

Inferring residue level hydrogen deuterium exchange with ReX

Oliver M. Crook *¹, Nathan Gittens², Chun-wa Chung², and Charlotte M. Deane¹

¹*Department of Statistics, University of Oxford, Oxford, UK*

²*Structural and Biophysical Sciences, GlaxoSmithKline R&D, Stevenage, UK*

April 8, 2024

Abstract

Hydrogen-Deuterium Exchange Mass-Spectrometry (HDX-MS) has emerged as a powerful technique to explore the conformational dynamics of proteins and protein complexes in solution. The bottom-up approach to MS uses peptides to represent an average of residues, leading to reduced resolution of deuterium exchange and complicates the interpretation of the data. Here, we introduce ReX, a method to infer residue-level uptake patterns leveraging the overlap in peptides, the temporal component of the data and the correlation along the sequence dimension. This approach infers statistical significance for individual residues by treating HDX-MS as a multiple change-point problem. By fitting our model in a Bayesian non-parametric framework, we perform parameter number inference, differential HDX confidence assessments, and uncertainty estimation for temporal kinetics. Benchmarking against existing methods using a three-way proteolytic digestion experiment shows our method's superior performance at predicting unseen HDX data. Moreover, it aligns HDX-MS with the reporting standards of other structural methods by providing global and local resolution metrics. Using ReX, we analyze the differential flexibility of BRD4's two Bromodomains in the presence of I-BET151 and quantify the conformational variations induced by a panel of seventeen small molecules on LXR α . Our analysis reveals distinct residue-level HDX signatures for ligands with varied functional outcomes, highlighting the potential of this characterisation to inform mode of action analysis.

1 Introduction

Understanding the conformational dynamics of proteins and the conformational changes induced by their interactions is key to determining their function. One method to explore such dynamics is hydrogen-deuterium exchange mass-spectrometry (HDX-MS), which tracks the rate of exchange of amide hydrogens with deuterium in solution (Orengo *et al.*, 1999; Engen, 2009; Chalmers *et al.*, 2011; Houde *et al.*, 2011; Masson *et al.*, 2017; Sauve *et al.*, 2018). This rate is impacted by various factors such as secondary (and higher order) structures, as well as solvent accessibility (Katta and

*oliver.crook@stats.ox.ac.uk

Chait, 1993; Konermann *et al.*, 2011; Kammari and Topp, 2020; Jia *et al.*, 2020). It is fundamentally based on the concept of local folding and unfolding, as suggested in the Linderstrøm-Lang theory (James *et al.*, 2021). The numerous applications of HDX-MS can include: the characterisation of binding interfaces of antibody-antigen complexes by mapping of their epitopes and paratopes (Tsai *et al.*, 2020); examining structural stabilisation by small molecules and their corresponding allosteric effects (Kielkopf *et al.*, 2019); and the effect of amino-acid mutations on transporters (Jia *et al.*, 2020) and applications to vaccine design (Gorman *et al.*, 2016).

The majority of HDX-MS studies involve a bottom-up approach to mass-spectrometry (Masson *et al.*, 2017). In this widely employed strategy, proteins undergo on-line proteolytic digestion via a non-specific protease—often pepsin or nepenthesin-II—immobilized to a column, usually yielding good sequence coverage by reporting deuterium uptake at the peptide fragment level. Although the selection of protease can be tailored to specific applications, inherent limitations emerge due to the resolution loss at the peptide level, obstructing a nuanced interpretation of the molecular phenomena under scrutiny. Hence, bottom-up HDX-MS leads to distinct scenarios that remain unaddressed by current statistical methodology. First, combining information from overlapping peptides (often referred to as redundancy) and replicates lacks a clear approach. Second, instances where overlapping peptides exhibit a mixture of protection and de-protection effects require a means of resolution, which is currently addressed by somewhat arbitrary approaches such as data omission (Pascal *et al.*, 2012). Third, differences in the length of the overlapping peptides introduces interpretation issues, especially when contrasting modest, consistent effects on longer peptides against more variable effects on shorter ones. We refer to this phenomena as the length-bias issue. Finally, these challenges need to be considered whilst also accounting for the temporal nature of HDX data and correlations along the sequence dimension.

Projecting peptide-level data to residue-level data is an apparent solution to the previously outlined challenges, and several strategies have been proposed in this direction (Babić and Smith, 2017; Babić *et al.*, 2019; Althaus *et al.*, 2010; Zhang *et al.*, 2012; Zhang, 2020; Hamuro and Zhang, 2018; Kan *et al.*, 2013, 2019; Liu *et al.*, 2011; Claesen and Politis, 2018; Skinner *et al.*, 2019; Saltzberg *et al.*, 2017; Salmas and Borysik, 2021; Smit *et al.*, 2021; Gessner *et al.*, 2017; Keppel and Weis, 2014). These proposed approaches can be divided into non-optimisation approaches (e.g. (Babić and Smith, 2017)), such as taking a running average along the sequence dimension (Keppel and Weis, 2014), and optimisation-based approaches where a model is developed and then fit to the data by minimising the discrepancy between the model and data (e.g. (Gessner *et al.*, 2017; Smit *et al.*, 2021; Salmas and Borysik, 2021; Skinner *et al.*, 2019)). Such models can incorporate physically relevant constraints, such as prohibiting negative deuterium uptake. While a residue-level model, in principle, could overcome the aforementioned issues, it introduces its own set of challenges. The main obstacle is the underdetermination of an appropriate model, implying a scenario where model parameters outnumber data points. Further, HDX suffers from a more extreme form of underdetermination that cannot be mitigated by acquiring additional data. This arises due to the potential countless combinations of deuterium uptake for two adjacent residues, each with unit redundancy, as long as the sum of the uptake is consistent with the observed data. All of these drawbacks become exacerbated when one wishes to study multiple protein states and perform differential HDX-MS experiments. Here, it becomes paramount to associate a measure of statistical confidence to the results to avoid false positives. Currently, there is no appropriate

statistical framework that attempts to tackle these issue.

In response to these outstanding methodological challenges, we developed ReX - a statistical method complemented by an associated software package, designed to achieve residue-resolved HDX. ReX combines a likelihood model, which models the deuterium uptake per residue, with a prior change-point model that permits correlations or jumps between the parameters of adjacent residues. ReX is a non-parametric Bayesian model (see ([Crook *et al.*, 2022a](#)) for examples), meaning it does not fix the number of parameters, allowing model complexity to expand with the quality and volume of data. The Bayesian framework allow us to associate probability distributions to pertinent quantities (such as model parameters) that are updated in the light of data ([Crook *et al.*, 2022a](#)). This updated distribution, termed the posterior distribution, allows us to quantify the uncertainty in our model inferences.

Despite a number of computational approaches available to obtain residue-level HDX from overlapping peptide-level HDX data, a comprehensive benchmarking of such approaches is absent from the literature. Directly conducting residue-level HDX via nuclear magnetic resonance (NMR) or electron transfer dissociation (ETD) has its own set of drawbacks ([Wei *et al.*, 2014](#); [Stofella *et al.*, 2022](#)). In particular, NMR has low sensitivity whilst ETD has poor efficiency and typically needs sample specific optimisation. Consequently, comparisons leveraging different experimental platforms are likely to only report on qualitative differences. Here, we propose a simple and extensible benchmark using a triple digestion of Cytochrome C. In this approach, we test the ability of different methods to reconstruct the HDX-MS data from the unseen digestions and quantify the error. In this experiment, we expect the same underlying HDX results, but by using different peptide digestions to generate independent datasets allows for a critical examination of whether diverse methods yield consistent residue-resolved HDX-MS profiles.

In this manuscript, we first detail the ReX model specification and posterior inference. We proceed to explain and perform a benchmarking analysis, highlighting ReX's superior performance but also providing insights into the recommendations and potential pitfalls associated with alternative methods. Subsequently, we illustrate how ReX can evaluate the resolution and quality of HDX data, as well as evaluate the consistency of structure prediction methods such as AlphaFold ([Jumper *et al.*, 2021](#)). We further showcase ReX's capabilities through a variety of examples and extensive simulations. A comparative analysis of the stabilisation of the two bromodomains of BRD4 using I-BET151 demonstrates that the BC loop of the second bromodomain is more flexible than that of the first bromodomain. We then perform conformational signature analysis (CSA) on LXR α with seventeen small molecules at the residue-level. Here, we demonstrate the additional insights that can be drawn by quantifying uncertainty in the conformational landscape induced by the small molecules. The associated R-package (<https://github.com/ococrook/RexMS>) enables single-protein analysis (<https://ococrook.github.io/RexMS/articles/ReX.html>), differential analysis (<https://ococrook.github.io/RexMS/articles/DifferentialRexMS.html>) and conformational signature analysis (<https://ococrook.github.io/RexMS/articles/ConformationalSignatureAnalysis.html>). The package comes with a variety of visualisations including PDB structure visualisation. We also provide Shiny Application for interactive data visualisation and as a template so other may construct their own. To interact with the BRD4 and LXR α results visit the following web-pages: <https://olivercrook.shinyapps.io/BRD4-ReX/> and <https://olivercrook.shinyapps.io/ConformationalSignatureAnalysis/>.

2 Results

2.1 The ReX method and posterior inference

ReX is designed as a versatile statistical method aiming to facilitate the inference of residue-level HDX-MS across a range of experimental designs. It accommodates multiple protein states, including mutations, small molecules, complexes and more (see Fig. 1). Whilst replicates and many time-points are not a strict requirement of the input data, we advocate for them. Incorporating additional replicates and time-points, especially those spanning three or four orders of magnitude, not only helps ReX infer the model parameters but also enhances confidence and quality in the outcomes. To ensure consistency between peptides, ReX incorporates information from fully deuterated controls.

Central to ReX is a Bayesian change-point model, designed to perform sub-localisation of HDX uptake. We posit the existence of a latent (unobserved), residue-level process, analogous to assumptions entailed in a Hidden Markov Model (Eddy, 2004). For a given residue, there is a temporal uptake model parametrised as mixture between a stretched exponential (Weibull-type) model (Crook *et al.*, 2022b) and a standard exponential (logistic model). The mixture proportions are learned during the inference process, enabling the model to better adapt without necessitating pre-specification. Given that this model usually possesses an excessive number of parameters, regularization of the parameter space is performed by defining a prior model (Betancourt, 2021; Crook *et al.*, 2022a). The complexity of the inferred model is guided by the data, and thus, we non-parametrically specify the prior model as a change-point process (Green, 1995; Fearnhead, 2006). Here, parameters between segments of residues are similar or necessitate a jump. This implies that the number of change-points, their locations, and associated kinetic parameters must be inferred (see Fig. 1). While these quantities could be pre-specified, they are typically unknown, and hence, we include these quantities in the inference process. A prior Poisson process is employed to model the unknown (random) locations of these change-points along the data's residue axis (Kingman, 1992).

Inference in Bayesian models typically employs Markov-chain Monte-Carlo (MCMC) to draw samples from the posterior distribution of parameters, by using a Markov-Chain (Betancourt, 2017). However, a challenge arises when adding or removing a change-point which increases or decreases the dimension (i.e. the parameter count) of the model. Therefore, we developed an MCMC method that allows the Markov-chain to explore whilst increasing and decreasing the model's dimension. The Reversible Jump MCMC (RJ-MCMC), sometimes referred to as the trans-dimensional MCMC algorithm, facilitates such dimension switching (Green, 1995). This algorithm handles scenarios where the number of unknown entities is itself unknown. The intricacies are primarily in determining the transition probability when altering dimensions (see methods for more details).

Bridging the unobserved model to the data remains challenging. The observation model takes the residue-level model and adds up the residue values from the exchangeable residues covered by a peptide to derive peptide-level inferences which can be compared with the data. The residuals are posited to follow a Laplace distribution with the scale (variance) parameter multiplied by the length of the peptide. This approach down-weights the contribution of longer peptides, or more precisely, permits them to be more variable.

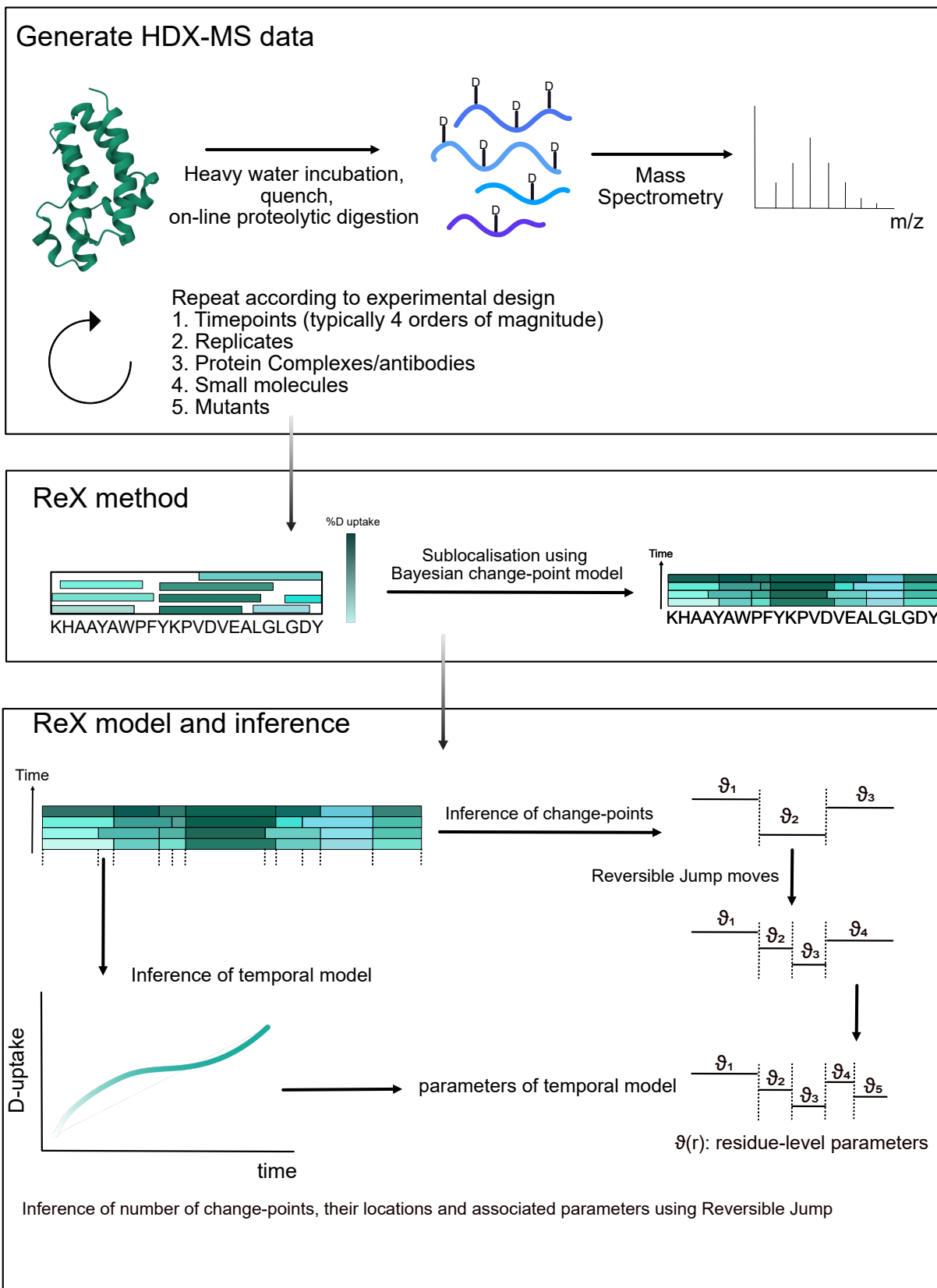


Figure 1: An Overview of the ReX method, model and inference.

Figure 1: ReX is applicable to general HDX-MS experiments performed using the bottom-up approach. Typical variations in workflow are allowed including changes to temperature, pH, buffer concentration. Experimental designs with at least three labelling time points and replicate measurements enhances the quality of ReX's outputs. The method can be used for any protein states such as protein complexes, antibodies, small molecules and mutants as examples. The goal of ReX is to take the peptide map generated by the experiment with deuterium reported on peptide fragments and infer (sub-localise) the deuterium exchange measurements at the level of residues. ReX uses a Bayesian change-point model on the parameters of the temporal kinetics. This non-parametric model is fitted using reversible jump Markov-chain Monte-Carlo.

2.2 An extensible benchmark for residue-level HDX-MS

In order to assess ReX, in comparison with other approaches, we devised a versatile and extensible experimental benchmark, designed to be straightforwardly implemented and extended by all HDX-MS practitioners. Taking cytochrome-C, HDX was conducted at four labelling time points (15,150,1500,15000s), initially undergoing digestion with nepenthesin II (Affipro) for three replicates. Subsequent digestions employing pepsin and aspergillopepsin (Type XIII) were performed in separate experiments. Each digest generates different peptide maps and we refer to this as a differential digestion experiment (see Fig. 2).

To computationally profile diverse approaches, residue-level uptake was initially inferred for each model under consideration, using data obtained from the nepenthesin II digest. Subsequently, the remaining two digests were ‘held-out’ and rendered invisible to the model, establishing our benchmark as a ‘leave-digests-out’ experiment. Each peptide from the ‘held-out’ data was considered in turn. We predicted their deuterium uptake by summing the already inferred residue values, thereby obtaining uptake ‘peptide-maps’ from both the predicted and observed data. To evaluate the performance of the model, we measured the deviation of the predicted and observed data (see Fig. 2), with the superior approaches having the lowest error. This benchmarking technique is adaptable to other digests, proteins, methods, and MS instruments, confirming its general applicability.

In order to compare to other methods, we searched the literature to find other open-source approaches. However, despite extensive treatment of the topic in the literature few methods are openly available. Only methods that were either simple enough to recode, or had available open-source code (ensuring both execution and veracity of the method) were benchmarked. Additionally, it was imperative that all benchmarked methods could report deuterium uptake, not merely differences or protection factors.

We successfully benchmarked several methods: the Pseudo-Inverse (P-Inv) approach (Babić and Smith, 2017), a least-squares based approach (LeastSq) (Gessner *et al.*, 2017), a weighted average approach (RFU-WA) (Keppel and Weis, 2014), and approaches based on the Fused-LASSO (PyHDX) (Skinner *et al.*, 2019; Smit *et al.*, 2021). Specifically, the weighted average approach calculates a residue-weighted average, considering the length of the peptides. P-Inv inverts a coupling matrix between residue and peptide-level uptakes, while LeastSq minimizes the deviation between an uptake model and observed data using a squared error loss function. PyHDX expands on this, incorporating an additional smoothness penalty between residues. Further details on all approaches are given in the methods.

The performance of each method under benchmark analysis is presented in Fig.3A, revealing ReX as a consistently accurate and physically reasonable approach for inferring deuterium uptake values. Furthermore, ReX exhibited the lowest Median Absolute Deviation (MAD) consistently in six out of eight examples, outperforming other methods and only being less accurate than the generalized-inverse approach (P-Inv) at one specific time point (150s) (See Fig.3A). Detailed examination revealed that P-Inv, in particular, generated unphysical kinetic curves and occasionally inferred negative uptake, violating physical principles and the monotonicity assumptions of deuterium uptake (Fig.3B). In contrast, ReX and the weighted average approach (RFU-WA) were the only methods that produced a monotonic deuterium uptake, with other methods sometimes presenting non-monotonic profiles (Fig.3C). An extension of our benchmark analysis, which considered

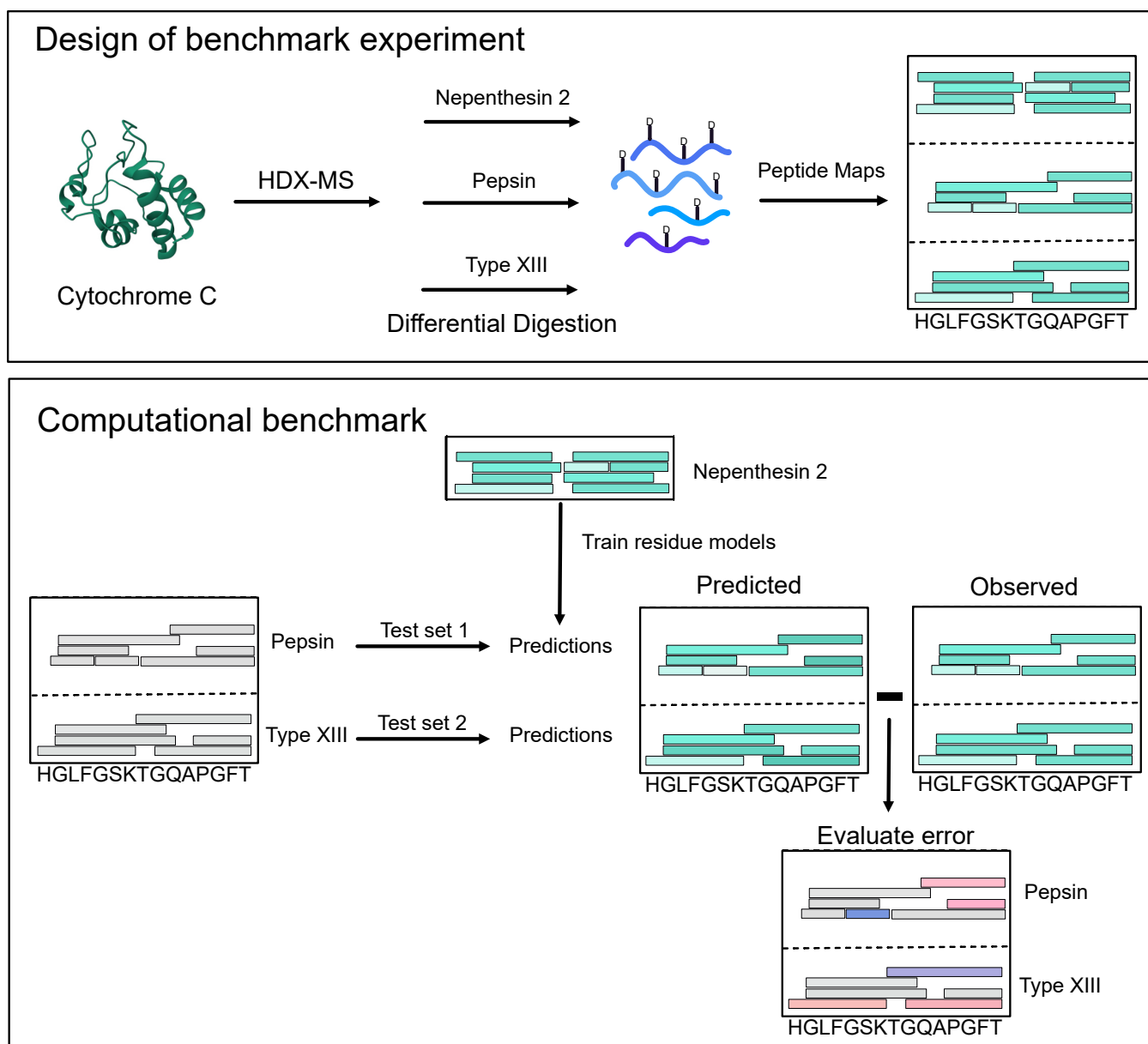


Figure 2: **Design of experimental and computational benchmark.** A triple-protease digest is performed on Cytochrome C to generate three different peptide maps. In turn the data from each digest is used to train residue level models which are then evaluated on the held-out data from the other digests.

alternative training data and downsampled Neptnesin data, confirmed ReX's superior performance in the largest number of comparisons, outperforming the next best method, PyHDX with $\lambda = 5$, across three additional benchmarking experiments, totalling 32 comparisons (Fig. 3D and Fig. S1). In instances where PyHDX excelled, ReX was the second-best method 80% of the time, solidifying its reliability. Ultimately, the challenge of all methods is identifying the correct amount of smoothness as overly smooth approaches (RFU-WA) and those with low amounts of smoothness (P-Inv) performed poorly. We also noted that the (P-Inv) approach become unstable at points of high-deuterium uptake.

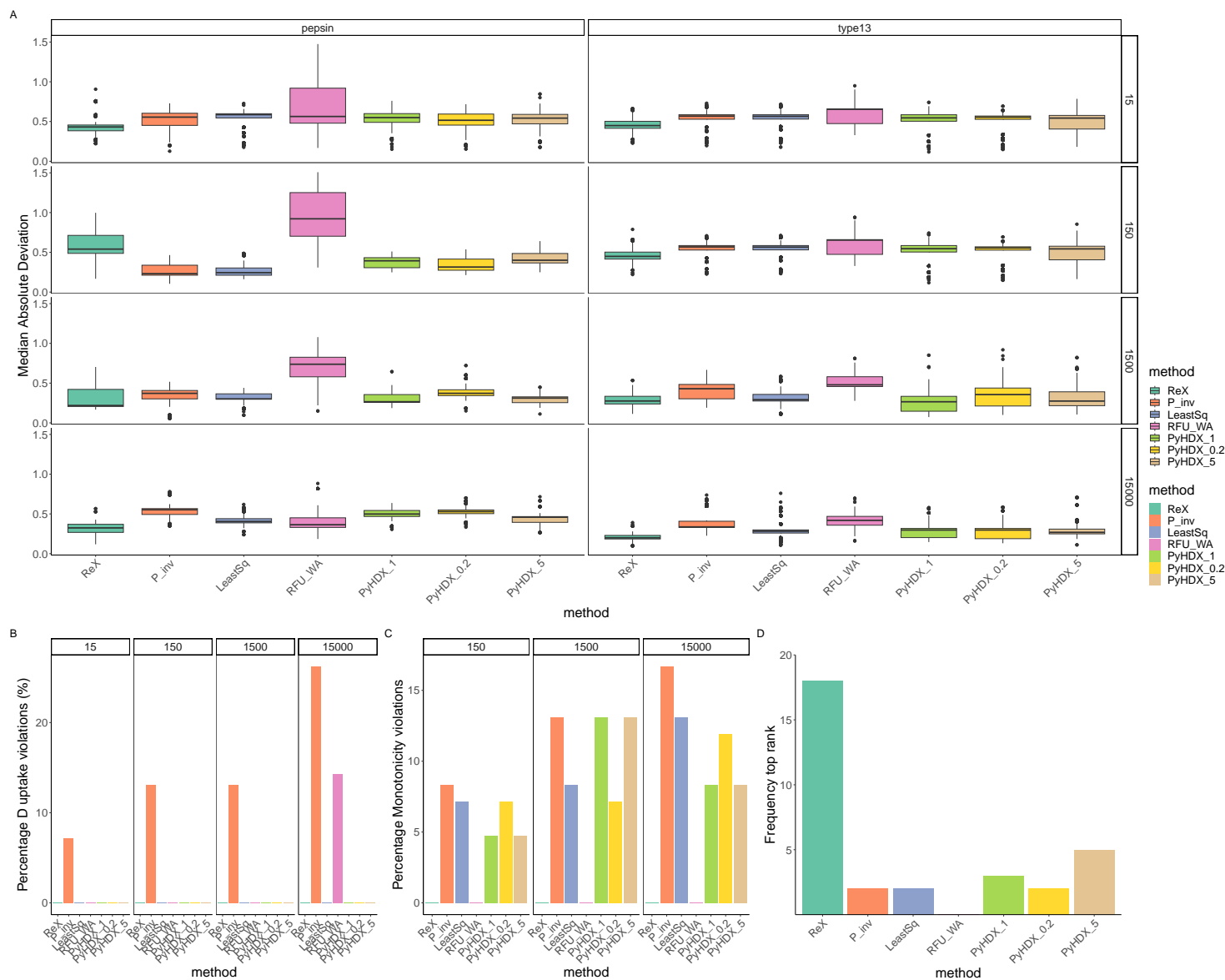


Figure 3: Benchmarking Results. (A) The bootstrap distributions of the median absolute deviations (MAD) are reported for the held-out digests faceted by each experimental time-point. Boxplots are in the Tukey format and represent approximate 95% confidence intervals. ReX has the lowest average error in six of the eight cases. (B) The percentage of residue-level uptakes that violate uptake constraints (uptake should be between zero and one). (C) The percentage of residue-level uptakes that are monotonicity violations. Deuterium uptake should be increasing or constant over time. (D) A summary of the ranking of the methods in all benchmarking scenarios (all digests and time-points). A method is top ranked if it has lowest average error.

2.3 Extracting features from ReX

Having established the accuracy of ReX, we extracted features from the model for subsequent downstream analysis. A current limitation of the HDX-MS literature is the routine absence of reported measures of resolution of the data. ReX outputs a scale or standard deviation parameter termed sigma (σ) and uses σ to model the deviation of its uptake fits from the observed data. Accordingly, σ is larger for more variable data. Hence, when we added additional standard deviations of noise to our Cytochrome C data and saw that σ increased in a close to linear fashion (see Fig. S2). This means that smaller values of σ indicate less variable data and large values of σ indicate greater variability. Consequently, σ is a viable global metric for data quality and resolution.

For a residue-level measure of data quality, we propose the average reconstructive error (ARE) - a measure of the model's ability in reconstructing observed data at the peptide level (see methods). Elevated ARE values for a residue implies consistently suboptimal modelling for overlapping peptides (Fig. S2). Significant variability in the value of the ARE highlights regions where fits are either of high or low quality. Moreover, the ARE is highest in the same region of the poorly modelled peptides in our benchmarking experiment. Furthermore, these results are consistent with adding additional variability to the data (see Fig. S2). Crucially, σ and the ARE are consistent, ensuring a credible interval constructed using σ of size $1 - \alpha$ means that at least $100 \times (1 - \alpha)\%$ of the data is covered by the credible interval (see Fig. S2). Hence, σ and ARE together are robust and communicable metrics for algorithmic faithfulness and dataset quality bringing it in-line with reporting standard in crystallography, electron microscopy and structural prediction.

We then examined four previously published HDX experiments on Alpha-Lactalbumin, Barnase, Enolase and Serium Amyloid Protein (SAP) (Salmas and Borysik, 2021), associating structural feature annotated via Uniprot (Consortium, 2015) and Free SASA with our model outputs (Mitternacht, 2016). We defined surface accessible (SA) as relative SA above 25% (Mitternacht, 2016). Though HDX cannot simply be reduced to these quantities, we found correlations across several of these annotations. Our results (presented in Fig. 4) are obtained by randomly resampling residues 1000 times and are reported as a distribution (see Fig. S3). Examining the mean number of change points for each of the structure features showed fewer change-points for non-surface accessible (NSA) regions compared with SA regions. We interpret fewer change-points (i.e. fewer jumps) as increased correlation in the sequence dimension. Hence, the result is likely because of a more stable hydrogen bond network in NSA regions of the proteins. Helices tended to have more change-points than strands perhaps indicative of the helices being generally able to concertina and more readily locally unfold and refold. Again, parallel strands appeared less stable to exchange than their anti-parallel counterparts, perhaps due to their weaker hydrogen bond angles (Tsutsumi and Otaki, 2011). When considering the mean uptake of these features, we found SA regions conferred a faster rate of exchange and higher maximum uptake. The effect of weaker hydrogen bonds of parallel strands was also observed with faster and higher uptake for parallel strands compared with the anti-parallel counterparts. Helices also appeared more amenable to deuterium exchange, suggesting perhaps a propensity to breath more.

Subsequently, we used ReX's features to predict the secondary structure of predicted structures of Cytochrome C. Using AlphaFold2 and multiple sequence alignment (MSA) subsampling (Wayment-Steele *et al.*, 2022), we generated 25 models of cytochrome C (See Fig. S11 and Supple-

mentary methods). By training random forest models with our HDX data, we achieved a median AUROC of 0.78 ([0.63, 0.88], 90% confidence interval) when predicting secondary structure (See Fig. S12 and supplementary methods). Prediction of SA and NSA were also obtained for each model and we computed the prediction confidence at each residue for both secondary structure and SA. Models with a confidence score (see supplementary methods) above 0.7 had a median RMSD from the crystal structure of 1.2Å, whereas models with a confidence score below 0.6 had a median RMSD of 4.1Å (see Fig. S13). This analysis suggests HDX could potentially triage AI generated protein structures.

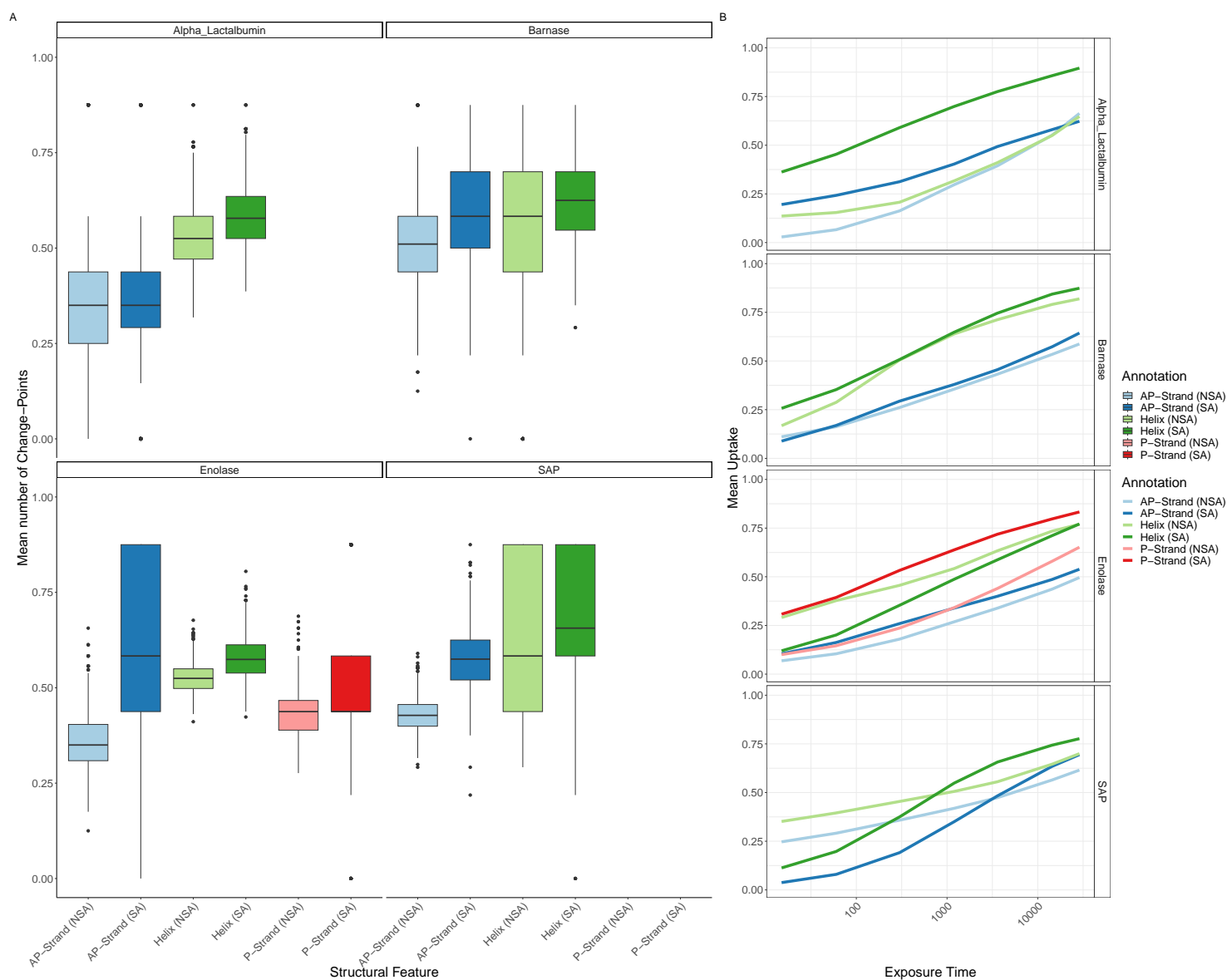


Figure 4: **ReX inferred features for structural annotations.** (A) The mean number of change-points stratified by structural annotation including strand-type, helix and solvent accessibility. The distributions are over random subsamples and boxplots are in the Tukey format and represent approximate 95% confidence intervals. The greater the number of change-points less correlated the uptake is along the sequence dimensions. (B) The same stratification as (A) but for the mean deuterium uptake over exposure time to heavy water.

2.4 Residue-Resolved differential HDX-MS with ReX

HDX-MS is a powerful technique in differential settings where conformational changes induced by interactions or perturbations can be revealed (Masson *et al.*, 2017). Details that are often challenging or impossible to obtain with other structural approaches are readily observable using HDX-MS (Engen and Komives, 2020). Having explored ReX’s abilities with single proteins, we then used it to quantify residue-level difference from differential HDX-MS experiments. ReX calculates residue-level probabilities of deuterium difference, enhancing the interpretation and visualisation of the resultant data. Given that HDX alone cannot currently de-convolute a binding site from conformational changes, we opt for an extensive simulation study to demonstrate ReX’s abilities.

Using our already generated cytochrome C data digested by nepenthesin II, we performed 120 simulations. This involved randomly selecting a residue from cytochrome-C and then subsequent reduction of deuterium uptake for all peptides covering that residue to specified percentages (80, 90, 95%) of the original uptake, with 80% representing the strongest protection effect. Each scenario was repeated four times across an additional ten random seeds representing a total of 120 simulations.

The challenge of quantifying individual residue differences arises from the need to account for the consistency, or lack thereof, among overlapping peptides. ReX, through a re-constructive approach, models the residue in each protein state under study, comparing each residue’s uptake with the APO (reference) state. We quantify the difference at the level of peptides by predicting the peptide-level values by summing over the posterior mean values for the appropriate residues. Subsequently, the total reconstructive error (TRE) (also effect size) is the value obtained by summing the effect for overlapping peptides. This summation approach, rather than average, gives more confidence to residues overlapping numerous peptides with a consistent direction of effect. For example, suppose four peptides overlap a residue and their deuterium difference is in random directions with consistent magnitude, then their sum will be close zero. In contrast, ten overlapping peptides each with the same direction of effect, but with small magnitude of change, will have greater value boosted by the summation of ten values. An associated probability calculation of confidence can be made (see methods for more details).

An illustrative analysis is shown in Fig. 5 for an 80% effect. Volcano plots, representing effect-size (TRE) vs Probability, are obtained for each time point with residues of interest located in the top corners (see Fig. 5A). However, volcano plots obscure spatial context, which can be re-introduced through line plots detailing effect size as a function of residue number, where transparency is scaled with probability. Our simulations induced a ripple effect where a clear protection effect is accompanied by a corresponding and adjacent de-protection effect. A subsequent protection effect but smaller is then observed proximally and so forth. When averaging the effect over time these effects are washed-out and the region of protection is clearly identified. Probabilities were transformed into an estimate of the expected false discovery rate (FDR), obtained by computing the average of the posterior error probabilities (see methods). For typical FDRs of interest ($FDR < 0.01$), we find that our probability estimates are calibrated but can underestimate the error at higher values of FDR or at specific time-points (see Fig. 5D). Averaging over the time dimension yielded consistent estimates of error (Fig. 5E). The complete set of examples is shown in the supplementary material, which comprises one week of simulations on 48 CPU cores.

An exhaustive evaluation of all simulations is detailed in the supplementary material, where initially the proximity of the randomly selected residue (for subsequent peptide perturbation) to the residue with highest probability of difference is evaluated. Typically, these were extremely close being within a couple of residues (see Fig. S4). Furthermore, it may not always be possible to improve this results if the residue was in a region of low redundancy. We then obtained the probability of this randomly selected residue having a deuterium difference. At an 80% effect, this randomly chosen residue had close to probability 1 in almost all examined scenarios with the probability reducing (as expected) for smaller protection effects. Our analysis confirms that our approach is well-calibrated and advocate for considering temporal averages if the effects and probabilities are small. Additional calibration plots are shown in Fig. S5.

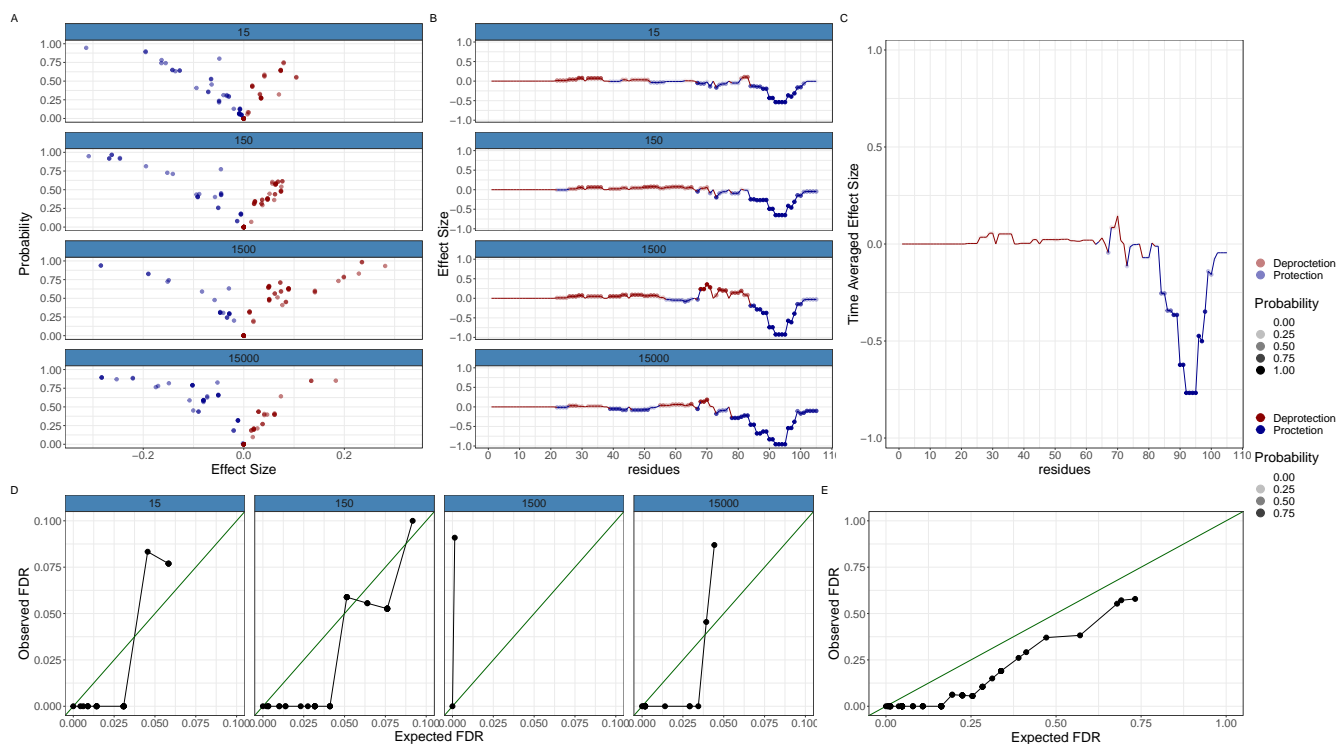


Figure 5: A simulated epitope mapping example. (A) Volcano plots stratified by time-points with probability of change in the y-axis and difference in Effect Size (ARE) in the x-axis. Protection is presented in blue and De-protection in red. (B) Line plots with difference in Effect Size (TRE) plotted as a function of residue. Protection is presented in blue and De-protection in red. The transparency is scaled with probability. (C) As in B but with the results averaged over time. (D) Calibration plots showing the expected false discovery rate (FDR) against the observed (FDR) are method is mostly conservative (points below the principal diagonal). (E) The same as (D) but averaged over time. The results are always conservative showing good calibration.

2.5 Comparative analysis of tandem Bromodomains in BRD4

The BET (Bromo- and Extra-Terminal domain) protein family, containing epigenetic readers such as BRD2, BRD3 and BRD4 (Chung *et al.*, 2011; Chung and Tough, 2012), have drawn significant attention due to their role in the recognition and binding to acetylated lysine residues on histones that regulate transcription (Filippakopoulos *et al.*, 2010). Inhibitors of the BET family of proteins have shown promise for a wide variety of therapeutic applications (Mujtaba *et al.*, 2007). These proteins contain tandem Bromodomains, where the functional role of each Bromodomain BD1 and BD2 has recently been clarified (Gilan *et al.*, 2020). Bromodomains have a characteristic four-helix bundle arranged in a left hand twist with high sequence similarity. For example, the majority of Bromodomains contain a canonical WPF shelf and a conserved asparagine (residue 140 in BD1 of BRD4). Sequence alignment of the two Bromodomains in BRD4 reveals high levels of conservation (See Fig. S10), yet subtle differences impact the respective domain's flexibility (Gilan *et al.*, 2020).

To understand the flexibility differences at residue-level between BD1 and BD2 of BRD4, we analysed triplicate HDX-MS experiments of BRD4 in the context of the non-specific inhibitor I-BET151 (Dawson *et al.*, 2011). Five labelling time-points (15,60,600,3600,14400s) were collected in triplicate in the presence of DMSO or I-BET151. We then applied ReX separately to each Bromodomain due to missing data in the proline-rich disordered region joining them.

Upon examining BD1, a strong stabilisation signature was observed at all time points (Fig. 6A) suggestive of molecule binding forming additional bonds and reducing HDX. The effect endured over extended labelling times, with minimal confident de-protection, indicating global stabilisation of the entire domain (Fig. 6B). Meanwhile, BD2 presented a an even more pronounced global stabilisation effect (Fig 6C). However, I-BET151 has a 13-fold higher domain selectivity for BD1 over BD2, attributed to a through-water interaction of the quinoline nitrogen to the side chain of Gln85 in BD1 (Petretich *et al.*, 2020). To gain further insights, we examined the time averaged effect in a line plot for both Bromodomains (Fig. 6D). The results are similar with marked stabilisation around the WPF shelves. However, the BC loop in BD2, proximal to the conserved asparagine, exhibited a much stronger stabilization signature than in BD1.

To rationalise this result, we examined the structure of the small molecule in the binding pocket and saw that I-BET 151 faces away from the BC loop (Fig. 6E) meaning that it unlikely that the molecule forms any new bonds in this region. Given that the selective inhibitor IBET-BD1 achieves selectivity by additionally binding to Asp144 via an appended pyrrolidine (Gilan *et al.*, 2020), we hypothesised that the BC loop, and the whole BD2, exhibits more intrinsic flexibility. Consequently, the enhanced stabilisation signature in BD2 might be attributed more to differential flexibility than to a significant disparities in the bonding network.

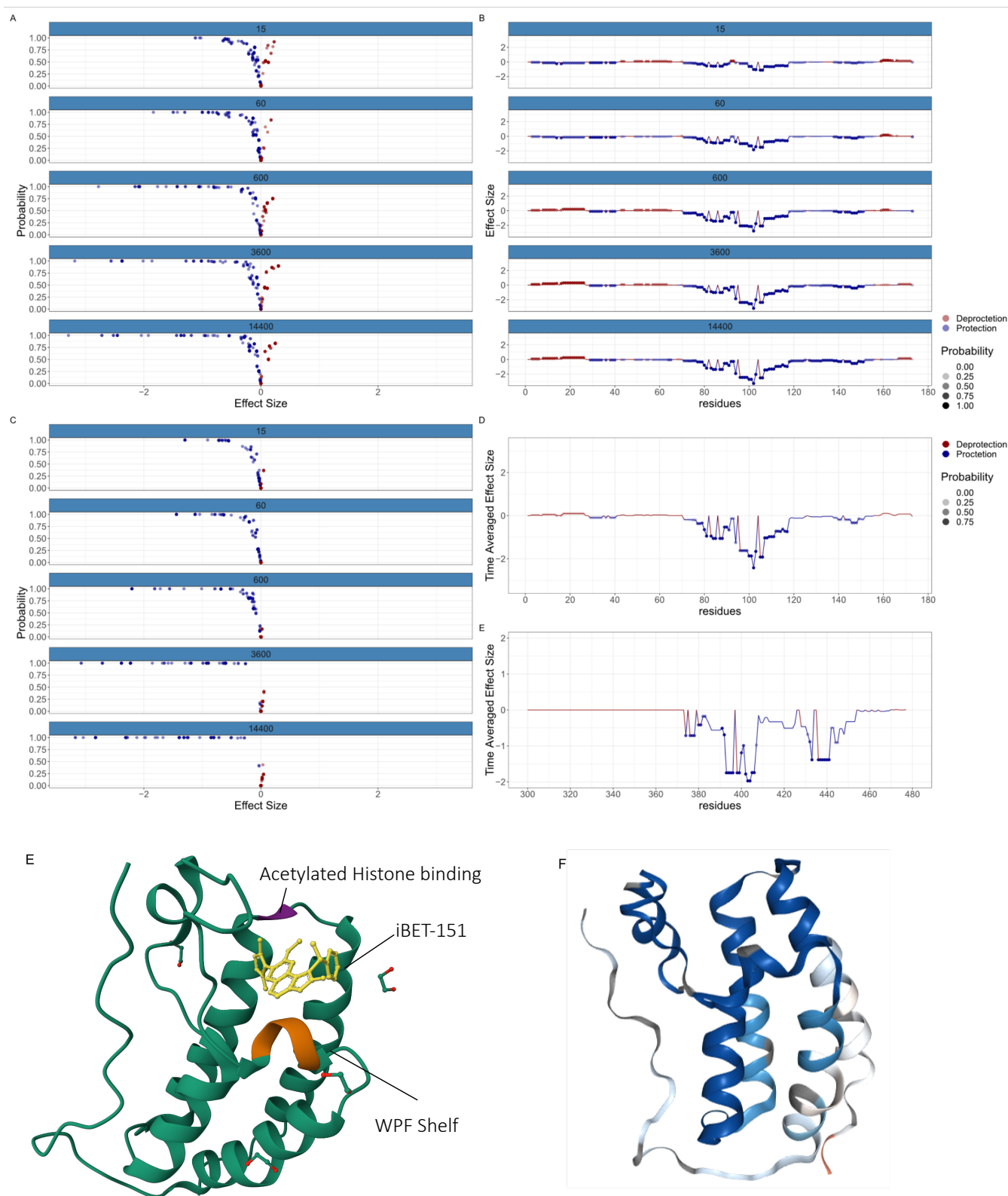


Figure 6: Comparative analysis of tandem Bromodomains in BRD4 using ReX (A) Volcano plots of Effect Size (TRE) against probability of difference of BD1 in BRD4 (B) The corresponding line plots associated with BD1 in BRD4. Transparency is scaled with probability. (C) As with (A) but for the BD2. (D) (Upper) A time averaged line plot showing protection effects as a function of residue for BD1 (lower) as with (upper) but for BD2, (E) iBET-151 in complex with BD1 of BRD4 with conserved asparagine and WPF shelf highlighted. The BC loop face away from the molecule. (F) A cartoon of BD1 in BRD4 with the HDX results plotted directly on the structure.

2.6 Conformational Signature Analysis of LXR α

LXR α is a member of the nuclear hormone receptors (NHR) family of transcription factors (Willy *et al.*, 1995; Janowski *et al.*, 1996). Upon binding with cholesterol derivatives, LXR α heterodimerizes with the retinoid X receptor to modulate gene expression (Hong and Tontonoz, 2014). Liver X receptors (LXRs) play a pivotal role in cholesterol homeostasis and are therapeutic targets for atherosclerotic cardiovascular disease (Venkateswaran *et al.*, 2000; Hong and Tontonoz, 2014). Induction of cholesterol transporters, such as ABCA1, by LXRs facilitates efflux of cholesterol from macrophages and subsequent excretion of cholesterol from the body (Wagner *et al.*, 2003). However, developing anti-atherogenic compounds has posed challenges because of their adverse lipogenic effects (Schultz *et al.*, 2000; Levin *et al.*, 2005), a consequence of LXRs regulating the expression of many lipogenic genes (Schultz *et al.*, 2000). Furthermore, it is believed that among LXRs it is LXR α that is primarily responsible for the lipogenic effects (Schultz *et al.*, 2000). Two therapeutic molecules WAY-252623 (Katz *et al.*, 2009) and BMS-852927 (Kick *et al.*, 2016) have reached clinical evaluation for their attenuated lipogenic effects and potent ABCA1 induction.

In order to understand how structural dynamics correlate with the *in vivo* pharmacology of the different available LXR compounds, Belorusova *et al.* (2019) conducted an HDX-MS study on seventeen molecules complexed with LXR α . These molecule were characterised by their *in vivo* beneficial and adverse effects using mouse models, then ranked based on their ABCA1 induction and induced plasma triglyceride (TG) levels. AZ5, AZ8, AZ9 had WAY-252623 exhibited low ABCA1 induction, with the remainder classified as being high inducers. Furthermore, AZ1, AZ876, T0901317, WAY-254011 and F1 were classed as lipogenic, AZ2-AZ5 were not classified, and the rest classified as non-lipogenic.

Given, the large number of molecules performing pairwise comparisons was infeasible; thus we (and the previous authors) applied multivariate statistical methods. Conformational Signature Analysis (CSA) was performed, whereby the HDX-MS data is formatted as a data matrix and treated statistically. Unlike the previous analysis, ReX was used to project the peptide-level data to residue-level data and additionally quantify uncertainty. The uniform treatment of all states enabled examination of the conformational landscape spanned by the perturbation induced by the ligands. Subsequently, we correlated this landscape with the *in vivo* determined pharmacological properties of the ligands.

First, ReX was employed and followed by principal component analysis (PCA) to visualise the high-dimensional data (see methods). PCA projects the data along orthogonal components, where the first component (PC1) has highest variance explained (largest eigenvalue) and subsequent components are then ordered by the amount of variance explained (see methods) (Holmes and Huber, 2018). The first three principal components are visualised across Fig. 7 A and B, with molecules colour-coded by molecular annotations. BMS-852927 was separated from all other molecule in both PC1 and PC2. This corresponded to considerably stronger stabilisation of helix 3 and the beta-sheets, and reduced stabilisation of helix 10-12. ABCA1 inducers separated broadly along PC1, whereas lipogenic molecules, though grouping together, presented more complex interpretation. The loadings, visualised directly on the structure in Fig. 7 G, indicate residues' contribution to the component variance with darker shades representing stronger contributions. As expected PC1 is driven by variation in beta-sheet and helix 3 stabilisation. PC2 variation is driven by residues on

helix 3 and 10, whilst PC3 variation is contributed to by residues on helix 9 and the loop from helix 9 to 10.

Adopting a flexible approach, functional annotations of the molecules were used in a supervised model. We employed orthogonal partial least squares discriminant analysis (OPLS-DA (Thevenot, 2016)), creating separate models for lipogenicity and ABCA1 induction (Fig. 7 C). OPLS-DA directly separates the data along a predictive component and subsequent orthogonal component(s). Associated residue weights (loadings) are depicted in the line plots in Fig. 7 E and F. The separation between high and low ABCA1 inducers was driven by residues on helix 3, the adjacent beta-sheet, and helix 5. Whereas separation along the lipogenic dimension involved larger contributions from helix 1, whether stabilisation and destabilisation depended on the molecule. Notably, the lipogenic molecules saw marked stabilisation of residues 214-219 and the residues 200-214 tended to be destabilised in non-lipogenic molecules. However, these effects are partially correlated with residues that contribute to ABCA1 induction, complicating the differentiation of the competing effects. Multidimensional scaling (MDS) on chemical (Tanimoto) similarity confirmed that molecules did not cluster, in chemical space, by their functional annotations (see Fig. S6).

Providing additional context, CSA was re-performed using ReX in the context of LXR α complexed with the second receptor interacting motif of the co-activator SRC1 (Belorusova *et al.*, 2019). Unliganded LXR α does not interact with the SRC1 peptide and all compounds induced binding of LXRs to the SRC1 peptide (Belorusova *et al.*, 2019). While separating the high and low ABCA1 inducers was more challenging, a clearer separation of the lipogenic classes was observed (see Supplementary Figs. S6, S7 and S8) with (18.3%) variance explained by the predictive component (as opposed to 15.1%). The ABCA1 separation was driven by similar residues as found without the co-activator peptide, with these residues also contributing to the class separation of the lipogenic molecules, but to a lesser extent. Moreover, residues on helix 10-12 gave a high contribution, which was not observed in the ABCA1 results or without the SRC1 peptide. Results from helix 1 changed in this context, with all ligands contributing to strong stabilization of this region. Lipogenic molecules appeared to enhance the interaction with SRC1 (mean $K_d = 1.2 \mu\text{M}$) with a weak agonist effect from the non-lipogenic molecules (Belorusova *et al.*, 2019). In addition, the co-activator peptide interacts with helix 12 supporting the marked reduction of HD exchange. Hence, the results might be explained by lipogenic molecules encouraging higher differential recruitment of the SRC1 peptide, explaining their HD results, while ABCA1 induction seems independent of this effect. Therefore, the therapeutic effects of LXR α ligands may hinge on the conformations they stabilize. Agonist ligands encourage helix 12 folding into the active state, creating a co-activator-binding surface and recruiting SRC1 (Belorusova *et al.*, 2019). Stabilisation of helix 3, 5, and corresponding beta sheets, which are the interaction sites of co-repressors induced high-levels of ABCA1. Perhaps the LXR co-repressors are disassociated from the complex during these conformations leading to elevated ABCA1 levels (Belorusova *et al.*, 2019).

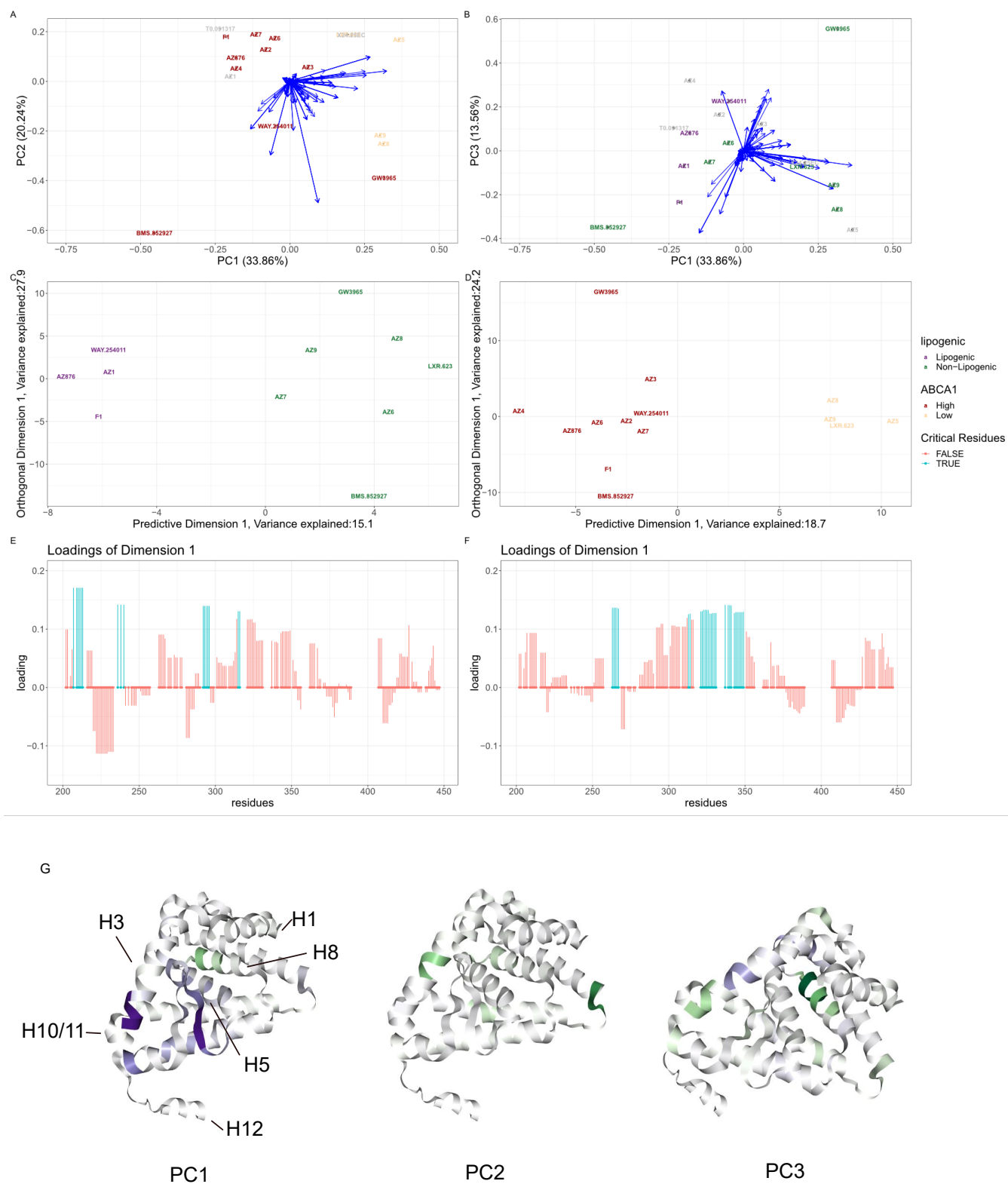


Figure 7: Conformational Signatures of LXR α .

Figure 7: (A) PCA plots of HDX-MS signatures per ligand. Ligands are coloured by ABCA1 induction and blue arrow represent residue contributions. (B) As with (A) but for PC1 versus PC3. The ligands are coloured by lipogenic annotation. (C) An OPLS-DA plot with lipogenic as the outcome variable (D) An OPLS-DA plot with ABCA1 induction as a outcome variable (E) The loading plot corresponding to the predictive dimension of (C) larger loadings indicate greater contribution. (F) A loading plot corresponding to the predictive dimension of (D). (G) PCA loading of (A) and (B) plotted directly on the protein structure of LXR α . Purple indicates a positive loading (positive correlation with the original variable) and Green indicates a negative loading (negative correlation with the original variable)

2.7 Uncertainty quantification of CSA

In the analysis of LXR α above, we have predominantly used point estimates, specifically means, from our Bayesian non-parametric approach ReX. While beneficial, this approach does not exploit the full potential of the methodology. Quantifying uncertainty in these analyses is paramount if they are to replace mouse experiments or be used for downstream pharmaceutical decisions.

To visualise the uncertainty in the conformational landscape, we project 50 samples from the posterior distribution (possible summaries) into PCA coordinates aligned using the procrustes transform (Nguyen and Holmes, 2019; Crook *et al.*, 2022c). Fig. 8a is particularly illuminating, confirming that BMS-852927 and GW3975 are positioned in distinct regions of the plot and the resulting contours suggest that these conformations are unlikely to overlap those induced by other molecules. Other molecules are also distinct or overlap somewhat, suggesting they induced similar conformations. In fact, numerous molecules, either annotated as high ABCA1 inducers or unannotated, overlap within the same spatial region, hinting that two unannotated molecules (AZ1, T0901317) could be potent ABCA1 inducers. All lipogenic molecules (and some non-lipogenic) occupy this spatial region, indicating that avoiding this part of the conformational landscape may be key to circumventing lipogenic side effects.

Exploring the uncertainty in the loadings, visualized in Fig. 8b, reveals the confidence levels attributed to specific residues contributing to variation in the conformational landscape. Caution is advised when interpreting residues where the loading contribution changes sign. This does not inherently indicate a change in the direction of protection or de-protection but if the weight does intersect zero this means that we are unsure of a non-zero contribution. However, the majority of residues contributions remain confident leaving conclusions unchanged. The C-terminal helix region displays substantial uncertainty, warranting further investigation in the context of the co-activator peptide (Supplementary Fig. S9). The quantiles, visualised in structure space and exemplified by plotting a 95% quantile directly onto the crystal structure (Fig. 8c,d,e), exhibit minor changes between the upper and lower quantiles, yet helix 3 and beta-sheets remain highlighted across the 95% confidence interval. Hence, uncertainty quantification strengthens our results and is a unique benefit of ReX.

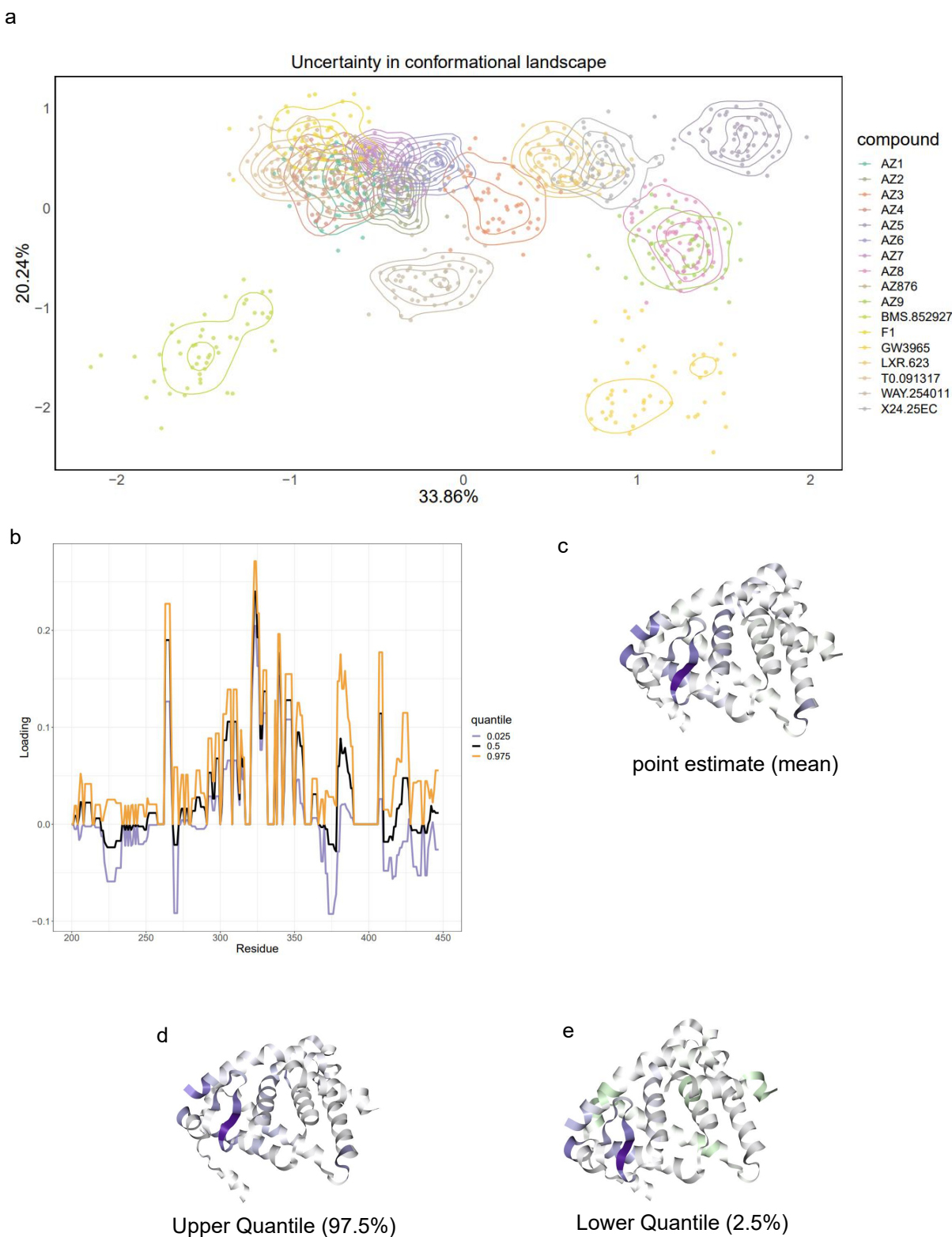


Figure 8: **Quantifying conformational uncertainty.** (A) A PCA (PC1 vs PC2) plot with uncertainty contours representing the posterior distribution of the locations of molecules in PCA coordinates. (B) The corresponding quantiles in the loadings of the principal components (C) The loading plotted directly on the protein structure corresponding to the mean estimate of the loadings. (D) As with (C) but the upper quantile of the loading (E) as with (D) but with the lower quantile of the loadings.

3 Discussion

HDX-MS offers a powerful approach to probe the conformational dynamics of proteins, yet the bottom-up approach to MS has inherent interpretative challenges stemming from peptide redundancy, length-bias, and contradictions. To address these challenges, we introduced ReX, a Bayesian non-parametric change-point model tailored to infer residue-level HDX patterns. Uniquely, ReX adapts to data quality and size, offers quantifiable measures of data resolution, and, importantly, quantifies uncertainty, distinguishing it from existing approaches. Its ability to correct length-bias, combined with a data-informed inference of smoothness along the residue dimension, sets ReX apart. As a flexible approach, ReX is able to model single proteins, differential experiments or large-scale comparisons via supervised or unsupervised conformational signature analysis.

Using a three-way differential digestion experiment, ReX was shown to be a highly accurate approach, whilst adhering to physical constraints of HD exchange. Furthermore, the analysis demonstrated the importance of smoothing in the residue dimension. Through a variety of case-studies, we demonstrated the insights that ReX obtained. For instance, considering BRD4 Bromodomains in the context of I-BET151, we determined the differential flexibility of the BC loop in the two Bromodomains. Further, CSA on seventeen molecules complexed with LXR α highlighted ReX's ability in linking conformational signatures with *in vivo* determined pharmacological properties. Hence, ReX offers a promising avenue to streamline further experimental investigations with additional confidence obtained from uncertainty quantification. Specifically, we found that the ligand BMS-852927 induced a distinct conformation of LXR α and the lack of stabilisation of helix 12, helix 3 and the beta-sheets could explain its therapeutic profile.

While ReX efficiently models centroided spectra, EX1 dynamics might necessitate a dual application of ReX, each tailored to a distinct population's centroid. Looking ahead, integrating ReX with alternative structural models, such as cryo-EM, and introducing 3D structure-derived constraints or correlations could further improve its performance.

4 Methods

4.1 Likelihood model

We first describe our proposed model as a Bayesian generative model (see (Crook *et al.*, 2022a) for a general description and mass-spectrometry examples). We let $y_i(t)$ be the observed deuterium incorporation process where i indexes peptides from $i = 1, \dots, N$. We assume there is an underlying residue level process that builds $y_i(t)$. More specifically, the observed peptide data are considered as noisy observations of some function of the unobserved residue processes. This unobserved (or latent) residue process is derived specified using the following functional form:

$$\mu_r = (1 - \pi_r)(1 - \exp(-b_r t^{p_r}) + \pi_r(1 - \exp(-d_r t))), \quad (1)$$

which defines the unobserved process for residue r , where residues are indexed $r = 1, \dots, R$. We remark that this functional form is more elaborate than previous approaches, which is to allow a very general approach. The mixture between two competing models is defined by mixing proportions $\pi_r \in (0, 1)$. These proportions are the relative preference for the two most popular models. The first component describes the stretched exponential of Weibull kinetics used by a number of studies and the second component uses the model compatible with protection factors as a single exponential per residue. The mixing components are learnt from the data and so concentrate on the preferred model or an interpolation between the two. We collect the parameters using $\theta(r) = (\pi_r, p_r, b_r, q_r)$, which is a function of residues.

The peptide process is derived from the residue process as the summation of residue-level processes, using a Laplace error model, and hence:

$$y_i \sim \mathcal{L} \left(\sum_{\mathcal{R}_i} \mu_r, l_i \sigma^2 \right), \quad (2)$$

$$\mathcal{R}_i := \{r : \text{start}_i + 2 \leq r < \text{end}_i, r \neq \text{Pro}\},$$

where l_i is the number of exchangeable amides for peptide i . This encodes a so-called length bias, allowing the longer peptides to be more variable (or contribute less to the overall fit). We assume that variance parameter arises from the same distribution for all residues. This assumption on σ^2 allows us to borrow power to better estimate the variance for each residue (and hence each peptide). For convenience, we write $\mu_i = \sum_{\mathcal{R}_i} \mu_r$.

4.2 Non-parametric prior model

It is pertinent to note that at the moment our model is over-specified, by which we mean that there are more parameters than there typically observed data. However, even if $N > R$, our model may still be difficult to identify. The model becomes identifiable when the redundancy for every residue is greater than one. This over-specification is necessary to complete the model, with the issue being alleviated by careful prior modelling and quantifying uncertainty.

We employ a non-parametric prior model, which is a class of flexible prior models which allow the number of parameters in the model to be inferred using the data. The prior model is the model of the likelihood parameters; that is, the model that describes $\theta(r)$. In this way if more data is

observed the model is able to adapt and introduce complexity. To reduce the effective dimension of the model we use a change-point model. Here, the parameters are specified as constant segments between points of change, where the values of the parameters, the locations of the change-points and the number of change-points are quantities to be inferred from the data. We note that whilst the prior model is hence piece-wise constant the posterior distribution (the distribution of the parameters having observed the data) is not necessarily in the same model family.

For a more precise mathematical description of the model, we specify these change-points using a *Poisson process prior* with rate $\lambda/(R - 1)$. The choice of rate ensure the mean number of change-points is λ (Kingman, 1992). From standard analysis of Poisson process, the number of change-points, K , is Poisson distributed $K \sim \text{Pois}(\lambda)$. We denote the vector of of change points as $\tau = (\tau_1, \dots, \tau_K)$, which are *a priori* uniformly distribution and ordered in $(1, R)$, by properties of the *Poisson process prior*.

There are $K + 1$ intervals to consider between each change-point and the endpoints. We let, $\theta_k = (\pi_k, p_k, b_k, q_k)$ be the parameters in the k^{th} interval, $k = 1, \dots, K + 1$ and let $\theta = (\theta_1, \dots, \theta_{K+1})$. Hence, we can write the prior function as

$$\theta(r) = \sum_{k=1}^{K+1} \theta_k \mathbb{I}(\tau_{k-1} \leq r < \tau_k), \quad (3)$$

from which it is obvious to see that the prior model is a piece-wise constant process. The prior model specification of the model is complete once we specify priors for the individual parameters for which we place the following hierarchical structure

$$\begin{aligned} d_k &\sim \mathcal{G}(d_{shape}, d_{rate}) \\ b_k &\sim \mathcal{G}(b_{shape}, b_{rate}) \\ \pi_k &\sim \mathcal{B}(\pi_{shape_1}, \pi_{shape_2}) \\ p_k &\sim \mathcal{B}(p_{shape_1}, p_{shape_2}) \\ \log \sigma &\sim \mathcal{N}(m, v^2) \\ d_{rate} &\sim \mathcal{G}(d_\alpha, d_\beta) \\ b_{rate} &\sim \mathcal{G}(b_\alpha, b_\beta). \end{aligned} \quad (4)$$

This model structure involves a hyper-prior structure which allows borrowing of information between the model parameters.

4.3 Derivation of posterior distribution

For clarity and to aid understanding the nuances of the model, we derive the posterior distribution of our model. By standard computations, using repeated application of Bayes' Theorem:

$$p(K, \tau, \theta, \sigma | y) \propto p(y | \tau, \theta, K, \sigma) p(\theta | \tau, K) p(\tau | K) p(K) p(\sigma) \quad (5)$$

First, we recall that

$$p(K) = \text{Pois}(K; \lambda), \quad (6)$$

and the change-points τ are uniform and ordered in $(1, R)$, so

$$p(\tau|K) = \frac{K!}{(R-1)^K} \quad (7)$$

as there are $K!$ permutations of uniforms that give the same τ . Given K , the τ state space $\Omega_{\tau,K}$ is

$$\Omega_{\tau,K} = \{\tau_1, \dots, \tau_k \in (1, R) : \tau_1 < \tau_2 < \dots < \tau_k\}. \quad (8)$$

The prior for θ is

$$\begin{aligned} p(\theta|\tau, K) &= p(\pi|\tau, K)p(b|\tau, K)p(d|\tau, K)p(p|\tau, K) \\ &= \prod_{k=1}^{K+1} p(\pi_k|\tau, K)p(b_k|\tau, K)p(d_k|\tau, K)p(p_k|\tau, K) \end{aligned} \quad (9)$$

with the state space $\Omega_{\theta,K+1} = ((0, 1) \times (0, \infty) \times (0, \infty) \times (0, 1))^{K+1}$. For the observation model, we recall that:

$$\begin{aligned} y_i &\sim \mathcal{L}\left(\sum_{\mathcal{R}_i} \mu_r, l_i \sigma^2\right), \\ \mu_r &= (1 - \pi_r)(1 - \exp(-b_r t^{p_r}) + \pi_r(1 - \exp(-d_r t))) \\ \mathcal{R}_i &:= \{r : \text{start}_i + 2 \leq r < \text{end}_i, r \neq \text{Pro}\}, \end{aligned} \quad (10)$$

Hence, we see that

$$p(y|\tau, \theta, N, \sigma)\pi(\sigma) = \prod_{i=1}^N \mathcal{L}\left(y_i; \sum_{\mathcal{R}_i} \mu_r, l_i \sigma^2\right) \mathcal{N}(\log \sigma; m, v^2). \quad (11)$$

The parameter space for fixed K is Ω_K , where

$$\Omega_K = \Omega_{\theta,K+1} \times \Omega_{\tau,K} \times \Omega_{\sigma}, \quad (12)$$

where $\Omega_{\sigma} = (0, \infty)$. The full parameter space Ω^* for the posterior $\pi(K, \tau, \theta, \sigma|y)$ is

$$\Omega^* = \bigcup_{K=0}^{\infty} \bigcup_{(\tau, \theta, \sigma) \in \Omega_K} (\tau, \theta, \sigma, K). \quad (13)$$

4.4 Derivation of reversible-jump Markov-chain Monte-Carlo sampler

If the number of parameters in the model was fixed, we could employ a Markov-chain Monte-Carlo algorithm to sample from the posterior distribution of the model (Brooks *et al.*, 2011). However, we are required to explore an infinite dimensional state-space or in other words our model is allowed to change dimension. That is the dimension of the model is to be inferred from the data which allows our model powerful flexibility. As a results, we require Monte-Carlo moves that allow moves between models of smaller and greater dimension. The number of parameters is essentially controlled by the number of change-points. That is the increasing the number of change-points increases the number of parameters in the model.

This is classically referred to as a class of statistical problems where "the number of things we do not know is something we do not know". To sample from the posterior distribution of our model, we need an MCMC algorithm that allows for a varying dimension model. One such an approach is referred to as reversible jump Markov-chain Monte-Carlo (RJMCMC), sometimes referred to as trans-dimensional MCMC. We briefly introduce RJMCMC in the general setting before specialising to our case and refer to the seminal work of [Green \(1995\)](#).

4.5 Preliminaries on RJMCMC

Suppose that we have a countable collection of models over which we wish to perform inference $\{\mathcal{M}_k, k \in \mathcal{K}\}$. Each model \mathcal{M}_k has a vector of parameters $\theta_k \in \Omega^{n_k}$, where the dimension n_k is allowed to vary for each model. Bayesian model choice for observed data y , can be written using the following hierarchical structure:

$$p(k, \theta_k, y) = p(k)(\theta_k|k)p(y|k, \theta_k). \quad (14)$$

For convenience, we set $x = (k, \theta_k)$. For given k , $x \in \{k\} \times \Omega^{n_k} =: \Omega_k$ and x varies over $\Omega^* = \bigcup_{\mathcal{K}} \Omega_k$. Now, we need to construct a Markov transition Kernel $P(x, dx')$ that is aperiodic, irreducible and satisfies detailed balance. In the case of transition between parameter subspace with different dimensions, some care is needed. Consider a Markov chain where the current state is x , and propose to make a move of type m that transitions to state dx' with probability $q_m(x, dx')$. Not all states with be accessible from x is a single move and so $q_m(x, \Omega^*) = 0$ for many m . Such a move is then accepted with the following probability

$$\alpha_m(x, x') = \min \left\{ 1, \frac{\pi(dx') q_m(x', dx)}{\pi(dx) q_m(x, dx')} \right\}. \quad (15)$$

However, it is still unclear how to construct moves between different models. Suppose $\rho_{k,k'}$ is the probability of proposing a move to model k' when the current model is k . Typically, $\rho_{k,k'} = 0$ if $|k - k'| > 1$ for simplicity, but still ensures all states are accessible and therefore the Markov-chain will be irreducible. The reversible jump MCMC algorithm can be stated as follows

- Suppose $X_t = (\theta_k, k)$ is the current state of the Markov-chain. State X_{t+1} is generated as follows:
- Set $k' = k + 1$ with probability $\rho_{k,k+1}$ and otherwise let $k' = k - 1$.
- Simulate $u \sim g_{k,k'}(\cdot)$, where $\dim(u) = k'$
- Set $(\theta_{k+1}, u') = \psi(\theta_k, u)$, where ψ is a differentiable bijection.
- If $k' = k + 1$, i.e. dimension increasing, we compute:

$$\alpha(\theta_{k+1}, k|\theta_k, k) = \min \left\{ 1, \frac{\pi(\theta_{k+1}, k+1|y)\rho_{k+1,k}}{\pi(\theta_k, k|y)\rho_{k,k+1}g_{k,k+1}(u)} \left| \frac{\partial\psi_{k,k+1}(\theta, u)}{\partial(\theta, u)} \right| \right\} \quad (16)$$

- If $k' = k - 1$, i.e. dimension decreasing, we compute:

$$\alpha(\theta_{k-1}, k-1|\theta_k, k) = \min \left\{ 1, \frac{\pi(\theta_{k-1}, k-1|y)\rho_{k-1,k}g_{k-1,k}(u')}{\pi(\theta_k, k|y)\rho_{k,k+1}} \left| \frac{\partial\psi_{k,k-1}(\theta, u)}{\partial\theta} \right| \right\} \quad (17)$$

- Sample $v \sim U[0, 1]$, and if $\alpha \geq v$, then set $X_{t+1} = (\theta_{k'}, k')$ and otherwise set $X_{t+1} = (\theta_k, k)$

It is often easier to compute the acceptance probability by considering the probability distribution of $\theta_{k'}, k' | \theta_k, k$, which we call Q . The acceptance probability is then

$$\alpha(\theta_{k'}, k' | \theta_k, k) = \min \left\{ 1, \frac{\pi(\theta_{k'}, k' | y) Q(\theta_k, k | \theta_{k'}, k')}{\pi(\theta_k, k | y) Q(\theta_{k'}, k' | \theta_k, k)} \right\}. \quad (18)$$

We can then perform a MCMC move between models of varying dimensions.

4.6 Contrusion of RJMCMC

Here, we explicitly construct the RJMCMC sampler for our *Poisson process prior* model. We proceed by first describing the moves and then computing the transition probabilities

4.6.1 Adding a dimension

We first consider adding a dimension with probability ρ . We then choose an interval to break/split by sampling $j \sim U\{1, 2, \dots, K+1\}$ and let the new change point be $\tilde{\tau} \sim U(\tau_{j-1}, \tau_j)$. We now have two new intervals for which we need to generate parameters and so we simulate $\tilde{\theta}_1, \tilde{\theta}_2 \sim p(\theta)$. We then set $k' = k+1$, $\tau' = (\tau_1, \dots, \tau_{j-1}, \tilde{\tau}, \tau_j, \dots, \tau_K)$ and $\theta' = (\theta_1, \dots, \theta_{j-1}, \tilde{\theta}_1, \tilde{\theta}_2, \theta_{j+1}, \theta_{K+1})$

4.6.2 Removing a dimension

We delete a dimension with probability $1 - \rho$. Choose a change-point to delete by sampling $j \sim U\{1, \dots, N\}$ and sample new parameters for the merged interval from $\tilde{\theta} \sim p(\theta)$. We then set $k' = k-1$, $\tau' = (\tau_1, \dots, \tau_{j-1}, \tau_{j+1}, \dots, \tau_K)$, $\theta' = (\theta_1, \dots, \theta_{j-1}, \tilde{\theta}, \theta_{j+2}, \theta_{K+1})$.

4.6.3 Transition probabilities

For adding dimensions, we compute the following transition probabilities, which follow by book-keeping and independence:

$$\begin{aligned} & Q(\theta', \tau', K' | \theta, \tau, K) \\ &= \{\text{Probability of new dimesnion}\} \times \{\text{Choose interval}\} \times \{\text{Choose } \tilde{\tau}\} \times \{\text{Choose } \tilde{\theta}_1, \tilde{\theta}_2\} \\ &= \rho \frac{1}{K+1} \frac{1}{\tau_j - \tau_{j-1}} p(\tilde{\theta}_1) p(\tilde{\theta}_2) \end{aligned} \quad (19)$$

and similarly for the reverse case we see that:

$$Q(\theta, \tau, K | \theta', \tau', K') = (1 - \rho) \frac{1}{K+1} p(\theta) \quad (20)$$

For removing dimensions, we compute the following transition probabilities

$$\begin{aligned} & Q(\theta', \tau', k' | \theta, \tau, k) \\ &= \{\text{Probability of removing a dimension}\} \times \{\text{Choose a change-point}\} \times \{\text{Choose } \tilde{\theta}\} \\ &= (1 - \rho) \frac{1}{k+1} p(\tilde{\theta}) \end{aligned} \quad (21)$$

For the reverse situation, we note that $K' = K$, and compute

$$Q(\theta, \tau, K | \theta', \tau', K') = \rho \frac{1}{K} \frac{1}{\tau_{j+1} - \tau_{j-1}} p(\tilde{\theta}_1) p(\tilde{\theta}_2) \quad (22)$$

Furthermore, we need to compute the posterior

$$\pi(\theta', K', \tau' | y) \propto p(y | \tau', \theta', K', \sigma) \pi(\theta' | \tau', K') \pi(\tau' | K') \pi(K') \quad (23)$$

and the formulas are given earlier.

4.7 MCMC moves for other parameters

Given the current iteration of the MCMC algorithm at model K (i.e. K change-points), we the parameters θ and σ are updated using a random-walk Metropolis-Hastings algorithm (Chib and Greenberg, 1995).

4.8 Other Methods

In this section we detail other possible methods for determining residue-level uptake data.

4.8.1 Generalised inversion approach

Let $U_{residue}$ be a $R \times 1$ vector of unknown residue level uptakes for a fixed exposure time. Let $U_{peptide}$ be the $N \times 1$ vector of peptide level uptakes for a fixed exposure time. Let C be a coupling matrix of dimension $N \times R$, which means all entries are 0 apart from the $(i, j)^{th}$ entry which is 1 if residue j has exchangeable amide for peptide i . Then it follows that

$$U_{peptide} = CU_{residue}. \quad (24)$$

In general C is rectangular is we can solve for unknown $U_{residue}$ by performing $U_{residue} = C^\dagger U_{peptide}$, where C^\dagger is the generalised inverse. The main limitations of this approach are that this inversion is unstable, there is no reason why the temporal component would be consistent, uptake values are not guaranteed to be 0 and 1 and no smoothing between uptake value between neighbouring residues.

4.8.2 Least-squares approach

Consider the notation of the following section, then $U_{residue}$ can be solved by considering the following minimisation problem

$$\min \|U_{peptide} - CU_{residue}\|_2, \quad (25)$$

constraining the entries of $U_{residue,j}$ to be in $[0, 1]$. This problem can solved by gradient-descent. This approach ensure the residue uptake values are between 0 and 1, however all the other limitation remain. Furthermore, the minimum in not necessarily unique or global.

4.8.3 Fused-LASSO approach

To introduce smoothness between residue; that is, constraining the uptake between neighbouring residues to be similar we can introduce a fusion penalty. Consider the following optimisation problem:

$$\min \|U_{peptide} - CU_{residue}\|_2 + \lambda |U_{residue,j} - U_{residue,j+1}|_1, \quad (26)$$

constraining the entries of $U_{residue,j}$ to be in $[0, 1]$. The extra penalty ensures that neighbouring residues are similar. This is the approach implemented by PyHDX (Smit *et al.*, 2021) and is similar to (Skinner *et al.*, 2019) and is part of a more general strategy called the fused LASSO. It is unclear how to select λ and because of double-dipping if λ is chosen using the data then computing p -value for difference after applying this approach leads to inflated false positives. This is referred to as the selective inference problem.

4.8.4 Residue-averaging approach

Here, we simply take the average deuterium uptake over peptides for the respective residues. Consider the matrix $U_{peptide}$ and normalise the columns by the number of exchangeable amides for the relevant peptide to construct $\tilde{U}_{peptide}$. The residue level uptake is then the row-average of the non-zero entries of $\tilde{U}_{peptide}$, which is a vector of length R . This approach suffers from the same issues as the other approaches. However, it also over smoothness because it cannot quickly adapt to sudden changes.

4.9 ReX methodology

4.9.1 Basic quantities

In this section, we detail a number of additional quantities available from analysis with our approach other than estimate of residue level uptakes and parameters. Firstly, for each observed peptide y_i , we can extract the posterior mean $\hat{\mu}_i$ from our model:

$$\hat{\mu}_i = \frac{1}{M} \sum_{m=1}^M \mu_{im}, \quad (27)$$

where $m = 1, \dots, M$ indexes over MCMC iterations. This is the prediction of that peptide from the model. Slight variations are of course possible and modelled by deviation from the mean using σ . We can then compute the reconstructive error (RE) using the following formula:

$$\text{RE}_i = y_i - \hat{\mu}_i, \quad (28)$$

which can be normalised by the l_i . To define a per residue error, we consider the collection of peptide N_r overlapping residue r . We then defined the (signed) average reconstructive error (ARE)

and the total reconstructive error (TRE) as following:

$$\begin{aligned}\text{Signed ARE}_r &= \frac{1}{n_r} \sum_{j \in N_r} \frac{y_j - \hat{\mu}_j}{l_j}, \\ \text{ARE}_r &= \frac{1}{n_r} \sum_{j \in N_r} \frac{|y_j - \hat{\mu}_j|}{l_j}, \\ \text{TRE}_r &= \sum_{j \in N_r} \frac{y_j - \hat{\mu}_j}{l_j},\end{aligned}\tag{29}$$

where n_r is the redundancy at residue r . We can see that any of these quantities would be large if the data are poorly modelled at residue r and these quantities give us an idea of which residue are contributing most to the error. TRE_r is particularly important in the context of HDX-MS because if several peptides overlapping a residue all with the same direction of error/difference, such a difference is then consistent. This is a natural way to exploit the redundancy of the data.

4.9.2 Probabilistic estimates

We note that σ is an estimate of the deviation at any residue r . Therefore, we may compute the following probabilities:

$$\begin{aligned}P(\text{Signed ARE}_r > 0) &\approx \frac{1}{m} \sum \mathbb{I}(\text{Signed ARE}_r > \tau^{(m)}), \tau^{(m)} \sim \mathcal{L}(0, \sigma^{(m)}) \\ P(\text{ARE}_r > 0) &\approx \frac{1}{m} \sum \mathbb{I}(\text{ARE}_r > |\tau^{(m)}|), \tau^{(m)} \sim \mathcal{L}(0, \sigma^{(m)}) \\ P(|\text{TRE}_r| > 0) &\approx \frac{1}{m} \sum \mathbb{I}(|\text{TRE}_r| > |\tau^{(m)}|), \tau^{(m)} \sim \sum_{N_r} \mathcal{L}(0, \sigma^{(m)})\end{aligned}\tag{30}$$

These values gives us an estimate of errors at a particular residue. In practice, we may wish to control the false discovery rate (FDR) but given that in a typical application we do not know the false positives, we have to estimate this quantity. We can estimate the expected false discovery rate (EFDR), in the following manner. Consider a set of ordered probabilities for a particular phenomena such as deuterium difference p_1, \dots, p_n . For a particular threshold τ , we estimate the EFDR as follows:

$$\text{EFDR}(\tau) = \frac{\sum_{i=1}^n (1 - p_i) \mathbb{I}(p_i \geq \tau)}{\sum_{i=1}^n \mathbb{I}(p_i \geq \tau)}\tag{31}$$

Typically, we wish to control the EFDR at a particular level, say 0.05, however it is not clear which threshold correspond to this value. To do this, we simply take the smallest p_i such that the following holds: $\text{EFDR}(p) < 0.05$ for $p > p_i$ and $\text{EFDR}(p) > 0.05$ for $p < p_i$. It is easy to find this value by trying sequentially smaller values of the observed p_i .

4.9.3 Differential Analysis

HDX-MS is frequently applied in the differential setting where by two proteins states are compared. Examples include epitope mapping using an antigen with and without antibody or proteins with ligands, modifications, mutations etc. To perform differential analysis, we fit our model on the

differential state being considered, for example the antigen-antibody complex data. We then using a reconstructive approach by computing the difference between the peptides observed in the APO data but using the posterior quantities from the state of interest. Large and consistent differences indicate confident deuterium difference between the two states. By using TRE_r and the associated probabilities of a difference, we can quantify the residues which are different between the two conditions. The comparison need not be the APO data but any protein state of interest. It is important to note that using this approach explicitly makes use of the redundancy of the data. Consistent differences in the same direction on an overlapping group of peptides will be assigned greater confidence because the quantities are summed rather than averaged.

4.9.4 Conformational Signature Analysis

At a high-level, conformational signature analysis (CSA) is for applications where pair-wise comparison between all protein states is unfeasible. Typically, a large-panel of small molecules or antibodies may be compared using HDX-MS. Several goals can be achieved by employing CSA. Firstly, we wish to quantify the conformational landscape; that is, what are the modes of variation that the perturbations to the protein induce. In what ways are perturbations similar in the variation they induce to the protein. Which residues are the largest contributors to this variation? If the perturbations have associated functional annotations (such as the molecular property of a ligand) can these be discriminated and can we quantify the residue associations? Given the complexity of these analysis, quantifying uncertainty is of importance to avoid spurious associations.

ReX provides functionality to answer these questions. We first employ ReX to project peptide-level to residue-level data, using TRE_r for all states $c = 1, \dots, C$ under consideration. These data are then combined into a single data matrix of conditions by residues X . That is the $(c, r)^{th}$ entry of X is the TRE_r for state c . For unsupervised analysis (data without external annotations), we employ principal components analysis (PCA). This projects the matrix X using an orthogonal linear transform into a new space. The directions taken for each subsequent component explain the greatest variation of the data whilst remaining orthogonal. This projection can be written as following

$$S = XL, \quad (32)$$

where S is the $C \times R$ score matrix and L is the $R \times R$ loading matrix. Typically, the data is visualised using a 2D scatter plot using the coordinates described by S . The first column of S describes the first principal component (PC) and so on. Dimensionality reduction is typically achieved by only keeping a small number of the columns of S , which explain a large amount of the variation in the data. L is the $R \times R$ loading matrix which is the projection matrix in this case. Taking columns of L and scaling by the variances, called loadings, define the important contributing variables to the variation quantified by the PCs. In this case the loadings are associated with residues and the magnitude (whether positive or negative) indicate the importance of that residue. The loadings for each PC can be visualised directly on the protein structure to aid interpretation. Note that a negative or positive loading does not necessarily indicate protection or de-protection

If each condition is associated with a external annotation such as a molecular property of a compound, we can add an annotation vector y . In this case, we can employ a discriminate analysis (DA) method. Similarly to PCA, we find the directions of the data that most discriminate the

molecular properties using partial least squares discriminate analysis (PLS-DA) or the orthogonal extensions (OPLS-DA) (Thevenot, 2016). In the latter, the dimensions are separated into predictive and orthogonal dimensions. The annotation vector can be categorical (antagonist vs agonist) or continuous (EC_{50}). As before, the loadings can be projected onto the protein structure to interpret the results.

To perform uncertainty quantification, we notice that we began our analysis with using the posterior means from the ReX output. Instead, we can repeat the aforementioned analysis using samples from the posterior distribution. In this way, we build up a distribution of locations for each projection of our state and a distribution of loadings. We align the different projections to the projection from the posterior means using the Procrustes transform.

4.10 RexMS R-package

The software package is implemented in R and can be found at the following url: <https://github.com/ococrook/RexMS>. Documentation is found within the package and additional tutorials are found within the package website <https://ococrook.github.io/RexMS/>. The tutorial cover single protein analysis <https://ococrook.github.io/RexMS/articles/ReX.html>, differential analysis <https://ococrook.github.io/RexMS/articles/DifferentialRexMS.html>, and conformational signature analysis <https://ococrook.github.io/RexMS/articles/ConformationalSignatureAnalysis.html>. Conformational signature analysis is implemented in both unsupervised and supervised mode, in addition the supervised mode allows both categorical (e.g. a functional label) and continuous (e.g. binding affinity). There can be missing values in the labels too if you wish to infer outcome given the conformational signature.

4.11 Benchmark experiment of HDX-MS of Cytochrome C

Horse heart cytochrome c (Sigma C7752) was prepared at a stock concentration of 80 μ M in an equilibrium buffer of 50 mM MOPS, 150 mM NaCl, pH7.2. For peptide mapping experiments, 6 μ L of 6.67 μ M cytochrome c (40 pmol) was diluted in 54 μ L equilibrium buffer. For labelling experiments 6 μ L of 3.33 μ M (20 pmol) cytochrome c in 54 μ L labelling buffer (50 mM MOPS, 150 mM NaCl, pD7.2 (pHread = 6.8)). The samples were labelled at 15, 150, 1500 and 15000 s at 20°C, performed in triplicate for experiments with nepenthesin-2 digestion, and in singlicate for aspergillopepsin and pepsin digestions. The reaction was quenched in an equal volume of quench buffer (400 mM sodium phosphate, 6 M guanidine hydrochloride pH 2.2, 2% formic acid) for 1 min at 0.5 *degree* C and injected into a 100 μ L sample loop and digested on one of three protease columns: nepenthesin-2 (2.1 mm \times 20 mm, Affipro), aspergillopepsin (2.1 x 20 mm, AffiPro) or pepsin (2.1 x 30 mm, Waters Enzymate BEH).

For 100% D labelling experiments (fully deuterated control), the reaction was performed offline, where 6 μ L of 6.67 μ M cytochrome c (40 pmol) was diluted offline in 54 μ L of 6 M d5-guanidinium deuteriochloride prepared in MOPS labelling buffer (50 mM MOPS, 150 mM NaCl, pD7.2 (pHread = 6.8)) and labelled for 15 min at 20 ° C. The reaction was then quenched by manual addition of an equal volume of 400 mM sodium phosphate, pH2.2, 2% formic acid for 1 min at 0.5° C prior to sample injection.

The resultant peptides were collected on a pre-column trap (UPLC BEH C18 Vanguard, Waters) for 4 min with 0.2% formic acid and 0.03% TFA in H₂O at a flow rate of 100 μ L/min, and separated by liquid chromatography (1.7 μ m UPLC BEH C18 column, 1.0 \times 50 mm dimensions, 130 pore size, Waters) for 12 min at a flow rate of 20 μ L/min at 0.5 ° C, over a gradient of 11–40% of 0.2% formic acid in MeCN before ramping to 98% for a further 3 min. A Leucine Enkephalin solution was used as lock mass.

Data was acquired on a Waters Cyclic IMS series mass spectrometer, operated in V-mode (resolution mode) using single-pass acquisition in the data-independent HD-MS acquisition mode for labelling experiments, and HD-MSE acquisition mode for peptide mapping experiments.

Peptide mapping experiment data was processed in ProteinLynx Global Server v3.0.2. Identified peptides were first filtered using a precursor maximum error of \pm 7 ppm, a PLGS score of 7, and sequence length of 5-20 amino acids. A second LARS-type filtering system was then employed to further triage peptides, where a minimum LARS score of 5 was set based on seven soft filtering thresholds: a minimum intensity 10,000, minimum products of 6, minimum products per amino acid of 0.5, minimum consecutive products of 1, minimum sum intensity of identified products of 2000, minimum peptide score of 8, as well as an identification in \geq 2 peptide mapping experiments performed. HDX-MS labelling experiments were processed and manually curated in HDEaminer (v3.3) against the filtered peptide list to determine deuterium uptakes.

To correct between uptake difference between digest experiments, even after normalisation by a full deuterated control, a correction factor was obtained by taking the peptides that were common in both experiments and computing the ratio of the medians of all these common peptides per time point. Deuterium uptakes for all peptides were then multiplicatively adjusted by this correction factor per time point to put them on the same scale.

References

- Althaus, E. et al. (2010). Computing h/d-exchange rates of single residues from data of proteolytic fragments. *BMC bioinformatics*, **11**, 1–12.
- Babić, D. et al. (2017). Localization improvement of deuterium uptake in hydrogen/deuterium exchange in proteins. *Journal of Chemometrics*, **31**(3), e2876.
- Babić, D. et al. (2019). Resolution of protein hydrogen/deuterium exchange by fitting amide exchange probabilities to the peptide isotopic envelopes. *Rapid Communications in Mass Spectrometry*, **33**(15), 1248–1257.
- Belorusova, A. Y. et al. (2019). Structural analysis identifies an escape route from the adverse lipogenic effects of liver x receptor ligands. *Communications biology*, **2**(1), 431.
- Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.
- Betancourt, M. (2021). Towards a principled bayesian workflow. https://github.com/betanalpha/knitr_case_studies/tree/master/principled_bayesian_workflow.

- Brooks, S. et al. (2011). *Handbook of markov chain monte carlo*. CRC press.
- Chalmers, M. J. et al. (2011). Differential hydrogen/deuterium exchange mass spectrometry analysis of protein–ligand interactions. *Expert review of proteomics*, **8**(1), 43–59.
- Chib, S. et al. (1995). Understanding the metropolis-hastings algorithm. *The american statistician*, **49**(4), 327–335.
- Chung, C.-w. et al. (2012). Bromodomains: a new target class for small molecule drug discovery. *Drug Discovery Today: Therapeutic Strategies*, **9**(2-3), e111–e120.
- Chung, C.-w. et al. (2011). Discovery and characterization of small molecule inhibitors of the bet family bromodomains. *Journal of medicinal chemistry*, **54**(11), 3827–3838.
- Claesen, J. et al. (2018). 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.
- Consortium, U. (2015). Uniprot: a hub for protein information. *Nucleic acids research*, **43**(D1), D204–D212.
- Crook, O. M. et al. (2022a). Challenges and opportunities for bayesian statistics in proteomics. *Journal of Proteome Research*, **21**(4), 849–864.
- Crook, O. M. et al. (2022b). Empirical bayes functional models for hydrogen deuterium exchange mass spectrometry. *Communications biology*, **5**(1), 1–10.
- Crook, O. M. et al. (2022c). Semi-supervised non-parametric bayesian modelling of spatial proteomics. *The annals of applied statistics*, **16**(4).
- Dawson, M. A. et al. (2011). Inhibition of bet recruitment to chromatin as an effective treatment for mll-fusion leukaemia. *Nature*, **478**(7370), 529–533.
- Eddy, S. R. (2004). What is a hidden markov model? *Nature biotechnology*, **22**(10), 1315–1316.
- Engen, J. R. (2009). Analysis of protein conformation and dynamics by hydrogen/deuterium exchange ms.
- Engen, J. R. et al. (2020). Complementarity of hydrogen/deuterium exchange mass spectrometry and cryo-electron microscopy. *Trends in biochemical sciences*, **45**(10), 906–918.
- Fearnhead, P. (2006). Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and computing*, **16**, 203–213.
- Filippakopoulos, P. et al. (2010). Selective inhibition of bet bromodomains. *Nature*, **468**(7327), 1067–1073.
- Gessner, C. et al. (2017). Computational method allowing hydrogen-deuterium exchange mass spectrometry at single amide resolution. *Scientific Reports*, **7**(1), 3789.

- Gilan, O. et al. (2020). Selective targeting of bd1 and bd2 of the bet proteins in cancer and immunoinflammation. *Science*, **368**(6489), 387–394.
- Gorman, J. et al. (2016). Structures of hiv-1 env v1v2 with broadly neutralizing antibodies reveal commonalities that enable vaccine design. *Nature structural & molecular biology*, **23**(1), 81–90.
- Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, **82**(4), 711–732.
- Hamuro, Y. et al. (2018). High-resolution hdx-ms of cytochrome c using pepsin/fungal protease type xiii mixed bed column. *Journal of The American Society for Mass Spectrometry*, **30**(2), 227–234.
- Holmes, S. et al. (2018). *Modern statistics for modern biology*. Cambridge University Press.
- Hong, C. et al. (2014). Liver x receptors in lipid metabolism: opportunities for drug discovery. *Nature reviews Drug discovery*, **13**(6), 433–444.
- Houde, D. et al. (2011). The utility of hydrogen/deuterium exchange mass spectrometry in biopharmaceutical comparability studies. *Journal of pharmaceutical sciences*, **100**(6), 2071–2086.
- James, E. I. et al. (2021). Advances in hydrogen/deuterium exchange mass spectrometry and the pursuit of challenging biological systems. *Chemical Reviews*.
- Janowski, B. A. et al. (1996). An oxysterol signalling pathway mediated by the nuclear receptor lxr α . *Nature*, **383**(6602), 728–731.
- Jia, R. et al. (2020). Hydrogen-deuterium exchange mass spectrometry captures distinct dynamics upon substrate and inhibitor binding to a transporter. *Nature communications*, **11**(1), 1–10.
- Jumper, J. et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, **596**(7873), 583–589.
- Kammari, R. et al. (2020). Effects of secondary structure on solid-state hydrogen–deuterium exchange in model alpha-helix and beta-sheet peptides. *Molecular Pharmaceutics*, **17**(9), 3501–3512.
- Kan, Z.-Y. et al. (2013). Protein hydrogen exchange at residue resolution by proteolytic fragmentation mass spectrometry analysis. *Proceedings of the National Academy of Sciences*, **110**(41), 16438–16443.
- Kan, Z.-y. et al. (2019). Exms2: An integrated solution for hydrogen–deuterium exchange mass spectrometry data analysis. *Analytical chemistry*, **91**(11), 7474–7481.
- Katta, V. et al. (1993). Hydrogen/deuterium exchange electrospray ionization mass spectrometry: a method for probing protein conformational changes in solution. *Journal of the American Chemical Society*, **115**(14), 6317–6321.

- Katz, A. et al. (2009). Safety, pharmacokinetics, and pharmacodynamics of single doses of lxr-623, a novel liver x-receptor agonist, in healthy participants. *The Journal of Clinical Pharmacology*, **49**(6), 643–649.
- Keppel, T. R. et al. (2014). 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), 547–554.
- Kick, E. K. et al. (2016). Discovery of highly potent liver x receptor β agonists. *ACS medicinal chemistry letters*, **7**(12), 1207–1212.
- Kielkopf, C. S. et al. (2019). Hdx-ms reveals orthosteric and allosteric changes in apolipoprotein-d structural dynamics upon binding of progesterone. *Protein Science*, **28**(2), 365–374.
- Kingman, J. F. C. (1992). *Poisson processes*, volume 3. Clarendon Press.
- Konermann, L. et al. (2011). Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chemical Society Reviews*, **40**(3), 1224–1234.
- Levin, N. et al. (2005). Macrophage liver x receptor is required for antiatherogenic activity of lxr agonists. *Arteriosclerosis, thrombosis, and vascular biology*, **25**(1), 135–142.
- Liu, T. et al. (2011). Quantitative assessment of protein structural models by comparison of h/d exchange ms data with exchange behavior accurately predicted by dxcorex. *Journal of the American Society for Mass Spectrometry*, **23**(1), 43–56.
- Masson, G. R. et al. (2017). An overview of hydrogen deuterium exchange mass spectrometry (hdx-ms) in drug discovery. *Expert opinion on drug discovery*, **12**(10), 981–994.
- Mitternacht, S. (2016). Freesasa: An open source c library for solvent accessible surface area calculations. *F1000Research*, **5**.
- Mujtaba, S. et al. (2007). Structure and acetyl-lysine recognition of the bromodomain. *Oncogene*, **26**(37), 5521–5527.
- Nguyen, L. H. et al. (2019). Ten quick tips for effective dimensionality reduction. *PLoS computational biology*, **15**(6), e1006907.
- Orengo, C. A. et al. (1999). From protein structure to function. *Current opinion in structural biology*, **9**(3), 374–382.
- Pascal, B. D. et al. (2012). Hdx workbench: software for the analysis of h/d exchange ms data. *Journal of the American Society for Mass Spectrometry*, **23**(9), 1512–1521.
- Petretich, M. et al. (2020). Domain-selective targeting of bet proteins in cancer and immunological diseases. *Current Opinion in Chemical Biology*, **57**, 184–193.
- Salmas, R. E. et al. (2021). Hdxmodeller: an online webserver for high-resolution hdx-ms with auto-validation. *Communications Biology*, **4**(1), 199.

- Saltzberg, D. J. et al. (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.
- Sauve, V. et al. (2018). Mechanism of parkin activation by phosphorylation. *Nature structural & molecular biology*, **25**(7), 623–630.
- Schultz, J. R. et al. (2000). Role of lxr α in control of lipogenesis. *Genes & development*, **14**(22), 2831–2838.
- Skinner, S. P. et al. (2019). Estimating constraints for protection factors from hdx-ms data. *Biophysical Journal*, **116**(7), 1194–1203.
- Smit, J. H. et al. (2021). Probing universal protein dynamics using hydrogen–deuterium exchange mass spectrometry-derived residue-level gibbs free energy. *Analytical Chemistry*, **93**(38), 12840–12847.
- Stofella, M. et al. (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), 813–822.
- Thevenot, E. A. (2016). ropls: Pca, pls (-da) and opl s (-da) for multivariate analysis and feature selection of omics data. *R package version*, **1**(0).
- Tsai, Y.-C. I. et al. (2020). Single-domain antibodies as crystallization chaperones to enable structure-based inhibitor development for rbr e3 ubiquitin ligases. *Cell chemical biology*, **27**(1), 83–93.
- Tsutsumi, M. et al. (2011). Parallel and antiparallel β -strands differ in amino acid composition and availability of short constituent sequences. *Journal of chemical information and modeling*, **51**(6), 1457–1464.
- Venkateswaran, A. et al. (2000). Control of cellular cholesterol efflux by the nuclear oxysterol receptor lxr α . *Proceedings of the National Academy of Sciences*, **97**(22), 12097–12102.
- Wagner, B. L. et al. (2003). Promoter-specific roles for liver x receptor/corepressor complexes in the regulation of abca1 and srebp1 gene expression. *Molecular and cellular biology*, **23**(16), 5780–5789.
- Wayment-Steele, H. K. et al. (2022). Prediction of multiple conformational states by combining sequence clustering with alphafold2. *BioRxiv*, pages 2022–10.
- Wei, H. et al. (2014). Hydrogen/deuterium exchange mass spectrometry for probing higher order structure of protein therapeutics: methodology and applications. *Drug discovery today*, **19**(1), 95–102.
- Willy, P. J. et al. (1995). Lxr, a nuclear receptor that defines a distinct retinoid response pathway. *Genes & development*, **9**(9), 1033–1045.

Zhang, Z. (2020). Complete extraction of protein dynamics information in hydrogen/deuterium exchange mass spectrometry data. *Analytical chemistry*, **92**(9), 6486–6494.

Zhang, Z. et al. (2012). Improved protein hydrogen/deuterium exchange mass spectrometry platform with fully automated data processing. *Analytical chemistry*, **84**(11), 4942–4949.

5 acknowledgements

We thank Geoff Nichols for useful suggestions on the modelling. The GPT-4 API was used for language suggestions.