandi and Weis 2017) (these are highly valuable for fundamental studies, e.g. to study how H/D exchange is dependent on the buffer conditions). HDX can be monitored at the level of the intact protein (*global* HDX); it is worth noting here that global HDX-MS has been applied to study structured oligonucleotides (Largy and Gabelica 2020) and a software, OligoR (not reviewed here), has been developed to analyse these data

(Largy and Ranz 2023). In order to obtain higher spatial resolution (*local* HDX), a "bottom-up" approach is generally implemented: the protein is digested, and the mass spectra of the proteolytic peptides are acquired. While measuring the mass shifts of the intact peptides yields data on the incorporation of deuterium per peptide, MS/MS fragmentation using collision-induced dissociation (CID) is used to confirm the sequence of the peptide without deuteration, but it scrambles hydrogen and deuterium within peptides and is therefore not useful for determining of exchange sites at the single amino acid (residue) level. Before digestion, the exchange must be quenched at low pH (~ 2.5) and temperature (~ 0°C) to minimize *back-exchange*, which corresponds to a partial loss of deuterium label. Pepsin is the most used enzyme for protein digestion in HDX-MS experiments because it is active at acidic pH, although other enzymes have been used (such as the fungal proteases XIII and XVIII (Cravello et al. 2003), nepenthesin (Kenji et al. 2005), *Aspergillus niger* prolyn endoprotease (Tsiatsiani et al. 2017), rice field eel pepsin and aspergillopepsin (Ahn et al. 2013)). These enzymes cleave the protein into peptides, producing non-predictable yet reproducible patterns of overlapping peptides; the use of multiple enzymes can increase the spatial resolution, which is determined by the digestion pattern. The proteolytic peptides are separated by rapid reversed-phase liquid chromatography (LC) with a gradient time of ~10 minutes (possibly holding the column close to 0°C), ionized with electrospray ionization (ESI), and eluted into the mass spectrometer. Alternative experimental setups for local HDX-MS, not covered in this review, include the fragmentation of the intact labelled protein ("top-down") or of the proteolytic peptides ("middle-down") using electron capture dissociation (ECD), electron transfer dissociation (ETD) (Wollenberg et al. 2020) or ultraviolet photo-dissociation (UVPD) (Mistarz et al. 2018).

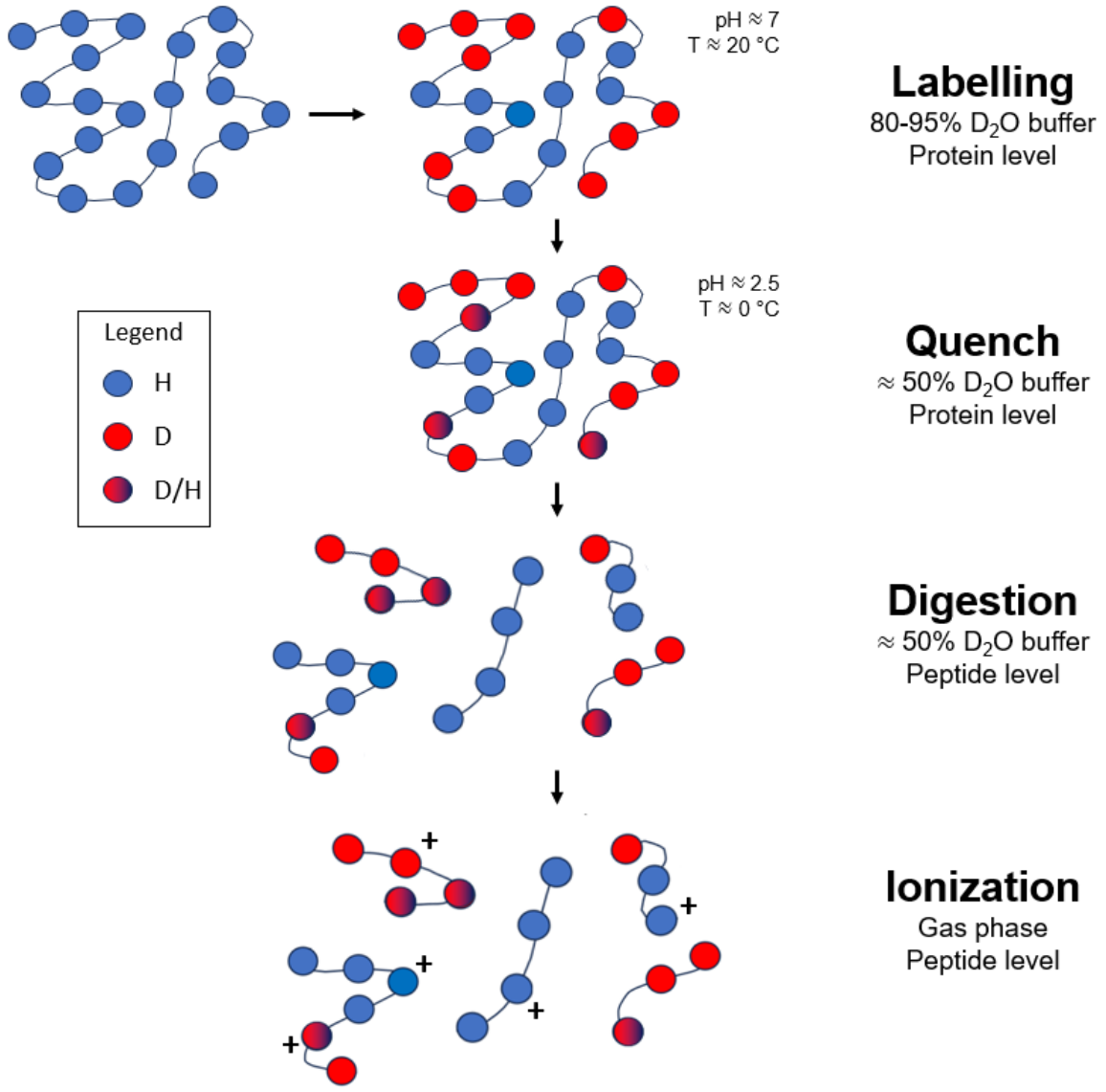


Figure 2.1. Typical experimental workflow of an HDX-MS experiment.

Typical experimental workflow of an HDX-MS experiment: 1) labelling: the undeuterated protein (blue) is diluted in a 80-95% deuterated buffer where the HDX occurs at the protein level for times ranging from milliseconds to hours; deuterated residues are shown in red; 2) quench: the exchange, still occurring at the protein level, is minimized by lowering the temperature (to ~0°C) and the pH (to ~2.5), back-exchange can occur at the protein level (blue/red beads); 3) digestion: the protein is digested, from this point forward-exchange and back-exchange (red-blue circles) occur at the peptide level; 4) ionization: the proteolytic peptides are ionized and eluted in the mass spectrometer.

The raw data of the peptide-level experiment comprises the time evolution of the mass spectra of proteolytic peptides of the protein, i.e. their mass shifts. A comprehensive tool for HDX-MS data analysis would 1) identify a list of proteolytic peptides assigned to mass spectra in the raw control (undeuterated) data, 2) assign peaks in the labelled (deuterated) raw spectra of the identified peptides, 3) identify peptides showing a bimodal spectrum (see **section 2.5.3**), 4) calculate the mass increase of each peptide, 5) correct for back-exchange, 6) increase the spatial resolution, ideally to residue level (protection factor analysis), and 7) localize and quantify statistically significant difference in the uptake pattern of two (or more) experimental conditions (differential analysis). Steps 1, 2 and 4 are generally conducted using vendor-specific software tools (namely PLGS and DynamX for Waters instruments, BioPharma Finder and HDExaminer for Thermo Fisher Scientific instruments), and the results are then exported to perform further analysis. Back in 2006, HX-Express (Weis, Engen, and Kass 2006) was one of the first software tools for HDX-MS data analysis. Since then, several platforms have been developed, such as HDX workbench (Pascal et al. 2012), Hydra (Slysz et al. 2009), Hexicon (Lindner et al. 2014), and ExMS (Kan et al. 2011) that have been previously reviewed (Claesen and Burzykowski 2017). In response to the recommendations for performing, interpreting and reporting HDX-MS experiments published by the international community in 2019 (Masson et al. 2019), several methods have been implemented with the goal of providing a standard and comprehensive framework for data visualization and differential analysis. Moreover, stand-alone computational methods have been developed to tackle the most common challenges provided by HDX-MS data, such as corrections for back-exchange, deconvolution of EX1/EX2 kinetics and protection factor analysis.

The purpose of this paper is to review the recent tools (both commercial and open-source) available for the analysis of continuous labelling, *local* HDX-MS data. First, we evaluate the capability of *comprehensive* software (by comprehensive, we mean a tool ideally able to cover all 7 points mentioned above) of providing a standardized framework for qualitative data visualization and quantitative data analysis for

differential experiments (when two or more experimental conditions are compared). Most biochemical experiments have this differential nature, as they compare two or more states of a protein (e.g. mutation, ligand binding or free against complex). In this common scenario, the data analysis workflow is divided into two parts: a commercial instrument-dependent software is first used to pre-process the experimental data, then a third-party open-source software is used for statistical data analysis. In most scenarios, this analysis is sufficient to answer the research question. Here, we particularly focus on more advanced tools where much more information contained in the data can be extracted. We review and discuss stand-alone programs implementing unique features for 'non-standard' analysis, such as automated peptide search (**section 2.5.1**), multimodal analysis (**section 2.5.3**), and protection factor analysis (**section 2.5.4**). The software and methods reviewed in this paper are listed in **Table 2.1**. Note that the Figures in this paper have not been created by one of the reviewed methods but by our own Python scripts.

Table 2.1. List of software tools and methods reviewed in this paper.

An updated list of software, publications and other resources for HDX-MS data analysis is available at the following link: https://github.com/hadexversum/HDX-MS-resources.

Software	Access to raw data			
	Automated Peptide Search (section 2.5.1)	Multimodal Analysis (section 2.5.2)	Differential Analysis (section 2.5.3)	High Resolution HDX-MS (section 2.5.4)
Claesen et al. (Claesen et al. 2021)			✓	
DECA (Lumpkin and Komives 2019)			✓	✓
deMix (Na et al. 2019)	✓	✓		
Deuteros 2.0 **(Lau et al. 2020a)**			✓	
ExMS2 **(Kan et al. 2013, 2019)**	✓	✓		✓
ExPfact **(Skinner et al. 2019; Stofella et al. 2022)**				✓
HaDeX (Puchała et al. 2020)			✓	
HD-e Xplosion (Zhang et al. 2021)			✓	

HX-Express (Guttman et al. 2013)	✓	✓			
Hdflex (Seetaloo, Kish, and Phillips 2022)				✓	✓
HDXAnalyzer (Liu et al. 2011)				✓	
HDXModeller (Salmas and Borysik 2021)					✓
Hdxstats (Crook, Chung, and Deane 2022)				✓	✓
HR-HDXMS (Gessner et al. 2017)					✓
Mass Spec Studio (Filandr et al. 2024; Raval et al. 2021; Rey et al. 2014; Ziemianowicz et al. 2020)	✓	✓	✓		
MEMHDX (Hourdel et al. 2016)				✓	
PyHDX (Smit et al. 2020)					✓
ReX (Crook et al. 2024)				✓	✓
Saltzberg et al. (Saltzberg et al. 2017)					✓

2.3 Theoretical background

When a protein is diluted in a solution containing deuterium oxide (D_2O), its amide hydrogens spontaneously exchange with deuterium (D). It is fair to say that all the hydrogens (H) of the protein are exchanging. However, the labelling timescales that can be probed with a typical HDX-MS experiment range from milliseconds to hours. In the light of this, carbon-bound aliphatic and aromatic hydrogens exchange far too slowly to be detected, while side chain acidic and basic hydrogens and polar –OH, –SH and $–NH_2$ groups exchange too fast, and therefore they rapidly back-exchange into hydrogen during the LC-MS analysis and are lost before detection (Englander et al. 1996; Hamuro 2021b).

Amide hydrogens are fully "exchange competent" ("open" state NH_{op}) when they are surface exposed and not engaged in secondary structure (i.e., they do not form hydrogen bonds other than with water). Some residues are structurally protected against exchange ("closed" state NH_{cl}), but local fluctuations (defined by the opening

and closing rates k_{op} and k_{cl}) can expose them to solvent-enabled deuteration and subsequently undergo exchange to form the deuterated state (ND) (Englander et al. 2016). As a consequence, HDX of a single amide hydrogen can be modelled as a two-step process (Linderstrøm-Lang model) (Linderstrøm-Lang 1955):

Eq. 2.1
$$NH_{cl} \underset{k_{cl}}{\overset{k_{op}}{\rightleftharpoons}} NH_{op} \xrightarrow{k_{int}} ND_{op} \underset{k_{op}}{\overset{k_{cl}}{\rightleftharpoons}} ND_{cl}$$

The intrinsic exchange rate k_{int} corresponds to the exchange rate of the residue in a completely unfolded structure. It depends on chemical properties of the buffer (pH, temperature and ionic strength) as well as the amino acid itself and the neighbouring residues (Al-Naqshabandi and Weis 2017; Bai et al. 1993; Connelly et al. 1993; Molday et al. 1972). HDX-MS is a kinetic experiment, with the ultimate goal of determining the rates of exchange defined in **Eq. 2.1**.

The exact analytical solution for the model in **Eq. 2.1** is a double exponential (Hvidt and Nielsen 1966). Under the so-called native approximation for a mostly folded peptide backbone ($k_{op} \ll k_{cl}$, i.e. the amide residue is mostly in the closed, protected state) and the EX2 regime ($k_{int} \ll k_{cl}$, i.e. the exchange is slow compared to the local structural dynamics), the deuteration of a single residue (d) – considering the deuterated residue either in the ND_{op} or ND_{cl} state – can be approximated as a single exponential:

Eq. 2.2
$$d(t) = 1 - e^{-\frac{k_{int}}{P}t}$$

The pseudo (pre)equilibrium constant $P \equiv k_{cl}/k_{op}$ is known as protection factor and encodes dynamic properties of the protein (Hamuro 2024): several microscopic models have been developed aiming to connect the structure of a protein to its protection factors; the most known model, often addressed as 'phenomenological model', describes the protection factor of a residue as the linear combination of heavy contacts (i.e. the number of atoms in the proximity of the amide not belonging to neighbouring residues in the primary sequence) and hydrogen bonds (Best and Vendruscolo 2006;

Vendruscolo et al. 2003). These models have already been reviewed by Devaurs et al. (Devaurs et al. 2022) and will not be discussed here.

Under denaturing conditions and for intrinsically disordered proteins, the amide backbone is largely exposed and the exchange kinetics may follow the so-called EX1 regime (occurring when $k_{cl} \ll k_{int}$) (Weis, Wales, et al. 2006). The deuterium uptake of a single residue can be approximated to occur in a single step with a rate k_{op}:

Eq. 2.3 $$d(t) = 1 - e^{-k_{op}t}$$

The presence of EX1 or EX2 kinetics (or their coexistence, known as EXX kinetics) can be discriminated in the raw HDX-MS data by the emergence of a bimodal pattern of the isotopic distribution in the mass spectrum of the peptide (see **section 2.5.3**) (Ferraro, Lazo, and Robertson 2004). However, this bimodal pattern is not guaranteed to occur in EX1 conditions. Indeed, when EX1 conditions are met, the exact analytical solution of the Linderstrøm-Lang model (**Eq. 2.1**) provides a fast-exchange kinetics per residue but no explanation for the bimodal pattern for the peptide. The explanation of the bimodal pattern stands in the cooperativity between residues, which is exclusive to peptide-level HDX-MS data and cannot be monitored by NMR experiments: under EX1 conditions, the probabilities of closing (k_{cl}) and exchanging (k_{int}) are such that, if subsets of residues open cooperatively, it is likely that most of (or all) the residues exchange, forming the second, fully exchanged population of the distribution (Zhang 2020). Other factors, discussed in **section 2.5.3**, may also lead to a bimodal pattern.

2.4 Connecting theory and HDX-MS data

HDX-MS experiments usually detect the deuterium uptake of a protein through its proteolytic peptides. Before performing any kind of analysis, pre-processing of the raw mass spectra is required to identify these peptides from the LC-MS/MS runs. This peptide search is performed on a digested control sample (without deuterium labelling). Identification of proteolytic peptides follows similar procedures as bottom-

up proteomics, albeit for non-tryptic peptides in the case of HDX and is generally performed using commercial software included with the instrument: PLGS and DynamX for Waters instruments, BioPharma Finder and HDExaminer for Thermo Fisher Scientific/Trajan (**section 2.4.7**). Next, the quality of the isotopic envelopes of each peptide (and charge state) is checked manually to verify assignments and eliminate false identifications. The major drawbacks of the semi-automated peptide search provided by these software and alternative strategies are discussed in **section 2.5.1**.

The peptide list is generally reported in a coverage map (**Figure 2.2**), where peptides are depicted as horizontal bars and visualized across the sequence of the protein. The quality of the dataset can be quantified mainly with 3 parameters: the number of peptides, the sequence coverage (the percentage of residues of the protein covered by the proteolytic peptides) and the redundancy (the "overlap", defined as average number of proteolytic peptides available per covered residue).

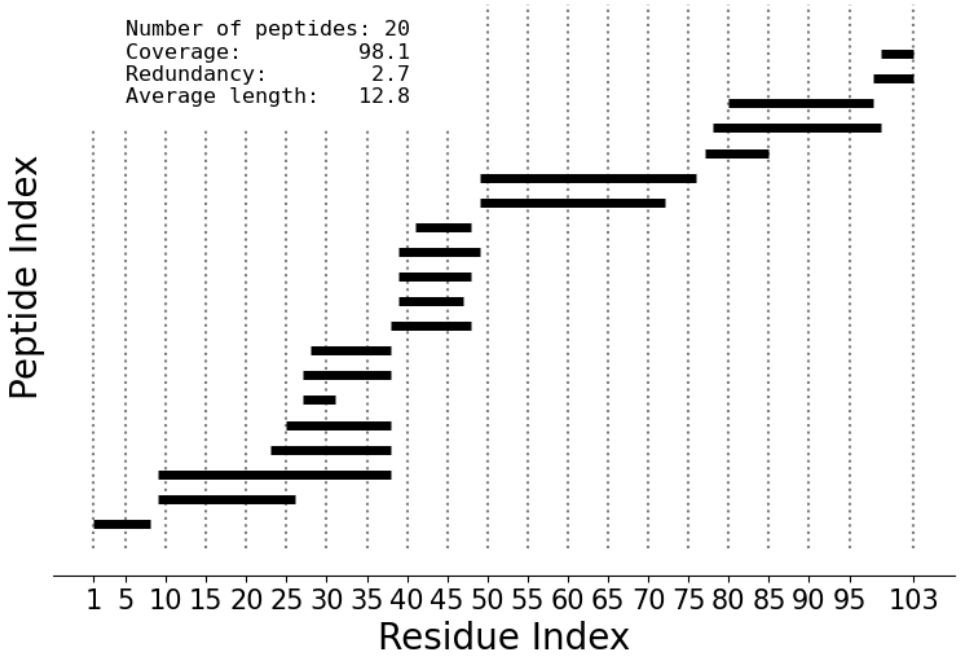


Figure 2.2. Example of a typical coverage map.

Horizontal bars represent proteolytic peptides localized along the sequence of the protein. Number of peptides, sequence coverage, redundancy, and average length (number of amino acids) are reported. Data showing the proteolytic pattern of cholera toxing B acquired by the author.

2.4.1 Absolute and fractional uptake

After the generation of a peptide list and the manual or automated assignment of isotopic envelopes at different labelling times, the intensity-weighted average m/z of the isotopic envelope of the peptide is recorded as a function of time (**Figure 2.3**).

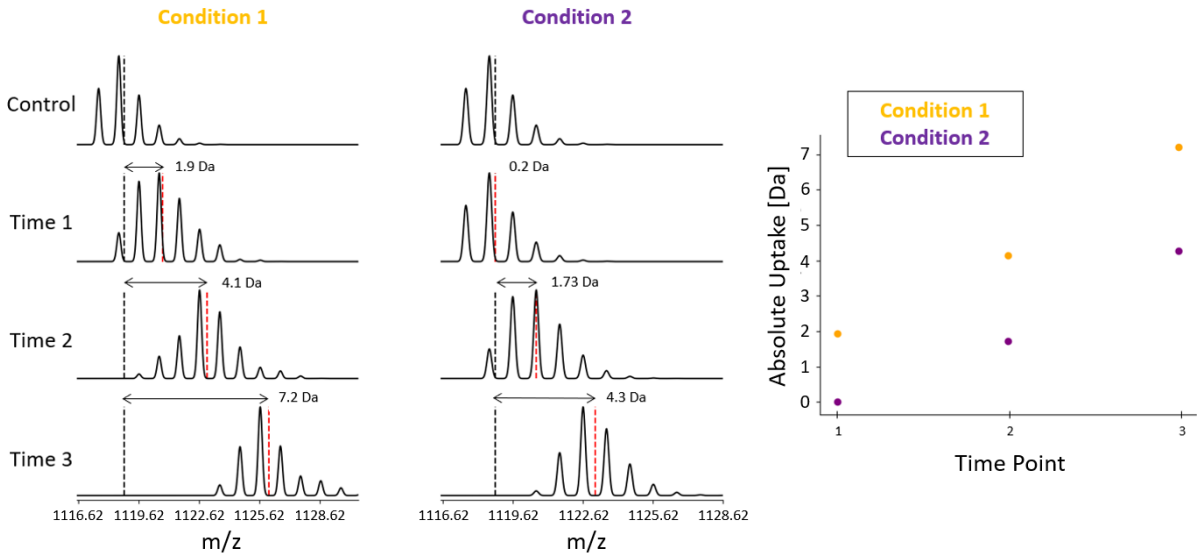


Figure 2.3. Isotopic envelopes of proteolytic peptides for different experimental conditions under analysis (simulated data).

Data shown for visualization purposes only for one peptide under two arbitrary experimental conditions: condition 1 (orange) and condition 2 (purple). At increasing labelling times, the isotopic envelope shifts towards higher values of m/z. The centroid of the isotopic envelope, i.e. the intensity-weighted average, is monitored. The absolute uptake (right) is defined as the difference between the centroid of the envelope at a specific time and the centroid of the control (fully protonated) envelope. The absolute uptake of different conditions is compared.

The measured m/z value (m_z) at a specific charge state z is converted into a mass scale (m) using the following formula:

Eq. 2.4 $$m = m_z \times z - z$$

The mass increase (or absolute uptake) is defined as the difference of the mass of the peptide at labelling time t (m(t)) and the mass of the peptide in the control sample (m_0):

Eq. 2.5 $$\text{Absolute Uptake (Da)} = m(t) - m_0$$

If the mass of the fully deuterated peptide (m_{FD}) is measured (Peterle, Wales, and Engen 2022), the absolute uptake is commonly converted to the fractional deuterium uptake (D_{Frac}):

Eq. 2.6 $$D_{Frac} = \frac{m(t) - m_0}{m_{FD} - m_0}$$

The measured fully deuterated sample often does not match the theoretical fully deuterated mass, which corresponds to the number of exchangeable amides (i.e. excluding prolines and the first/second residues (Hamuro 2021b)). This discrepancy arises because back-exchange (i.e. deuterium loss) can happen at different stages along the experimental workflow (**section 2.4.2**). When a fully deuterated sample is available, the fractional uptake in **Eq. 2.6** represents the conventional back-exchange correction.

2.4.2 Back-exchange

The Linderstrøm-Lang model (**Eq. 2.1**) considers "forward" HDX (i.e. H to D) to be an irreversible process, which is true only during the labelling phase (before quenching), when the protonated protein is exchanging within a 100% deuterated buffer, and when further processing steps from the quench onwards are neglected. This is not the case in typical HDX-MS experiments, where the protein is diluted resulting in an 80-95% deuterated buffer. While higher dilution factors could reduce reverse exchange and more closely align with the theoretical model, they are often impractical because the resulting protein concentrations might fall below the detection limit of the mass spectrometer. During the HDX-MS experimental workflow, there are several steps at which *back-exchange*, i.e. partial loss of deuterium label, can occur (**Figure 2.1**). The deuterium labelling is performed in a highly (yet not purely) deuterated buffer (80-

95%), and therefore *reverse exchange* (D to H, deuterium/hydrogen exchange) is occurring at the native protein level (e.g. with a 5% probability in a 95% deuterated buffer). From the quench onwards, the deuterated solution is mixed with a water-based quench buffer (generally at a 1:1 ratio): forward exchange and back-exchange are competing mechanisms occurring at the protein level from quench to digestion and at the peptide level afterwards. Additional back-exchange can occur during ionization and in the gas phase before detection in the mass spectrometer. To minimize back-exchange after the labelling phase, the temperature of the solution should be decreased (even below 0 °C) by placing the reversed-phase column for peptide separation in a refrigerated unit; but the digestion unit is usually kept at higher temperature to ensure efficient digestion (Wales et al. 2017). However minimized, it is not possible to completely remove back-exchange from the HDX-MS workflow, and therefore a proper quantification of back-exchange levels and consequent data normalization are highly desirable, but currently still lacking.

Most differential studies (i.e. where two or more experimental conditions of the protein are compared) do not perform any back-exchange correction and instead compare the absolute uptake (**Eq. 2.5**) of the same proteolytic peptides derived from different biological states of the protein, under the same experimental (technical) conditions. This procedure is correct only under the assumption that the extent of back-exchange is the same in the two experimental conditions, such that the denatured intact protein (from quench to digestion) and the peptides (from digestion onwards) are fully unstructured or retain similar residual structure. The validity of this assumption is not straightforward: for example, Sheff et al (Sheff et al. 2013) have shown that proteolytic peptides can retain residual structure in the LC column, hence different protein conformations may induce different back-exchange levels. Beyond differential studies, a proper back-exchange correction is essential if absolute and quantitative biophysical properties are required (such as exchange rates or protection factors, see **section 2.5.4**).

When implemented, the *standard* approach to correct for back-exchange is the acquisition of a fully deuterated sample and the normalization of absolute uptake values into fractional uptake (**Eq. 2.6**). There exist different protocols to acquire fully deuterated samples. For example, a fully deuterated control can be acquired by leaving the protein to deuterate for a time that is *long enough* to see the plateau in the kinetic uptake curves (e.g. for 12 hours). In many settings, researchers decide to avoid this strategy because many proteins are unstable for such long times (their partial denaturation resulted in lower intensities in the chromatogram, or their aggregation causes false protection), and in rare cases membrane proteins retain regions so protected that their exchange after 12 hours is negligible. A second strategy to acquire a fully deuterated control consists of diluting the protein in a deuterated buffer containing high levels of denaturant (e.g. 4 M urea) and leaving it overnight to exchange. A third strategy consists in performing offline digestion and deuteration of the proteolytic peptides. In the absence of a published study which systematically compares the results of the different strategies to acquire a fully deuterated sample, we recommend either of the latter two approaches.

As an alternative to the *standard* approach, the software DECA (Lumpkin and Komives 2019) was implemented around the need for developing a back-exchange correction. The authors identified two distinct forms of back-exchange that can influence deuteration (**Figure 2.4**): they called these "global back-exchange" that occurs at the level of the intact, but denatured protein (from quench to digestion on the pepsin column), and "local back-exchange" that acts at the level of the peptide (from digestion to the point of injection into the mass spectrometer) (Zhang, Zhang, and Xiao 2012). Back-exchange causes the deuterium uptake curve to plateau at a value lower than the theoretical fully deuterated mass.

The authors of DECA (Lumpkin and Komives 2019) also identify a "long exposure effect", which causes later time points to slowly deviate from the fully deuterated plateau. There are a number of possible reasons which can cause such effects, the most

obvious being related to the stability of the protein at longer time points where aggregation could lead to apparent protection against exchange. In addition, ambient moisture can also lead to deuterium loss in the sample causing a drop from the deuterated plateau. During a multi-day series of time points and replicates, protein samples may end up being kept at 0 °C for several days in the autosampler; but issues arising from this can be addressed by careful experimental design (e.g. mixing replicates of different time points randomly, or regularly replacing the protein sample with fresh aliquots). Protein stability tests done prior to HDX analysis are also helpful. In addition, the DECA paper describes an experimental artefact which can be misinterpreted as an additional form of back-exchange, caused by different liquid handling procedures at short and long time points. For example, when a LEAP robot is used for time points below 2 minutes, the mixing syringes skip a step, and this results in a slightly lower back-exchange. DECA allows to correct for global and local back-exchange by the application of a scaling factor, as well as accounting for this long exposure effect by the application of a universal, linear correction to all peptides. A recent paper by Wrigley et al. (Wrigley et al. 2024) expanded on the subject, confirming that automated liquid handling procedures can indeed introduce a large variability to the measured deuteration. While liquid handlers provide excellent efficiency with respect to manual pipetting, the number of steps involving syringe operations with small liquid volumes that occur during an HDX-MS experiment can be source for volumetric errors, which can cause minor differences in the final deuterium concentration or in pH, and in turn can be sufficient to cause significant differences in the uptake curves of peptides. These robot-related issues can be resolved by tracking the performances of the liquid handler over the different operations performed during the workflow and consequently optimizing the robot methods (e.g. changing the needle position or depth) to reduce the variability in the measured deuteration.

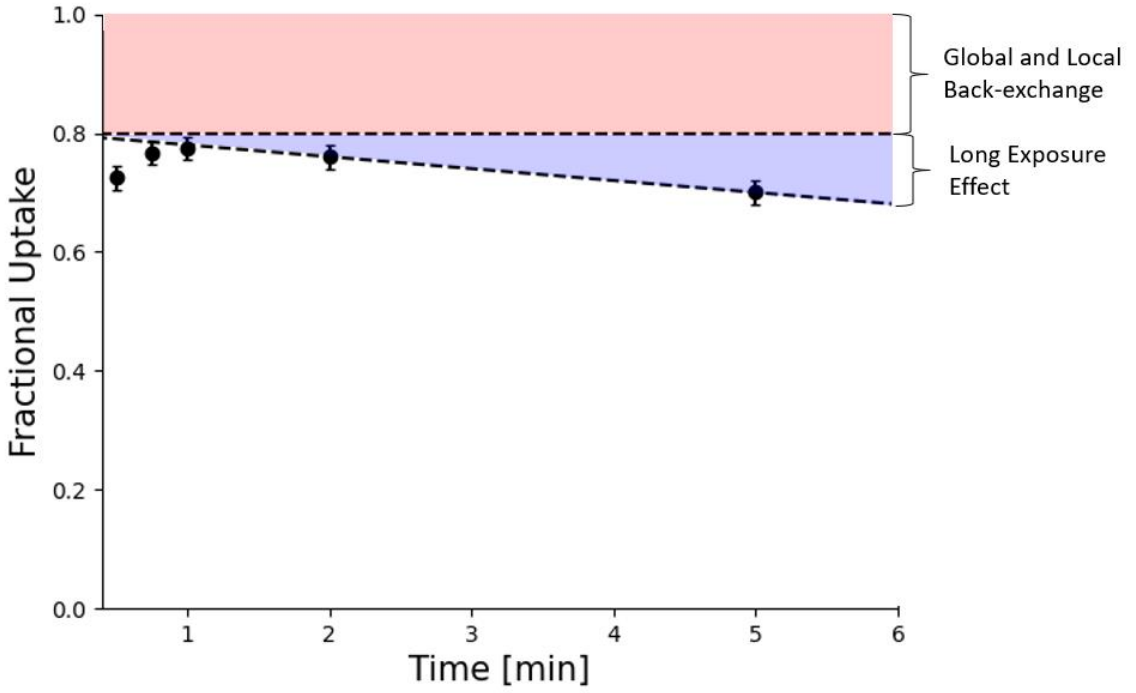


Figure 2.4. Back-exchange correction applied by the software DECA (Lumpkin and Komives 2019).

The global and local back-exchange correction (red) produces a peptide-dependent plateau, resulting from both protein-dependent back-exchange (occurring from sample dilution into labelling buffer to digestion) and peptide-dependent back-exchange (occurring from digestion to detection in the mass spectrometer). The long exposure effect is an apparent back-exchange correction suggested by DECA which consists of a linear correction that is universally applied to all peptides. Adapted with permission from data published by the authors of DECA **(Lumpkin and Komives 2019)**.

The corrections for back-exchange mentioned in this section underscore our limited knowledge of the phenomenon, and several questions remain unsolved: what is the best strategy to acquire a fully deuterated sample? What percentage of forward- and back-exchange is occurring during deuterium labelling, during the quench procedure, during digestion and in the gas phase, respectively? Can we reduce or eliminate any of these contributions, or at least control them so that they can be quantified accurately? Fundamental studies are needed to systematically answer these questions, for example studying the behaviour of model proteins while varying the deuterium percentages in

the quench buffer, or by replacing the water-based LC solutions with deuterium-based equivalents. While a proper back-exchange correction offers minor advantages for differential studies, it becomes crucial when integrating experimental data with modelling (i.e. for the methods described in **section 2.5.4**) as the *standard* back-exchange correction may yield inaccuracies in that it assumes all resides in a peptide back-exchange to the same extent.

2.4.3 Replicates

The reported mass increase (**Eq. 2.5**) or fractional uptake (**Eq. 2.6**) is averaged over the available number of replicates, generally limited to 3 or 4. The main factor limiting the number of replicates in HDX-MS experiments is the cost associated with additional sample consumption and instrument runs. The error associated with the experimental measure is either the standard deviation or the standard error (standard deviation divided by the square root of the number of replicates). Replication allows to assess whether the observed differences are likely to occur by chance or not (Vaux, Fidler, and Cumming 2012), and to ensure the reliability of the conclusions drawn from the observed data. Increasing the number of replicates results in a more precise inference regarding differences between groups (Oberg and Vitek 2009).

Not all replicates are equivalent. In the context of HDX-MS experiments (as well as proteomics and other biophysical techniques), replicates can be divided into two categories (Engen and Wales 2015): *biological* replicates, which can derive from i) independent protein expression or isolation from source tissue and ii) steps prior to the addition of deuterium (e.g. incubation with a ligand or membrane), and *technical* replicates, which can be in turn subdivided into three subcategories: i) *labelling* replicates, corresponding to independent deuterium additions to the same protein stock material, testing sample conditions during labelling (timing, pH, temperature) and LC-MS parameters, ii) *analysis* replicates, which are repeated LC-MS injections of the identically labelled sample, testing the variables from the point of injection into the

LC-MS system, and iii) *processing* replicates, that are software-based replicates on the same set of data that test the computational parameters and data processing reliability.

These types of replicates have a well-defined hierarchy: inferences drawn from biological replicates are more powerful than inferences made from technical replicates (Moroco and Engen 2015). Currently available software are not able to account for these differences. Multiple approaches are being used to account for technical and biological replicates. A commonly applied approach is averaging the deuteration values of the technical replicates within each biological replicate. This strategy yields consistent estimates, but gives incorrect uncertainty estimates, leading in a differential study to a higher number of false findings. Another approach is to analyse each biological replicate separately. Such an approach ignores the dependencies of the technical replicates within a biological replicate and ignores relevant biological variation, limiting generalization and replication of results. Both approaches should be avoided as they do not properly acknowledge the data structure. A third approach is the use of statistical models and tests that account for the level of replication (technical/biological), such as linear mixed models, which are described in more detail in **section 2.5.2.3**. It is worth mentioning that most published studies report only technical replicates. As a general recommendation, biological replicates should be prioritized over technical replicates whenever possible. When the number of biological replicates is limited (e.g. when there are only two), collecting data from both biological replicates, along with multiple technical replicates, allows for more robust inferences than relying on technical replicates alone. In the latter case, implementing a mixed effects model is essential to appropriately account for the level of replication.

2.4.4 Charge state effect

Many peptides can be found in electrospray ionization mass spectra with more than one charge state, and the apparent deuterium uptake behaviour of the same peptide at different charge states can show systematic differences. This is a well-known but rarely reported effect (Salisbury, Liu, and Agar 2014), and is caused by back-exchange post-

ionization in the electrospray source and gas phase of the mass spectrometer. Guttman et al (Guttman et al. 2016) demonstrated that this charge state offset, which occurs to different extents on different instruments, is due to non-uniform gas-phase exchange with water vapour within the ion optics of the instrument. For example, such back-exchange in a Waters Synapt G2-Si can be eliminated by adjusting the settings of the StepWave ion guide (mainly DC offset potential and the travelling wave height and velocity). There are two policies implemented by the available software packages: 1) the mass increase (or fractional uptake) is reported as an average over the available charge states of a peptide; 2) only the mass increase (or fractional uptake) of the most intense charge state is reported. We note here that neither option is ideal as the first one is not able to account for the possible systematic difference in deuterium loss between charge states, and the second introduces a *selection bias* in the analysis. A third alternative, which probably represents the best option, is to analyse different charge states individually and check that the results are consistent across the different charge states; in this latter strategy replication may be a problem as not all charge states are found for each replicate or condition.

2.4.5 Linderstrøm-Lang model for peptide-level data

The Linderstrøm-Lang model (**Eq. 2.1**) describes HDX at the level of the single amino acid. However, most HDX-MS experiments detect the deuterium uptake of a protein via its proteolytic peptides. For this reason, HDX-MS data are *coarse-grained*: they monitor the behaviour of entities (peptides) that are smaller than the whole system (protein) but bigger than the smallest resolvable unit (amino acid).

For a peptide with N exchangeable residues (i.e. excluding prolines), the deuterium uptake (D) of the proteolytic peptide can be written, using the Linderstrøm-Lang model, as the sum of the uptake d_i of its residues:

Eq. 2.7 $$D(t) = \sum_{i=2}^{N} d_i = \sum_{i=2}^{N} \left(1 - e^{-\frac{k_{int,i}}{P_i}t}\right)$$

The first amino acid ($i = 1$) is excluded from the contributing residues because its amide hydrogen is lost upon digestion. Sometimes, depending on the sequence of the proteolytic peptide, also the second residue should be excluded, assuming it rapidly back-exchanges during the quench step and the deuteration is lost (Walters et al. 2012). In certain sequences, such as those containing histidines, the back-exchange rate for even a middle amide can be so fast that all deuterium will be nearly lost by the time the peptide is analysed, and therefore the amide will not contribute to the overall deuterium measurement. A paper by Hamuro nicely summarized the expected deuterium loss for different sequence contexts (Hamuro 2021b).

One of the challenges for the analysis of HDX-MS data is to retrieve single residue information (i.e., the individual protection factors) from peptide-level data. In statistics, this problem is defined as *underdetermined*: the number of parameters to be estimated is greater than the number of experimental data points (Stofella et al. 2022). In the case of an isolated peptide formed by N residues, whose exchange has been monitored at K time points, we can distinguish two scenarios (assuming for simplicity that experimental error is negligible): i) when $N > K$ (which is the most common case as the average peptide length is ~ 10 amino acids and the HDX is generally detected at between 3 to 5 time points), the consequence of under-determination is that there are multiple solutions (i.e. patterns of protection factors) in agreement with experimental data; ii) when $N < K$, there is one only solution in agreement with experimental data, but the extracted protection factors cannot be assigned to a specific residue (indeed, **Eq. 2.7** does not account for the order of the residues). Using the complementary information contained in overlapping peptides helps reducing the multiplicity of the solutions, up to the point that single residue information can be in principle obtained in an *ideal* dataset where all peptides differ by one amino acid only (Kan et al. 2013). This is usually not the case, and other approaches have been used instead. Different methods aiming to extract protection factors from HDX-MS data are discussed in detail in **section 2.5.4**.

2.4.6 Visualization of pre-processed data for one condition

Pre-processed data are generally visualized through *uptake curves* (**Figure 2.5A**), reporting either the (average) mass increase (Eq. 2.5, in Da) or the (average) fractional uptake (**Eq. 2.6**, in % of the maximum) at different labelling times of proteolytic peptides. Heat maps (**Figure 2.5B**) can be generated to visualize the time evolution of the uptake at a pseudo-residue resolution along the sequence of the protein. Generally, the deuterium uptake of a residue at a specific labelling time is calculated as an equal fraction of the average over the mass increase (or fractional uptake) of the peptides covering the amino acid position. However, this calculation varies from software to software. For example, DECA (Lumpkin and Komives 2019) generates heat maps by assigning residues to the most representative peptide available (i.e. the shortest). An alternative approach uses weighted averaging, where peptide uptake values are weighted by the inverse of the peptide's length (Keppel and Weis 2015). If the structure of the protein is available, the pseudo-residue uptake provided by the heat map can be mapped onto a 3D structure (**Figure 2.5C**). Many pieces of software provide a PyMol script to generate these plots. HDX-Viewer (Bouyssié et al. 2019) is an online tool that was developed to provide an easy-to-use interface to visualize deuteration within the structure of the protein. It is worth stressing that these representations are useful tools to map experimental data onto protein models, but they can be misleading as the high resolution achieved is artificial.

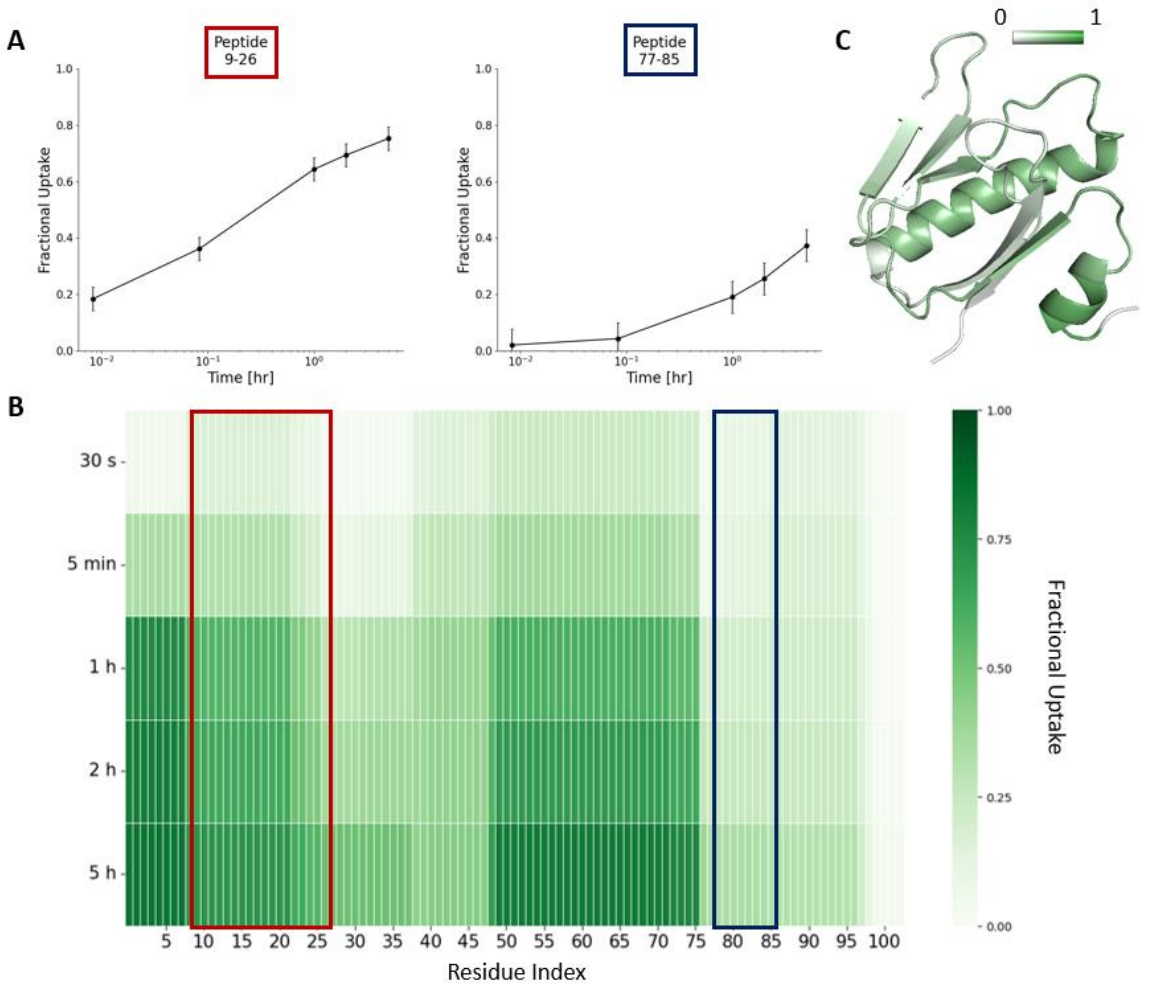


Figure 2.5. Visualization of pre-processed data for one condition.

(A) The uptake curve shoes the fractional uptake as a function of the labelling time. Average and standard deviation are displayed. (B) The heat map shows the fractional uptake as a function of the labelling time along the sequence of the protein. The fractional uptake of each residue is the equal fraction of the average fractional uptake of the proteolytic peptides covering that specific amino acid position. (C) The heat map is projected onto the protein structure at labelling time 1 hour. Data showing the HDX of cholera toxin B acquired by the author.

Alternative visualization tools involving multivariate analysis, such as principal component analysis (PCA) or spectral mixture analysis (SMA), can be used to check the quality of the data (e.g. to see if samples from the same condition or time group together) (Claesen et al. 2021), and have been proven useful to show whether

compounds with similar in vivo properties were forming statistically distinct clusters (Belorusova et al. 2019).

2.4.7 Pre-processing data with commercial software

It is needless to say that a correct pre-processing of data is a crucial step in the data analysis workflow of HDX-MS experiments as poorly curated data can lead to incorrect biological conclusions. In the context of HDX-MS data, we define as data "pre-processing" the steps of the data analysis workflow that start from the raw data and a list of potential peptides, and lead to the generation of deuterium uptake curves. Most HDX-MS publications utilize commercial software to pre-process raw data, namely DynamX for Waters instruments and HDExaminer for Thermo Scientific/Trajan. These two programmes share common features: they require knowledge of the protein sequence, in the form of a peptide list (whose generation is later discussed in **section 2.5.1**), and the raw HDX-MS data as inputs. They enable the identification and assignment of the undeuterated and labelled isotopic envelopes, to calculate the absolute uptake of peptides, to visualize the data through coverage maps, uptake plots and heat maps; and they return a spreadsheet containing the information about the uptake of peptides over time. They mostly differ on how the user can interact with the isotopic envelopes identified by the software and edit them manually. They also share the same limitations: they do not perform statistical analysis, back-exchange correction or fully deuterated normalization, which must be done using third-party software packages. While these commercial software packages are critical for pre-processing raw data, they must be integrated with other software packages to achieve comprehensive and publishable results.

2.5 Computational tools for HDX-MS data analysis

2.5.1 Automated peptide search and identification

At the beginning of the data analysis pipeline, it is necessary to identify peptides and assign peaks in the raw mass spectra. As already mentioned, the identification of

proteolytic peptides is generally performed using commercial software included with the instrument, and the quality of the isotopic envelopes of each peptide is then checked manually. Peptides showing saturation (the intensity of the peptide signal exceeds the instrument's dynamic range, altering the shape of the isotopic envelope), multimodal behaviour which can be due to EX1 or EXX kinetics (see **section 2.5.3**), carryover (peptides retained on the fluidics system from the previous sample injection) or ambiguous assignment (e.g. due to the presence of different envelopes in the same m/z range) can be kept or rejected depending on the practitioner. There are no clear guidelines on how to perform these assignments, and policies vary from group to group. Consequently, this pre-processing step is time-consuming and user dependent. Moreover, a major disadvantage of commercial software packages is that they do not allow to export the isotopic envelopes but only the average m/z values, making it hard to retrieve information about the detailed characteristics of the assigned mass spectra of the peptides.

Tools have been developed to tackle the drawbacks mentioned above. ExMS2 (Kan et al. 2019) proposes an automated peptide validation pipeline to speed up the peptide quality checks. This requires as input a peptide list generated by SEQUEST/Bioworks (alternatively Proteome Discoverer or MassLynx) from the control sample (undeuterated, all-H protein). Each peptide is associated with its chromatographic retention time (RT) and its m/z value. For each peptide in the list, ExMS2 selects the MS scans within the known RT window and compares the experimental spectra with calculable mass spectrometric information, such as monoisotopic mass, charge state, and isotopic peak positions. The process is repeated for each sample at the different time points available. ExMS2 records m/z values and relative intensities for each isotopic peak to define the shape of the isotopic envelope. The recorded peptides are validated through 12 quality tests (six performed on a peptide level and six on a multi-peptide level), for example checking if the overall peak intensity is above a certain threshold or if the peak is within the possible m/z range for a peptide in the list. Peptides failing one or more tests are flagged and can be manually inspected.

Mass Spec Studio (Raval et al. 2021; Rey et al. 2014; Ziemianowicz et al. 2020) first proposed HXpipe (peptide identification and peptide evaluation) as a tool for automated peptide search and validation. Two searches are performed independently: i) an MS/MS search, which looks for peptides in the LC-MS/MS files using one of two available search engines (MS-GF+ or OMSSA+); ii) an MS search, which uses a peak picker that scans within the LC-MS/MS data to create a library of chromatographic features, which are then compared with theoretical isotopic distributions calculated using Senko's Averagine model (Senko, Beu, and McLaffertycor 1995) for peptides. The results from the two searches are then combined together. A new module, named AutoHX (Filandr et al. 2024), has been implemented into Mass Spec Studio, to facilitate (and ideally remove) the manual inspection of the peptide search. AutoHX leverages the information contained in the deuterated fragment peptides to i) validate the identity of the peptide and ii) confirm the deuteration level of the precursor peptide by checking that the deuterium content of the peptide fragments has a linear relationship with the fragment length. This automatic authentication and validation, which exploits MS/MS data and uses deuterium-scrambled CID or HCD fragments as surrogates that confirm the identity and the deuteration value of any given peptide, yields objective results with known certainty, rather than biased results with unknown certainty provided by a traditional approach, which uses MS data only and is followed by laborious manual (i.e. user dependent) validation.

In the previous paragraphs, we reviewed different software packages designed for automated peptide search and identification from raw LC-MS data. These tools bridge the gap between researchers and raw data, facilitating the pre-processing and validating peptide-peak assignments along with deuteration values at specific labelling time points. While they serve as a viable alternative to the commercial software described in **section 2.4.7**, their adoption is limited mostly due to a lack of know-how outside of the group of researchers which generated them. Although their documentation is generally robust, we believe that additional tutorials and workshops for HDX users would help with their broader adoption in the community, and this has

also been suggested at the recent conference of the International Society of HDX-MS in April 2024 in Monterey (California, USA). Among the software reviewed, Mass Spec Studio stands out as the most comprehensive; integrating automated peptide search and identification with subsequent workflow steps, such as differential analysis (**section 2.5.2**) and multimodal analysis (**section 2.5.3**). It is crucial for software developers to consider the integration of diverse data types (e.g. tools for analysing ExD or UVPD fragment data) and ensure easy access to processed data (e.g., straightforward export of processed isotopic envelopes). Likewise, instrument manufacturers should be encouraged to enable the export of HDX-MS datasets with key information such as the isotope patterns and charge states of peptides.

2.5.2 Differential analysis

The analysis of HDX-MS data generally relies on a side-by-side comparison of two (or more) conditions (e.g. a protein in absence or presence of a ligand). For each proteolytic peptide, the difference in deuterium content obtained from the different experimental conditions is classified as *significant*, or not, using thresholding and/or statistical tests and models. Differences in the uptake pattern of peptides highlight regions of the protein where a structural perturbation has occurred (binding site, allosteric change, etc.). There are two strategies to analyse differential HDX-MS data. The first (and most used) looks at the difference in deuterium content *at a given time point*: manual thresholding (**section 2.5.2.1**), simple hypothesis tests (**section 2.5.2.2**), or linear models (**section 2.5.2.3**). The second approach compares deuterium uptake curves (**section 2.5.2.5**). These two strategies, summarized in **Table 2.2**, are described in this section.

Table 2.2. List of software to analysis differential HDX-MS data.

Differential Analysis	
Name	**Approach**
DECA (Lumpkin and Komives 2019)	T-test
Deuteros 2.0 (Lau et al. 2020a)	Linear Model
HaDeX (Puchała et al. 2020)	T-test

HD-eXplosion (Zhang et al. 2021)	T-test
HDflex (Seetaloo et al. 2022)	T-test
HDXAnalyzer (Liu et al. 2011)	Linear Model
Hdxstats (Crook et al. 2022)	Functional analysis
Mass Spec Studio (Raval et al. 2021; Rey et al. 2014; Ziemianowicz et al. 2020)	T-test
MEMHDX (Hourdel et al. 2016)	Mixed model

2.5.2.1 Manual thresholding

One approach used to analyse differential HDX-MS data consists of defining a manual threshold for the difference in deuterium content between conditions (Iacob et al. 2013; Ständer et al. 2021; Wei et al. 2012). This threshold is set to a predefined value (generally 0.5 Da (Houde, Berkowitz, and Engen 2011)) or based on the standard deviation of the data (e.g. using the pooled standard deviation (Hageman and Weis 2019)). If the difference in deuterium content at a specific time point exceeds this threshold, the peptide is classified as *different.* This approach ignores the variability of the peptide-deuteration levels and can therefore lead to false findings (Claesen et al. 2021; Hageman and Weis 2019). For example, defining a strict threshold to reduce the number of false positives leads to ignoring small yet biologically relevant differences in deuteration (i.e. false negatives), while a generous threshold limits the number of false negatives, but results in many false positives. We therefore advise against manual thresholding approaches and advocate the use of statistical methods to test for differences, as they account for the variability of measured deuterium levels and thus control the number of false findings.

2.5.2.2 Simple hypothesis testing

In simple hypothesis testing, a null hypothesis (H_0) is compared against an alternative hypothesis (H_a). In differential HDX-MS, the null hypothesis commonly states that there is no difference in the deuterium content of a peptide between two or more conditions, while the alternative hypothesis claims that there is a difference. Statistical tests are used to test the null hypothesis, i.e., to reject or not to reject the null hypothesis, by

calculating a test statistic. Student's *t*-test is commonly used when one wants to compare the means between conditions/groups. Student's *t*-test is in essence a signal-to-noise ratio test, where the difference in the average deuterium content is divided by a nuisance parameter, which is a function of the variability of the data. The larger this ratio, the more likely the null hypothesis can be rejected in favour of the alternative hypothesis. The exact value (*critical value*) required to reject the null hypothesis depends on the number of observations and the specified significance level (α). Generally, a *p*-value is reported instead of the critical value. If this *p*-value is smaller or equal than α, the null hypothesis can be rejected. When more than two conditions have to be compared, an *F*-test which tests if at least one mean is different from the others. Student's *t*-test and the *F*-test both assume that the data is normally distributed. If this is not the case, non-parametric alternatives, i.e. the Wilcoxon signed-rank test or Mann-Whitney *U*-test and the Kruskal-Wallis test can be used. Note that when the underlying assumptions of the parametric tests are true, the non-parametric test statistics are less powerful than their parametric counterparts, i.e. they identify less differences in deuteration that are truly different as statistically significant.

Differential HDX-MS experiments are generally done with a limited number of replicates. As a consequence, the variability of the deuterium content of a peptide is harder to estimate accurately. This can potentially lead to more false findings, i.e., more false positives and/or false negatives. Claesen et al. (Claesen et al. 2021) proposed using moderated *t*- and *F*-statistics instead of Student's *t*-test and *F*-statistics. These test statistics borrow information from other peptides with similar deuteration values to reliably estimate the standard error of the mean, resulting in a lower number of false findings.

2.5.2.3 Linear regression models

Although simple hypothesis testing is a convenient way to test for differential hydrogen/deuterium exchange (per peptide), uniting all hypothesis tests in a linear regression model allows to directly estimate differences between the different groups

or conditions (protein states). Additionally, it allows to correct for other factors (*confounders*) that could have an effect on the deuteration.

In a linear regression model, the response variable or dependent variable (**y**) is a linear function of one or more explanatory or independent variables (**X**):

Eq. 2.8 $$\mathbf{y} = \boldsymbol{\beta}\mathbf{x} + \boldsymbol{\varepsilon}$$

where ε is the residual error and follows a normal distribution ($\boldsymbol{\varepsilon} \sim \mathrm{N}(\mu = 0, \sigma^2)$), and $\boldsymbol{\beta}$ are regression coefficients that are derived from the data and express the effect of the explanatory variables. The reader would be familiar with the simplest case of a straight line with slope m and intercept q: $y = mx + q$, which is equivalent to **Eq. 2.8** in the case $\mathbf{x} = \binom{1}{x}$ and $\boldsymbol{\beta} = \binom{q}{m}$. In the latter case, a linear regression model fits a line to the data and allows to evaluate the effect (m) of the explanatory variable x on the response variable (y). The same linear regression model in **Eq. 2.8** to compare means of different groups and check whether they are significant different from each other by testing the estimates for $\boldsymbol{\beta}$ with a t-test. The advantage of a linear model over a t-test is that it can account for more than one explanatory variable at a time.

Deuteros 2.0 (Lau et al. 2020a) and HDX-Analyzer (Liu et al. 2011) implemented the following multiple regression-model, where the absolute deuteration of a peptide **D** is modelled as a function of the explanatory variables **Time** (labelling time) and **State** (biological state of the protein):

Eq. 2.9 $$\mathbf{D} = \boldsymbol{\alpha} + \boldsymbol{\beta} \times \mathbf{Time} + \boldsymbol{\delta} \times \mathbf{State} + \boldsymbol{\gamma} \times (\mathbf{Time} \times \mathbf{State}) + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\alpha}$ represent the intercepts of the model, $\boldsymbol{\beta}$ the regression coefficients for the labelling timepoints, $\boldsymbol{\delta}$ the regression coefficients for the different conditions/states, $\boldsymbol{\gamma}$ the regression coefficients for the interaction of state and time, and $\boldsymbol{\varepsilon}$ the residual errors of the model. In this model, **Time** and **State** are categorical variables, i.e., characteristics that are not quantifiable. In other words, if we have three time points, they are treated as time point number one, two and three (rather than, for example, 30

seconds, 5 minutes and 1 hour) – as a consequence, interchanging the time points would not affect the results. This regression model can be used to test whether changes in the deuterium-uptake of a peptide are associated to changes in state and/or time, and the interaction between state and time. Note that including the labelling time points as a continuous variable (rather than categorical variables) is possible, but time would have to be transformed to account for the non-linear relationship between labelling time and deuteration uptake, or a non-linear regression model would be required. The proposed multiple regression model can also be extended by adding other (categorical) variables, for example, charge state.

Depending on the experimental design, HDX-MS data can have correlated and/or repeated measures, for example, when an experiment is run in different batches or when both technical and biological replicates are acquired (see **section 2.4.3**). In the latter case, for example, we expect data from within the same biological replicate to be more similar to data between different biological replicates. The linear model, as defined in (**Eq. 2.8** and **Eq. 2.9**), can be updated to a linear mixed effects model to account for the correlation present in the data:

Eq. 2.10 $$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}$ represent the residual error of the model and follows a normal distribution ($\boldsymbol{\varepsilon} \sim N(0, \sigma_{\varepsilon}^2)$), $\mathbf{u}$ is an unknown vector of random effects and also follows a normal distribution ($\boldsymbol{u} \sim N(0, \sigma_{u}^2)$), and $\mathbf{Z}$ is a design matrix for the random effects. The random effects, $\mathbf{u}$, account for the correlation that is present in the data.

To clarify the content of **Eq. 2.10**, we now provide two examples where using a mixed model is advisable in the context of HDX-MS experiments.

Suppose we performed an experiment with 3 biological replicates, and 3 technical replicates per biological replicate (i.e. 9 experiments). The design matrix $\mathbf{Z}$ in **Eq. 2.10** indicates which observations come from which biological replicate. For a peptide i at time point j, we can assign $\boldsymbol{uZ}_{ij} = 1$ for all technical replicates of the first biological

replicate and similarly $\boldsymbol{uZ}_{ij} = 2$ and $\boldsymbol{uZ}_{ij} = 3$ for the technical replicates acquired from the second and third biological replicate).

Alternatively, suppose that the same protein was studied under different experimental conditions in three different laboratories, and we wanted to combine all measurements into a single dataset to perform a separate meta-analysis. The different protocols implemented by the different groups (for sample handling, automation of the LEAP robot, different parameters for the LC-MS gradient, etc.) introduce random fluctuations to the deuterium uptake value of the same peptide under the same experimental condition. Differences in uptake between conditions (for the same peptide at the same time point) are systematic and should be visible, but combining the results from the different laboratories without considering this as a source of random effects might introduce a bias into the outcomes of the experiment. For example, a peptide with significant differences correctly detected (i.e. a true positive) by the three different laboratories might be misclassified as non-significant if all measurements were combined (**Figure 2.6**). A mixed model can deconvolve the effect of the standard deviation of the different laboratories on the standard deviation of the combined dataset.

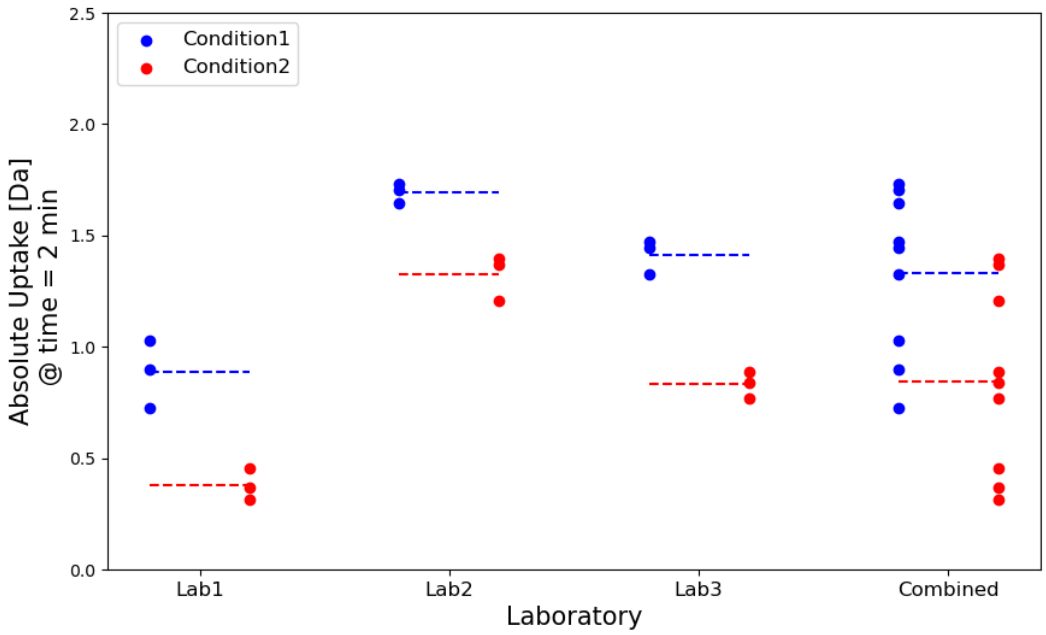


Figure 2.6. Example of random effects affecting the outcomes of an analysis (simulated data). The absolute uptake of a proteolytic peptide of a protein has been measured at labelling time 2 minutes under two different experimental conditions (Condition1, blue; Condition 2, red) in three different laboratories. Each laboratory classifies the difference in uptake to be significant for the peptide (p-values are 0.0066 for Lab1, 0.0048 for Lab2, 0.0005 for Lab3). When the results are combined together into a single dataset, the peptide is no longer significant (p-value = 0.0183).

MEMHDX (Hourdel et al. 2016) is the only software that implements a mixed model for the analysis of HDX-MS data. Here, time and experimental condition represent fixed effects and the replicated or repeated measures are considered as a random effect, meaning that each technical replicate is assigned to a *different* random effect. The GUI version of the software only allows the user to perform a traditional differential analysis, and it does not allow the cross-experiment statistics described in the experiments above, which can however be performed using the multiple statistical packages in R, such as nlme (Pinheiro, Bates, and R Core Team 2022).

2.5.2.4 Multiple testing or multiplicity

Hypothesis tests (such as the *t*-test) are prone to false positive results when multiple comparisons are performed simultaneously, i.e., comparing peptides across conditions at each time-point separately. In order to control the probability of finding false positives, several multiple testing or multiplicity correction approaches have been proposed that adjust the p-value (Dudoit and Laan 2008). The best-known multiple testing correction method is the Bonferroni correction, which divides the significant threshold α by the number of comparisons m, therefore the adjusted significance threshold reads $\alpha^* = \alpha/m$. However, the Bonferroni method is very conservative (Claesen et al. 2021), i.e. it leads to a very high number of false negatives. Another well-known approach is the Benjamini-Hochberg procedure(Benjamini et al. 2001), which is less conservative than the Bonferroni approach. Hageman and Weis proposed a *hybrid* approach that combines *t*-tests with manual thresholding to correct for multiplicity (Hageman and Weis 2019): the difference in deuterium content between two conditions is classified as statistically significant if two conditions are met simultaneously: i) the *p*-value returned by the *t*-test is smaller than the significance level (α) and ii) the difference in deuteration is greater than a pre-defined threshold. This hybrid approach is implemented in HaDeX (Puchała et al. 2020), HD-eXplosion (Zhang et al. 2021) and Mass Spec Studio (Rey et al. 2014).

2.5.2.5 Comparing deuterium uptake curves

Crook et al. introduced a novel approach to the analysis of HDX-MS data in the framework of functional analysis (Crook et al. 2022). Experimental uptake curves of peptides are fitted with a Weibull model (also referred to as stretched exponential) of the form:

Eq. 2.11 $$D(t) = a\left(1 - e^{-bt^q}\right) + d$$

where the parameter d represents the mass at time 0 (no exchange; undeuterated), which is inferred from the data; a controls the value at which the exchange reaches a

plateau (maximum incorporation); b, the exchange rate constant, which models the exchange kinetics; q refers to additional factors that are deflecting the uptake curve from a single exponential behaviour. The stretched exponential in **Eq. 2.11** approximates the multi-exponential behaviour derived from the Linderstrøm-Lang model (**Eq. 2.7**).

The Weibull model (**Eq. 2.11**) is fitted with experimental data from two conditions and tools of functional analysis are implemented to assess whether the curves are significantly different. The underlying null hypothesis of functional analysis is that the same parameters can fit experimental curves from both conditions. The alternative hypothesis is that two independent models describe better the data. P-values are returned by an F test, and multiple testing corrections (see **section 2.5.2.4**) can be applied (as for t-tests, linear models and mixed models).

When using a linear model, time is modelled as a categorical variable: changing the order of time points does not affect the results of the analysis. With a mixed model, the random effect can account for the correlation present between time points for a given peptide. The major advantage of the functional model implemented is the possibility of explicitly modelling the deuterium content as a function of time, which allows to incorporate intrinsic exchange rates of the residues forming a peptide. This comes at the cost of acquiring a relatively large number of *informative* time points (early/late, spacing) to properly sample the uptake curve of each peptide.

2.5.2.6 Visualization of differential analysis

The tools to visualize data for a single condition (uptake plots, heat maps and 3D structure visualization, see **section 2.4.6**) can also be used to qualitatively visualize the results of a differential analysis (**Figure 2.7**). Differential heat maps show the difference in uptake between 2 conditions rather than the mass increase of a single condition. These differences are often mapped onto a 3D protein structure, with a colour scheme showing regions in white without significant differences, in blue those

that are more protected in the target condition and in red regions that are less protected.

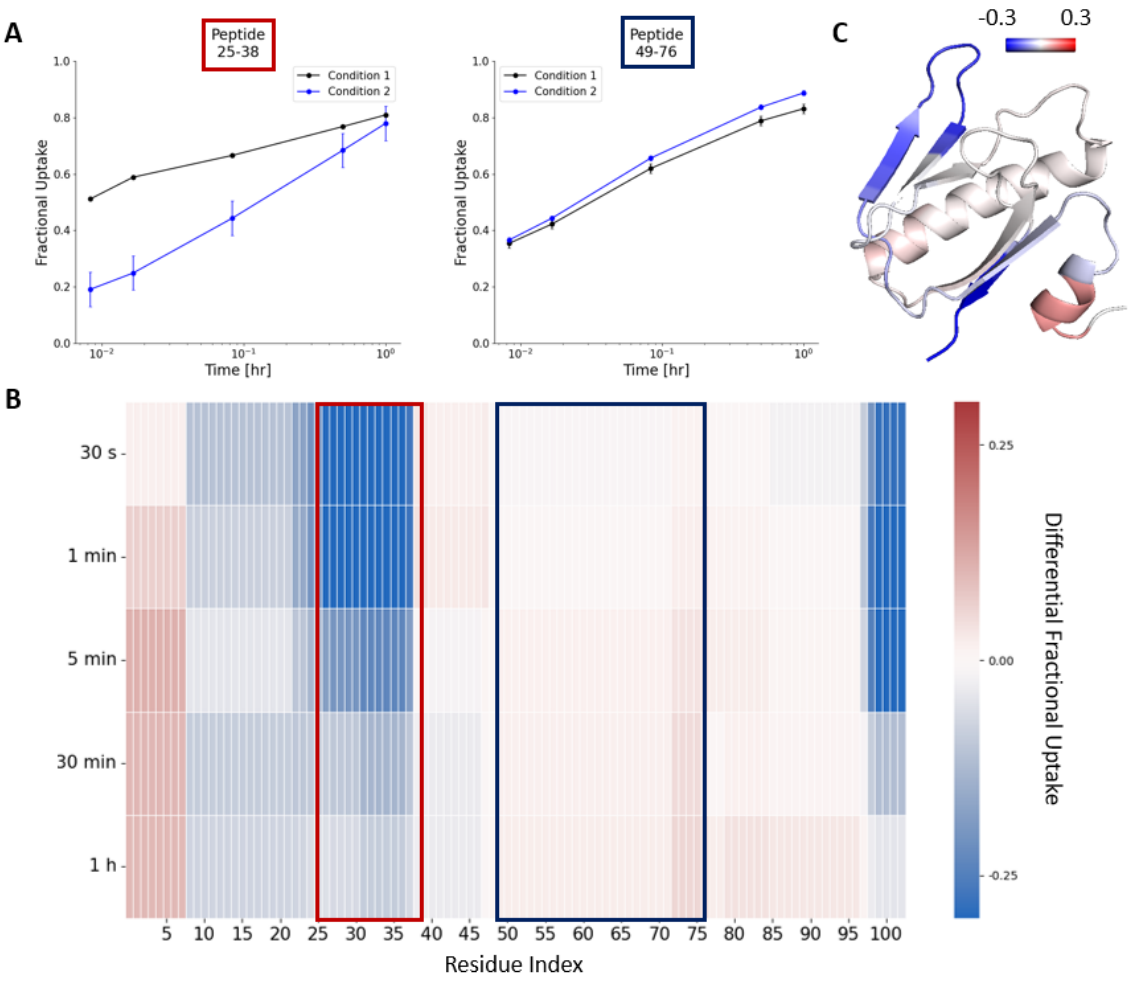


Figure 2.7. Qualitative visualization of HDX-MS differential analysis.

(A) Uptake plots for peptides covering residues 25-38 and 49-76 are shown for two different experimental conditions. Differing curves highlight structural changes in this area of the protein. (B) Differential heat maps show the difference in uptake between two conditions as a function of the labelling time and along the sequence of the protein. Blue regions identify areas where Condition 2 is more protected than Condition 1; red regions correspond to areas where Condition 2 is less protected than Condition 1. (C) The differential heat map is mapped onto the 3D structure of the protein at labelling time 5 minutes. Data showing the HDX of cholera toxing B in presence/absence of sugar GM1os (acquired by the author).

The plots in **Figure 2.7** do not show the results provided by the statistical test used. The results of a differential analysis are generally reported in publications using Woods plots: proteolytic peptides are visualized across the sequence of the protein with horizontal bars and positioned along the y-axis according to the difference in uptake between two conditions; peptides showing statistically significant differences are highlighted (in blue or red). The statistical significance can be defined either by a single threshold on the p-value (pvalue $< \alpha$) or by a double threshold on the p-value and on the difference in uptake. If a t-test, a linear model or a mixed model is used, each time point will be visualized on a different Woods plot; if functional analysis is implemented, the results of the whole time-course will be displayed in a single Woods plot.

The volcano plot is an alternative tool to visualize the results of a differential analysis. Each proteolytic peptide is a point in the plot: the horizontal axis represents the difference in uptake between conditions; the vertical axis shows $-\log(\mathrm{p-value})$, which can be considered a measure of the statistical significance (the p-value depends on the statistical test implemented) : the higher the differences between conditions, the lower the p-value, and therefore the higher $-\log(\mathrm{p-value})$. The volcano plot is ideal to visualize statistically significant peptides using the double threshold (on the p-value and on the difference in uptake), but it does not visualize the location of the peptide along the protein sequence.

The Manhattan plot is another alternative tool to directly visualize the p-values returned by the chosen statistical test along the sequence of the protein. In this plot, the horizontal axis represents the peptide index, while the vertical axis shows the statistical significance ($-\log(\mathrm{pvalue})$). Alternatively, as shown in **Figure 2.8**, peptides can be visualized as horizontal bars positioned along the sequence of the protein.

Woods plots, volcano plots and Manhattan plots (examples are shown in **Figure 2.8**) are all valid options to show the results of a differential analysis. We find the Woods plots to be more *complete* as they show directly the difference in uptake and the position of the perturbation, and indirectly the statistical significance. Volcano plots

show the difference in uptake and the statistical significance but fail to directly localize changes along the sequence of the protein; Manhattan plots can localize the differences and show the statistical significance but fail to show the difference in uptake.

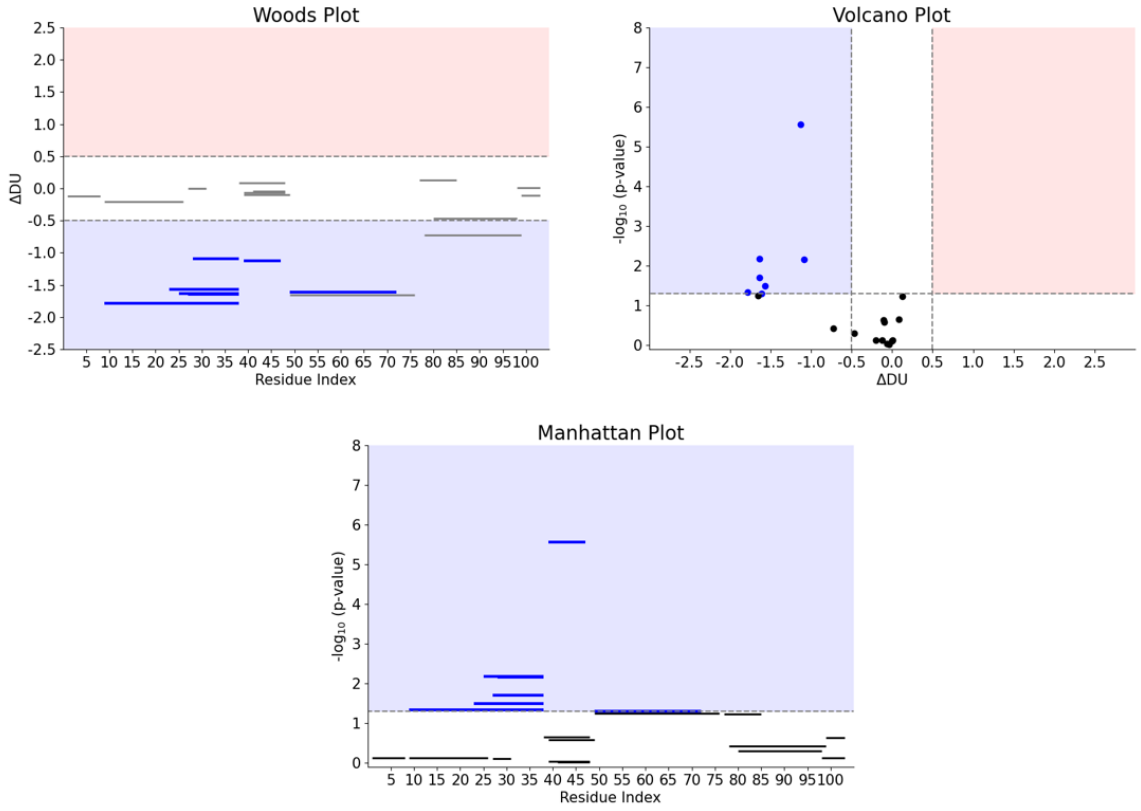


Figure 2.8. Quantitative visualization of the results of differential analysis (simulated data). Quantitative visualization of the results of differential analysis on the same dataset using Woods plots, Volcano plots and Manhattan plots. See main text for plot description.

2.5.2.7 Which statistical test to choose?

In **section 2.5.2**, we reviewed several strategies implemented in software to analyse differential HDX-MS data: i) manual thresholding, ii) simple hypothesis testing (t-test), iii) linear regression model, iv) mixed models, and iv) functional analysis-based strategies. We strongly suggest avoiding manual thresholding as it fails to control for false positives. While simple hypothesis testing is not inherently flawed, it can be easily generalized into a linear regression model. A t-test is limited to comparing one

explanatory variable at a time, whereas a linear model can account for multiple variables – such as labelling time, experimental condition, and charge state – making it more suitable for HDX-MS data. The choice between linear models, mixed models or functional analysis-based strategies depends on the dataset available and on the specific research question. Mixed models are ideal in cases where data are not independent, such as when both biological and technical replicates are available, when meta-analysis of datasets from different research groups needs to be performed, or when newly acquired data on one variant of a protein have to be compared with older data, possibly collected by different researchers but using the same instrument. In most scenarios, a linear model is the most pragmatic solution to assess statistically significant differences between two conditions. The function analysis-based strategies, on the other hand, are powerful when kinetic information is needed, but require an informed selection of time points to accurately estimate the underlying deuterium uptake curve.

2.5.3 Multimodal analysis

Sometimes the presence of EX1, mixed EX1/EX2 (also known as EXX) behavior, or the coexistence of multiple conformational states of a protein, can cause the isotopic envelope to assume a multimodal shape (**Figure 2.9**). Pure EX1 kinetics can produce two isotopic envelopes with fixed m/z values (the fully protonated and fully deuterated) but with variable intensities (the intensity of the fully deuterated envelope increases and the undeuterated one decreases accordingly over time). The coexistence of EX1 and EX2 kinetics (EXX) is also characterized by the presence of bimodal isotopic envelopes, with the first population gradually shifting towards higher m/z values (as in the EX2 regime) and the second associated to the fully deuterated spectrum (as in a pure EX1 kinetics); in this mixed regime, an intensity shift to the higher-deuterated state is observed. It is also common to find a multimodal behavior with two populations that can both undergo EX2 kinetics, which is associated with two distinct conformations of the protein that are not inter-exchanging (James et al. 2022; Oganesyan, Lento, and

Wilson 2018). In such cases, the modes of the bimodal spectrum should be deconvolved before comparing the intensity-weighted average of the individual populations with the statistical models described in **section 2.5.2**. In other words, two values of deuterium uptake are needed to fit the isotopic distribution properly and to quantify the fraction of molecules following EX1 or EX2 behavior (or, analogously, the population in either conformational state). Note that EX1 kinetics is a rare phenomenon and should not be confused with carryover (Fang et al. 2011).

To perform a multimodal analysis, the raw mass spectra of the proteolytic peptides are needed in order to obtain the full isotopic distributions. We note here that retrieving such raw spectra is not trivial: the majority of the tools described here require the csv output files generated by DynamX (for Waters instruments) or HDExaminer (for Thermo Fisher Scientific), which only contain information on the intensity-weighted average of the isotopic envelope that have been automatically assigned and manually curated. Manually analyzing raw data is very time-consuming and error prone, even for a dataset with a limited number of samples. One can also use tools from MS-based proteomics and/or MS-based metabolomics to extract the needed information from the raw spectra. However, these tools cannot be used out-of-the-box and are therefore not very user-friendly for the inexperienced user. A third option is to implement a method from scratch that takes as input the raw files, implements a peptide search, carries out automated or manual mass spectrum assignments, and stores information on the shape of the isotopic envelopes. The latter strategy has been developed by several groups which, being able to interface with raw data, have developed methods to study the bimodal behaviour caused by EXX kinetics or by the coexistence of multiple conformations.

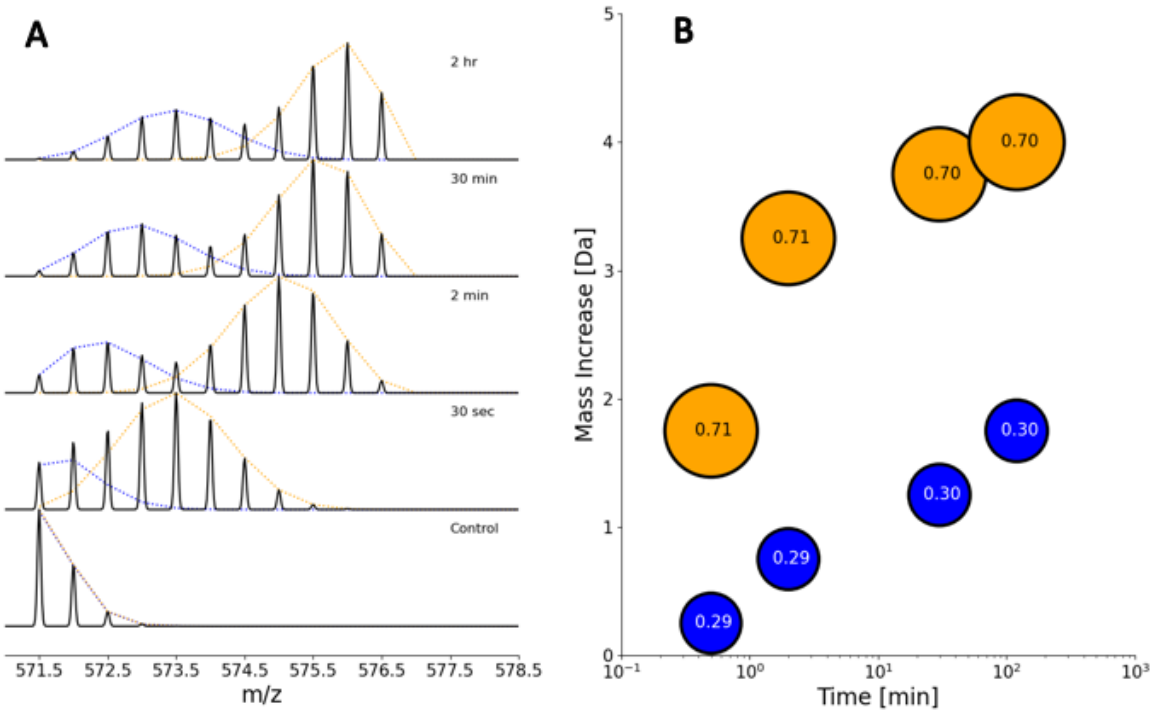


Figure 2.9. Multimodal behaviour from two coexisting conformations (simulated data).
(A) The time evolution of the isotopic envelope of a peptide with bimodal behaviour is shown. Both populations follow an EX2 kinetics, and therefore identify two distinct conformations of the protein that are not inter-exchanging. The bimodal distribution is fitted with two binomial distributions, and the mass increase and ratio of the two populations is recorded. (B) The mass increase of the two populations is shown as a function of time. The size of the scatter points is proportional to the fraction of molecules following the specific population.

Mass Spec Studio (Raval et al. 2021) and HX-Express (Guttman et al. 2013), for example, can identify peptides showing bimodal behavior in the isotopic distribution. These software packages allow fitting experimental spectra with a double binomial distribution and to extract the associated parameters, namely the center of mass of the two sub-distributions and their relative intensities.

ExMS2 (Kan et al. 2019) can detect peptides showing multimodal behavior through a "unimodality check" introduced in the latest version of the software to assess the quality of peptide selection. These peptides are flagged and can be further studied by a

module named “Multimodal analysis”. The isotopic envelope of the peptide can be fitted with several functions (varied binomials, uniform binomial and gaussian(s), gaussians, or reference shapes – in case a control sample displaying the pure sub-spectrum of one population is available). The multimodal behavior can be detected, and the parameters extracted through the fitting procedure can be used to determine the fraction of sample following EX1 or EX2 regime.

deMix (Na et al. 2019) is a recent method aiming to tackle the issue of discriminating different populations when a bimodal distribution appears due to mixed EX1 and EX2 behavior in HDX-MS data. The deuterated isotopic distribution (of every peptide and at every time point) is fitted with a separate binomial distribution. An optimal value for deuteration d_A is calculated. If the deuterated distribution is not explained enough by d_A, then bimodal analysis is performed. The top two-scoring deuteration values d_A and d_B are calculated. The resulting bimodal distribution is fitted with experimental data to determine how each species is populated. deMix reports two values of deuteration only if the error of the bimodal distribution is significantly improved and if the weight factor for the least abundant species is greater than 10%.

Here we presented various strategies for analyzing HDX-MS data of peptides exhibiting multimodal behavior, which can arise from several factors, including carryover, coexistence of different non-interacting protein conformations, EX1 kinetics, or mixed EXX kinetics. The methods discussed here enable robust deconvolution of the extent of deuteration of each population, but they do not inform the user *per se* about what causes the bimodal behaviour. If bimodality is known to be due to carryover, these methods allow for its correction (rather than redoing the experiment). In the case of coexisting protein conformations (both following EX2 kinetics), they help determining the fraction of molecules in each conformation. When the relative intensities of both populations are sufficiently high, this enables the study of the exchange kinetics of both conformations. In the rare instances of pure EX1 kinetics or mixed EX1/EX2 kinetics, these methods allow determination of the fraction of fast- and slow-exchanging

molecules. The weakness in the latter scenario is the unclear application of this information. Indeed, peptides showing EX1 or EXX kinetics are generally excluded from differential analysis. Sometimes, for example, standard EX2 kinetics might be observed for one protein state, while pure EX1 kinetics is observed for another. It is true that the emergence of EX1 or mixed EXX kinetics can qualitatively assess protein disorder, but quantitatively assess statistically significant differences between different states and integrating these data into modelling remain unresolved challenges.

2.5.4 Protection factor analysis

The Linderstrøm-Lang model (**Eq. 2.1**) describes HDX as a phenomenon occurring at the level of the single residue. The exchange kinetics follows an exponential law with an exponent that, in the EX2 limit, depends on the intrinsic exchange rate and on the protection factor (**Eq. 2.2**). The intrinsic exchange rate represents the rate that the same type of residue (amino acid) has in a completely unfolded structure. The protection factor of the residue depends on the local structure of the protein surrounding the residue. Retrieving protection factors from HDX-MS data would enable to connect the experimental data with microscopic properties that can be inferred from atomistic modelling and MD simulations. Indeed, protection factors can be measured for labelled residues of a protein through HDX-NMR (Dempsey 2001). However, the information provided by HDX-MS is coarse-grained to the peptide level and underdetermined (see **section 2.4.5**), and extracting protection factors (or exchange rates) at the resolution of the single amide from HDX-MS data is not trivial.

The spatial resolution of HDX-MS data can be increased experimentally. On one hand, different proteases (Ahn et al. 2013; Cravello et al. 2003; Kenji et al. 2005; Tsiatsiani et al. 2017) or multi-enzyme strategies have shown to be beneficial in increasing peptide overlaps (Sobott 2020). On the other hand, MS/MS fragment data can be exploited. Among the fragmentation techniques available, collision induced dissociation (CID) has the drawback of favoring H/D scrambling within the peptide (protons and deuterium atoms are mobile within the peptide). Alternative dissociation techniques, such as

ECD/ETD (more generally ExD) and UVPD, have been proven to increase spatial resolution while minimizing H/D scrambling (Mistarz et al. 2018; Wollenberg et al. 2020). However, reaching single residue resolution for the whole protein with these fragmentation techniques is still challenging, mainly for two reasons: sensitivity (the intensity of peptides and fragments vary significantly due to the broad specificity of pepsin, the fragmentation and ESI efficiencies) and protein size (the proportion of inter-residue cleavages decreases with the protein size) (Rand et al. 2014).

Advanced data analysis strategies can be used to computationally increase spatial resolution of peptide-level HDX-MS data or to estimate protection factors (computational tools for such purpose are listed in **Table 2.3**).

To increase resolution, DECA (Lumpkin and Komives 2019) implements a computational method named Overlapping Peptide Segmentation (OPS). OPS exploits the overlapping of peptides to assign better-resolved uptake values to non-overlapping areas (**Figure 2.10**). When two peptides have a common terminus (e.g. peptide A covering residues 10-15 and peptide B covering residues 10-19), the absolute uptake of a smaller peptide defined by the non-overlapping residues of the observed peptides (i.e. an artificial peptide C covering residues 16-19) is calculated as the difference in absolute uptake of the bigger peptides (if peptide A has absolute uptake 3.5 Da and peptide B 5.5 Da, the uptake of peptide C is set to 5.5 – 3.5 = 2.0 Da). Because of error propagation, the error associated with the uptake of these artificial peptides is bigger than the original. For this reason, OPS should not be repeated more than once.

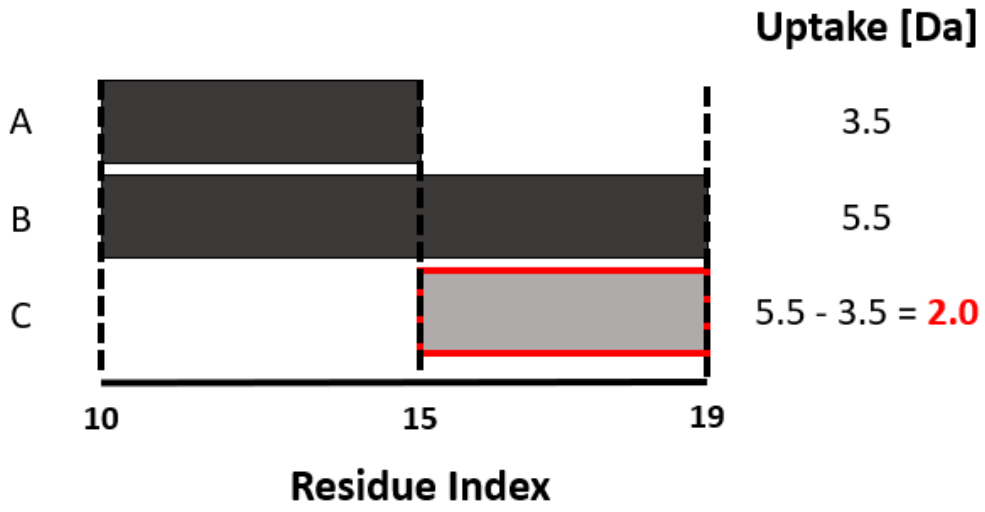


Figure 2.10. Example of Overlapping Peptide Segmentation.
Example of Overlapping Peptide Segmentation (OPS) for two peptides A and B, covering respectively residues 10-15 and 10-19 and with absolute uptake 3.5 and 5.5 Da. OPS generates an artificial peptide C covering residues 15-19 with absolute uptake 2.0 Da.

HDfleX (Seetaloo et al. 2022) and hdxstats (Crook et al. 2022) fit peptide level data with a stretched exponential (**Eq. 2.11**). The fit returns a peptide-level exchange rate that can be used to obtain a pseudo (peptide-level) protection factor. HDflex (Seetaloo et al. 2022) has the unique capability of analyzing peptide- and ETD fragment- level data simultaneously. The uptake curve of the peptide/fragment is divided by the number of exchangeable sites, so that the uptake curve of a residue is an average over the available peptides and fragments covering that specific residue. The combination of ETD data and this "data flattening" procedure thus allows an improvement in spatial resolution beyond the peptide level.

ExMS2 (Kan et al. 2019) contains a module named HDSite to extract protection factors. Here, the estimation of amide exchange rates can be performed using two different strategies. In the envelope-based method, the isotopic envelopes calculated by ExMS2 are fitted at each time point to calculate the deuteration of the residues, exploiting the overlapping of peptides. The uptake of each residue is then fitted with a single

exponential (**Eq. 2.2**) to extract the exchange rate of the single amide. Alternatively, HDSite attempts to directly fit the amide exchange rates for a whole set of peptides and exchange times. These two methods provide better results depending on the dataset.

ExPfact (Skinner et al. 2019) is a computational method aiming to extract protection factors at the resolution of the single amide and relies on the information encoded in the intensity-weighted average of the isotopic envelopes. The time-dependent uptake of each peptide is fitted simultaneously with **Eq. 2.7** and the values of the protection factors are adjusted to minimize the difference between predicted and experimental values (i.e. a cost function). Because of underdetermination, the solution is not unique (the existence of a multiplicity of solutions is known as 'degeneracy'): different sets of protection factors have the same agreement with experimental data. To attenuate the degeneracy of the solutions, the fitting algorithm is coupled to a *regularizer*, i.e. an additional term in the cost function that favors the finding of smooth patterns of protection factors (this can be interpreted as an assumption that adjacent residues do not "jump" in protection). ExPfact calculates alternative sets of protection factors, where each set is the result of a minimization procedure starting from a randomized initial guess. To further reduce the degeneracy of the solutions, a clustering algorithm (based on a mixture of multi-variate Gaussian distributions) is applied and ExPfact returns a discrete number of families of solutions. Each element of each family is a set of protection factors in agreement with experimental data. The additional information contained in the isotopic distribution can be used *a posteriori* to rank sets of protection factors (Stofella et al. 2022). HDSite and ExPfact have been cross-validated with HDX-NMR data.

HDXModeller (Salmas and Borysik 2021) implements a strategy very similar to ExPfact: a minimization procedure is repeated multiple times starting from a random initial guess; the software introduces a correlation matrix as an auto-validation tool to estimate the accuracy of the modelled protection factor of individual amino acids.

PyHDX (Smit et al. 2020) uses a machine learning framework to perform the fitting directly in a free energy landscape (the connection between the protection factor and the free energy is: $P = e^{\Delta G/RT}$, where ΔG represents the difference in free energy between the open and closed states); the problem of underdetermination is mitigated because one specific initial guess is selected and the (stochastic) fitting algorithm is coupled with a regularizer.

Also HR-HDXMS (Gessner et al. 2017) implements non-linear programming to estimate HDX exchange rates at single amino acid resolution. The degeneracy is moderated by choosing a data-oriented initial guess for the exchange rates: Overlapping Peptide Segmentation (OPS) is used to artificially increase spatial resolution and the deuteration of subpeptides is fitted with an exponential model to obtain a rate constant. This rate constant is used as initial guess for all the residues belonging to the subpeptide considered.

A Bayesian framework to estimate protection factors from HDX-MS data was first proposed by Saltzberg et al. (Saltzberg et al. 2017). The Bayesian approach estimates the probability of a particular model, given all the information about the modelled system, including prior knowledge of the system, experimental data on the system and models of experimental noise. In other words, the output of a Bayesian approach is the probability distribution of an exchange rate is calculated, not a specific value. The problem of selecting an initial guess is translated into selecting an initial probability distribution. An uninformative Jeffrey's prior (which corresponds to a uniform probability distribution) is applied to each individual exchange rate constant to represent a lack of information on the bounds and distribution of the parameter. Best scoring solutions are clustered, and mean values and standard deviations are reported.

ReX (Crook et al. 2024) is a new strategy, proposed by the same authors of Hdxstats (Crook et al. 2022), to infer residue-level rates from HDX-MS data. "ReX combines a likelihood model, which models the deuterium per residue, with a prior change-point model that permits correlations or jumps between the parameters of adjacent residues"

(Crook et al. 2024). HDX is modelled as a latent process (i.e., unobserved) occurring at the level of the single amino acid. The exchange of each residue is modelled as a mixture of a stretched exponential (**Eq. 2.11**) and a standard exponential (**Eq. 2.2**) – the proportion of the mixture is learned during the inference process. If every residue was considered as a separate entity, the model would have too many parameters to be fitted to experimental data. To overcome this issue, a change-point model (Fearnhead 2006; Green 1995) is implemented, which allows the parameters between segments of residues be either similar or discontinuous (jump). The number of change points (*where* the jump occurs) is determined via a specific Markov Chain Monte Carlo (MCMC) algorithm, known as Reverse Jump Markov Chain Monte Carlo (RJMCMC) (Green 1995), that allows the number of change points to be variable (i.e. not fixed a priori).

Table 2.3. List of software packages for high-resolution HDX-MS data analysis at the peptide level.

High resolution HDX-MS	
Name	**Strategy**
DECA (Lumpkin and Komives 2019)	Overlapping Peptide Segmentation
HDflex (Seetaloo et al. 2022)	Stretched exponential
Hdxstats (Crook et al. 2022)	Stretched exponential
ExMS2 (Kan et al. 2013, 2019)	Isotopic envelope fitting
ExPfact (Skinner et al. 2019; Stofella et al. 2022)	Intensity-weighted average fitting
pyHDX (Smit et al. 2020)	Intensity-weighted average fitting
HDXModeller (Salmas and Borysik 2021)	Intensity-weighted average fitting
Saltzberg et al. (Saltzberg et al. 2017)	Intensity-weighted average fitting
HR-HDXMS (Gessner et al. 2017)	Intensity-weighted average fitting
ReX (Crook et al. 2024)	Change-point model

Protection factor analysis requires knowledge of the contributions of each amide to the overall, observed deuterium incorporation into a peptide. In this section, we reviewed various strategies developed to increase spatial resolution from peptide-level HDX-MS data, which can be grouped into five classes (**Table 2.3**): i) overlapping peptide segmentation, ii) stretched exponential, iii) isotopic envelope fitting, iv) intensity-weighted average fitting, v) change-point model. We discourage the use of overlapping peptide segmentation as it has been shown that subtractive methods for improving

spatial resolution in HDX-MS data often yield inaccurate predictions as they neglect different levels of back-exchange for peptides of different lengths (Sheff et al. 2013). Fitting individual uptake curves with a stretched exponential can be useful to obtain a qualitative parameter describing the kinetics of a specific peptide, but this parameter is barely connected with the parameters of the Linderstrøm-Lang model (opening/closing rate or protection factor). The same limitation applies to the change-point model. We believe the most effective strategies to achieve single-residue resolution from peptide-level experimental data are the isotopic envelope fitting provided by ExMS2 and the intensity-weighted average fitting provided by ExPfact. These are the only two methods that have been cross-validated with NMR experiments, demonstrating a strong correlation between the protection factors derived from both techniques. A reference dataset analysing the HDX of a model protein with both NMR and MS would significantly aid the development of these methods. The main drawback associated to these strategies is that the results are highly dependent on the quality of the HDX-MS dataset, which is determined by the number of peptides and redundancy provided by the coverage map, as well as by the number and distribution of labelling time points. Additionally, the limited understanding of back-exchange and of EX1/EXX kinetics are holding back the development of these methods, which remain an active area of research. While they have shown promising results in inferring single residue resolution from peptide level data, a protocol to perform a 'protection factor analysis' for HDX-MS data has yet to be established. To encourage the use of the tools described here across the community, software developers should prioritize the creation of user-friendly graphical interfaces, comprehensive documentation, and tutorials.

2.6 Concluding remarks

The growing popularity of HDX-MS spurred the recent development of several data analysis tools, which are described here alongside more basic (commercial) software. We took the different steps of the data analysis workflow of HDX-MS as a guide and discussed how the pre-processing of raw data, which is generally performed with

commercial software, can be now performed with alternative open source platforms, allowing the user to better interact with the raw data. The curation of HDX-MS is however still lacking in some aspects of a complete theoretical understanding, for example in a proper correction for back-exchange.

We discussed differential experiments, where HDX-MS enables the relative and qualitative comparison of exchange patterns under different experimental conditions to pinpoint perturbations along a protein's sequence. While statistical analysis and data visualization for differential HDX-MS experiments are now well-established, there are still some nuanced aspects that warrant attention. First and foremost is the critical choice of an appropriate statistical test for comparing exchange curves across different states. We advocate for the use of statistical tests (t-tests, functional analysis, linear models, or mixed models) over manual thresholding. The rationale behind this choice is that the latter approach provides no control over false positives and false negatives. Additionally, we encourage the use of multiple testing corrections. The selection of the most appropriate statistical test is contingent upon the experimental design's architecture. For experiments encompassing both technical and biological replicates, mixed models emerge as the optimal choice. Conversely, if only one type of replication is available and no specific information about the average exchange rate is required, then the linear model represents the simplest and most pragmatic alternative. Functional analysis offers the advantage of modelling the time variable and providing quantitative insights into exchange kinetics at the cost of needing many time points to model the non-linear relation adequately. Second, it is worth noting the well-established observation that the deuteration of a peptide can be influenced by its charge state. This phenomenon, which arises from back-exchange occurring during the gas phase, remains incompletely understood, and necessitates a careful treatment of different charge states to avert spurious discoveries. When performing a differential analysis, it is important to compare the same charge state for the different experimental conditions available. When multiple charge states have been detected, it is important to check that the same results (protection/deprotection) are consistent among the

different charge states. This can relatively easily be tackled by adding an extra variable (the charge state) to the linear model implemented in **Eq. 2.9.**

We discussed how conventional differential analysis approaches should be coupled with deconvolution tools when dealing with peptides exhibiting EX1 or EXX kinetics. Several methodologies have been devised to address these scenarios, enabling the analysis of the bimodal behaviour of isotopic envelopes and the extraction of information regarding deuteration and the fractions of the two modes involved. However, it is important to acknowledge that such analysis hinges on the availability of raw mass spectra, which can be challenging to obtain. Moreover, it remains unclear how the information derived from bimodal distributions can be interpreted and integrated with protein modelling, and care should be taken when coming across such complex peptide spectra, and their precise cause established. We also discussed how HDX-MS holds promise beyond its utility in differential experiments. It affords the opportunity to delve into exchange kinetics at the single-residue level, making it an ideal candidate for validating ab-initio models or predictions of protein structure. Numerous techniques have been proposed for extracting protection factors from HDX-MS data, but a universally accepted standard for protection factor analysis has yet to be established.

What should ideal HDX-MS software look like in five years' time? Ultimately, the goal of HDX-MS software development is to enable researchers to obtain a deeper understanding of protein dynamics, functions and interactions. Therefore, the software should remove HDX-MS experimental idiosyncrasies and express the information contained in the data in the form of more physical descriptions of protein dynamics. These physical descriptors can take many forms, for example as outputs which are already established, such as protection factors or Gibbs free energies. Due to the richness of HDX-MS datasets we anticipate that future software development can give more detailed insights into hydrogen-bond networks and protein allostery, identify regions of local cooperative unfolding, or generalize functional patterns from series of protein mutations. These physical descriptions of protein dynamics could then function

as input for downstream bioinformatics methods, in the form of constraints for molecular dynamics simulations or as training data for predictive artificial intelligence (AI), taking deep learning approaches such as Alphafold (Jumper et al. 2021) beyond static predictions of protein structure and instead offer functional information based on protein dynamics. For example, predictive AI models could learn from HDX-MS data how to identify allosteric regulation in *de novo* designed proteins.

In general, while it is important that software serves the direct needs of the HDX-MS community itself, in the form of statistical testing and dataset quality validation, we envision that future software development will facilitate dissemination of novel insights towards broader audiences and allow for increased interfacing with neighbouring fields.

To work towards these goals, the software should perform the following basic steps. The software would accept the protein sequence and undeuterated raw data as input, performing robust peptide search and identification to generate a coverage map. This search would support not only peptide-level data but also fragment-level data from CID, ExD, or UVPD fragmentation. In the next step, the software would process the deuterated raw data, automatically detecting the isotopic envelopes of previously identified peptides. Since a large body of user-annotated peptides datasets are readily available, we anticipate that AI models can be trained on these data and provide further automation and validation in this critical step, increasing both throughput and accuracy. The identified peptides could then be exported in a single operation as isotopic envelopes in a standardized format. It would also perform accurate back-exchange correction (or the best available correction based on future research), showing users how the correction modifies raw input data and provide feedback on confidence and potential experimental artifacts. The software would deconvolute peptide spectra exhibiting multimodal behaviour, enabling researchers to export results for further analysis of EX1/EXX kinetics.

In "Differential Analysis" mode, researchers could select the most appropriate statistical test for their experimental design and research question. The software would then generate publication-quality Woods plots, Manhattan plots, and volcano plots to highlight statistically significant changes across the protein sequence. Protein structural information could be uploaded to the software, either obtained from experimental methods or Alphafold predictions. The software could feature one or multiple modelling options or fitting strategies, such as “Protection Factor Analysis”, where users are guided through steps and various modelling parameters, and the software would evaluate the dataset’s quality and providing a confidence level for the final predictions. The estimated pattern of protection factors or other modelling output could then be mapped onto the uploaded protein structure and presented as an integrative structural and functional output. There should be a strong focus on accessibility, providing comprehensive documentation and a user-friendly graphical user interface. Data processing best-practice and the effect of user-configurable settings and tuneable parameters such as thresholds and how they influence output and confidence should be clearly explained through tutorials or other forms of documentation. Publication of source code under a permissive license is required for other researchers to validate and review the processing pipelines as well collaborate and iterate on published works.

Experimental researchers would focus on the experiment, and the software would provide real-time results and suggestions to guide their decisions, while computational researchers would be able to download online datasets from a standardized repository, rapidly perform the same analysis performed in published papers, and easily and exhaustively export all the information they need to improve the implemented methods or to propose alternative solutions to tackle the remaining challenges.

Chapter 3. High Resolution Hydrogen-Deuterium Protection Factors From Sparse MS Data Validated By NMR Measurements

Published as: M Stofella, SP Skinner, F Sobott, J Houwing-Duistermaat and E Paci, "High-Resolution Hydrogen-Deuterium Protection Factors from Sparse Mass Spectrometry Data Validated by Nuclear Magnetic Resonance Measurements", J Am Soc Mass Spectrom, 2022, 33, 5, 813-822, doi: https://doi.org/10.1021/jasms.2c00005

3.1 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 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 unit-less 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) 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.

3.2 Introduction

Hydrogen-deuterium exchange (HDX) is the spontaneous exchange of covalently bonded hydrogens of a protein with deuterium in solution (Englander et al. 2016). In his pioneering work, Lindestrøm-Lang probed the phenomenon through density gradient tubes (Linderstrøm-Lang 1955). Since then, nuclear magnetic resonance (NMR) has been the leading technique to probe HDX until the early 2000s (Dempsey 2001), when mass spectrometry (MS) has emerged as an alternative with many advantages (no sample size limitations, no labelling required, low protein concentration, low costs, highly automated processing), counting an increasing number of applications in fundamental biophysics and applied biotechnology (Deng et al. 2016; James et al. 2021; Masson et al. 2019). 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 properties of single amino acids (Hamuro 2021b).

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 hetero-nuclear multidimensional NMR spectra (Dempsey 2001). MS measures directly the mass variation as a function of exchange time of peptides obtained by proteolytic digestion (Masson et al. 2019). 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 (Hamuro 2021a). Most current approaches use collision-induced dissociation (CID) for MS/MS fragmentation of peptides, but due to 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 of alternative MS/MS methods which minimize H/D scrambling during fragmentation (Mistarz et al. 2018; Pan et al. 2009; Pan, Han, and Borchers 2012).

Several computational strategies have been proposed to extract single-residue protection factors from peptide-level HDX-MS data (Babić, Kazazić, and Smith 2019; Gessner et al. 2017; Kan et al. 2013; Salmas and Borysik 2021; Saltzberg et al. 2017; Smit et al. 2020; Zhang 2020; Zhang et al. 2012). Here we demonstrate the advantages of a method recently proposed (Skinner et al. 2019) 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 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 structure and dynamics of proteins. Since the seminal work from Linderstrøm-Lang (Linderstrøm-Lang 1955), HDX has been modelled 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:

Eq. 3.1
$$NH_{cl} \underset{k_o}{\overset{k_c}{\rightleftharpoons}} NH_{op} \xrightarrow[D_20]{k_{int}} ND$$

Here k_o and k_c are referred as *opening* and *closing* rates, respectively, which let the residue switch from an exchange-incompetent state (i.e., in a closed or folded state NH_{cl}) to an exchange-competent state (i.e., in 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 pH, temperature of the solution and side chains

of the two adjacent amino acids (Bai et al. 1993; Connelly et al. 1993; Englander et al. 1996; Molday et al. 1972; Nguyen, Mayne, Phillips, and Englander 2018).

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

Eq. 3.2 $$k_{obs} = \frac{k_{int} k_o}{k_{int} + k_c}$$

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

Eq. 3.3 $$d(t, P) = 1 - e^{-\frac{k_{int}}{P} t}$$

where the opening equilibrium constant $P \equiv {k_c}/{k_o}$, known as protection factor, is linked to dynamic properties of the residue by definition; moreover, several studies have shown a correlation between the protection factors of a protein and its structure (Best and Vendruscolo 2006; Vendruscolo et al. 2003). 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 (Ferraro et al. 2004).

In HDX-NMR experiments, the proton signal decays exponentially as deuteration occurs because deuterium is ^{1}H-NMR silent, and the experimental curves can be fitted with a **Eq. 3.3** to obtain P (Barnes et al. 2019; Fitzkee, Torchia, and Bax 2011).

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

Eq. 3.4 $$D(t, \{P_i\}) = \frac{1}{N} \sum_{i=1}^{N} \left(1 - e^{-\frac{k_{int,i}}{P_i} t} \right)$$

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 (Skinner et al. 2019).

The possibility of estimating individual protection factors from HDX-MS data depends on four factors (Hamuro 2021a): (i) peptide overlap, (ii) time point resolution, (iii) time window coverage, (iv) experimental error. (i) The protection factor of an individual amino acid can be in principle 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 number of overlapping peptides (Kan et al. 2013; Skinner et al. 2019). 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 3.1A**); for two overlapping peptides formed by N1 and N2 residues, respectively, and with Nc residues in common, there are $(N1 - Nc)! \cdot (N2 - Nc)! \cdot Nc!$ alternative solutions (**Supplementary Figure 3.1B**). 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 3.1C**). While the observed rates span several orders of magnitude depending on experimental conditions (pH, temperature), 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. 3.4**) 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. 3.4** 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 (Bishop 2006): $N_{exp} \approx QN_{res}$, where $N_{exp} = N_{times}$ for an isolated peptide. (iii) To properly sample the multi-exponential uptake of a peptide (**Eq. 3.4**), 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. sub-second) is now possible, with recent developments giving access to millisecond timescales (Kish et al. 2019; Rob et al. 2012; Svejdal et al. 2019). 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, $N_{exp} = N_{times} \times N_{pep}$. Experimentally, the overlap of peptides depends on the choice of the protease, which is limited due to 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 the mean of the measurements, i.e. 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 datasets. HDSite (Kan et al. 2013, 2019) 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. 3.3**) 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 dataset and prediction is limited by the number of peptides available and their overlap. Analogously to HDSite, there are other methods aiming to extract single residue information from HDX-MS data by fitting the isotopic envelopes of peptides (Babić et al. 2019; Zhang 2020; Zhang et al. 2012). A method more recently proposed (ExPfact) (Skinner et al. 2019) simultaneously fits the uptake curves of contiguous overlapping peptides with multi-exponential curves (**Eq. 3.4**), determining all alternative patterns of protection factors compatible with experimental data. This method can be applied to any dataset, 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 (Smit et al. 2020), HDXModeller (Salmas and Borysik 2021), HR-HDXMS (Gessner et al. 2017) and HDX Workbench (Saltzberg et al. 2017).

In this paper, we analyse a dataset previously published (Moulick, Das, and Udgaonkar 2015), 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 (Skinner et al. 2019), 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 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.

3.3 Methods and Materials

3.3.1 Dataset

The measurements analysed here were previously published (Moulick et al. 2015) and probed the 103 residues mouse prion protein at pH 4 and temperature 25°C at different urea concentrations (see **section 3.6.1** for experimental details). To ensure the validity of the EX2 approximation (and thus of **Eq. 3.3** and **Eq. 3.4**), 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 temperature 0°C and the protein was digested by pepsin, providing a dataset (**Figure 3.1**) which 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 five seconds to 24 hours, and the experiment was conducted in triplicate.

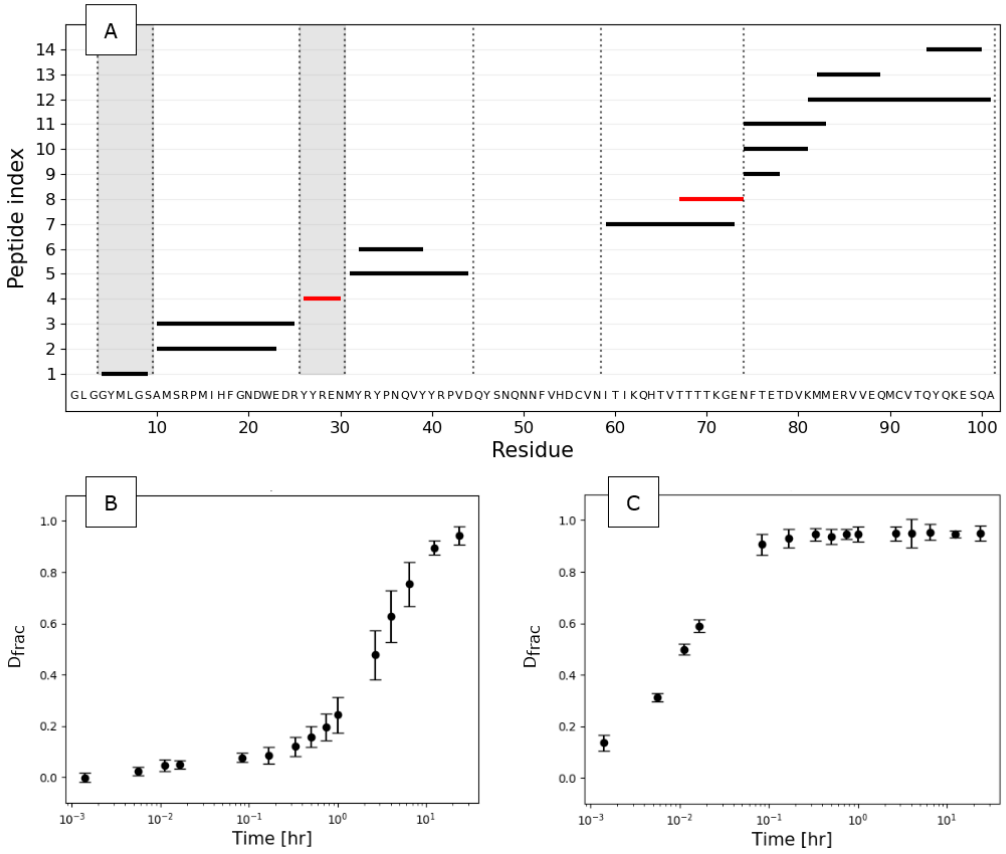


Figure 3.1. The HDX-MS dataset previously published in (Moulick et al. 2015).

A) The coverage map localizes the 14 peptides identified after pepsin digestion. The six regions covered by isolated (grey) 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 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 $D_{0\%}$ sample:

Eq. 3.5 $$D = \frac{D_t - D_{0\%}}{D_{FD} - D_{0\%}}$$

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 datasets. Since NMR experiments were performed only once, we assumed that the protection factors provided by NMR represent their true values.

3.3.2 Prediction of protection factors

ExPfact is a computational method aiming to extract protection factors at the resolution of the single amide (Skinner et al. 2019). Considering regions covered by contiguous overlapping peptides one by one, the method finds the 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. 3.3**), 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. 3.3** (D_j^{pred}) and the set of protection factors $\{P_i\}$ was adjusted to minimize the cost function

Eq. 3.6
$$C(\lambda, \{P_i\}) = \underbrace{\sum_j \sum_k w_{jk} \left[D_j^{pred}(t_k, \{P_i\}) - D_j^{exp}(t_k) \right]^2}_{SSR} + \underbrace{\lambda \sum_i (\ln (P)_{i-1} - 2\ln (P)_i + \ln (P)_{i+1})^2}_{\text{Penalty term}}$$

The cost function in **Eq. 3.6** consists in a regular term, the sum of squared residuals (SSR), which depends on the experimental data, and on a penalty term, which was introduced to avoid overfitting and, given the correlation between exchange rates and structure of the protein (Best and Vendruscolo 2006; Vendruscolo et al. 2003), 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 3.2**). Following the recommendations for the propagation of error in HDX-MS data (Weis 2021), a pooled standard deviation can be associated to 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 3 – then it is more accurate to consider the weights as the inverse of the standard deviation. The cost function in **Eq. 3.6** 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 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 5,000 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 (Scrucca et al. 2016). The histograms of the predicted protection factors, which are often multimodal (**Figure 3.2**), are combined into an 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 3.3**). 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.

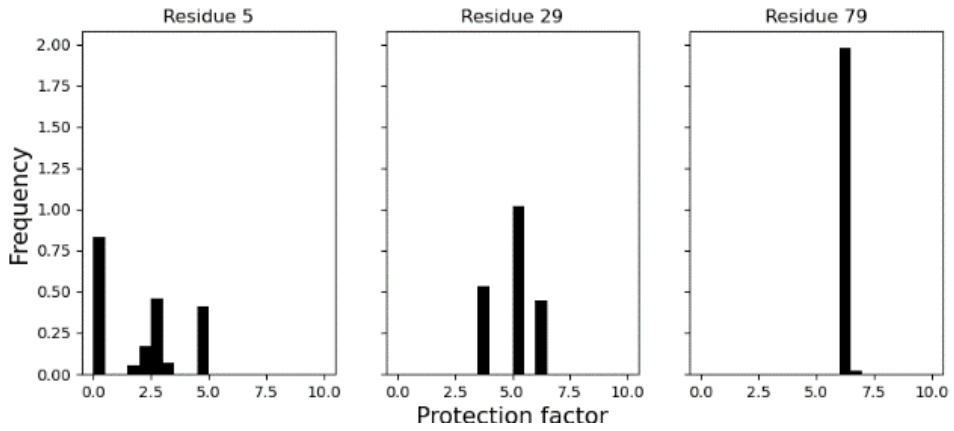


Figure 3.2. Histograms of protection factors predicted for selected residues.
Histograms of protection factors predicted for residues 5 (left), 29 (centre) and 79 (right) from 5,000 minimizations. In most cases, histograms are multimodal distributions: 3 modes can be identified for residue 5; 4 modes for residue 29.

3.3.3 Performances

One minimization procedure requires on average 12.7 seconds on the dataset here analysed (processor: Intel® Xeon® W-1290P 3.7GHz) 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 4 cores is sufficient to complete 5,000 minimizations in less than 5 hours. We recommend running ExPfact overnight, setting up the number of minimizations and the tolerance parameter according to the computational power available.

3.3.4 Prediction of isotopic envelopes

For a peptide, the fractional deuterium uptake at time t (**Figure 3.1B-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.

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:

Eq. 3.7 $$\Pi(k,t) = \sum_{\substack{A\subset\{1,\dots,n\}\\ |A|=k}} \prod_{i\in A} d_i(t) \prod_{j\in\{1,\dots,n\}\setminus A} \left(1 - d_j(t)\right)$$

Eq. 3.7 can be build following these considerations: (i) the probability of a residue to exchange is a function of time and is given by **Eq. 3.3**; (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 a 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. 3.7** to the fully protonated envelope of the peptide (calculated using the python library pyOpenMS (Röst et al. 2014)). 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 $\pi_i\Pi(k=0,t) + \pi_{i-1}\Pi(k=1,t) + \cdots + \pi_{i-N}\Pi(k=N,t) = \sum_{j=0}^{N} \pi_{i-j}\Pi(j,t)$, 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. 3.7** was applied until deuteration time t, i.e. towards higher m/z values, using the intrinsic exchange rates calculated at temperature 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 one because the deuteration in **Eq. 3.7** does not account for back-exchange. Back-exchange occurs at the protein level in the labelling 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. 3.7** towards protonation, i.e. towards lower m/z values, using the same set of $\{P_i\}$ and the intrinsic exchange rates calculated at temperature 0°C and pH 2.4 for a deuterated protein in a protonated buffer. This back-exchange correction was applied for an "effective back-exchange time" τ minimizing the difference between the predicted and the experimental shape. The underlying assumption is that back-exchange can be modelled analogously to in-exchange (i.e., using the multi-exponential in **Eq. 3.4**). 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 3.3**.

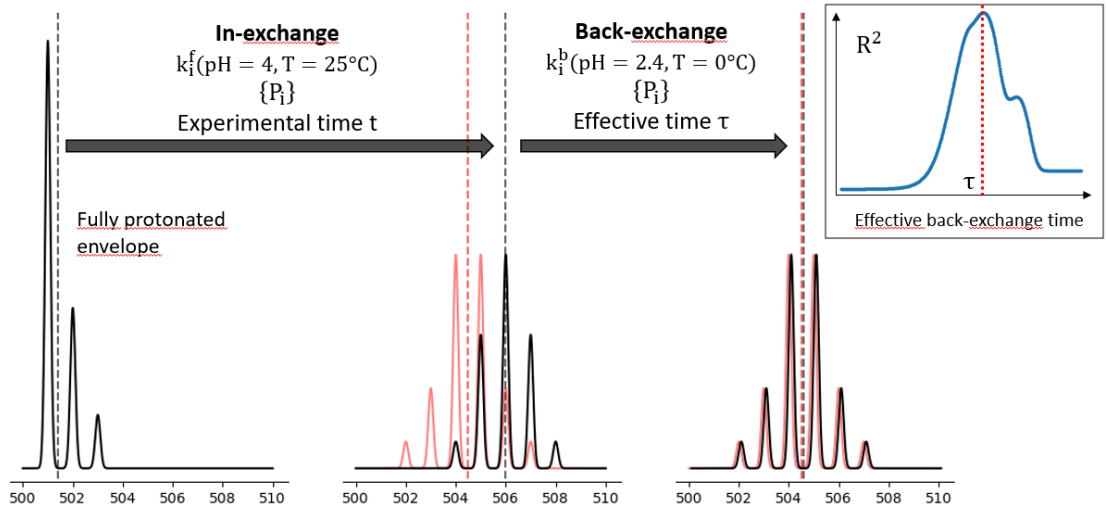


Figure 3.3. Schematic representation of the calculations for the reproduction of the experimental isotopic envelope.

The fully protonated envelope can be calculated from the knowledge of the peptide sequence. The isotopic envelope at time t is evaluated applying the evolution in **Eq. 3.7** to the fully protonated envelope, using a specific pattern of protection factors $\{\boldsymbol{P_i}\}$ and the intrinsic "forward" exchange rates $\boldsymbol{k_i^f}$ calculated at pH 4 and temperature 25°C for a protonated protein in a deuterated buffer. The in-exchange predicts

an envelope (black) which 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. 3.7** is applied towards protonation, using the same $\{\boldsymbol{P_i}\}$ and the intrinsic "back" exchange rates $\boldsymbol{k_i^b}$ 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).

3.4 Results

The protection factors derived from HDX-MS measurements probing the mouse prion protein in its native state are shown in **Figure 3.4A** with their associated error. The value(s) and the error(s) associated to 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), while multiple clusters are found in the two regions (residues 5-9 and 27-30) in which proteolytic peptides do not overlap (**Figure 3.4B-C**). In both regions, the NMR protection factors fall within 1σ (one standard deviation) of the MS estimation. The predicted profiles of protection factors reflect known structural properties of the protein. Indeed, higher protection against exchange is observed at helices α1 (residues 21-30) and α3 (residues 77-101), with completely unprotected residues surrounding Cys91, which forms a disulphide bond with Cys56. Lower protection is also observed in the loop between α2 and α3 (residues 72-76).

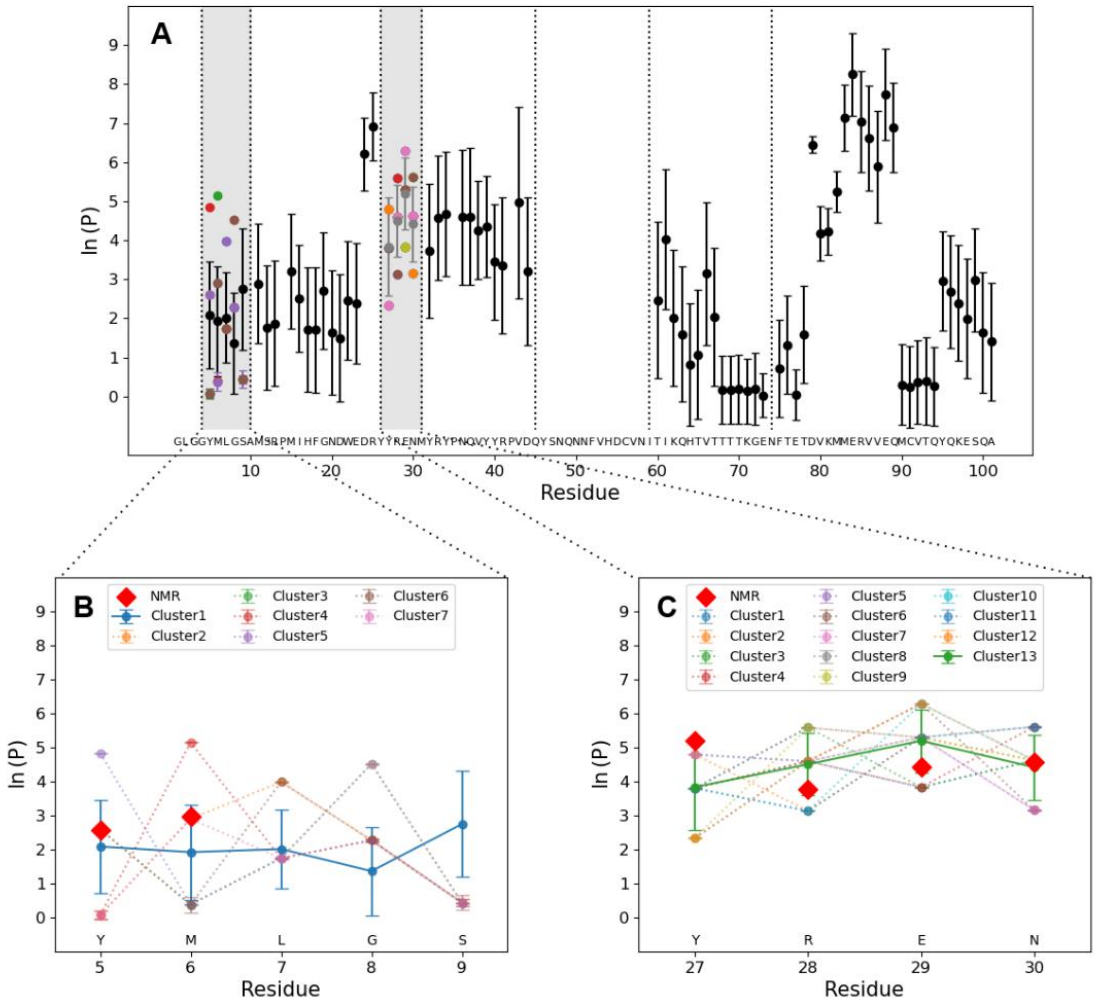


Figure 3.4. Estimated protection factors of the mouse prion protein.

A) Protection factors corresponding to cluster means of 5,000 least square solutions obtained by minimizing the cost function in **Eq. 3.6**. Vertical dashed lines show regions covered by isolated (grey) or overlapping peptides (compare with **Figure 3.1**). 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 colours. 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.

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 3.1**), where the clustering algorithm identified seven clusters (**Figure 3.4**). We randomly select a set of $\{P_i\}$ from each cluster and predict the isotopic envelope as discussed in the Methods section. The outcomes (**Figure 3.5**) 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.

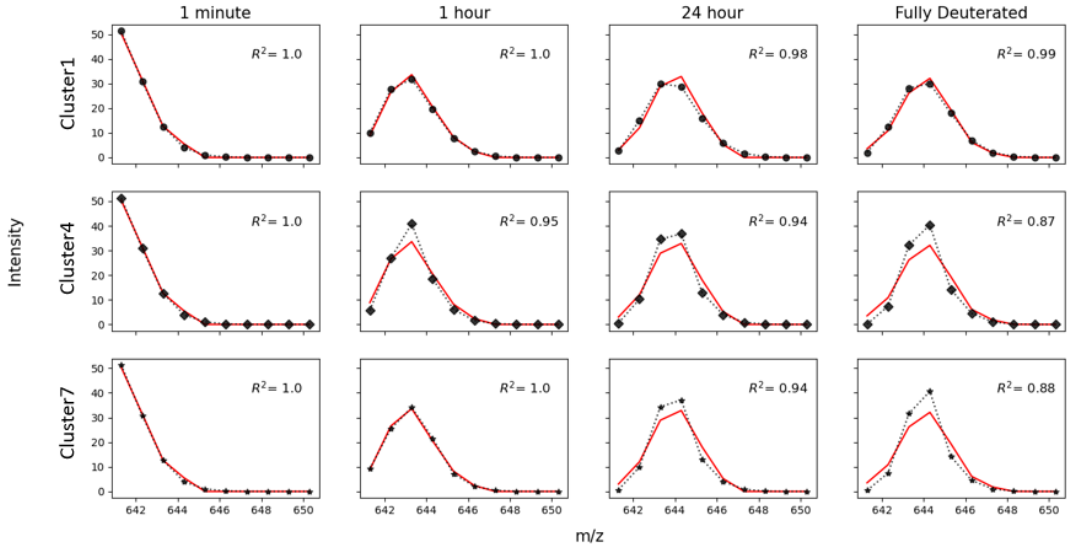


Figure 3.5. Prediction of isotopic envelopes.

Starting from the fully protonated envelope of peptide 1 (sequence YMLGSA), the evolution in **Eq. 3.7** is applied towards deuteration at times 1 minute (column 1), 1 hour (column 2), 24 hours (column 3) and infinite time (column 4) using intrinsic exchange rates calculated at pH 4 and temperature 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 applying **Eq. 3.7** towards protonation, using intrinsic exchange rates calculated at pH 2.4 and temperature 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 with the experimental envelopes (red lines). The agreement is evaluated using the $\boldsymbol{R^2}$.

Protection factors for 27 residues covered by the MS dataset 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 3.4**). 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 3.5**). The region covered by peptide 4 (**Figure 3.1**) remains underdetermined (**Figure 3.4C**) because the experimental isotopic envelopes for this peptide were not available. In this region, we selected cluster 13 as final pattern of protection factors because it showed compatibility with NMR measurements. The comparison between protection factors extracted by MS and NMR (**Figure 3.6A-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 3.6D**) when the outlier residues 25, 91 and 94 (which are not compatible within 3σ) are not considered.

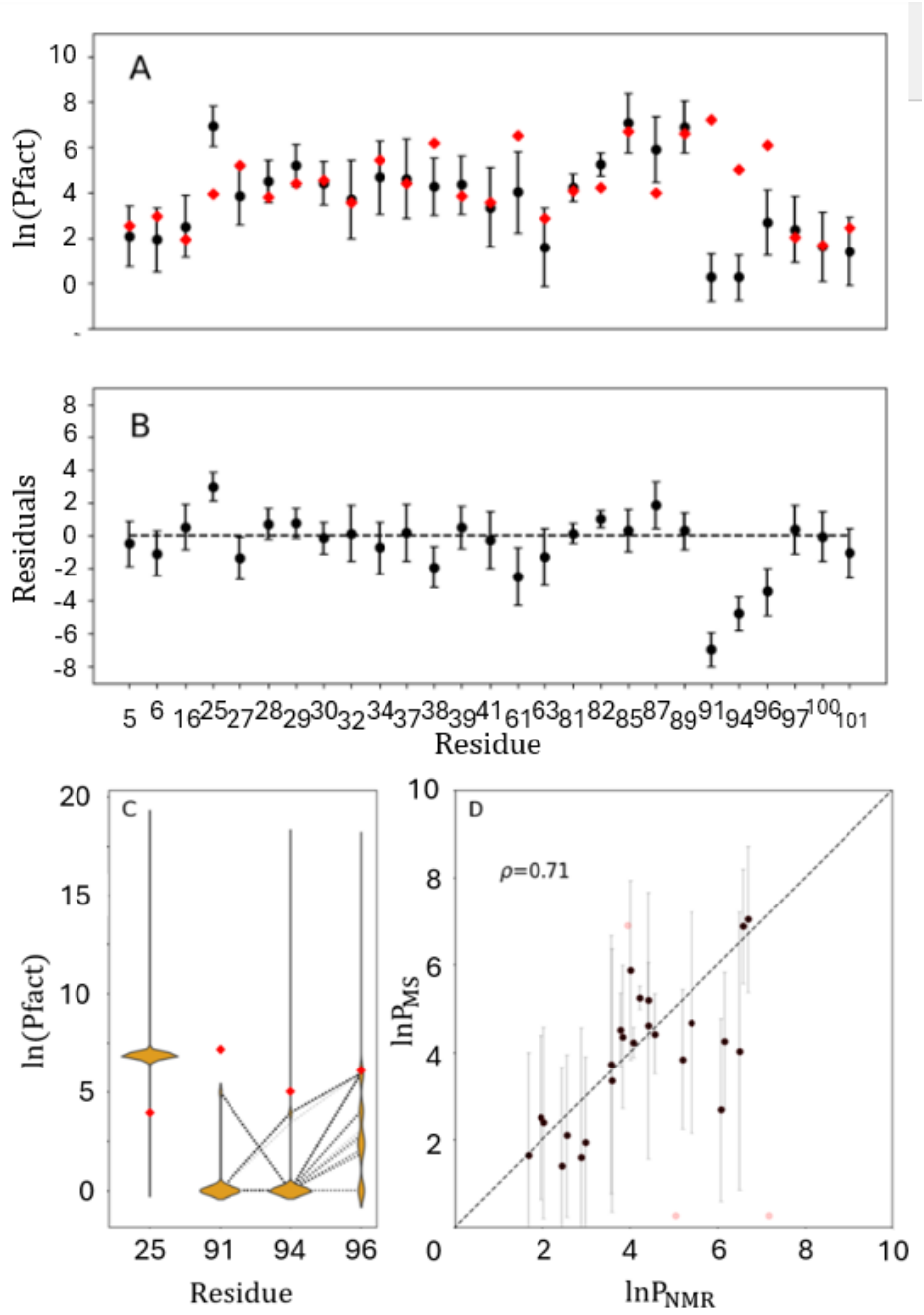


Figure 3.6. Comparison of protection factors from HDX-MS and HDX-NMR experiments.
A) Clusters of protection factors extracted from HDX-MS data (black dots with error bars) are compared with NMR measurements (red diamonds) for every amino acid covered by both datasets. B) Residuals of protection factors from HDX-MS and HDX-NMR experiments. C) Marginal probability distribution of protection factors derived from 5,000 minimization procedures for residues 25, 91, 94 and 96. D)

Correlation between protection factors extracted by NMR and MS; Pearson's correlation coefficient $\boldsymbol{\rho} = \mathbf{0.71}$ excluding outliers (red).

3.4 Discussion

Despite the only partial coverage provided by the MS dataset (**Figure 3.1**), we showed how alternative patterns of protection factors with similar agreement with experimental data can be accurately derived at the resolution of single amino acid (**Figure 3.4A**). 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 an ambiguity on the final estimate of the protection factors, but at least one of the clusters identified in these regions is compatible with protection factors derived from NMR measurements (**Figure 3.4B-C**). Nonetheless, the ambiguity could be completely removed for one of these two regions exploiting the supplementary information contained in the shape of the experimental isotopic envelope (**Figure 3.5**). 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 3.6**), validating the method. The four discrepancies shown by residues 25, 91, 94 and 96 provide insight into the limitations of the datasets 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 NMR measures (**Figure 3.6C**). The GMM clustering algorithm selects the final number of components based on the minimum $\mathrm{BIC} = \mathrm{k}\ln(\mathrm{n}) - 2\ln(\hat{\mathrm{L}})$, 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 favour 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 artefact is even more evident when we look at protection factors of residue 94, which has a multimodal probability distribution with 4 modes; moreover, one of the modes is compatible with the NMR measure. 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, **Supplementary Figure 3.3**) would find the low-intensity peaks approaching NMR measures shown in **Figure 3.6C** 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 neighbours. 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 3.6C** 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 measures did not fit the uptake curves of the HDX-MS dataset. To prove this, we constrained protection factors in the region 75-101 to their NMR value (when available) during the least-squares minimization, while the remaining protection factors are adjusted to minimize the cost function in **Eq. 3.6**. 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 3.7**). The analysis of these discrepancies 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 dataset) or too large (in the MS dataset). The disagreement is however localized in a specific region of the protein and could be therefore caused by artifacts either in the NMR or MS experiment. In the absence of additional measurements, these results cannot be interpreted further.

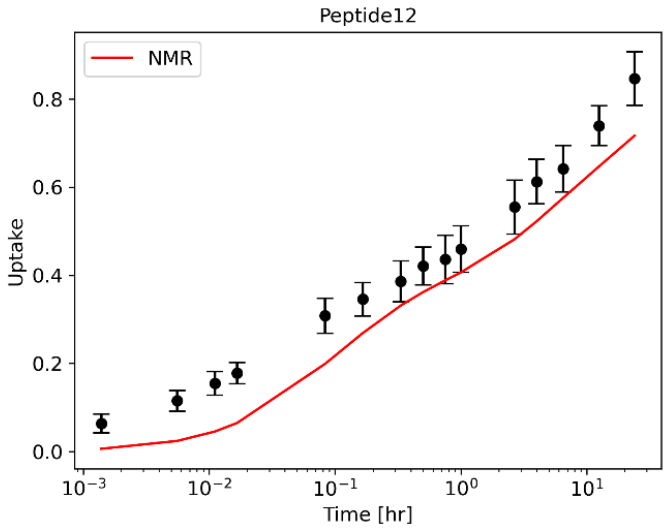


Figure 3.7. Deuterium uptake prediction for peptide 12 using an optimized set of protection factors with constrained NMR values.

In the region covering residues 76-101, the protection factor of 11 residues was measured by NMR. These values are fixed, while the remaining protection factors are optimized to minimize the cost function in **Eq. 3.6**. The resulting prediction (red line) is not compatible with MS data.

3.5 Conclusions

In this paper, we applied ExPfact (Skinner et al. 2019) to a previously published dataset probing the HDX of the same protein under the same experimental conditions by both MS and NMR (Moulick et al. 2015). The novelties introduced with respect to the previous publications are (i) the validation of the method via a comparison with NMR data, which is often neglected in related papers (Babić et al. 2019; Gessner et al. 2017; Kan et al. 2013; Salmas and Borysik 2021; Saltzberg et al. 2017; Smit et al. 2020; Zhang 2020; Zhang et al. 2012); (ii) the prediction of the experimental isotopic envelope of peptides (via the back-exchange correction) as a further 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, extended documentation).

The approach demonstrated here enables the quantitative analysis of any HDX-MS dataset (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) emerges from the isotopic distribution of peptides, single residue information can be extracted via other methods (Na et al. 2019; Zhang 2020). The information extracted in the two regimes is different. For the exchange in EX2 conditions, a protection factor can be exctracted: this is a unit-less 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) (Englander et al. 2016). 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 $[time]^{-1}$ and can be expressed through the Eyring equation (Eyring 1935) as proportional to Gibb's free energy of activation: $k_o = (k_B T/h)\exp\left(-\frac{\Delta G_o^{\ddagger}}{RT}\right)$ (where k_B is the Boltzmann constant and h is the Planck's constant) (Hamuro 2021b). ExPfact aims to extract protection factors from HDX-MS because they encode structural information of the protein, and is consequently limited to the study of datasets with peptides showing EX2 behaviour.

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 photo-dissociation (UVPD) (Mistarz et al. 2018; Pan et al. 2009, 2012) – 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 (Kish et al. 2019; Rob et al. 2012; Svejdal et al. 2019) combined with a unified computational approach encompassing the knowledge acquired in the last decade (Kan et al. 2013; Salmas and Borysik 2021; Skinner et al. 2019; Smit et al. 2020) 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 (Jumper et al. 2021).

3.6 Supplementary Material

3.6.1 Experimental details

This section aims to provide experimental procedures adopted by the authors who first published the dataset. The text is adapted with permission from R. Moulick, R. Das and J. B. Udgaonkar (2015) "Partially unfolded forms of the prion protein populated under misfolding-promoting conditions", Journal of Biological Chemistry.

Protein expression and purification. The full-length recombinant mouse prion protein, moPrP(23-231), encoded in the pET-17b(+) plasmid was expressed in *Escherichia coli* BL21(DE3) CodonPlus (Stratagene) cells and purified as described previously (Jain and Udgaonkar 2008). The protein was lyophilized and stored at −20 °C. The concentration of the protein was determined by absorbance measurements at 280 nm using an extinction coefficient of 62,160 M^{-1} cm^{-1}. The recombinant moPrP(23–231) lacks the first 22-residue signal sequence that is cleaved off in the formation of mature protein *in vivo*.

Chemicals and buffers. All the experiments utilized buffers containing 20 mM sodium acetate and variable concentrations of urea (obtained from USB Corp.) in the range 0–2 M. Urea was deuterated by dissolving it in D_2O, flash freezing the solution, and lyophilizing it. This cycle was repeated three times to ensure complete deuteration of the urea. The deuteration or exchange buffer consisted of 20 mM sodium acetate dissolved in D_2O adjusted to pD 4 ($pD = pD_{read} + 0.4$) using DCl. All solutions were filtered using 0.22-μm Millipore syringe filters before use. The concentrations of urea stock solutions were determined prior to use by refractive index measurements using an Abbe refractometer. All chemicals used were obtained from Sigma (unless mentioned otherwise).

Peptide Map of moPrP. Lyophilized protein was dissolved in Milli-Q water at pH 2.5 to a final concentration of 3.5 μM. This sample was injected into the HDX module (Waters) where the sample was digested by pepsin (at a 50 μl/min flow rate), and the

peptic fragments were separated on a C_{18} reverse phase chromatography column using a gradient of 3–40% acetonitrile before being fed into a Waters Synapt G2 HD mass spectrometer. All columns were kept at 4 °C, and the mobile phases were chilled on ice. The fragments were identified using Protein Global Lynx software. The fragments obtained corresponded to all structured parts of the protein including the loops connecting the secondary structures. However, no peptide could be obtained for a stretch of 14 residues (168–181) corresponding to the N-terminal segment of α2.

Deuteration of moPrP. Lyophilized protein was dissolved in deuteration buffer (20 mM sodium acetate dissolved in D_2O at pD 4) to a final concentration of ~20 μM. This stock solution was heated to 65 °C for 10 min to unfold and deuterate the protein, immediately kept on ice for 15 min for refolding, and then kept at room temperature. The stock solution was concentrated using a 10-kDa Centricon filter unit from Millipore to 180 μM. The mass of the deuterated protein was checked using a Synapt G2 HD mass spectrometer, and the protein was found to be completely deuterated.

HDX-MS of moPrP. Lyophilized protein was dissolved in 20 mM sodium acetate, pH 4 to a final concentration of 20 μM. To initiate exchange, 25 μl of protonated protein were mixed with 475 μl of deuteration buffer in the presence of 0–2 M urea for varying lengths of time, at 25 °C. The reaction was quenched by mixing with ice-cold 500 mM glycine, pH 2.4, and the solution was desalted into ice-cold water, pH 2.5 using a Sephadex G-25 HiTrap desalting column in conjunction with an ÄKTA Basic HPLC. The desalted samples were injected into the HDX module (Waters) coupled to a nanoACQUITY UPLC. A gradient of 3–40% acetonitrile (0.1% HCOOH) at a flow rate of 40 μl/min was used for elution of the protein from an analytical C_{18} reverse phase chromatography column in 10 min. The extent of exchange was determined by measuring the increase in the mass of the protein in the Synapt G2 HD mass spectrometer.

For experiments in which HDX was followed by pepsin digestion, the protein was dissolved to a final concentration of 180 μM in 20 mM sodium acetate, pH 4. For online

pepsin digestion, a flow rate of 50 µl/min of water (0.05% HCOOH) was used. The peptides eluted were collected using a peptide trap column, washed to remove salt, and eluted as above. All columns were kept at 4 °C in the cold chamber of the HDX module. The peptides separated on the column were detected using the Synapt G2 HD mass spectrometer. The mass spectrometer parameters were set as follows: source temperature, 35 °C; desolvation temperature, 100 °C; capillary voltage, 2.8 kV. To determine the number of exchanged deuteriums that are lost due to back-exchange during sample processing after quenching of the exchange reaction, 25 µl of completely deuterated protein were diluted in 475 µl of deuteration buffer, the reaction was quenched as above, and the sample was then processed in an identical way.

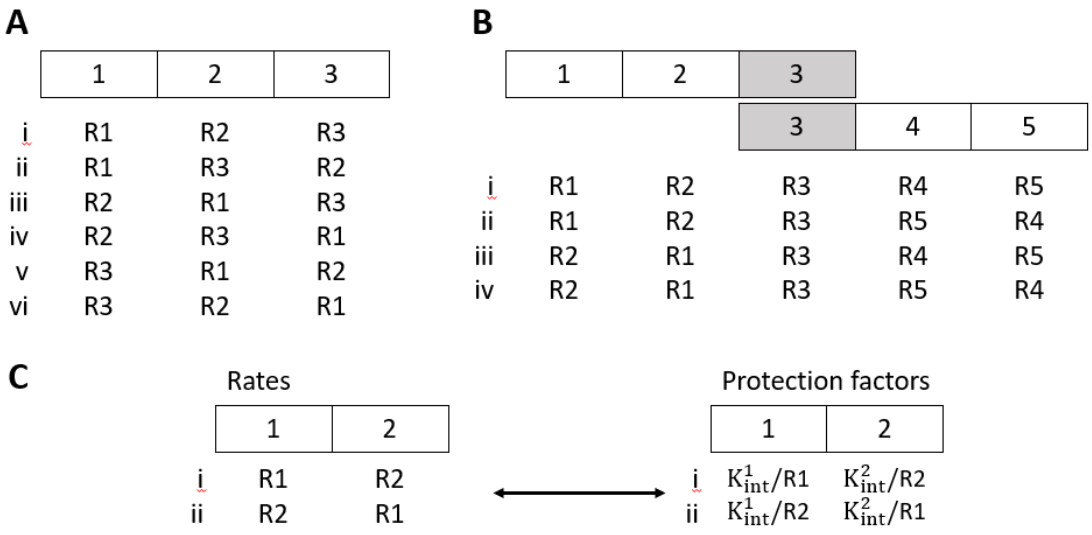
Sequential backbone assignment of the NMR spectrum of moPrP. Two-dimensional ^{15}N heteronuclear single quantum coherence, HNCO, HNCA, HNCACO, HNCOCACB, and HNCACB resonance experiments were carried out on 300 µM moPrP in 20 mM sodium acetate at pH 4 (in the presence of 5% D_2O) for the backbone assignment. Data were processed using NMRPipe, and the assignment was done using the NMR data visualization and assignment software Sparky. The NMR spectrum of the protein collected at pH 4 in this study is indistinguishable from that collected at pH 4.5.

HDX-NMR. A ^{1}H-^{15}N two-dimensional selective optimized flip angle short transient heteronuclear multiple quantum coherence spectrum was collected as the reference unexchanged sample spectrum by dissolving lyophilized protein in 20 mM sodium acetate, pH 4, 95% H_2O, 5% D_2O buffer to a final concentration of 300 µM. To monitor exchange, lyophilized protein was dissolved in 20 mM sodium acetate (95% D_2O), pD 4 to a final concentration of 300 µM (exchange sample). A series of ^{1}H-^{15}N two-dimensional selective optimized flip angle short transient heteronuclear multiple quantum coherence spectra of the exchange sample were collected up to 40 days following an initial dead time of 10 min. The two-dimensional spectra were collected on a Bruker 800-MHz spectrometer with 1024 × 256 ($t_1 \times t_2$) time points totaling up to an acquisition time of 5 min and processed using NMRPipe and Sparky. The NMR

spectrum of the unexchanged sample was corrected for intensity differences arising due to differences in matching, tuning, and shimming by multiplying with the ratio of the intensity of the Cys^{213} amide (an isolated peak that exchanges on a 1-month time scale) in the unexchanged spectrum to that of the Cys^{213} amide at 10 min of exchange. This was the intensity-corrected unexchanged spectrum. The signal intensity for each residue in all exchange spectra was normalized with the signal intensity of that residue in the intensity-corrected unexchanged spectrum. The resulting decrease in the normalized signal intensity with increasing time of exchange was converted into a percent hydrogen occupancy *versus t* plot and fit to an exponential decay equation to yield an observed exchange rate. This rate was used to calculate the free energy of opening of structure to exchange using the equation $\Delta G_{op} = RT\ln(k_{int}/k_{obs})$. The intrinsic rates of exchange for individual residues were obtained as described elsewhere (Bai et al. 1993).

Supplementary Figure 3.1. When measurements are not affected by experimental error, estimating protection factors is a combinatorial problem.

A) For an isolated peptide of N residues, the uptake curve is a sum of N exponentials and there are N! possible patterns of protection factors. For N=3, a best fit always extract three rates R1, R2 and R3, however these rates cannot be assigned to specific residues, therefore all the N!=6 assignments (i) to (vi) are possible. B) For two overlapping peptides formed by N1=N2=3 residues with Nc=1 common residue as shown, it is possible to unambiguously assign the rate of residue 3, say R3. However, it is not possible to assign rates R1 and R2 to residues 1 and 2 and, analogously, rates R4 and R5 to residues 4 and 5. There are (N1-Nc)!(N2-Nc)!Nc!=4 possible patterns of protection factors, namely options (i) to (iv). C) Assignment of rates does not yield to the same result in terms of protection factors. For an isolated peptide of N=2 residues, a best fit always provides rates R1 and R2, e.g. $\boldsymbol{R1 = 1s^{-1}}$ and $\boldsymbol{R1 = 10s^{-1}}$. In terms of rates, the N!=2 alternative solutions are $\{\boldsymbol{R1}, \boldsymbol{R2}\} = \{\boldsymbol{1}, \boldsymbol{10}\}$ and $\{\boldsymbol{R2}, \boldsymbol{R1}\} = \{\boldsymbol{10}, \boldsymbol{1}\}$. If the amino acids corresponding to residues 1 and 2 are different, they have different intrinsic exchange rates, e.g. $\boldsymbol{k_{int}^1 = 10s^{-1}}$ and $\boldsymbol{k_{int}^2 = 100s^{-1}}$. In terms of protection factors, the 2 alternative solutions are $\left\{\frac{\boldsymbol{10}}{\boldsymbol{1}}, \frac{\boldsymbol{100}}{\boldsymbol{10}}\right\} = \{\boldsymbol{10}, \boldsymbol{10}\}$ and $\left\{\frac{\boldsymbol{10}}{\boldsymbol{10}}, \frac{\boldsymbol{100}}{\boldsymbol{1}}\right\} = \{\boldsymbol{1}, \boldsymbol{100}\}$.



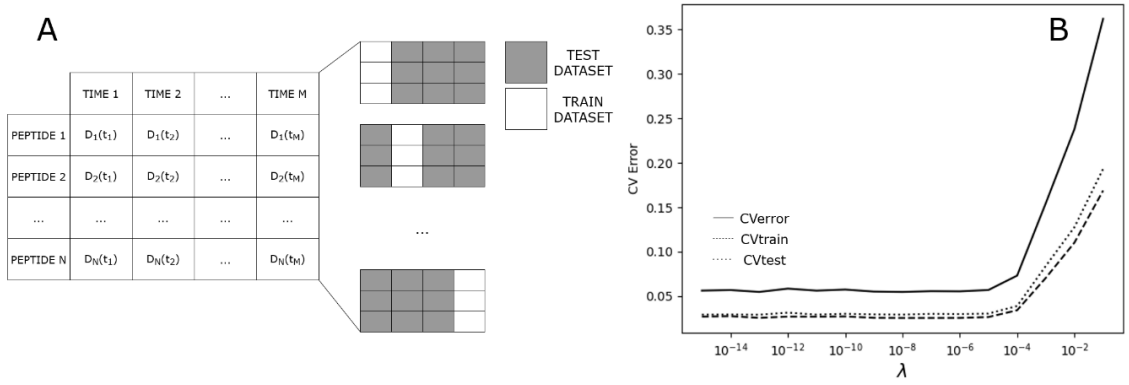
Supplementary Figure 3.2. The penalized cost function and leave-one-out cross validation.

Protection factors are adjusted to minimize the cost function (**Eq. 3.6** in main text):

$$\mathbf{C}(\boldsymbol{\lambda}, \{\mathbf{P_i}\}) = \underbrace{\sum_{\mathbf{j}} \sum_{\mathbf{k}} \mathbf{w_{jk}} \left[\mathbf{D_j^{pred}}(\mathbf{t_k}, \{\mathbf{P_i}\}) - \mathbf{D_j^{exp}}(\mathbf{t_k}) \right]^2}_{\mathbf{SSR}} + \underbrace{\boldsymbol{\lambda} \sum_{\mathbf{i}} (\mathbf{ln}\,(\mathbf{P})_{\mathbf{i-1}} - \mathbf{2ln}\,(\mathbf{P})_{\mathbf{i}} + \mathbf{ln}\,(\mathbf{P})_{\mathbf{i+1}})^2}_{\mathbf{Penalty\ term}}$$

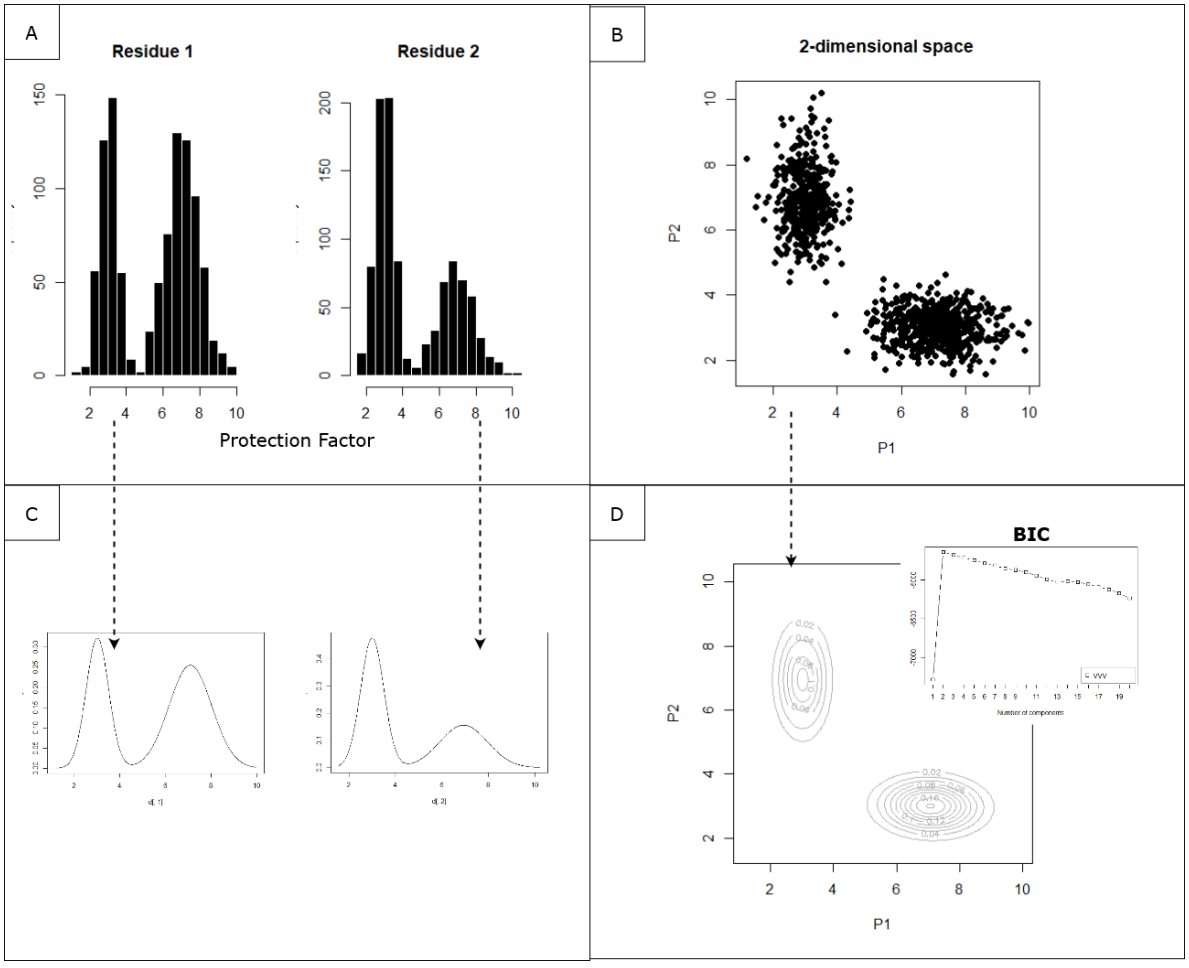
The cost function consists in a regular term, i.e. the sum of squared residuals (SSR), which depends on the experimental data, and a penalty term, which depends on the estimated parameters (i.e., the protection factors). The penalty term is built to minimize variations between protection factors of neighbouring residues (the functional form in the penalty term is the Newton approximation of the second derivative of a discrete function: $\frac{\mathbf{d^2x}}{\mathbf{dt^2}} \approx \mathbf{x_{i-1} - 2x_i + x_{i+1}}$). The relative contribution of the penalty term with respect to the regular term is determined by penalty constant λ.

We applied leave-one-out cross-validation (CV) to determine the penalty constant λ. The MS dataset was divided into a training dataset formed by 14 of the 15 time points available (for all peptides) and a test dataset composed by the remaining time point. This splitting procedure was repeated by leaving out one time point at a time; as a result, 15 training (and, analogously, test) datasets were generated (A). Starting from an initial guess on the protection factors (to ensure the reproducibility of the results, the protection factor of every residue except prolines was initialized to 1), the cost function was minimized for each training dataset. The minimized value of the cost function and the estimated set of protection factors was recorded; the sum of the cost function over all training datasets ($\mathbf{CV_{train}}$) was evaluated. The estimated set of protection factors was used to predict the fractional deuterium uptake of the test dataset (i.e., the remaining time points), and the cost function was calculated for every test dataset. The sum of the cost function over all test datasets ($\mathbf{CV_{test}}$) was evaluated. The CV error $\mathbf{CV_{error}}(\boldsymbol{\lambda}) = \mathbf{CV_{train}}(\boldsymbol{\lambda}) + \mathbf{CV_{test}}(\boldsymbol{\lambda})$ was calculated for λ ranging from $\mathbf{10^{-15}}$ to $\mathbf{10^{-1}}$. The value of λ corresponding to the minimum CV error was selected: $\boldsymbol{\lambda} = \mathbf{1} \times \mathbf{10^{-8}}$ (B).



Supplementary Figure 3.3. Visual representation of the clustering algorithm in 2 dimensions (i.e. for a 2-residue peptide).

The histograms of the protection factors that fit deuterium uptake curves (not shown) are shown in panel A. The application of a univariate clustering algorithm would identify two components per amino acid (panel C). The univariate approach provides four approximate solutions for which each amino acid assumes protection factors ~3 and ~7 (namely $\{\mathbf{3},\mathbf{3}\},\{\mathbf{7},\mathbf{3}\},\{\mathbf{3},\mathbf{7}\},\{\mathbf{7},\mathbf{7}\}$). The multivariate algorithm identifies two clusters (panel B), showing that solutions $\{\mathbf{3},\mathbf{3}\}$ and $\{\mathbf{7},\mathbf{7}\}$ are not possible. The clustering algorithm fits a mixture of 1 to 99 2-dimensional Gaussians to the 2-dimensional probability distribution (panel D) and calculates the BIC (Bayesian Information Criterion) associated to a specific number of Gaussian components (insert in panel D). The number of components with highest BIC is 2. Therefore, while the solution is underdetermined, the multivariate approach is less underdetermined than the univariate.



Chapter 4. Recalibrating Protection Factors Using Millisecond Hydrogen/Deuterium Exchange Mass Spectrometry

Submitted as: M Stofella, N Seetaloo, AN St John, E Paci, JJ Phillips and F Sobott, "Recalibrating protection factors using millisecond hydrogen/deuterium exchange mass spectrometry", Anal Chem, 2024

4.1 Abstract

Hydrogen/deuterium exchange mass spectrometry (HDX-MS) is a powerful technique to interrogate protein structure and dynamics. With the ability to study almost any protein without size limit, including intrinsically disordered ones, HDX-MS has shown fast growing importance as a complement to structural elucidation techniques. Current experiments compare two or more related conditions (sequences, interaction partners, excipients, conformational states etc.) to determine statistically significant differences at a number of fixed time points and highlight areas of changed structural dynamics in the protein. The work presented here builds on the fundamental research performed in the early days of the technique, and points towards establishing HDX-MS as an *absolute* and *quantitative,* rather than *relative* and *qualitative,* measurement. We performed millisecond HDX-MS experiments on a mixture of three unstructured peptides (angiotensin, bradykinin and atrial natriuretic peptide amide rat). We compared experimental deuterium uptake curves with theoretical ones predicted using established exchange rate calculations, which for an unstructured peptide depend on its primary sequence and on the pH and temperature at which the labelling was performed. Our aim was to uncover issues with the commonly used reference compound and suggest an alternative: we show that exchange rate calculations are more accurate when tri-alanine peptide is used as a reference instead of longer and partially structured poly-DL-alanine (PDLA). Through molecular dynamics (MD) simulations, we confirmed the high helical propensity of PDLA peptides, which need as

few as 15 residues to form one (or more) stable helix. The change of reference compound largely explains discrepancies between predictions and experiments previously reported. An accurate recalibration of intrinsic exchange rate calculations is crucial to enable kinetic modelling of the exchange process and to ultimately allow HDX-MS to move towards a direct link with atomistic structural models.

4.2 Introduction

Protein structural dynamics and order-disorder transitions play an important, but often overlooked role in the human proteome, with approximately one third of proteins being partly or fully disordered (Deiana et al. 2019). Well-studied examples include the important tumour suppressor p53 and the abundant, Parkinson's disease-related protein α-synuclein, which are known to undergo conformational transitions when interacting with DNA sequences and lipid bilayers, respectively. Moreover, the structural and functional behaviour of intrinsically disordered proteins (IDPs) and their interactions in the crowded cellular environment often depend on biophysical parameters such as pH, dielectric properties, ion concentrations and macromolecular crowding (Theillet et al. 2014). For example, IDPs can undergo conformational changes (Wohl, Jakubowski, and Zheng 2021) and even liquid-liquid phase separation (Wang et al. 2021) at different salt concentrations. *In vivo*, proteins are solvated in diverse environments which can vary considerably within and between cells, e.g. with high intracellular concentrations of specific metal ions (up to 20 mM for Mg^{2+}) and pH ranging from neutral in cytosol (6.8-7.2) to acidic conditions in endosomes and lysosomes (as low as 4.5).

While high-resolution structural techniques such as x-ray crystallography and cryo-electron microscopy can capture very detailed images of individual molecular states, the characterisation of structural dynamics and intrinsic disorder under near-physiological conditions remain challenging. Over the last decade, hydrogen/deuterium exchange mass spectrometry (HDX-MS) has emerged as a powerful technique to fingerprint structural and dynamic properties of proteins (Engen

et al. 2021; Englander et al. 2016; Masson et al. 2019; Trabjerg et al. 2018; Vinciauskaite and Masson 2023) in different solvent environments. HDX-MS utilizes the spontaneous exchange between amide backbone hydrogen atoms and deuterium in the solvent, which increases the mass of the protein and can be monitored by mass spectrometry to detect changes in the degree of hydrogen bonding per amino acid and determine local structural dynamics. HDX-MS data retain information about the exchange of a protein at peptide-level resolution (5-10 amino acids). The same phenomenon (HDX) can be monitored at the level of the single residue using nuclear magnetic resonance (NMR) spectroscopy (Dempsey 2001). Traditional HDX-NMR experiments probe exchange time scales of seconds/minutes (Chu and Pielak 2023; Kuwata et al. 2003; Olofsson et al. 2009) but can be pushed down to a sub-millisecond range (Kateb, Pelupessy, and Bodenhausen 2007). However, HDX-MS is more versatile, allowing the study of small molecules (Damont et al. 2021) as well as MDa complexes (Raval et al. 2021), it requires lower amounts of sample and is compatible with many buffers and solution conditions, therefore enabling studies in different chemical environments such as integral membrane proteins in lipid nanodiscs (Martens and Politis 2020) and biotherapeutics in formulations with added excipients (Wood et al. 2020). There is now fast growing interest in HDX-MS in the biopharma industry for the characterization of biotherapeutic molecules (Masson et al. 2017) and epitope mapping (Sun et al. 2021). More recently, fast (millisecond) HDX-MS has emerged as a key technique (Al-Naqshabandi and Weis 2017; Kish et al. 2023; Svejdal et al. 2019) to study weak (or fast cycling) binding interactions, allosteric effects, and dynamics of unstructured sequences in intrinsically disordered proteins (IDPs) (Seetaloo and Phillips 2022) that are associated with cancer and amyloid-related neurodegenerative diseases. It is desirable in such cases to map out conformational landscapes rather than just determining individual structures, and to understand the factors (which might be environmental rather than intrinsic to the sequence itself) which govern transitions between different states; a formidable challenge which is well addressed by fast HDX-MS approaches aided by ensemble calculations using advanced computational methods (Jia et al. 2023; Salmas, Harris, and Borysik 2023).

In HDX-MS, the backbone amide hydrogen-exchange rate is an important and highly sensitive measure of a protein's structural dynamics (Linderstrøm-Lang 1955). To accurately assess differences in the exchange kinetics, it is necessary to distinguish the impact of the chemical environment from that of the protein itself and its structural changes. The kinetics of exchange depends on the acidity/basicity of the respective amide protons which is determined by the sequence, i.e. the nature of each amino acid and its nearest neighbours (Bai et al. 1993; Connelly et al. 1993; Molday et al. 1972; Nguyen, Mayne, Phillips, and Walter Englander 2018), as well as on structural properties of the protein which define the 3D microenvironment of an amino acid – mainly dictated by hydrogen bonding, electrostatics and solvent accessibility (Devaurs et al. 2022). Exchange rates also depend on chemical properties of the solvent which determine the mobility and activity of protons (H^+/D^+), which in turn is intimately linked with the availability of hydroxide ions as the actual catalytic agents initiating HDX (pH/pD, temperature and ionic strength). While effects of pH and temperature on OH^-/OD^+ activity can in principle be predicted by calculations, salt effects are usually not explicitly considered. An approach adopting an empirical buffer correction has been recently proposed, where a reporter peptide is used to detect differences in exchange cause by the introduction of additives in the buffer (Toth et al. 2017). A theoretical framework enabling the prediction of intrinsic exchange rates as a function of salt type and concentration is however lacking. Published empirical calculations of *intrinsic* exchange rates, which refer to the exchange rate of a residue in a completely unfolded chain, take some of these factors into account, and they are usually calibrated based on what is assumed to be a fully unstructured sequence. Current practice in HDX-MS relies on relative measurements of two or more states in direct comparison, and it interprets the differential exchange pattern at the peptide level *qualitatively*, based on statistical significance. This falls well short of what the method could in principle achieve with proper calibration. If true and correctly calibrated intrinsic rates were available which take salt effects and accurate back-exchange estimations into account, "absolute" H/D exchange levels could be measured directly instead of differences between conditions. With such knowledge of *quantitatively* correct exchange rates, sets of protection factors

could be determined which are meaningful across separate experiments and different solvent environments. That would in turn then also facilitate the use of such information in integrative structural modelling approaches, with the ultimate goal to combine HDX-MS and molecular dynamics (MD) for elucidation of protein structural ensembles. Here, we make some key steps towards this goal.

In this work, we performed millisecond HDX-MS experiments on a mixture of unstructured peptides to test the validity of the commonly used intrinsic exchange rate estimates provided by the Englander group (Bai et al. 1993; Connelly et al. 1993; Molday et al. 1972; Nguyen, Mayne, Phillips, and Walter Englander 2018), with the aim to determine an appropriate unstructured reference sequence. The assumption that the peptides are unstructured was validated by circular dichroism (CD) spectroscopy. The exchange of unstructured peptides is too fast to be detected in a 'standard' HDX-MS instrument, where the minimum acquisition time is 20-30s. The access to the millisecond time scale is proven here to be crucial to determine correct intrinsic exchange rates and how they are influenced by the presence of salt. Our findings revealed that intrinsic exchange calculations are more accurate when a three-alanine peptide (3-Ala) reference is used instead of the standard poly-DL-alanine (PDLA), which retains some residual structure. We used MD simulations to confirm the high structural propensity of PDLA peptides, which had already been reported by several computational and experimental studies (Chakrabartty, Kortemme, and Baldwin 1994; Hinck 2022; Ingwall et al. 1968; Kuczera et al. 2021; López-Llano, Campos, and Sancho 2006; Rohl, Fiori, and Baldwin 1999). Therefore, our results corroborate the validity of the established intrinsic exchange rate calculations when recalibrated using a proper unstructured reference such as 3-Ala.

4.3 Methods

4.3.1 Theoretical framework

In principle, HDX provides information at the resolution of the single amino acid. Indeed, the Linderstrøm-Lang model (Linderstrøm-Lang 1955) describes the exchange of each residue as a two-step process, the first guided by local fluctuations of the protein, the second by the chemistry of the individual residue and the surrounding solvent. The observed rate of exchange

Eq. 4.1 $$k_{obs} = \frac{k_{int}}{P}$$

is defined as the ratio between the *intrinsic* exchange rate k_{int}, representing the exchange rate of the amino acid in a completely unfolded chain, and the protection factor P, which can be interpreted as the "degree of protection" of the residue induced by the structure of the protein. Intrinsic exchange rates have been studied in the early days of HDX and their dependence on pH, temperature (Scrosati, Yin, and Konermann 2021; Tajoddin and Konermann 2022) and side chains of the neighbouring residues is widely accepted (Bai et al. 1993; Connelly et al. 1993; Molday et al. 1972; Nguyen, Mayne, Phillips, and Walter Englander 2018). On the other hand, the protection factor encodes structural properties of the residue within the protein (Englander et al. 2016): several microscopic models have been developed to link the structure of a protein with its protection factors, with satisfying outcomes (Devaurs et al. 2022). Retrieving a well-defined biophysical parameter, such as the protection factor, from HDX-MS experiments allows linking data with atomistic models of protein structure and dynamics obtained from complementary techniques, such as NMR, cryo-EM or molecular dynamics (MD) simulations. Differential (i.e. relative and qualitative) HDX-MS data are extremely useful to locate the effect of a perturbation, but they make predictions of structural properties and correlation with other experiments rather challenging (Hamuro 2021a).

Isolating the effect of chemistry (k_{int}) is crucial to derive absolute structural information (P) from the observed data (k_{obs}). Even in differential studies, omitting the deconvolution of these two effects can introduce a bias in the results or, worse, can lead to the wrong conclusions, mostly when studying conformational changes of proteins under different buffer conditions, e.g. when dealing with temperature- or pH-driven conformational changes (Li et al. 2014; Tajoddin and Konermann 2020). Consider that a minor change in pH can cause differences in the uptake curves that can be misclassified as significant structural changes. For example, we used the Englander intrinsic exchange rates to calculate that a difference in pH of 0.1 is sufficient to generate differences > 0.5 Da in the uptake curve of an unstructured peptide with sequence AAAAAAAAAA at temperature 300 K (**Supplementary Figure 4.1**).

One of the main challenges associated with deriving quantitative information (such as the absolute protection factors, rather than their relative differences) from HDX-MS data is the deconvolution of the peptide level data provided by the experiment into single residue information (Kan et al. 2013). We have recently developed a computational method that exploits the additional information contained in the isotopic envelope to extract (most of) the protection factors of a protein from HDX-MS data (Skinner et al. 2019), and have shown that our estimates correlate well with NMR measurements (Stofella et al. 2022). Our method relies on the accuracy of the intrinsic exchange rates, which we assumed to be correct and constant (for a given sequence at a fixed pH and temperature), following the empirical estimates developed by the Englander group (Bai et al. 1993; Connelly et al. 1993; Molday et al. 1972; Nguyen, Mayne, Phillips, and Walter Englander 2018). We decided to further challenge our assumption by studying the exchange of unstructured peptides, taking advantage of the recent developments in the acquisition of millisecond HDX-MS data (Kish et al. 2023; Seetaloo and Phillips 2022).

The intrinsic exchange rate estimates from the Englander group assume that the exchange rate of a residue in a completely unfolded structure depends mainly on three

factors: i) temperature, ii) pD of the labelling buffer and iii) side chains of the neighbouring residues (Bai et al. 1993; Connelly et al. 1993; Molday et al. 1972; Nguyen, Mayne, Phillips, and Walter Englander 2018). Additional factors, such as the reported dependency of the intrinsic exchange rate on salt concentration (Kim and Baldwin 1982), are neglected. The temperature dependence follows the Arrhenius law, which is valid within the range of temperatures 0-60 °C provided that the protein structure remains stable, while it needs to be adjusted for higher temperatures (Tajoddin and Konermann 2020). The dependence of the intrinsic exchange rate on the pD ($pD = pH_{read} + 0.4$) has a V-shaped curve, with a minimum at pD 2.55 (this value is averaged over all amino acids). The dependence of the intrinsic rate on the neighbouring side chains was empirically determined by studying all 20 naturally occurring amino acids with dipeptide models and comparing their exchange rates with polyalanine models (Connelly et al. 1993). In their original paper (Bai et al. 1993), Bai et al. used NMR to determine the reference values for the left (L) and right (R) isomers of an alanine dipeptide (N-Ac-Ala-N'MA), for the internal NH of a blocked alanine tripeptide (N-Ac-Ala-Ala-Ala-N'MA) and for a racemic poly-DL-alanine (PDLA) with degree of polymerization 28 (which represented the average length of the polypeptides). The reference rates were measured in presence of 0.5 M KCl and then extrapolated to "low salt concentration" (Bai et al. 1993). In a follow-up study, the Englander group adjusted the reference values for PDLA (at low salt concentration) by a factor of 1.35, after comparing the exchange of PDLA peptides of different lengths with apolipoprotein C3, which was assumed to be completely unstructured (Nguyen, Mayne, Phillips, and Walter Englander 2018). However, several studies have criticised the validity of these calculations because they could not match the predictions with experimental data: the experimental uptake was found to be faster than the predicted one, which for a fully unstructured reference should be the fastest exchange possible on that amino acid (at a fixed pH and temperature) (Al-Naqshabandi and Weis 2017; Del Mar et al. 2005; Keppel and Weis 2013; Mori, van Zijl, and Shortle 1997; Zhang et al. 2012).

4.3.2 Materials

Deuterium oxide (99.9% D_2O) was purchased from Goss Scientific (catalogue number: DLM-4). The peptide mixture (**Supplementary Table 4.1**) contained three peptides (10 μM each): angiotensin (A9202, Sigma Aldrich), bradykinin (90834, Sigma Aldrich) and ANP (atrial natriuretic peptide) Amide Rat (SCP0022, Sigma Aldrich). We performed circular dichroism (CD) spectroscopy experiments to validate the assumption that the peptides are completely unfolded (**Supplementary Figure 4.2**).

4.3.3 Hydrogen/deuterium exchange experiments

Hydrogen-deuterium exchange (HDX) was performed using a fully-automated, millisecond HDX labelling and online quench-flow instrument, ms2min (Applied Photophysics, UK) (Kish et al. 2023; Seetaloo and Phillips 2022), connected to an HDX manager (Waters, USA). The peptide mixture (**Supplementary Table 4.1**) in the equilibrium buffer (20 mM Tris, pH 7.40) was delivered into the labelling mixer and diluted 20-fold with labelling buffer (20 mM Tris, pH_{read} 7.00) at 20˚C, initiating HDX at 95% deuteration. The labelling times depended on the varying length of mixing loops in the sample chamber and the flow rate of the carrier buffer. The peptides were labelled for a range of times from 50 ms to 5 min. Immediately post-labelling, the sample was mixed 1:1 with quench buffer (100 mM Tris, pH = 2.55 for the mixture of equilibration and quench buffer) in the quench mixer to minimize any further exchange. The sample was loaded into the HPLC injection loop of the ms2min and sent to the HDX manager. The peptides were trapped on a VanGuard 2.1 x 5 mm ACQUITY BEH C18 column (Waters, USA) for 3 minutes at 7000-9000 psi and separated on a 1 × 100 mm ACQUITY BEH 1.7 μm C18 column (Waters, USA) with a 4 min. linear gradient of acetonitrile (15-40%) supplemented with 0.1% formic acid. The eluted peptides were analysed on a Synapt G2-Si mass spectrometer (Waters, Wilmslow, UK). An MS-only method with a low collisional activation energy was used: fragmentation was not needed as we wanted to study the exchange of intact peptides with known sequence.

Up to 4 technical replicates were collected. Deuterium incorporation into the peptides was measured in DynamX 3.0 (Waters, USA).

4.3.4 Data processing and analysis

The evolution of the isotopic envelope of the three peptides was monitored as a function of time. We calculated the experimental fractional deuterium uptake as

Eq. 4.2 $$D_{Frac}^{Exp}(t) = \frac{D(t) - D_0}{D_{Max} - D_0}$$

where D(t) is the centroid (intensity-weighted average) of the isotopic envelope of the peptide at time t, D_0 is the centroid of the fully protonated envelope (no exchange) and D_{Max} is calculated as the centroid of the maximally deuterated envelope (after 5 minutes labelling when the uptake reached a plateau). The experimental fractional uptake was averaged over the replicates available, with the error associated with experimental measurements as the pooled standard deviation (**Supplementary Figure 4.3**).

The theoretical fractional uptake was calculated using a sum of exponentials:

Eq. 4.3 $$D_{Frac}^{Theor}(t) = \frac{1}{N-1}\sum_{i=2}^{N}\left(1 - e^{-k_{int,i}t}\right)$$

where N is the number of exchangeable residues in the peptide (prolines are excluded) and $k_{int,i}$ is the intrinsic exchange rate of residue i. Note that the first residue is excluded from the sum because of the lack of an amide at the N terminus. To calculate the intrinsic exchange rate, we used a Python script (available at https://github.com/pacilab/exPfact)(Skinner et al. 2019) adapted from the spreadsheet of the Englander Lab (https://hx2.med.upenn.edu/download.html). The intrinsic exchange rate of one residue can be predicted from the knowledge of temperature, pH and side chains of the neighbouring residues (Bai et al. 1993; Connelly et al. 1993; Molday et al. 1972; Nguyen, Mayne, Phillips, and Walter Englander 2018), and was calculated using either polyalanine (PDLA) or the internal amide hydrogen of an alanine tri-peptide (3-Ala) as references (**Table 4.1**).

Table 4.1. Hydrogen/deuterium (HD) and deuterium/hydrogen (DH) exchange rate constants for alanine-based reference molecules at 293 K.

The values were empirically determined in previous work**(Bai et al. 1993; Connelly et al. 1993; Molday et al. 1972; Nguyen, Mayne, Phillips, and Walter Englander 2018)** by fitting experimental curves depicting the V-shaped dependence of the exchange rate of these reference molecules on the pD with the equation: $\boldsymbol{k_{ex} = k_A 10^{-pD} + k_B 10^{(pD - pk_D)} + k_W}$. Reference parameters for PDLA are available for forward (H to D) and reverse (D to H) exchange; the reference parameters for 3-Ala are available for forward exchange only.

Name	Exchange	$\log(k_A)$ ($M^{-1}\,min^{-1}$)	$\log(k_B)$ ($M^{-1}\,min^{-1}$)	$\log(k_W)$ (min^{-1})
3-Ala	HD	2.04	10.36	-1.5
PDLA	HD	1.62	10.05	-1.5
PDLA	DH	1.40	10.00	-1.6

The agreement between experimental (**Eq. 4.2**) and predicted (**Eq. 4.3**) fractional uptake was evaluated using the sum of squared residuals (SSR) over the J time points available:

Eq. 4.4 $$SSR = \sum_{j=1}^{J} \left(D_{Frac}^{Theor}(t_j) - D_{Frac}^{Exp}(t_j) \right)^2$$

4.3.5 Molecular Dynamics Simulations

Racemic poly-alanine peptides (50% D-alanine, 50% L-alanine, alternated) were constructed in PyMOL version 2.5.2. Acetyl and amide caps were added to neutralise each terminal charge. D-alanine residues were introduced manually by exchanging the Hα and methyl group (containing the Cβ, Hβ1, Hβ2 and Hβ3 atoms). Parameter and topology files were obtained using Tleap (Case et al. 2022); the ff19SB (Tian et al. 2020) and TIP3P force-fields for peptide and water molecules, respectively. Each peptide was solvated with a water box extending at least 12.0 Å away from any peptide atom. Potassium and chloride ions were added to obtain a concentration of 0.5 M KCl (Machado and Pantano 2020). Hydrogen mass repartitioning was carried out with ParmEd (Case et al. 2022) to facilitate a time-step of 4 fs. Each system was minimised using AMBER with 2500 steps of the steepest descent followed by 2500 steps of the conjugate gradient algorithm, or until convergence. A harmonic restraint was applied

to peptide atoms during minimisation and a 9 Å non-bonded interaction cut off distance was used. After minimisation, equilibration molecular dynamics (MD) was carried out using pmemd (Salomon-Ferrer et al. 2013) with a 1 fs time-step in the NVT ensemble, during which the temperature was slowly increased from 0 K to 293 K for 125 ps using Langevin dynamics with a collision frequency of 1 ps^{-1}. All bonds apart from those containing hydrogen were constrained using the SHAKE algorithm (Ryckaert, Ciccotti, and Berendsen 1977). Production runs followed equilibration dynamics for 200 ns using an increased time-step of 4 fs in the NPT ensemble where a 1 atm pressure was maintained using a Monte Carlo barostat. Snapshots were saved every 100 ps during the production runs and secondary structure propensity was calculated as an average over the snapshots using the DSSP algorithm (Kabsch and Sander 1983).

4.4 Results

4.4.1 Intrinsic exchange rate predictions are more accurate when 3-Ala is used as a reference

Experimental data showing the fractional uptake (**Eq. 4.2**) of angiotensin, bradykinin and ANP, assumed to be unstructured following CD experiments (**Supplementary Figure 4.2**), in absence of salt are shown in **Figure 4.1**. We predicted the fractional uptake (**Eq. 4.3**) of the peptides using the intrinsic exchange rates calculations by Englander (Bai et al. 1993; Connelly et al. 1993; Molday et al. 1972; Nguyen, Mayne, Phillips, and Walter Englander 2018), using either 3-Ala or PDLA as reference. The monoisotopic mass detected for ANP (**Supplementary Table 4.1**) reflects the formation of a disulfide bond between residues C_4 and C_{15}, so the parameters for cystine (and not reduced cysteine) were used in the intrinsic exchange rate calculations. To reproduce the uptake of bradykinin, we had to make some assumptions on the configuration of the prolines. Prolines do not exchange because they do not have an amide hydrogen, but their cis/trans isomerisation affects the exchange rate constants of neighbouring residues. Indeed, different parameters are tabulated in the intrinsic exchange rate calculations for trans or cis proline. The deuterium uptake curves

predicted by alternative bradykinin conformations are shown in **Supplementary Figure 4.4**. Ion mobility studies have shown that the most probable conformation corresponds to trans-Pro_2, trans-Pro_3, cis-Pro_7 (**Supplementary Table 4.2**) (Pierson et al. 2013). The deuterium uptake predicted for this conformation is shown in **Figure 4.1**.

The predictions were compared with experimental data and the agreement was evaluated using the Sum of Squared Residuals (SSR, **Eq. 4.4**). The SSR was 0.092 for angiotensin, 0.096 for trans-trans-cis bradykinin and 0.067 for ANP when PDLA was used as reference. Switching the reference from PDLA to 3-Ala reduced the SSR by approximately one order of magnitude: 0.011 for angiotensin, 0.022 for bradykinin and 0.010 for ANP, values compatible with the pooled standard deviation $\sigma_{pooled} = 0.041$. Our experimental measurements showed a faster exchange than the theoretical exchange of fully unstructured peptides when PDLA was used as reference, while they matched the predictions much better when 3-Ala was used as reference.

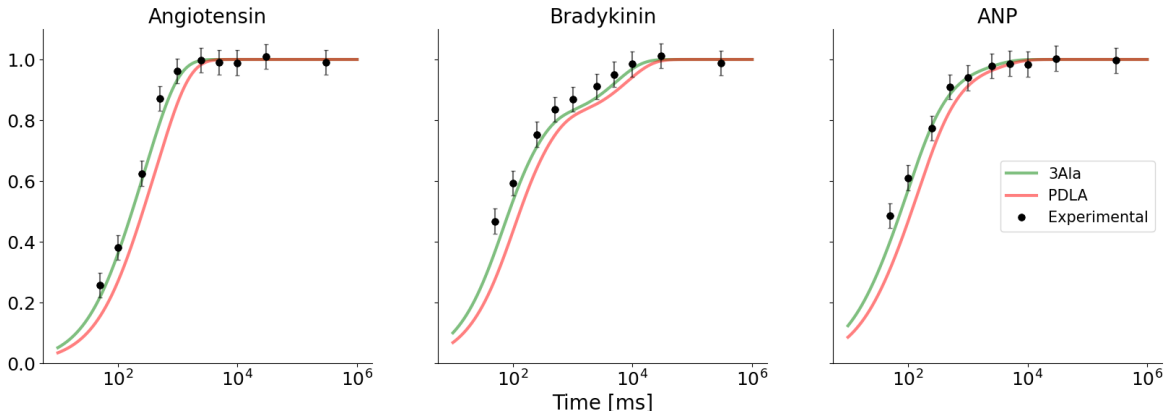


Figure 4.1. Hydrogen-deuterium exchange of angiotensin, bradykinin and ANP.
The experimental fractional uptake curves (black) are compared with theoretical deuterium uptake calculated using the intrinsic exchange rates calculations from Englander using 3-Ala (green) or PDLA (red) as reference. The error associated to the experimental measurements is the pooled standard deviation.

4.4.2 3-Ala rather than PDLA is a suitable unstructured reference

Molecular dynamics simulations of racemic PDLA (50% L-alanine, 50% D-alanine, alternated) highlighted its structural propensity. To replicate the experimental conditions used by Bai et al. (Bai et al. 1993), we simulated the behaviour of PDLA in presence of 0.5 M KCl. We performed simulations for PDLA peptides of increasing lengths (from 4 to 40 amino acids, with steps of 4) and measured the secondary structure propensity using the DSSP algorithm (Kabsch and Sander 1983). The average helical propensity per amino acid over the simulation time is reported in **Supplementary Figure 4.5**. The results in **Figure 4.2** show the helical propensity averaged over the amino acids as a function of the peptide length. The simulations highlight that a few alanine residues are sufficient to form helical conformations, with double (**Figure 4.2B**) or triple helical bundles (**Figure 4.2C**) forming at increasing peptide lengths. These results confirm our hypothesis, already supported by several experimental and computational findings (Chakrabartty et al. 1994; Hinck 2022; Ingwall et al. 1968; Kuczera et al. 2021; López-Llano et al. 2006; Rohl et al. 1999), that PDLA is not a suitable unstructured reference.

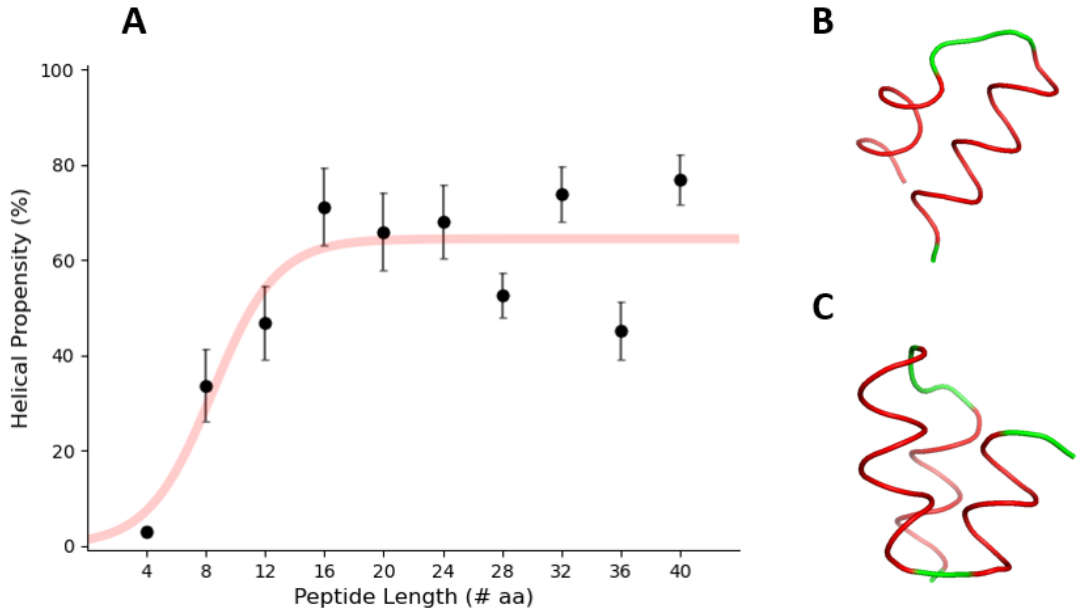


Figure 4.2. Structural propensity of PDLA peptides of increasing length from Molecular Dynamics simulations.

A) The helical propensity, calculated using the DSSP algorithm and averaged over the amino acids of the peptide, is shown as a function of the peptide length. The error bars associated to the helical propensity are the standard deviations. Snapshots were taken every 100 ps of simulation. B-C) Snapshots of a double helical bundle from the simulations of PDLA with 24 residues (B) and a triple helical bundle for PDLA with 32 residues (C). Simulations were performed by Dr Alex St John and analysed by the author.

Using PDLA as reference, several studies have observed that the intrinsic exchange rate calculations predicted an exchange *slower* than the experimental exchange of unstructured peptides or proteins (Del Mar et al. 2005; Keppel and Weis 2013; Mori et al. 1997; Zhang et al. 2012). This is in principle not possible because intrinsic exchange rates should describe the exchange of a fully unstructured peptide, i.e. the fastest exchange possible for a given amino acid sequence at a given temperature and pH. However, the calculations used in these studies i) used PDLA as reference instead of 3-Ala, and ii) did not account for minor corrections in the reference parameters that were introduced later (Nguyen, Mayne, Phillips, and Walter Englander 2018). For example, Al-Naqshabandi and Weis showed that intrinsic exchange rate calculations were not

able to reproduce the experimental curves for model peptides or unstructured proteolytic peptides of intrinsically disordered proteins (Al-Naqshabandi and Weis 2017). We compared the experimental exchange of a subset of these peptides with the deuterium uptake calculated using either PDLA or 3-Ala as reference (**Figure 4.3**). The exchange kinetics predicted using PDLA as reference (red curve) were either equal or *slower* than the experimental exchange, even after introducing the corrections implemented by Nguyen et al. (Nguyen, Mayne, Phillips, and Walter Englander 2018) for the intrinsic exchange rate calculations. When we predicted the exchange using 3-Ala as reference (green curve), the predicted mass increase was either equal or *faster* than the experimental data. This strengthened the fact that intrinsic exchange rate calculations are more accurate when 3-Ala is used as a reference instead of PDLA.

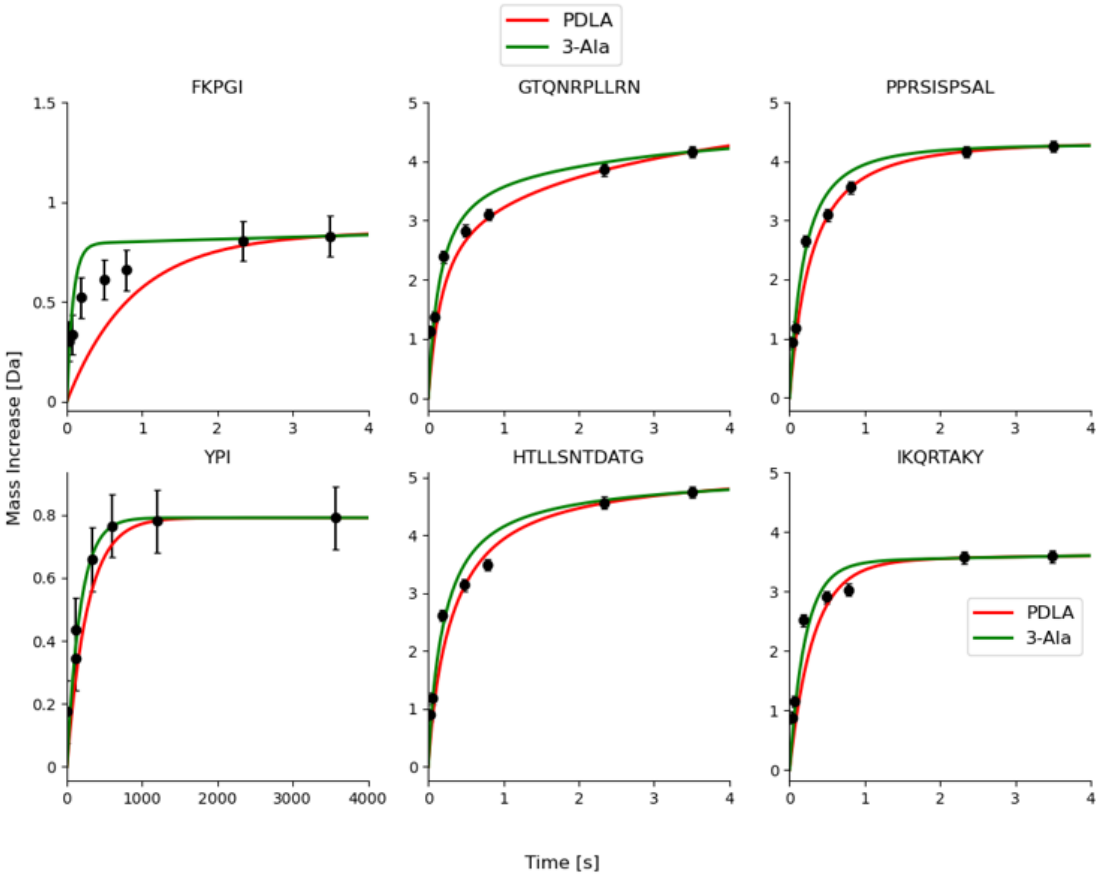


Figure 4.3. Experimental and predicted HDX data of unstructured peptides previously published.

Experimental HD exchange data of unstructured peptides (black) previously published by Al-Naqshabandi and Weis **(Al-Naqshabandi and Weis 2017)** is compared with the deuterium uptake calculated by us, using PDLA (red) or 3-Ala (green) as reference and accounting for the corrections published by Nguyen et al. (Nguyen, Mayne, Phillips, and Walter Englander 2018). Data points were extracted from figures published in **(Al-Naqshabandi and Weis 2017)** using a plot digitizer, and a default error of $\pm\mathbf{0.1}$ Da was assigned to experimental measurements. Mass increase is shown instead of fractional uptake to facilitate direct comparison with the original paper. The cis-Pro assumption is required for peptide FKPGI.

4.4.3 The intrinsic exchange rate depends on the ionic strength of the buffer

The results in **Figure 4.3** show that the predicted exchange (with 3-Ala as the correct reference, green) is always faster than the experimental uptake for these unstructured peptides. But why is the observed exchange *slower* than the prediction? We suggest that this can be explained by the presence of salt in the buffer used in these experiments (100 mM NaCl) (Al-Naqshabandi and Weis 2017). The dependence of the intrinsic exchange rate on the salt type and concentration had already been reported (Kim and Baldwin 1982). Bai et al. measured the k_{int} for all amino acids in presence of 0.5 M KCl "to shield possible charge effects" (Bai et al. 1993) and they extrapolated the values at "low salt concentration" by comparing their results with data previously published in absence of salt (Molday et al. 1972). Only the parameters at low salt concentration were reported in the well-known spreadsheet used for intrinsic exchange rate calculations (https://hx2.med.upenn.edu/download.html) and used in the follow-up study by Nguyen et al. (Nguyen, Mayne, Phillips, and Walter Englander 2018). Moreover, in a recent paper Toth et al. proposed the use of a reporter peptide to experimentally evaluate the effect of different buffer conditions on the exchange (Toth et al. 2017). We confirm here this additional dependence of the intrinsic exchange rate on the concentration (i.e. the ionic strength) of the salts in the labelling buffer. We measured the exchange of angiotensin, bradykinin and ANP in presence of 150 mM NaCl (**Figure 4.4**). The experimental curves were fitted with a stretched exponential ($D_{Frac} = 1 - e^{-k_{obs}t^q}$) and showed that the introduction of salt in the buffers slows down the exchange (subtly, yet significantly) of the peptides.

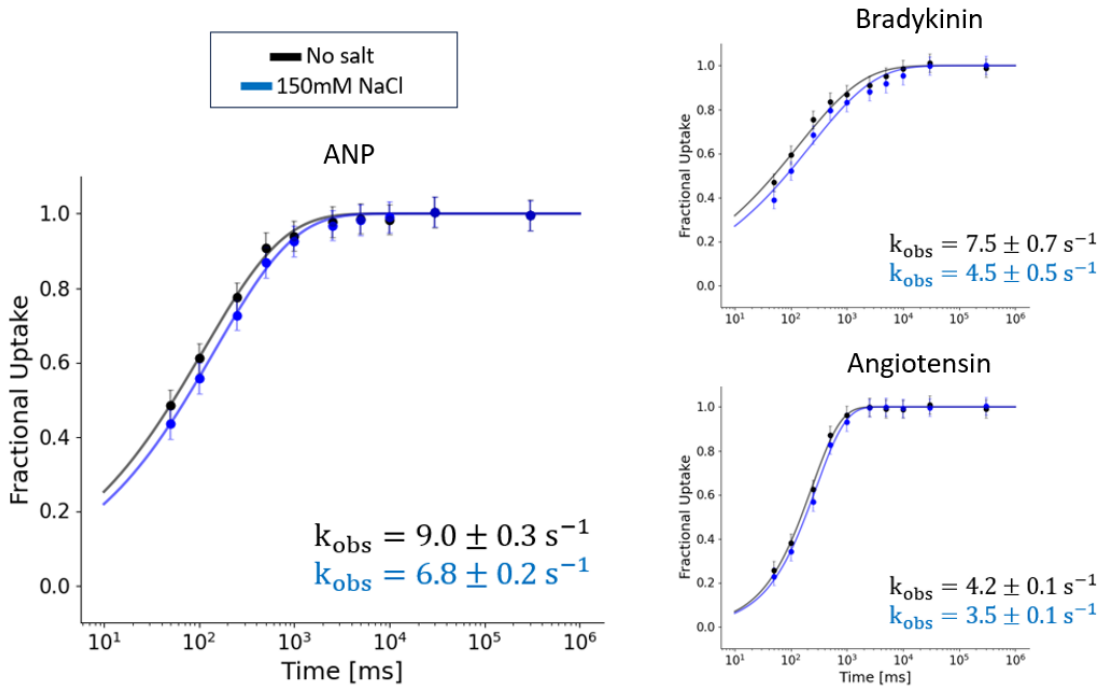


Figure 4.4. Fractional uptake of angiotensin, bradykinin and ANP in absence or presence of salt. Fractional uptake of angiotensin, bradykinin and ANP in absence (black) or presence of 150 mM NaCl (blue). The error associated to experimental measurements is the pooled standard deviation. Experimental data are fitted with a stretched exponential model and the observed rates $\boldsymbol{k_{obs}}$ are reported.

4.5 Discussion

HDX-MS measures an observable (deuterium incorporation) that is related to structural properties of the protein (Devaurs et al. 2022; Marzolf, Seffernick, and Lindert 2021; Peacock and Komives 2021) and it has been proven powerful in deriving data-driven structural models in combination with reweighting techniques and computational modelling (Calvaresi et al. 2024; Jia et al. 2023; Salmas et al. 2023). To achieve this goal, it is crucial to separate the effect of the buffer on the exchange pattern from that of the structure of the protein. Here we take key steps towards establishing a framework for a quantitative, "absolute" analysis of H/D exchange rates, which would

ultimately enable a direct connection between the HDX-MS dataset and the 3D protein model. We identify the fully unstructured tri-alanine peptide (3-Ala) as a more suitable reference peptide for intrinsic exchange rate calculations than the commonly used racemic poly-DL-alanine peptide (PDLA; **Figure 4.1** and **Figure 4.3**). Both references were already published in the original paper by Bai et al. (Bai et al. 1993), but only PDLA has been used in later studies. We show that PDLA is partially structured (**Figure 4.2**) and therefore cannot be used as a fully unprotected reference. After switching to a suitable reference (3-Ala), the observed exchange is either compatible with or slower (but never faster) than predicted, for the peptides in our mixture and for a set of disordered peptides previously published (Al-Naqshabandi and Weis 2017). We further confirm that intrinsic exchange rates have a subtle additional dependence on salt concentration which has so far not been considered in the theoretical k_{int} predictions. The addition of salts slows down the exchange, which can explain the remaining discrepancy that the experimental curves sometimes show *slower* (and not equal) exchange than predicted.

PDLA cannot be used as a fully unstructured reference because it displays slower exchange than 3-Ala as well as some other peptides, and we show that this is due to it having a high helical propensity above a length of ca. 10-15 amino acids (**Figure 4.2**). PDLA has been used as a standard at different temperatures and different pH in a publication by Linderstrøm-Lang et al. more than 65 years ago (Berger and Linderstrøm-Lang 1957). At the time, the authors remarked that "the slow exchange may be explained by a stabilization of the helix due to internal nonpolar bonds between the methyl groups of the side chains". Hence, we are actually not surprised that PDLA turned out to be unsuitable for exchange rate calculations, as it retains some protection despite its mixed D/L stereochemistry. Even Bai et al. in their original paper stated that "the NH and CαH resonances of the PDLA sample showed some substructure, apparently intrinsic to interactions of the D and L residues" (Bai et al. 1993). The structural propensity of PDLA was also reported by Frushour and Koenig using Raman spectroscopy: "When PDLA is dissolved in water, the spectra suggest that short α-

helical segments are formed upon dissolution" (Frushour and Koenig 1975). PDLA has probably been preferred to 3-Ala in the context of HDX-MS experiments because the exchange of 3-Ala is too fast to be detected with a "standard", i.e. manual or robotic workflow which is generally able to detect time points at or above 20-30 s. This highlights the importance of the millisecond timescale for fundamental studies of HDX. As a side technical note, PDLA peptides exhibit high hydrophobicity, making their purification quite challenging. This reinforces the argument that PDLA is not a good reference model.

4.6 Conclusions

HDX-MS has the potential ability to retrieve absolute structural information of a protein, in principle at the resolution of the single amide – either experimentally (Sobott 2020) or via an approach we described earlier (Stofella et al. 2022) – which is critical to get a robust correlation between HDX data and atomistic models of protein structure and dynamics. To achieve this goal, it is essential to separate the effect of solution chemistry from the effects of sequence and structure on the HDX pattern, and to use a correct, fully unstructured reference. Taken together these considerations enable us to obtain an accurate and precise estimate of the intrinsic exchange rate k_{int}.

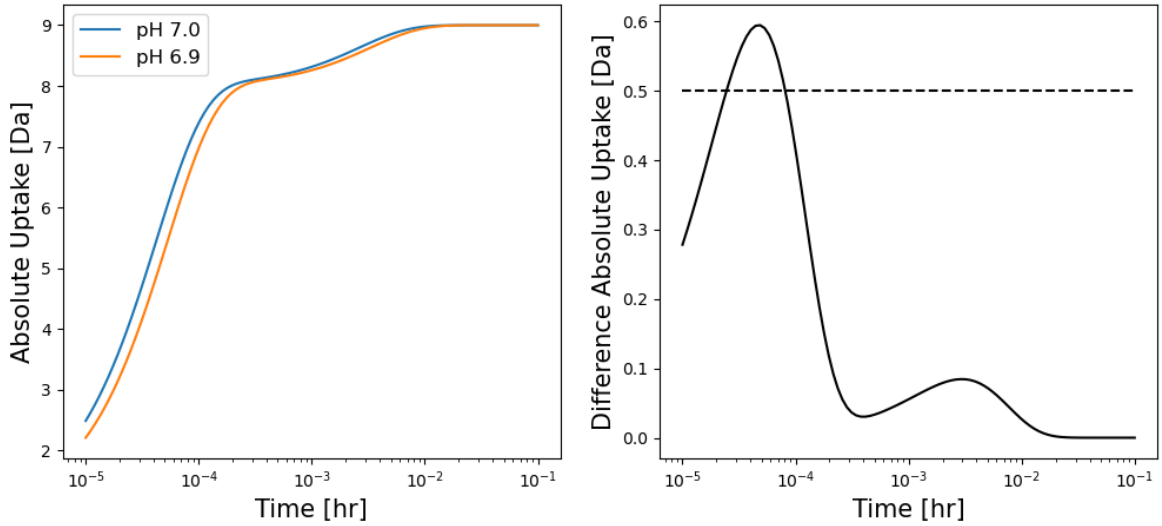
The empirical predictions for the intrinsic exchange rate developed by the Englander Lab have proven useful since their first publication in 1993, but their validity has been questioned by several studies probing the HDX of intrinsically disordered proteins, with some observed rates exceeding the supposedly fastest possible rate based on calibration with PDLA. We showed that these k_{int} calculations are more accurate when a tri-alanine peptide (3-Ala) is used as a reference instead of PDLA, because the latter is not a completely unstructured peptide. To perform these calculations, we therefore suggest to use the rate constants for 3-Ala (reported in **Table 4.1**) in the Englander spreadsheet (https://hx2.med.upenn.edu/download.html) or, alternatively, to use our Python script available on GitHub (https://github.com/pacilab/exPfact).

The exchange kinetics of unstructured peptides is also a function of ionic strength. The presence of salt (NaCl) at 150 mM slows down the exchange, and therefore the predictions mentioned above are not necessarily accurate when salt is present. In any case, the exchange predicted by the intrinsic rate should represent the fastest exchange possible for a given amino acid sequence at a given temperature and pH. We plan to further investigate such effects to determine a *salt correction factor*. Thus, we envisage in future that a combined correction for temperature, pH and salts will be possible, which will also allow to define intrinsic rates for forward exchange more stringently as amide exchange rates of an individual amino acid within a given sequence, under standardized conditions of pH, temperature and salt. This will serve to robustly deconvolve the true intrinsic rate of the covalent chemical structure, as determined by the protein sequence, from the extrinsic environmental conditions and from the effects of protein structural dynamics – the information that we ultimately want to reveal.

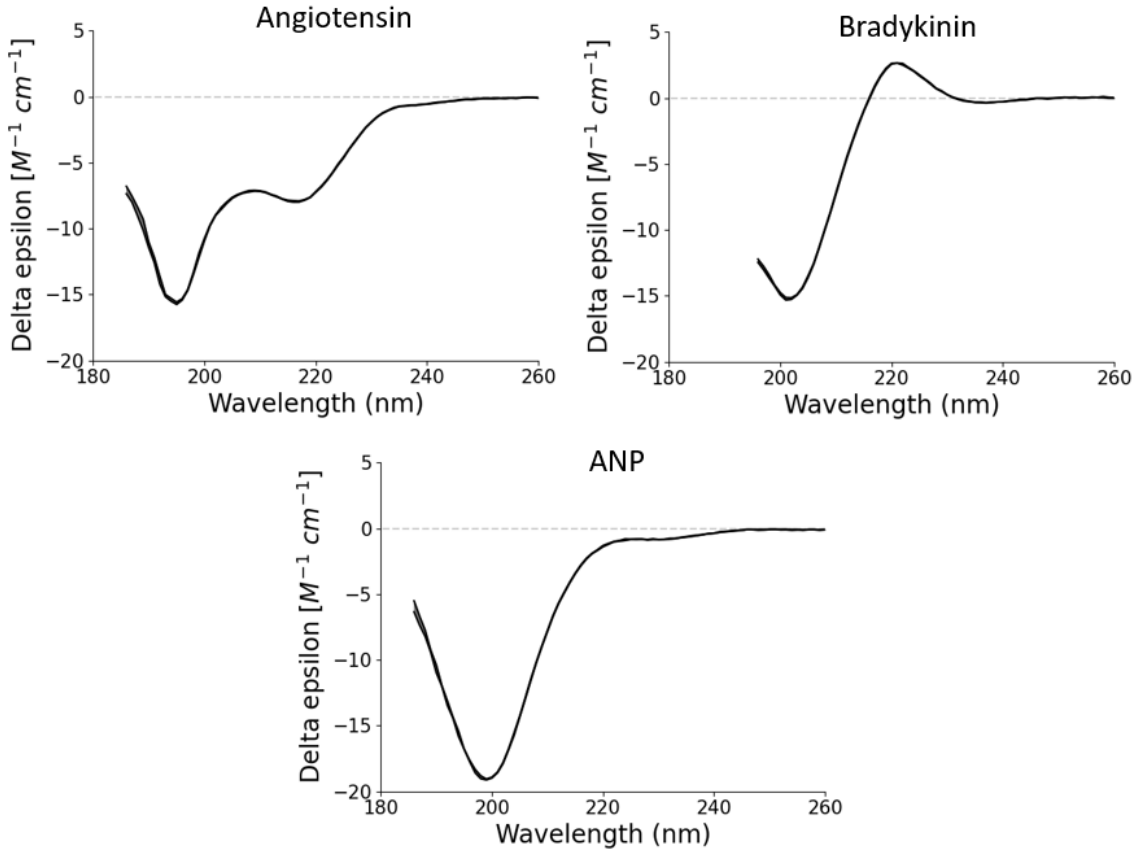
4.7 Supplementary Material

Supplementary Figure 4.1. Calculations showing that a minor change in pH can cause differences in the uptake curves that can be misclassified as significant.

We used intrinsic Englander's exchange rate calculations to determine the exchange rates of a completely unfolded poly-alanine peptide (sequence: AAAAAAAAAA) at temperature 300 K and at pH 7.0 (blue line) and pH 6.9 (orange line). The absolute uptake curves are calculated using Equation 2 (without normalization). The difference between the two curves is evaluated and is found to be > 0.5 Da for time scales ranging from ~72 ms to ~360 ms.

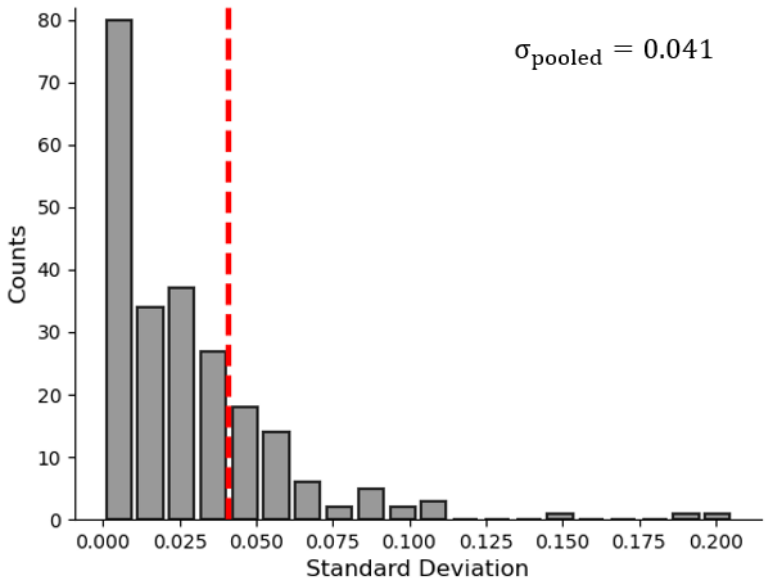


Supplementary Figure 4.2. Circular dichroism spectra of the peptides in the peptide mixture.
Circular dichroism (CD) spectra of peptides in the peptide mixture (**Supplementary Table 4.1**). CD experiments were performed on an Applied Photophysics Chirascan instrument (temperature 5°C), acquiring wavelengths in the range 180-250 nm. Wavelengths associated with voltages above 700V were cropped as they are not reliable (there are not enough photons to have a statistically relevant measure). All spectra show a minimum at around 200 nm, which indicates that the primary behaviour of all peptides is that of a random coil. ANP does not have additional peaks, suggesting that it is the most unstructured peptide in the mixture (despite being the longest). The spectrum of bradykinin resembles poly-proline II in a conformation that "maximizes favourable interactions with the solvent" (Rucker and Creamer 2002). The spectrum of angiotensin has a second minimum at 218 nm which generally represents a beta-like structure; given the length of the angiotensin peptide (7 residues), this second peak most probably indicates a couple of residues with beta-like psi/phi angles preferences. It is not uncommon to find "conformational preferences" in unstructured peptides (Smith et al. 1996).



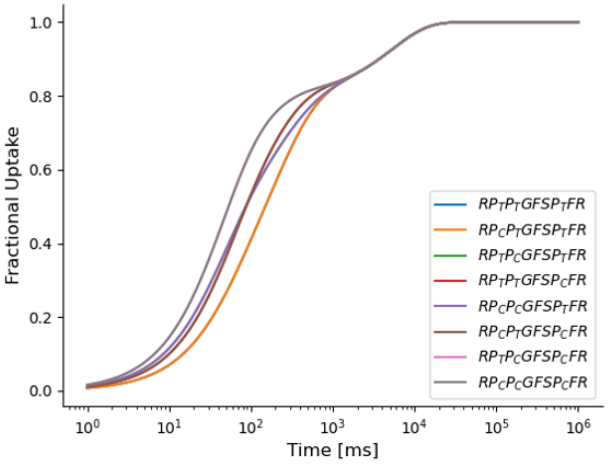
Supplementary Figure 4.3. Distribution of standard deviations.

Distribution of standard deviations of the fractional uptake from the entire dataset, i.e. considering measurements from all replicates, time points and conditions. The red vertical dashed line represents the pooled standard deviation $\sigma_{\text{pooled}} = 0.041$



Supplementary Figure 4.4. The effect of proline conformations on the H/D exchange of bradykinin.

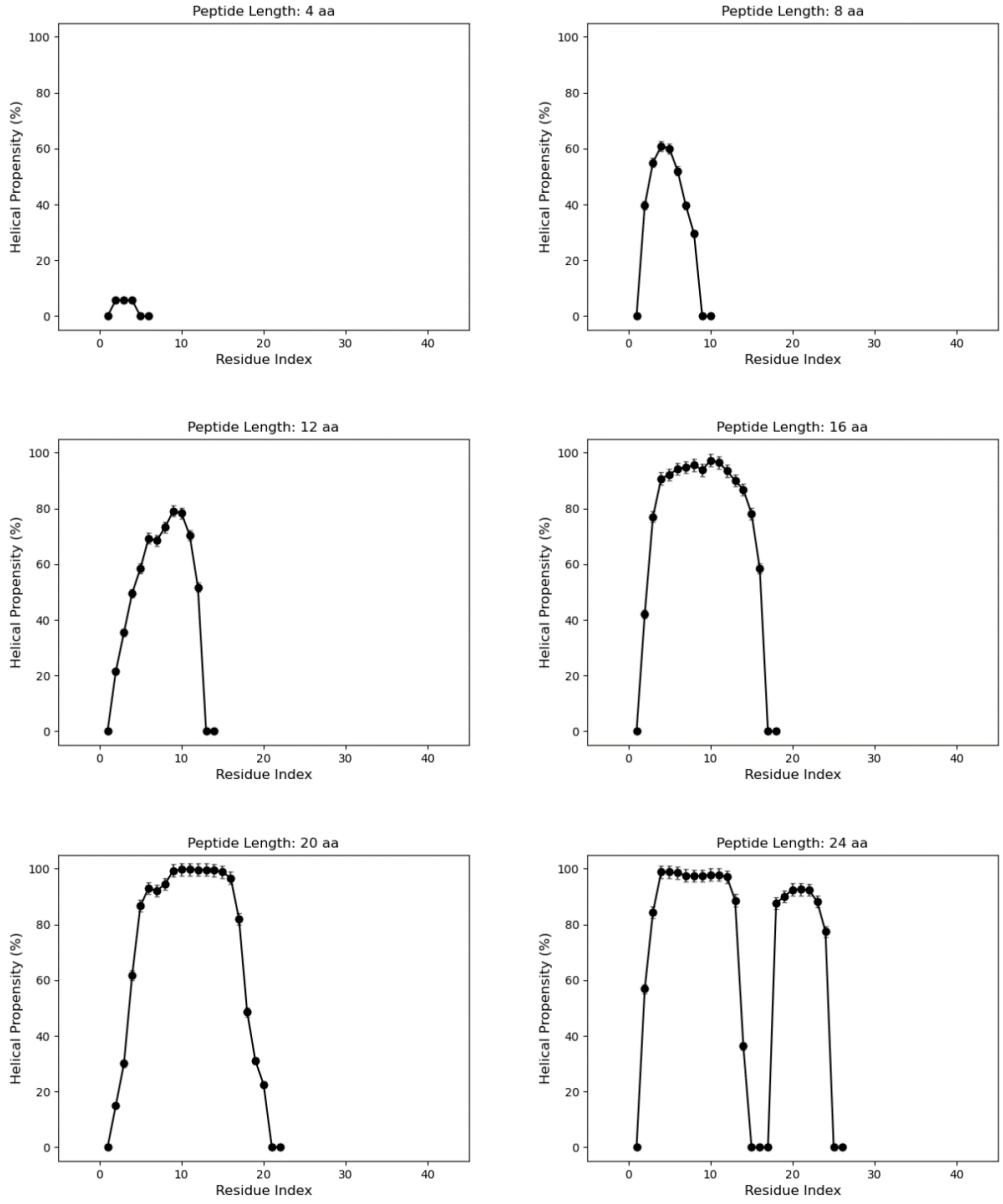
The fractional uptake calculated using 3-Ala as reference for different combinations of trans (PT) or cis (PC) prolines in the sequence of bradykinin. The table on the right summarizes the agreement with the experimental data shown in Figure 4.1 via the sum of squared residuals (SSR). Independently of the conformation chosen, the SSR decreases when 3-Ala is used as reference instead of PDLA. The conformation depicted in Figure 4.1 is highlighted in red.

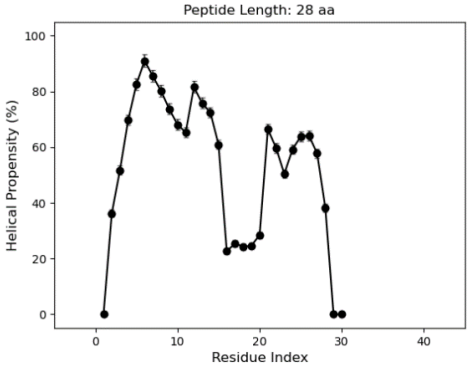


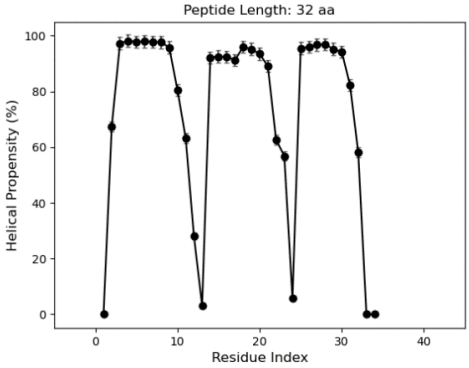
Sequence	SSR (3-Ala) SHOWN	SSR (PDLA) NOT SHOWN
$RP_TP_TGFSP_TFR$	0.107	0.245
$RP_TP_TGFSP_CFR$	**0.022**	**0.096**
$RP_TP_CGFSP_TFR$	0.027	0.100
$RP_CP_TGFSP_TFR$	0.107	0.244
$RP_CP_CGFSP_TFR$	0.027	0.100
$RP_CP_TGFSP_CFR$	0.022	0.096
$RP_TP_CGFSP_CFR$	0.008	0.023
$RP_CP_CGFSP_CFR$	0.008	0.023

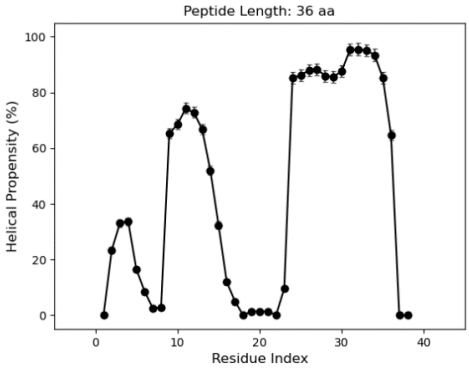
Supplementary Figure 4.5. Structural propensity of PDLA peptides of increasing lengths from Molecular Dynamics simulations.

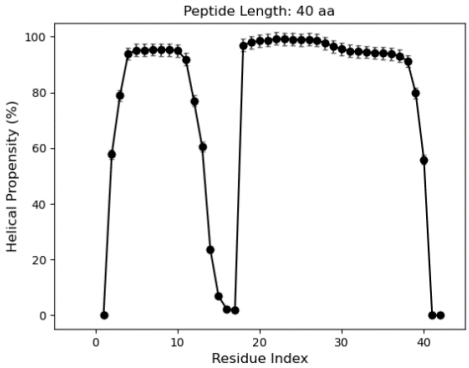
Snapshots were acquired every 100 ps of simulation and the secondary structural propensity of the peptide was calculated using the DSSP algorithm **(Kabsch and Sander 1983)**. Average values are reported for helical propensity as a function of the amino acid index. Note that the residues at the first and last index corresponds to the acetyl and amide caps of the peptide.




Peptide Length: 28 aa
Helical Propensity (%)
Residue Index


Peptide Length: 32 aa
Helical Propensity (%)
Residue Index


Peptide Length: 36 aa
Helical Propensity (%)
Residue Index


Peptide Length: 40 aa
Helical Propensity (%)
Residue Index

Supplementary Table 4.1. Peptide mixture.

The peptide mixture was formed by three peptides: angiotensin, bradykinin and ANP 4-23 Amide rat. The sequences of the peptides and the measured monoisotopic mass are reported together with the Sigma Aldrich code. No information on the proline configuration of bradykinin is provided by the supplier. The monoisotopic mass detected for ANP (C-terminal amide) differs by -2 Da from the sequence mass (1595.75 Da), reflecting the formation of a disulfide bond between residues C4 and C15.

Peptide	Sequence	Experimental Monoisotopic Mass (Da)	Sigma Aldrich Code
Angiotensin	DRVYIHP	889.455	A9202
Bradykinin	RPPGFSPFR	1059.550	90834
ANP	RSSCFGGRIDRIGAC-NH_2 [Cys4-Cys15]	1593.750	SCP0022

Supplementary Table 4.2. Main conformations of bradykinin as determined by IM-MS.

Main conformations of bradykinin as determined by ion mobility-mass spectrometry (IM-MS) (Pierson, Valentine, and Clemmer 2010). While it is not possible to derive the stereochemistry of prolines from HDX-MS data, the effect of prolines on neighbouring residues is a secondary yet important element to be considered when dealing with intrinsic exchange rate calculations (**Supplementary Figure 4.4**). Although the trans configuration of prolines is usually considered to be more stable, the assumption of cis-prolines in unstructured peptides is not unrealistic (Kienlein, Zacharias, and Reif 2024). Indeed, it was reported that "the cis-peptidyl-prolyl (cis-Pro) conformations in unfolded polypeptide chains are populated to significantly higher levels [with respect to folded proteins]" (Alderson et al. 2018). Using IM-MS, Pierson et al. have shown that multiple (up to 10) bradykinin conformers can coexist (Pierson et al. 2011) with different combinations of cis/trans prolines (Pierson et al. 2013). Among the most abundant isomers, named A, B and C, the C state accounts for ~80% of the population (Pierson et al. 2010). The C conformation corresponds to the bradykinin isomer trans-Pro2, trans-Pro3, cis-Pro7 (Pierson et al. 2013), which we assumed for the curve depicted in **Figure 4.1**. While a mixture of conformers is likely to exist in solution, it is realistic to assume that the HDX-MS data would capture the exchange of the most abundant conformer. Nevertheless, other conformations of bradykinin still provide good agreement with the experimental fractional uptake of bradykinin (**Supplementary Figure 4.4**). Importantly, the use of 3-Ala as reference instead of PDLA improves the agreement with experimental data regardless of the bradykinin conformation considered.

Bradykinin Configuration	**Pro2**	**Pro3**	**Pro7**	**Abundance (quasi-equilibrium)**
a	*cis*	*cis*	*cis*	2 %
b	*cis*	*trans*	*trans*	16 %
c	*trans*	*trans*	*cis*	80 %

Chapter 5. Inferring Single-Residue Resolution From Peptide-Level Hydrogen/Deuterium Exchange Mass Spectrometry Data

Drafted as: M Stofella, A Grimaldi, M Batchelor, R Bayliss, B Schiffrin, A Zhuravleva, E Paci and F Sobott, "Inferring single-residue resolution from peptide-level hydrogen/deuterium exchange mass spectrometry data", 2024

5.1 Abstract

Hydrogen/deuterium exchange (HDX) is a spontaneous process in which amide hydrogens of proteins are replaced with deuterium in solution, providing insights into protein structure and dynamics. HDX can be probed at single-residue resolution by nuclear magnetic resonance (NMR) experiments, though this technique is limited to smaller proteins (< 50 kDa). Mass spectrometry (MS) extends HDX studies to larger proteins but offers lower peptide-level resolution. The challenge of extracting single-residue information (such as protection factors) from HDX-MS data can be framed as an optimization problem, where the goal is to find the minima of a rugged cost function landscape. We conducted a comparative study of HDX kinetics on two proteins – ubiquitin, a well-structured model protein, and GB1-TACC3, which contains disordered regions – under identical experimental conditions, using both MS and NMR. We show that single-residue resolution can be inferred from peptide-level HDX-MS data, with a strong correlation to NMR-derived protection factors, considered the reference standard. We propose a protection factor analysis that addresses the inherent ambiguity of peptide-level data either by performing a random search to broadly explore the rugged cost function landscape or by incorporating an informed initial guess into the optimization process, thus focusing on physically meaningful minima. We also show that dataset quality, particularly redundancy, is critical to achieving reliable

single-residue resolution, with higher redundancy yielding more precise predictions. Our findings show that peptide-level HDX-MS, when supported by sufficient data redundancy and temporal sampling, can effectively provide single-residue insights comparable to NMR, expanding the applicability of HDX-MS to absolute quantitative studies of both structured and disordered proteins.

5.2 Introduction

Hydrogen/deuterium exchange (HDX) is a spontaneous phenomenon that occurs in proteins when they are in a solution containing deuterium oxide (Englander et al. 2016; Hamuro 2021b; Linderstrøm-Lang 1955). The amide hydrogen of each residue (except proline) exchanges with deuterium in solution, with a rate that depends on the chemical environment (Bai et al. 1993; Connelly et al. 1993; Molday et al. 1972; Nguyen, Mayne, Phillips, and Englander 2018) (pH, temperature, ionic strength of the solution and side chains of the neighbouring amino acid) as well as on structural and dynamical properties of the protein (mainly dictated by the amount of hydrogen bonding and heavy contacts that the amide hydrogen forms with surrounding residues) (Best and Vendruscolo 2006; Hamuro 2024; Radou et al. 2014; Vendruscolo et al. 2003).

Historically, the characterization of the HDX of proteins goes back several decades but received a boost in the 1990s and early 2000s thanks to the pioneering work of the Englander group (Englander 2023) and the advent of 2D nuclear magnetic resonance (NMR) (Dempsey 2001; Englander and Mayne 2014). In HDX-NMR experiments, a protein bearing ^{1}H is exposed to a solvent containing deuterium. As the exchangeable hydrogen atoms on the protein are replaced with (non-observable) ^{2}D atoms from solvent, the intensity of the observable ^{1}H signals decays over time (Kleckner and Foster 2011). The measurement of individual pseudo-first order exchange rate constants is limited only by the resolution of the amide signals themselves, i.e. the ability to uniquely assign all non-overlapping resonances or cross peaks in multidimensional NMR spectra (Dempsey 2001). The exchange kinetics time scale that HDX-NMR can measure goes typically from minutes to hours, but can be as short as

microseconds (Kateb et al. 2007). HDX-NMR has been mainly used to study protein folding intermediates (Englander 2000). Applications of HDX-NMR are hampered by limitations in protein size (< 50 kDa) and the costs associated with expressing and purifying ^{15}N labelled proteins. With its ability to overcome these limitations, mass spectrometry (MS) rose as a valid alternative to study the HDX of proteins, leveraging the increase in mass of the protein due to the exchange of protein hydrogens with deuterium in solution. In a typical HDX-MS experiment (Masson et al. 2019), a protein sample is diluted into a solution containing deuterium oxide and left for an increasing amount of time, until the exchange is quenched (by lowering the pH and temperature of the buffer). The protein is then digested by an acid protease, peptides are separated rapidly by liquid chromatography (LC) and injected into the mass spectrometer. MS has widened the range of biological systems for which HDX can be studied: today, relatively large systems such as integral membrane proteins are routinely analysed (Martens and Politis 2020), and studies have been performed to investigate the HDX of ribosomal nascent chains (Wales et al. 2022). Typical HDX-MS experiments cover time scales ranging from seconds to hours, with recent developments giving access to the millisecond timescale (Kish et al. 2023; Rob and Wilson 2009; Seetaloo and Phillips 2022; Svejdal et al. 2019; Wilson and Konermann 2003). The main drawback of HDX-MS experiments over HDX-NMR is spatial resolution: HDX-NMR probes the exchange of the protein at the level of the single residue, while HDX-MS measures deuterium incorporation at the level of the proteolytic peptides (typically 5-20 amino acids), making it challenging to connect the experimental data with molecular models or to cross-compare it with other experimental techniques (Jia et al. 2023; Kan et al. 2013; Lau et al. 2020b; Skinner et al. 2019; Stofella et al. 2022).

The Linderstrøm-Lang model (Linderstrøm-Lang 1955) describes the HDX of a residue as a two-step process, where local and global fluctuations, governed by opening/closing rates k_{op}/k_{cl}, allow the amide hydrogen to switch from a closed state NH_{cl} to an open state NH_{op}, where deuteration occurs with intrinsic exchange k_{int}:

Eq. 5.1 $$NH_{cl} \underset{k_{cl}}{\overset{k_{op}}{\rightleftarrows}} NH_{op} \xrightarrow{k_{int}} ND$$

Under the native ($k_{cl} \gg k_{op}$) and EX2 ($k_{int} \ll k_{cl}$) approximations, which hold for most residues and proteins, the deuterium uptake of each residue $d(t)$ can be modelled as a single exponential

Eq. 5.2 $$d(t) = 1 - e^{-k_{obs}t}$$

with exchange rate

Eq. 5.3 $$k_{obs} = \frac{k_{int}}{P}$$

The intrinsic exchange rate k_{int} depends on the nature of the amino acid and the chemical environment surrounding the residue, and can be predicted if the sequence of the protein is known together with the pH and temperature at which the labelling is performed (Bai et al. 1993; Connelly et al. 1993; Molday et al. 1972; Nguyen, Mayne, Phillips, and Englander 2018). The protection factor P in the Linderstrøm-Lang model is the ratio of the opening and close rates ($P \equiv k_{cl}/k_{op}$). Attempts to predict NMR measured protection factors from the knowledge of the folded structure of a protein and the chemical environment of each exchanging amide hydrogen have been moderately successful. For example, estimations based on counting the heavy contacts and hydrogen bonds made by the amide (Best-Vendruscolo model) correlate relatively well with the experiment: for example, residues that are deeply buried inside the protein generally have larger protection factors, exposed residues show lower values.

As HDX-NMR experiments monitor the decrease in ^1H signal (S_{NMR}) upon deuteration of individual residues, the raw data can be fitted with an exponential decay analogous to **Eq. 5.2**: $S_{NMR}(t) = e^{-k_{obs}t}$. Using **Eq. 5.3**, the protection factor of the residue P can be obtained by dividing the intrinsic rate k_{int}, which can be predicted, with the observed rate k_{obs}.

HDX-MS experiments monitor the deuterium uptake of a protein via its proteolytic peptides. The deuterium uptake D(t) of a peptide can be written as the sum of the contributions of its *n* exchangeable residues (i.e. excluding prolines and the first residue)

Eq. 5.4 $$D(t, \{P_i\}) = \frac{1}{n}\sum_i \left(1 - e^{-\frac{k_{int,i}}{P_i}t}\right)$$

where $k_{int,i}$ and P_i are the intrinsic exchange rate and the protection factors of residue i. In the ideal scenario of a 2-residue long peptide, **Eq. 5.4** is equivalent to **Eq. 5.2** (as the first residue is non-observable), and a protection factor can be easily derived. For this reason, several approaches have been proposed to experimentally increase the spatial resolution of HDX-MS data: the combination of multiple proteases (active at quench conditions, i.e. at low pH) can increase the digestion efficiency (Cravello et al. 2003; Kenji et al. 2005; Tsiatsiani et al. 2017), and the integration of fragmentation techniques that avoid H/D scrambling (such as ECD or ETD) has been proved useful to obtain sub-peptide resolution (Rand et al. 2014; Sobott 2020; Wollenberg et al. 2020).

However, typical HDX-MS experiments report peptides formed by 5-20 residues, hence obtaining protection factors from HDX-MS data by fitting **Eq. 5.4** to the experimental deuterium uptake of peptides is an underdetermined problem. In statistics, a problem is underdetermined when the number of parameters to be estimated is greater than the number of experimental data points. The consequence of underdetermination is degeneracy, i.e. the existence of multiple different solutions (i.e. patterns of protection factors) equally in agreement with experimental data. If we fit one peptide at a time, there is an infinite number of solutions in agreement with experimental data. If all peptides are fitted simultaneously, the degeneracy is reduced, i.e. there is a smaller infinity of solutions in agreement with experimental data (Skinner et al. 2019; Stofella et al. 2022).

These alternative solutions (sets of protection factors in agreement with experimental HDX-MS data) can be found by finding the minima of the cost function χ

Eq. 5.5 $$\chi(\{P_i\}) = \sum_j \sum_k \left[D_j^{pred}(t_k, \{P_i\}) - D_j^{exp}(t_k)\right]^2$$

where the index j runs over all peptides and k over all time points, $D_j^{exp}(t_k)$ is the experimental uptake of peptide j at time t_k, $D_j^{pred}(t_k, \{P_i\})$ is the deuterium uptake of peptide i at time t_k predicted using a set of protection factors $\{P_i\}$ in **Eq. 5.4**.

The cost function $\chi(\{P_i\})$ is a function of the protection factors which is zero when the measured peptide kinetics is exactly reproduced by the predicted uptake calculating using $\{P_i\}$. In the case of error-free data, the zeros of such cost function exist. Since HDX-MS data are affected by experimental errors, the solutions to our problem, i.e. the sets of protection factors in agreement with experimental HDX-MS data, correspond to the minima (rather than the zeros) of the cost function. Of course, only one minimum corresponds to the *true* set of protection factors of the protein.

The minimum of a function can be identified with the implementation of minimization algorithms, such as least-squares. If more minima are present, as in the case of the cost function in **Eq. 5.5**, the minimization can be repeated multiple times, starting from different initial guesses, to explore the different minima (**Figure 5.1**). We have recently developed a method, ExPfact, which simultaneously fits the information contained in the deuterium uptake curves of overlapping peptides to explore the different minima of the cost function in **Eq. 5.5**. Starting only from the knowledge of the uptake curves, the sequence of the protein, the temperature and pH at which the deuterium labelling was performed, ExPfact returns all possible patters in agreement with experimental HDX-MS data.

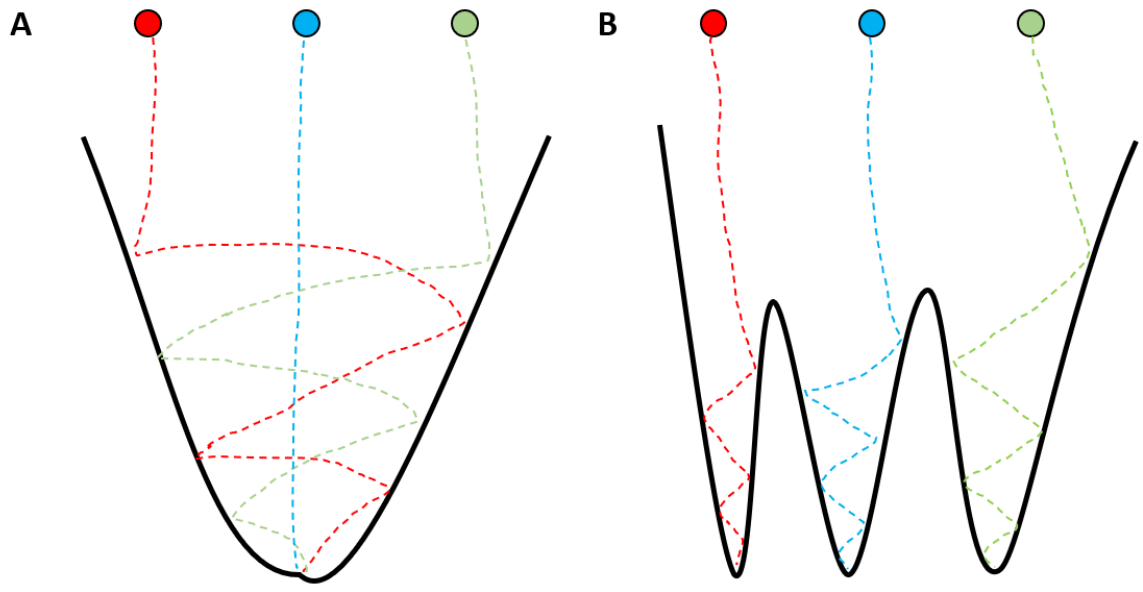


Figure 5.1. Schematic representation of a minimization procedure.
Schematic representation of a minimization procedure when different initial guesses are selected when one minimum (A) or multiple minima (B) are present.

Once the different minima have been identified, the problem translates into finding the *true* one (which does not necessarily correspond to the global minimum): the reliability of a prediction of protection factors from HDX-MS data relies on the availability of a reference benchmark NMR dataset. Here, we report on the measurement of two proteins by MS and NMR, keeping the experimental conditions the same as similar as possible. Two proteins were chosen for this study: ubiquitin, a well-characterized model protein (76 residues) that is known to exchange on a timescale that can be easily probed by both HDX-MS and NMR experiments (Bougault et al. 2004), and an original GB1-TACC3 complex (91 residues) which fuses the small model protein GB1 (59 residues) with a short disordered peptide sequence TACC3 (32 residues).

In this paper, we show that ambiguous protection factors can be inferred from peptide-level HDX-MS data by minimizing a cost function. The reliability of the predictions depends on the quality of the HDX-MS dataset, which can be defined by its redundancy (i.e. the average number of peptides covering each residue) and adaptive temporal

sampling (i.e. the number, spread and value of the labelling time points). The ambiguity of the prediction can be dramatically reduced if the search of the minima of the cost function is performed using an informed guess based on a known or a model structure. In this way the search of the minima of the cost function χ disfavors "unphysical" solutions (i.e. the search is biased against buried residues having low P or exposed residues having a high *P*).

5.3 Materials and methods

5.3.1 Materials

15-N labelled ubiquitin was purchased from UBPBio (product code: E1130), while unlabelled ubiquitin was purchased from Sigma Aldrich (product code: U6253). GB1-TACC3 was expressed and purified (see **section 5.3.2**). Deuterium oxide (99.8% D, product code: 1003690110) and d_4-urea (product code: 176087) were purchased from Sigma Aldrich. $^{15}NH_4Cl$ (99%, NLM-467) and ^{13}C D-glucose (99%, CLM-1396) were purchased from Goss Scientific. All other chemicals were of analytical grade or higher. Minimal media for $^{15}N/^{13}C$ expression of GB1-TACC3 consisted of 2 g/L of $^{15}NH_4Cl$ and 4 g/L of ^{13}C D-glucose in 50 mM Na_2HPO_4, 25 mM KH_2PO_4, 20 mM NaCl, 2 mM $MgSO_4$, 0.2 mM $CaCl_2$, 0.01 mM $FeSO_4$, supplemented with micronutrients and vitamins (BME vitamins, Sigma-Aldrich). The solution was syringe filtered (0.2 μm) prior to use.

5.3.2 Expression and purification of GB1-TACC3

The expression and purification of GB1-TACC3, as well as the NMR peak assignments (section 5.3.3) were performed by Dr Matthew Batchelor.

GB1-TACC3 is the small 'model' protein GB1 (Huth et al. 1997) fused to a short, disordered peptide sequence from human TACC3 (Uniprot: Q9Y6A5, residues Gly545 to Arg573). This hybrid sequence was cloned into a pETM6T1 vector which contains a TEV protease-cleavable N-terminal His-NusA tag. GB1-TACC3 was expressed in BL21 (DE3) RIL *Escherichia coli* cells. Cells were transformed using the heat-shock method.

50 μg/mL kanamycin was used to select and maintain the pET vector; 35 μg/mL chloramphenicol was used for routine maintenance of the RIL vector. A single colony was grown overnight in 50 mL of LB at 37 °C. This starter culture was used to seed 1 L of LB media. Cells were grown to an OD of ~0.6. For unlabelled expression, cells were then induced by addition of 1 mM ITPG and incubated overnight at 20 °C. For NMR-labelled proteins, LB cultures were grown to an OD ~ 0.6, cells were pelleted at 2500*g,* resuspended in PBS buffer to remove residual LB, pelleted again and resuspended in 250 mL of minimal media. After transferral to an autoclaved 2.5 L flask, cells were incubated at 20 °C for a further 2 h at 200 rpm and then overnight expression was induced with 1 mM IPTG. Cells were harvested by centrifugation and pellets stored at -80 °C.

Cell pellets were thawed and resuspended in TBS (20 mM tris, 150 mM NaCl, 2 mM b-mercaptoethanol, pH 8) supplemented with a c0mpleteTM Mini EDTA-free Protease Inhibitor Cocktail tablet (Roche). Cells were lysed by sonication and then clarified by centrifugation at 40,000*g* for 45 min. Protein was purified from the lysate using Ni-affinity chromatography with a HisTrap column; eluting from the column using a 0–500 mM gradient of imidazole. Two aliquots (~ 1 mg) of a solution of His-tagged wild type TEV protease (~ 28.5 kDa) stored in 50% glycerol were added to the collected NusA-His-(TEV)-GB1-TACC3 fractions (with an estimated 0.035:1 molar ratio of TEV:protein), and the solution was dialysed in TBS buffer at 4 °C overnight. The solution was subjected to Ni-affinity subtraction step to remove the cleaved His-NusA tag and His-tagged TEV protease. GB1-TACC3 protein was further purified and buffer-exchanged using size-exclusion chromatography (SEC) with a Superdex 16/600 S75 column (GE Healthcare). The final size exclusion buffer contained 20 mM $(K/H)_3PO_4$, 150 mM NaCl, pH 6.5. GB1-TACC3 was concentrated, flash frozen in liquid nitrogen and stored at -80 °C.

5.3.3 NMR peak assignments

Peak assignments for ubiquitin were available at 20°C. Spectra for assignment of GB1-TACC3 were recorded on a 600 MHz Oxford Instruments spectrometer equipped with a 5 mm Bruker QCI-P cryoprobe and a Bruker Avance III HD console, with data acquisition achieved using Topspin. Spectra were recorded using a 0.8 mM sample of $^{15}N/^{13}C$-labelled GB1-TACC3 sample (20 mM $(K/H)_3PO_4$, 150 mM NaCl, 5% D_2O, pH 6.5) at 10 °C. HNCO, HNcaCO, HNCA, HNcoCA, HNcaCB and HNcocaCB (HNC) triple-resonance assignment spectra and a HBHAcoNH and $^{1}H–^{13}C$ HSQC spectrum were recorded. Spectra were generated through processing data with NMRPipe/NMRDraw (Delaglio et al. 1995). CCPNmr Analysis v2.5 (Vranken et al. 2005) was used for peak assignments. Full *ab initio* assignment of the $^{1}H–^{15}N$ HSQC spectrum was achieved; the GB1 assignment subsequently being cross-referenced with known assignments. For both ubiquitin and GB1-TACC3, peak positions were manually tracked to transfer assignments to HSQC spectra recorded at 4 °C (temperature at which the HDX experiments were performed).

5.3.4 HDX-NMR experiments

To acquire the control sample (non-deuterated), 300 µL of protein solution (ubiquitin or GB1-TACC3) at concentration ~300 µM in equilibration buffer (20 mM tris, pH 7.0) was placed into an NMR tube. For the deuterated samples, 500 µL of protein solution (ubiquitin or GB1-TACC3) at concentration ~300 µM were concentrated up to ~3 mM by centrifugation at 14,000g for 20 min using Amicon® Ultra-0.5 3K centrifugal filter devices (Millipore, product code: UFC5003BK). 50 µL of sample were then diluted into 450 µl of labelling buffer (20 mM tris, pD 7.0) inside an NMR tube. The NMR tube was then transferred into a 950 MHz Bruker Ascend Aeon ™ NMR spectrometer. For the deuterated sample, 2D ^{1}H-^{15}N HSQC spectra were acquired as follows: 20 spectra were acquired for 4 scans (short acquisition: 4 min 34 s each), followed by 30 spectra acquired for 16 scans (medium acquisition: 17 min 30 s each), followed by 10 spectra acquired for 64 scans (long acquisition: 1 h 9 min 12 s each). Between the dilution of

labelling buffer in the protein solution and the beginning of the first acquisition a dead time of 21 min 39 s occurred for ubiquitin and 16 min 43 s for GB1-TACC3; this was caused by the time required to manual inject the sample and the labelling buffer into the NMR tube, manual transferring to the cooled sample changer, automatic transfer from the sample changer to the spectrometer, calibration and tuning of the spectrometer. The intensities as a function of labelling time were extracted using the Relaxation Analysis module in CCPNmr (Skinner et al. 2016).

5.3.5 HDX-MS experiments

HDX-MS experiments were carried out using an automated HDX robot (LEAP Technologies, Fort Lauderdale, FL, USA) coupled to an M-Class Acquity LC and HDX manager (Waters Ltd., Wilmslow, Manchester, UK). 6 µL of protein solution containing ~20 µM protein in equilibration buffer (20 mM tris, pH 7.0) was added to 54 µL deuterated buffer (20 mM tris in D_2O, pD 7.0) and incubated at 4 °C for 0.5, 1, 2, 5, 10, 30, 60, 120, 300 and 480 min. Following the labelling reaction, samples were quenched by adding 40 µL quench buffer (100 mM tris, 2 M urea) to 40 µL of the labelled solution giving a final quench pH = 2.5. 60 µL of quenched sample was passed through an immobilised ethylene-bridged hybrid (BEH) pepsin column (Waters Ltd., Wilmslow, Manchester, UK) at 200 µL min^{-1} (20 °C) and a VanGuard Pre-column Acquity UPLC BEH C18 (1.7 µm, 2.1 mm × 5 mm, Waters Ltd., Wilmslow, Manchester, UK) for 3 min. The resulting peptic peptides were transferred to a C18 column (75 µm × 150 mm, Waters Ltd., Wilmslow, Manchester, UK) and separated by gradient elution of 0–40% MeCN (0.1% v/v formic acid) in H_2O (0.3% v/v formic acid) over 12 min at 40 µL min^{-1}. Trapping and gradient elution of peptides were performed at 0°C. The HDX system was interfaced to a Synapt G2Si mass spectrometer (Waters Ltd., Wilmslow, Manchester, UK). $HDMS^E$ and dynamic range extension modes (data independent analysis (DIA) coupled with ion mobility separation) were used to separate peptides prior to CID fragmentation in the transfer cell. Fully deuterated samples were acquired as follows: 30 µL of protein sample at ~20 µM was dried in a vacuum concentrator for 15 min; the

protein was then reconstituted in a fully deuterated buffer (same as the labelling buffer, but with addition of 4M D4-urea); the solution was vortexed for 30 s and left to incubate overnight on ice before injection into the mass spectrometer via the robot (using the same procedure described earlier). Experiments were replicated up to 4 times. HDX data were analysed using PLGS (v3.0.2) and DynamX (v3.0.0) software supplied with the mass spectrometer. The fractional uptake of peptides was calculated:

Eq. 5.6 $$D(t) = \frac{M(t) - M(0)}{M(FD) - M(0)}$$

where $M(t)$ is the mass of the peptide at labelling time t, $M(0)$ is the mass of the peptide in the control sample (without deuteration), and $M(FD)$ is the mass of the peptide in the fully deuterated control.

5.3.6 Estimating protection factors from HDX-MS data

Protection factor analysis was performed using ExPfact (Skinner et al. 2019; Stofella et al. 2022)**.** According to the Linderstrøm-Lang model (Linderstrøm-Lang 1955), the deuterium uptake $D(t)$ of a peptide formed by N exchangeable residues (i.e. excluding the first residue of the peptide and all prolines), can be written as a sum of single exponentials (**Eq. 5.4**). In analogy to the HDX-NMR data analysis, intrinsic exchange rates were calculated using Englander's predictions (Bai et al. 1993; Connelly et al. 1993; Molday et al. 1972; Nguyen, Mayne, Phillips, and Englander 2018). ExPfact implements a non-linear fitting algorithm that simultaneously fits the multiexponential in **Eq. 5.4** to the fractional uptake (**Eq. 5.6**) of all peptides and adjusts the protection factors of the protein to minimize a cost function (squared difference between predicted and experimental data).

Inferring single-residue resolution from peptide-level HDX-MS data is an underdetermined problem (the number of parameters to be estimated is greater than the number of experimental points available), and this causes ExPfact to return ambiguous solutions (yet equally in agreement with experimental data) depending on

the starting point selected for the minimization algorithm. Different initial guesses were used as a starting point for the fitting algorithm:

i) Random search (Salmas and Borysik 2021; Skinner et al. 2019): 10,000 sets of protection factors were randomly generated with the constraint $0 \leq \ln(P) \leq 20$ and the best one (i.e. the one with lowest cost function χ) was used as initial guess for the fitting algorithm. This random search is repeated 100 times and returned ambiguous sets of protection factors in agreement with experimental HDX-MS data. Average values and standard deviations were reported.

ii) Structurally driven: the initial guess was extracted from a protein structure, either a PDB structure (Berman et al. 2000) or an AlphaFold model (Jumper et al. 2021), using the phenomenological model (Best and Vendruscolo 2006; Radou et al. 2014; Vendruscolo et al. 2003) (see **section 5.3.7**). 100 sets of protection factors are generated in the surroundings of the calculated set of protection factors by allowing a Gaussian fluctuation on the estimated $\ln(P_i)$ ($\sim N(\mu = \ln(P_i), \sigma = 0.5)$). A minimization was performed starting from each of these sets. Average values and standard deviations were reported.

iii) Unstructured (Smit et al. 2020): all protection factors were initially set to 1, which corresponded to the hypothesis that the protein is completely unstructured. Analogously to the structurally driven initial guess, 100 sets of protection factors were generated in the surroundings of the set of protection factors by allowing a Gaussian fluctuation on the estimated value and a minimization was performed starting from each of these sets. Average values and standard deviations were reported.

iv) Data driven (Gessner et al. 2017): the deuterium uptake curves of peptides were individually fitted with a single exponential model. The extracted rate was used as initial guess for all the residues belonging to the peptide. An average rate was considered when a residue was covered by multiple peptides. Analogously to the structurally driven and unstructured initial

guesses, 100 sets of protection factors were generated in the surroundings of the set of protection factors by allowing a Gaussian fluctuation on the estimated value and a minimization was performed starting from each of these sets. Average values and standard deviations were reported.

6.3.7 Comparing protection factors from MS and NMR

Protection factors estimated by HDX-MS experiments (using either strategy) were compared with the reference set of protection factors obtained from NMR using Pearson's correlation coefficient (Kirch 2008). To account for the error bars in the protection factor estimates from MS, which arise from multiple minimizations, correlation coefficients were calculated for the individual runs and an average value was reported. Alternatively, a weighted correlation coefficient could be calculated considering the inverse of the standard deviation as weight for the individual protection factor. Following the traditional approach to interpret correlation coefficients (Schober, Boer, and Schwarte 2018), we classify $\rho > 0.9$ as "very strong correlation", $0.7 < \rho \leq 0.9$ as "strong correlation", $0.4 < \rho \leq 0.7$ as "moderate correlation", $0.1 < \rho \leq 0.4$ as "weak correlation", and $\rho \leq 0.1$ as "negligible correlation". The expected correlation between protection factors extracted from the two experiments is *strong* rather than *very strong*. This is partly due to differences in the time ranges each technique probes: HDX-MS captures exchange over a period from 30 s to 8 h, while HDX-NMR covers exchange from approximately 20 minutes (dead time) up to 48 h. Moreover, the experimental data suffer different levels of back-exchange: HDX-NMR experiments are only affected by back-exchange during the labelling phase, whereas HDX-MS experiments have additional back-exchange occurring at the protein level from quench to digestion and at the peptide level from digestion onwards. While these experimental differences might introduce a bias in the absolute value of the estimated protection factors, similar patterns should be consistently identified by both experiments (for instance, both techniques should indicate higher protection factors in structured regions).

5.3.8 Protection factors from PDB structure

The structure used for ubiquitin was acquired by x-ray diffraction (PDB code: 1UBQ) (Vijay-Kumar, Bugg, and Cook 1987), while the structure for GB1-TACC3 was predicted using AlphaFold (Jumper et al. 2021). The protection factor of residue i in protein conformation X can be estimated using the Best-Vendruscolo model (Best and Vendruscolo 2006; Radou et al. 2014; Vendruscolo et al. 2003):

Eq. 5.7 $$\ln P_i^{sim}(X) = \beta_c N_i^c(X) + \beta_h N_i^h(X)$$

where N_i^c and N_i^h represent the number of contacts with heavy atoms (i.e. non hydrogens) and hydrogen bonds residue i is involved in. The parameters $\beta_c = 0.35$ and $\beta_h = 2$ were previously optimized on a dataset of 7 proteins for which experimental protection factors were available (Best and Vendruscolo 2006).

5.4 Results

5.4.1 Reference protection factors from NMR

The NMR spectra show a decrease of the residue peak intensities as a function of the labelling time (**Figure 5.2A**). Experimental data were fitted with a single exponential model (**Figure 5.2B**). The Linderstrøm-Lang model (Linderstrøm-Lang 1955) expresses the extracted rate k_{obs} as the ratio of the intrinsic exchange rate k_{int} and the protection factor P_i (**Eq. 5.3**). The intrinsic exchange rate was calculated using Englander's predictions (Bai et al. 1993; Connelly et al. 1993; Molday et al. 1972; Nguyen, Mayne, Phillips, and Englander 2018), and the protection factors of the assigned residues were extracted from the experimental rates (**Figure 5.2B** and **Supplementary Figure 5.2**). The protection factor pattern extracted for ubiquitin correlates qualitatively well with the secondary structure of the protein (**Figure 5.2C**): higher protection factors are associated with beta sheets and alpha helices, while lower protection factors are associated with loops. For ubiquitin, we checked that the extracted protection factors agreed with previously published NMR data (Bougault et

al. 2004), finding a correlation coefficient $\rho = 0.88$ (**Supplementary Figure 5.3**). Protection factors for GB1-TACC3 were obtained analogously (**Supplementary Figure 5.4** and **Supplementary Figure 5.5**). Protection factors from HDX-NMR were used as reference to benchmark protection factors extracted from HDX-MS data.

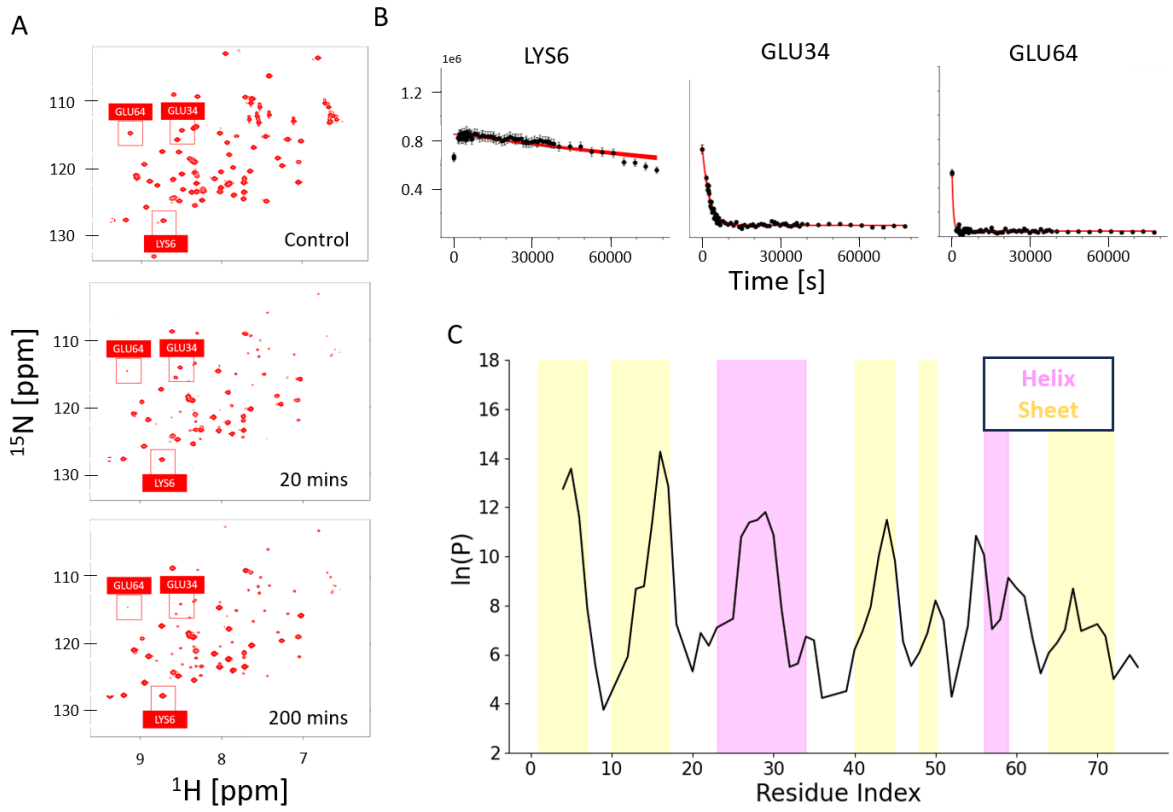


Figure 5.2. Results for HDX-NMR for ubiquitin.
(A) Top panel: Raw NMR data: 2D HSQC spectra shown for the control sample (no deuteration), and after 20 min and 200 min of incubation with labelling buffer. The assignments for LYS6, GLU34 and GLU64 are highlighted. (B) Intensities extracted from the raw NMR data for LYS6 (slow exchanging), GLU34 (medium exchanging) and GLU64 (fast exchanging) are shown as a function of the labelling time. The curve is fitted with a single exponential model (red line). (C) Protection factors extracted from HDX-NMR data compared with helices (pink) and sheets (yellow).

5.4.2 Peptide-level HDX-MS

The HDX-MS experiments provided the coverage maps reported in **Figure 5.3** for the two proteins, reporting an average redundancy of 10.9 for ubiquitin and 13.4 for GB1-

TACC3. Fractional uptake curves of all peptides are shown for ubiquitin (**Supplementary Figure 5.6**) and GB1-TACC3 (**Supplementary Figure 5.7**).

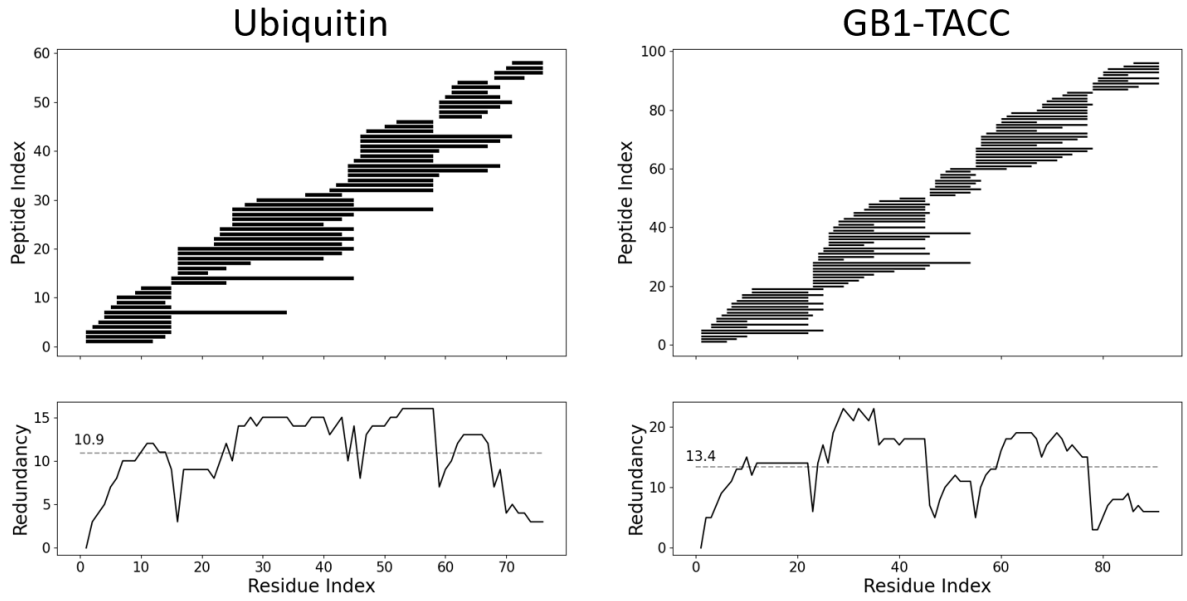


Figure 5.3. Coverage map and per-residue redundancy provided by the HDX-MS experiments. Coverage map and per-residue redundancy provided by HDX-MS experiments for ubiquitin (left) and GB1-TACC3 (right). For ubiquitin, 58 peptides were identified, with a coverage of 100% and an average redundancy of 10.9; for GB1-TACC3, 96 peptides were identified, with a coverage of 100% and an average redundancy of 13.4.

5.4.3 Compatibility between MS and NMR

We evaluated the overall compatibility between MS and NMR experiments. To do so, we used the protection factors extracted by NMR as initial guess for the minimization algorithm, obtaining a correlation coefficient $\rho = 0.82$ for ubiquitin and $\rho = 0.90$ for GB1-TACC3 (**Supplementary Figure 5.8**). We compared this cross-technique correlation obtained for ubiquitin ($\rho = 0.82$) with the cross-laboratory correlation obtained by comparing our NMR protection factors with the ones obtained by Bougault et al (Bougault et al. 2004), which yielded $\rho = 0.88$ (**Supplementary Figure 5.3**). Ideally, one would expect a *perfect* correlation coefficient ($\rho \sim 1.00$), but experimental errors lower this value. Thus, these correlation coefficients ($\rho = 0.82$ for ubiquitin and

$\rho = 0.90$ for GB1-TACC3) serve as a 'maximum correlation reference', i.e what quality of prediction to expect in a best-case scenario.

5.4.4 Agreement between Best-Vendruscolo model and NMR

We assessed the agreement between the protection factors estimated using the phenomenological approximation (**Eq. 5.7**) and the reference NMR measurements. The results (**Supplementary Figure 5.9**) reported a correlation coefficient $\rho = 0.61$ for ubiquitin and $\rho = 0.75$ for GB1-TACC3. These relatively low values are consistent with previously reported studies (Claesen and Politis 2019) which have underscored the limitations of the phenomenological approximation in accurately interpreting the physical meaning of the protection factors of a protein.

5.4.5 Strategy 1: random search

We performed the protection factor analysis for HDX-MS data using no prior information. The results in **Figure 5.4** show that the minimization, initialized with different random initial guesses, converged to ambiguous sets of protection factors (grey lines). The minimization was repeated 100 times and average protection factors were reported with their standard deviations (black line). The standard deviations represent a measure of the underdetermination of the problem: the more ambiguous the protection factor, the higher the standard deviation. The protection factors obtained by HDX-MS data using this random search strategy were compared with protection factors extracted by NMR (red line), reporting a strong correlation in both cases: $\rho = 0.78$ for ubiquitin and $\rho = 0.88$ for GB1-TACC3.

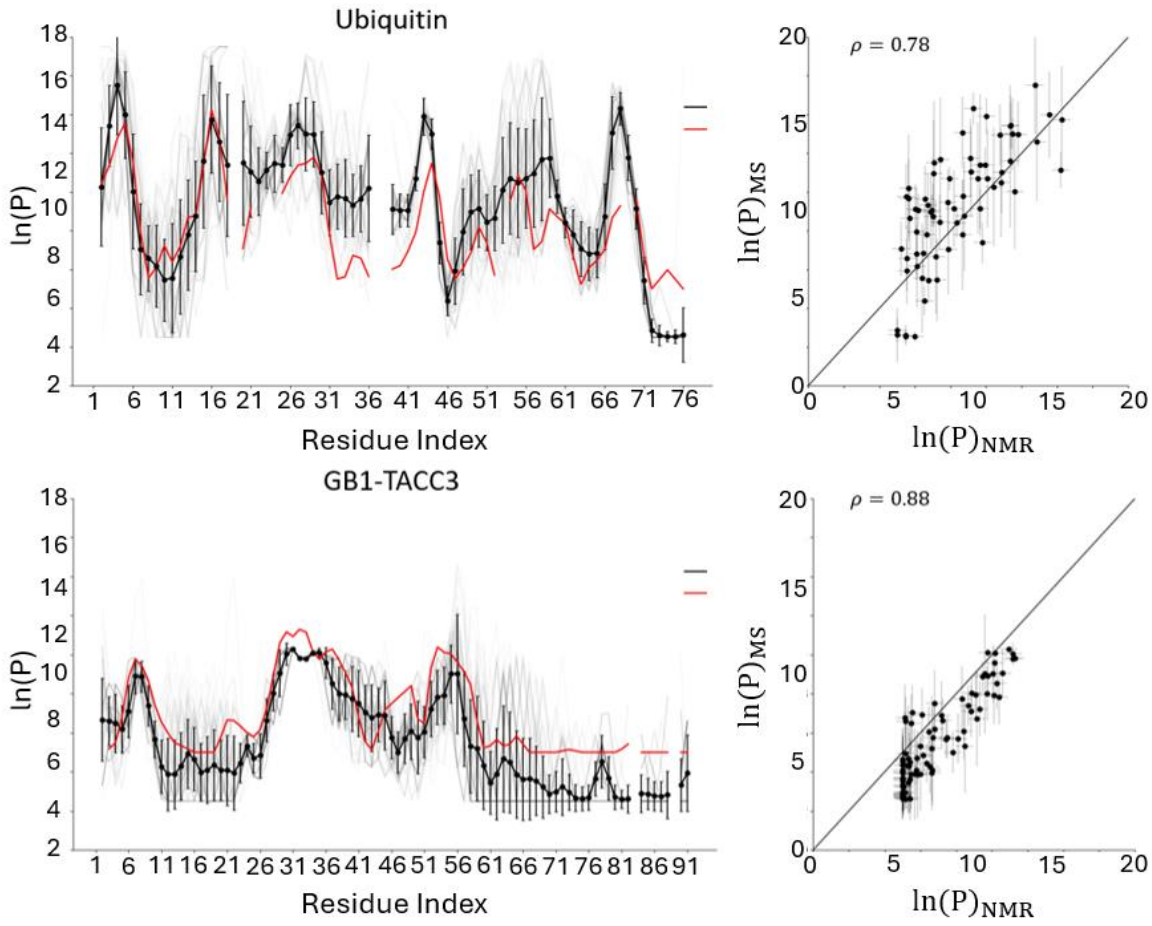


Figure 5.4. Protection factors estimated from HDX-MS data using a random initial guess. Protection factors estimated from HDX-MS data using a random initial guess (black) are compared with protection factors extracted from NMR data (red). The fitting procedure was repeated 100 times (grey lines), and average values and standard deviation are reported. On the right, a correlation plot shows a correlation coefficient $\boldsymbol{\rho = 0.78}$ for ubiquitin (top) and $\boldsymbol{\rho = 0.88}$ for GB1-TACC3 (bottom).

5.4.6 Strategy 2: structurally driven initial guess

Starting from a 3D structure of the protein, we estimated the protection factors of the protein using the Best-Vendruscolo model (**Eq. 5.7**). Allowing a Gaussian error on the estimated protection factor $\ln(P_i)$ ($\sim N(\ln(P_i), \sigma = 0.5)$), we randomly sampled 20 initial guesses in the surrounding of the set estimated by the protein structure. These initial estimates were then optimized on the HDX-MS data, and average and standard

deviations were reported (**Figure 5.5**). The standard deviations associated with the protection factors estimated from HDX-MS data using the structurally driven initial guess are smaller than the ones provided by the random search shown in **Figure 5.4**, showing that the choice of an initial guess helps avoiding unphysical minima, hence reducing the ambiguity on the estimated pattern of protection factors. The protection factors extracted by HDX-MS data using the structurally driven initial guess were compared with protection factors from NMR, reporting a correlation coefficient $\rho = 0.82$ for ubiquitin and $\rho = 0.83$ for GB1-TACC3. The choice of an initial guess reduces the ambiguity on the optimized pattern of protection factors (grey lines).

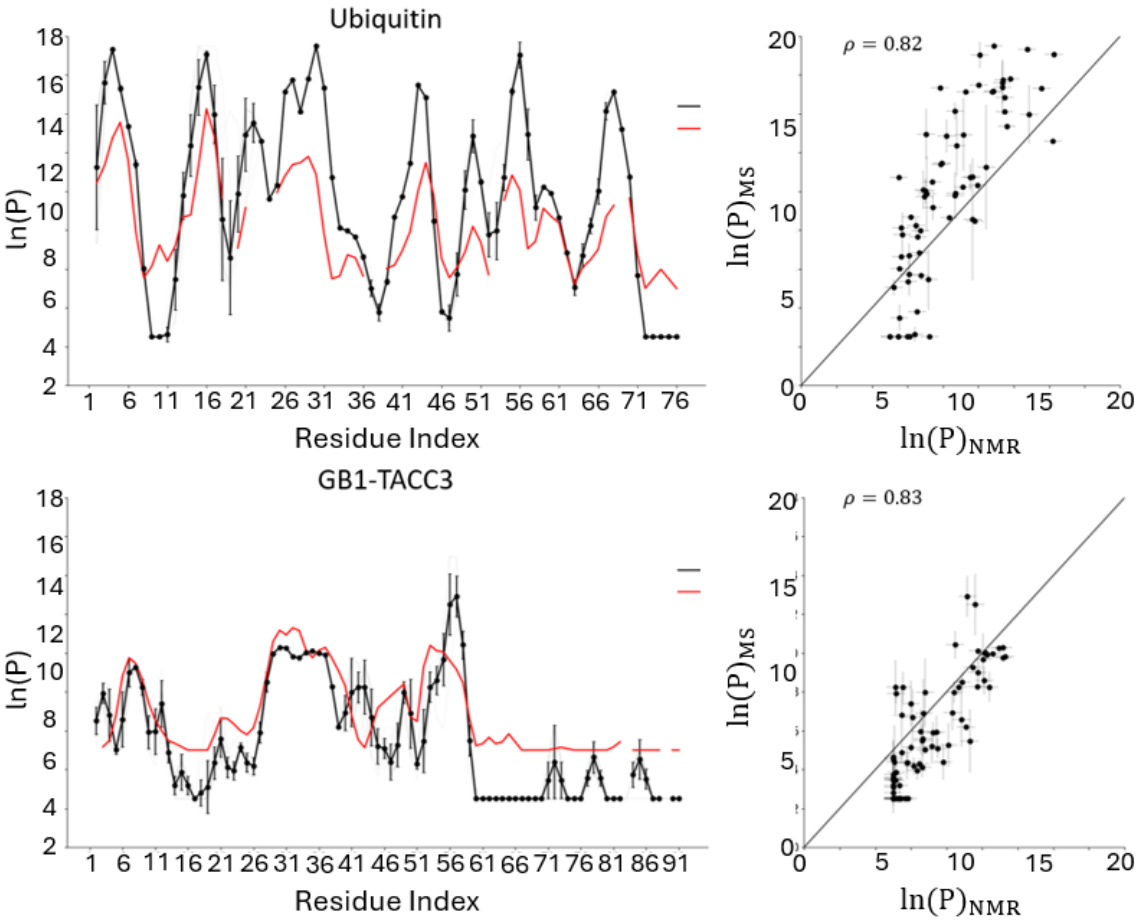


Figure 5.5. Protection factors estimated from HDX-MS data using a structure driven initial guess.

Protection factors estimated from HDX-MS data using a structurally driven initial guess (black) are compared with protection factors extracted from NMR data (red). The fitting procedure was repeated 20 times (grey lines), allowing a $\pm\mathbf{0.5}$ Gaussian error on the estimated provided by the phenomenological approximation, and average values and standard deviation are reported. On the right, a correlation plot shows a correlation coefficient $\boldsymbol{\rho = 0.82}$ for ubiquitin (top) and $\boldsymbol{\rho = 0.83}$ for GB1-TACC3 (bottom).

5.4.7 Strategy 3 and 4: unstructured and data-driven initial guess

We tested the behaviour of the algorithm when starting from two additional initial guesses: the unstructured initial guess and the data-driven initial guess (see **section 5.3.6**). By comparison with reference NMR data, the unstructured initial guess reported a correlation coefficient $\rho = 0.78$ for ubiquitin and $\rho = 0.81$ for GB1-TACC3, while the data-driven initial guess showed $\rho = 0.78$ for ubiquitin and $\rho = 0.76$ for GB1-TACC3. The results from the four strategies reported in all cases a strong correlation with NMR data.

The distribution of the cost function χ (**Eq. 5.5**) before and after the minimization for the four different approaches is reported in **Figure 5.6**. The shape of the distribution after minimization shows the choice of an initial guess reduces the ambiguity on the final solution provided by the minimization algorithm by exploring a more localized area of the landscape of the cost function, with the sharpest peak provided by the structurally driven initial guess.

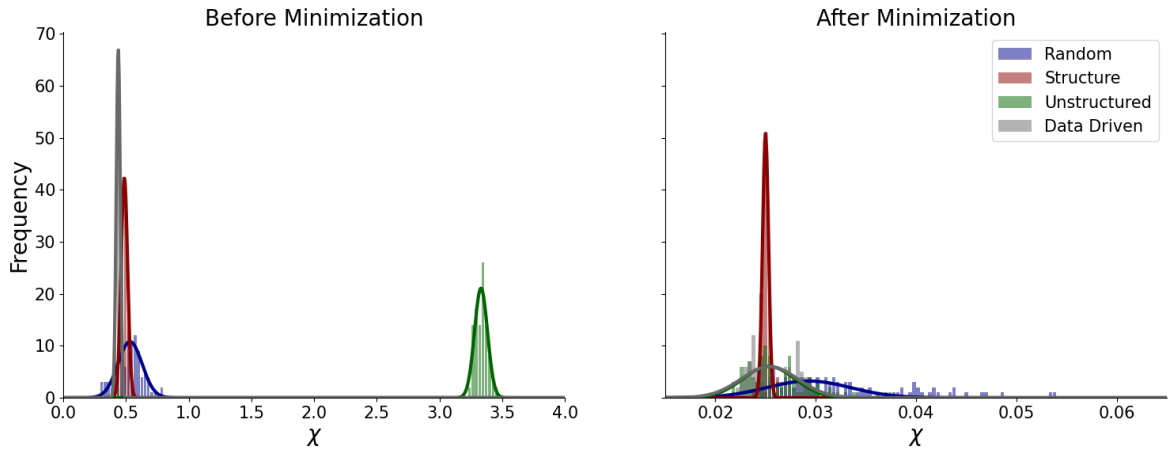


Figure 5.6. Frequency distribution of the cost function χ before and after minimization.
Frequency distribution of the cost function χ (**Eq. 5.5**) before and after minimization using four the four different approaches to evaluate an initial guess described in Methods: random (blue), structurally driven (red), unstructured (green) and data driven (grey).

5.4.8 The structurally driven initial guess provides robust estimates

We tested which of the four approaches was more robust with respect to the quality of the HDX-MS dataset. To vary the quality of the HDX-MS dataset, we removed peptides while maintaining the average length of peptides within a maximum variation of one amino acid. This allowed to generate several sub-datasets, containing 20, 40, 60 and 80% of the peptides of the full datasets of ubiquitin and GB1-TACC3. The protection factors extracted from the different sub-datasets using the four approaches were then compared with NMR data (**Figure 5.7**).

The results (**Supplementary Figure 5.10**) showed that the approach that best correlated with NMR data is the structurally driven initial guess (average correlation coefficient over the sub-datasets: $\bar{\rho} = 0.78$), followed by the random search ($\bar{\rho} = 0.75$). The data-driven initial guess and the unstructured initial guess ($\bar{\rho} = 0.69$ for both) provided less robust results upon peptide removal. As expected, the correlation between protection factors obtained from HDX-MS data and reference NMR protection factors generally improves with the number of peptides available (**Figure 5.7A**). We converted the percentage of peptides into a redundancy scale and plotted the correlation coefficients for both proteins in every sub-dataset (**Figure 5.7B**). We used a linear model to fit the results and concluded that a redundancy of 4 is sufficient to have a correlation coefficient greater than 0.70.

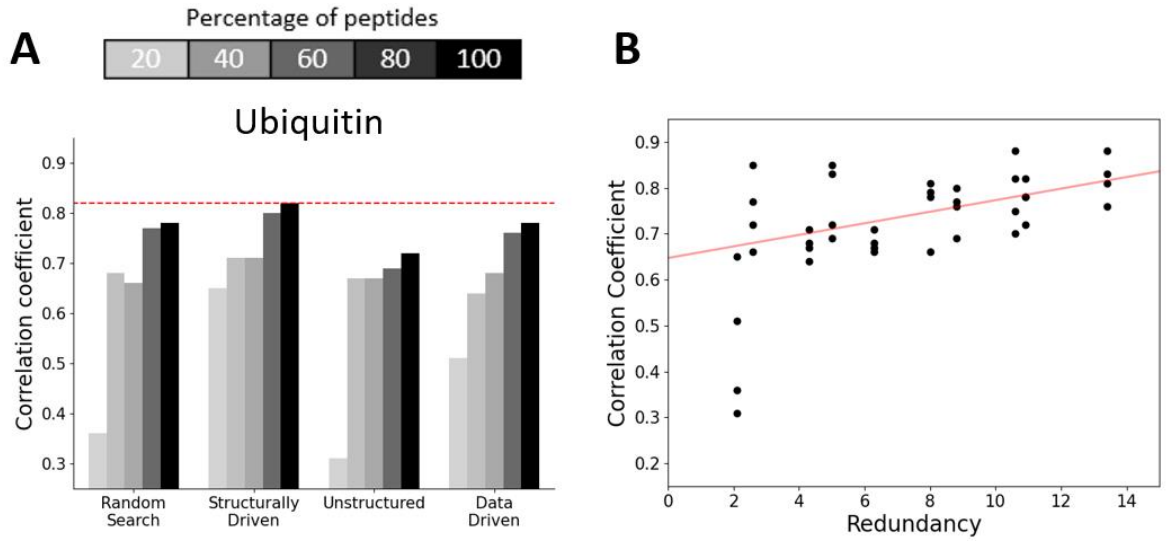


Figure 5.7. Comparison of protection factors extracted by HDX-MS data against reference NMR protection factors as a function of the peptide map quality.

(A) Correlation coefficients obtained from HDX-MS data using 20/40/60/80/100% peptides of the dataset (grey-scale) are shown while varying the initial guess on the protection factors (random search, structurally driven, unstructured, data driven). Results are shown for ubiquitin (compare with Supplementary Figure 5.9). (B) The correlation coefficients are shown as a function of the redundancy of the sub-dataset considered. Results for both proteins are combined and fitted with a linear model (red line).

5.4.9 Benchmarking against the number of time points

In analogy to the previous section, we analysed the outcomes as a function of the number of time points available. Starting from a rich dataset with 10 timepoints, we generated several sub-datasets where time points were dropped (first one at a time, then two at a time, then three, up to 8). The redundancy was preserved by using datasets with all overlapping peptides for these timepoints. We applied the minimization using the structurally driven initial guess and we compared the extracted protection factors with the NMR reference. The results for ubiquitin showed that the average correlation over the combinations increased with the number of time points considered (as expected), and 4 time points were sufficient to have a correlation

coefficient > 0.75 (**Figure 5.8A**). Which time points are more important? To show this, we calculated the average correlation coefficient over all possible combinations removing at least that specific point (**Figure 5.8B**). The results for ubiquitin show that the correlation coefficient significantly decreases when the longer time points are removed. What part of the protein's sequence is more affected by omitting earlier/later time points? We selected two scenarios: i) removing the first 5 time points, ii) removing the last 5 time points, and we calculated the difference between the protection factors extracted using the full dataset or the sub-dataset in either scenario. For ubiquitin (**Figure 5.8C**), the structured parts of the proteins, especially the α-helix covering residues 23-34 and the β-sheet covering residues 40-45, were the most affected when later time points were removed (scenario ii). The removal of the earlier timepoints (scenario i) does not have much influence on the predictions.

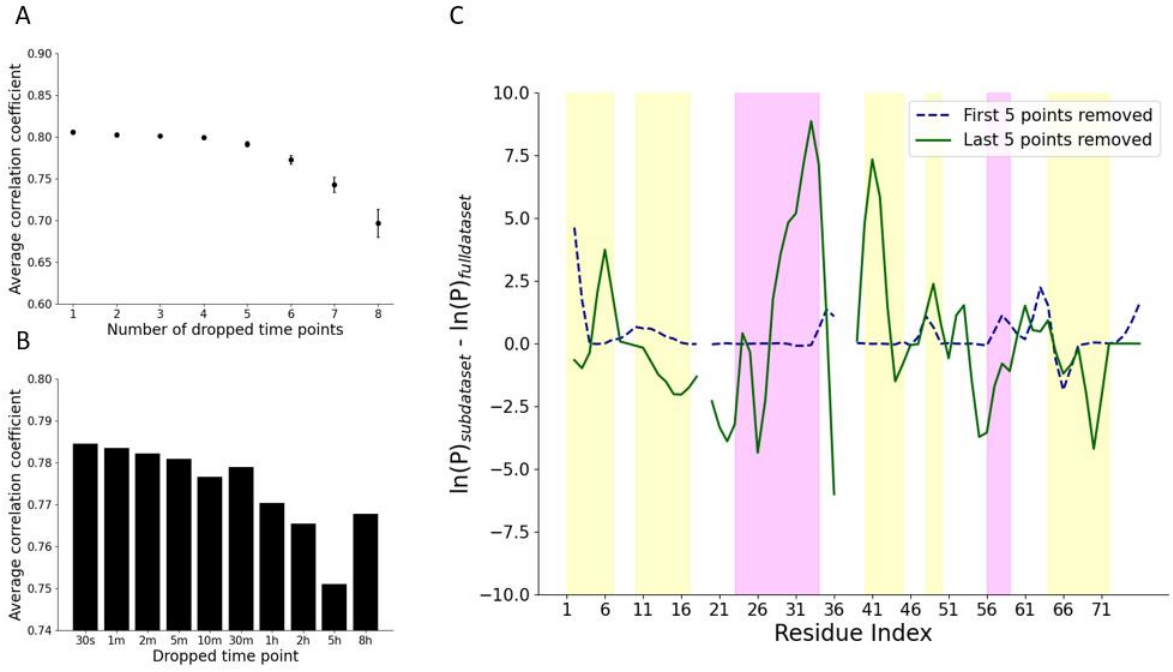


Figure 5.8. Dependence of the quality of protection factor predictions from HDX-MS data for ubiquitin as a function of the number and type of time points in the dataset.

The structure driven initial guess was used with the full dataset (100% peptides). (A) The correlation coefficient is shown as a function of the number of dropped time points. The average value is calculated

over all possible combinations dropping that specific number of time points. (B) The correlation coefficient is depicted as a function of the specific dropped time point. The value is averaged over all combinations containing the dropped time point. (C) The difference between the extracted protection factors using the full dataset (10 time points) and two scenarios is shown: i) first five time points removed (blue), ii) last five time points removed (green).

As testified by the results for GB1-TACC3 (**Supplementary Figure 5.11**), the choice of the optimal set of time points is highly protein/structure dependent. For example, the disordered part of GB1-TACC3 (covering residues 63-91) has already completely exchanged after the earliest time point collected by the HDX-MS experiment (30 s) and is therefore not properly sampled. The Best-Vendruscolo model (**Eq. 5.7**) can be used prior to the experiment to guide the decision on the labelling time points to be collected, for example to predict the optimal shortest time point. Using the protection factors calculated from the protein structure and the intrinsic exchange rates predicted using Englander's calculations, the deuterium uptake curve of each residue can be predicted, and a $t_{0.3}$ (i.e. the time at which the residue deuteration reaches 0.3) can be estimated (**Figure 5.9**). We showed the $t_{0.3}$ (rather than a $t_{0.5}$) as this choice of minimal time point would allow to properly sample $\sim 2/3$ of the uptake curve (rather than $1/2$). For ubiquitin, approximately 60% of the residues reach their $t_{0.3}$ within the time range sampled by our HDX-MS experiment (with 10 time points between 30 s and 8 h), and only one residue (namely Ala46) has a $t_{0.3}$ lower than 30 s. For GB1-TACC3, the predictions show that approximately one third of the protein (the disordered region covering residues 63-91) has a $t_{0.3}$ lower than 30 s, and therefore millisecond HDX is required to properly sample the behaviour of this specific protein region.

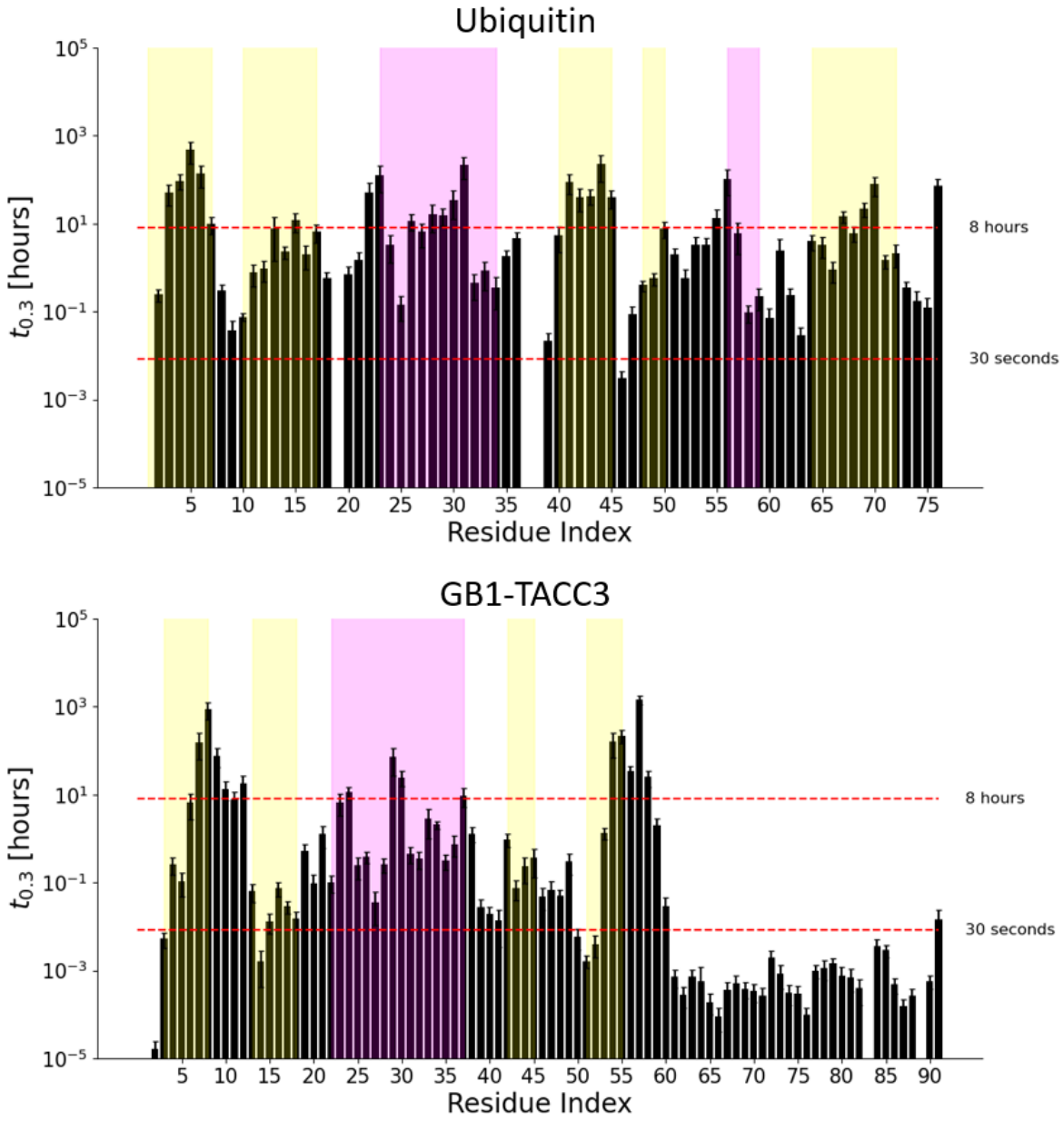


Figure 5.9. Optimum time point predictions.

The deuterium uptake curve for each residue is calculated using the protection factors extracted by the protein structure using the Best-Vendruscolo model (**Eq. 5.7**) and Englander's intrinsic exchange rate calculations. The $\boldsymbol{t_{0.3}}$, i.e. the time at which the uptake curve reaches 0.3, is shown as a function of the residue index for ubiquitin (top) and GB1-TACC3 (bottom). A Gaussian error of $\pm \mathbf{0.5}$ units on the protection factors extracted by the protein structure is allowed. The time range sampled by the HDX-MS experiments performed in this paper is shown with red dashed lines.

5.4.10 Example of single residue resolution

We show that the strategy provided to acquire protection factors from HDX-MS data is capable of highlighting residue-level differences such as jumps in the protection factors along the peptide sequence, in agreement with NMR. The deuterium uptake curve of a single proteolytic peptide may contain information about the dynamical properties of that specific amino acid sequence within the protein. In the example peptide taken from the ubiquitin dataset and shown in **Figure 5.10A**, the measured fractional uptake increases from 54% at 30 s to 64% at 8 h,, which suggests that ~ 50% of the residues exchange fast (before 30 s) and ~ 40% exchange slow (after 8 h). The hypothesis is compatible with the location of the peptide within the structure of ubiquitin (**Figure 5.10B**), as it covers a loop (fast exchanging) and parts of two β-sheets (slow exchanging). A simultaneous fitting of all peptides, as implemented here using the structure driven approach, is able to capture sub-peptide features: residue level jumps in the protection factors can be identified, as for residues 13-15, or swapped, e.g. residues 10 and 11 (**Figure 5.10C**).

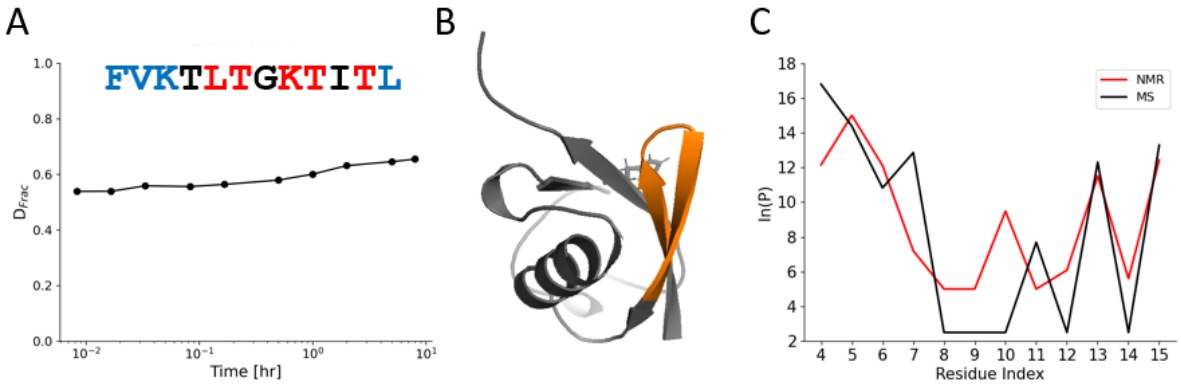


Figure 5.10. Example of sub-peptide resolution intrinsically contained in HDX-MS data.
(A) A sample peptide, covering residues 4-15 of ubiquitin, has a fractional uptake of 0.54 after 30 s, which increases to 0.64 after 8 h, suggesting that there is a subset of fast-exchanging residues and a subset that is slow-exchanging. The sequence of the peptide is shown, with HDX-NMR fast-exchanging residues highlighted in blue, slow-exchanging residues in red, medium-exchanging residues highlighted in black. (B) Structurally, the peptide (orange) covers a loop (fast-exchanging) and two sheets (slow-exchanging).

(C) Protection factors for residues 4-15 extracted from HDX-MS data using a structure driven initial guess (black) are compared with protection factors extracted by NMR (red).

5.5 Discussion

We have studied the HDX kinetics of two related proteins, well-structured ubiquitin and GB1-TACC3 with a disordered sequence, under identical experimental conditions (pH 7.0 and temperature 4°C) and cross-compared the results obtained from peptide-level MS data with residue-level NMR data. We performed a protection factor analysis that we had previously published (Skinner et al. 2019; Stofella et al. 2022) to robustly show that single residue information can be inferred from peptide-level HDX-MS data. The protection factors obtained from HDX-MS data were found to be equivalent to (i.e. with a strong correlation with) those derived from NMR experiments (which we consider to be the "true" reference). Obtaining single residue information from peptide level data is inherently an underdetermined problem, and our method provides the alternative ambiguous solutions that align with experimental data. We showed that the introduction of an informal initial guess into the minimization algorithm can reduce this ambiguity, and that the accuracy and precision of the predictions are influenced by the quality of the dataset, which is mainly defined by its redundancy and its capacity to effectively sample the uptake curves of peptides – a property we refer to as "adaptive sampling".

The problem of extracting protection factors from HDX-MS data translates into finding the minima of the cost function in **Eq. 5.5**. The redundancy provided by the HDX-MS experiments is sufficiently high to extract a unique pattern of protection factors, reporting a strong correlation ($\rho > 0.75$) when compared to NMR data, independently of the starting point of the minimization procedure. Indeed, we tested four different initial guesses (random, structure driven, unstructured and data-driven, see **section 5.3.6**). For ubiquitin, we found a correlation coefficient $\rho = 0.78$ for the random search, the unstructured initial guess and the data-driven initial guess, and $\rho = 0.82$ for the structurally informed initial guess. For GB1-TACC3, $\rho = 0.88$ for the random search,

$\rho = 0.83$ for the structurally informed guess, $\rho = 0.76$ for the unstructured initial guess, and $\rho = 0.81$ for the data-driven initial guess. The strong agreement and its independence on the initial guess are due to the high redundancies of the HDX-MS datasets (**Figure 5.3**): the more overlapping peptides are available, the more the cost function landscape assumes a single funnel shape (**Figure 5.1A** and **Figure 5.6**).

With lower redundancies, the different initial guesses are not equivalent. We checked how gradually removing peptides from the coverage maps of the two proteins influenced the outcomes of these four strategies. The results (**Figure 5.7** and **Supplementary Figure 5.10**) showed that the structurally driven initial guess and the random search strategies provided more robust predictions than the other two approaches. The fewer peptides are available, the more rugged the cost function landscape becomes (**Figure 5.1B**). The presence of multiple minima might force a specific initial guess to converge to a local minimum which does not correspond to the true NMR reference. The random search and the structure driven initial guess provide better agreement with NMR data because the former is able to widely sample the cost function landscape, while the latter samples a physically meaningful local minimum.

Hence, we showed that it is possible to infer single-residue information from peptide-level HDX-MS data. However, these predictions depend on the quality of the dataset available. But how can we define the quality of an HDX-MS dataset? In other words, when should the results of such analysis be considered robust? For the two proteins available, we quantified that a redundancy > 4 was required to achieve a strong correlation ($\rho > 0.70$) between MS and NMR data (**Figure 5.7B**). As the sequences of our two test proteins contained many common structural motifs including an extended disordered region, we are confident that we can generalize this statement. We recommend to limit this kind of analysis to datasets with redundancy > 4, or to those regions of the protein with redundancy > 4. This poses some limitations on the biological systems for which single residue resolution can be extracted from peptide level data. For example, high redundancy can routinely be achieved with relatively

small yet intriguing systems such as intrinsically disordered proteins (Keppel and Weis 2015; Parson, Jenkins, and Burke 2022; Seetaloo and Phillips 2022; Stephens et al. 2020; Zhu et al. 2015) or recombinant antigens (Calvaresi et al. 2024), and protection factors can be used to derive dynamic models from experiments. At the same time, many current and challenging HDX studies probe the exchange of large systems such as integral membrane proteins (Martens and Politis 2020) or ribosomes (Wales et al. 2022), for which it is not easy to meet the threshold on redundancy. In these cases, single residue resolution is most likely to be either inconclusive or limited to well-covered subregions of the protein. The adoption of novel, more efficient acid proteases for digestion and the integration of alternative fragmentation techniques into the experimental HDX-MS workflow are however likely to help with the challenge of achieving routine, high-resolution data in more complicated systems (Trabjerg et al. 2018).

It is well understood that longer time points are more valuable for characterisation of secondary structure elements in proteins such as α-helices and β-sheets, whereas short time points are essential for loops and disordered regions, which potentially even require msec timepoints (Seetaloo and Phillips 2022). Adaptive temporal sampling, i.e. the use of well-matched time points for the system under investigation, is a promising approach to high-quality data and more efficient experimental workflows. We used the opportunity here to check how the results are influenced by the number, spread and type (early/late) of time points available in the dataset (**Figure 5.8**). For the ubiquitin dataset, where most of the peptides' uptake curves were properly sampled, we showed that 4 time points were sufficient to achieve a strong correlation with NMR data. Rather than the number of time points alone, the quality of the dataset is defined by the choice of the appropriate time points, and this is protein (structure) dependent: later time points were found to be more important for ubiquitin, while earlier time points for GB1-TACC3 (**Supplementary Figure 5.11**) – in agreement with the more disordered nature of the latter. One advantage of the structurally informed approach is that the protection factors estimated using the Best-Vendruscolo model (**Eq. 5.7**) can be used to calculate

an expected minimal time of exchange (**Figure 5.9**). The labelling time points can be chosen in the light of these predictions: for example, the C-terminal disordered region of GB1-TACC3 has completely exchanged after 30 s (the earliest time point acquired here), and the exchange time predictions suggest that millisecond HDX-MS data are needed (Kish et al. 2023; Rob et al. 2012; Svejdal et al. 2019) to properly sample this area of the protein.

Given the high coverage provided both in the ordered and disordered regions, the GB1-TACC3 complex proves to be an ideal reference for the study of disordered areas with fast HDX-MS instruments. The exchange time predictions depend – obviously - on the intrinsic exchange rates, which are amongst others a function of pH and temperature. These two parameters can be optimized *a priori*, within the stable conformational range of the protein, to accelerate or decelerate the exchange kinetics of a specific region of interest so that it fits into the monitored time range.

The data treatment performed here assumes that all residues of the protein follow the EX2 regime. The alternative scenario, known as EX1, is extremely rare and prone to misinterpretation (Fang et al. 2011; Hodge, Benhaim, and Lee 2020). However, it has been observed for intrinsically disordered proteins, and we cannot exclude that GB1-TACC3 might follow such regime on the millisecond time scale. The observed exchange rate for a residue in the EX1 regime is the opening exchange rate k_{op} (**Eq. 5.1**), hence the protection factor is not a well-defined quantity (Hamuro 2021b). An integration of EX1 and EX2 kinetic information is possible only when moving the focus from protection factors to even finer (i.e. at higher resolution) parameters: the opening, closing and intrinsic exchange rates. However, experimental techniques are far from achieving such resolution and also computational studies are limited on this account (Craig et al. 2011; Park et al. 2015; Persson and Halle 2015).

The fact that single residue resolution can be achieved from HDX-MS experiments does not diminish the importance of HDX-NMR experiments for fundamental understanding of the HDX mechanism. For example, the Englander intrinsic exchange rate predictions

(Bai et al. 1993; Connelly et al. 1993; Molday et al. 1972; Nguyen, Mayne, Phillips, and Englander 2018), which predict the exchange rate of a residue in a completely unfolded structure at a given pH and temperature, do not account for the effect of the ionic strength on the exchange. Predicting the dependence of the intrinsic exchange rate on the salt type and concentration would help the deconvolution of chemical and structural effects in experiments probing conformational changes of the same protein in different environments (Toth et al. 2017). HDX-NMR, more than MS, is the ideal candidate to study such dependence. As a second example, the Best-Vendruscolo model (**Eq. 5.7**) is only one of several models that have been developed to connect the structure of a protein to its protection factors (Devaurs et al. 2022). These models have been optimized over datasets containing a relatively small number of proteins, seven in the case of the phenomenological approximation. The largest dataset available, Start2Fold, contains the protection factors of 38 proteins (Pancsa et al. 2016). The acquisition of a larger dataset containing the protection factors of hundreds of proteins would shed light on how other factors, such as electrostatics, affect the protection against exchange. Moreover, a larger dataset would enable the use of machine learning tools to predict protection factors from the structure of a protein. Even in this second case, HDX-NMR, more than MS, would be the ideal candidate.

5.6 Conclusions

We showed that it is possible to retrieve single residue information (protection factors) from peptide level HDX-MS data that has a strong correlation ($\rho > 0.75$) with NMR data studying the same protein under the same conditions.

In the light of our results, we propose two strategies to perform a protection factor analysis for HDX-MS data. If a protein structure model is available, protection factors can be estimated and used as *informed* initial guess to fit experimental data; if the structure is not known, a random search can be performed.

The accuracy of these predictions depends on the quality of the HDX-MS dataset, and we recommend performing such analysis only for the parts of the protein where the redundancy is higher than 4. As for the number of time points, the importance rests in their ability to properly sample the uptake curves of peptides rather than in their number. If the structure of the protein is available, we showed how the choice of the labelling time points can be guided by the protection factors predicted by the phenomenological approximation.

We hope that the dataset published here, which probes the exchange kinetics of two different proteins with both HDX-MS and HDX-NMR experiments under the same experimental conditions, will become a reference for all researchers aiming to publish methods to extract protection factors from HDX-MS data. The dataset is available in the PRIDE database [5].

There are still several areas that need improvements to enhance the quantitative analysis of HDX-MS, such as the integration of molecular dynamics (MD) simulations into the workflow – more than isolated structures – or the incorporation of EX1 kinetics analysis, which has not been considered here. Additionally, there are barriers in the fundamental understanding of the phenomenon (HDX) that are hindering the development and widespread use of single-residue resolution methods, such as the limited understanding of back-exchange and the restricted access to the isotopic distribution of peptides.

[5] At the writing time of this Thesis (September 2024), the dataset is not available yet on PRIDE.

5.7 Supplementary Material

Supplementary Figure 5.1. Protein sequences.

Ubiquitin

MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL

EDGRTLSDYN IQKESTLHLV LRLRGG

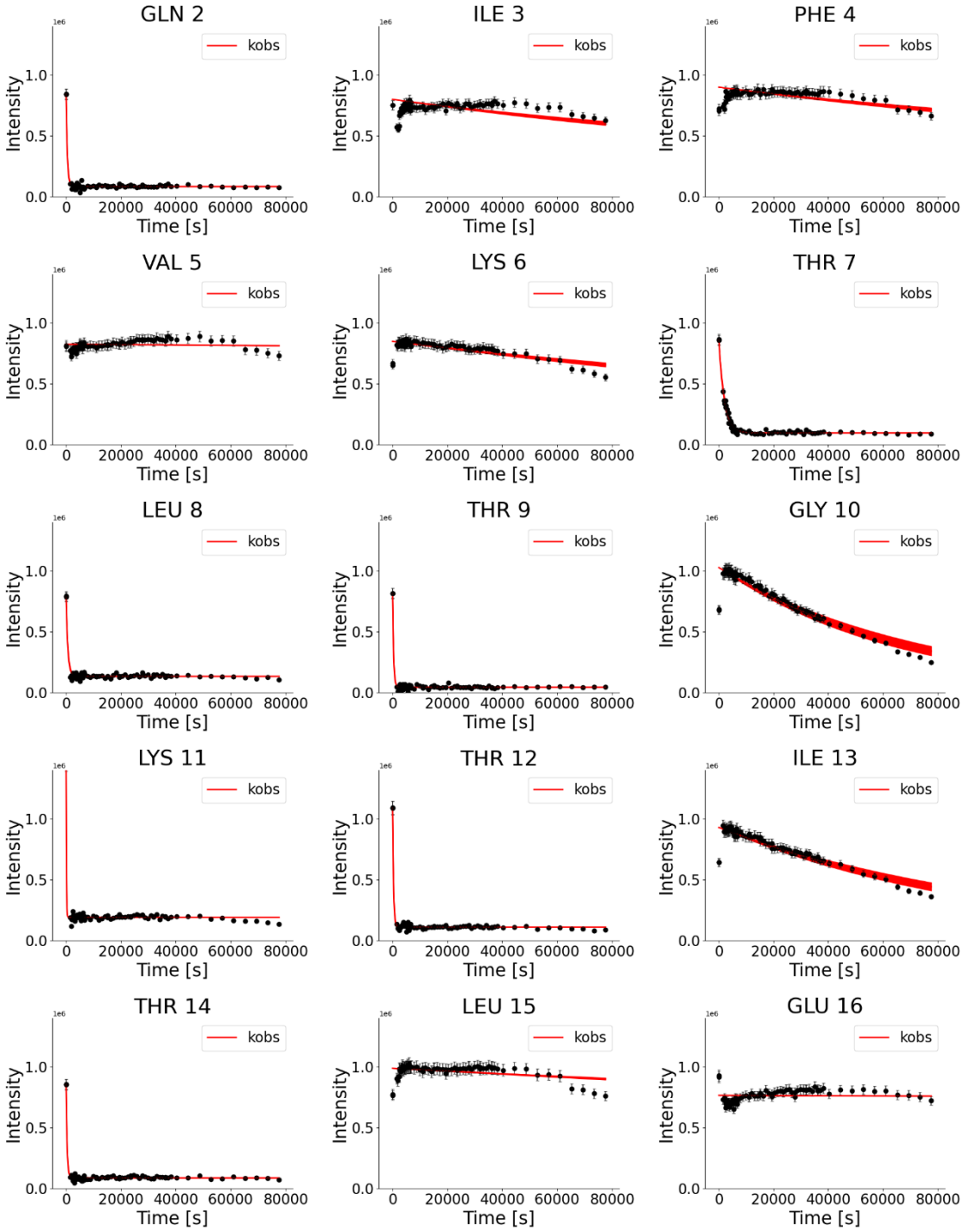
GB1-TACC3

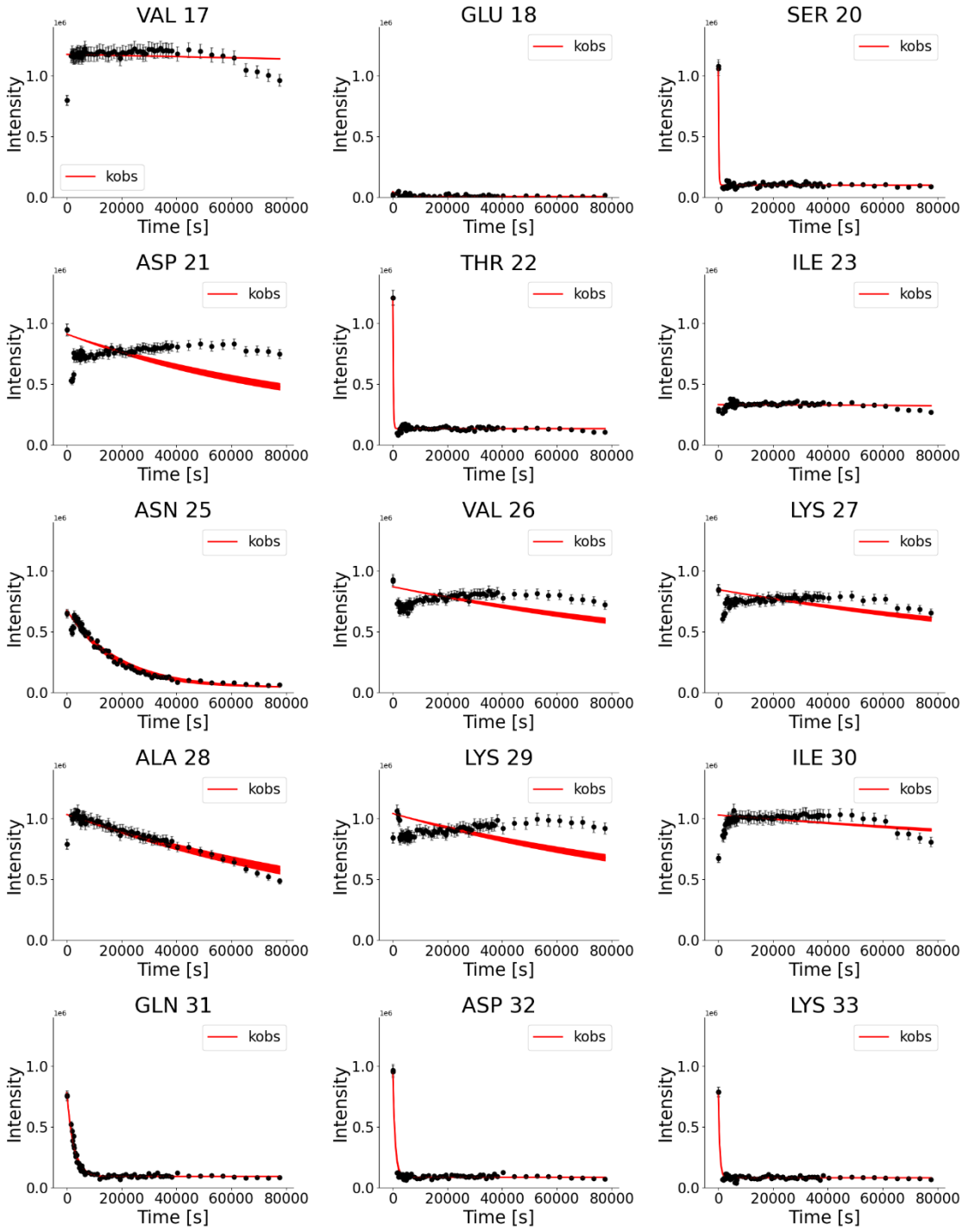
GAMAQYKLAL NGKTLKGETT TEAVDAATAE KVFKQYANDN GVDGEWTYDD

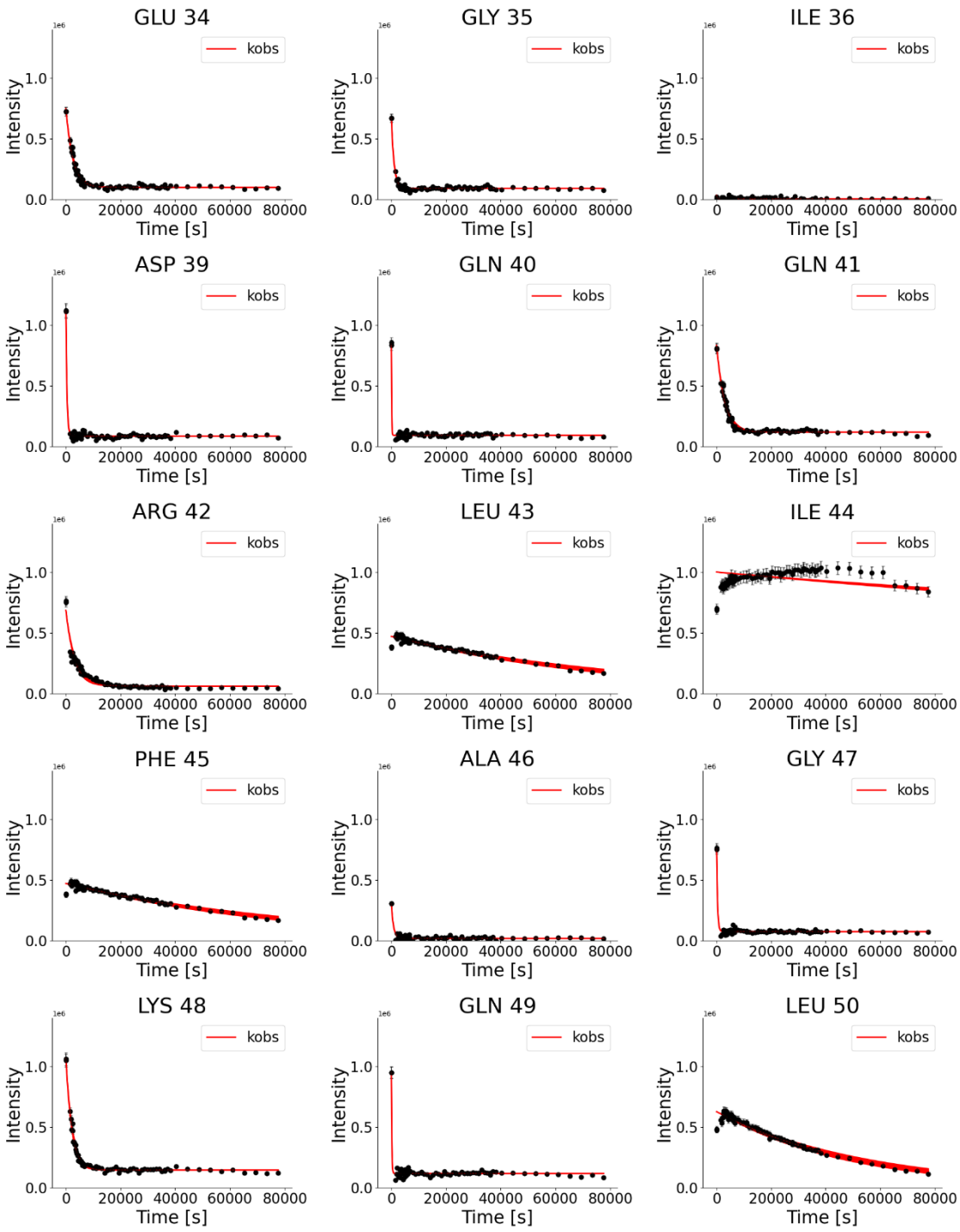
ATKTFTVTET SDGTSSFKES ALRKQSLYLK FDPLLRDSPG R

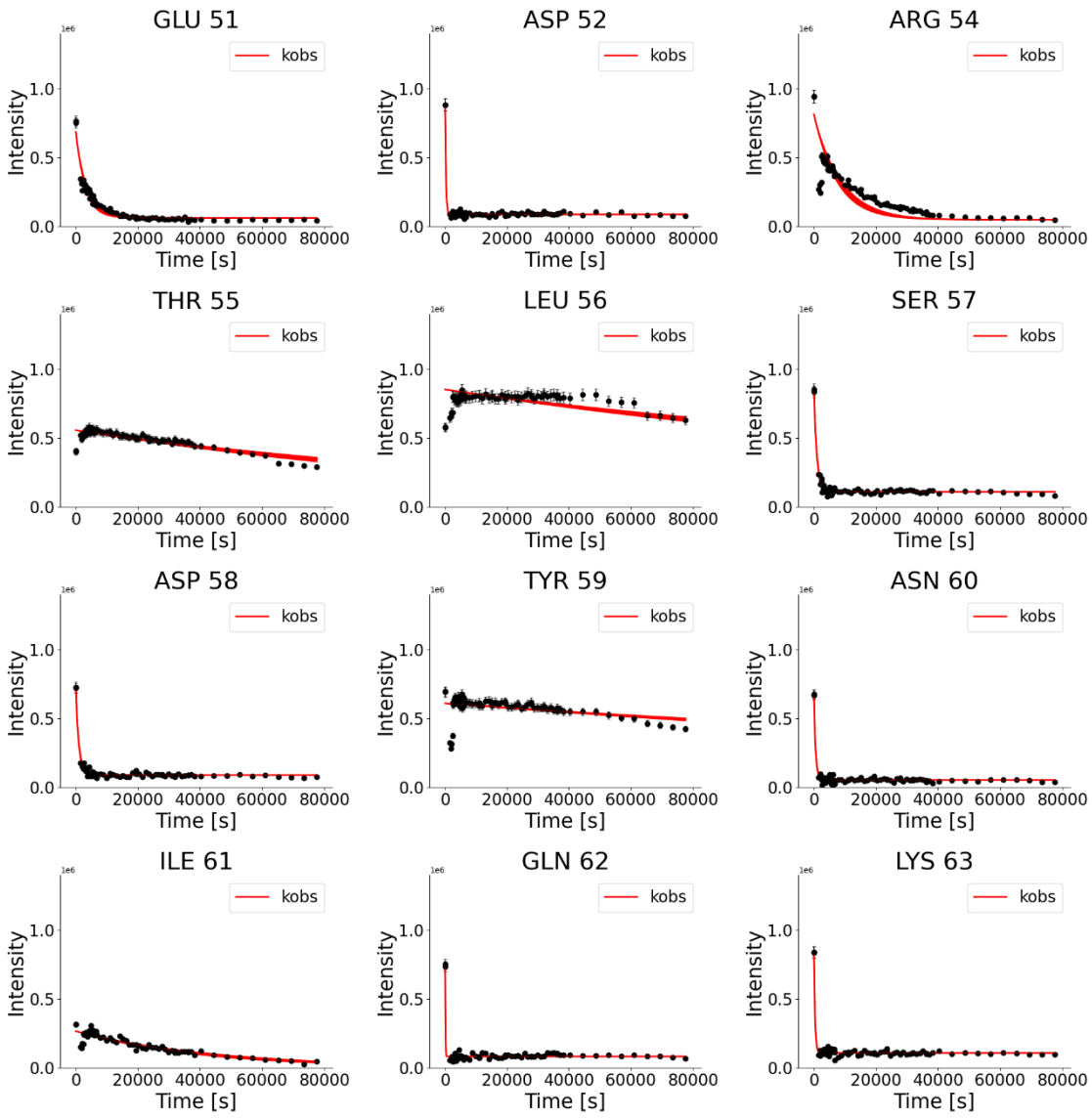
Supplementary Figure 5.2. HDX-NMR curves for ubiquitin.

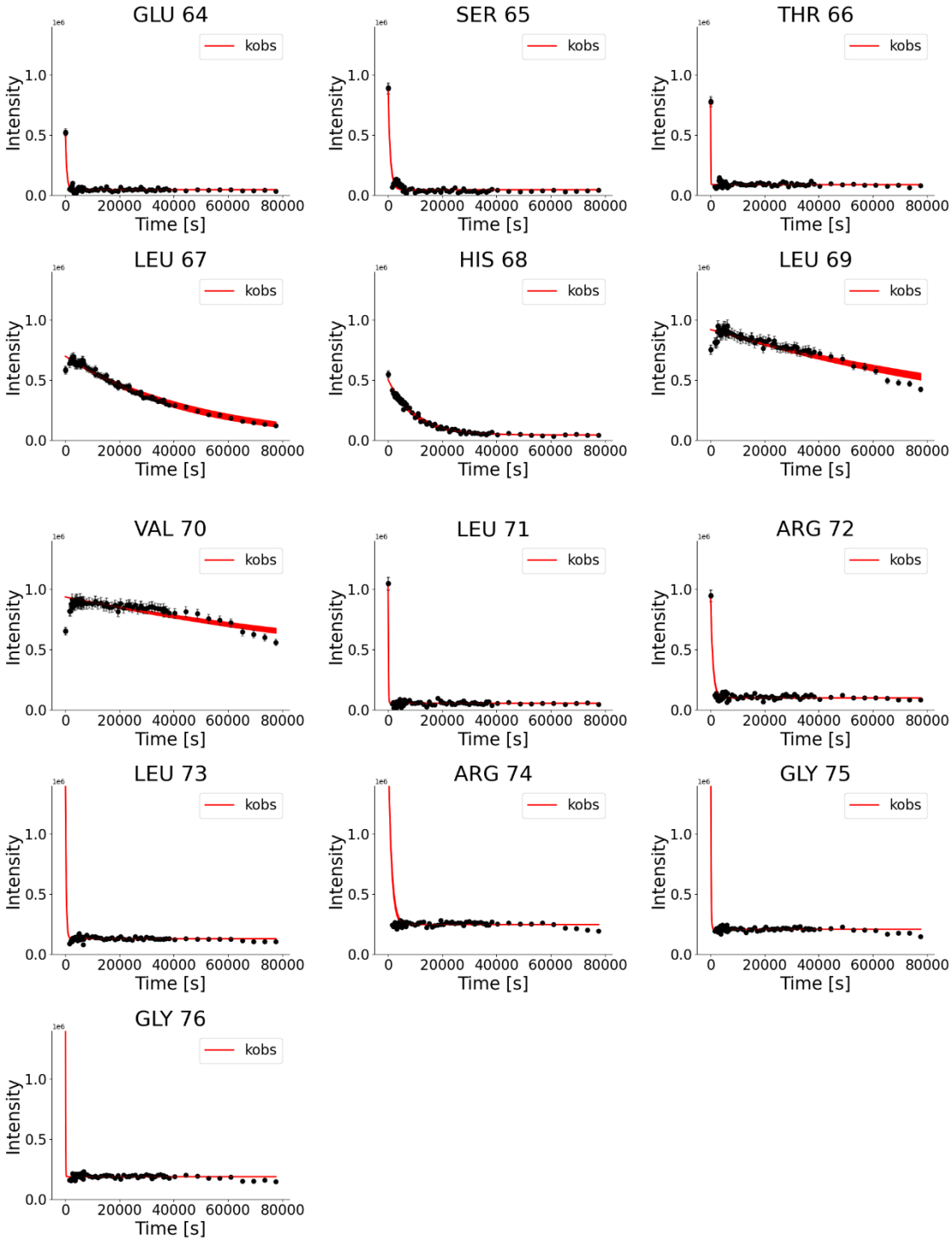
Intensities extracted from 2D HSQC spectra shown as a function of the labelling time for every assigned residue in the control spectrum of ubiquitin. Experimental data are fitted with a single exponential model (red line).




VAL 17
GLU 18
SER 20
ASP 21
THR 22
ILE 23
ASN 25
VAL 26
LYS 27
ALA 28
LYS 29
ILE 30
GLN 31
ASP 32
LYS 33
kobs
Intensity
Time [s]
1e6
0.0
0.5
1.0
0
20000
40000
60000
80000

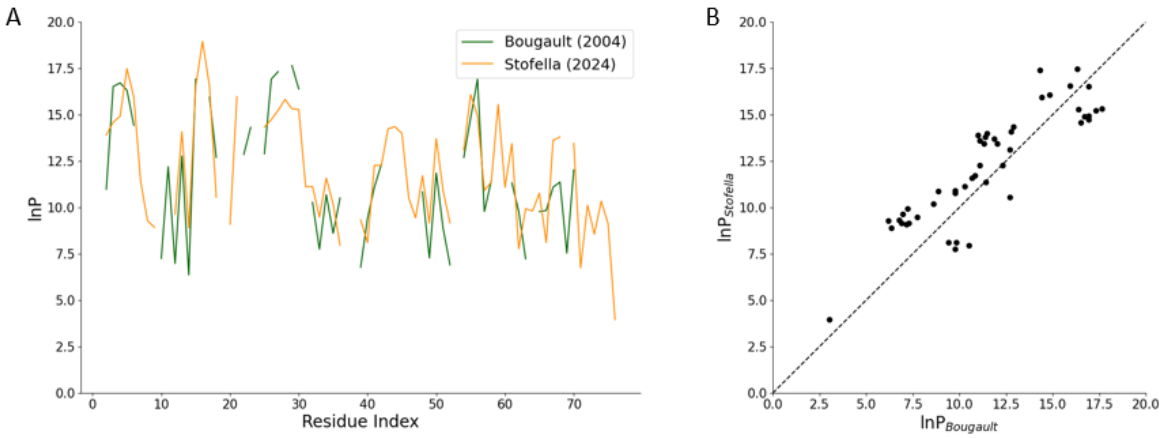

GLU 34
GLY 35
ILE 36
ASP 39
GLN 40
GLN 41
ARG 42
LEU 43
ILE 44
PHE 45
ALA 46
GLY 47
LYS 48
GLN 49
LEU 50
kobs
Intensity
Time [s]
1e6
0.0
0.5
1.0
0
20000
40000
60000
80000


GLU 51
ASP 52
ARG 54
THR 55
LEU 56
SER 57
ASP 58
TYR 59
ASN 60
ILE 61
GLN 62
LYS 63
kobs
Intensity
Time [s]
1e6
0.0
0.5
1.0
0
20000
40000
60000
80000


GLU 64
SER 65
THR 66
LEU 67
HIS 68
LEU 69
VAL 70
LEU 71
ARG 72
LEU 73
ARG 74
GLY 75
GLY 76
kobs
Intensity
Time [s]
1e6
0.0
0.5
1.0
0
20000
40000
60000
80000

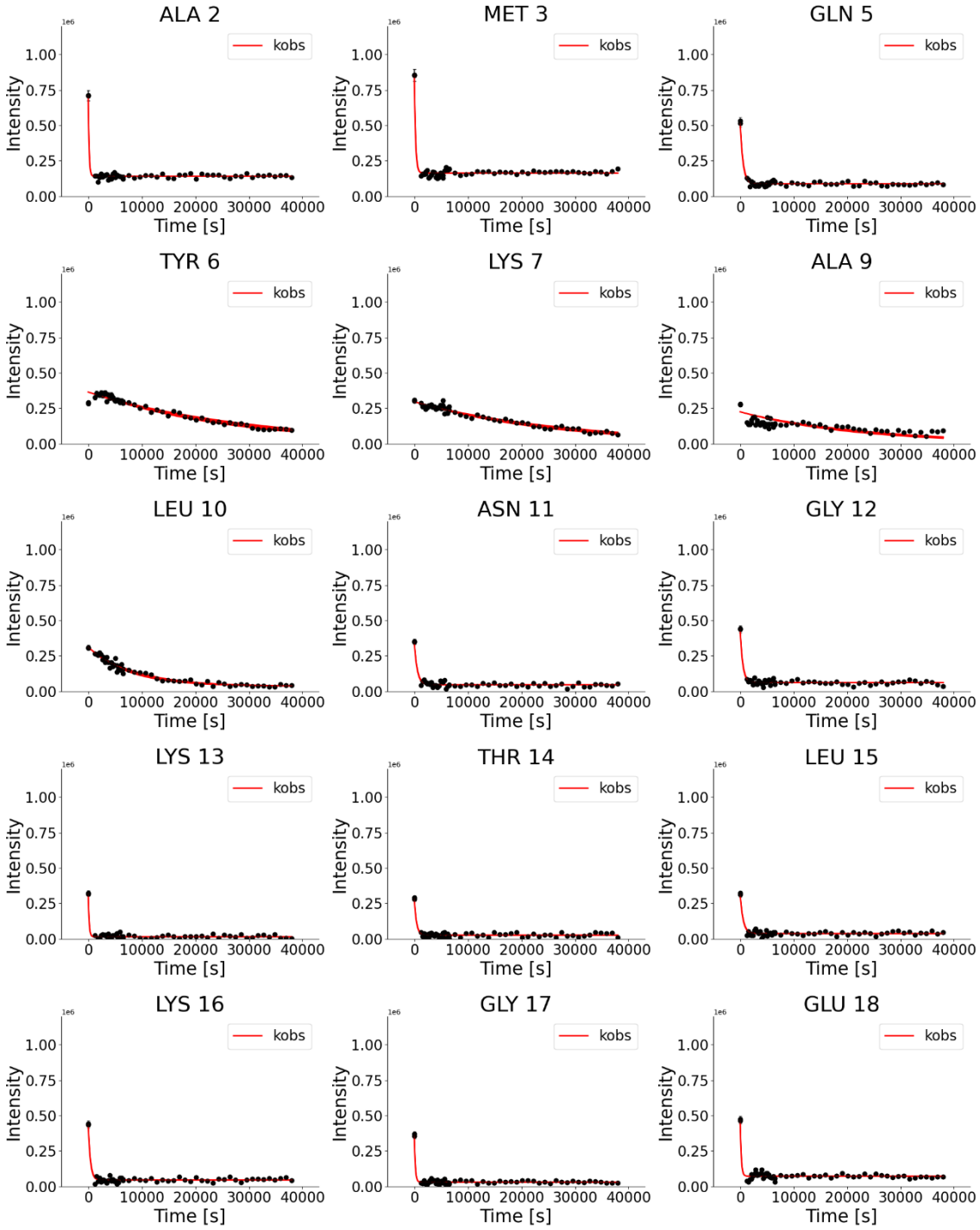
Supplementary Figure 5.3. Comparison of ubiquitin protection factors from HDX-NMR data with literature.

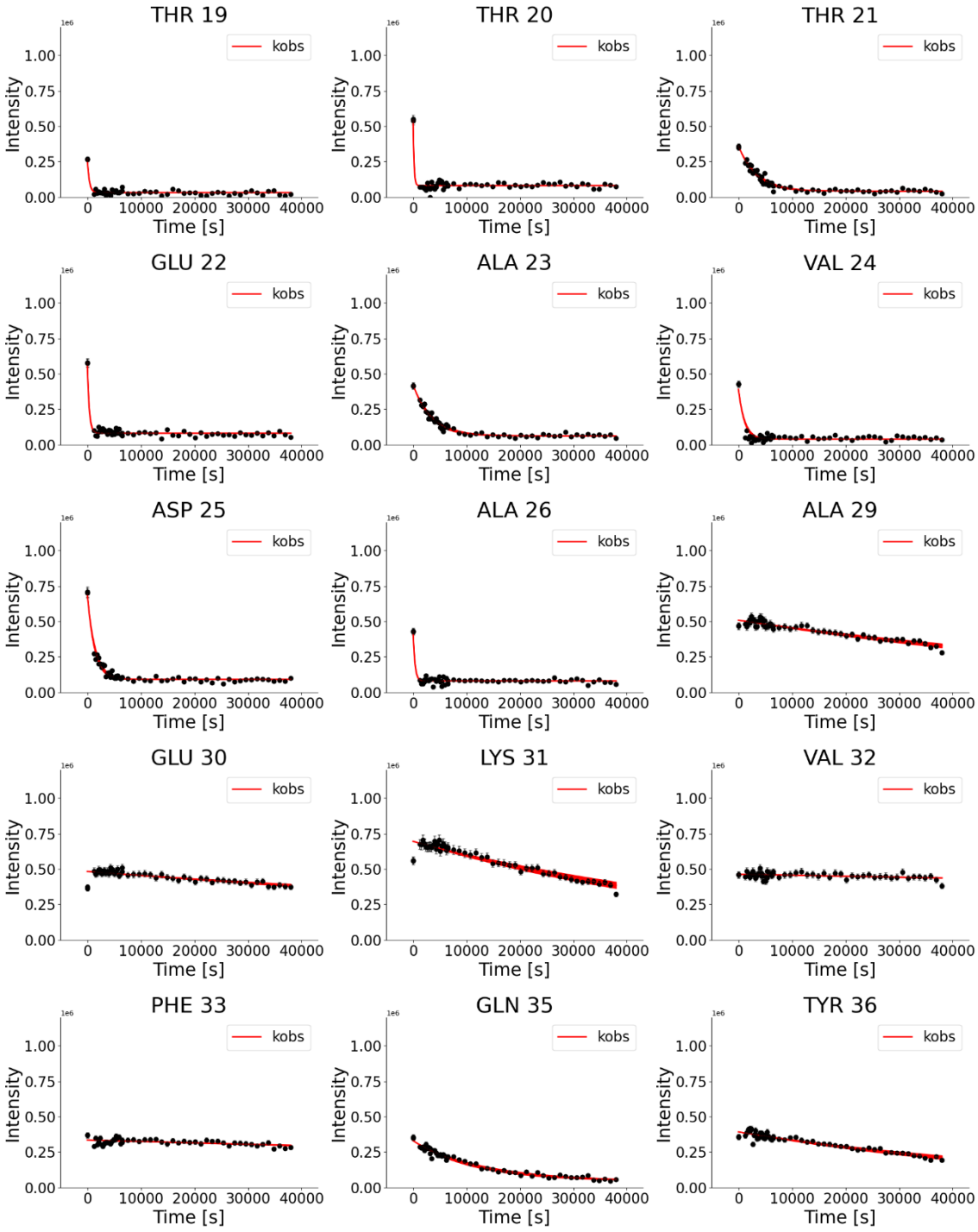
A) Ubiquitin protection factors extracted from HDX-NMR data (orange line) at pH 7.0 and temperature 4°C are compared with protection factors estimated by Bougault et al. (2004) at pH 6.2 and temperature 25°C (green light). B) Correlation plot, showing a correlation coefficient $\boldsymbol{\rho = 0.88}$.

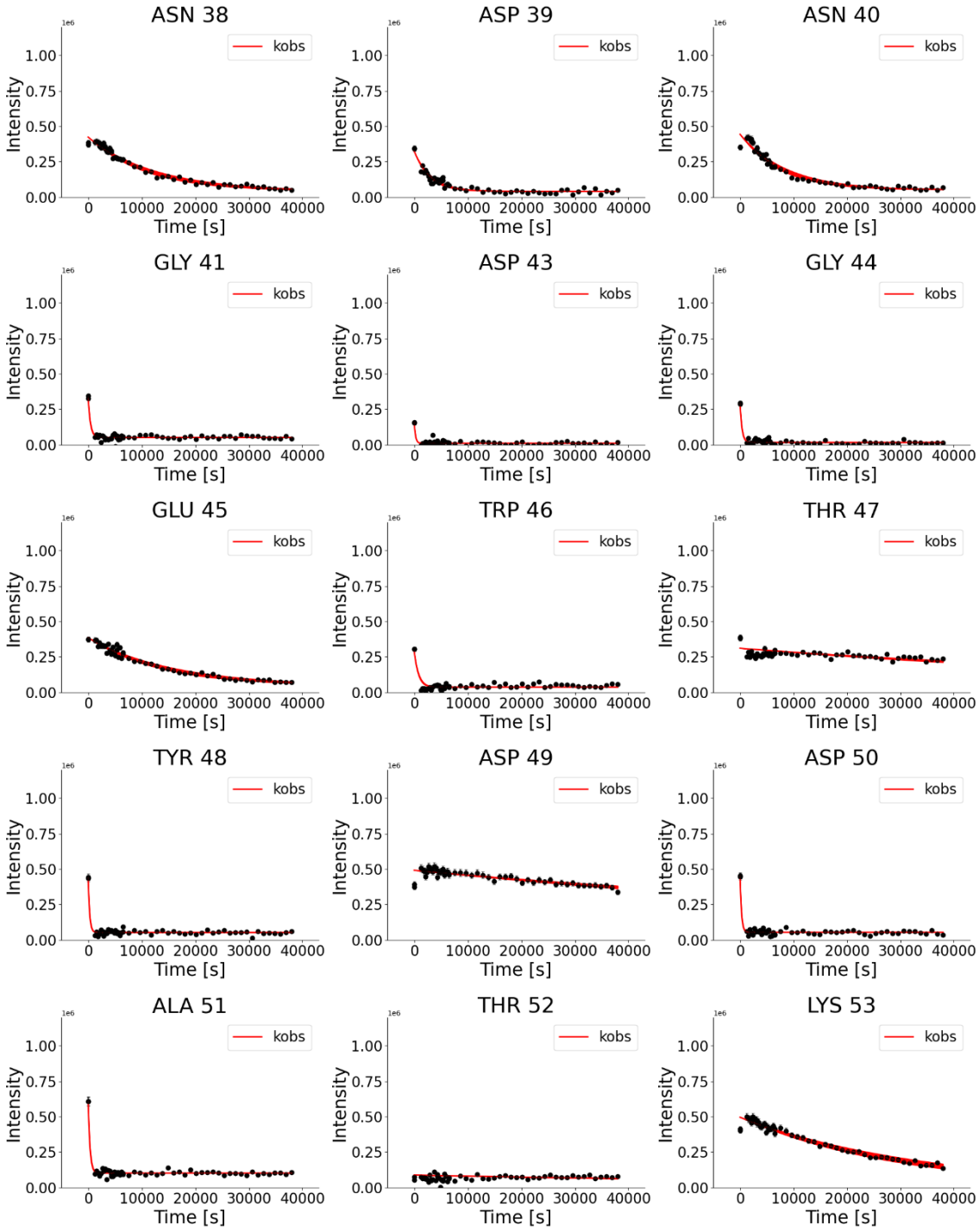


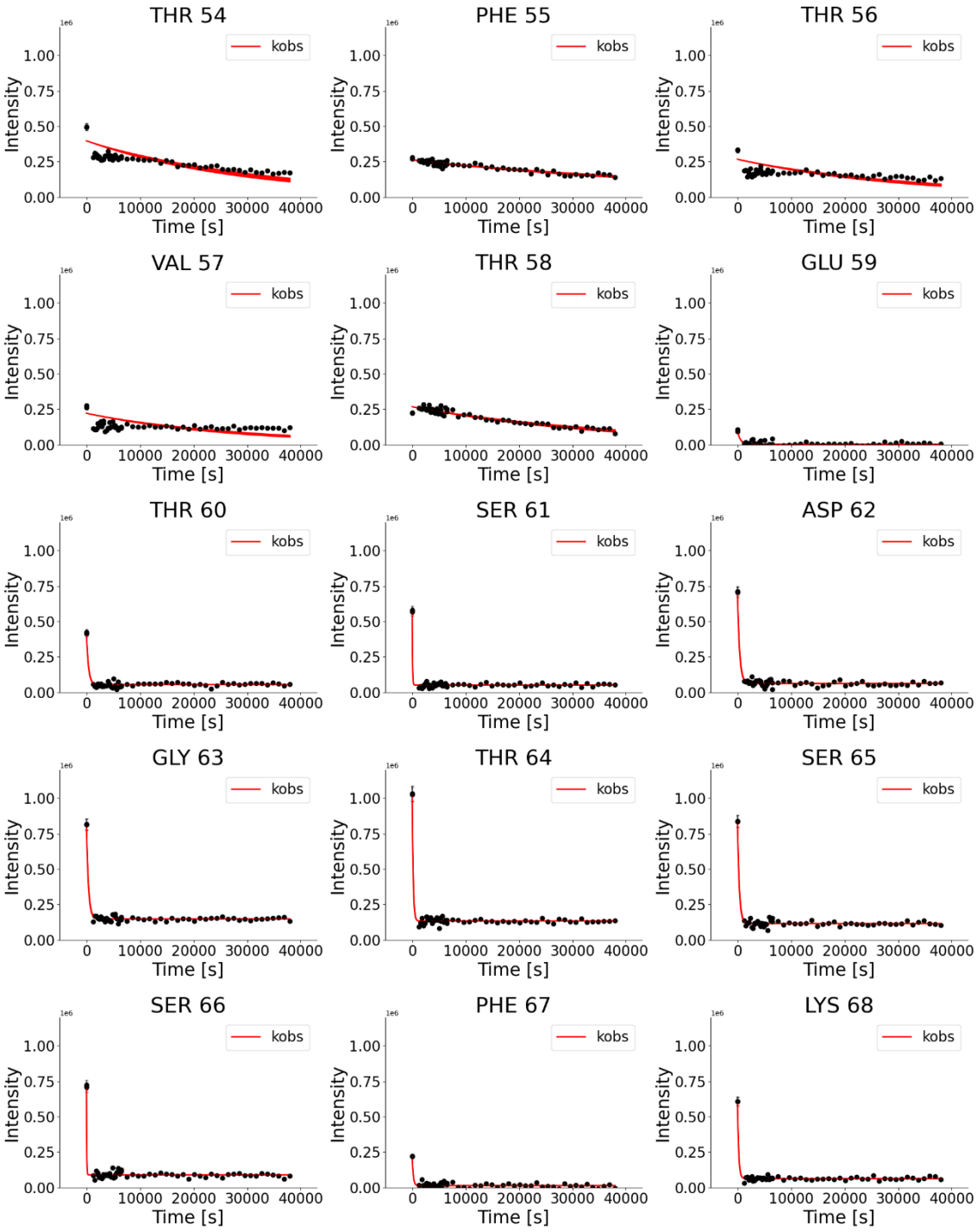
Supplementary Figure 5.4. HDX-NMR curves for GB1-TACC3.

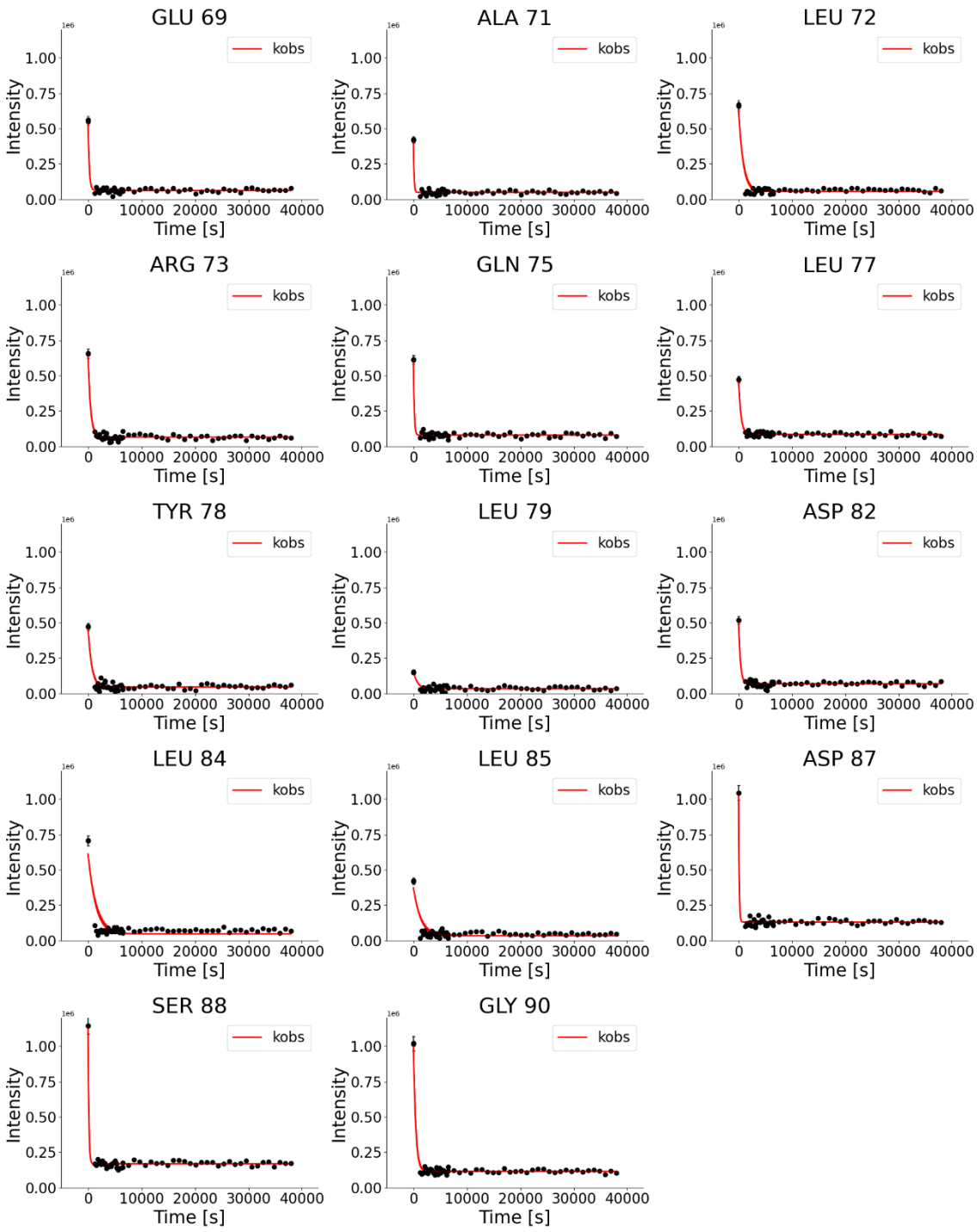
Intensities extracted from 2D HSQC spectra shown as a function of the labelling time for every assigned residue in the control spectrum of ubiquitin. Experimental data are fitted with a single exponential model (red line).



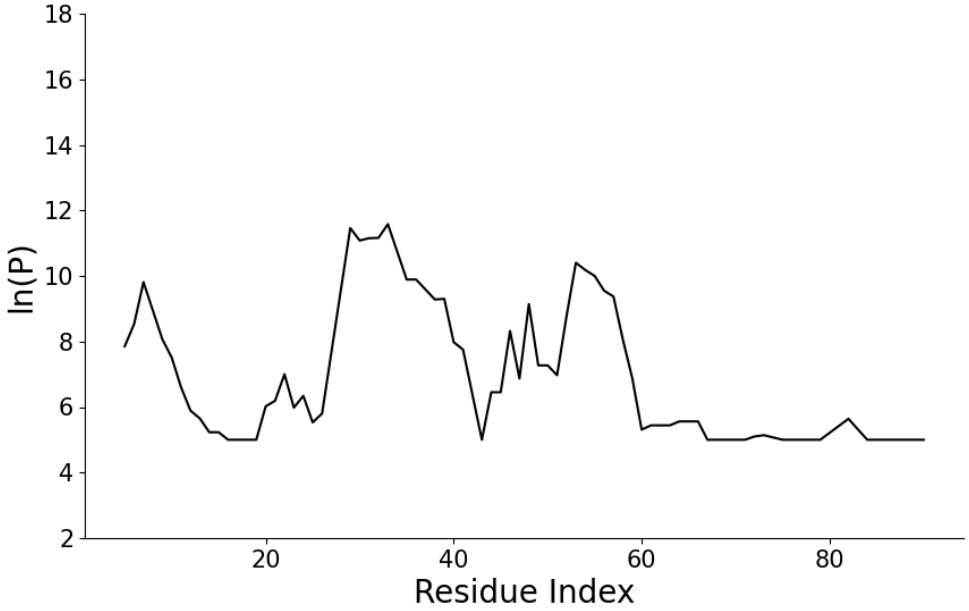

THR 19
THR 20
THR 21
GLU 22
ALA 23
VAL 24
ASP 25
ALA 26
ALA 29
GLU 30
LYS 31
VAL 32
PHE 33
GLN 35
TYR 36
kobs
Intensity
Time [s]
1e6
0.00
0.25
0.50
0.75
1.00
0
10000
20000
30000
40000


ASN 38
ASP 39
ASN 40
GLY 41
ASP 43
GLY 44
GLU 45
TRP 46
THR 47
TYR 48
ASP 49
ASP 50
ALA 51
THR 52
LYS 53
kobs
Intensity
Time [s]
1e6
0.00
0.25
0.50
0.75
1.00
0
10000
20000
30000
40000


THR 54
PHE 55
THR 56
VAL 57
THR 58
GLU 59
THR 60
SER 61
ASP 62
GLY 63
THR 64
SER 65
SER 66
PHE 67
LYS 68
kobs
Intensity
Time [s]
1e6
0.00
0.25
0.50
0.75
1.00
0
10000
20000
30000
40000

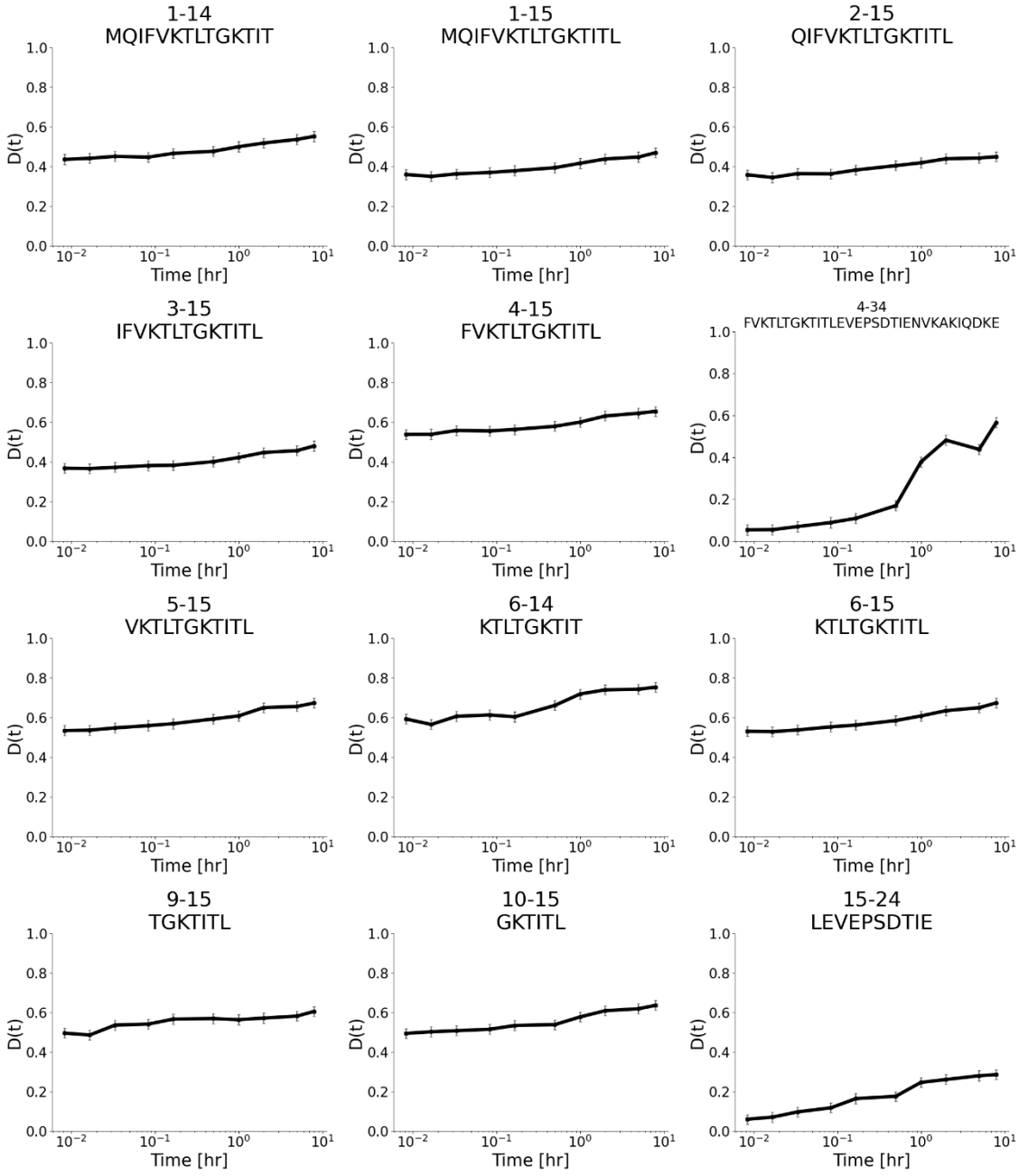

GLU 69
ALA 71
LEU 72
ARG 73
GLN 75
LEU 77
TYR 78
LEU 79
ASP 82
LEU 84
LEU 85
ASP 87
SER 88
GLY 90
kobs
Intensity
Time [s]
1e6
0.00
0.25
0.50
0.75
1.00
0
10000
20000
30000
40000

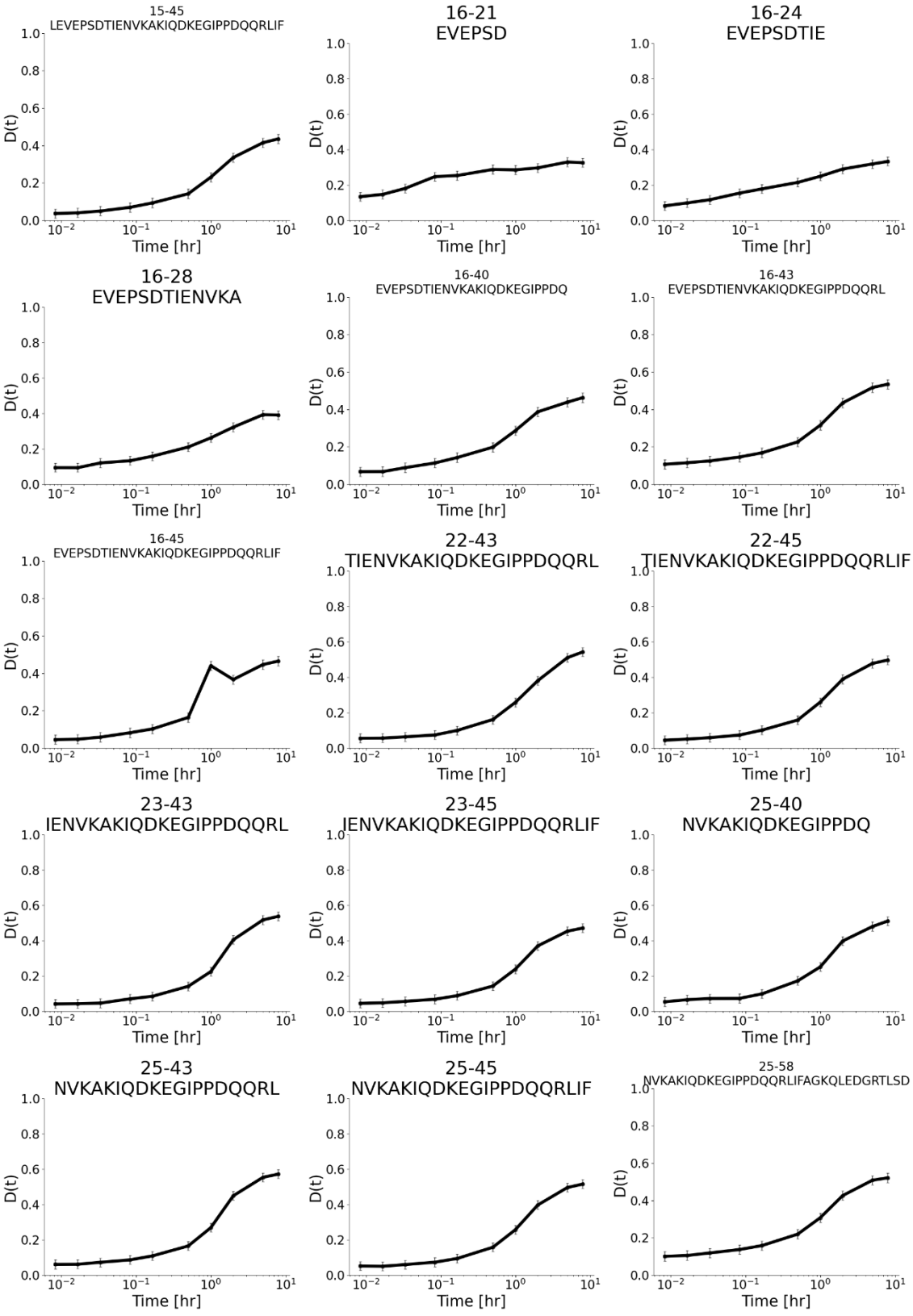
Supplementary Figure 5.5. Protection factors extracted for GB1-TACC3 from HDX-NMR data. Protection factors extracted for GB1-TACC3 from HDX-NMR data and used as reference for the MS/NMR comparison.

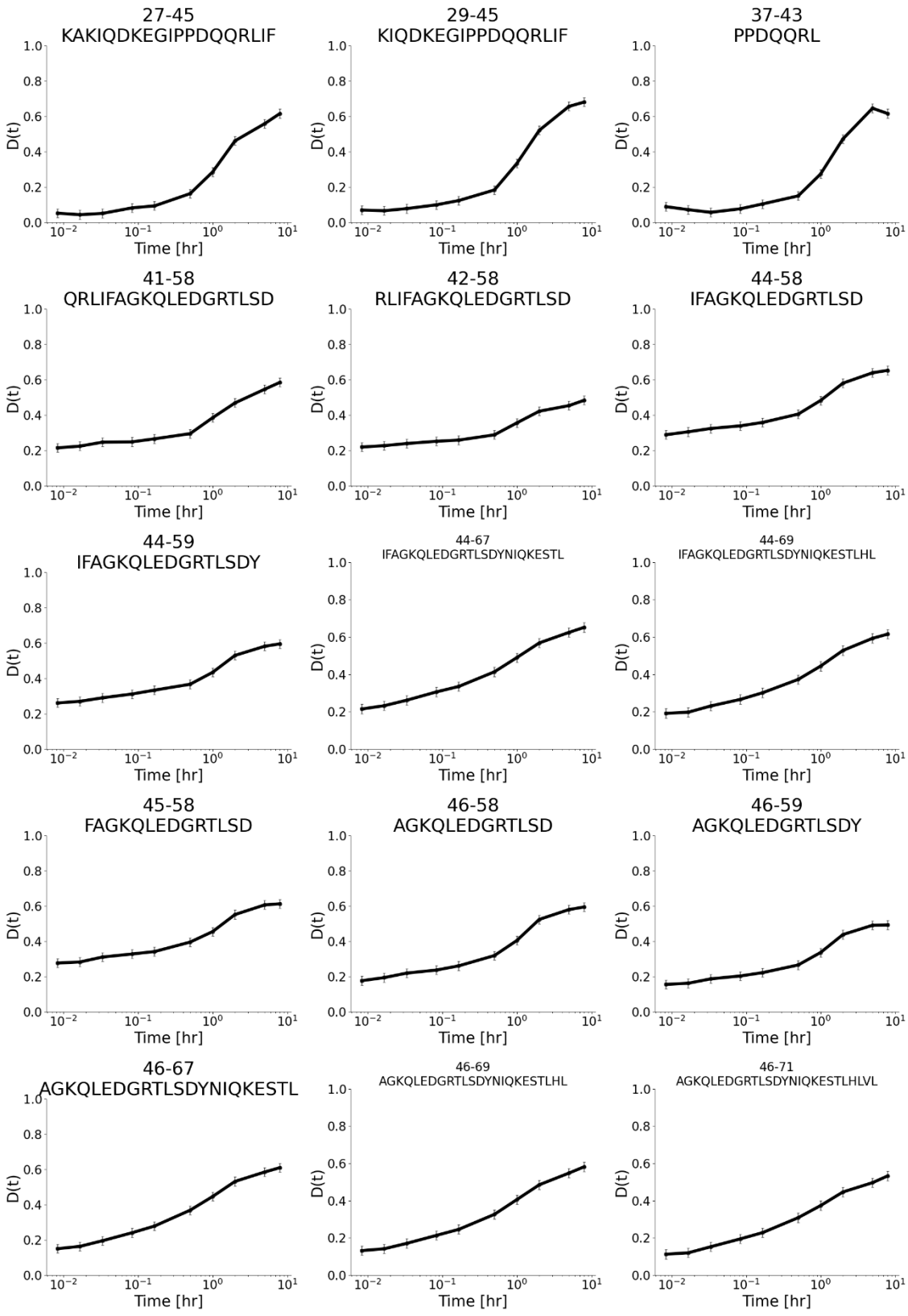


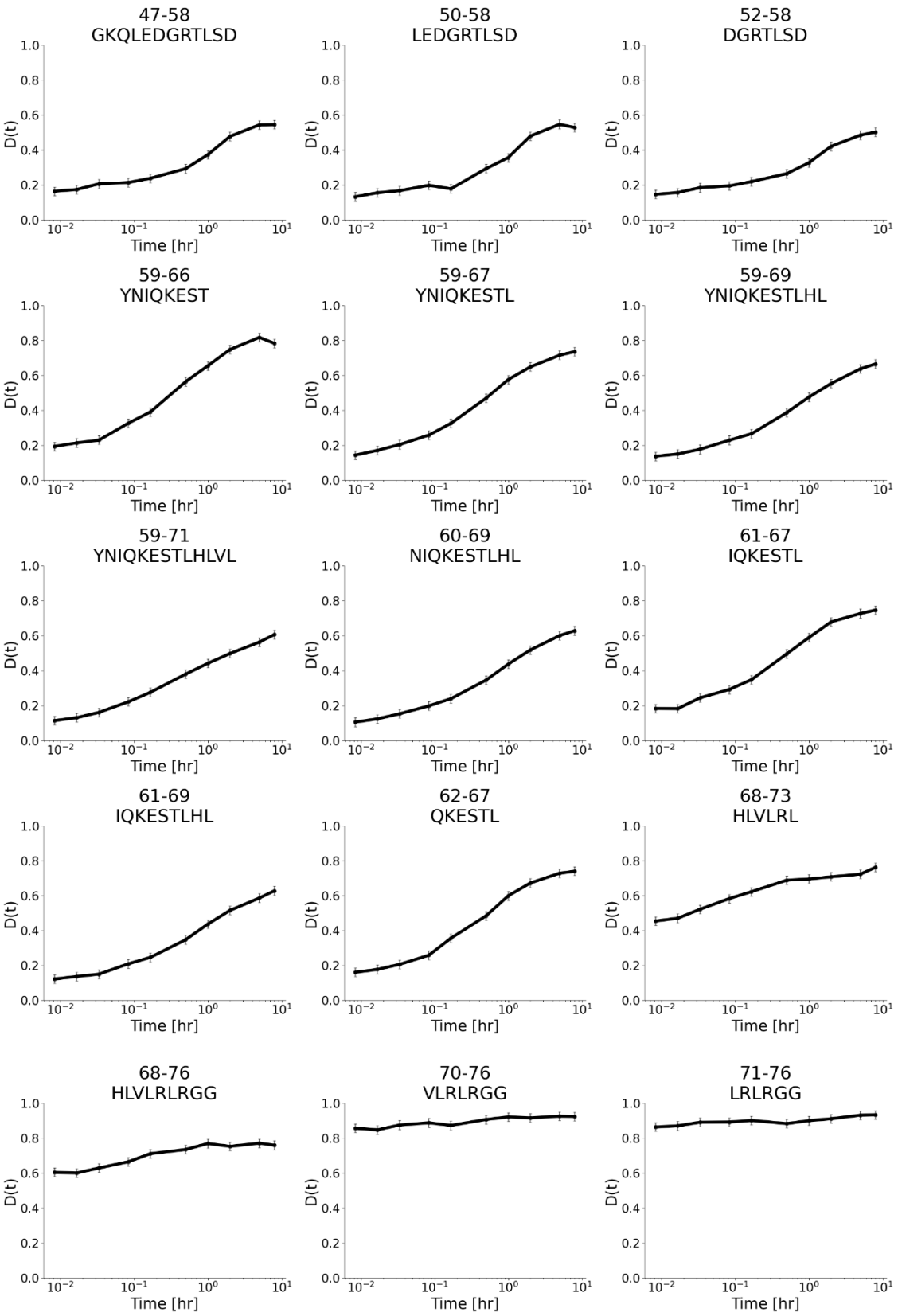
Supplementary Figure 5.6. Fractional uptake for all peptides identified by the HDX-MS experiments for ubiquitin.

The error associated to experimental measurements is the pooled standard deviation $\boldsymbol{\sigma = 0.025}$.



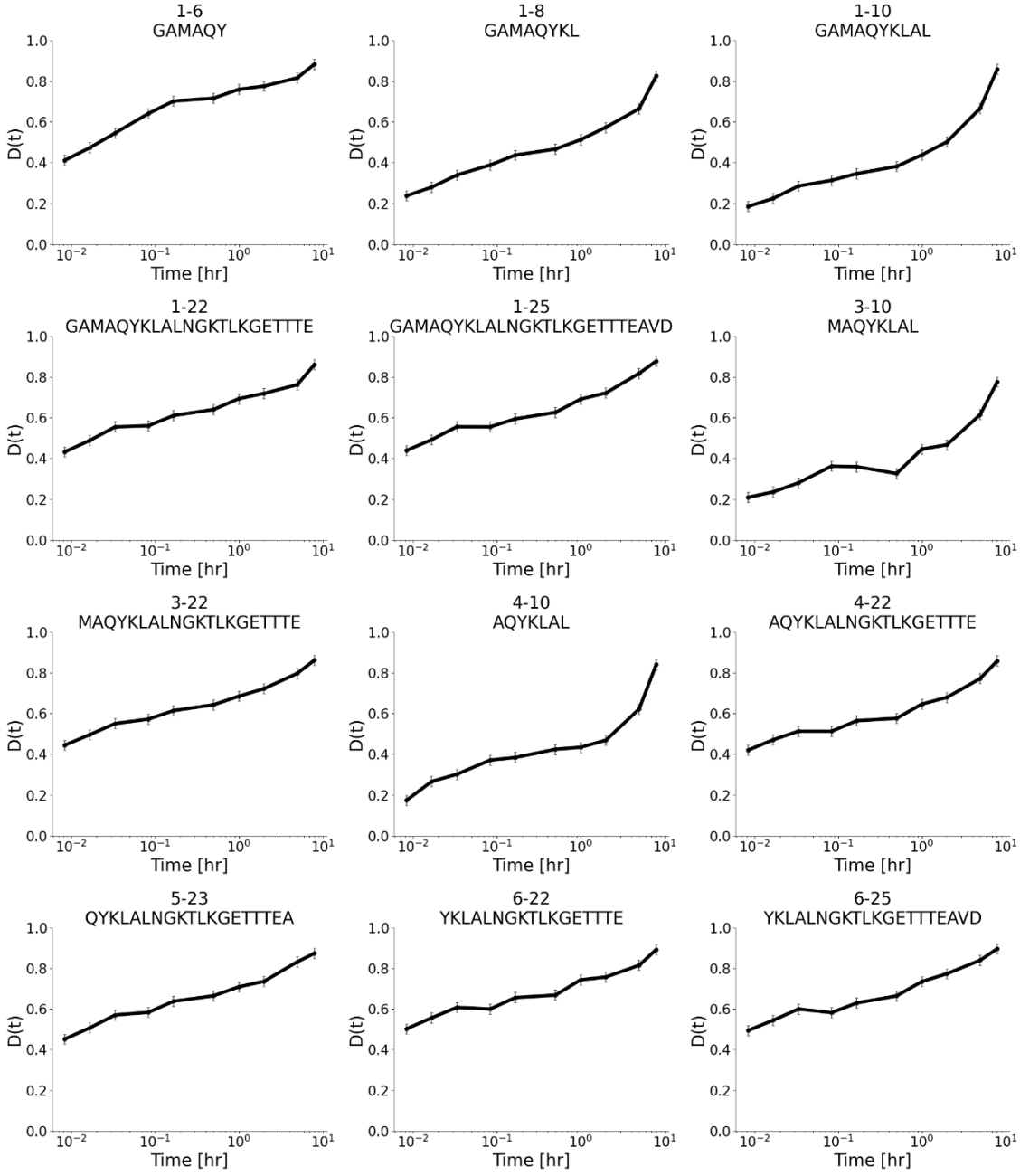

15-45
LEVEPSDTIENVKAKIQDKEGIPPDQQRLIF
16-21
EVEPSD
16-24
EVEPSDTIE
16-28
EVEPSDTIENVKA
16-40
EVEPSDTIENVKAKIQDKEGIPPDQ
16-43
EVEPSDTIENVKAKIQDKEGIPPDQQRL
16-45
EVEPSDTIENVKAKIQDKEGIPPDQQRLIF
22-43
TIENVKAKIQDKEGIPPDQQRL
22-45
TIENVKAKIQDKEGIPPDQQRLIF
23-43
IENVKAKIQDKEGIPPDQQRL
23-45
IENVKAKIQDKEGIPPDQQRLIF
25-40
NVKAKIQDKEGIPPDQ
25-43
NVKAKIQDKEGIPPDQQRL
25-45
NVKAKIQDKEGIPPDQQRLIF
25-58
NVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSD
D(t)
Time [hr]

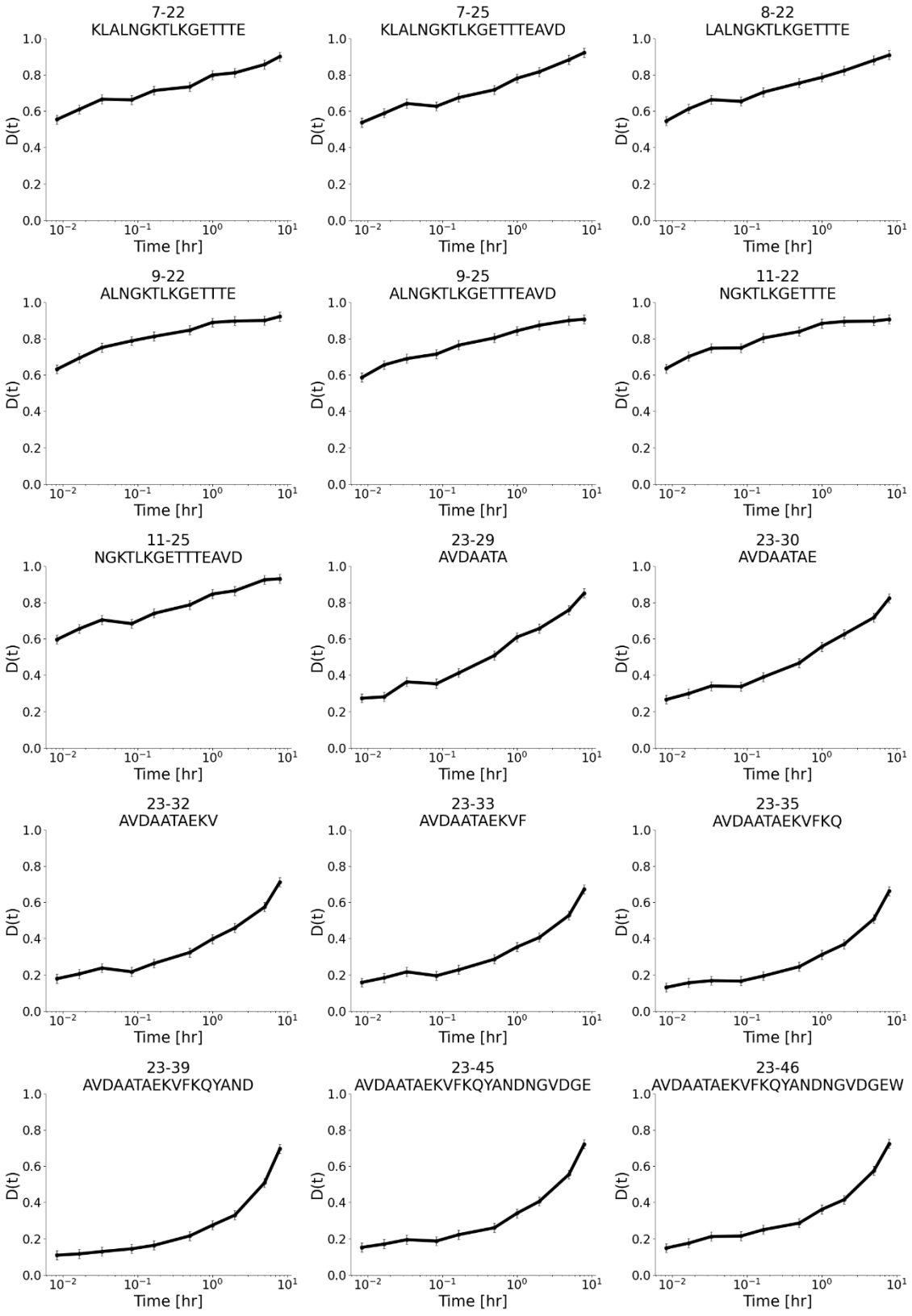

27-45
KAKIQDKEGIPPDQQRLIF
29-45
KIQDKEGIPPDQQRLIF
37-43
PPDQQRL
41-58
QRLIFAGKQLEDGRTLSD
42-58
RLIFAGKQLEDGRTLSD
44-58
IFAGKQLEDGRTLSD
44-59
IFAGKQLEDGRTLSDY
44-67
IFAGKQLEDGRTLSDYNIQKESTL
44-69
IFAGKQLEDGRTLSDYNIQKESTLHL
45-58
FAGKQLEDGRTLSD
46-58
AGKQLEDGRTLSD
46-59
AGKQLEDGRTLSDY
46-67
AGKQLEDGRTLSDYNIQKESTL
46-69
AGKQLEDGRTLSDYNIQKESTLHL
46-71
AGKQLEDGRTLSDYNIQKESTLHLVL
D(t)
Time [hr]

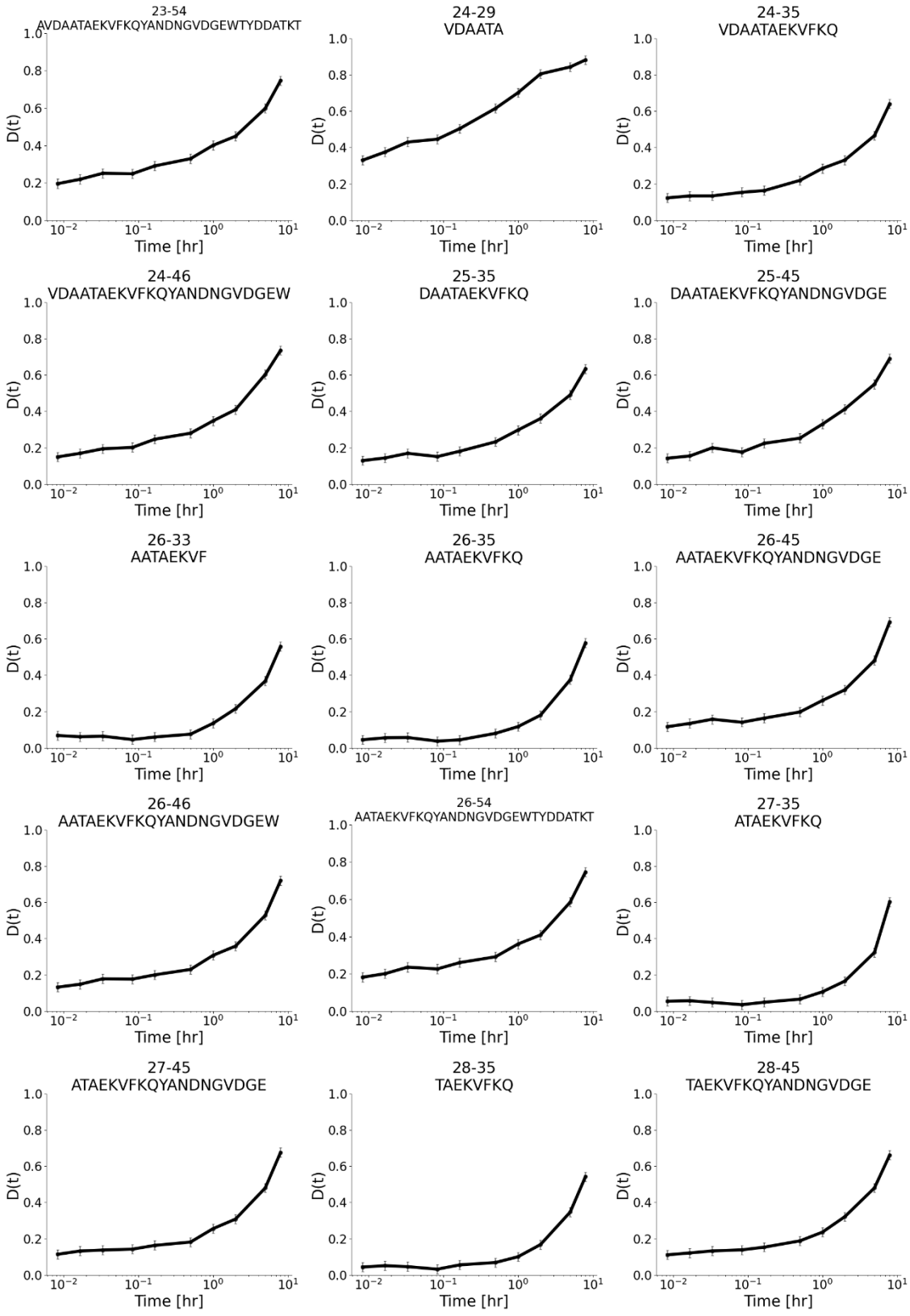

47-58
GKQLEDGRTLSD
50-58
LEDGRTLSD
52-58
DGRTLSD
59-66
YNIQKEST
59-67
YNIQKESTL
59-69
YNIQKESTLHL
59-71
YNIQKESTLHLVL
60-69
NIQKESTLHL
61-67
IQKESTL
61-69
IQKESTLHL
62-67
QKESTL
68-73
HLVLRL
68-76
HLVLRLRGG
70-76
VLRLRGG
71-76
LRLRGG
D(t)
Time [hr]

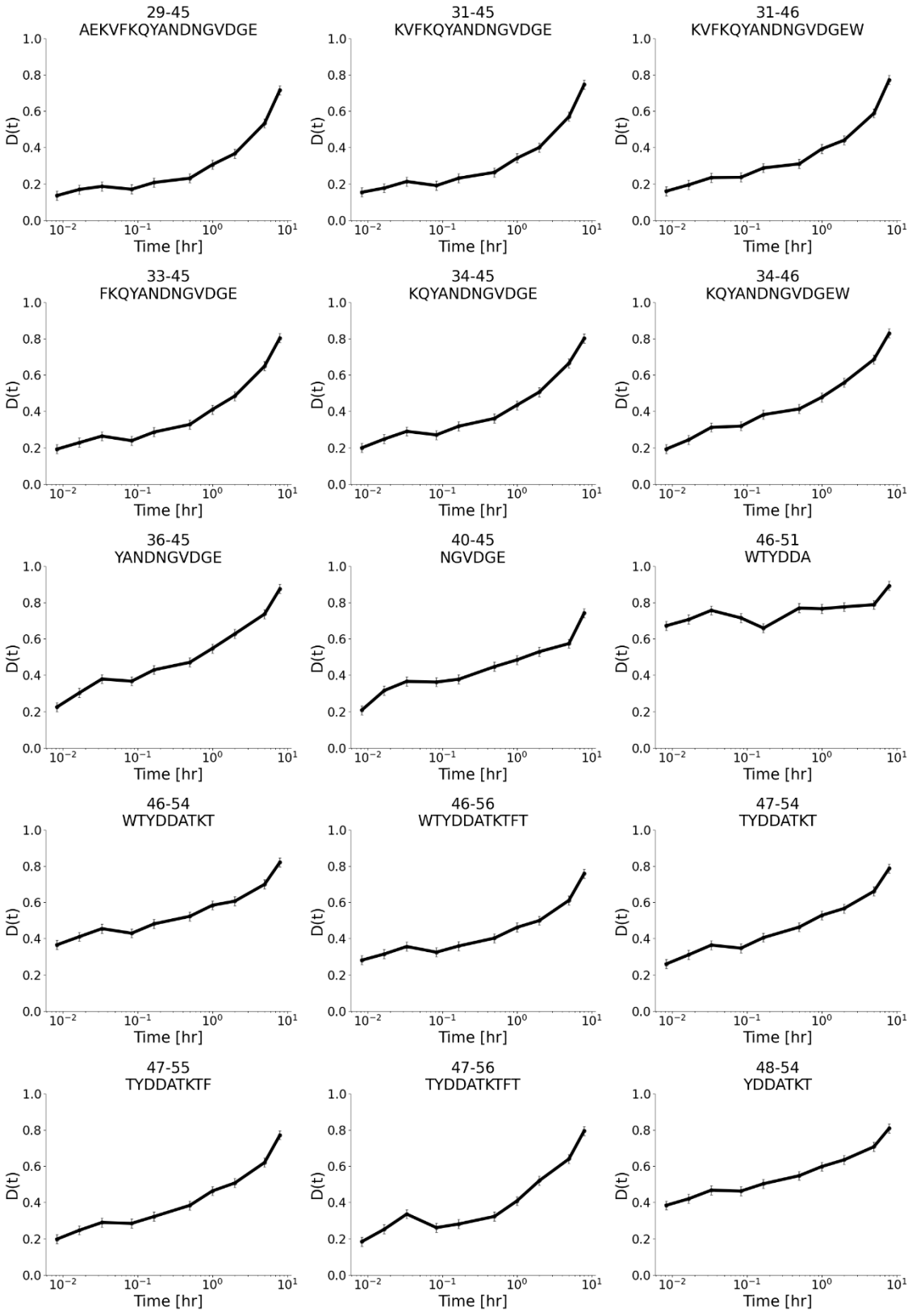
Supplementary Figure 5.7. Fractional uptake for all peptides identified by the HDX-MS experiments for GB1-TACC3.

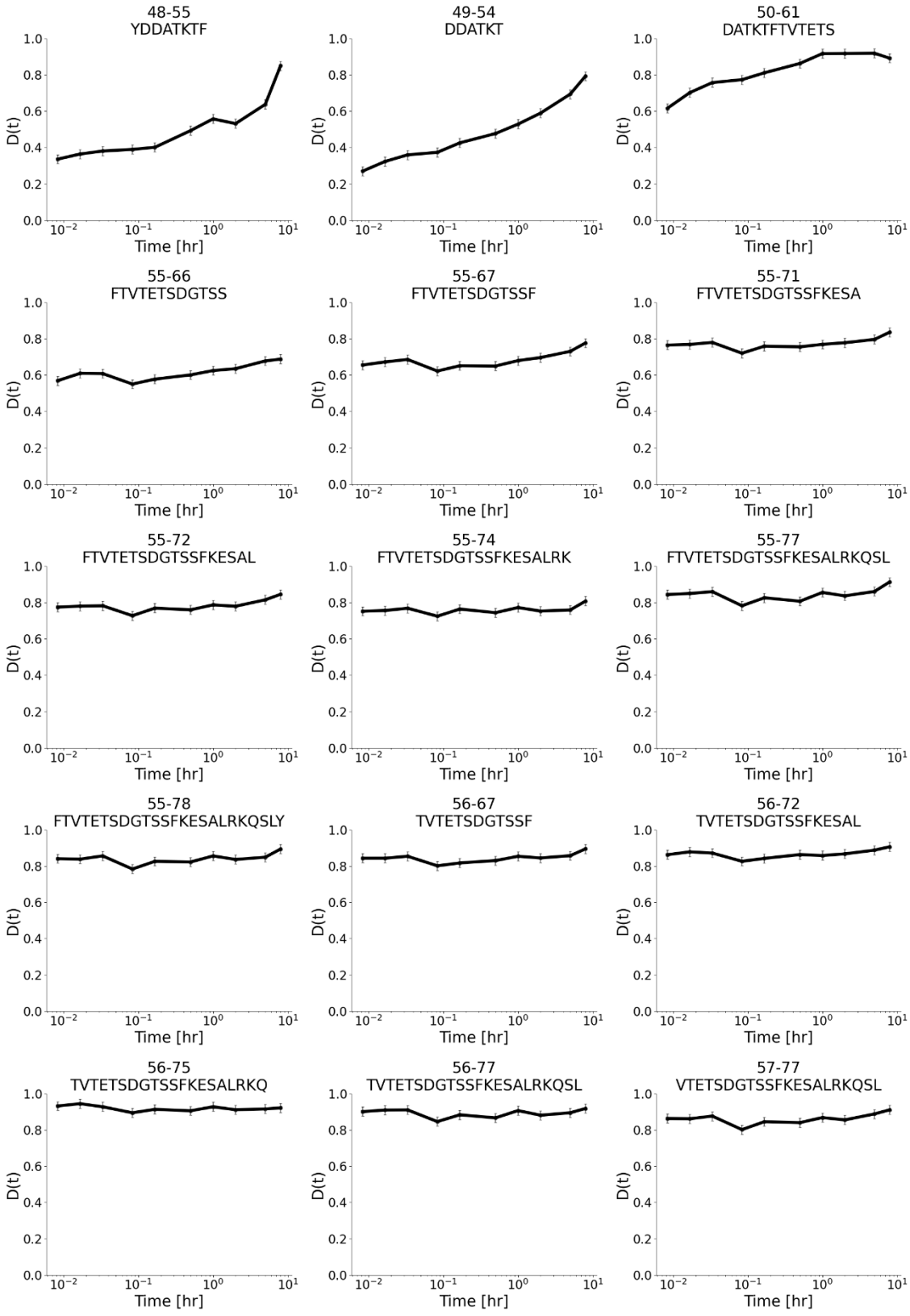
The error associated to experimental measurements is the pooled standard deviation $\boldsymbol{\sigma = 0.025}$.

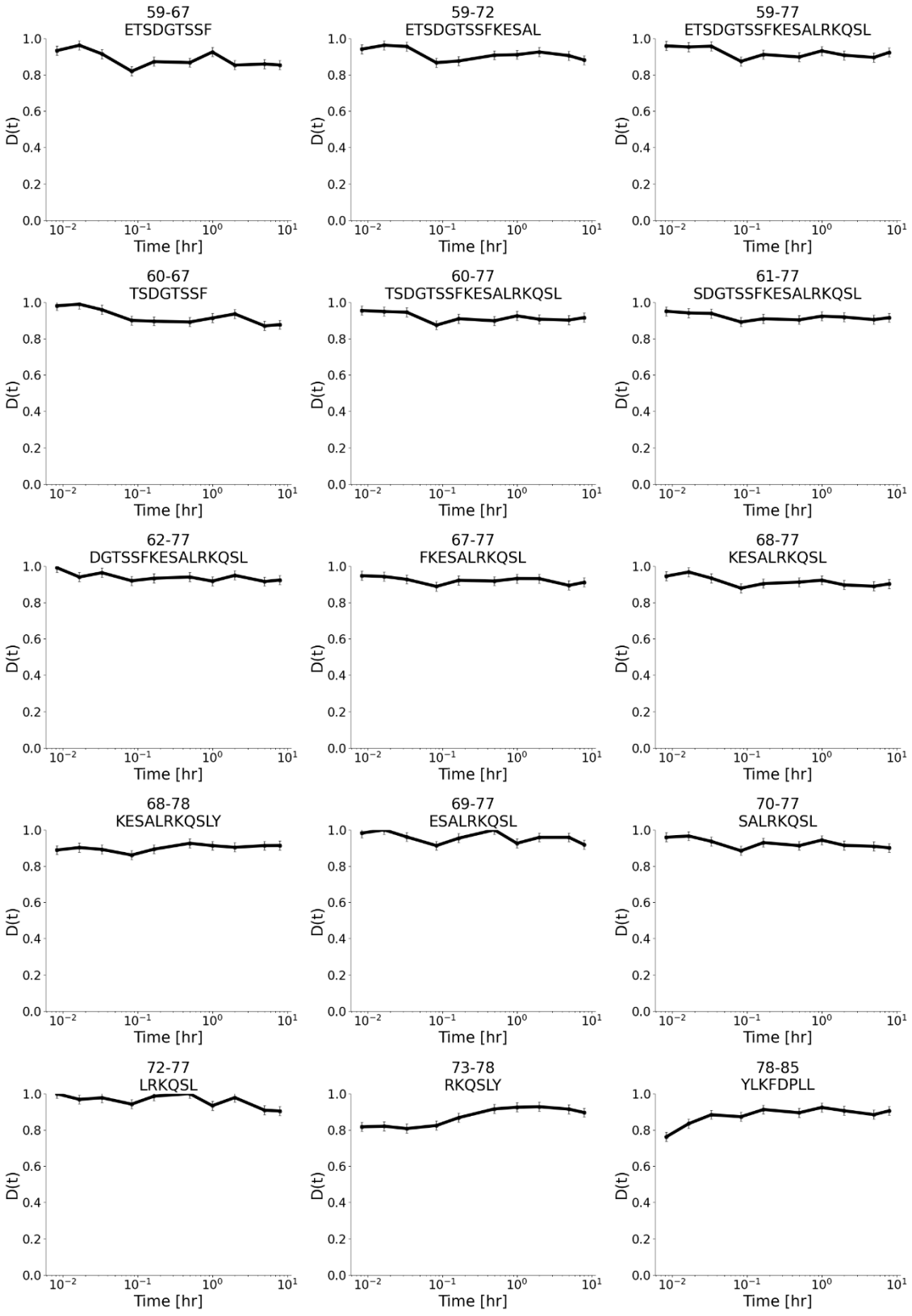


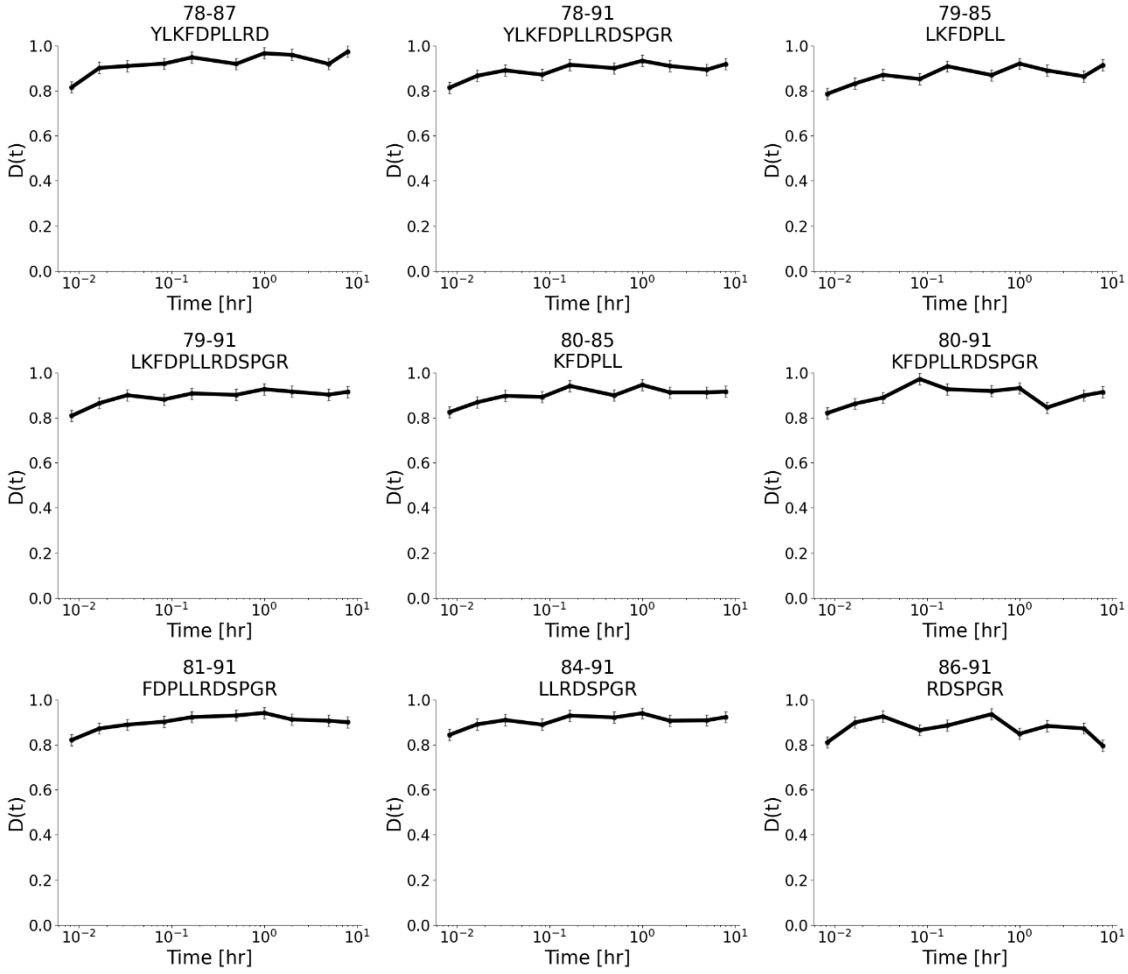

7-22
KLALNGKTLKGETTTE
7-25
KLALNGKTLKGETTTEAVD
8-22
LALNGKTLKGETTTE
9-22
ALNGKTLKGETTTE
9-25
ALNGKTLKGETTTEAVD
11-22
NGKTLKGETTTE
11-25
NGKTLKGETTTEAVD
23-29
AVDAATA
23-30
AVDAATAE
23-32
AVDAATAEKV
23-33
AVDAATAEKVF
23-35
AVDAATAEKVFKQ
23-39
AVDAATAEKVFKQYAND
23-45
AVDAATAEKVFKQYANDNGVDGE
23-46
AVDAATAEKVFKQYANDNGVDGEW
D(t)
Time [hr]


23-54
AVDAATAEKVFKQYANDNGVDGEWTYDDATKT
24-29
VDAATA
24-35
VDAATAEKVFKQ
24-46
VDAATAEKVFKQYANDNGVDGEW
25-35
DAATAEKVFKQ
25-45
DAATAEKVFKQYANDNGVDGE
26-33
AATAEKVF
26-35
AATAEKVFKQ
26-45
AATAEKVFKQYANDNGVDGE
26-46
AATAEKVFKQYANDNGVDGEW
26-54
AATAEKVFKQYANDNGVDGEWTYDDATKT
27-35
ATAEKVFKQ
27-45
ATAEKVFKQYANDNGVDGE
28-35
TAEKVFKQ
28-45
TAEKVFKQYANDNGVDGE
D(t)
Time [hr]


29-45
AEKVFKQYANDNGVDGE
31-45
KVFKQYANDNGVDGE
31-46
KVFKQYANDNGVDGEW
33-45
FKQYANDNGVDGE
34-45
KQYANDNGVDGE
34-46
KQYANDNGVDGEW
36-45
YANDNGVDGE
40-45
NGVDGE
46-51
WTYDDA
46-54
WTYDDATKT
46-56
WTYDDATKTFT
47-54
TYDDATKT
47-55
TYDDATKTF
47-56
TYDDATKTFT
48-54
YDDATKT
D(t)
Time [hr]

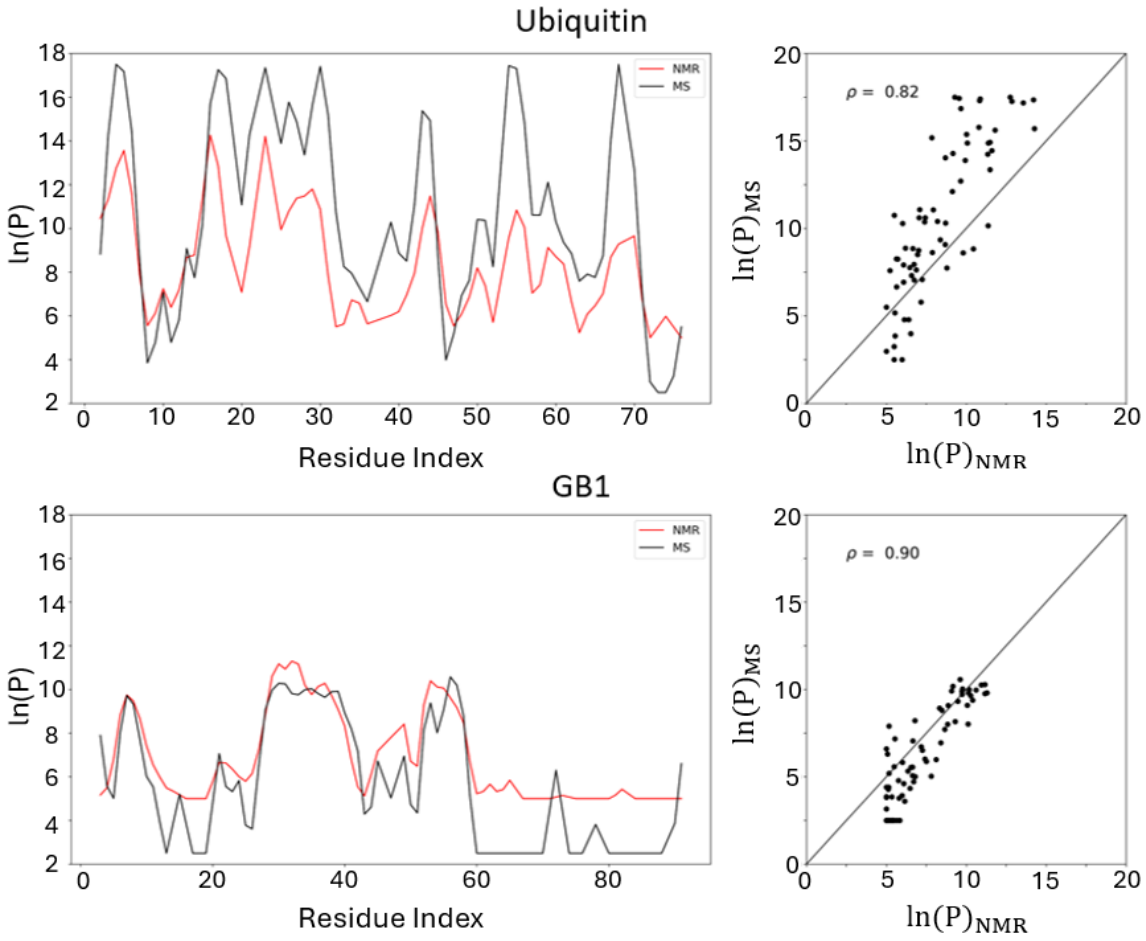

48-55
YDDATKTF
49-54
DDATKT
50-61
DATKTFTVTETS
55-66
FTVTETSDGTSS
55-67
FTVTETSDGTSSF
55-71
FTVTETSDGTSSFKESA
55-72
FTVTETSDGTSSFKESAL
55-74
FTVTETSDGTSSFKESALRK
55-77
FTVTETSDGTSSFKESALRKQSL
55-78
FTVTETSDGTSSFKESALRKQSLY
56-67
TVTETSDGTSSF
56-72
TVTETSDGTSSFKESAL
56-75
TVTETSDGTSSFKESALRKQ
56-77
TVTETSDGTSSFKESALRKQSL
57-77
VTETSDGTSSFKESALRKQSL
D(t)
Time [hr]


59-67
ETSDGTSSF
59-72
ETSDGTSSFKESAL
59-77
ETSDGTSSFKESALRKQSL
60-67
TSDGTSSF
60-77
TSDGTSSFKESALRKQSL
61-77
SDGTSSFKESALRKQSL
62-77
DGTSSFKESALRKQSL
67-77
FKESALRKQSL
68-77
KESALRKQSL
68-78
KESALRKQSLY
69-77
ESALRKQSL
70-77
SALRKQSL
72-77
LRKQSL
73-78
RKQSLY
78-85
YLKFDPLL
D(t)
Time [hr]


78-87
YLKFDPLLRD
78-91
YLKFDPLLRDSPGR
79-85
LKFDPLL
79-91
LKFDPLLRDSPGR
80-85
KFDPLL
80-91
KFDPLLRDSPGR
81-91
FDPLLRDSPGR
84-91
LLRDSPGR
86-91
RDSPGR
D(t)
Time [hr]

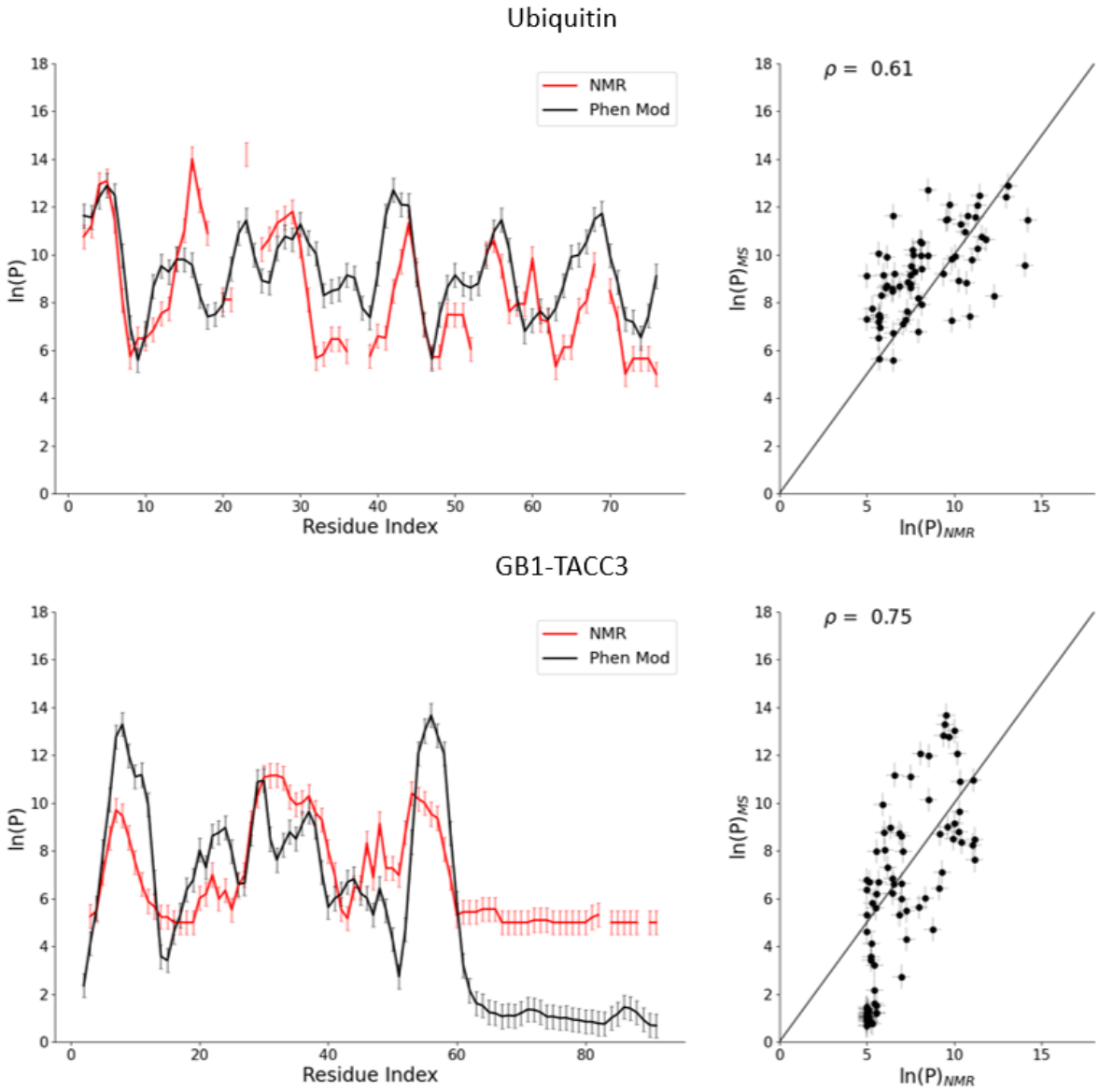
Supplementary Figure 5.8. Protection factors extracted by HDX-MS data using the reference NMR protection factors as initial guess.

Protection factors extracted by HDX-MS data (black) using the reference NMR protection factors as initial guess are compared with the NMR protection factors themselves (red). Results are shown for ubiquitin (top) and GB1-TACC3 (bottom). Correlation plots show a correlation coefficient $\boldsymbol{\rho = 0.82}$ for ubiquitin and $\boldsymbol{\rho = 0.90}$ for GB1-TACC3.



Supplementary Figure 5.9. Protection factors extracted from HDX-MS data using the structure driven initial guess.

Comparison between protection factors estimated using the Best-Vendrucolo model in **Eq. 5.7** (black) and the protection factors derived from NMR measurements (red) for ubiquitin (top) and GB1-TACC3 (bottom). Correlation plots report a correlation coefficient $\boldsymbol{\rho = 0.61}$ for ubiquitin and $\boldsymbol{\rho = 0.75}$ for GB1-TACC3.



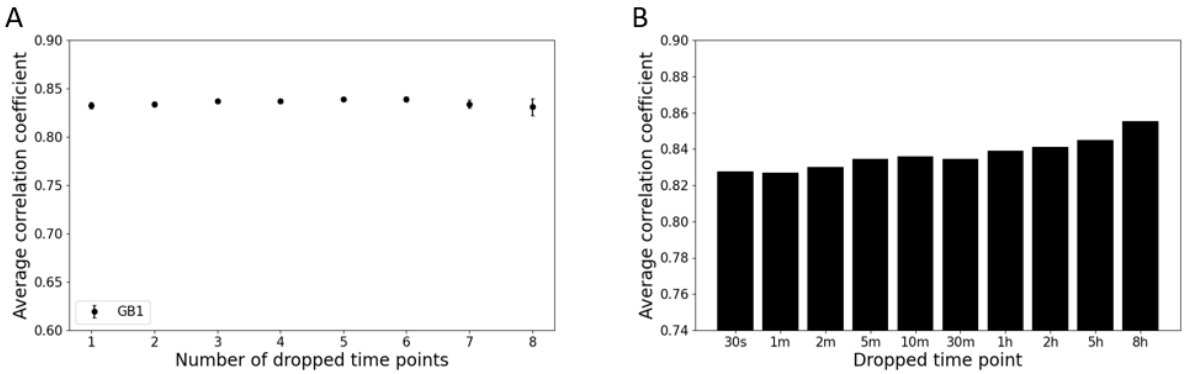
Supplementary Figure 5.10. Dependence of the correlation coefficients between protection factors extracted by NMR and MS on the number of peptides available.

Correlation coefficients between protection factors from NMR and protection factors extracted from HDX-MS data using four different approaches (structure driven initial guess, random search, data driven initial guess and unstructured initial guess) for different sub-datasets, containing 20, 40, 60, 80 and 100% of the peptides identified for ubiquitin and GB1-TACC3.

	Ubiquitin					GB1-TACC3					AVG
	20%	40%	60%	80%	100%	20%	40%	60%	80%	100%	
Structure	**0.65**	**0.71**	**0.71**	**0.80**	**0.82**	**0.77**	**0.83**	**0.81**	**0.82**	**0.83**	**0.78**
Random	**0.36**	**0.68**	**0.66**	**0.77**	**0.78**	**0.85**	**0.85**	**0.79**	**0.88**	**0.88**	**0.75**
Data driven	0.51	0.64	0.68	0.76	0.78	0.72	0.72	0.66	0.70	0.76	0.69
Unstructured	0.31	0.68	0.66	0.77	0.78	0.66	0.69	0.78	0.75	0.81	0.69

Supplementary Figure 5.11. Dependence of the protection factor analysis of HDX-MS data for GB1-TACC3 as a function of the number of time points in the dataset.

The structure driven initial guess was used, and the full dataset (100% peptides) was used. (A) The correlation coefficient is shown as a function of the number of dropped time points. The average value is calculated over all possible combinations dropping that specific number of time points. (B) The correlation coefficient is depicted as a function of the dropped time point. The value is averaged over all combinations containing the dropped time point.



Chapter 6. General Discussion

6.1 From physics to molecular biology: an unexpected meeting

I heard about hydrogen/deuterium exchange mass spectrometry for the first time in September 2018. At the time, I had just finished my BSc in Physics at the University of Trieste with a dissertation about Molecular Dynamics simulations, where I implemented a small old school (it was written in FORTRAN) software to perform simulations of constrained particles interacting through a Lennard-Jones potential. In the final section of the dissertation, I cited the work of a certain **Prof Emanuele Paci**, who was explaining Atomic Force Microscopy (AFM) experiments (Bustamante and Yan 2022; Hughes and Dougan 2016; Petrosyan, Narayan, and Woodside 2021) (the first biophysical technique I fell in love with) with Molecular Dynamics simulations (Best et al. 2003; Neelov et al. 2006; Paci and Karplus 2000). To my surprise, during the first week of my MSc in Applied Physics at the University of Bologna, a talk by that professor Paci was announced. During his lecture, he spoke about his work trying to predict the protection factors of a protein from a known structure (Radou et al. 2014; Vendruscolo et al. 2003), and his ideas on how to estimate protection factors from sparse and underdetermined mass spectrometry data.

One year later, I was working with him at the University of Leeds. Thanks to an Erasmus scholarship, I was able to write my MSc dissertation. Emanuele gave me a paper that he had recently published (Skinner et al. 2019), which proposed a method, ExPfact, to estimate the alternative patterns of protection factors in agreement with experimental HDX-MS data. The paper mostly focused on synthetic data and showed that even with ideal error free data, it is not possible to estimate a unique set of protection factors from peptide-level HDX-MS data, but it is possible to reduce the number of possible solutions to a finite number of sets. My main objective was to master the code (written by **Simon Skinner**) and the method, and applying it to a real-world dataset probing glycogen

phosphorylase that had previously been acquired by **Dr Jonathan Phillips** in Exeter (Kish et al. 2019).

In November 2019, Emanuele asked me to follow him in a meeting with a researcher who had asked his opinion on limited proteolysis (LiP) mass spectrometry data (Cappelletti et al. 2021; Fontana et al. 2004; Schopper et al. 2017). That was my first meeting with **Prof Frank Sobott**. Frank showed us some data acquired by **Dr Rani Moons** (at the time PhD student at the University of Antwerp), where they were monitoring the unspecific digestion of α-synuclein by proteinase K as a function of digestion time. The question that Rani and Frank had was how to identify statistically significant changes in the cleavage pattern of α-synuclein in its monomeric, oligomeric (with addition of EGCG or DA) or fibrillar state. Emanuele saw an interesting analogy with HDX-MS data, where the deuteration of the proteolytic peptides is replaced by their label-free quantification (LFQ) intensity (Cox et al. 2014; Tyanova, Temu, and Cox 2016). My task was to implement a pair-wise comparison of conditions to provide a list of peptides with statistically significant changes. After selecting the subset of peptides present in both conditions under analysis, we compared the intensities with a Welch's t-test. To reduce the false discovery rate, we used a hybrid significant testing approach implemented in the context of HDX-MS data (Hageman and Weis 2019). We visualized the results via volcano plots (**Figure 6.1**), where a significant threshold α is applied on the p-value ($\alpha = 0.01$) and on the difference in LFQ intensity (greater than the pooled standard deviation). Using this strategy, we were able to identify and localize within the sequence of α-synuclein statistically significant changes concerning the oligomeric states with respect to the monomeric state, while the cleavage pattern of the monomeric and fibrillar state was similar.

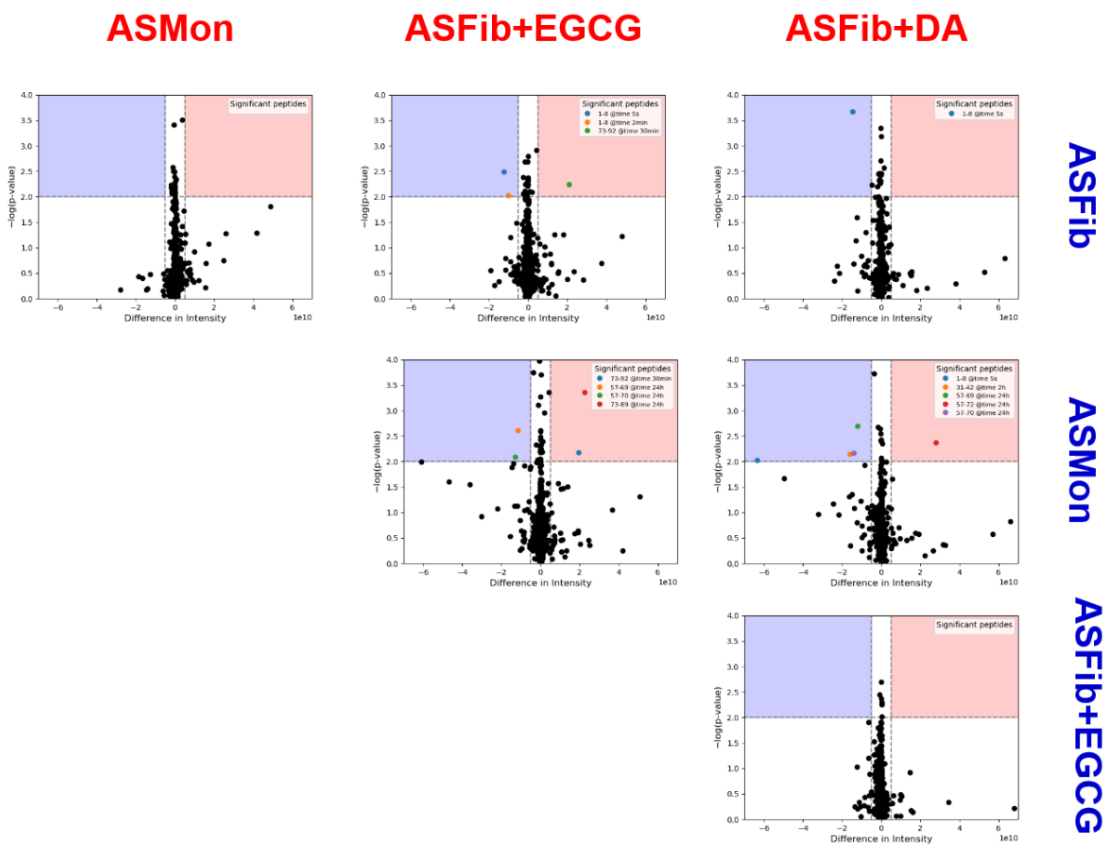


Figure 6.1. LiP-MS of α-synuclein.

Volcano plots identifying significant peptides in a pair-wise comparison of four different experimental conditions: monomeric (ASMon), fibrillar (ASFib) and two oligomeric states with addition of EGCG (ASFig+EGCG) or DA (ASFib+DA). LFQ intensities of peptides identified in both conditions under analysis were compared with a Welch's t-test. The difference in intensity is plotted against the p-value (returned by the t-test). A significance threshold was set at $\boldsymbol{\alpha = 0.01}$. To correct for false discovery rate, we set a significance threshold on the difference in intensity (horizontal axis) at the value of the pooled standard deviation. Data acquired by Dr Rani Moons and analysed by the author.

I was blessed by the fact that Frank and Emanuele had an open shared PhD position to work on these same topics. I finished my MSc degree, applied for the position, and in October 2020 I moved to Leeds to officially start my PhD.

6.2 Validating ExPfact between lockdown and a connection with India

When I started the PhD, I was already mastering the computational method implemented by Simon and Emanuele to estimate protection factors from HDX-MS data (Skinner et al. 2019). Starting from any HDX-MS dataset, we were able to extract alternative patterns of protection factors in agreement with experimental data. The first question that Emanuele asked me to answer was straightforward: are we extracting the *true* protection factors or meaningless random numbers? Knowing that NMR was able to detect the HDX of proteins at the level of individual residues, I started an extensive literature search: I was looking for a published dataset containing HDX data for the same protein with both NMR and MS experiments. In my mind, I already had the idea to acquire such a specific dataset myself, but the pandemic kept me from accessing any laboratory during the first year of the PhD.

I was able to find only one dataset, acquired by **Dr Roumita Moulick** at the Tata Institute for Fundamental Research (Bengaluru, India), that was probing the HDX of the mouse prion protein using both experiments (Moulick et al. 2015). I contacted the authors, which were very keen on sharing their data. **Chapter 3** is the result of this exchange. The HDX-MS dataset had a scarce overlap (or redundancy, see **Figure 3.1**), indeed the authors were originally using NMR experiments to retrieve information about areas of the protein not covered by mass spectrometry (and vice versa). Nonetheless, we applied ExPfact to estimate the different patterns of protection factors in agreement with HDX-MS data (**Figure 3.4**). We found that at least one of the solutions was in agreement with NMR data. Where multiple solutions were available, we needed to find a way to select the *best* solution. To do so, we decided to exploit the additional information contained in the shape of the isotopic envelope of the peptide (rather than the intensity-weighted average only). Simulating the time evolution of the isotopic envelope starting from the alternative sets of protection factors (with the protocol described in **section 3.3.4**), we found that some sets provided a better prediction of the

isotopic envelope than others (**Figure 3.5**). We therefore proposed this two-step strategy to estimate protection factors from HDX-MS data: i) calculate the alternative solutions in agreement with experimental data; ii) rank these alterative patterns based on their ability to predict the isotopic envelope of the peptide. Using this strategy, we found a promising (yet not outstanding) correlation coefficient $\rho = 0.71$ between our protection factors estimated from HDX-MS data and NMR experiments **(Figure 3.6**).

I make a jump ahead in time, but it is worth mentioning here that I managed to present the results of this project at the International HDX-MS conference (April 2022, London) through a flash oral presentation and a poster. I was so surprised by the interest elicited by our research that the idea of acquiring an ad-hoc high-quality dataset for testing our protection factor analysis method (as well as others) became stronger in my mind. I wanted to provide the HDX-MS community with a dataset probing one (or possibly more) proteins with both MS and NMR that could be used to test the computational methods aiming to extract single-residue information from peptide-level HDX-MS data against *true* NMR data. It was during the same conference that several conversations led me to believe that the only way to extract accurate protection factors from HDX-MS data was to combine this data type with an alternative technique, and I immediately thought about Molecular Dynamics simulations. Can we overcome the underdetermination of HDX-MS data by integrating the results of MD simulations?

6.3 Into the laboratory

Going back to my first academic year in Leeds (2020/2021), while waiting for Covid restrictions to decline and allow me to start some activity in the laboratory, I focused on extensively reviewing the theory and data analysis tools of HDX-MS experiments. That is when I started writing the comprehensive review of the state-of-the-art tools to analyse HDX-MS data now contained in **Chapter 2**.

It was only in October 2021, more than one year into my PhD, that I was finally able to access the laboratory and perform some HDX-MS experiments. A premise is due here:

as a physicist, I had never used a pipette, performed a dilution or made a buffer before. This is why I decided to dedicate my first week on the instrument (Waters Synapt G2-Si) to study the HDX of a model protein (myoglobin) by repeating the procedures for sample and buffer preparation. I acquired a dataset with a total of 18 replicates: 6 technical replicates from 3 manipulation replicates (see **section 2.4.3**). Using traditional HDX-MS tools for data analysis, namely DynamX and Deuteros 2.0 (Lau et al. 2020a), I expected to find no differences between the different manipulation replicates. Surprisingly, this was not the case: several false positives had been identified. At first, I strongly doubted my experimental skills, but I decided to speak with **Dr Jeanine Houwing-Duistermaat** (then at the Departments of Statistics, School of Mathematics, University of Leeds) about my problem, and she suggested to try and use linear mixed models to model the different sample procedures as random effects (Pinheiro et al. 2022). Surprisingly for me, unsurprisingly for her, this solved the issue. Different types of replicates introduce different kinds of errors.

I think that my first experimental week taught me something unexpected yet important. I wanted to learn basic laboratory skills; I turned out to learn new statistical tools. This project led to the birth of the statistical section in **Chapter 2** (**section 2.5.2**): I wanted to learn more about the statistics needed to analyse HDX-MS, and in particular how the appropriate statistical test should be chosen depending on the experimental design implemented. This focus of my project on statistical aspects was also boosted by a collaboration with **Dr Rene Frank**, who asked Frank's opinion about and help on the statistical meta-analysis of proteomics data (Andrews et al. 2022).

6.4 Testing intrinsic exchange rates with millisecond HDX

During the literature review that led to **Chapter 2**, I read several papers that were criticizing the intrinsic exchange predictions described in **section 1.1.1** (Al-Naqshabandi and Weis 2017; Del Mar et al. 2005; Keppel and Weis 2013; Mori et al. 1997; Zhang et al. 2012). The ability of our method (ExPfact) to extract protection factors from HDX-MS data highly depends on the accuracy of these predictions, as they

are modelled as known constants. Therefore, we decided to test this hypothesis by studying the HDX of unstructured peptides. However, peptides exchange in the millisecond regime, and we did not have the capability to perform fast HDX experiments in the facility. Frank managed to organize a meeting with **Dr Jonathan Phillips** and **Dr Neeleema Seetaloo** at the University of Exeter, who kindly hosted me as a visitor in July 2022.

This collaboration yielded the results discussed in **Chapter 4**. We decided to "create" a mixture of peptides with increasing sequence length (**Supplementary Table 4.1**) and we studied their HDX behaviour in absence of salt. We compared the experimental curves with the theoretical ones obtained using the intrinsic exchange rate calculations provided by the Englander group (**section 1.1.1**), obtaining the results shown by the red curve in **Figure 4.1**. The results were consistent with the criticisms raised by the above-mentioned papers: the experimental deuterium uptake curve exchanged (slightly) faster than the predicted one. In principle, this should be impossible, as the intrinsic exchange rate should represent the rate of a residue in a completely unfolded structure. I had to read the original paper by Bai et al. (1993) several times before realizing that they had originally measured the reference exchange rate constants (**Table 4.1**) for two unstructured model systems: a three-alanine (3-Ala) peptide and a poly-DL-alanine (PDLA) peptide. In our version of the Englander spreadsheet, we were only considering the reference values of PDLA. I tried to switch the reference from PDLA to 3-Ala and re-calculated the predicted uptake curves, finding the green curve in **Figure 4.1**. The predictions were better when using the 3-Ala reference instead of PDLA. Why? Did we have a rationale to suggest using 3-Ala instead of PDLA?

The PDLA model used by Bai et al. (1993) had a degree of polymerization dp = 28, which means that the authors had studied an alanine-based peptide mixture with an average length of 28 residues. Emanuele mentioned some unpublished simulation studies that he had performed on poly-valine peptides where he could see that after a certain number of residues (around 50) the system tended to form β-sheets instead of

α-helices. We questioned whether the PDLA model with such numbers of residues would retain some residual structure, which would be responsible for the slower exchange. I asked Dr Alex St John to run some short simulations of poly-alanine peptides at increasing length and calculate their secondary structure propensity. These results, shown in **Figure 4.2**, validated our hypothesis: 8-12 alanine residues are enough to form α-helices. This results were supported by several experimental and simulation studies that we found in literature (Chakrabartty et al. 1994; Hinck 2022; Ingwall et al. 1968; Kuczera et al. 2021; López-Llano et al. 2006; Rohl et al. 1999). We tried to order some PDLA peptides (with a length of approximately 28 residues), but all peptide synthesis companies replied saying the peptide was too hydrophobic to be produced. To us, this was a further red flag that PDLA was probably not the best choice as an unstructured reference for HDX experiments. To further test our hypothesis, we asked **Prof David Weis** if he could share with us some data he had recently published about the exchange of unstructured or disordered peptides (Al-Naqshabandi and Weis 2017). He kindly sent us the deuterium uptake curves of several peptides, which we compared with the theoretical predictions, using either 3-Ala or PDLA as reference. The results (**Figure 4.3**) showed that the predicted curves either matched or were faster than the experimental ones. This confirmed our hypothesis that 3-Ala is a better unstructured reference than PDLA because the latter retains some residual structure. However, it raised a further question: why was the predicted uptake faster than the experimental one?

During the time spent in Exeter, we managed to perform some experiments probing the exchange of the peptide mixture under different salt concentrations (150, 500 and 1000 mM NaCl and CsCl), and we noticed that the introduction of salt in the buffer slowed down the exchange **(Figure 4.4**). In their work, Al-Naqshabandi and Weis (2017) performed HDX-MS experiments in a buffer containing 100 mM NaCl. We suggested that this could be the answer: the experimental exchange in **Figure 4.3** is slower than the predicted one because the presence of salt in the experimental buffer is slowing down the exchange. It was hard to quantify whether the discrepancy was completely

explained by this factor as the dependence of the intrinsic exchange rate on salt is not well studied and cannot be predicted. We show here the dependence of HDX kinetics of bradykinin (one of the peptides in the mixture) on the concentration of CsCl (**Figure 6.2**). Our data suggest that high salt concentrations ($\geq 150\, mM$) slow down the exchange (**Figure 6.2A**), reducing the observed exchange rate of the peptide by approximately 50% (**Figure 6.2B**). A differential analysis (using the buffer with no salt as reference state) showed that the differences in deuteration introduced by varying salt concentrations are minor yet statistically significant (**Figure 6.2C-E**), especially at shorter time points (i.e. before the plateau is reached). Similar observations were made for the other peptides. These findings motivated us to further study this salt dependence, especially in the range between 0 and 150 mM. We planned a second visit to the University of Exeter in February 2024, but technical issues with the instrument prevented us from acquiring a conclusive dataset.

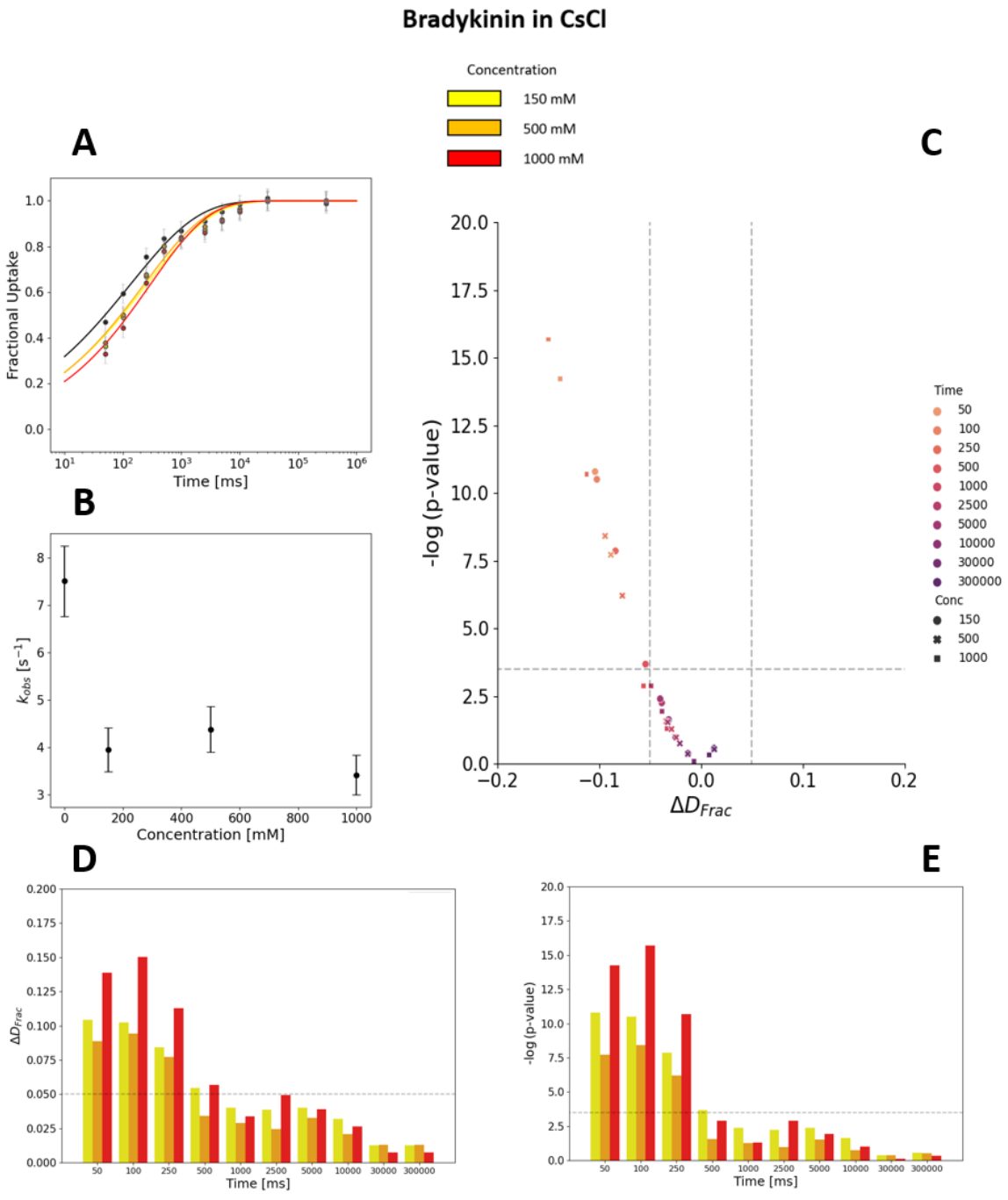


Figure 6.2. The intrinsic exchange rate's dependence on salt concentration for bradykinin in CsCl. A) Experimental data showing the fractional uptake of bradykinin in absence of salt (black) or in presence of 150 mM (yellow), 500 mM (orange) or 1 M (red) CsCl. Experimental data are fitted with a stretched exponential $\mathbf{1} - \boldsymbol{e}^{-(\boldsymbol{k_{obs}t})^{\boldsymbol{\beta}}}$. Error bars are not shown for visualization purposes. B) The fitted

k_{obs} are shown as a function of concentration. The dashed line shows the results of a best fit with a parabolic curve. C) Volcano plot: the horizontal axis shows the difference between the fractional uptake of a data point – at a specific concentration (150 mM CsCl – circles; 500 mM CsCl – crosses; 1 M CsCl – squares) and a specific time (different colours show different times, darkest points are longest time points) – and the fractional uptake of the same time in absence of salt. The vertical axis shows the negative logarithm of the p-value returned by the linear model in Equation 4and adjusted with false discovery rate (FDR) correction. Differences are considered significant when $|\Delta \boldsymbol{D_{Frac}}| > \mathbf{5}\%$ (vertical dashed lines) and p-value < 0.01 (horizontal dashed line). D-E) Projection of the volcano plot on the horizontal axis (D) or the vertical axis (E). Differences are shown as a function of time (horizontal axis) for different concentrations (150 mM – yellow; 500 mM – orange; 1 M – red) either in terms of $\Delta \boldsymbol{D_{Frac}}$ (D) or –log(p-value) (E). Thresholds (horizontal dashed lines) are analogous to (C).

6.5 Building up experience with HDX-MS projects

The third year of my PhD was characterized by several side projects that helped me building up experience with HDX-MS experiments (also thanks to the endless nights spent troubleshooting the instrument). In collaboration with **Prof Bruce Turnbull**, **Dr James Ross** and **Dr Alex St John**, we performed HDX-MS experiments on Cholera Toxin B (CTB) to check the binding site of different sugars (GM1os, Lewis-x and lactosyl azide) to the protein (Au et al. 2024; Ross et al. 2019). In collaboration with **Dr Qian Wu** and **Tiago Moreira**, we performed HDX-MS experiments on BRCT, the C terminus domain of BRCA1, a DNA-repair protein that serves as a marker of breast cancer susceptibility (Wu 2020). In collaboration with **Dr Julian Streit** and **Prof John Christodoulou** from UCL (London, UK), we attempted to study the HDX kinetics of ribosomal nascent chains (Burridge et al. 2021; Wales et al. 2022). In collaboration with **Prof Andrew Quigley** and **Dr Peter Harrison** at the Diamond Light Source Ltd (Oxford, UK), we performed HDX-MS experiments with the aim of mapping nanobody binding epitopes on integral membrane proteins (IMPs) (Martens and Politis 2020).

During the latter project, I had some time to test the idea that I had in mind since the HDX-MS conference in London: combining HDX-MS data and MD simulations to achieve single residue resolution. The idea was relatively simple: perform a short fully atomistic MD simulation of the membrane protein (PglL, **Figure 6.3A**), use the Best-Vendruscolo

model (**Eq. 1.11**) to calculate its protection factors, and use that estimate as initial guess to fit experimental HDX-MS data using ExPfact. I managed to calculate an optimized set of protection factors for the protein alone and the protein bound to a nanobody (NB32). I then performed a residue-level differential analysis, simply subtracting the protection factors of the bound state to those of the unbound state (**Figure 6.3C**). Major differences ($\Delta \ln(P) > 2$) were identified in the region between residues 260-300 (corresponding to the nanobody binding site). Qualitatively, the results correlated well with the results provided by a Woods plot, obtained using Deuteros 2.0 (**Figure 6.3B**).

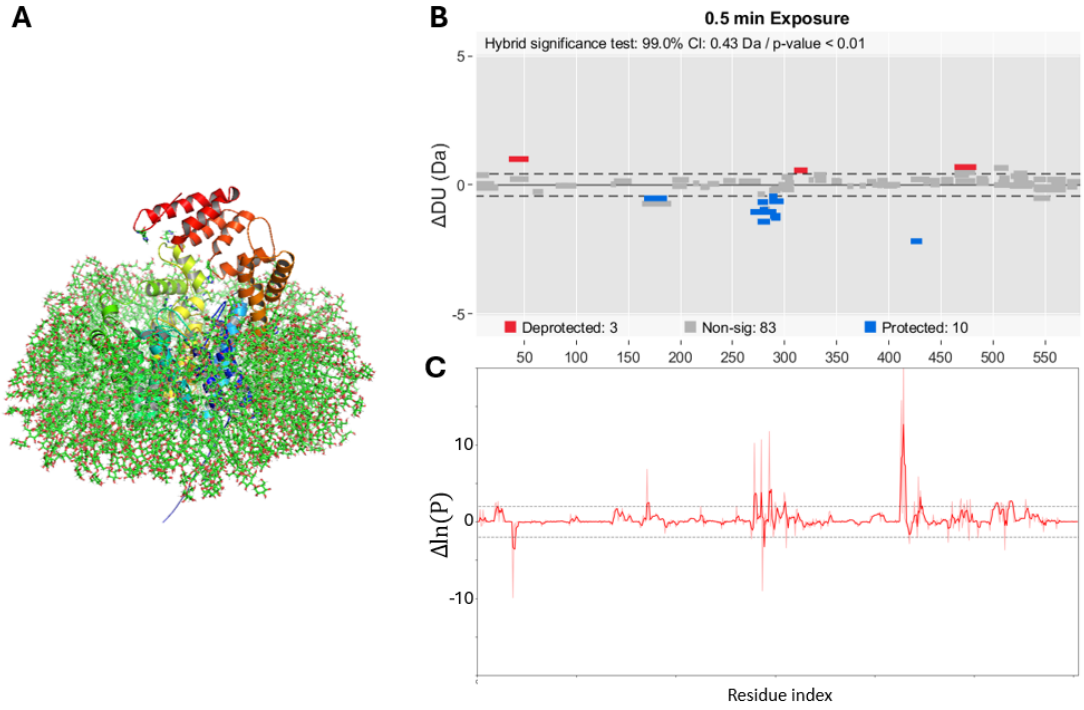


Figure 6.3. Protection factor analysis for membrane protein PglL provides results consistent with traditional differential analysis.

A) The structure of the protein PglL, studied with HDX-MS experiments and fully atomistic MD simulations. HDX-MS experiments were performed for the protein alone and the protein bound to a nanobody (NB32). B) Woods plot obtained by a traditional differential analysis of HDX-MS data, identifying the binding site within residues 260-300. C) Differential analysis performed by subtracting the protection factors obtained by ExPact for the bound/unbound state, providing similar results to the

traditional approach (at least qualitatively). The protein was purified by Dr Peter Harrison, data were acquired and analysed by the author.

6.6 Our protection factor analysis for HDX-MS data

The promising results provided by the protection factor analysis performed on PglL, which at an initial stage was only another test of our methodology in a real-world scenario, gave a boost to the idea that the underdetermination of HDX-MS data could be overcome by combining the experimental data with MD simulations. In September 2023, after a discussion with **Dr Anastasia Zhuravleva**, we finally decided to acquire a dataset probing the HDX of two proteins with both MS and NMR. Our ultimate goal was to test our methodology, highlighting its strengths and weaknesses in deriving protection factors from HDX-MS data. The results of this collaboration are outlined in **Chapter 5**. We tested ExPfact in different scenarios, the most relevant being i) absence of prior information on the protein, and ii) availability of known protein structure.

In the absence of information on the protein, we run ExPfact as already presented in **Chapter 3**. A minimization procedure is repeated multiple times, allowing to estimate the alternative patterns of protection factors in agreement with HDX-MS data. If the peptide overlap is high enough, the underdetermination is solved, the average protection factors can be reported as final estimates, providing a strong correlation with NMR data (**Figure 5.4**). Instead, if the overlap is not sufficient to overcome underdetermination and multiple solutions are identified, these can be ranked exploiting the information encoded in the isotopic envelopes of the peptides (as described in **section 3.3.4**).

If a protein structure is known, we can initialize the minimization procedure using a physical guess provided by the Best-Vendruscolo model (**Eq. 1.11**). This allows to explore only a local minimum of the cost function landscape (**Figure 5.1** and **Figure 5.6**) and to avoid the finding of unphysical solutions (e.g. assigning high protection factors to residues in a loop). If the overlap is high enough, the estimated protection

factors show a strong correlation coefficient with NMR data, similar to the one provided by the random search (**Figure 5.5**). The real advantage of this structure driven approach with respect to the random search is that it is more robust upon peptide removal: the correlation with NMR data is less dependent on the number of peptides available in the HDX-MS dataset (**Figure 5.7**). In the world of AlphaFold (Jumper et al. 2021), this approach can be achieved using a protein model rather than an experimental structure.

These two approaches represent my answer to the research question of this Thesis: can we infer single-residue resolution (i.e. protection factors) from peptide-level HDX-MS data? The answer is: yes, with limitations. "Yes" because we have shown that from a high-quality HDX-MS dataset (high redundancy, high number of spread time points) we can estimate protection factors with a strong correlation with NMR. "With limitations" because the final estimates still depend on parameters that need further research to be addressed, as described in the following **section 6.7**.

6.7 Open challenges in HDX-MS

The ability of extracting accurate protection factors from HDX-MS data is limited by a number of factors that could be addressed by additional fundamental research on the phenomenon and the technique.

6.7.1 Controlling redundancy

The digestion pattern can be optimized, e.g. by using a cocktail of different proteases or optimizing the composition of the quench buffer, but there is no guarantee to reach 100% coverage with a sufficiently high redundancy. The integration of fragmentation techniques that avoid H/D scrambling, such as ECD, ETD or UVPD, has been proven useful to increase the spatial resolution (hence the redundancy) of HDX-MS data, especially for studying some areas of interest of the protein system at high resolution (Rand 2013). However, this will not completely solve the problem of deriving single-residue resolution information, especially when studying larger protein complexes,

because the proportion of cleavage sites decreases with the protein size (Rand et al. 2014), which translates into limited spatial resolution. Building on the methods proposed in this Thesis, single-residue information obtained directly from fragmentation techniques can be easily integrated in the workflow. Protection factors extracted from ECD/ETD/UVPD data, even when available only for a limited number of residues, can serve as additional constraints in the minimization procedure, limiting the search to those patterns of protection factors that are consistent with both peptide-level and fragment-level experimental data.

6.7.2 Adaptive temporal sampling

The number and spread of time points available depends on the materials available (protein, buffers etc.), on the experimental time available on the instrument (in other words, the budget available) as well as on the instrument itself, whether it is a traditional instrument or a millisecond setup. Ideally, one should be able to sample the whole deuterium uptake curve of every peptide for a quantitative analysis. More than how many time points are needed, the question is what time points are needed. In an attempt to answer this question, we have provided a tool to estimate the exchange time range for a protein (**Figure 5.9**), which could be used to guide the choice of the labelling time points.

6.7.3 Validity of the intrinsic exchange rate predictions

To extract protection factors from experimental HDX-MS data (as well as HDX-NMR data), it is important to deconvolve the effect that the buffer (pH, temperature, ionic strength, viscosity etc.) has on the exchange from that of the protein structure. In other words, the intrinsic exchange rates need to be known. We have shown that the k_{int} predictions are reliable in absence of salt. The more complex the buffer composition, the more these predictions need further testing. With HDX-MS applications tending to probe more and more complex systems – there are studies attempting to understand

the behaviour of proteins in-vivo (Lin et al. 2022) – the effect of the intrinsic exchange rate becomes more and more crucial even for differential studies.

6.7.4 Availability of isotopic envelopes

We have shown that the time evolution of the isotopic envelope contains more information than the intensity-weighted average, but this information is generally difficult to export from commercial software. Open-source software packages such as Mass Spec Studio (**section 2.5.1**) started to provide direct access to the raw data.

6.7.5 Back-exchange

Another important assumption of our method (ExPfact) is that a normalization of the intensity-weighted average with a fully deuterated sample is sufficient to correct for back-exchange. However, back-exchange occurs at the residue level, and the same residue in different proteolytic peptides may have different back-exchange levels (Sheff et al. 2013).

We have attempted to model back-exchange using the dataset probing the millisecond HDX of unstructured peptides (**Chapter 4**). We simulated the time evolution of the isotopic envelope using a procedure similar to the one described in **section 3.3.4:** i) we calculated the fully protonated envelope (**Figure 6.4A**); ii) we predicted the shape of the isotopic envelope at experimental time t ($t = 500\,ms$ is shown in **Figure 6.4B**) under labelling buffer conditions ($pD = 7.0$, $T = 20\,°C$); iii) we adjusted the shape of the isotopic envelope for reverse-exchange occurring during labelling because experiments were performed in a 95% deuterated buffer (**Figure 6.4C**); iv) we calculated the shape of the isotopic envelope for an effective back-exchange time ranging from 0 to 50 minutes; v) we compared the predicted envelope with the experimental one using R^2 (**Figure 6.4E**) and we recorded the effective back-exchange time with best agreement τ (**Figure 6.4D**).

We estimated the effective back-exchange time τ for the three peptides in our mixture from all experimental labelling timepoints (**Figure 6.4F**). Assuming that it is independent of labelling time, the average τ (red line in **Figure 6.4F**) was 7.98 ± 1.59 minutes for angiotensin, 9.77 ± 0.46 minutes for bradykinin and 7.73 ± 0.84 minutes for ANP. The effective back-exchange time of the peptides should in principle be the sum of the trapping time (3 minutes) and their experimentally determined, peptide-specific retention times (blue lines in **Figure 6.4F**) which are defined as the period after trapping up to MS detection. The average back-exchange observed is greater than expected in all cases, by between approximately 1.5 to 3 minutes. This discrepancy can be due to i) inaccuracy of the reference exchange rate constants for reverse exchange (the reference constants of 3-Ala for back-exchange are not available, see **Table 4.1**), ii) experimental uncertainty on the intensities of the experimental isotopic envelopes (here averaged across the replicates), iii) experimental uncertainty on the retention time (retention time tolerance was set to 0.5 minutes in DynamX) iv) additional dead times from quench/mixing and v) back-exchange occurring during ionization and in the gas-phase. Reducing these discrepancies would more accurately reproduce the time evolution of the isotopic envelope of an unstructured peptide from the knowledge of sequence and retention time of the peptide together with pH, temperature and ionic strength of the buffer. Nevertheless, we showed that within current limitations, it is in principle possible to produce a forward model for peptide level exchange that can then be used to isolate the 'labelling phase' kinetics of the protein ensemble from the 'analysis phase' kinetics of protein and peptide ensembles.

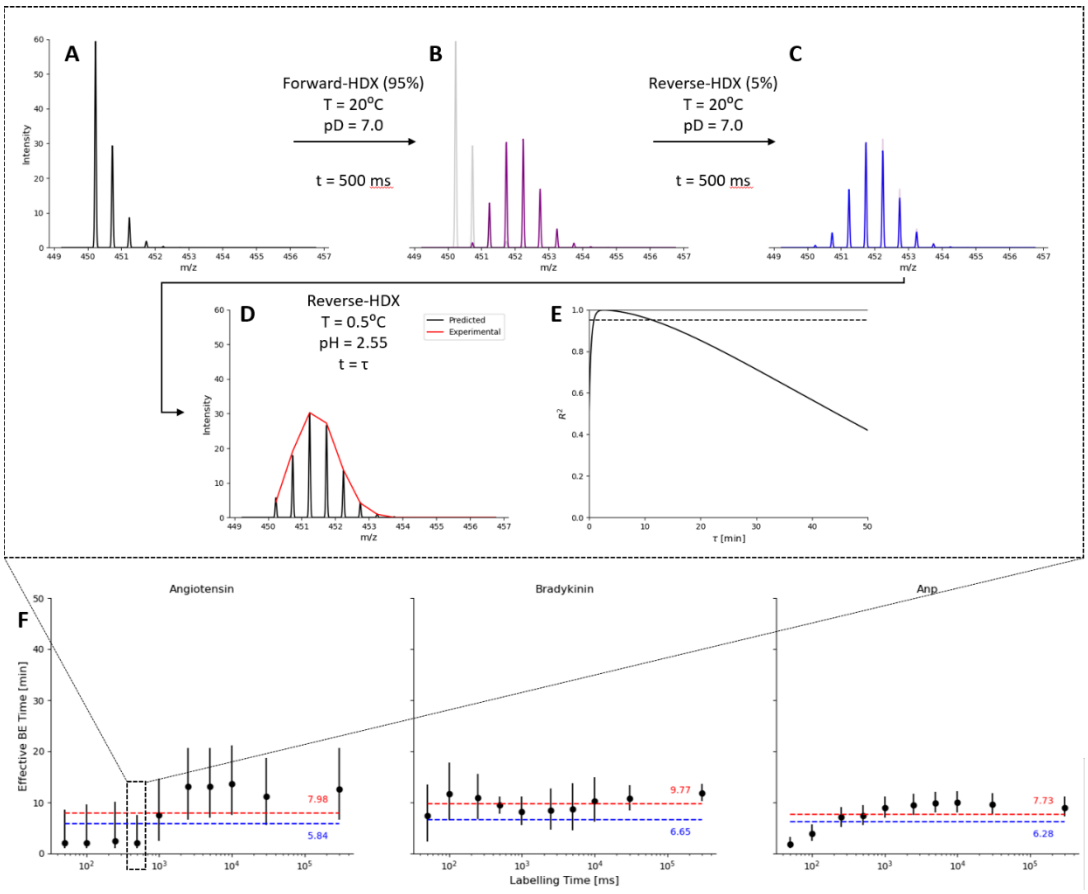


Figure 6.4. Isotopic envelope calculations to estimate the effective back-exchange time.

Insert: example of isotopic envelope calculations for angiotensin at time t = 500 ms. A) The fully protonated envelope (black) is calculated. B) The fully protonated envelope (black) evolves towards deuteration under labelling conditions (pD 7.0, T=20˚C) for a labelling time t (purple). C) The resulting isotopic envelope (purple) evolves towards protonation under the same labelling conditions because the experiment was performed in 95% deuterated solution. D) The resulting envelope evolves towards protonation under quench conditions (pH = 2.55, T = 0.5 ˚C) for an effective time $\boldsymbol{\tau}$. The predicted isotopic envelope (black) is compared with the experimental one (red) E) The agreement between the predicted and measured isotopic envelope is evaluated with $\boldsymbol{R^2}$ and recorded for $\boldsymbol{\tau}$ ranging from 0 to 50 minutes. The dashed line represents $\boldsymbol{R^2 = 0.95}$. F) Effective back-exchange times for angiotensin, bradykinin and ANP estimated from all labelling times available. The black dot represents the $\boldsymbol{\tau}$ maximizing the $\boldsymbol{R^2}$; error bars represent the interval of $\boldsymbol{\tau}$ in which $\boldsymbol{R^2 > 0.95}$. The average estimated $\boldsymbol{\tau}$ (red dashed line) are compared with the expected back-exchange times (blue line) calculated as the sum of the trapping time (3 minutes) and the retention times of the peptides.

In several discussions with **Dr Jonathan Phillips** (University of Exeter) and **Dr Jochem Smit** (KU Leuven, Belgium), we agreed that systematic studies are needed to understand what percentage of forward- and back-exchange is occurring during the different stages of the experimental workflow (labelling, quench, digestion, gas-phase). He proposed experiments on unfolded peptides and model proteins where the back-exchange levels are tested by varying the deuterium percentage in the quench buffer, or by replacing the water-based LC solutions with deuterium-based equivalents.

6.7.6 Integration of EX1 information

Protection factors are a well-defined quantity only in the EX2 regime. Despite being a rare phenomenon, some protein systems (especially intrinsically disordered proteins) show a multimodal behaviour, characteristic of EX1 (or mixed EXX) kinetics (**section 2.5.3**). I had a minor encounter with these type of data in a collaboration with **Dr Alice Colyer** and **Dr Anton Calabrese**. They showed me the isotopic envelopes of a peptide which was following EX2 kinetics in one experimental condition and EXX kinetics in a second state. They asked me if I could extract some information out of these data, and I used the methods described in **section 2.5.3** to quantify the fractions of population belonging to the two different modes of the isotopic envelope.

The problem with this analysis is that it is an end in itself. We have discussed how the exchange rates that can be extracted from EX1 or EX2 kinetics have different meanings (**section 1.1.2**). Apart from stating that a specific peptide follows EX1 (or mixed EXX) kinetics, there is no straightforward way to connect this result with the level of protection of the rest of the protein. It might be evidence of the presence of a disordered region of the protein, but how does the protection of this region relate to that of adjacent regions? The information that we extract are different quantities: we cannot compare apples and oranges. The only option (that I personally see) is to integrate information from EX1 (or mixed EXX) kinetics with EX2 kinetics is by determining the closing, opening and intrinsic rates for each residue (**Eq. 1.6**). Although obtaining these rates is not feasible with HDX-MS data and is at least challenging with HDX-NMR

experiments, MD simulations may offer some valuable insights (Craig et al. 2011; Park et al. 2015; Persson and Halle 2015).

6.7.7 Integration with alternative labelling techniques

As mentioned in **section 1.3**, HDX is only one of several labelling techniques that can be coupled with mass spectrometry, with LiP (Fontana et al. 2004), FPOP (Johnson et al. 2019) and chemical cross-linking (Leitner et al. 2010) as main alternative approaches. Each of these methods offers insights on features such as protein dynamics, surface accessibility, flexibility and molecular dimension (Mitra 2021). Adopting a more integrated approach that combines multiple MS labelling methods could enable a deeper investigation of protein systems, yielding a more comprehensive understanding both locally and globally, and providing a more accurate structural characterization (Cornwell et al. 2018; Li et al. 2019; Oliva 2024; Zhang et al. 2019).

6.7.8 Integration with modelling

High resolution protein structures can be experimentally determined using X-ray christallography, cryogenic electron microscopy and nuclear magnetic resonance. However, these techniques provide only a snapshot of the protein, which instead is a dynamic entity. The power of HDX-MS experiments is to complement this structural information with dynamic information of the protein system (Hamuro 2024).

We have discussed (**section 1.1.3**) and shown (**Chapter 5**) that information extracted from HDX-MS data can be linked to MD simulations (or structural information of the protein) via the Best-Vendruscolo model (or similar approaches). The potential of this integrative method has been highlighted by implementing ensemble-reweighting strategies at both peptide-level resolution (Jia et al. 2023) and single-residue resolution (Kihn et al. 2024). We believe that the method presented in this Thesis for estimating protection factors from peptide-level data (along with the alternative approaches described in **section 2.5.4**) will help advance HDX-MS from a differential qualitative experiment to an absolute quantitative one. Ultimately, this shift will enable HDX-MS to

address the key biological questions it was originally designed to answer. While we do not focus on the utility of differential HDX-MS experiments for localizing binding sites or detecting conformational changes, we hope to have convinced the reader that HDX-MS has far greater potential than is commonly appreciated.

6.8 HDX and artificial intelligence: a controversial opinion

Throughout my PhD journey, I have often been asked: "Have you considered applying machine learning to your problem?". My response has always been (perhaps stubbornly), "Yes, but it is not something I am particularly passionate about."

Let me explain.

I fully recognize that the next frontier of HDX, like most of modern science, is entangled with machine learning and artificial intelligence. These technologies are undeniably powerful. Indeed, I have applied them in this very work – using a clustering algorithm in **Chapter 3**, AlphaFold in **Chapter 5**, and even turning to ChatGPT for proof-reading parts of this thesis. I acknowledge that AI is not just revolutionizing science; it is transforming every facet of life – sometimes for the better, as in healthcare, education and transportation, and at times with more concerning effects, as seen in privacy, warfare and disinformation. While the impact of these changes may be subject to debate, the magnitude of this transformation is undeniable.

Yet, my fascination with science was never rooted in the ability to predict outcomes with perfect accuracy. What drew me in, and what continues to captivate me, is the pursuit of understanding *why* something happens. That is a question artificial intelligence cannot fully answer (yet?). I can confidently predict that in the coming years, we will see publications leveraging machine learning algorithms to extract protection factors from HDX-MS data, most likely achieving much stronger correlations with NMR data than the methods described here. In fact, with enough NMR data on hundreds or thousands of proteins, we could train a neural network to map the

relationship between protein structure and its protection against exchange, giving us near-perfect accuracy.

But even then, the essential question will remain: *why*? Why does this particular conformation confer protection? What is the underlying mechanism? Theories like the Linderstrøm-Lang model or the Best-Vendruscolo phenomenological approximation, despite their limitations, attempt to provide some of these answers, offering a glimpse into the underlying reasons behind structural protection. From a scientific perspective, I would rather understand 50% of *why* something is happening than predict 100% of what will happen without any insight into the cause. For me, the joy of science lies in the journey toward understanding, not just in arriving at predictions.

References

Abdolvahabi, Alireza, Jennifer L. Gober, Richard A. Mowery, Yunhua Shi, and Bryan F. Shaw. 2014. 'Metal-Ion-Specific Screening of Charge Effects in Protein Amide H/D Exchange and the Hofmeister Series'. *Analytical Chemistry* 86(20):10303–10. doi: 10.1021/ac502714v.

Ahadi, Elias, and Lars Konermann. 2011. 'Ejection of Solvated Ions from Electrosprayed Methanol/Water Nanodroplets Studied by Molecular Dynamics Simulations'. *Journal of the American Chemical Society* 133(24):9354–63. doi: 10.1021/ja111492s.

Ahn, Joomi, Min-Jie Cao, Ying Qing Yu, and John R. Engen. 2013. 'Accessing the Reproducibility and Specificity of Pepsin and Other Aspartic Proteases'. *Biochimica Et Biophysica Acta* 1834(6):1222–29. doi: 10.1016/j.bbapap.2012.10.003.

Alderson, T. Reid, Jung Ho Lee, Cyril Charlier, Jinfa Ying, and Ad Bax. 2018. 'Propensity for Cis-Proline Formation in Unfolded Proteins'. *Chembiochem : A European Journal of Chemical Biology* 19(1):37–42. doi: 10.1002/cbic.201700548.

Al-Naqshabandi, Mohammed A., and David D. Weis. 2017. 'Quantifying Protection in Disordered Proteins Using Millisecond Hydrogen Exchange-Mass Spectrometry and Peptic Reference Peptides'. *Biochemistry* 56(31):4064–72. doi: 10.1021/acs.biochem.6b01312.

Andrec, Michael, R. Blake Hill, and James H. Prestegard. 1995. 'Amide Exchange Rates in Escherichia Coli Acyl Carrier Protein: Correlation with Protein Structure and Dynamics'. *Protein Science: A Publication of the Protein Society* 4(5):983–93. doi: 10.1002/pro.5560040518.

Andrews, Byron, Alan E. Murphy, Michele Stofella, Sarah Maslen, Leonardo Almeida-Souza, J. Mark Skehel, Nathan G. Skene, Frank Sobott, and René A. W. Frank. 2022. 'Multidimensional Dynamics of the Proteome in the Neurodegenerative and Aging Mammalian Brain'. *Molecular & Cellular Proteomics* 21(2). doi: 10.1016/j.mcpro.2021.100192.

Anfinsen, Christian B. 1973. 'Principles That Govern the Folding of Protein Chains'. *Science* 181(4096):223–30.

Aprahamian, Melanie L., Emily E. Chea, Lisa M. Jones, and Steffen Lindert. 2018. 'Rosetta Protein Structure Prediction from Hydroxyl Radical Protein Footprinting Mass Spectrometry Data'. *Analytical Chemistry* 90(12):7721–29. doi: 10.1021/acs.analchem.8b01624.

Au, Cheuk W., Iain Manfield, Michael E. Webb, Emanuele Paci, W. Bruce Turnbull, and James F. Ross. 2024. 'The Mutagenic Plasticity of the Cholera Toxin B-Subunit Surface Residues: Stability and Affinity'. *Toxins* 16(3):133. doi: 10.3390/toxins16030133.

Babić, Darko, Saša Kazazić, and David M. Smith. 2019. 'Resolution of Protein Hydrogen/Deuterium Exchange by Fitting Amide Exchange Probabilities to the Peptide Isotopic Envelopes'. *Rapid Communications in Mass Spectrometry* 33(15):1248–57. doi: 10.1002/rcm.8460.

Back, Jaap Willem, Luitzen de Jong, Anton O. Muijsers, and Chris G. de Koster. 2003. 'Chemical Cross-Linking and Mass Spectrometry for Protein Structural Modeling'. *Journal of Molecular Biology* 331(2):303–13. doi: 10.1016/S0022-2836(03)00721-6.

Bai, Yawen, John S. Milne, Leland Mayne, and S. Walter Englander. 1993. 'Primary Structure Effects on Peptide Group Hydrogen Exchange'. *Proteins* 17(1):75–86. doi: 10.1002/prot.340170110.

Baldwin, Robert L. 2011. 'Early Days of Protein Hydrogen Exchange: 1954–1972'. *Proteins: Structure, Function, and Bioinformatics* 79(7):2021–26. doi: 10.1002/prot.23039.

Baldwin, Robert L. 2017. 'Clash between Energy Landscape Theory and Foldon-Dependent Protein Folding'. *Proceedings of the National Academy of Sciences* 114(32):8442–43. doi: 10.1073/pnas.1709133114.

Barnes, C. Ashley, Yang Shen, Jinfa Ying, Yasuharu Takagi, Dennis A. Torchia, James R. Sellers, and Ad Bax. 2019. 'Remarkable Rigidity of the Single α-Helical Domain of Myosin-VI As Revealed by NMR Spectroscopy'. *Journal of the American Chemical Society* 141(22):9004–17. doi: 10.1021/jacs.9b03116.

Belorusova, Anna Y., Emma Evertsson, Daniel Hovdal, Jenny Sandmark, Emma Bratt, Ingela Maxvall, Ira G. Schulman, Peter Åkerblad, and Eva-Lotte Lindstedt. 2019. 'Structural Analysis Identifies an Escape Route from the Adverse Lipogenic Effects of Liver X Receptor Ligands'. *Communications Biology* 2:431. doi: 10.1038/s42003-019-0675-0.

Benjamini, Yoav, Dan Drai, Greg Elmer, Neri Kafkafi, and Ilan Golani. 2001. 'Controlling the False Discovery Rate in Behavior Genetics Research'. *Behavioural Brain Research* 125(1):279–84. doi: 10.1016/S0166-4328(01)00297-2.

Berger, Arieh, and Kaj U. Linderstrøm-Lang. 1957. 'Deuterium Exchange of Poly-Dl-Alanine in Aqueous Solution'. *Archives of Biochemistry and Biophysics* 69:106–18. doi: 10.1016/0003-9861(57)90478-2.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. 'The Protein Data Bank'. *Nucleic Acids Research* 28(1):235–42. doi: 10.1093/nar/28.1.235.

Best, Robert B., Susan B. Fowler, José L. Toca Herrera, Annette Steward, Emanuele Paci, and Jane Clarke. 2003. 'Mechanical Unfolding of a Titin Ig Domain: Structure of Transition State Revealed by Combining Atomic Force Microscopy, Protein Engineering and Molecular Dynamics Simulations'. *Journal of Molecular Biology* 330(4):867–77. doi: 10.1016/S0022-2836(03)00618-1.

Best, Robert B., and Michele Vendruscolo. 2006. 'Structural Interpretation of Hydrogen Exchange Protection Factors in Proteins: Characterization of the Native State Fluctuations of CI2'. *Structure (London, England: 1993)* 14(1):97–106. doi: 10.1016/j.str.2005.09.012.

Bishop, Christopher. 2006. *Pattern Recognition and Machine Learning*. New York: Springer-Verlag.

Boesl, Ulrich. 2017. 'Time-of-Flight Mass Spectrometry: Introduction to the Basics'. *Mass Spectrometry Reviews* 36(1):86–109. doi: 10.1002/mas.21520.

Bougault, Catherine, Lianmei Feng, John Glushka, Eriks Kupče, and J. H. Prestegard. 2004. 'Quantitation of Rapid Proton-Deuteron Amide Exchange Using Hadamard Spectroscopy'. *Journal of Biomolecular NMR* 28(4):385–90. doi: 10.1023/B:JNMR.0000015406.66725.30.

Bouyssié, David, Jean Lesne, Marie Locard-Paulet, Renaud Albigot, Odile Burlet-Schiltz, and Julien Marcoux. 2019. 'HDX-Viewer: Interactive 3D Visualization of Hydrogen–Deuterium Exchange Data'. *Bioinformatics* 35(24):5331–33. doi: 10.1093/bioinformatics/btz550.

Boyd, Robert, and Arpád Somogyi. 2010. 'The Mobile Proton Hypothesis in Fragmentation of Protonated Peptides: A Perspective'. *Journal of the American Society for Mass Spectrometry* 21(8):1275–78. doi: 10.1016/j.jasms.2010.04.017.

Brand, Torsten, Eurico J. Cabrita, Gareth A. Morris, Robert Günther, Hans-Jörg Hofmann, and Stefan Berger. 2007. 'Residue-Specific NH Exchange Rates Studied by NMR Diffusion Experiments'. *Journal of Magnetic Resonance* 187(1):97–104. doi: 10.1016/j.jmr.2007.03.021.

Burkitt, William, and Gavin O'Connor. 2008. 'Assessment of the Repeatability and Reproducibility of Hydrogen/Deuterium Exchange Mass Spectrometry Measurements'. *Rapid Communications in Mass Spectrometry* 22(23):3893–3901. doi: 10.1002/rcm.3794.

Burridge, Charles, Christopher A. Waudby, Tomasz Włodarski, Anaïs M. E. Cassaignau, Lisa D. Cabrita, and John Christodoulou. 2021. 'Nascent Chain Dynamics and Ribosome Interactions within Folded Ribosome–Nascent Chain Complexes Observed by NMR Spectroscopy'. *Chemical Science* 12(39):13120–26. doi: 10.1039/D1SC04313G.

Bustamante, Carlos, and Shannon Yan. 2022. 'The Development of Single Molecule Force Spectroscopy: From Polymer Biophysics to Molecular Machines'. *Quarterly Reviews of Biophysics* 55:e9. doi: 10.1017/S0033583522000087.

Byrd, Emily J., Martin Wilkinson, Sheena E. Radford, and Frank Sobott. 2023. 'Taking Charge: Metal Ions Accelerate Amyloid Aggregation in Sequence Variants of α-Synuclein'. *Journal of the American Society for Mass Spectrometry* 34(3):493–504. doi: 10.1021/jasms.2c00379.

Calvaresi, Valeria, Lucia Dello Iacono, Sara Borghi, Enrico Luzzi, Alessia Biolchi, Barbara Benucci, Ilaria Ferlenghi, Ilaria Peschiera, Fabiola Giusti, Lucia E. Fontana, Zhong-Yuan Kan, Zaira Spinello, Marcello Merola, Isabel Delany, Kasper D. Rand, and Nathalie Norais. 2024. 'Structural Dynamics and Immunogenicity of the Recombinant and Outer Membrane Vesicle-Embedded Meningococcal Antigen NadA'. 2024.01.30.577382.

Cappelletti, Valentina, Thomas Hauser, Ilaria Piazza, Monika Pepelnjak, Liliana Malinovska, Tobias Fuhrer, Yaozong Li, Christian Dörig, Paul Boersema, Ludovic Gillet, Jan Grossbach, Aurelien Dugourd, Julio Saez-Rodriguez, Andreas Beyer, Nicola Zamboni, Amedeo Caflisch, Natalie de Souza, and Paola Picotti. 2021. 'Dynamic 3D Proteomes Reveal Protein Functional Alterations at High Resolution in Situ'. *Cell* 184(2):545-559.e22. doi: 10.1016/j.cell.2020.12.021.

Case, D. A., H. M. Aktulga, K. Belfon, I. Y. Ben-Shalom, J. T. Berryman, S. R. Brozell, D. S. Cerutti, T. E. Cheatham, G. A. Cisneros, V. W. D. Cruzeiro, T. A. Darder, R. E. Duke, G. Giambasu, M. K. Gilson, H. Gohlke, A. W. Goetz, R. Harris, S. Izadi, S. A. Izmailov, K. Kasavajhala, M. C. Kayman, E. King, A. Kovalenko, T. Kurtzman, T. S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, M. K. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K. A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, A. Shajan, J. Shen, C. L. Simmerling, N. R. Skrynnikov, J. Smith, J. Swails, R. C. Walker, J. Wang, J. Wang, H. Wei, R. M. Wolf, X. Wu, Y. Xiong, Y. Xue, D. M. York, S. Zhao, and P. A. Kollman. 2022. 'Amber 2022'.

Chakrabartty, Avijit, Tanja Kortemme, and Robert L. Baldwin. 1994. 'Helix Propensities of the Amino Acids Measured in Alanine-Based Peptides without Helix-Stabilizing Side-Chain Interactions.' *Protein Science: A Publication of the Protein Society* 3(5):843–52. doi: 10.1002/pro.5560030514.

Chu, I.-Te, and Gary J. Pielak. 2023. 'Using NMR-Detected Hydrogen-Deuterium Exchange to Quantify Protein Stability in Cosolutes, under Crowded Conditions *in Vitro* and in Cells'. *Magnetic Resonance Letters* 3(4):319–26. doi: 10.1016/j.mrl.2023.04.003.

Claesen, Jürgen, and Tomasz Burzykowski. 2017. 'Computational Methods and Challenges in Hydrogen/Deuterium Exchange Mass Spectrometry'. *Mass Spectrometry Reviews* 36(5):649–67. doi: https://doi.org/10.1002/mas.21519.

Claesen, Jürgen, Srinath Krishnamurthy, Andy M. Lau, and Anastassios Economou. 2021. 'Moderated Test Statistics to Detect Differential Deuteration in Hydrogen/Deuterium Exchange Mass Spectrometry Experiments'. *Analytical Chemistry* 93(49):16341–49. doi: 10.1021/acs.analchem.1c02346.

Claesen, Jürgen, and Argyris Politis. 2019. 'POPPeT: A New Method to Predict the Protection Factor of Backbone Amide Hydrogens'. *Journal of the American Society for Mass Spectrometry* 30(1):67–76. doi: 10.1007/s13361-018-2068-x.

Connelly, Gregory P., Yawen Bai, Mei-Fen Jeng, and S. Walter Englander. 1993. 'Isotope effects in peptide group hydrogen exchange'. *Proteins: Structure, Function, and Bioinformatics* 17(1):87–92. doi: 10.1002/prot.340170111.

Cornwell, Owen, Sheena E. Radford, Alison E. Ashcroft, and James R. Ault. 2018. 'Comparing Hydrogen Deuterium Exchange and Fast Photochemical Oxidation of Proteins: A Structural Characterisation of Wild-Type and ΔN6 B2-Microglobulin'. *Journal of The American Society for Mass Spectrometry* 29(12):2413–26. doi: 10.1007/s13361-018-2067-y.

Cox, Jürgen, Marco Y. Hein, Christian A. Luber, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. 2014. 'Accurate Proteome-Wide Label-Free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ'. *Molecular & Cellular Proteomics : MCP* 13(9):2513–26. doi: 10.1074/mcp.M113.031591.

Craig, Patricio O., Joachim Lätzer, Patrick Weinkam, Ryan M. B. Hoffman, Diego U. Ferreiro, Elizabeth A. Komives, and Peter G. Wolynes. 2011. 'Prediction of Native-State Hydrogen Exchange from Perfectly Funneled Energy Landscapes'. *Journal of the American Chemical Society* 133(43):17463–72. doi: 10.1021/ja207506z.

Cravello, Laetitia, David Lascoux, and Eric Forest. 2003. 'Use of Different Proteases Working in Acidic Conditions to Improve Sequence Coverage and Resolution in Hydrogen/Deuterium Exchange of Large Proteins'. *Rapid Communications in Mass Spectrometry* 17(21):2387–93. doi: 10.1002/rcm.1207.

Crook, Oliver M., Chun-wa Chung, and Charlotte M. Deane. 2022. 'Empirical Bayes Functional Models for Hydrogen Deuterium Exchange Mass Spectrometry'. *Communications Biology* 5(1):1–10. doi: 10.1038/s42003-022-03517-3.

Crook, Oliver M., Nathan Gittens, Chun-wa Chung, and Charlotte M. Deane. 2024. ‘Inferring Residue Level Hydrogen Deuterium Exchange with ReX’.

Damont, Annelaure, Anaïs Legrand, Chenqin Cao, François Fenaille, and Jean-Claude Tabet. 2021. ‘Hydrogen/Deuterium Exchange Mass Spectrometry in the World of Small Molecules’. *Mass Spectrometry Reviews* n/a(n/a):e21765. doi: 10.1002/mas.21765.

Deiana, Antonio, Sergio Forcelloni, Alessandro Porrello, and Andrea Giansanti. 2019. ‘Intrinsically Disordered Proteins and Structured Proteins with Intrinsically Disordered Regions Have Different Functional Roles in the Cell’. *PLOS One* 14(8):e0217889. doi: 10.1371/journal.pone.0217889.

Del Mar, Charyl, Eric A. Greenbaum, Leland Mayne, S. Walter Englander, and Virgil L. Woods. 2005. ‘Structure and Properties of α-Synuclein and Other Amyloids Determined at the Amino Acid Level’. *Proceedings of the National Academy of Sciences* 102(43):15477–82. doi: 10.1073/pnas.0507405102.

Delaglio, Frank, Stephan Grzesiek, Geerten W. Vuister, Guang Zhu, John Pfeifer, and Ad Bax. 1995. ‘NMRPipe: A Multidimensional Spectral Processing System Based on UNIX Pipes’. *Journal of Biomolecular NMR* 6(3):277–93. doi: 10.1007/BF00197809.

Dempsey, Christopher E. 2001. ‘Hydrogen Exchange in Peptides and Proteins Using NMR Spectroscopy’. *Progress in Nuclear Magnetic Resonance Spectroscopy* 39(2):135–70. doi: 10.1016/S0079-6565(01)00032-2.

Deng, Bin, Cristina Lento, and Derek J. Wilson. 2016. ‘Hydrogen Deuterium Exchange Mass Spectrometry in Biopharmaceutical Discovery and Development – A Review’. *Analytica Chimica Acta* 940:8–20. doi: 10.1016/j.aca.2016.08.006.

Devaurs, Didier, Dinler A. Antunes, and Antoni J. Borysik. 2022. ‘Computational Modeling of Molecular Structures Guided by Hydrogen-Exchange Data’. *Journal of the American Society for Mass Spectrometry* 33(2):215–37. doi: 10.1021/jasms.1c00328.

Dill, Ken A., and Hue Sun Chan. 1997. ‘From Levinthal to Pathways to Funnels’. *Nature Structural Biology* 4(1):10–19. doi: 10.1038/nsb0197-10.

Dill, Ken A., and Justin L. MacCallum. 2012. ‘The Protein-Folding Problem, 50 Years On’. *Science* 338(6110):1042–46. doi: 10.1126/science.1219021.

Dudoit, Sandrine, and Mark J. Laan. 2008. *Multiple Testing Procedures with Applications to Genomics*.

Ellis, Reginald J. 1993. ‘The General Concept of Molecular Chaperones’. Pp. 1–5 in *Molecular Chaperones*, edited by R. J. Ellis, R. A. Laskey, and G. H. Lorimer. Dordrecht: Springer Netherlands.

Engen, John R., Thomas Botzanowski, Daniele Peterle, Florian Georgescauld, and Thomas E. Wales. 2021. 'Developments in Hydrogen/Deuterium Exchange Mass Spectrometry'. *Analytical Chemistry* 93(1):567–82. doi: 10.1021/acs.analchem.0c04281.

Engen, John R., and Thomas E. Wales. 2015. 'Analytical Aspects of Hydrogen Exchange Mass Spectrometry'. *Annual Review of Analytical Chemistry (Palo Alto, Calif.)* 8:127–48. doi: 10.1146/annurev-anchem-062011-143113.

Englander, Joan J., D. B. Calhoun, and S. Walter Englander. 1979. 'Measurement and Calibration of Peptide Group Hydrogen-Deuterium Exchange by Ultraviolet Spectrophotometry'. *Analytical Biochemistry* 92(2):517–24. doi: 10.1016/0003-2697(79)90693-6.

Englander, S. Walter. 1963. 'A Hydrogen Exchange Method Using Tritium and Sephadex: Its Application to Ribonuclease'. *Biochemistry* 2:798–807.

Englander, S. Walter. 2000. 'Protein Folding Intermediates and Pathways Studied by Hydrogen Exchange'. *Annual Review of Biophysics* 29(Volume 29, 2000):213–38. doi: 10.1146/annurev.biophys.29.1.213.

Englander, S. Walter. 2023. 'HX and Me: Understanding Allostery, Folding, and Protein Machines'. *Annual Review of Biophysics* 52(Volume 52, 2023):1–18. doi: 10.1146/annurev-biophys-062122-093517.

Englander, S. Walter, and Neville R. Kallenbach. 1983. 'Hydrogen Exchange and Structural Dynamics of Proteins and Nucleic Acids'. *Quarterly Reviews of Biophysics* 16(4):521–655. doi: 10.1017/S0033583500005217.

Englander, S. Walter, and Leland Mayne. 1992. 'Protein Folding Studied Using Hydrogen-Exchange Labeling and Two-Dimensional NMR'. *Annual Review of Biophysics and Biomolecular Structure* 21:243–65. doi: 10.1146/annurev.bb.21.060192.001331.

Englander, S. Walter, and Leland Mayne. 2014. 'The Nature of Protein Folding Pathways'. *Proceedings of the National Academy of Sciences of the United States of America* 111(45):15873–80. doi: 10.1073/pnas.1411798111.

Englander, S. Walter, and Leland Mayne. 2017. 'The Case for Defined Protein Folding Pathways'. *Proceedings of the National Academy of Sciences* 114(31):8253–58. doi: 10.1073/pnas.1706196114.

Englander, S. Walter, Leland Mayne, Zhong-Yuan Kan, and Wenbing Hu. 2016. 'Protein Folding—How and Why: By Hydrogen Exchange, Fragment Separation, and Mass Spectrometry'. *Annual Review of Biophysics* 45:135–52. doi: 10.1146/annurev-biophys-062215-011121.

Englander, S. Walter, Tobin R. Sosnick, Joan J. Englander, and Leland Mayne. 1996. 'Mechanisms and Uses of Hydrogen Exchange'. *Current Opinion in Structural Biology* 6(1):18–23.

Eyring, Henry. 1935. 'The Activated Complex in Chemical Reactions'. *The Journal of Chemical Physics* 3(2):107–15. doi: 10.1063/1.1749604.

Fang, Jing, Kasper D. Rand, Penny J. Beuning, and John R. Engen. 2011. 'False EX1 Signatures Caused by Sample Carryover during HX MS Analyses'. *International Journal of Mass Spectrometry* 302(1):19–25. doi: 10.1016/j.ijms.2010.06.039.

Fearnhead, Paul. 2006. 'Exact and Efficient Bayesian Inference for Multiple Changepoint Problems'. *Statistics and Computing* 16(2):203–13. doi: 10.1007/s11222-006-8450-8.

Fenn, John B. 2003. 'Electrospray Wings for Molecular Elephants (Nobel Lecture)'. *Angewandte Chemie International Edition* 42(33):3871–94. doi: 10.1002/anie.200300605.

Ferraro, Debra M., Noel D. Lazo, and Andrew D. Robertson. 2004. 'EX1 Hydrogen Exchange and Protein Folding'. *Biochemistry* 43(3):587–94. doi: 10.1021/bi035943y.

Filandr, Frantisek, Vladimir Sarpe, Shaunak Raval, D. Alex Crowder, Morgan F. Khan, Pauline Douglas, Stephen Coales, Rosa Viner, Aleem Syed, John A. Tainer, Susan P. Lees-Miller, and David C. Schriemer. 2024. 'Automating Data Analysis for Hydrogen/Deuterium Exchange Mass Spectrometry Using Data-Independent Acquisition Methodology'. *Nature Communications* 15(1):2200. doi: 10.1038/s41467-024-46610-3.

Fitzkee, Nicholas C., Dennis A. Torchia, and Ad Bax. 2011. 'Measuring Rapid Hydrogen Exchange in the Homodimeric 36 kDa HIV-1 Integrase Catalytic Core Domain'. *Protein Science* 20(3):500–512. doi: https://doi.org/10.1002/pro.582.

Fontana, Angelo, Patrizia Polverino de Laureto, Barbara Spolaore, Erica Frare, Paola Picotti, and Marcello Zambonin. 2004. 'Probing Protein Structure by Limited Proteolysis.' *Acta Biochimica Polonica* 51(2):299–321. doi: 10.18388/abp.2004_3573.

Frushour, Bruce G., and Jack L. Koenig. 1975. 'Raman Spectra of D and L Amino Acid Copolymers. Poly-DL-Alanine, Poly-DL-Leucine, and Poly-DL-Lysine'. *Biopolymers* 14(2):363–77. doi: 10.1002/bip.1975.360140210.

Gessner, Chris, Wieland Steinchen, Sabrina Bédard, John J. Skinner, Virgil L. Woods, Thomas J. Walsh, Gert Bange, and Dionysios P. Pantazatos. 2017. 'Computational Method Allowing Hydrogen-Deuterium Exchange Mass Spectrometry at Single Amide Resolution'. *Scientific Reports* 7(1):3789. doi: 10.1038/s41598-017-03922-3.

Goswami, Devrishi, Srikripa Devarakonda, Michael J. Chalmers, Bruce D. Pascal, Bruce M. Spiegelman, and Patrick R. Griffin. 2013. 'Time Window Expansion for HDX Analysis of an Intrinsically Disordered Protein'. *Journal of the American Society for Mass Spectrometry* 24(10):1584–92. doi: 10.1007/s13361-013-0669-y.

Green, Peter J. 1995. 'Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination'. *Biometrika* 82(4):711–32. doi: 10.1093/biomet/82.4.711.

Griffiths, Jennifer. 2008. 'A Brief History of Mass Spectrometry'. *Analytical Chemistry* 80(15):5678–83. doi: 10.1021/ac8013065.

Guttman, Miklos, Thomas E. Wales, Dale Whittington, John R. Engen, Jeffery M. Brown, and Kelly K. Lee. 2016. 'Tuning a High Transmission Ion Guide to Prevent Gas-Phase Proton Exchange During H/D Exchange MS Analysis'. *Journal of The American Society for Mass Spectrometry* 27(4):662–68. doi: 10.1007/s13361-015-1330-8.

Guttman, Miklos, David D. Weis, John R. Engen, and Kelly K. Lee. 2013. 'Analysis of Overlapped and Noisy Hydrogen/Deuterium Exchange Mass Spectra'. *Journal of the American Society for Mass Spectrometry* 24(12). doi: 10.1007/s13361-013-0727-5.

Hageman, Tyler S., and David D. Weis. 2019. 'Reliable Identification of Significant Differences in Differential Hydrogen Exchange-Mass Spectrometry Measurements Using a Hybrid Significance Testing Approach'. *Analytical Chemistry* 91(13):8008–16. doi: 10.1021/acs.analchem.9b01325.

Hamuro, Yoshitomo. 2021a. 'Quantitative Hydrogen/Deuterium Exchange Mass Spectrometry'. *Journal of the American Society for Mass Spectrometry*. doi: 10.1021/jasms.1c00216.

Hamuro, Yoshitomo. 2021b. 'Tutorial: Chemistry of Hydrogen/Deuterium Exchange Mass Spectrometry'. *Journal of the American Society for Mass Spectrometry* 32(1):133–51. doi: 10.1021/jasms.0c00260.

Hamuro, Yoshitomo. 2024. 'Interpretation of Hydrogen/Deuterium Exchange Mass Spectrometry'. *Journal of the American Society for Mass Spectrometry*. doi: 10.1021/jasms.4c00044.

Hamuro, Yoshitomo, and Stephen J. Coales. 2018. 'Optimization of Feasibility Stage for Hydrogen/Deuterium Exchange Mass Spectrometry'. *Journal of the American Society for Mass Spectrometry* 29(3):623–29. doi: 10.1007/s13361-017-1860-3.

Heck, Albert J. R., and Robert H. H. van den Heuvel. 2004. 'Investigation of Intact Protein Complexes by Mass Spectrometry'. *Mass Spectrometry Reviews* 23(5):368–89. doi: 10.1002/mas.10081.

Hegde, R. S., J. A. Mastrianni, M. R. Scott, K. A. DeFea, P. Tremblay, M. Torchia, S. J. DeArmond, S. B. Prusiner, and V. R. Lingappa. 1998. 'A Transmembrane Form of the Prion Protein in Neurodegenerative Disease'. *Science (New York, N.Y.)* 279(5352):827–34. doi: 10.1126/science.279.5352.827.

Hernández, Griselda, Janet S. Anderson, and David M. LeMaster. 2009. 'Polarization and Polarizability Assessed by Protein Amide Acidity'. *Biochemistry* 48(27):6482–94. doi: 10.1021/bi900526z.

Hinck, Andrew P. 2022. 'On "Deuterium Exchange of Poly-Dl-Alanine in Aqueous Solution" by Arieh Berger and K. Linderstrøm-Lang'. *Archives of Biochemistry and Biophysics* 726:109185. doi: 10.1016/j.abb.2022.109185.

Hodge, Edgar A., Mark A. Benhaim, and Kelly K. Lee. 2020. 'Bridging Protein Structure, Dynamics, and Function Using Hydrogen/Deuterium-Exchange Mass Spectrometry'. *Protein Science* 29(4):843–55. doi: 10.1002/pro.3790.

Hoffmann, Edmond de, and Vincent Stroobant. 2007. *Mass Spectrometry: Principles and Applications*. John Wiley & Sons.

Houde, Damian, Steven A. Berkowitz, and John R. Engen. 2011. 'The Utility of Hydrogen/Deuterium Exchange Mass Spectrometry in Biopharmaceutical Comparability Studies'. *Journal of Pharmaceutical Sciences* 100(6):2071–86. doi: 10.1002/jps.22432.

Hourdel, Véronique, Stevenn Volant, Darragh P. O'Brien, Alexandre Chenal, Julia Chamot-Rooke, Marie-Agnès Dillies, and Sébastien Brier. 2016. 'MEMHDX: An Interactive Tool to Expedite the Statistical Validation and Visualization of Large HDX-MS Datasets'. *Bioinformatics* 32(22):3413–19. doi: 10.1093/bioinformatics/btw420.

Hughes, Megan L., and Lorna Dougan. 2016. 'The Physics of Pulling Polyproteins: A Review of Single Molecule Force Spectroscopy Using the AFM to Study Protein Unfolding'. *Reports on Progress in Physics* 79(7):076601. doi: 10.1088/0034-4885/79/7/076601.

Huth, Jeffrey R., Carole A. Bewley, Belinda M. Jackson, Alan G. Hinnebusch, G. Marius Clore, and Angela M. Gronenborn. 1997. 'Design of an Expression System for Detecting Folded Protein Domains and Mapping Macromolecular Interactions by NMR'. *Protein Science: A Publication of the Protein Society* 6(11):2359–64. doi: 10.1002/pro.5560061109.

Hvidt, Aase, and Kaj U. Linderstrøm-Lang. 1954. 'Exchange of Hydrogen Atoms in Insulin with Deuterium Atoms in Aqueous Solutions'. *Biochimica et Biophysica Acta* 14:574–75. doi: 10.1016/0006-3002(54)90241-3.

Hvidt, Aase, and Sigurd O. Nielsen. 1966. 'Hydrogen Exchange in Proteins'. *Advances in Protein Chemistry* 21:287–386. doi: 10.1016/s0065-3233(08)60129-1.

Iacob, Roxana E., George M. Bou-Assaf, Lee Makowski, John R. Engen, Steven A. Berkowitz, and Damian Houde. 2013. 'Investigating Monoclonal Antibody Aggregation Using a Combination of H/DX-MS and Other Biophysical Measurements'. *Journal of Pharmaceutical Sciences* 102(12):4315–29. doi: 10.1002/jps.23754.

Iavarone, Anthony T., and Evan R. Williams. 2003. 'Mechanism of Charging and Supercharging Molecules in Electrospray Ionization'. *Journal of the American Chemical Society* 125(8):2319–27. doi: 10.1021/ja021202t.

Ingwall, R. T., H. A. Scheraga, N. Lotan, A. Berger, and E. Katchalski. 1968. 'Conformational Studies of Poly-L-Alanine in Water'. *Biopolymers* 6(3):331–68. doi: 10.1002/bip.1968.360060308.

Iribarne, J. V., and B. A. Thomson. 1976. 'On the Evaporation of Small Ions from Charged Droplets'. *The Journal of Chemical Physics* 64(6):2287–94. doi: 10.1063/1.432536.

Jain, Shweta, and Jayant B. Udgaonkar. 2008. 'Evidence for Stepwise Formation of Amyloid Fibrils by the Mouse Prion Protein'. *Journal of Molecular Biology* 382(5):1228–41. doi: 10.1016/j.jmb.2008.07.052.

James, Ellie I., Taylor A. Murphree, Clint Vorauer, John R. Engen, and Miklos Guttman. 2021. 'Advances in Hydrogen/Deuterium Exchange Mass Spectrometry and the Pursuit of Challenging Biological Systems'. *Chemical Reviews*. doi: 10.1021/acs.chemrev.1c00279.

James, Ellie I., Taylor A. Murphree, Clint Vorauer, John R. Engen, and Miklos Guttman. 2022. 'Advances in Hydrogen/Deuterium Exchange Mass Spectrometry and the Pursuit of Challenging Biological Systems'. *Chemical Reviews* 122(8):7562–7623. doi: 10.1021/acs.chemrev.1c00279.

Jia, Ruyu, Richard T. Bradshaw, Valeria Calvaresi, and Argyris Politis. 2023. 'Integrating Hydrogen Deuterium Exchange–Mass Spectrometry with Molecular Simulations Enables Quantification of the Conformational Populations of the Sugar Transporter XylE'. *Journal of the American Chemical Society* 145(14):7768–79. doi: 10.1021/jacs.2c06148.

Johnson, Danté T., Luciano H. Di Stefano, and Lisa M. Jones. 2019. 'Fast Photochemical Oxidation of Proteins (FPOP): A Powerful Mass Spectrometry–Based Structural Proteomics Tool'. *Journal of Biological Chemistry* 294(32):11969–79. doi: 10.1074/jbc.REV119.006218.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 2021. 'Highly Accurate Protein Structure Prediction with AlphaFold'. *Nature* 596(7873):583–89. doi: 10.1038/s41586-021-03819-2.

Kabsch, Wolfgang, and Christian Sander. 1983. 'Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features'. *Biopolymers* 22(12):2577–2637. doi: 10.1002/bip.360221211.

Kafader, Jared O., Rafael D. Melani, Luis F. Schachner, Ashley N. Ives, Steven M. Patrie, Neil L. Kelleher, and Philip D. Compton. 2020. 'Native vs Denatured: An in Depth Investigation of Charge State and Isotope Distributions'. *Journal of the American Society for Mass Spectrometry* 31(3):574–81. doi: 10.1021/jasms.9b00040.

Kan, Zhong-Yuan, Leland Mayne, Palaniappan Sevugan Chetty, and S. Walter Englander. 2011. 'ExMS: Data Analysis for HX-MS Experiments'. *Journal of the American Society for Mass Spectrometry* 22(11). doi: 10.1021/jasms.8b03933.

Kan, Zhong-Yuan, Benjamin T. Walters, Leland Mayne, and S. Walter Englander. 2013. 'Protein Hydrogen Exchange at Residue Resolution by Proteolytic Fragmentation Mass Spectrometry Analysis'. *Proceedings of the National Academy of Sciences of the United States of America* 110(41):16438–43. doi: 10.1073/pnas.1315532110.

Kan, Zhong-yuan, Xiang Ye, John J. Skinner, Leland Mayne, and S. Walter Englander. 2019. 'ExMS2: An Integrated Solution for Hydrogen–Deuterium Exchange Mass Spectrometry Data Analysis'. *Analytical Chemistry* 91(11):7474–81. doi: 10.1021/acs.analchem.9b01682.

Kanu, Abu B., Prabha Dwivedi, Maggie Tam, Laura Matz, and Herbert H. Hill. 2008. 'Ion Mobility–Mass Spectrometry'. *Journal of Mass Spectrometry* 43(1):1–22. doi: 10.1002/jms.1383.

Karplus, Martin. 1997. 'The Levinthal Paradox: Yesterday and Today'. *Folding and Design* 2:S69–75. doi: 10.1016/S1359-0278(97)00067-9.

Kateb, Fatiha, Philippe Pelupessy, and Geoffrey Bodenhausen. 2007. 'Measuring Fast Hydrogen Exchange Rates by NMR Spectroscopy'. *Journal of Magnetic Resonance* 184(1):108–13. doi: 10.1016/j.jmr.2006.09.022.

Kaur, Upneet, Danté T. Johnson, Emily E. Chea, Daniel J. Deredge, Jessica A. Espino, and Lisa M. Jones. 2019. 'Evolution of Structural Biology through the Lens of Mass

Spectrometry'. *Analytical Chemistry* 91(1):142–55. doi: 10.1021/acs.analchem.8b05014.

Kebarle, Paul, and Udo H. Verkerk. 2009. 'Electrospray: From Ions in Solution to Ions in the Gas Phase, What We Know Now'. *Mass Spectrometry Reviews* 28(6):898–917. doi: 10.1002/mas.20247.

Kendrew, J. C., R. E. Dickerson, B. E. Strandberg, R. G. Hart, D. R. Davies, D. C. Phillips, and V. C. Shore. 1960. 'Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 A. Resolution'. *Nature* 185(4711):422–27. doi: 10.1038/185422a0.

Kenji, Takahashi, B. P. Athauda Senarath, Matsumoto Koji, Rajapakshe Sanath, Kuribayashi Masayuki, Kojima Masaki, Kubomura-Yoshida Nobuko, Iwamatsu Akihiro, Shibata Chiaki, and Inoue Hideshi. 2005. 'Nepenthesin, a Unique Member of a Novel Subfamily of Aspartic Proteinases: Enzymatic and Structural Characteristics'. *Current Protein & Peptide Science* 6(6):513–25.

Keppel, Theodore R., and David D. Weis. 2013. 'Analysis of Disordered Proteins Using a Simple Apparatus for Millisecond Quench-Flow H/D Exchange'. *Analytical Chemistry* 85(10):5161–68. doi: 10.1021/ac4004979.

Keppel, Theodore R., and David D. Weis. 2015. 'Mapping Residual Structure in Intrinsically Disordered Proteins at Residue Resolution Using Millisecond Hydrogen/Deuterium Exchange and Residue Averaging'. *Journal of the American Society for Mass Spectrometry* 26(4):547–54. doi: 10.1007/s13361-014-1033-6.

Kienlein, Maximilian, Martin Zacharias, and Maria M. Reif. 2024. 'Comprehensive Analysis of Coupled Proline Cis–Trans States in Bradykinin Using ωBP-REMD Simulations'. *Journal of Chemical Theory and Computation*. doi: 10.1021/acs.jctc.3c01356.

Kihn, Kyle C., Olivia Purdy, Vincent Lowe, Lenka Slachtova, Ally K. Smith, Paul Shapiro, and Daniel J. Deredge. 2024. 'Integration of Hydrogen–Deuterium Exchange Mass Spectrometry with Molecular Dynamics Simulations and Ensemble Reweighting Enables High Resolution Protein–Ligand Modeling'. *Journal of the American Society for Mass Spectrometry*. doi: 10.1021/jasms.4c00202.

Kim, Peter S., and Robert L. Baldwin. 1982. 'Influence of Charge on the Rate of Amide Proton Exchange'. *Biochemistry* 21(1):1–5. doi: 10.1021/bi00530a001.

Kirch, Wilhelm, ed. 2008. 'Pearson's Correlation Coefficient'. Pp. 1090–91 in *Encyclopedia of Public Health*. Dordrecht: Springer Netherlands.

Kish, Monika, Victoria Smith, Natasha Lethbridge, Lindsay Cole, Nicholas. J. Bond, and Jonathan J. Phillips. 2023. 'Online Fully Automated System for Hydrogen/Deuterium-Exchange Mass Spectrometry with Millisecond Time Resolution'. *Analytical Chemistry*. doi: 10.1021/acs.analchem.2c05310.

Kish, Monika, Victoria Smith, Sivaraman Subramanian, Frank Vollmer, Natasha Lethbridge, Lindsay Cole, Nicholas J. Bond, and Jonathan J. Phillips. 2019. 'Allosteric Regulation of Glycogen Phosphorylase Solution Phase Structural Dynamics at High Spatial Resolution'. *bioRxiv* 654665. doi: 10.1101/654665.

Kleckner, Ian R., and Mark P. Foster. 2011. 'An Introduction to NMR-Based Approaches for Measuring Protein Dynamics'. *Biochimica et Biophysica Acta* 1814(8):942–68. doi: 10.1016/j.bbapap.2010.10.012.

Konermann, Lars, Elias Ahadi, Antony D. Rodriguez, and Siavash Vahidi. 2013. 'Unraveling the Mechanism of Electrospray Ionization'. *Analytical Chemistry* 85(1):2–9. doi: 10.1021/ac302789c.

Konermann, Lars, Antony D. Rodriguez, and Jiangjiang Liu. 2012. 'On the Formation of Highly Charged Gaseous Ions from Unfolded Proteins by Electrospray Ionization'. *Analytical Chemistry* 84(15):6798–6804. doi: 10.1021/ac301298g.

Konijnenberg, A., A. Butterer, and F. Sobott. 2013. 'Native Ion Mobility-Mass Spectrometry and Related Methods in Structural Biology'. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1834(6):1239–56. doi: 10.1016/j.bbapap.2012.11.013.

Kuczera, Krzysztof, Robert Szoszkiewicz, Jinyan He, and Gouri S. Jas. 2021. 'Length Dependent Folding Kinetics of Alanine-Based Helical Peptides from Optimal Dimensionality Reduction'. *Life* 11(5):385. doi: 10.3390/life11050385.

Kuwata, Kazuo, Tomoharu Matumoto, Hong Cheng, Kuniaki Nagayama, Thomas L. James, and Heinrich Roder. 2003. 'NMR-Detected Hydrogen Exchange and Molecular Dynamics Simulations Provide Structural Insight into Fibril Formation of Prion Protein Fragment 106–126'. *Proceedings of the National Academy of Sciences* 100(25):14790–95. doi: 10.1073/pnas.2433563100.

Largy, Eric, and Valérie Gabelica. 2020. 'Native Hydrogen/Deuterium Exchange Mass Spectrometry of Structured DNA Oligonucleotides'. *Analytical Chemistry* 92(6):4402–10. doi: 10.1021/acs.analchem.9b05298.

Largy, Eric, and Matthieu Ranz. 2023. 'OligoR: A Native HDX/MS Data Processing Application Dedicated to Oligonucleotides'. *Analytical Chemistry* 95(25):9615–22. doi: 10.1021/acs.analchem.3c01321.

Lau, Andy M., Jürgen Claesen, Kjetil Hansen, and Argyris Politis. 2020a. 'Deuteros 2.0: Peptide-Level Significance Testing of Data from Hydrogen Deuterium Exchange Mass Spectrometry'. *Bioinformatics* (btaa677). doi: 10.1093/bioinformatics/btaa677.

Lau, Andy M., Ruyu Jia, Richard T. Bradshaw, and Argyris Politis. 2020b. 'Structural Predictions of the Functions of Membrane Proteins from HDX-MS'. *Biochemical Society Transactions* 48(3):971–79. doi: 10.1042/BST20190880.

Lee, Kitaik, and Francis J. O'Reilly. 2023. 'Cross-Linking Mass Spectrometry for Mapping Protein Complex Topologies in Situ'. *Essays in Biochemistry* 67(2):215. doi: 10.1042/EBC20220168.

Leitner, Alexander, Thomas Walzthoeni, Abdullah Kahraman, Franz Herzog, Oliver Rinner, Martin Beck, and Ruedi Aebersold. 2010. 'Probing Native Protein Structures by Chemical Cross-Linking, Mass Spectrometry, and Bioinformatics*'. *Molecular & Cellular Proteomics* 9(8):1634–49. doi: 10.1074/mcp.R000001-MCP201.

LeMaster, David M., Janet S. Anderson, and Griselda Hernández. 2009. 'Peptide Conformer Acidity Analysis of Protein Flexibility Monitored by Hydrogen Exchange'. *Biochemistry* 48(39):9256–65. doi: 10.1021/bi901219x.

Lento, Cristina, and Derek J. Wilson. 2022. 'Subsecond Time-Resolved Mass Spectrometry in Dynamic Structural Biology'. *Chemical Reviews* 122(8):7624–46. doi: 10.1021/acs.chemrev.1c00222.

Lermyte, Frederik. 2020. *Advanced Fragmentation Methods in Biomolecular Mass Spectrometry*. Royal Society of Chemistry.

Levinthal, Cyrus. 1968. 'Are There Pathways for Protein Folding?' *Journal de Chimie Physique* 65:44–45. doi: 10.1051/jcp/1968650044.

Li, Jing, Mykola V. Rodnin, Alexey S. Ladokhin, and Michael L. Gross. 2014. 'Hydrogen–Deuterium Exchange and Mass Spectrometry Reveal the pH-Dependent Conformational Changes of Diphtheria Toxin T Domain'. *Biochemistry* 53(43):6849–56. doi: 10.1021/bi500893y.

Li, Ke Sherry, Elizabeth T. Schaper Bergman, Brett R. Beno, Richard Y. C. Huang, Ekaterina Deyanova, Guodong Chen, and Michael L. Gross. 2019. 'Hydrogen-Deuterium Exchange and Hydroxyl Radical Footprinting for Mapping Hydrophobic Interactions of Human Bromodomain with a Small Molecule Inhibitor'. *Journal of the American Society for Mass Spectrometry* 30(12):2795–2804. doi: 10.1007/s13361-019-02316-1.

Liko, Idlir, Timothy M. Allison, Jonathan TS Hopper, and Carol V. Robinson. 2016. 'Mass Spectrometry Guided Structural Biology'. *Current Opinion in Structural Biology* 40:136–44. doi: 10.1016/j.sbi.2016.09.008.

Lin, Xiaoxuan, Adam M. Zmyslowski, Isabelle A. Gagnon, Robert K. Nakamoto, and Tobin R. Sosnick. 2022. 'Development of in Vivo HDX-MS with Applications to a TonB-Dependent Transporter and Other Proteins'. *Protein Science* 31(9):e4402. doi: 10.1002/pro.4402.

Linderstrøm-Lang, K. 1955. 'The pH-Dependence of the Deuterium Exchange of Insulin'. *Biochimica et Biophysica Acta* 18:308. doi: 10.1016/0006-3002(55)90084-6.

Lindner, Robert, Xinghua Lou, Jochen Reinstein, Robert L. Shoeman, Fred A. Hamprecht, and Andreas Winkler. 2014. 'Hexicon 2: Automated Processing of Hydrogen-Deuterium Exchange Mass Spectrometry Data with Improved Deuteration Distribution Estimation'. *Journal of the American Society for Mass Spectrometry* 25(6):1018–28. doi: 10.1007/s13361-014-0850-y.

Liu, Sanmin, Lantao Liu, Ugur Uzuner, Xin Zhou, Manxi Gu, Weibing Shi, Yixiang Zhang, Susie Y. Dai, and Joshua S. Yuan. 2011. 'HDX-Analyzer: A Novel Package for Statistical Analysis of Protein Structure Dynamics'. *BMC Bioinformatics* 12(1):S43. doi: 10.1186/1471-2105-12-S1-S43.

López-Llano, J., L. a. Campos, and J. Sancho. 2006. 'α-Helix Stabilization by Alanine Relative to Glycine: Roles of Polar and Apolar Solvent Exposures and of Backbone Entropy'. *Proteins: Structure, Function, and Bioinformatics* 64(3):769–78. doi: 10.1002/prot.21041.

Lumpkin, Ryan J., and Elizabeth A. Komives. 2019. 'DECA, A Comprehensive, Automatic Post-Processing Program for HDX-MS Data'. *Molecular & Cellular Proteomics : MCP* 18(12):2516–23. doi: 10.1074/mcp.TIR119.001731.

Machado, Matías R., and Sergio Pantano. 2020. 'Split the Charge Difference in Two! A Rule of Thumb for Adding Proper Amounts of Ions in MD Simulations'. *Journal of Chemical Theory and Computation* 16(3):1367–72. doi: 10.1021/acs.jctc.9b00953.

Martens, Chloe, and Argyris Politis. 2020. 'A Glimpse into the Molecular Mechanism of Integral Membrane Proteins through Hydrogen–Deuterium Exchange Mass Spectrometry'. *Protein Science* 29(6):1285–1301. doi: 10.1002/pro.3853.

Marzolf, Daniel R., Justin T. Seffernick, and Steffen Lindert. 2021. 'Protein Structure Prediction from NMR Hydrogen–Deuterium Exchange Data'. *Journal of Chemical Theory and Computation* 17(4):2619–29. doi: 10.1021/acs.jctc.1c00077.

Masson, Glenn R., John E. Burke, Natalie G. Ahn, Ganesh S. Anand, Christoph Borchers, Sébastien Brier, George M. Bou-Assaf, John R. Engen, S. Walter Englander, Johan Faber, Rachel Garlish, Patrick R. Griffin, Michael L. Gross, Miklos Guttman, Yoshitomo Hamuro, Albert J. R. Heck, Damian Houde, Roxana E. Iacob, Thomas J. D. Jørgensen, Igor A. Kaltashov, Judith P. Klinman, Lars Konermann, Petr Man, Leland Mayne, Bruce D. Pascal, Dana Reichmann, Mark Skehel, Joost Snijder, Timothy S. Strutzenberg, Eric S. Underbakke, Cornelia Wagner, Thomas E. Wales, Benjamin T. Walters, David D. Weis, Derek J. Wilson, Patrick L. Wintrode, Zhongqi Zhang, Jie Zheng, David C. Schriemer, and Kasper D. Rand. 2019. 'Recommendations for Performing, Interpreting and Reporting Hydrogen Deuterium Exchange Mass Spectrometry (HDX-MS) Experiments'. *Nature Methods* 16(7):595–602. doi: 10.1038/s41592-019-0459-y.

Masson, Glenn R., Meredith L. Jenkins, and John E. Burke. 2017. 'An Overview of Hydrogen Deuterium Exchange Mass Spectrometry (HDX-MS) in Drug Discovery'. *Expert Opinion on Drug Discovery* 12(10):981–94. doi: 10.1080/17460441.2017.1363734.

McAllister, Robert G., and Lars Konermann. 2015. 'Challenges in the Interpretation of Protein H/D Exchange Data: A Molecular Dynamics Simulation Perspective'. *Biochemistry* 54(16):2683–92. doi: 10.1021/acs.biochem.5b00215.

Mistarz, Ulrik H., Bruno Bellina, Pernille F. Jensen, Jeffery M. Brown, Perdita E. Barran, and Kasper D. Rand. 2018. 'UV Photodissociation Mass Spectrometry Accurately Localize Sites of Backbone Deuteration in Peptides'. *Analytical Chemistry* 90(2):1077–80. doi: 10.1021/acs.analchem.7b04683.

Mitra, Gopa. 2021. 'Emerging Role of Mass Spectrometry-Based Structural Proteomics in Elucidating Intrinsic Disorder in Proteins'. *Proteomics* 21(3–4):2000011. doi: 10.1002/pmic.202000011.

Molday, R. S., S. W. Englander, and R. G. Kallen. 1972. 'Primary Structure Effects on Peptide Group Hydrogen Exchange'. *Biochemistry* 11(2):150–58. doi: 10.1021/bi00752a003.

Möller, Ingvar R., Patrick S. Merkle, Dionisie Calugareanu, Gerard Comamala, Solveig Gaarde Schmidt, Claus J. Loland, and Kasper D. Rand. 2020. 'Probing the Conformational Impact of Detergents on the Integral Membrane Protein LeuT by Global HDX-MS'. *Journal of Proteomics* 225:103845. doi: 10.1016/j.jprot.2020.103845.

Mori, Susumu, Peter C. M. van Zijl, and David Shortle. 1997. 'Measurement of water–amide proton exchange rates in the denatured state of staphylococcal nuclease by a magnetization transfer technique'. *Proteins: Structure, Function, and Bioinformatics* 28(3):325–32. doi: 10.1002/(SICI)1097-0134(199707)28:3<325::AID-PROT3>3.0.CO;2-B.

Moroco, Jamie A., and John R. Engen. 2015. 'Replication in Bioanalytical Studies with HDX MS: Aim as High as Possible'. *Bioanalysis* 7(9):1065–67. doi: 10.4155/bio.15.46.

Moulick, Roumita, Ranabir Das, and Jayant B. Udgaonkar. 2015. 'Partially Unfolded Forms of the Prion Protein Populated under Misfolding-Promoting Conditions'. *The Journal of Biological Chemistry* 290(42):25227–40. doi: 10.1074/jbc.M115.677575.

Na, Seungjin, Jae-Jin Lee, Jong Wha J. Joo, Kong-Joo Lee, and Eunok Paek. 2019. 'deMix: Decoding Deuterated Distributions from Heterogeneous Protein States via HDX-MS'. *Scientific Reports* 9. doi: 10.1038/s41598-019-39512-8.

Narang, Dominic, Cristina Lento, and Derek J. Wilson. 2020. 'HDX-MS: An Analytical Tool to Capture Protein Motion in Action'. *Biomedicines* 8(7):224. doi: 10.3390/biomedicines8070224.

Neelov, Igor M., David B. Adolf, Tom C. B. McLeish, and Emanuele Paci. 2006. 'Molecular Dynamics Simulation of Dextran Extension by Constant Force in Single Molecule AFM'. *Biophysical Journal* 91(10):3579–88. doi: 10.1529/biophysj.105.079236.

Nguyen, David, Leland Mayne, Michael C. Phillips, and S. Walter Englander. 2018. 'Reference Parameters for Protein Hydrogen Exchange Rates'. *Journal of the American Society for Mass Spectrometry* 29(9):1936–39. doi: 10.1007/s13361-018-2021-z.

Nguyen, David, Leland Mayne, Michael C. Phillips, and S. Walter Englander. 2018. 'Reference Parameters for Protein Hydrogen Exchange Rates'. *Journal of the American Society for Mass Spectrometry* 29(9):1936–39. doi: 10.1007/s13361-018-2021-z.

Oberg, Ann L., and Olga Vitek. 2009. 'Statistical Design of Quantitative Mass Spectrometry-Based Proteomic Experiments'. *Journal of Proteome Research* 8(5):2144–56. doi: 10.1021/pr8010099.

Oganesyan, Irina, Cristina Lento, and Derek J. Wilson. 2018. 'Contemporary Hydrogen Deuterium Exchange Mass Spectrometry'. *Methods* 144:27–42. doi: 10.1016/j.ymeth.2018.04.023.

Oliva, Luis. 2024. 'Using HDX-MS and FPOP-MS/MS to Identify Structural Differences in Intrinsically Disordered Regions of NFkB and IκBα upon Binding'. *Biophysical Journal* 123(3):217a. doi: 10.1016/j.bpj.2023.11.1377.

Olofsson, Anders, Malin Lindhagen-Persson, Monika Vestling, A. Elisabeth Sauer-Eriksson, and Anders Öhman. 2009. 'Quenched Hydrogen/Deuterium Exchange NMR Characterization of Amyloid-β Peptide Aggregates Formed in the Presence of Cu2+ or Zn2+'. *The FEBS Journal* 276(15):4051–60. doi: 10.1111/j.1742-4658.2009.07113.x.

Osborne, H. Beverley, and Eliane Nabedryk-Viala. 1978. 'The Conformation of Membrane-Bound and Detergent-Solubilised Bovine Rhodopsin'. *European Journal of Biochemistry* 89(1):81–88. doi: 10.1111/j.1432-1033.1978.tb20898.x.

Paci, Emanuele, and Martin Karplus. 2000. 'Unfolding Proteins by External Forces and Temperature: The Importance of Topology and Energetics'. *Proceedings of the National Academy of Sciences* 97(12):6521–26. doi: 10.1073/pnas.100124597.

Pan, Jingxi, Jun Han, and Christoph H. Borchers. 2012. 'Top-down Hydrogen/Deuterium Exchange and ECD-Stitched FTICR-MS for Probing Structural Dynamics of a 29-

kDa Enzyme'. *International Journal of Mass Spectrometry* 325–327:130–38. doi: 10.1016/j.ijms.2012.06.021.

Pan, Jingxi, Jun Han, Christoph H. Borchers, and Lars Konermann. 2009. 'Hydrogen/Deuterium Exchange Mass Spectrometry with Top-Down Electron Capture Dissociation for Characterizing Structural Transitions of a 17 kDa Protein'. *Journal of the American Chemical Society* 131(35):12801–8. doi: 10.1021/ja904379w.

Pancsa, Rita, Mihaly Varadi, Peter Tompa, and Wim F. Vranken. 2016. 'Start2Fold: A Database of Hydrogen/Deuterium Exchange Data on Protein Folding and Stability'. *Nucleic Acids Research* 44(D1):D429–34. doi: 10.1093/nar/gkv1185.

Park, In-Hee, John D. Venable, Caitlin Steckler, Susan E. Cellitti, Scott A. Lesley, Glen Spraggon, and Ansgar Brock. 2015. 'Estimation of Hydrogen-Exchange Protection Factors from MD Simulation Based on Amide Hydrogen Bonding Analysis'. *Journal of Chemical Information and Modeling* 55(9):1914–25. doi: 10.1021/acs.jcim.5b00185.

Parson, Matthew A. H., Meredith L. Jenkins, and John E. Burke. 2022. 'Investigating How Intrinsically Disordered Regions Contribute to Protein Function Using HDX-MS'. *Biochemical Society Transactions* 50(6):1607–17. doi: 10.1042/BST20220206.

Pascal, Bruce D., Scooter Willis, Janelle L. Lauer, Rachelle R. Landgraf, Graham M. West, David Marciano, Scott Novick, Devrishi Goswami, Michael J. Chalmers, and Patrick R. Griffin. 2012. 'HDX Workbench: Software for the Analysis of H/D Exchange MS Data'. *Journal of the American Society for Mass Spectrometry* 23(9):1512–21. doi: 10.1021/jasms.8b04350.

Peacock, Riley B., and Elizabeth A. Komives. 2021. 'Hydrogen/Deuterium Exchange and Nuclear Magnetic Resonance Spectroscopy Reveal Dynamic Allostery on Multiple Time Scales in the Serine Protease Thrombin'. *Biochemistry* 60(46):3441–48. doi: 10.1021/acs.biochem.1c00277.

Persson, Filip, and Bertil Halle. 2015. 'How Amide Hydrogens Exchange in Native Proteins'. *Proceedings of the National Academy of Sciences* 112(33):10383–88. doi: 10.1073/pnas.1506079112.

Perutz, M. F., M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will, and A. C. North. 1960. 'Structure of Haemoglobin: A Three-Dimensional Fourier Synthesis at 5.5-A. Resolution, Obtained by X-Ray Analysis'. *Nature* 185(4711):416–22. doi: 10.1038/185416a0.

Peterle, Daniele, Thomas E. Wales, and John R. Engen. 2022. 'Simple and Fast Maximally Deuterated Control (maxD) Preparation for Hydrogen–Deuterium Exchange Mass

Spectrometry Experiments'. *Analytical Chemistry* 94(28):10142–50. doi: 10.1021/acs.analchem.2c01446.

Petrosyan, Rafayel, Abhishek Narayan, and Michael T. Woodside. 2021. 'Single-Molecule Force Spectroscopy of Protein Folding'. *Journal of Molecular Biology* 433(20):167207. doi: 10.1016/j.jmb.2021.167207.

Petruk, Ariel A., Lucas A. Defelipe, Ramiro G. Rodríguez Limardo, Hernán Bucci, Marcelo A. Marti, and Adrian G. Turjanski. 2013. 'Molecular Dynamics Simulations Provide Atomistic Insight into Hydrogen Exchange Mass Spectrometry Experiments'. *Journal of Chemical Theory and Computation* 9(1):658–69. doi: 10.1021/ct300519v.

Pierson, Nicholas A., Liuxi Chen, David H. Russell, and David E. Clemmer. 2013. 'Cis–Trans Isomerizations of Proline Residues Are Key to Bradykinin Conformations'. *Journal of the American Chemical Society* 135(8):3186–92. doi: 10.1021/ja3114505.

Pierson, Nicholas A., Liuxi Chen, Stephen J. Valentine, David H. Russell, and David E. Clemmer. 2011. 'Number of Solution States of Bradykinin from Ion Mobility and Mass Spectrometry Measurements'. *Journal of the American Chemical Society* 133(35):13810–13. doi: 10.1021/ja203895j.

Pierson, Nicholas A., Stephen J. Valentine, and David E. Clemmer. 2010. 'Evidence for a Quasi-Equilibrium Distribution of States for Bradykinin [M+3H]3+ Ions in the Gas Phase'. *The Journal of Physical Chemistry. B* 114(23):7777–83. doi: 10.1021/jp102478k.

Pinheiro, Josè, Douglas Bates, and R Core Team. 2022. 'Nlme: Linear and Nonlinear Mixed Effects Models'.

Plotkin, Steven S., and José N. Onuchic. 2002. 'Understanding Protein Folding with Energy Landscape Theory Part I: Basic Concepts'. *Quarterly Reviews of Biophysics* 35(2):111–67. doi: 10.1017/S0033583502003761.

Porter, Lauren L., and Loren L. Looger. 2018. 'Extant Fold-Switching Proteins Are Widespread'. *Proceedings of the National Academy of Sciences* 115(23):5968–73. doi: 10.1073/pnas.1800168115.

Prusiner, Stanley B. 1998. 'Prions'. *Proceedings of the National Academy of Sciences* 95(23):13363–83. doi: 10.1073/pnas.95.23.13363.

Puchała, Weronika, Michał Burdukiewicz, Michał Kistowski, Katarzyna A. Dąbrowska, Aleksandra E. Badaczewska-Dawid, Dominik Cysewski, and Michał Dadlez. 2020. 'HaDeX: An R Package and Web-Server for Analysis of Data from Hydrogen–Deuterium Exchange Mass Spectrometry Experiments'. *Bioinformatics* 36(16):4516–18. doi: 10.1093/bioinformatics/btaa587.

Radou, Gaël, Frauke N. Dreyer, Roman Tuma, and Emanuele Paci. 2014. 'Functional Dynamics of Hexameric Helicase Probed by Hydrogen Exchange and Simulation'. *Biophysical Journal* 107(4):983–90. doi: 10.1016/j.bpj.2014.06.039.

Rand, Kasper D. 2013. 'Pinpointing Changes in Higher-Order Protein Structure by Hydrogen/Deuterium Exchange Coupled to Electron Transfer Dissociation Mass Spectrometry'. *International Journal of Mass Spectrometry* 338:2–10. doi: 10.1016/j.ijms.2012.08.010.

Rand, Kasper D., Martin Zehl, and Thomas J. D. Jørgensen. 2014. 'Measuring the Hydrogen/Deuterium Exchange of Proteins at High Spatial Resolution by Mass Spectrometry: Overcoming Gas-Phase Hydrogen/Deuterium Scrambling'. *Accounts of Chemical Research* 47(10):3018–27. doi: 10.1021/ar500194w.

Raval, Shaunak, Vladimir Sarpe, Morgan Hepburn, D. Alex Crowder, Terry Zhang, Rosa Viner, and David C. Schriemer. 2021. 'Improving Spectral Validation Rates in Hydrogen–Deuterium Exchange Data Analysis'. *Analytical Chemistry* 93(9):4246–54. doi: 10.1021/acs.analchem.0c05045.

Rayleigh, Lord. 1882. 'On the Equilibrium of Liquid Conducting Masses Charged with Electricity'. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 14(87):184–86. doi: 10.1080/14786448208628425.

Rey, Martial, Vladimir Sarpe, Kyle M. Burns, Joshua Buse, Charles A. H. Baker, Marc van Dijk, Linda Wordeman, Alexandre M. J. J. Bonvin, and David C. Schriemer. 2014. 'Mass Spec Studio for Integrative Structural Biology'. *Structure* 22(10):1538–48. doi: 10.1016/j.str.2014.08.013.

Rob, Tamanna, Peter Liuni, Preet Kamal Gill, Shaolong Zhu, Naresh Balachandran, Paul J. Berti, and Derek J. Wilson. 2012. 'Measuring Dynamics in Weakly Structured Regions of Proteins Using Microfluidics-Enabled Subsecond H/D Exchange Mass Spectrometry'. *Analytical Chemistry* 84(8):3771–79. doi: 10.1021/ac300365u.

Rob, Tamanna, and Derek J. Wilson. 2009. 'A Versatile Microfluidic Chip for Millisecond Time-Scale Kinetic Studies by Electrospray Mass Spectrometry'. *Journal of the American Society for Mass Spectrometry* 20(1):124–30. doi: 10.1016/j.jasms.2008.09.005.

Rohl, Carol A., Wayne Fiori, and Robert L. Baldwin. 1999. 'Alanine Is Helix-Stabilizing in Both Template-Nucleated and Standard Peptide Helices'. *Proceedings of the National Academy of Sciences* 96(7):3682–87. doi: 10.1073/pnas.96.7.3682.

Ross, James F., Gemma C. Wildsmith, Michael Johnson, Daniel L. Hurdiss, Kristian Hollingsworth, Rebecca F. Thompson, Majid Mosayebi, Chi H. Trinh, Emanuele Paci, Arwen R. Pearson, Michael E. Webb, and W. Bruce Turnbull. 2019. 'Directed Assembly of Homopentameric Cholera Toxin B-Subunit Proteins into Higher-Order

Structures Using Coiled-Coil Appendages'. *Journal of the American Chemical Society* 141(13):5211–19. doi: 10.1021/jacs.8b11480.

Röst, Hannes L., Uwe Schmitt, Ruedi Aebersold, and Lars Malmström. 2014. 'pyOpenMS: A Python-Based Interface to the OpenMS Mass-Spectrometry Algorithm Library'. *PROTEOMICS* 14(1):74–77. doi: 10.1002/pmic.201300246.

Rucker, Adam L., and Trevor P. Creamer. 2002. 'Polyproline II Helical Structure in Protein Unfolded States: Lysine Peptides Revisited'. *Protein Science : A Publication of the Protein Society* 11(4):980–85.

Ryckaert, Jean-Paul, Giovanni Ciccotti, and Herman J. C. Berendsen. 1977. 'Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes'. *Journal of Computational Physics* 23(3):327–41. doi: 10.1016/0021-9991(77)90098-5.

Sabareesan, Ambadi Thody, Jogender Singh, Samrat Roy, Jayant B. Udgaonkar, and M. K. Mathew. 2016. 'The Pathogenic A116V Mutation Enhances Ion-Selective Channel Formation by Prion Protein in Membranes'. *Biophysical Journal* 110(8):1766–76. doi: 10.1016/j.bpj.2016.03.017.

Salisbury, Joseph P., Qian Liu, and Jeffrey N. Agar. 2014. 'QUDeX-MS: Hydrogen/Deuterium Exchange Calculation for Mass Spectra with Resolved Isotopic Fine Structure'. *BMC Bioinformatics* 15(1):403. doi: 10.1186/s12859-014-0403-1.

Salmas, Ramin E., Matthew J. Harris, and Antoni J. Borysik. 2023. 'Mapping HDX-MS Data to Protein Conformations through Training Ensemble-Based Models'. *Journal of the American Society for Mass Spectrometry*. doi: 10.1021/jasms.3c00145.

Salmas, Ramin Ekhteiari, and Antoni James Borysik. 2021. 'HDXmodeller: An Online Webserver for High-Resolution HDX-MS with Auto-Validation'. *Communications Biology* 4(1):1–8. doi: 10.1038/s42003-021-01709-x.

Salomon-Ferrer, Romelia, Andreas W. Götz, Duncan Poole, Scott Le Grand, and Ross C. Walker. 2013. 'Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald'. *Journal of Chemical Theory and Computation* 9(9):3878–88. doi: 10.1021/ct400314y.

Saltzberg, Daniel J., Howard B. Broughton, Riccardo Pellarin, Michael J. Chalmers, Alfonso Espada, Jeffrey A. Dodge, Bruce D. Pascal, Patrick R. Griffin, Christine Humblet, and Andrej Sali. 2017. 'A Residue-Resolved Bayesian Approach to Quantitative Interpretation of Hydrogen–Deuterium Exchange from Mass Spectrometry: Application to Characterizing Protein–Ligand Interactions'. *The Journal of Physical Chemistry B* 121(15):3493–3501. doi: 10.1021/acs.jpcb.6b09358.

Schober, Patrick, Christa Boer, and Lothar A. Schwarte. 2018. 'Correlation Coefficients: Appropriate Use and Interpretation'. *Anesthesia & Analgesia* 126(5):1763. doi: 10.1213/ANE.0000000000002864.

Schopper, Simone, Abdullah Kahraman, Pascal Leuenberger, Yuehan Feng, Ilaria Piazza, Oliver Müller, Paul J. Boersema, and Paola Picotti. 2017. 'Measuring Protein Structural Changes on a Proteome-Wide Scale Using Limited Proteolysis-Coupled Mass Spectrometry'. *Nature Protocols* 12(11):2391–2410. doi: 10.1038/nprot.2017.100.

Scrosati, Pablo M., Victor Yin, and Lars Konermann. 2021. 'Hydrogen/Deuterium Exchange Measurements May Provide an Incomplete View of Protein Dynamics: A Case Study on Cytochrome c'. *Analytical Chemistry* 93(42):14121–29. doi: 10.1021/acs.analchem.1c02471.

Scrucca, Luca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. 2016. 'Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models'. *The R Journal* 8(1):289–317.

Seetaloo, Neeleema, Monika Kish, and Jonathan J. Phillips. 2022. 'HDfleX: Software for Flexible High Structural Resolution of Hydrogen/Deuterium-Exchange Mass Spectrometry Data'. *Analytical Chemistry* 94(11):4557–64. doi: 10.1021/acs.analchem.1c05339.

Seetaloo, Neeleema, and Jonathan J. Phillips. 2022. 'Millisecond Hydrogen/Deuterium-Exchange Mass Spectrometry for the Study of Alpha-Synuclein Structural Dynamics Under Physiological Conditions'. *Journal of Visualized Experiments: JoVE* (184). doi: 10.3791/64050.

Senior, Andrew W., Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. 2020. 'Improved Protein Structure Prediction Using Potentials from Deep Learning'. *Nature* 577(7792):706–10. doi: 10.1038/s41586-019-1923-7.

Senko, Michael W., Steven C. Beu, and Fred W. McLaffertycor. 1995. 'Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions'. *Journal of the American Society for Mass Spectrometry* 6(4):229–33. doi: 10.1016/1044-0305(95)00017-8.

Serdyuk, Igor N., Nathan R. Zaccai, and Joseph Zaccai. 2007. *Methods in Molecular Biophysics: Structure, Dynamics, Function*. Cambridge University Press.

Shan, Yibing, Anton Arkhipov, Eric T. Kim, Albert C. Pan, and David E. Shaw. 2013. 'Transitions to Catalytically Inactive Conformations in EGFR Kinase'. *Proceedings*

of the National Academy of Sciences 110(18):7270–75. doi: 10.1073/pnas.1220843110.

Sheff, Joey G., Martial Rey, and David C. Schriemer. 2013. 'Peptide–Column Interactions and Their Influence on Back Exchange Rates in Hydrogen/Deuterium Exchange-MS'. *Journal of the American Society for Mass Spectrometry* 24(7):1006–15. doi: 10.1007/s13361-013-0639-4.

Skinner, John J., Woon K. Lim, Sabrina Bédard, Ben E. Black, and S. Walter Englander. 2012. 'Protein Dynamics Viewed by Hydrogen Exchange'. *Protein Science* 21(7):996–1005. doi: 10.1002/pro.2081.

Skinner, John J., Wookyung Yu, Elizabeth K. Gichana, Michael C. Baxa, James R. Hinshaw, Karl F. Freed, and Tobin R. Sosnick. 2014. 'Benchmarking All-Atom Simulations Using Hydrogen Exchange'. *Proceedings of the National Academy of Sciences* 111(45):15975–80. doi: 10.1073/pnas.1404213111.

Skinner, Simon P., Rasmus H. Fogh, Wayne Boucher, Timothy J. Ragan, Luca G. Mureddu, and Geerten W. Vuister. 2016. 'CcpNmr AnalysisAssign: A Flexible Platform for Integrated NMR Analysis'. *Journal of Biomolecular NMR* 66(2):111–24. doi: 10.1007/s10858-016-0060-y.

Skinner, Simon P., Gael Radou, Roman Tuma, Jeanine J. Houwing-Duistermaat, and Emanuele Paci. 2019. 'Estimating Constraints for Protection Factors from HDX-MS Data'. *Biophysical Journal* 116(7):1194–1203. doi: 10.1016/j.bpj.2019.02.024.

Sljoka, Adnan, and Derek Wilson. 2013. 'Probing Protein Ensemble Rigidity and Hydrogen–Deuterium Exchange'. *Physical Biology* 10(5):056013. doi: 10.1088/1478-3975/10/5/056013.

Slysz, Gordon W., Charles AH Baker, Benjamin M. Bozsa, Anthony Dang, Andrew J. Percy, Melissa Bennett, and David C. Schriemer. 2009. 'Hydra: Software for Tailored Processing of H/D Exchange Data from MS or Tandem MS Analyses'. *BMC Bioinformatics* 10(1):162. doi: 10.1186/1471-2105-10-162.

Smit, Jochem H., Srinath Krishnamurthy, Bindu Y. Srinivasu, Spyridoula Karamanou, and Anastassios Economou. 2020. 'PyHDX: Derivation and Visualization of Protection Factors from Hydrogen-Deuterium Exchange Mass Spectrometry at near Residue Resolution'. *bioRxiv* 2020.09.30.320887. doi: 10.1101/2020.09.30.320887.

Smith, Lorna J., Klaus M. Fiebig, Harald Schwalbe, and Christopher M. Dobson. 1996. 'The Concept of a Random Coil: Residual Structure in Peptides and Denatured Proteins'. *Folding and Design* 1(5):R95–106. doi: 10.1016/S1359-0278(96)00046-6.

Sobott, Frank. 2020. 'Structural Studies Using Electron-Based Fragmentation Methods and Chemical Labelling of Proteins'. doi: 10.1039/9781839161056-00072.

Sosnick, Tobin R., and Doug Barrick. 2011. 'The Folding of Single Domain Proteins--Have We Reached a Consensus?' *Current Opinion in Structural Biology* 21(1):12–24. doi: 10.1016/j.sbi.2010.11.002.

Sowole, Modupeola A., John A. Alexopoulos, Yi-Qiang Cheng, Joaquin Ortega, and Lars Konermann. 2013. 'Activation of ClpP Protease by ADEP Antibiotics: Insights from Hydrogen Exchange Mass Spectrometry'. *Journal of Molecular Biology* 425(22):4508–19. doi: 10.1016/j.jmb.2013.08.005.

Ständer, Susanne, Laura R. Grauslund, Maria Scarselli, Nathalie Norais, and Kasper Rand. 2021. 'Epitope Mapping of Polyclonal Antibodies by Hydrogen–Deuterium Exchange Mass Spectrometry (HDX-MS)'. *Analytical Chemistry* 93(34):11669–78. doi: 10.1021/acs.analchem.1c00696.

Stephens, Amberley D., Maria Zacharopoulou, Rani Moons, Giuliana Fusco, Neeleema Seetaloo, Anass Chiki, Philippa J. Woodhams, Ioanna Mela, Hilal A. Lashuel, Jonathan J. Phillips, Alfonso De Simone, Frank Sobott, and Gabriele S. Kaminski Schierle. 2020. 'Extent of N-Terminus Exposure of Monomeric Alpha-Synuclein Determines Its Aggregation Propensity'. *Nature Communications* 11(1):2820. doi: 10.1038/s41467-020-16564-3.

Stofella, Michele, Simon P. Skinner, Frank Sobott, Jeanine Houwing-Duistermaat, and Emanuele Paci. 2022. 'High-Resolution Hydrogen–Deuterium Protection Factors from Sparse Mass Spectrometry Data Validated by Nuclear Magnetic Resonance Measurements'. *Journal of the American Society for Mass Spectrometry* 33(5):813–22. doi: 10.1021/jasms.2c00005.

Suckau, D., Y. Shi, S. C. Beu, M. W. Senko, J. P. Quinn, F. M. Wampler, and F. W. McLafferty. 1993. 'Coexisting Stable Conformations of Gaseous Protein Ions.' *Proceedings of the National Academy of Sciences* 90(3):790–93. doi: 10.1073/pnas.90.3.790.

Sun, Haofeng, Lingyun Ma, Leyu Wang, Peng Xiao, Hongmei Li, Min Zhou, and Dewei Song. 2021. 'Research Advances in Hydrogen–Deuterium Exchange Mass Spectrometry for Protein Epitope Mapping'. *Analytical and Bioanalytical Chemistry* 413(9):2345–59. doi: 10.1007/s00216-020-03091-9.

Svejdal, Rasmus R., Eleanor R. Dickinson, Drago Sticker, Jörg P. Kutter, and Kasper D. Rand. 2019. 'Thiol-Ene Microfluidic Chip for Performing Hydrogen/Deuterium Exchange of Proteins at Subsecond Time Scales'. *Analytical Chemistry* 91(2):1309–17. doi: 10.1021/acs.analchem.8b03050.

Tajoddin, Nastaran N., and Lars Konermann. 2020. 'Analysis of Temperature-Dependent H/D Exchange Mass Spectrometry Experiments'. *Analytical Chemistry* 92(14):10058–67. doi: 10.1021/acs.analchem.0c01828.

Tajoddin, Nastaran N., and Lars Konermann. 2022. 'Structural Dynamics of a Thermally Stressed Monoclonal Antibody Characterized by Temperature-Dependent H/D Exchange Mass Spectrometry'. *Analytical Chemistry* 94(44):15499–509. doi: 10.1021/acs.analchem.2c03931.

Tartaglia, Gian Gaetano, Andrea Cavalli, and Michele Vendruscolo. 2007. 'Prediction of Local Structural Stabilities of Proteins from Their Amino Acid Sequences'. *Structure* 15(2):139–43. doi: 10.1016/j.str.2006.12.007.

Theillet, Francois-Xavier, Andres Binolfi, Tamara Frembgen-Kesner, Karan Hingorani, Mohona Sarkar, Ciara Kyne, Conggang Li, Peter B. Crowley, Lila Gierasch, Gary J. Pielak, Adrian H. Elcock, Anne Gershenson, and Philipp Selenko. 2014. 'Physicochemical Properties of Cells and Their Effects on Intrinsically Disordered Proteins (IDPs)'. *Chemical Reviews* 114(13):6661–6714. doi: 10.1021/cr400695p.

Tian, Chuan, Koushik Kasavajhala, Kellon A. A. Belfon, Lauren Raguette, He Huang, Angela N. Migues, John Bickel, Yuzhang Wang, Jorge Pincay, Qin Wu, and Carlos Simmerling. 2020. 'ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution'. *Journal of Chemical Theory and Computation* 16(1):528–52. doi: 10.1021/acs.jctc.9b00591.

Toth, Ronald T. IV, Brittney J. Mills, Sangeeta B. Joshi, Reza Esfandiary, Steven M. Bishop, C. Russell Middaugh, David B. Volkin, and David D. Weis. 2017. 'Empirical Correction for Differences in Chemical Exchange Rates in Hydrogen Exchange-Mass Spectrometry Measurements'. *Analytical Chemistry* 89(17):8931–41. doi: 10.1021/acs.analchem.7b01396.

Trabjerg, Esben, Zeinab E. Nazari, and Kasper D. Rand. 2018. 'Conformational Analysis of Complex Protein States by Hydrogen/Deuterium Exchange Mass Spectrometry (HDX-MS): Challenges and Emerging Solutions'. *TrAC Trends in Analytical Chemistry* 106:125–38. doi: 10.1016/j.trac.2018.06.008.

Truhlar, Stephanie M. E., Carrie H. Croy, Justin W. Torpey, Julia R. Koeppe, and Elizabeth A. Komives. 2006. 'Solvent Accessibility of Protein Surfaces by Amide H/2H Exchange MALDI-TOF Mass Spectrometry'. *Journal of The American Society for Mass Spectrometry* 17(11):1490–97. doi: 10.1016/j.jasms.2006.07.023.

Tsiatsiani, Liana, Michiel Akeroyd, Maurien Olsthoorn, and Albert J. R. Heck. 2017. 'Aspergillus Niger Prolyl Endoprotease for Hydrogen–Deuterium Exchange Mass Spectrometry and Protein Structural Studies'. *Analytical Chemistry* 89(15):7966–73. doi: 10.1021/acs.analchem.7b01161.

Tyanova, Stefka, Tikira Temu, and Juergen Cox. 2016. 'The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics'. *Nature Protocols* 11(12):2301–19. doi: 10.1038/nprot.2016.136.

Urey, Harold C., F. G. Brickwedde, and G. M. Murphy. 1932. 'A Hydrogen Isotope of Mass 2 and Its Concentration'. *Physical Review* 40(1):1–15. doi: 10.1103/PhysRev.40.1.

Vaux, David L., Fiona Fidler, and Geoff Cumming. 2012. 'Replicates and Repeats—What Is the Difference and Is It Significant?' *EMBO Reports* 13(4):291–96. doi: 10.1038/embor.2012.36.

Vendruscolo, Michele, Emanuele Paci, Christopher M. Dobson, and Martin Karplus. 2003. 'Rare Fluctuations of Native Proteins Sampled by Equilibrium Hydrogen Exchange'. *Journal of the American Chemical Society* 125(51):15686–87. doi: 10.1021/ja036523z.

Verma, Sunita., Steven C. Pomerantz, Satinder K. Sethi, and James A. McCloskey. 1986. 'Fast Atom Bombardment Mass Spectrometry Following Hydrogen-Deuterium Exchange'. *Analytical Chemistry* 58(14):2898–2902. doi: 10.1021/ac00127a002.

Vijay-Kumar, Senadhi, Charles E. Bugg, and William J. Cook. 1987. 'Structure of Ubiquitin Refined at 1.8Åresolution'. *Journal of Molecular Biology* 194(3):531–44. doi: 10.1016/0022-2836(87)90679-6.

Vinciauskaite, Vanesa, and Glenn R. Masson. 2023. 'Fundamentals of HDX-MS'. *Essays in Biochemistry* 67(2):301–14. doi: 10.1042/EBC20220111.

Vranken, Wim F., Wayne Boucher, Tim J. Stevens, Rasmus H. Fogh, Anne Pajon, Miguel Llinas, Eldon L. Ulrich, John L. Markley, John Ionides, and Ernest D. Laue. 2005. 'The CCPN Data Model for NMR Spectroscopy: Development of a Software Pipeline'. *Proteins: Structure, Function, and Bioinformatics* 59(4):687–96. doi: 10.1002/prot.20449.

Wales, Thomas E., Keith E. Fadgen, Michael J. Eggertson, and John R. Engen. 2017. 'Subzero Celsius Separations in Three-Zone Temperature Controlled Hydrogen Deuterium Exchange Mass Spectrometry'. *Journal of Chromatography. A* 1523:275–82. doi: 10.1016/j.chroma.2017.05.067.

Wales, Thomas E., Aleksandra Pajak, Alžběta Roeselová, Santosh Shivakumaraswamy, Steven Howell, F. Ulrich Hartl, John R. Engen, and David Balchin. 2022. 'Resolving Chaperone-Assisted Protein Folding on the Ribosome at the Peptide Level'. 2022.09.23.509153.

Walters, Benjamin T., Alec Ricciuti, Leland Mayne, and S. Walter Englander. 2012. 'Minimizing Back Exchange in the Hydrogen Exchange-Mass Spectrometry Experiment'. *Journal of the American Society for Mass Spectrometry* 23(12):2132–39. doi: 10.1021/jasms.8b04189.

Wang, Bin, Lei Zhang, Tong Dai, Ziran Qin, Huasong Lu, Long Zhang, and Fangfang Zhou. 2021. 'Liquid–Liquid Phase Separation in Human Health and Diseases'. *Signal Transduction and Targeted Therapy* 6(1):1–16. doi: 10.1038/s41392-021-00678-1.

Wang, Boshen, Alan Perez-Rathke, Renhao Li, and Jie Liang. 2018. 'A General Method for Predicting Amino Acid Residues Experiencing Hydrogen Exchange'. Pp. 341–44 in *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*.

Wei, Hui, Joomi Ahn, Ying Qing Yu, Adrienne Tymiak, John R. Engen, and Guodong Chen. 2012. 'Using Hydrogen/Deuterium Exchange Mass Spectrometry to Study Conformational Changes in Granulocyte Colony Stimulating Factor upon PEGylation'. *Journal of the American Society for Mass Spectrometry* 23(3):498–504. doi: 10.1007/s13361-011-0310-x.

Wei, Hui, Jingjie Mo, Li Tao, Reb J. Russell, Adrienne A. Tymiak, Guodong Chen, Roxana E. Iacob, and John R. Engen. 2014. 'Hydrogen/Deuterium Exchange Mass Spectrometry for Probing Higher Order Structure of Protein Therapeutics: Methodology and Applications'. *Drug Discovery Today* 19(1):95–102. doi: 10.1016/j.drudis.2013.07.019.

Weis, David D. 2016. *Hydrogen Exchange Mass Spectrometry of Proteins: Fundamentals, Methods, and Applications*. John Wiley & Sons.

Weis, David D. 2021. 'Recommendations for the Propagation of Uncertainty in Hydrogen Exchange-Mass Spectrometric Measurements'. *Journal of the American Society for Mass Spectrometry*. doi: 10.1021/jasms.0c00475.

Weis, David D., John R. Engen, and Ignatius J. Kass. 2006. 'Semi-Automated Data Processing of Hydrogen Exchange Mass Spectra Using HX-Express'. *Journal of the American Society for Mass Spectrometry* 17(12):1700–1703. doi: 10.1016/j.jasms.2006.07.025.

Weis, David D., Thomas E. Wales, John R. Engen, Matthew Hotchko, and Lynn F. Ten Eyck. 2006. 'Identification and Characterization of EX1 Kinetics in H/D Exchange Mass Spectrometry by Peak Width Analysis'. *Journal of the American Society for Mass Spectrometry* 17(11):1498–1509. doi: 10.1021/jasms.8b02565.

Wilson, Derek J., and Lars Konermann. 2003. 'A Capillary Mixer with Adjustable Reaction Chamber Volume for Millisecond Time-Resolved Studies by Electrospray Mass Spectrometry'. *Analytical Chemistry* 75(23):6408–14. doi: 10.1021/ac0346757.

Wohl, Samuel, Matthew Jakubowski, and Wenwei Zheng. 2021. 'Salt-Dependent Conformational Changes of Intrinsically Disordered Proteins'. *The Journal of Physical Chemistry Letters* 12(28):6684–91. doi: 10.1021/acs.jpclett.1c01607.

Wójcik, J., K. Ruszczyńska, I. Zhukov, and A. Ejchart. 1999. 'NMR Measurements of Proton Exchange between Solvent and Peptides and Proteins'. *Acta Biochimica Polonica* 46(3):651–63. doi: 10.18388/abp.1999_4137.

Wollenberg, Daniel T. Weltz, Stuart Pengelley, Jeppe Christian Mouritsen, Detlev Suckau, Christian Isak Jørgensen, and Thomas J. D. Jørgensen. 2020. 'Avoiding H/D Scrambling with Minimal Ion Transmission Loss for HDX-MS/MS-ETD Analysis on a High-Resolution Q-TOF Mass Spectrometer'. *Analytical Chemistry* 92(11):7453–61. doi: 10.1021/acs.analchem.9b05208.

Wood, Victoria E., Kate Groves, Adam Cryar, Milena Quaglia, Paul Matejtschuk, and Paul A. Dalby. 2020. 'HDX and In Silico Docking Reveal That Excipients Stabilize G-CSF via a Combination of Preferential Exclusion and Specific Hotspot Interactions'. *Molecular Pharmaceutics* 17(12):4637–51. doi: 10.1021/acs.molpharmaceut.0c00877.

Wrigley, Michael S., Hayley Blockinger, H. M. Emranul Haque, Sachini P. Karunaratne, and David D. Weis. 2024. 'Optimization of a Hydrogen Exchange-Mass Spectrometry Robotic Liquid Handler Using Tracers'. *Analytical Chemistry* 96(4):1522–29. doi: 10.1021/acs.analchem.3c04186.

Wu, Qian. 2020. 'Guardians of the Genome: DNA Damage and Repair'. *Essays in Biochemistry* 64(5):683–85. doi: 10.1042/EBC20200109.

Yu, Hongtao. 1999. 'Extending the Size Limit of Protein Nuclear Magnetic Resonance'. *Proceedings of the National Academy of Sciences of the United States of America* 96(2):332–34.

Zehl, Martin, Kasper D. Rand, Ole N. Jensen, and Thomas J. D. Jørgensen. 2008. 'Electron Transfer Dissociation Facilitates the Measurement of Deuterium Incorporation into Selectively Labeled Peptides with Single Residue Resolution'. *Journal of the American Chemical Society* 130(51):17453–59. doi: 10.1021/ja805573h.

Zhang, Mengru Mira, Brett R. Beno, Richard Y. C. Huang, Jagat Adhikari, Ekaterina G. Deyanova, Jing Li, Guodong Chen, and Michael L. Gross. 2019. 'An Integrated Approach for Determining a Protein-Protein Binding Interface in Solution and an Evaluation of HDX Kinetics for Adjudicating Candidate Docking Models'. *Analytical Chemistry* 91(24):15709. doi: 10.1021/acs.analchem.9b03879.

Zhang, Naifu, Xiaohe Yu, Xinchao Zhang, and Sheena D'Arcy. 2021. 'HD-eXplosion: Visualization of Hydrogen–Deuterium Exchange Data as Chiclet and Volcano Plots with Statistical Filtering'. *Bioinformatics* 37(13):1926–27. doi: 10.1093/bioinformatics/btaa892.

Zhang, Zhongqi. 2020. ‘Complete Extraction of Protein Dynamics Information in Hydrogen/Deuterium Exchange Mass Spectrometry Data’. *Analytical Chemistry* 92(9):6486–94. doi: 10.1021/acs.analchem.9b05724.

Zhang, Zhongqi, Aming Zhang, and Gang Xiao. 2012. ‘Improved Protein Hydrogen/Deuterium Exchange Mass Spectrometry Platform with Fully Automated Data Processing’. *Analytical Chemistry* 84(11):4942–49. doi: 10.1021/ac300535r.

Zhu, Shaolong, Agnesa Shala, Alexandr Bezginov, Adnan Sljoka, Gerald Audette, and Derek J. Wilson. 2015. ‘Hyperphosphorylation of Intrinsically Disordered Tau Protein Induces an Amyloidogenic Shift in Its Conformational Ensemble’. *PLOS ONE* 10(3):e0120416. doi: 10.1371/journal.pone.0120416.

Ziemianowicz, Daniel S., Vladimir Sarpe, D. Alex Crowder, Troy J. Pells, Shaunak Raval, Morgan Hepburn, Atefeh Rafiei, and David C. Schriemer. 2020. ‘Harmonizing Structural Mass Spectrometry Analyses in the Mass Spec Studio’. *Journal of Proteomics* 225:103844. doi: 10.1016/j.jprot.2020.103844.