Development of hydrogen/deuterium exchange mass spectrometry simulations to enable accurate protein structure and complex selection

Harris, Matthew

Awarding institution:
King's College London

Development of hydrogen/deuterium exchange mass spectrometry simulations to enable accurate protein structure and complex selection

by Matthew J. Harris

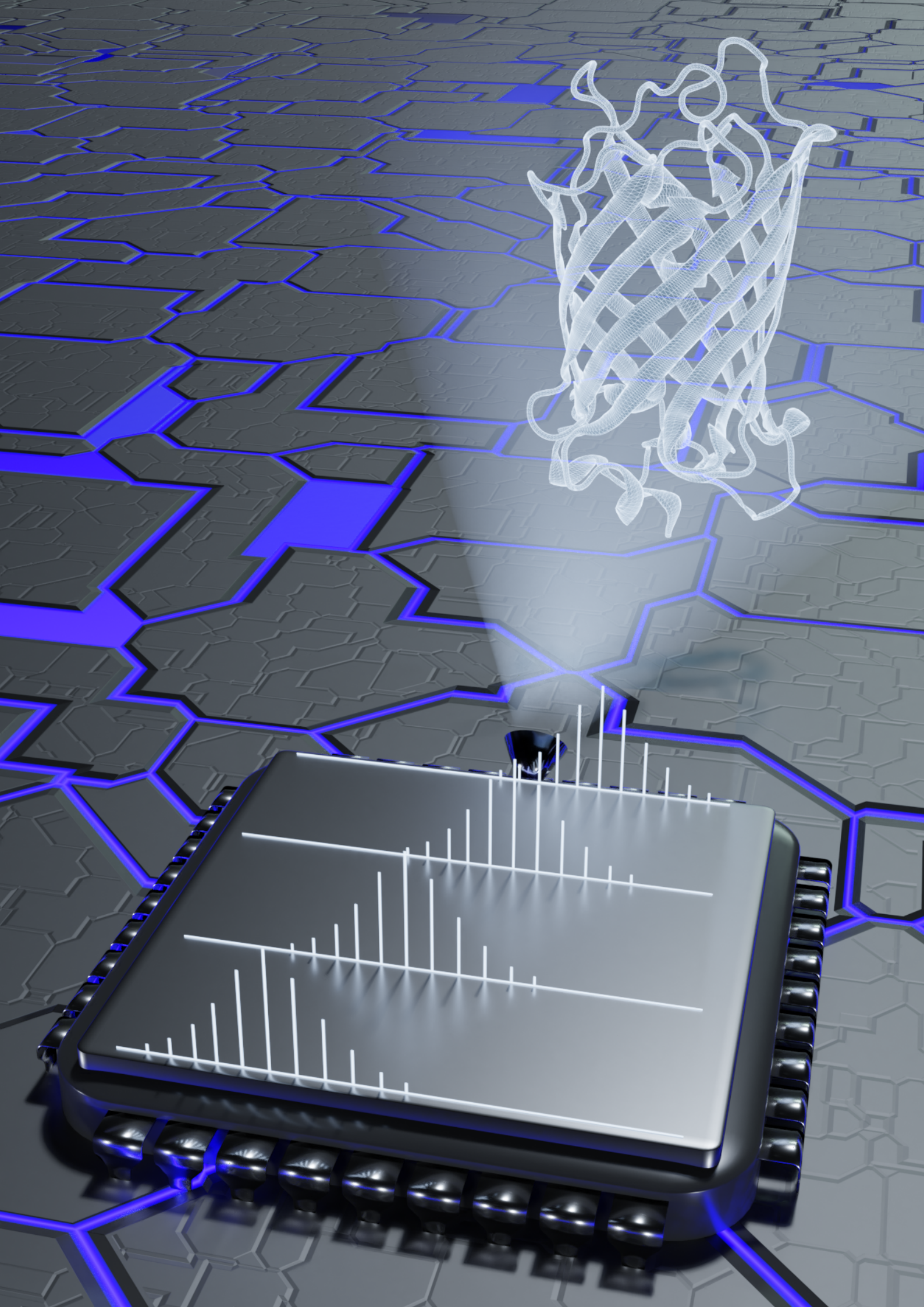
King's College London

Department of Chemistry

School of Natural and Mathematical Sciences

A thesis submitted in partial fulfilment of the requirements for the degree of 'Doctor of Philosophy' in Biochemistry

August 2020



Abstract

Hydrogen deuterium exchange coupled to mass spectrometry (HDX-MS) is a powerful and sensitive technique for the analysis of protein dynamics in solution. HDX-MS reports on the transient uptake of deuterium by a protein in bulk deuterated solvent due to the phenomenon of hydrogen exchange, whereby labile hydrogen atoms can spontaneously exchange for other hydrogens or deuterium in solution. HDX-MS is typically used to inform about the location of binding interfaces between proteins and other proteins/small molecules by comparing exchange profiles of a protein in its bound and unbound states, where regions of comparatively reduced deuterium uptake are indicative of a binding surface. HDX-MS is however not typically used to inform on the structure of proteins by itself due to a lack of understanding about the factors that cause changes in exchange rate.

In this thesis, we set out to develop a method that could accurately discriminate between native and non-native protein structures using nothing but the protein's experimental HDX-MS profiles. This was achieved by coupling state of the art mass spectrometry to computational chemistry techniques in order to develop a method that could accurately calculate HDX-MS profiles from *in silico* three-dimensional structures and compare these calculated profiles to experimental data in order to classify the structures as being either native or non-native. We achieved reasonable classification accuracy using this method over the course of this thesis with the potential for much greater accuracy with subsequent research and development. We also took the first steps towards modifying this methodology to work on classifying binary complex structures as well as individual protein structures.

In addition to our primary focus on structure classification, we also undertook a more traditional HDX-MS side project involving the determination of the location of the binding interface between the enzyme dUTPase and its inhibitor Stl. We successfully characterised this interaction and helped develop a model of the mechanism of inhibition based on our data.

The work presented in this thesis is extensive in its breadth and variety, incorporating a diverse range of different techniques spanning multiple scientific disciplines. From classical biochemical approaches such as the manipulation of DNA, cell culture and the production of proteins to analytical chemistry in the form of HDX and native mass spectrometry and computational chemistry and computer science techniques such as Molecular Dynamics, protein docking and Python programming.

Acknowledgements

Matthew J. Harris would like to thank Antoni Borysik for his supervision and guidance throughout this project. Additional thanks go to Rivka Isaacson for her supervision, Ramin Salmas for his help with the computational chemistry portions of this thesis, Rebecca Beavil and Ruth Rose for the production of various proteins and Teikichi Ikura for the supply of the initial barnase and barstar plasmids and protocols. Thanks also go to Heather Findlay and the Booth group for assistance with protein production and everyone in the Borysik and Politis groups, past and present, for their advice and support throughout this project. Funding was provided by the BBSRC and the Waters Corporation.

Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

φ11DUT φ11 dUTPase

φNM1DUT φNM1 dUTPase

AIRs Ambiguous Interaction Restraints

AUC Area Under Curve

BEX Back Exchange Control

Bn Barnase

Bs Barstar

CCS Collisional Cross Section

dTTP deoxythymidine triphosphate

dUMP deoxyuridine monophosphate

dUTP deoxyuridine triphosphate

dUTPase deoxyuridine triphosphate nucleotidohydrolase

FPR False Positive Rate

GFP Green Fluorescent Protein

GFP-nb Green Fluorescent Protein nanobody

GFP-nbmin Green Fluorescent Protein nanobody minimizer

hDUT human dUTPase

HDX-MS Hydrogen/Deuterium Exchange Mass Spectrometry

HPC High Performance Computing

IEX In Exchange Control

IMS Ion Mobility Spectrometry

LC Liquid Chromatography

m/z mass to charge ratio

MD Molecular Dynamics

PDB Protein Data Bank

PPI Protein-Protein Interaction

RFU Relative Fractional Uptake

RMSD Root Mean Square Deviation

RMSE Root Mean Square Error

ROC Receiver Operating Characteristic

SaPI *Staphylococcus aureus* pathogenicity islands

SEC Size Exclusion Chromatography

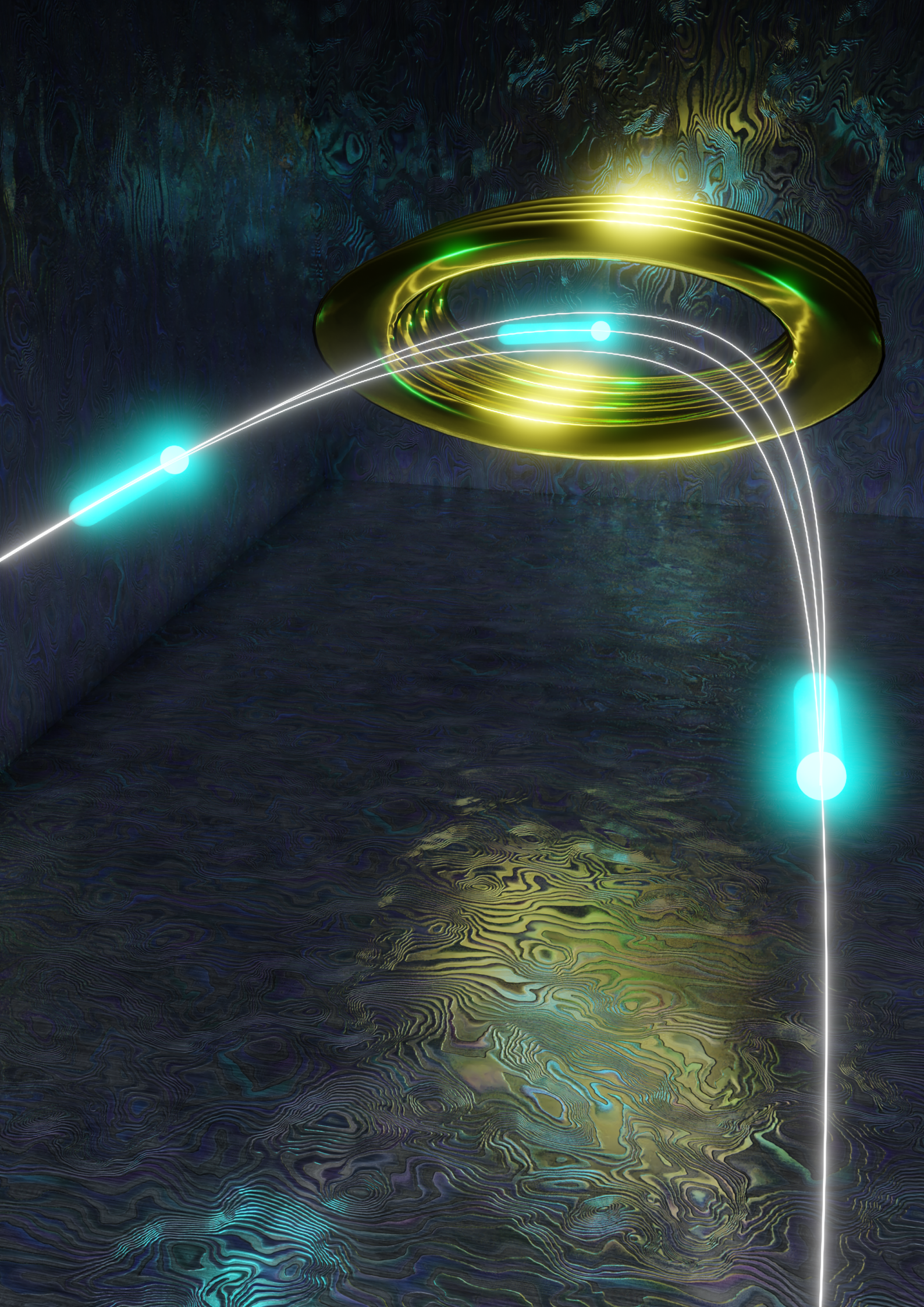
SEC-SAXS Size-Exclusion Chromatography in line with Small-Angle X-ray Scattering

SEM Standard Error of the Mean

Stl *Staphylococcus* pathogenicity island repressor protein

ToF Time of Flight

TPR True Positive Rate



1 Introduction

1.1 Statement of the problem and the need for this study

Since the discovery of the first protein structure, that of myoglobin in 1958 by John Kendrew [1], scientists have understood that knowledge of three-dimensional fold is of vital importance to the understanding of the function and behaviour of proteins. Indeed it is not an understatement to assert that the three-dimensional structure is the single most important piece of information a researcher can have about their particular protein of choice. This importance derives from the fact that a protein's function is almost entirely dependent on the three-dimensional fold and therefore investigating researchers will want to obtain knowledge of the structure if at all possible. Unfortunately, due to the large amount of time and effort required to solve structures experimentally, such as gold-standard techniques like x-ray crystallography, NMR and, increasingly, cryo-electron microscopy, the vast majority of discovered proteins do not have structures associated with them [2]. This problem is exacerbated even further when one considers the realm of protein-protein complexes, of which there are an estimated 500,000 binary complexes in humans alone [3], where even if structures of the individual component proteins have been solved, the likelihood is that the complex as a whole has not been. This wholesale lack of structural data is due to the inverse relationship between throughput and resolution, namely that higher resolution methods such as x-ray crystallography have lower throughput but relatively higher throughput methods such as cryo-electron microscopy have lower resolution. Therefore, the development of high resolution, high throughout methods for determining protein structures are of great interest to the scientific community.

Hydrogen Deuterium Exchange monitored by Mass Spectrometry (HDX-MS) is one such technique that has the potential to fill this desirable niche thanks to recent advances in methodology that demonstrate the ability of HDX-MS profiles to be calculated from three-dimensional structures [4]. In this thesis, we present a body of work that builds upon this new method in order to enable native monomeric structures to be selected from a background of decoys. Work was also undertaken to adapt the method for use with binary protein-protein interactions (PPI). Techniques employed in the development of this novel approach fall into three broad categories: Molecular Biology, involving cell culture, protein production and purification; Analytical Chemistry, involving HDX and native mass spectrometry; and Computational Chemistry, involving programming, molecular dynamics simulations and protein-protein docking.

1.2 Overview of Mass Spectrometry

Mass spectrometry is a powerful analytical technique involving the deflection of charged ions by a magnetic field. Identification of the first principles which would go to become mass spectrometry began in 1886 by Eugen Goldstein [5] with his discovery of positive ion beams he termed "canal rays". Subsequent experimentation with these rays resulted in the development of a much more recognisable form of mass spectrometry by J. J. Thomson in 1912 [6] in which isotopes of neon with different mass to charge ratios (m/z) were separated from each other and correctly identified. In all types of mass spectrometry, ions are generated by some type of ion source, whereupon they enter the mass spectrometer under a vacuum. In classical sector mass spectrometry [7–10], the ions then encounter an orthogonal

magnetic field which causes the path of the ion to bend, with the degree of bending being proportional to the ion's m/z. Those ions with higher ratios are influenced by the magnetic field less and so bend a lesser degree compared to ions with lower ratios. The ions subsequently collide with a detector, with higher m/z ions making contact at one end and lower m/z ions making contact at the other. In this way, the detector can ascertain the m/z's of all the ions in the original sample and present the data as a function of m/z and signal intensity in counts per second.

Instead of the position of impact on the detector being employed to determine the ion's m/z, the mass spectrometer utilised in the experiments presented in this thesis uses the time between when the ions enter the mass analyser and when they strike the detector at a known distance. This is known as Time of Flight (ToF) MS [11]. This separation according to time instead of distance is achieved by trapping packets of ions and accelerating them at the same time in a direction that is orthogonal to their original trajectory, after which they are pushed into a reflectron which corrects minor fluctuations in kinetic energy and reverses their direction of travel [12], setting the ions on a collision path with the detector. Ions with a lower m/z are imparted greater initial velocity by the pusher and so penetrate further into the reflectron, increasing the amount of time they take to reach the detector and vice versa. In this way, an ion's m/z can be correlated to its time of flight. The differences between sector and ToF mass spectrometers are illustrated in Figure 1.1.

Our mass spectrometer utilises the Electrospray Ionisation (ESI) [13–15] technique in order to generate ions. In this method, a high voltage is applied to the liquid sample in order to create an ionized aerosol. This aerosol is then dispersed as a fine spray of charged droplets, whereupon solvent evaporation and subsequent ejection of the contained ions occurs [16]. ESI is considered to be a "soft" ionisation technique because relatively little energy is transferred to the analyte and so limited fragmentation occurs. This makes ESI an ideal technique for studying proteins in their native state as liquid-phase information is retained in the gas-phase. Additionally, ESI's propensity to produce multiply charged ions enables even very large proteins to be detected by relatively low-range mass analysers [17].

1.3 Hydrogen Deuterium Exchange Mass Spectrometry

1.3.1 Theory of HDX-MS

HDX-MS is a technique first developed in 1991 by Viswanatham Katta & Brian T. Chait [18] which leverages the phenomenon of hydrogen exchange, whereby labile hydrogens covalently bound to amino acids can spontaneously exchange out with other hydrogens in the surrounding solution. In addition, exchange can also occur between hydrogen and its isotopes: deuterium (^{2}H or D) [19] and even tritium (^{3}H or T) [20], which have molecular weights approximately 2x and 3x greater than that of hydrogen respectively. The increased molecular weights of deuterium and tritium means their presence (or absence) within a protein can be measured by a mass spectrometer of sufficient resolution, allowing these isotopes to be utilised as transient labels. Tritium is radioactive and so rarely used, hence making deuterium the most common isotope employed in modern exchange experiments. These typically involve the dissolution of a protein in an excess of deuterated buffer and the measurement of deuterium uptake by MS over a variety of time points (Figure 1.2).

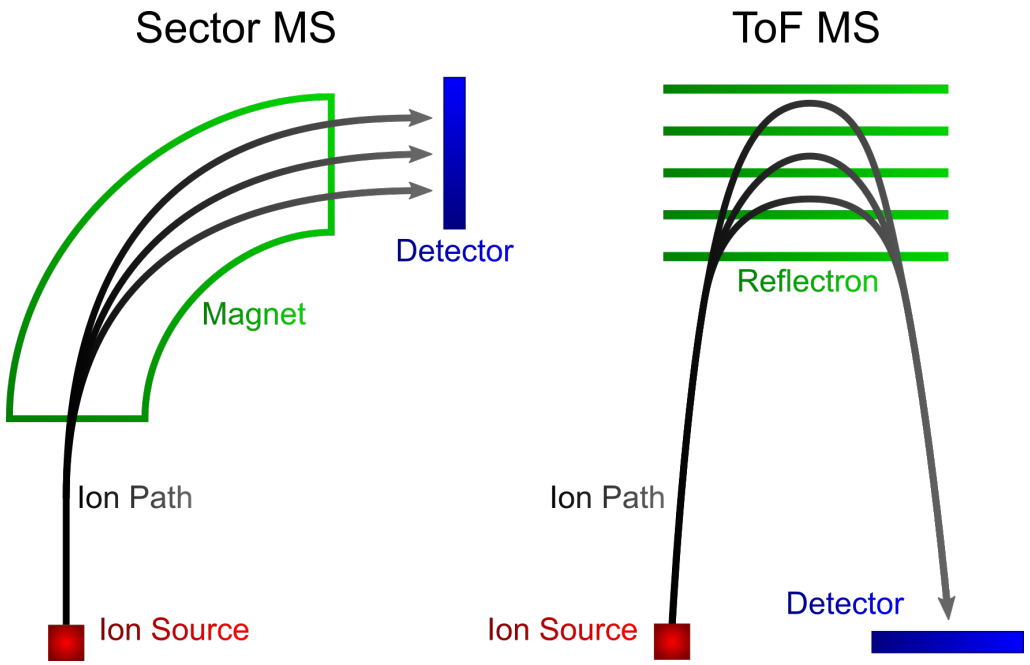


Figure 1.1: Differences between Sector and ToF MS. Simplified illustration of a Sector MS (left) and a ToF MS (right). Different colours represent various components: red – ion source, black – ion path, green – method by which ions with differing m/z's are separated, blue – detector. In Sector MS, ions with different m/z's are separated by a magnet bending their path. Ions with higher m/z ratios are influenced by the magnetic field less and so experience less path bending compared to ions with lower ratios. The position of impact on the detector is interpreted and correlated to their m/z. In ToF MS, position of impact on the detector is the same for all ions regardless of m/z. Ions are released simultaneously as a packet and pushed into a reflectron which reverses their direction of travel and sets them on a collision course with the detector. Ions with lower m/z ratios are imparted greater initial velocity by the pusher and so penetrate further into the reflectron, increasing the amount of time they take to reach the detector and vice versa. Therefore the time taken between the release of a packet and the arrival of an ion can be correlated to its m/z.

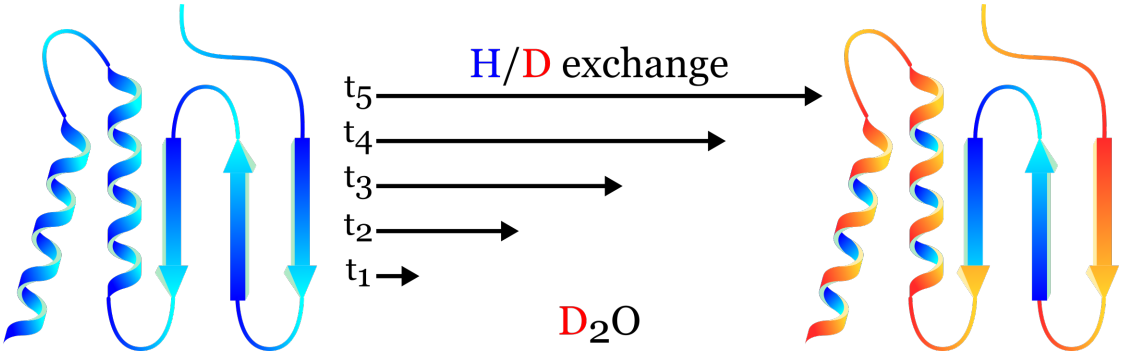


Figure 1.2: Basic principle of HDX. A protonated protein (blue) is dissolved in an excess of deuterated buffer. Over a variety of labelling time points (t1-t5), HDX will occur with exchangeable protons swapped for deuterons, causing certain residues of the protein to become deuterated (red). The amount of HDX increases with longer labelling time points.

For any given amino acid, there are three different types of hydrogens which can exchange: α -carbon, side chain and backbone amide. The exchange rate of α -carbon hydrogens is extremely slow and so no measurable exchange occurs within the time frame of a typical experiment. The exact opposite phenomena occurs in the case of side chain hydrogens, whereby because their optimal quench conditions (explained later in this section) are different from those typically used in a HDX experiment, any side chain deuterium uptake will have fully back-exchanged for hydrogen by the time of measurement [21]. The extreme timescales of these two different hydrogen positions make them impractical for use in experiments, as in both cases the net measurable amount of deuterium incorporation will be zero. By comparison, backbone amide hydrogens are perfect for deuterium uptake measurements as they are sensitive to changes in their environment and are also highly quenchable, enabling their use as probes of protein conformation and dynamics. As only one amide hydrogen is present per residue in polypeptide chains, there is therefore conveniently one measurable hydrogen exchange location per amino acid, with the exception of proline which has no amide hydrogen when part of a polypeptide. One additional caveat is that the N-terminal amino acid generally cannot be measured due to rapid back-exchange and there is debate as to whether this caveat should be applied to additional (i.e. N+1 etc.) residues as well [22].

The specific amount of HDX that a folded protein will undergo in any given timeframe is given by the observed exchange rate constant, k_{HX} which is influenced by the intrinsic chemical exchange rate of an unstructured polypeptide, k_{ch}, as well as the protection factor, PF (Equation 1.1).

$$k_{HX} = \frac{k_{ch}}{PF} \tag{1.1}$$

PF cannot be calculated directly and must instead be calculated by rearranging Equation 1.1 to give Equation 1.2:

$$PF = \frac{k_{ch}}{k_{HX}} \tag{1.2}$$

k_{ch} can be calculated using Equation. 1.3 [23]:

$$k_{ch} = k_{H^+}(A_{left} \cdot A_{right})[H^+] + k_{OH^-}(B_{left} \cdot B_{right})[OH^-] \tag{1.3}$$

Where A_{left}, A_{right} and B_{left}, B_{right} refer to side-chain-specific acid or base factors respectively. The various factors influencing both k_{ch} and PF are described in Figure 1.3. HDX is highly pH dependant with rate constants determined for both acid and base catalysed HDX for an unstructured polypeptide showing that base catalysis is by far the more important of the two. Plotting the log of k_{ch} vs. pH results in a distinctive v-shaped curve (Figure 1.4 A) with a characteristic minima around pH 2.5-3.0, the point at which rates of acid and base catalysed HDX are equal. HDX is also dependant to a lesser degree on temperature [24]. Increasing the temperature alters the water ionisation constant, KW, which increases the concentration of OH^-, thus increasing the amount of base catalysed exchange. When plotting the exchange rate vs. temperature, an exponential curve is produced (Figure

1.4 B).

For a folded protein dissolved in bulk D_2O solvent, HDX occurs by way of transient unfolding events which break internal H-bonds and allow for H-bonding to the solvent instead. Therefore the protein is in a constant state of flux between a closed exchange-prohibited state and an open exchange-capable state. The mechanism by which hydrogen is exchanged for deuterium in this way is described in Equation 1.4 [25]:

$$cl(H) \underset{k_{cl}}{\overset{k_{op}}{\rightleftharpoons}} op(H) \xrightarrow[D_2O]{k_{ch}} op(D) \underset{k_{op}}{\overset{k_{cl}}{\rightleftharpoons}} cl(D) \tag{1.4}$$

Where cl and op denote the protein in its closed and open state respectively, H and D denote the protein in its protonated and deuterated state respectively and k_{cl} and k_{op} denote the closing and opening rate constants respectively. Unfolding can occur either locally (termed "breathing motions") or globally with different modes of HDX occurring depending on which type of unfolding event occurs. If it is local unfolding of specific subsections of the protein, k_{cl} is typically much faster than k_{ch} and therefore the rate of HDX occurs according to Equation 1.5:

$$k_{HX} = \frac{k_{op}}{k_{cl}} \cdot k_{ch} \tag{1.5}$$

This is referred to as "EX2" kinetics and is the most common mode of HDX as the native state of most proteins is quite stable at physiological conditions. An alternative path may be taken however if global unfolding of the protein occurs. In this situation, k_{cl} is typically much slower than k_{ch} and therefore the rate of HDX occurs according to Equation 1.6:

$$k_{HX} = k_{op} \tag{1.6}$$

This is referred to as "EX1" kinetics [26]. These two different exchange pathways can be detected by HDX-MS as they produce characteristically different spectra when comparing samples with varying degrees of deuteration. For EX2 kinetics, the spectra appear to migrate to higher m/z values with increasing deuteration as a unimodal distribution whereas for EX1 kinetics, the spectra form a bimodal distribution with one mode equal to the lowest amount of deuterium incorporation and a second mode equal to the maximum amount of deuterium incorporation. The relative populations of these two modes in any given sample depends on its degree of deuteration. EX1 and EX2 kinetics are summarised in Figure 1.5.

Measuring the deuterium uptake of a protein as a whole returns global-level data which has only limited use for clarifying most conformation and dynamics questions. HDX-MS becomes a much more powerful tool when the protein in question is digested into peptides using an appropriate protease such as pepsin, as was first achieved in 1993 by Zhongqi Zhang & David Smith [27]. Pepsin is commonly used because it has high activity around typical HDX quench pH values (approx. 2.5) and is also active around experimental temperature values (approx. 20 °C). Pepsin also benefits from being a relatively

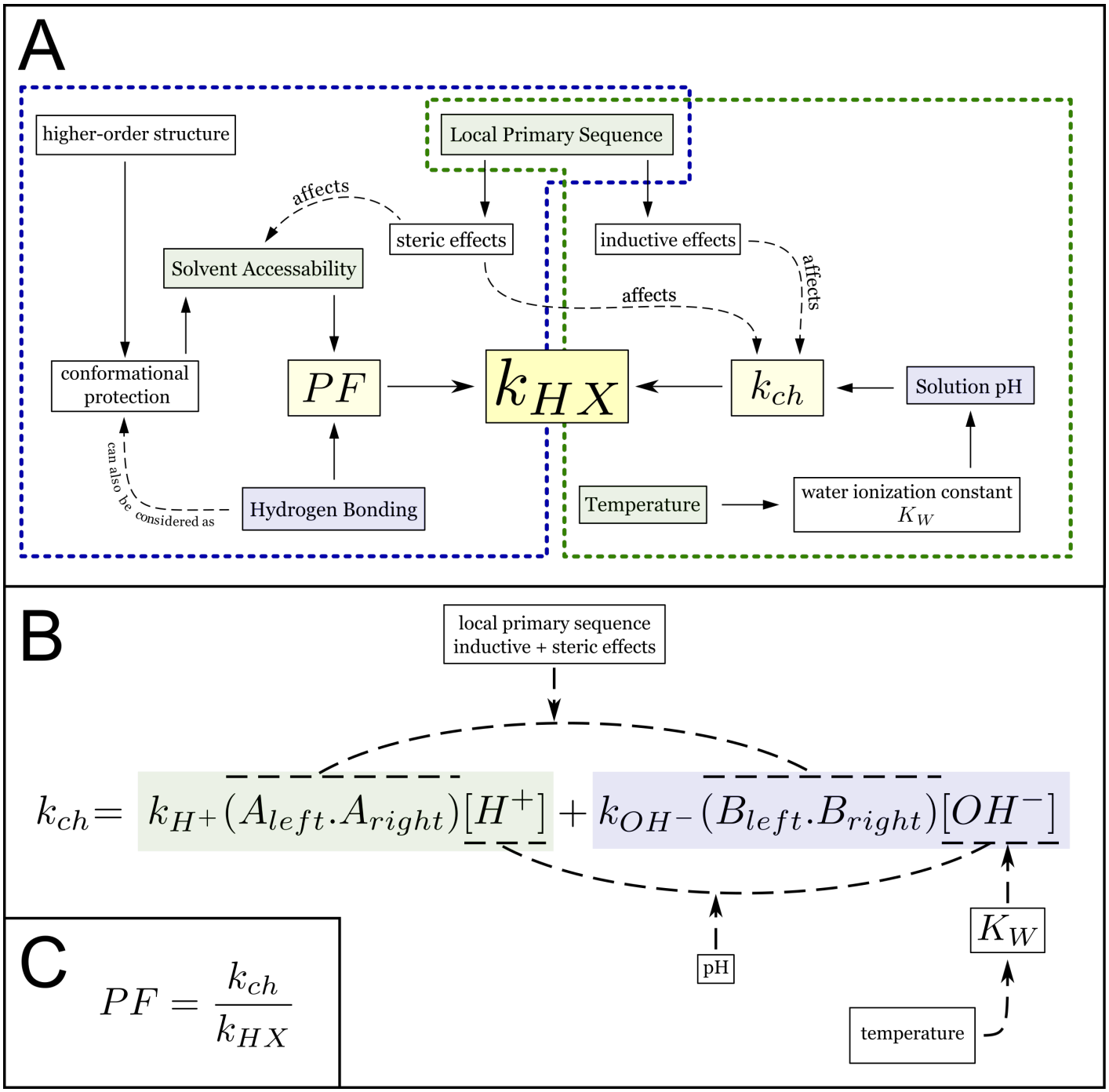


Figure 1.3: Factors influencing exchange rates in a folded protein. Blue highlights designate factors with a relatively greater degree of influence compared to green highlights which designate factors with a relatively lesser degree of influence. The importance of these factors as highlighted is confined (in the case of (A)) to be relative to other factors contained within the same dotted boundary and not across boundaries. **(A)** The exchange rate (k_{HX}) of an amino acid is influenced by two overall factors: the chemical exchange rate (k_{ch}) and the protection factor (PF). Primary factors that affect PF include: hydrogen bonding, solvent accessibility & local primary sequence. Primary factors that affect k_{ch} include: solution pH, temperature & local primary sequence. **(B)** k_{ch} can be calculated by determining the rates of acid and base catalysed exchange. The ways in which the various factors described in (A) influence the rates of exchange are indicated. **(C)** PF cannot be determined directly and must instead be calculated from $\frac{k_{ch}}{k_{HX}}$.

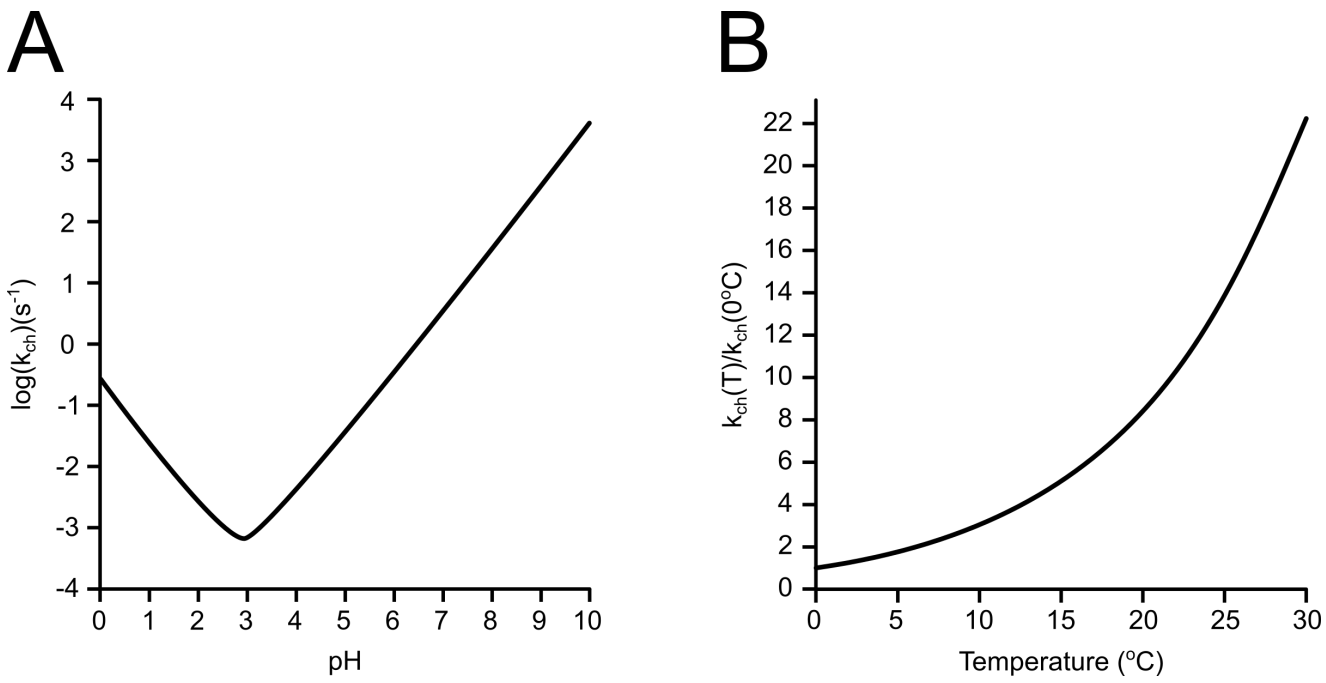


Figure 1.4: Influence of pH/temperature on intrinsic chemical exchange. (A) Graph showing how k_{ch} is affected by sample pH, displaying a characteristic minima around pH 2.5-3.0, the point at which rates of acid and base catalysis are equal. **(B)** Graph showing how k_{ch} is affected by sample temperature with a minima at 0 °C and an exponential rise with increasing temperature values. Graphs taken from [21].

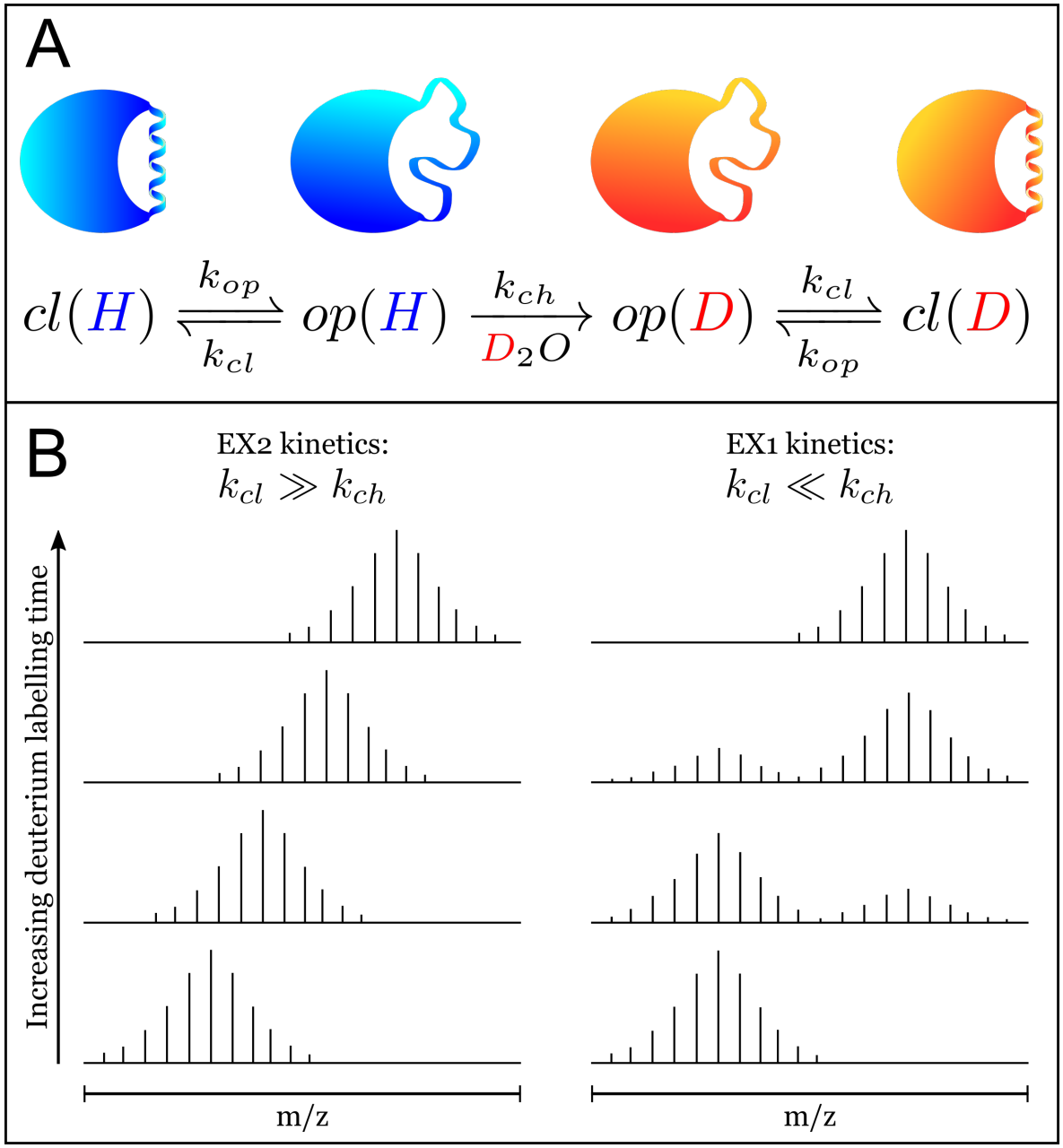


Figure 1.5: EX1 & EX2 kinetics. (A) Equation describing the mechanism by which HDX takes place. Diagrams describe the state of the protein: closed (helix), open (random coil), protonated (blue), deuterated (red) above the corresponding place in the equation. **(B)** Example mass spectra showing how EX1 & EX2 kinetics can be visually distinguished. EX2 kinetics produces a unimodal distribution that advances to higher m/z values with labelling time whereas EX1 kinetics produces a bimodal distribution, the relative populations of which tend towards the higher m/z local maxima with labelling time.

non-specific cleaver, although with a preference for large aromatics, allowing for a diverse range of overlapping peptides to be generated regardless of the protein's sequence [28]. After digestion, deuterium uptake data can then be calculated for the individual peptides instead of the protein as a whole, enabling local-level data to be obtained which allows information to be narrowed down from the protein level to the peptide level (and with some additional data processing, the pseudo-residue level by averaging exchange values across multiple overlapping peptides). Having peptide or potentially residue level uptake data allows the localisation of any observed differences to the specific sites within the protein which gave rise to them, as opposed to global HDX which is unable to provide such information.

When HDX-MS is employed as a difference method, it can be used to elucidate the conformation and dynamics of the protein, in particular it has the capacity to reveal interfacial information. When measured as part of a protein-ligand "bound" data set, peptides located at the interface between the protein and its ligand tend to have increased protection due to an increase in hydrogen bonding and/or a reduction in solvent accessibility. The result of this is that these peptides will have a reduced k_{HX} and so will uptake less deuterium compared to these same peptides when measured as part of a free "unbound" data set. Constructing a bound vs. unbound difference plot for all peptides (Figure 1.6) highlights those regions of the protein that experience an uptake decrease, which can be indicative of a binding interface.

1.3.2 Typical HDX-MS workflow

In a typical HDX-MS experiment, a small amount of the protonated protein sample is diluted into a large excess of deuterated labelling buffer at around ambient temperature and a pH value appropriate for the protein, typically around pH 7, initiating the HDX process. Labelling then occurs for a set amount of time before the sample is transferred into a quench buffer which retards subsequent exchange. HDX is very sensitive to pH and sensitive to a lesser extent to temperature and thus the quench buffer is designed in order to slow HDX rates as much as possible during protein digestion and subsequent chromatography. This is done by reducing the pH of the reaction down to a value of approx. 2.5, the point at which the rate of HDX is lowest, reducing the overall exchange rate by approx. 4 orders of magnitude compared to the rate at pH 7. In addition, the temperature of the reaction is lowered to 1 °C, reducing the overall exchange rate by approximately 14-fold compared to the rate at 25 °C. When these pH and temperature changes are combined, the HDX rate is slowed by more than 5 orders of magnitude [21] which allows sufficient time for subsequent processing steps to occur before the deuterium labels completely back exchange for hydrogen.

After quenching, the sample is digested on-line by an immobilised protease (typically pepsin) with the resultant peptides separated by liquid chromatography (LC) before being transferred into the gas phase by ESI. All of the experiments in this thesis were performed using a Waters Synapt G2Si HDMS coupled to an Acquity UPLC M-Class system with HDX and automation. The Synapt allows for the removal of neutral contaminants by an ion transfer device and m/z filtering by quadrupole as well as two different primary forms of separation to be used in series: the first is Ion Mobility Spectrometry (IMS) which separates ions based on their Collisional Cross Section (CCS) & m/z using a drift tube. The second (and principal) separation type is ToF whereby ions are separated by their m/z using an electric field of

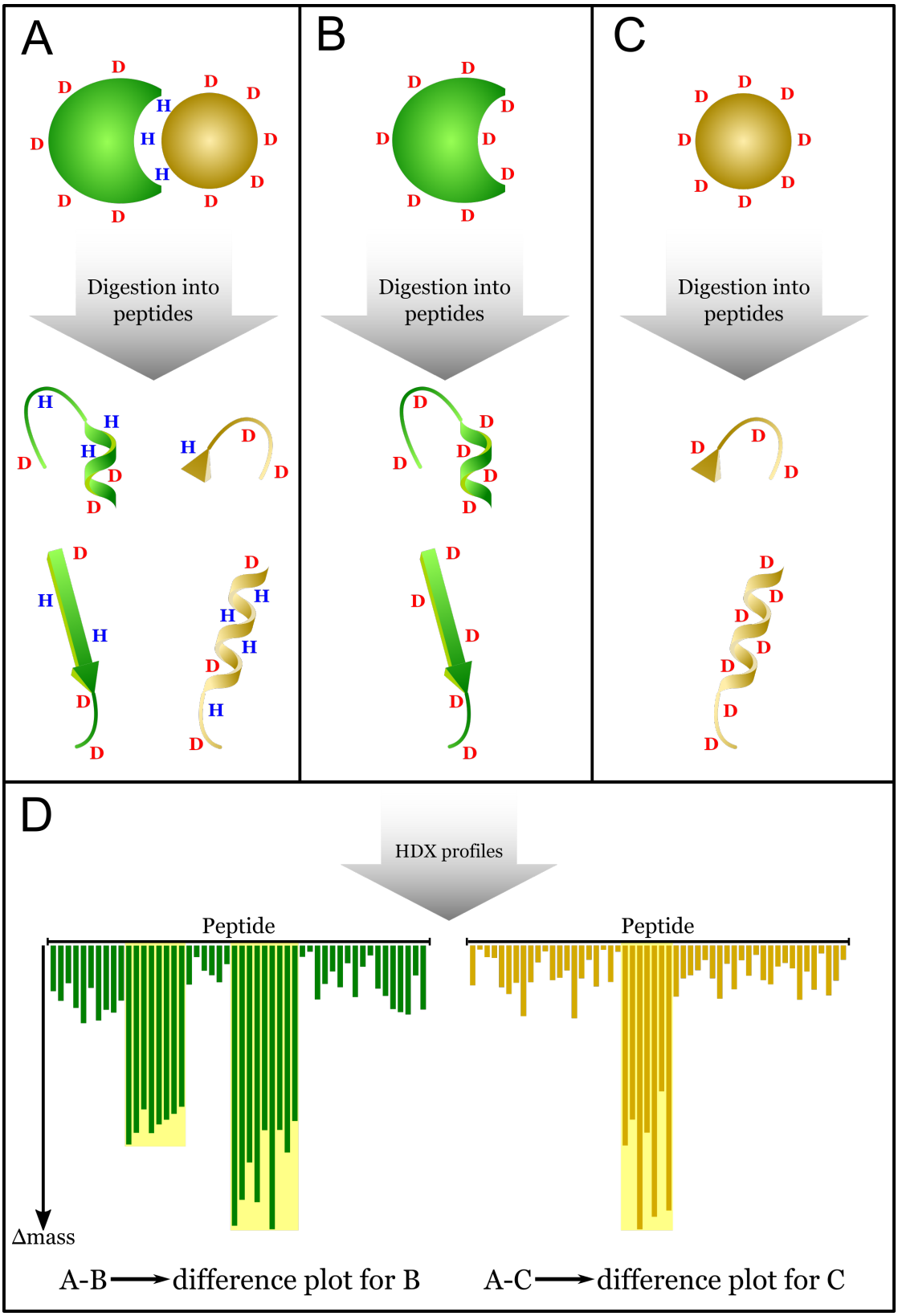

A
B
C
Digestion into peptides
Digestion into peptides
Digestion into peptides
D
HDX profiles
Peptide
Peptide
Δmass
A-B → difference plot for B
A-C → difference plot for C

Figure 1.6: Elucidation of the location of protein binding sites using HDX-MS. Two proteins (green & yellow) are known to bind together but the location of the binding interface is unknown. HDX-MS can determine this by collecting 3 data sets: a bound data set with the proteins mixed together (A), and 2 unbound data sets with just the individual proteins on their own (B & C). In these 3 data sets, the proteins are incubated in buffer containing bulk D_2O solvent for a certain amount of time, causing hydrogens (blue "H") to be exchanged for deuterium (red "D"). In the bound data set, an increased PF at the binding interface prevents the same degree of deuterium uptake from occurring in this region vs. the same regions in the unbound data set. The proteins in all data sets are then digested into peptides and characterised by MS, enabling isotope uptake to be localised to the peptide level. HDX profiles (D) are then generated by subtracting the level of deuterium uptake for each peptide in the bound data set vs. the appropriate unbound data set in order to determine its degree of uptake difference. Regions which show a deuterium uptake difference exceeding a certain threshold (yellow highlights) can be used as evidence for the location of the binding interface, although allosteric effects must also be considered.

known strength propelling them towards the detector, with the arrival time of the ions corresponding to their m/z. After data has been collected for the analyte in both the bound and unbound states at multiple time points, the data is analysed in order to determine the amount of deuterium incorporation that has occurred (vs. an un-deuterated reference) for each peptide. Any uptake differences between the states can then be visualised and further data processing carried out.

During an experiment, back exchange of deuterium for hydrogen occurs because all subsequent liquid phase steps after incubation occur in protonated solvent. Even with the rate of exchange slowed to its minimum practical value, a substantial amount of back exchange can occur during this time. Therefore, it is usually worthwhile running a back exchange control (BEX) during an experiment which consists of a sample that has been fully deuterated, usually by incubation in D_2O for a period of several days/weeks. As all the peptides in this sample should have a Relative Fractional Uptake (RFU) of 1 i.e. all theoretically exchangeable sites should have exchanged, any that are missing can be used to correct the normal labelling samples for back exchange using Equation 1.7 [27]:

$$D_{corr} = \frac{m_{expt} - m_{ref}}{m_{100} - m_{ref}} \cdot N \tag{1.7}$$

Where D_{corr} is deuterium uptake of the peptide corrected for back exchange, m_{expt} is mass of the peptide, m_{ref} is the mass of the peptide in the non-deuterated reference, m_{100} is the mass of the peptide in the BEX and N is the theoretical number of exchangeable sites in the peptide. Back exchange correction is not technically required if the experiment is only concerned with the acquisition of difference data, as both the bound and unbound states should experience the same level of back exchange. However, if researchers are interested in obtaining absolute values of RFU, then back exchange correction is the minimum that must be carried out. An additional step that can be taken is to correct for in exchange as well. In exchange controls (IEX) test the ability of the quench to retard further exchange and consist of an un-deuterated sample being immersed in a deuterated quench. If the quench completely prevented any additional exchange from occurring, the mass of IEX peptides should be identical to those of the non-deuterated reference samples. Therefore, any difference between the reference masses and the IEX masses can be attributed to the quench not fully preventing further exchange from happening. Data from the BEX and the IEX can be combined in order to correct for any extraneous exchange using Equation 1.8:

$$D_{corr} = \frac{m_{expt} - m_0}{m_{100} - m_0} \cdot N \tag{1.8}$$

Where m_0 is the mass of the peptide in the IEX. When working with RFU values instead of mass values, Equation 1.8 can be simplified slightly to from Equation 1.9:

$$RFU_{corr} = \frac{RFU_{expt} - RFU_0}{RFU_{100} - RFU_0} \tag{1.9}$$

Where RFU_{corr} is deuterium uptake of the peptide corrected for back exchange, RFU_{expt} is RFU of the peptide, RFU_0 is the RFU of the peptide in the IEX, RFU_{100} is the RFU of the peptide in the

BEX.

1.4 Work undertaken in this thesis

1.4.1 Overview

The ability to highlight binding interfaces is one of the great applications of HDX-MS: it can provide evidence of the location of interfaces between a protein and a ligand for which there is no complex structure available. It should be noted however that HDX-MS is not typically used to determine the specific three-dimensional fold of a protein or protein-protein complex, merely to localise where along the amino acid sequence binding is suspected to take place. This weakness is what we are attempting to remedy with this thesis. HDX-MS is very high throughput compared to techniques such as x-ray crystallography, with experiments and data analysis able to be completed in days instead of months or years and, thanks to aforementioned recent advances in methodology, also has the potential to be used for structure determination as well.

The potential to calculate HDX profiles from three-dimensional structures, including from models such as docking poses has been demonstrated previously [4]. The capacity to calculate HDX-MS data from structures is useful as it facilitates varying models to be differentiated on the basis of their comparison to experimental HDX-MS data. This method also permits the generation of a quantitative metric which can be used to evaluate whether a given structure (especially *in silico* models such as decoys and docking poses) are native or not. This ability to assess HDX data quantitatively rather than the usual qualitative appraisal is a marked advantage of this new technique as it reduces user interpretation as a potential source of error. However, while the foundation stones for using this approach to assess native structures have been described, a significant amount of development is still required for approaches such as these to become mainstream. The primary focus of this PhD has therefore been on the continued development of the methodology presented in [4] to allow the characterisation of *in silico* native protein structures for *ab initio* applications. In order to accomplish this, we had the following aims and objectives:

- To develop a quantitative metric in order to allow the classification of protein structures as either native or non-native
- The acquisition of additional protein data sets in order to properly benchmark subsequent methodological developments
- The development of several computational tools with which we could automate the processing and analysis of data sets
- To accurately classify monomeric protein decoys as being either native or non-native in nature
- To accurately classify binary PPI docking poses as being either native or non-native in nature

As part of a wholly separate project, we also conducted HDX-MS analysis on the interaction of two proteins: dUTPase & Stl, in order to determine the location of their binding site for the purposes of developing a mechanism by which they interact.

1.4.2 Development of a quantitative metric to enable structure classification

First, we developed a method to quantify the ability of HDX-MS to discriminate between native and non-native protein conformations based on the approach of Vendruscolo & Paci *et al.* [29] and Best & Vendruscolo [30] to estimate protection factors from decoy sets. This was accomplished primarily using Equation 1.10:

$$lnP_i^{sim} = N_i^C \beta_C + N_i^H \beta_H \tag{1.10}$$

Where lnP_i^{sim} is the natural logarithm of the simulated protection factor of residue i, N_i^C is the number of heavy atoms, N_i^H is the number of hydrogen bond acceptors within certain distance cut-offs from the backbone amide and β is an empirically determined scaling factor that is independently applied to both N_i^C and N_i^H. The capabilities of this method were evaluated on the peptide level using the simulated protection factors to calculate HDX-MS outputs of proteins and their assemblies and then compare these calculations to experimental data obtained in-house to generate Root Mean Square Error (RMSE) values for each structure in a decoy set. Each decoy was then compared to the native structure to generate Root Mean Square Deviation (RMSD) values. The ability of HDX-MS to identify native structures (defined as those with an RMSD value of $\leq$ 2.5 Å) from a background of decoys was analysed quantitatively based on their performance in a binary classification system via the construction of Receiver Operating Characteristic (ROC) curves in order to provide insight into the use of HDX-MS for protein modelling. Our intentions for this work were to develop the method to correctly classify the input structures at a rate significantly above that of random chance. This work is detailed in Harris *et al.* 2018 [31], included in full in Appendix L.

In this paper, we demonstrated that HDX-MS data simulated directly from atomic structures can be highly diagnostic of a protein's native fold, even if the PFs underpinning this data are poorly defined. Interestingly, this diagnostic capacity was actually higher for data calculated from crystal structures then for data calculated from an ensemble. While perhaps surprising, this observation underpins much of our work going forward due to the substantial throughput increase gained by not having to generate an ensemble for every protein we wanted to analyse. Data generated for the analytes alpha lactalbumin and barnase indicated that high peptide redundancy was arguably one of the most important factors for accurate determination of fold; proposed to be due to the extra constraints such redundancy places on lnP calculations. However despite our success at identifying the native fold for alpha lactalbumin and (to a lesser extent) barnase, we found that when the same technique was applied to homo-protein assemblies such as enolase it did not fare as well. We proposed that this could be due to significant differences in the HDX behaviour of protein complexes and the fact that Equation 1.10 was never optimised for use with large multi-chain proteins. Additionally, for homo-protein assemblies, knowledge of peptide redundancy and coverage in the native interface can only be had with the aid of a high-resolution structure. This is not a challenge for hetero-proteins however, as the degree of peptide sampling in the native interface can be inferred directly from associated HDX-MS difference data without the need for any structural reference. The production and purification of barnase as well as experimental HDX data acquisition and analysis and related writings were all done by the author.

1.4.3 Acquisition of additional protein data sets

The second step undertaken in this thesis was in response to a limitation identified in our paper, namely that more protein data sets were required in order to properly evaluate the technique. Therefore, we set about producing/acquiring a number of additional proteins with which we could generate additional data sets for testing purposes, with the goal being to acquire enough to thoroughly test subsequent methodological advances. These proteins were barnase (Bn), barstar (Bs), Green Fluorescent Protein (GFP), GFP-nanobody (GFP-nb) & GFP-nanobody minimiser (GFP-nbmin); all of which had the additional property of being involved in a binary PPI (Bn:Bs, GFP:GFP-nb & GFP:GFP-nbmin), a feature that was necessary for additional work we would carry out adapting our method to work on binary PPIs.

1.4.4 Development of computational tools for process automation

In order to improve upon the techniques detailed in the paper, Ramin Salmas from the Borysik group developed two tools that would be fundamental to our work. The first was HDXmodeller which takes experimental peptide-level RFU data and returns residue-resolved protection factors [32, 33]. The second tool was HDXsimulator which predicts the veracity of protein models by calculating RFU as well as lnP data and comparing them to experimental/modelled values, RFU produced by HDX-MS and lnP produced by HDXmodeller, allowing for discrimination between native and non-native structures using both RFU and lnP as metrics.

HDXmodeller was developed primarily by Ramin Salmas and Antoni Borysik with limited testing undertaken by the author, however HDXsimulator was co-developed as part of the work carried out in this thesis. The third step undertaken was therefore as part of the process of mapping out the boundaries of modelling protein conformation by HDX-MS using HDXsimulator. In order to do this, we conducted a comprehensive study where we sought to determine HDXsimulator's capabilities and limitations, a process which involved the comparison of calculated data against data with errors deliberately introduced in order to test how the binary classification system would respond. Our goal was to try and determine the relationship between the degree of error and the data set's eventual classification score, as this understanding would be essential to making improvements to the program in the future. This goal was expedited by the development of an automation pipeline involving numerous Python and Bash scripts which greatly sped up the collection of data sets.

1.4.5 Classifying monomeric protein decoy structures

The last step in this process was therefore using HDXsimulator to select for native structures against a backdrop of decoys for each of the protein data sets mentioned previously. This process used part of the same pipeline that was developed for the aforementioned mapping step and substituted out deliberately error-laden RFU and lnP values for experimental/modelled ones produced earlier. RFU and lnP values calculated for the decoys were then compared against these experimental/modelled values in order to determine which among the decoys could be considered a native structure and which a non-native one, with confidence being judged by the binary classification system. In doing this we hoped to see the method correctly classifying the input decoy structures of the various different proteins at a

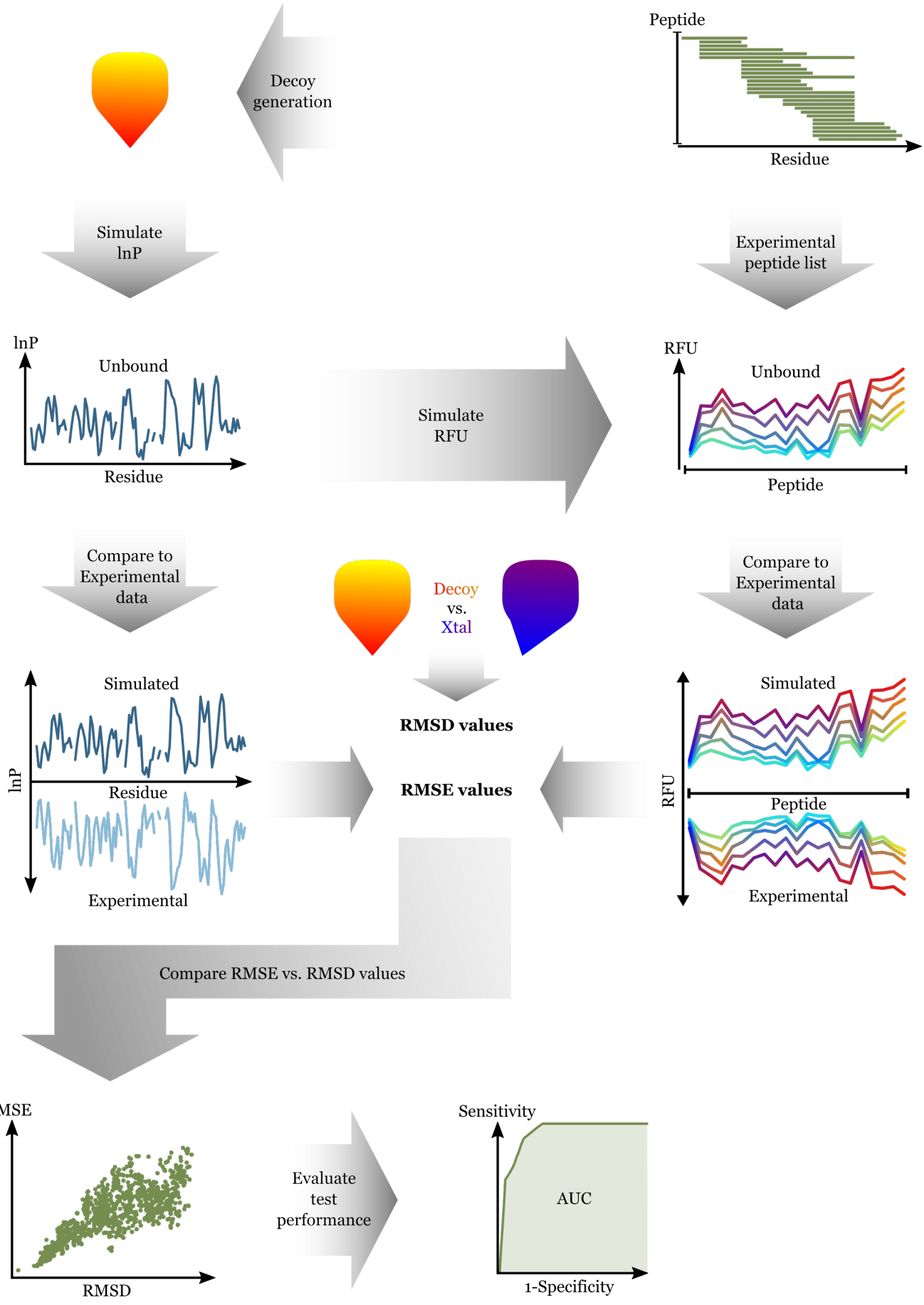

Decoy generation
Peptide
Residue
Simulate lnP
Experimental peptide list
lnP
Unbound
Residue
Simulate RFU
RFU
Unbound
Peptide
Compare to Experimental data
Decoy
vs.
Xtal
Compare to Experimental data
Simulated
lnP
Residue
Experimental
RMSD values
RMSE values
Simulated
RFU
Peptide
Experimental
Compare RMSE vs. RMSD values
RMSE
RMSD
Evaluate test performance
Sensitivity
AUC
1-Specificity

Figure 1.7: Pipeline for the prediction of protein structure. A set of decoys (yellow-red) are first generated using a method such as Rosetta or 3DRobot. lnP values for each decoy are then calculated from the unbound (dark blue line graph) HDX data set using the procedure outlined in [4]. RFU values can then be calculated for each decoy (multi-coloured line graphs) from the simulated lnPs and the experimental HDX peptide list for the same protein. Comparing the simulated lnPs and RFU values to experimental values allows the calculation of RMSE metrics. Technique validation is accomplished by also calculating RMSD values between each decoy and the crystal structure (blue-purple) and visualising each docking pose's RMSE vs. RMSD values in a scatter plot where those decoys with the low RMSE values should also have low RMSD values. Test performance can be quantified by performing ROC curve analysis and generating an Area Under Curve (AUC, green shading).

rate significantly above that of random chance. This pipeline is illustrated in Figure 1.7.

1.4.6 Classifying binary PPI docking structures

As a natural progression to this primary focus on determining an individual protein's fold, we also began, but did not finish, developing a pipeline to do the same for protein complexes. This is the reason why all of the additional proteins used in this work form part of a binary PPI: Bn:Bs, GFP:GFPnb & GFP:GFPnbmin; so that their data could pull double duty for both methods. HDX profiles of the two proteins in both their bound and unbound forms were collected previously. Next, Molecular Dynamics (MD) simulations were carried out in order to relax the individual crystal structures of the proteins that would serve as the starting points for the pipeline, followed by protein-protein docking to generate the range of potential complex structures. These docking poses would then be treated as described in Figure 1.8 in order to quantitatively assess their accuracy. In doing this we hoped to see the method correctly classifying the input docking poses of the various different binary PPIs at a rate significantly above that of random chance.

1.4.7 Investigating the interaction of dUTPase with Stl

In addition to method development, we also carried out a substantial amount of work on a completely unrelated project: analysing the interaction between the proteins dUTPase and Stl. This was a collaborative project between the Borysik group and the Vértessy group from the Budapest University of Technology and Economics and revolved around the much more traditional application of HDX-MS: identifying interaction surfaces. Our aim with this project was to use this information to develop a mechanism by which the interaction of dUTPase with Stl could occur.

A visual summary of all the work undertaken in this thesis is available in Figure 1.9 and the various constituent parts will now be explained in detail.

1.5 HDXsite: tools for the analysis of HDX-MS data

HDXsite (`https://hdxsite.nms.kcl.ac.uk/`) is website hosting a suite of webserver-based tools developed by the Borysik group for the analysis of HDX-MS data. It contains two major programs: HDXmodeller & HDXsimulator as well as a collection of ancillary tools under the banner of HDXutilities that can be useful for the submission of data to and analysis of data from the two main programs.

1.5.1 HDXmodeller

HDXmodeller is a tool for generating high-resolution information from low-resolution HDX data. The program accepts peptide-level RFU data typical of proteolytically cleaved HDX-MS experiments and returns residue-resolved modelled lnPs along with a range of statistical outputs for validation. The primary validation metric is a co-variance matrix which calculates pair-wise correlation coefficients for each replicate performed across the entire run and outputs the mean of these values as an R-matrix score. Such a score can be used to inform on the predicted accuracy of the modelled data, with values

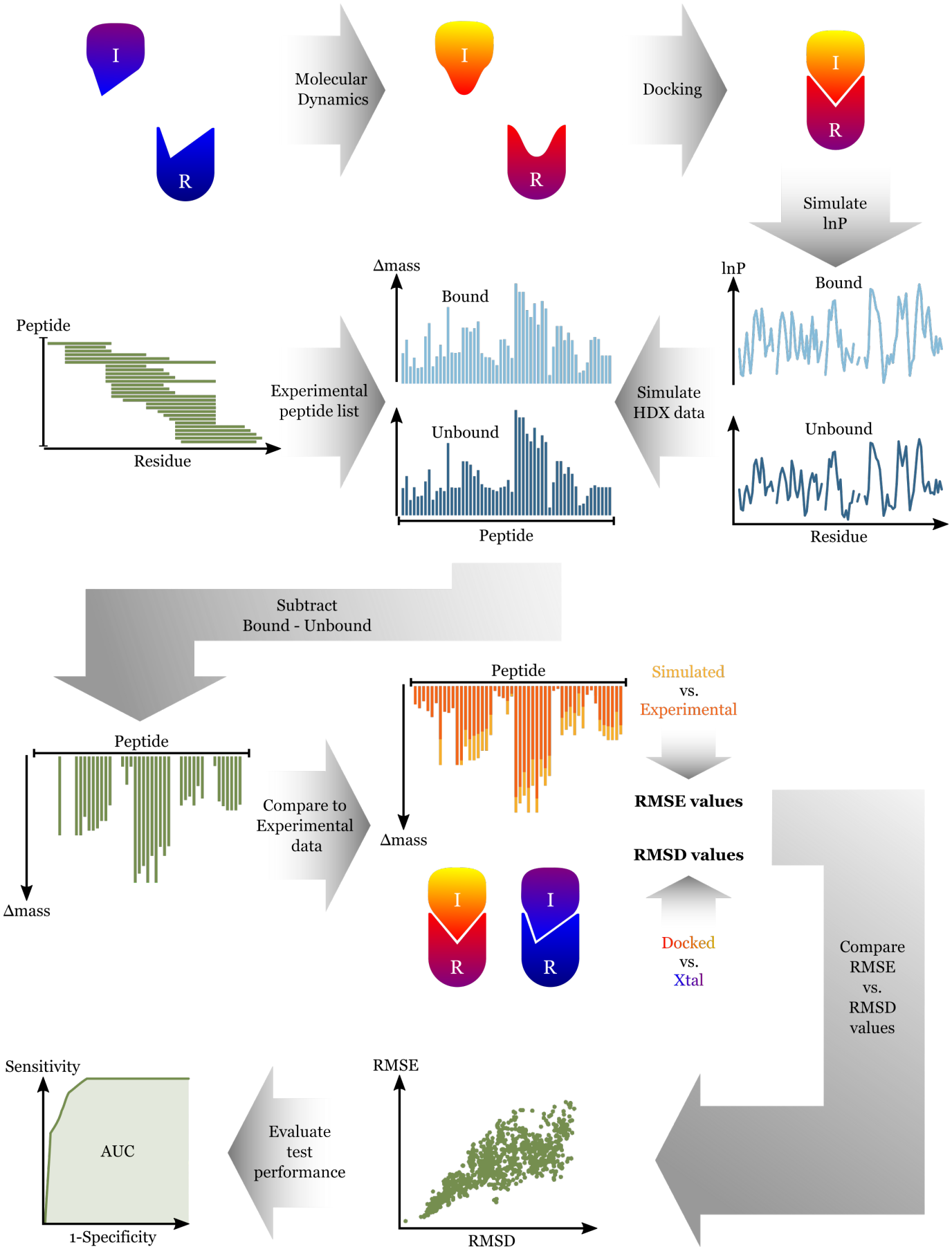

I
R
Molecular Dynamics
I
R
Docking
I
R
Simulate lnP
lnP
Bound
Unbound
Residue
Δmass
Bound
Unbound
Peptide
Simulate HDX data
Peptide
Residue
Experimental peptide list
Subtract Bound - Unbound
Peptide
Δmass
Compare to Experimental data
Peptide
Δmass
Simulated vs. Experimental
RMSE values
RMSD values
Docked vs. Xtal
I
R
I
R
Compare RMSE vs. RMSD values
RMSE
RMSD
Evaluate test performance
Sensitivity
AUC
1-Specificity

Figure 1.8: Prediction of protein-protein complex structure using docking and HDX-MS. The crystal structures of the two proteins involved in the binary PPI under investigation (blue-purple R & I) are first relaxed by MD (Figure 1.10) to ensure no bias is introduced into subsequent steps. The proteins are then docked together (as in Figure 1.11) using Ambiguous Interaction Restraints (AIRs) generated by HDX data (yellow-red R & I). lnP values are then calculated for both the bound (light blue line graph) and unbound (dark blue line graph) forms of each pose of the docked complex from atomic coordinates using the procedure outlined in [4]. Simulated lnP values are then combined with experimental HDX peptide lists for the same proteins in order to generate simulated HDX data for both the bound (light blue bar graph) and unbound (dark blue bar graph) forms of both proteins. Bound simulated data is then subtracted from unbound simulated data in order to construct a simulated difference plot. Comparing the simulated (yellow bar graph) difference plots to an experimental (orange bar graph) difference plot allows the calculation of RMSE values, the metric that will be used to score the docking poses. Technique validation is accomplished by also calculating RMSD values between each docked pose and the crystal structure (blue-purple I & R) and visualizing each docking pose's RMSE vs. RMSD values in a scatter plot where those poses with the low RMSE values should also have low RMSD values. Test performance can be quantified by performing ROC curve analysis and generating an AUC (green shading).

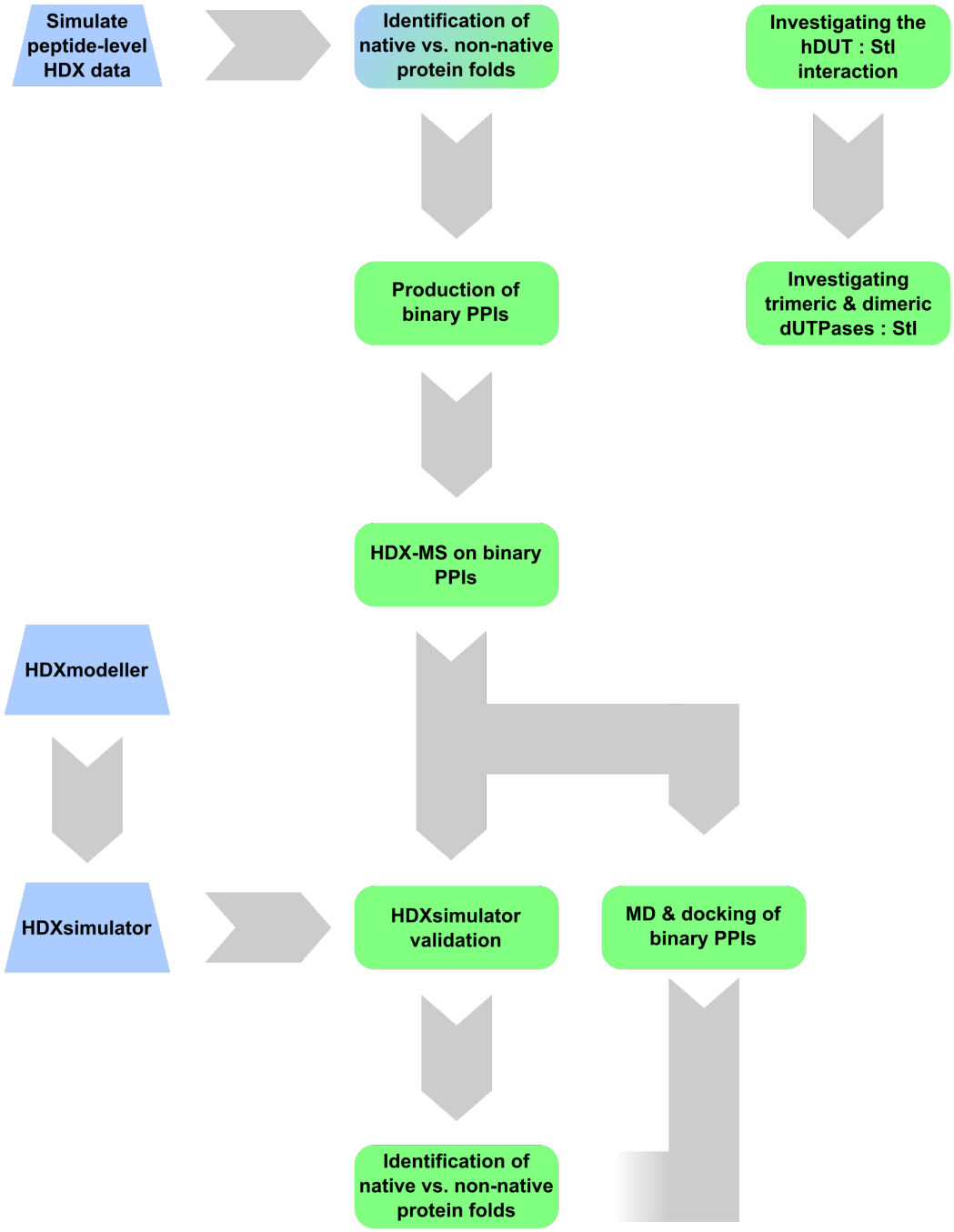


Figure 1.9: Schematic overview of work undertaken in this thesis. Diagram detailing the various different stages of this PhD and how they relate to one another. Stages represented in green were undertaken by the author, stages represented in blue were undertaken by others. The blue-green stage represents a collaborative effort. The incomplete arrow represents the future work that was not able to be completed during this thesis.

≥ 0.7 considered to be of high accuracy, values 0.5-0.69 considered to be of fair accuracy and values $<$ 0.5 considered to be of low accuracy.

1.5.2 HDXsimulator

HDXsimulator is a tool for predicting the veracity of protein models by calculating HDX data and comparing it to experimental values, allowing for discrimination between native and non-native structures. The program accepts a number of three-dimensional models such as decoys and returns calculated lnP and RFU values for each structure. HDXsimulator compares calculated lnP values with those reported by HDXmodeller and calculated RFU values with those experimentally determined by the original HDX-MS experiments to generate RMSE metrics for each model. For the purposes of testing the efficacy of HDXsimulator, it also compares the three-dimensional coordinates of the models themselves with a native structure to generate an RMSD metric for each model. Theoretically, if lnP and RFU calculation are accurate, models with the lowest RMSE values should be closest to the native structure. Having an RMSD value for each model allows for quantitative validation via the construction of a ROC plot, which tests the diagnostic ability of a binary classifier system i.e. whether a pose is native or non-native. ROC plots graph the True Positive Rate (TPR or "sensitivity") vs. the False Positive Rate (FPR or "1-specificity") with the scoring metric being the Area Under the Curve (AUC). An AUC value above 0.5 classifies the test as being better than a random guess with values of 0.7-0.8 considered to be acceptable, values of 0.8-0.9 considered to be excellent and values $>$ 0.9 considered to be outstanding [34]. Values below 0.5 indicate that the method assigns the classification incorrectly, i.e. a value of 0.1 suggests that native structures are almost always being classified as non-native etc.

1.6 Using Python to accelerate the acquisition of data

Using HDXsimulator to model residue-level lnPs from three-dimensional structures requires a large amount of data processing steps and is very time consuming to run manually, with an average time-per-data set of over a day. As this tool was still in development, it was known that a large number of data sets would need to be run and so would benefit greatly from as much automation as was feasible in order to reduce the time required to produce each data set. In order to accomplish this, we used the programming language Python as well as Bash shell scripts to facilitate intermediary data processing steps and so reduce both time and human error.

Python is powerful and readable object-orientated programming language first released in 1991 by Guido van Rossum that is comparable to Perl and Java. Object-orientated means that the language is organised around manipulating data objects which can have unique attributes and behaviours as opposed to being organised around the functions and logic required to manipulate those objects. Python is notable in the realm of programming languages in that its syntax is comparatively easy to read, making it an excellent first language to learn for those without a computer science background. Python can be run on any operating system, making it a very portable language and also features the ability to extend its functionality by adding new modules which can even be coded in another language. These attributes, combined with a large standard library that supports many common programming tasks has led to Python becoming the *de facto* language of choice for scientists interested in applying computational strategies to natural science problems.

1.7 Generating protein structure decoys

1.7.1 Rosetta

Rosetta is an extensive software suite that includes many algorithms for computational modelling and analysis of protein structures. Of particular note for the work we conducted on decoy generation are the "Abinitio" and "Relax" tools [35–40] which we used in combination with each other in order to generate decoy sets for all the individual proteins presented in this thesis.

Abinitio is a tool for *de novo* structure prediction that consists of a coarse-grained, fragment-based search through conformational space using a knowledge-based "centroid" score function that favours protein-like features. It takes as input the protein's sequence and two fragment files containing short 3-mer & 9-mer backbone fragments that are randomly inserted at all positions during the calculation. These fragments files were created using Robetta's Fragment Libraries tool (`http://old.robetta.org/fragmentsubmit.jsp`) [40, 41]. Optionally, the crystal structure of the protein can be included to generate additional scoring information from the decoys. Abinitio was used to generate the vast bulk of the decoys for each protein system, however a problem we found (as is common to *de novo* methods) was that few of the structures it generated were native, even with a high number of decoys. In order to enrich the supply of native structures to be approx. 2 % of the total, we employed "Relax" which is an all-atom refinement tool using Rosetta's full-atom force-field. We took the most native-like of the Abinitio structures and used Relax to generate additional conformers which formed the bulk of our native structures.

1.7.2 3DRobot

In order to see the effects that different types of decoy generation would have on our eventual output, we also created protein decoy sets using the tool "3DRobot" (`https://zhanglab.ccmb.med.umich.edu/3DRobot/`) [42], a program devoted to automated generation of diverse and well-packed protein structure decoys. 3DRobot takes a crystal structure as input and then identifies diverse structural scaffolds from a non-redundant PDB library, followed by restraint-free fragment reassembly simulations in order to construct a series of diverse full-length models. Finally, the models are further refined at the atomic level by a two-step iterative energy minimisation procedure in order to improve the hydrogen bonding networks and steric overlaps within the decoys. 3DRobot has several theoretical advantages over *de novo*-based approaches such as Rosetta. One is that 3DRobot allows the user to define a custom range of RMSDs over which the decoys can deviate from the original structure, as opposed to Rosetta in which the RMSDs of the decoys vs. the original is mostly random. Another advantage of 3DRobot is that the generated decoys are spread over a much more linear range of RMSDs, whereas those generated by Rosetta tend to cluster around certain values. This can make sampling certain RMSD values difficult with Rosetta but with 3DRobot, every one is equally populated.

1.8 Molecular Dynamics simulations

1.8.1 Theory of Molecular Dynamics

MD simulations are a computational method for the prediction and analysis of the movements of atoms over time, achieved by solving the physical equations governing inter-atomic interactions [43]. MD has found uses all over the scientific spectrum but particularly in the realm of structural biology because of its ability to elucidate important biochemical mechanisms such as protein folding, conformational change upon ligand binding, perturbations induced by mutation/post-translational modification etc. [44]. The movement, or trajectory, of each individual atom in the simulation is calculated over a typical time-step of a few femtoseconds, allowing extremely high temporal resolution models of protein motions to be generated. As a consequence, MD simulations are very computationally expensive, requiring the throughput of a workstation or High Performance Computing (HPC) cluster to be viable. Indeed, the modern zeitgeist of macromolecular MD simulations comprising many tens of thousands of atoms has only been brought about by the vast increase in computational power since their origins. The first simulations of a few hundred gas atoms were performed in the late 1950s [45] and it would take another 20 years before technology advanced to a sufficient degree to allow the first MD simulation of a small protein (58 residues) for a short time (8.8 ps) [46]. Modern simulations can, by comparison, calculate the trajectories of many hundreds of amino acids over many hundreds of nanoseconds, enabling their current use cases.

The interactions that atoms in a simulation will experience are defined and calculated using a mathematical model known as a "force field". They typically contain information regarding factors such as ideal bond length, bond angle, electrostatic and van der Waals interactions etc. Force fields are typically created by a mixture of theoretical calculations and experimental data and so are inherently approximate in nature with different force fields being specialised for specific purposes. As our knowledge of the various factors that influence inter-atomic interactions has improved over time, so too have force fields become better at modelling them, with substantial improvement seen over the last decade [47].

Running simulations on proteins *in silico* instead of experiments *in vivo* or *in vitro* offers a number of distinct advantages. One is that MD simulations record the exact position and velocity of every atom at every time-step, an impossible task for any experimental technique. Another is that because the whole process is entirely artificial, every aspect of the simulation can be exactly controlled, from the components that make up the simulation to the conditions under which it is run. This is in comparison to experimental techniques where there is always an element of doubt as it is impossible to know for certain the exact make up of a sample (ligands/mutations/post-translational modifications etc.) or maintain conditions perfectly.

In order to run MD simulations, a number of steps need to be taken. The first is modelling in any missing residues/atoms (including hydrogens) that may not be present in the original structure. Next the system is solvated by adding a waterbox of appropriate size around the protein. Thirdly, periodic boundary conditions are set up which allows the finite waterbox to act like an infinite system whereby anything that exits the waterbox through one side enters it again from the opposite side.

Lastly, decisions need to be made about various parameters used to run the simulations such as which force field to use and system temperature, as well as which physical parameters are allowed to vary over the simulation's duration and those that must be kept constant. Once these set up steps have been completed, the MD simulation can begin, usually with an Equilibration run which is carried out to remove the effects of adding water, ions etc. around the protein. This is achieved over a number of time steps in order to prevent simulation instability. Once equilibrium has been reached, the Production run (the main MD simulation) can begin, usually lasting tens or hundreds of nanoseconds depending on the research question being asked. The steps involved in setting up and running an MD simulation are illustrated in Figure 1.10.

1.8.2 Solution Builder & NAMD

Two programs were used to run the MD simulations: CHARMM-GUI's Solution Builder tool (`http://www.charmm-gui.org/?doc=input/solution`) was used to generate input files and NAMD then took those files and ran the simulations. CHARMM-GUI [48] is a web server allowing the user to interactively build complex systems and prepare their inputs with established simulation protocols for biomolecular simulations. The Solution Builder tool [49] is specialised to generate input files for MD simulations of proteins in aqueous solvent environments. Solution Builder was used to model missing residues and add hydrogens to structures, solvate them in a waterbox of appropriate size, determine periodic boundary conditions and lastly to set the force field as well as other simulation setting such as temperature. These input files were then accepted by a local install of NAMD. NAMD [50] is a parallel MD program designed for high-performance simulation of large biomolecular systems. This program was used for equilibration of the system and also for the main production run.

1.9 Protein-protein docking

1.9.1 Theory of protein-protein docking

Protein-protein docking is the prediction of the structure of a complex from the structures of the individual proteins [51]. The process of docking involves taking two or more proteins and/or small molecules and orientating them together into a realistic approximation of the native complex. There are many different docking programs available such as PatchDock/FiberDock [52–54], RosettaDock [55,56] & HADDOCK [57,58] etc., all of which work in subtly different ways but which usually share the same general steps. These comprise a global search, followed by a local search and finally evaluation and analysis. The first step is a coarse-grain rigid-body docking simulation in which one protein is kept static and the other rotated around it with 6 degrees of freedom (translations and rotations around the x, y & z axes) in order to generate a range of approximate orientations or poses. The second phase involves flexible refinement whereby the poses generated by the rigid-body docking have their side chain orientations perturbed in order to simulate the conformational changes that typically occur upon binding. Once docking has been completed, various analysis and evaluation steps can be taken involving a range of both automated and manual methods (Figure 1.11).

One of the great challenges facing all docking protocols is how to score the various poses they generate at various stages i.e. which are the most native-like and which are the least. Scoring functions

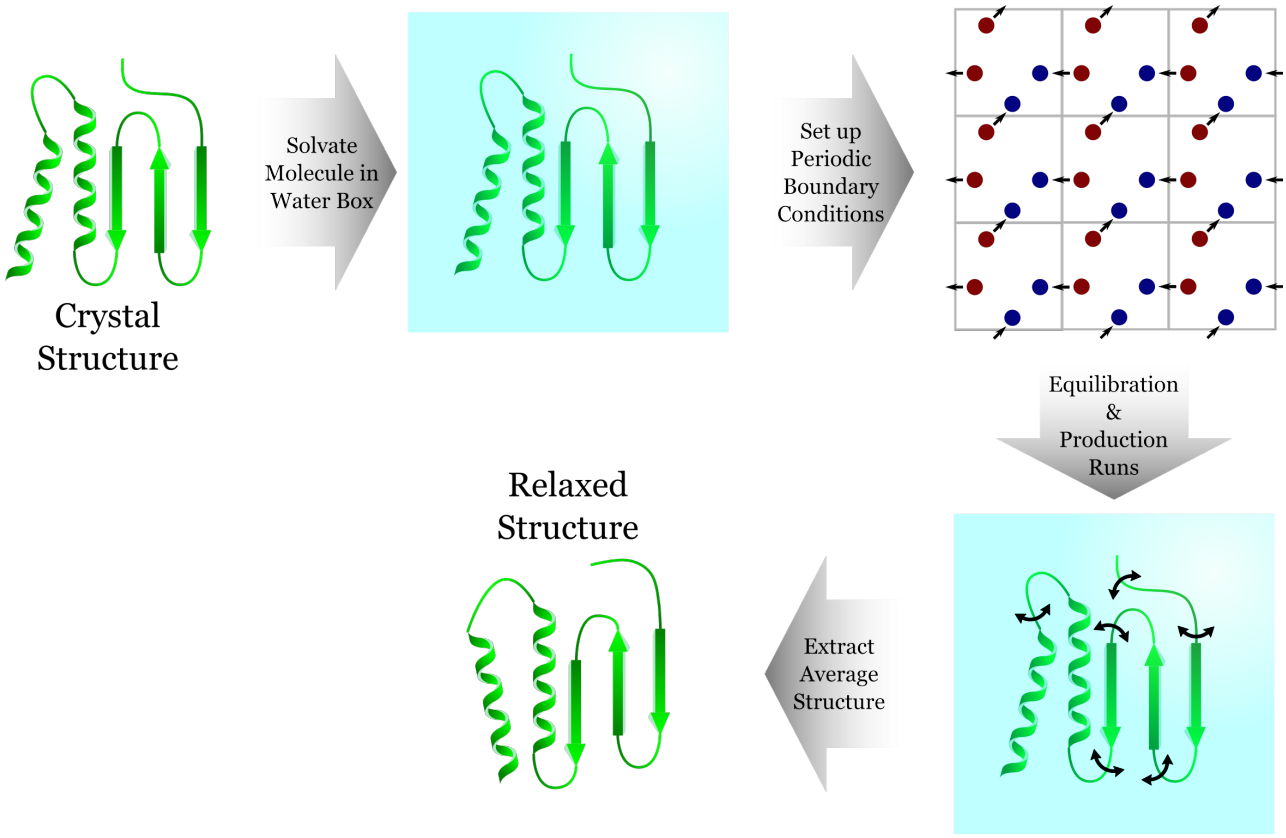


Figure 1.10: Molecular Dynamics simulation set up. The structure in question has any missing atoms modelled in and is then solvated in a water box of appropriate size and periodic boundary conditions put in place in order for the finite box to be treated like one of infinite size. An Equilibration run is then carried out to remove the effects of adding water, ions etc. around the protein. This is done over a number of time steps in order to prevent simulation instability. After equilibration, the Production run is then carried out for a length of time appropriate to the biological question, typically in the order of nanoseconds. In the case of obtaining a relaxed structure, the average RMSD of all the frames vs the starting frame is calculated and the frame with the lowest value extracted as the relaxed structure.

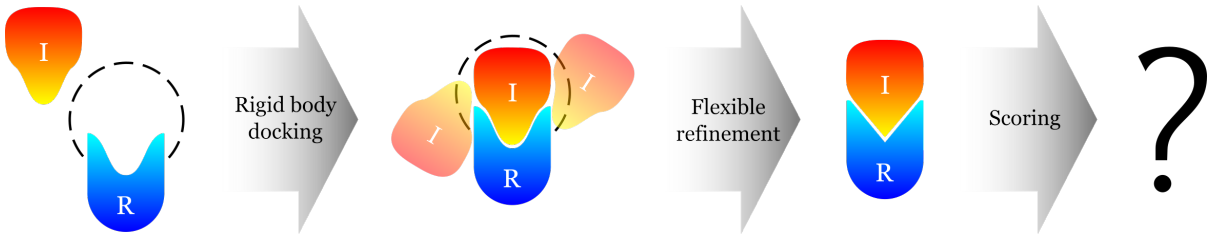


Figure 1.11: Overview of docking methodology. Docking programs attempt to orientate two (or more) unbound proteins, termed the receptor (blue, R) and the inhibitor (red, I), into their native complex conformation. Coarse-grained rigid body docking is first carried out by immobilising the receptor and manoeuvring the inhibitor around it in order to generate a large number of possible orientations (translucent representations). If any binding interface location information is already known (e.g. HDX-MS or other biochemical techniques), some docking programs can be biased towards the native structure by marking the surfaces predicted to interact (dotted circle). Flexible refinement is then carried out on a subset of the highest scoring rigid body docking poses. This involves perturbing side chain orientations in order to approximate the movements that occur in proteins upon binding. Each refined pose is then given a score with those with the highest score being theoretically the most native-like. The scoring metric is the area of greatest weakness within the field of docking.

generally take into account a number of different parameters such as steric complementarity, van der Waals interactions, electrostatic interactions, hydrogen bonds etc. Theoretically, those docking poses with the highest scores should be the most native-like and vice versa. However, if one compares these highest scoring poses with experimentally determined structures using RMSD as a metric, there is often little agreement between those with the highest score and those with the lowest RMSD. Therefore we can see that the scoring of docking poses is a critical weakness within the field, one that is being constantly improved upon through the testing of new methodologies using blind trials in the Critical Assessment of Predicted Interactions (CAPRI) meetings [59].

1.9.2 HADDOCK

The docking program used for this study was HADDOCK (High Ambiguity Driven protein-protein DOCKing) [57], an information-driven program that distinguishes itself from *ab initio* docking methods in that it encodes information from identified or predicted protein interfaces in the form of Ambiguous Interaction Restraints (AIRs) to drive the docking process. AIRs are a list of residues that have been identified as being part of the interaction surface which are then incorporated into HADDOCKs scoring method, thereby biasing generated poses towards those where the "marked" surfaces come together. Marked residues are defined as being either "active" or "passive". Active residues are those experimentally identified to be involved in the interaction between the two proteins and are also determined to be solvent accessible. Passive residues are all solvent accessible surface neighbours of the active residues. An AIR is defined as an ambiguous intermolecular distance between the two proteins where an active residue of one comes within a certain distance cut off of an active or a passive residue of the other. Using this system, passive residues can satisfy the restraints of the partner protein but cannot originate restraints themselves. Experimental HDX-MS data was used to define the AIRs for each protein-protein interaction, however it should be noted that the current implementation of AIRs within HADDOCK is a binary "on-off" classification, meaning that the complex shape of the HDX difference plot is lost.

HADDOCK's docking protocol consists of three stages. First is rigid body docking, followed by semi-flexible simulated annealing where the interface is perturbed, and finally refinement in an explicit solvent layer to improve energetics and therefore scoring. The build of HADDOCK used in this thesis was the webserver version of HADDOCK2.2 (`http://haddock.science.uu.nl/services/HADDOCK2.2/`) [58].

1.10 Methodological pipelines

1.10.1 Pipeline process for the prediction of protein structure

Now that individual component parts have been explained, an in-depth description of the pipeline as a whole will be given, as carried out on a single protein of a binary PPI. First, a candidate binary PPI with a solved complex structure must be selected, sourced and, if necessary, produced in-house. HDX-MS experiments for both the bound and unbound states are then carried out on each protein at a variety of time points and with appropriate controls to correct for extraneous exchange. Once these experiments have been deemed to have acquired data of sufficient quality, the computational steps can begin.

A set of decoy structures are generated using a tool such as Rosetta or 3DRobot and lnPs/RFU data is then simulated for each residue/peptide in each decoy in the data set in the unbound state using HDXsimulator. This data is then compared against the experimentally modelled lnPs calculated by HDXmodeller as well as the experimental RFU data determined by HDX-MS in order to generate residue-level as well as peptide-level RMSE metrics for that decoy. The three-dimensional coordinates of the atoms in the decoy itself can then be compared to the three-dimensional coordinates of the atoms in the protein's structure in order to generate an RMSD metric for that decoy. These two metrics can be visualised in a scatter plot of RMSE vs. RMSD (on both the lnP and RFU levels), where those poses with low RMSE values should also have low RMSD values. In a real-life application, the RMSD metric will not exist and poses will need to be ranked by RMSE alone. The purpose of having RMSD at this juncture is to allow the ability of RMSE to correctly distinguish between poses to be assessed and the underlying methodology behind the RMSE metric to be improved until it can correctly rank poses. The performance of RMSE can be judged through the construction of ROC curves which allow quantitative testing of how well RMSE can distinguish between native and non-native structures.

1.10.2 Pipeline process for the prediction of protein-protein complex structure

Complex selection and analysis by HDX-MS is as stated previously, after which the computational phase begins. The first stage is MD simulations to separate and relax the individual protein chains of the solved complex to generate pseudo-unbound structures. This is to prevent the complex structure from influencing subsequent steps, as when this methodology is applied to a real-life scenario, the structure of the complex will not exist. However, the structures of the individual monomers will need to exist, either solved experimentally or determined by the process described above. After relaxation, docking is then performed with surfaces identified to be involved in binding by HDX-MS marked so as to initially bias the docking towards the native structure using AIRs; initially rigid-body docking, followed by flexible refinement on the top scoring subset of poses. This is the extent to which we have developed this pipeline thus far. Future developments will see a similar approach to the one developed for individual protein's structures applied to the complexes in order to enable native complex structures to be distinguished from a background of non-native poses. These comparisons will make use of simulated vs. experimental difference plots in order to distinguish between poses, instead of lnP/RFU values.

Clearly the veracity of both these pipelines relies primarily on our ability to accurately calculate lnPs from structures. The current methodology, as described in the aforementioned paper [4], provides evidence that utilising RMSE values in this way has considerable merit, however there is room for significant improvement, which is in part as a result of the limited data set of proteins the method was initially trained against, as well as being bound by low resolution peptide-level data. Therefore, a broader depth and variety of training data sets and increased resolution to the residue-level enabled by HDXmodeller should accordingly increase its accuracy. Binary PPIs were chosen as the training data sets because of their ability to pull double duty for both pipelines as well as their high importance to cellular life as well as the comparative level of difficulty in benchmarking a new technique vs. relatively simple heterodimers compared to much more intricate multi-meric complexes. The principal binary PPI employed in this study was the barnase-barstar system with additional data sets collected for the GFP-

GFP nanobody system. It should be noted that, while the increased number of data sets presented here are a good start in the process of benchmarking this technique, it is likely that far more data sets than can be collected by one group will be required for this method to achieve its true potential. Therefore a long term goal is to open up the submission of test data sets to the HDX-MS research community at large and so allow benchmarking to occur with potentially hundreds of proteins.

1.11 Protein interactions investigated in this study

1.11.1 Barnase:Barstar

Barnase and barstar are two small proteins (12.4 kDa & 10.2 kDa respectively) that have been studied extensively since the 1960s as model proteins for folding studies. Produced by *Bacillus amyloliquefaciens*, barnase is an extracellular ribonuclease which functions as a weapon against competing bacteria in the local environment while barstar acts as its intracellular inhibitor [60]. As a ribonuclease, barnase catalyses the degradation of RNA and would therefore be lethal to the producing cell if not pacified in some way before secretion occurs. Thus in order to prevent apoptosis, *B. amyloliquefaciens* also produces barstar which binds extremely tightly ($K_d = 1 \times 10^{-14}$ M [61]) to barnase to form a heterodimer [62] that competitively inhibits barnase's RNase activity until it is safely secreted from the cell. This system was chosen to be the basis of our work because it has been extensively characterised in the literature, including a substantial amount of kinetics data covering both the wild type proteins as well as various mutant-wild type and mutant-mutant interactions which shows that significant kinetics changes are seen when key residues are mutated [61]. Furthermore, neither protein has any disulphide bonds which allows for easier digestion by proteases (and so greater peptide coverage), the overall fold of both proteins have been shown to be very resistant to mutations, nor does the interaction require any additional ligands [60].

For these reasons, we decided to test two different barnase-barstar HDX interaction profiles, the BnWT:BspWT interaction and the BnH102A:BsY29F mutant interaction. This provided two benchmarking data sets while enabling us to use essentially the same production, purification and mass spectrometry protocols for both. Kinetics of the interactions of barnase, barstar and several different mutants are thoroughly described in the literature [61], with the BnH102A:BsY29F interaction displaying a substantially higher K_d (3.1×10^{-9} M) than the BnWT:BspWT interaction. It was for this reason that these mutants were chosen for study because they could be used to report on the effects of K_d on the shape/magnitude of the HDX profiles generated for the barnase:barstar system.

1.11.2 GFP:GFP-nanobody/minimizer

Green Fluorescent Protein is perhaps one of the most famous proteins ever discovered. Thanks to its alluring glow, GFP has found uses all over the domain of biochemistry, particularly in the field of fluorescence microscopy [63] but also as a marker in reporter assays [64] as well as purely commercial applications such as the creation of glow-in-the-dark pets. GFP is a 26.9 kDa protein produced by the jellyfish *Aequorea victoria* and was discovered by Osamu Shimomura *et al.* in 1962 [65]; earning him and two collaborators the 2008 Nobel Prize in Chemistry. GFP-nanobody is a 13.8 kDa protein that forms a tightly bound ($K_d = 1.4 \times 10^{-9}$ M) 1:1 complex with GFP [66]. Nanobodies are antibody frag-

ments (also called single-domain antibodies) consisting of a single variable domain (VHH) cloned and isolated from heavy-chain only antibodies found in members of the *camelidae* family [67]. Nanobodies retain the full antigen binding capacity and specificity of their primogenitors as well as their stability; additionally they also benefit from the advantages of being relatively small in size such as easier transformation into bacterial cells for bulk production and (in a clinical setting) better permeability in tissues.

In addition to GFP-nb, the interaction of a second nanobody with GFP: GFP-nanobody minimizer was also investigated. GFP-nbmin is so called because upon binding it reduces the fluorescence of the GFP protein by approx. 75 % [68], and it also binds in a completely different orientation to the regular nanobody (although with some regional overlap) with a comparable K_d (4.5×10^{-9} M). Thus it was reasoned that the HDX profile of the GFP:GFP-nbmin interaction would show significant differences compared to the GFP:GFP-nb interaction and so make it a worthy addition to the benchmarking data set.

1.12 Studying the interaction of dUTPase with Stl

1.12.1 Overview

In addition to utilising HDX-MS for the unorthodox purpose of determining protein structure, HDX-MS was also used during this PhD for the far more traditional goal of informing on the location of a protein-ligand interaction in a biological system and so allowing inferences to be made about the system through analysis of the data. This system was the interaction of Stl with dUTPases from various different species and was a collaborative project between the Borysik group and the Vértessy group from the Budapest University of Technology and Economics.

Deoxyuridine triphosphate nucleotidohydrolase (dUTPase) is an enzyme involved in DNA synthesis where it acts to prevent the misincorporation of uracil into DNA [69]. Uracilation can cause various types of DNA damage such as double-strand breaks, deletion and chromosomal breaks [70] and so in healthy cells is prevented through the constant hydrolysis of deoxyuridine triphosphate (dUTP) (which can be readily incorporated into DNA by DNA polymerase) into deoxyuridine monophosphate (dUMP) via the enzyme dUTPase (Figure 1.12). dUMP can be further processed into deoxythymidine triphosphate (dTTP), allowing for the correct incorporation of thymine into DNA by DNA polymerase [71]. Therefore, dUTPase has a dual role both in keeping the dUTP pool low and the dTTP pool high in order to facilitate the correct functioning of DNA.

Staphylococcus pathogenicity island repressor protein (StlSaPIbov1 or simply Stl) is the master repressor of the highly mobile *Staphylococcus aureus* pathogenicity islands (SaPI), which play an important role in *S. aureus* toxicity [72]. Stl has previously been shown to interact with dUTPases from the φ11, 80α and φNM1 helper phage where complexation results in the disruption of the Stl-DNA interaction, allowing for the transcription of repressed SaPI genes [73]. Fortuitously, the binding of dUTPase with Stl also causes inhibition of dUTPase's enzymatic activity, with the dUTP substrate being unable to bind dUTPase in the presence of Stl. Interestingly, neither Stl nor dUTP can bind to dUTPase if the other is present first, indicating the involvement of the dUTPase active site in dUTPase:Stl complexation [74], where it is hypothesized that Stl can only bind when dUTPase is in an "open", substrate-free

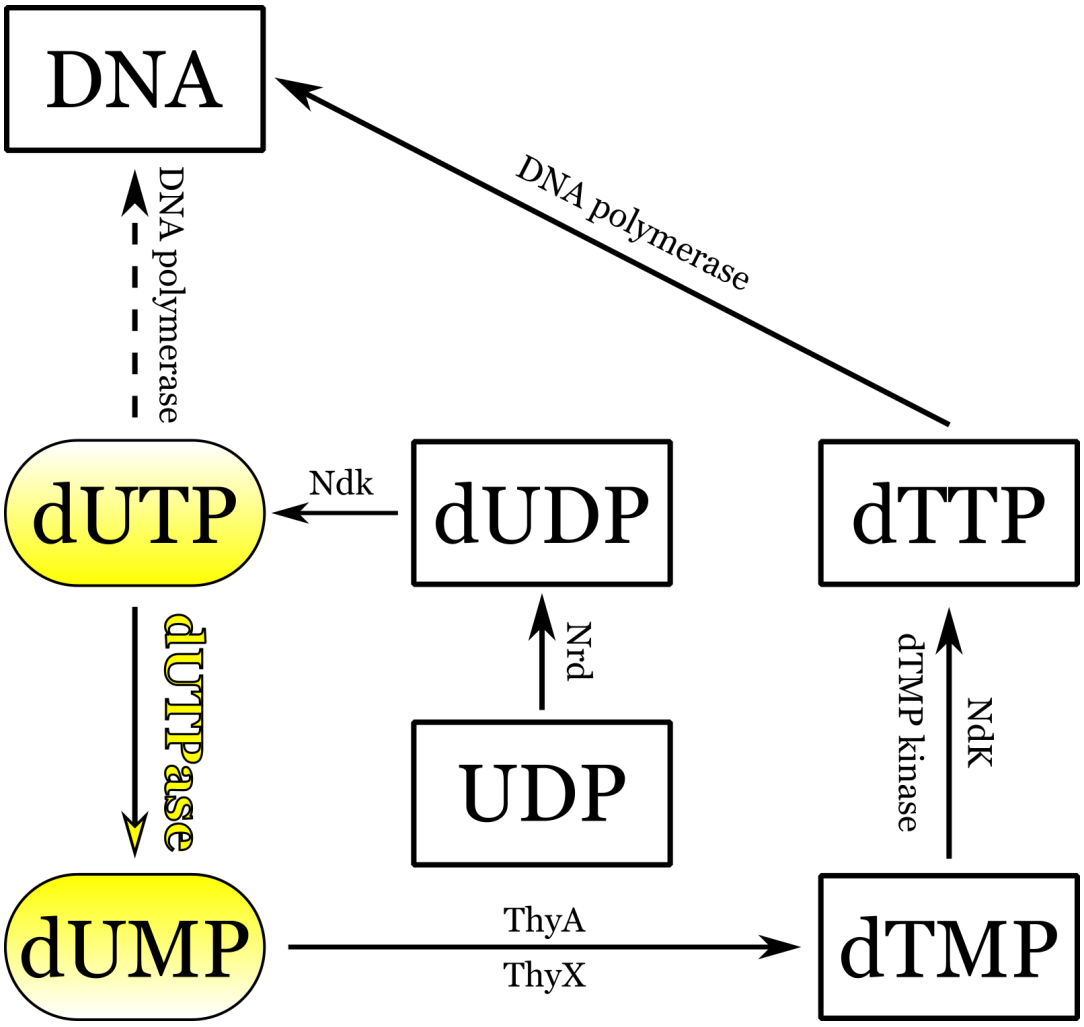


Figure 1.12: Role of dUTPase in preventing uracil misincorporation into DNA. Simplified diagram showing the steps involved in dTTP production from UDP, allowing for correct incorporation of thymine into DNA. Ancillary steps represented as square boxes, while steps impacted by dUTPase represented as ovals with a yellow highlight. Arrows represent chemical reactions; associated enzymes are named next to them. The enzyme dUTPase itself is highlighted in yellow. Dotted line represents the misincorporation of uracil into DNA which is prevented by the action of dUTPase converting dUTP into dUMP.

conformation. Stl has been shown to have quite substantial functional plasticity, with an inhibitory effect being documented on dUTPases from multiple different species, even those that share relatively little sequence or structural homology [73, 75–77].

The interaction between dUTPase and Stl is of particular interest to the scientific community because both proteins are involved in mechanisms which mediate disease. dUTPase participates both in terms of disease in humans, i.e. cancer, and disease-causing organisms, as its correct function is vital for their survival, [70] and Stl is directly responsible for the expression of toxic *S. aureus* genes as well as being involved in *S. aureus* resistance by playing a key role in horizontal gene transfer [78].

There are currently no crystal structures available for any dUTPase:Stl complex, hence the study of the interaction between these two proteins could benefit significantly from analysis by a technique that can provide information on where binding occurs, in terms of the peptides involved on both sides. The goal of this collaboration with the Vértessy group was therefore to study the interaction of Stl from *S. aureus* with dUTPases from a variety of different species using HDX-MS in order to better understand how the inhibition of dUTPase's catalytic function by Stl is mediated. The use of a technique that provides location data enabled the collection of hitherto unavailable information about the various dUTPase:Stl interactions and thus will contribute a substantial amount towards our understanding of this unusual system. For the purpose of this thesis, we will describe the two different facets of our collaboration with the Vértessy group that lead to published works: the first being the interaction of human dUTPase with Stl [79] and the second being the functional plasticity of Stl interacting with both homotrimeric and homodimeric dUTPases [77].

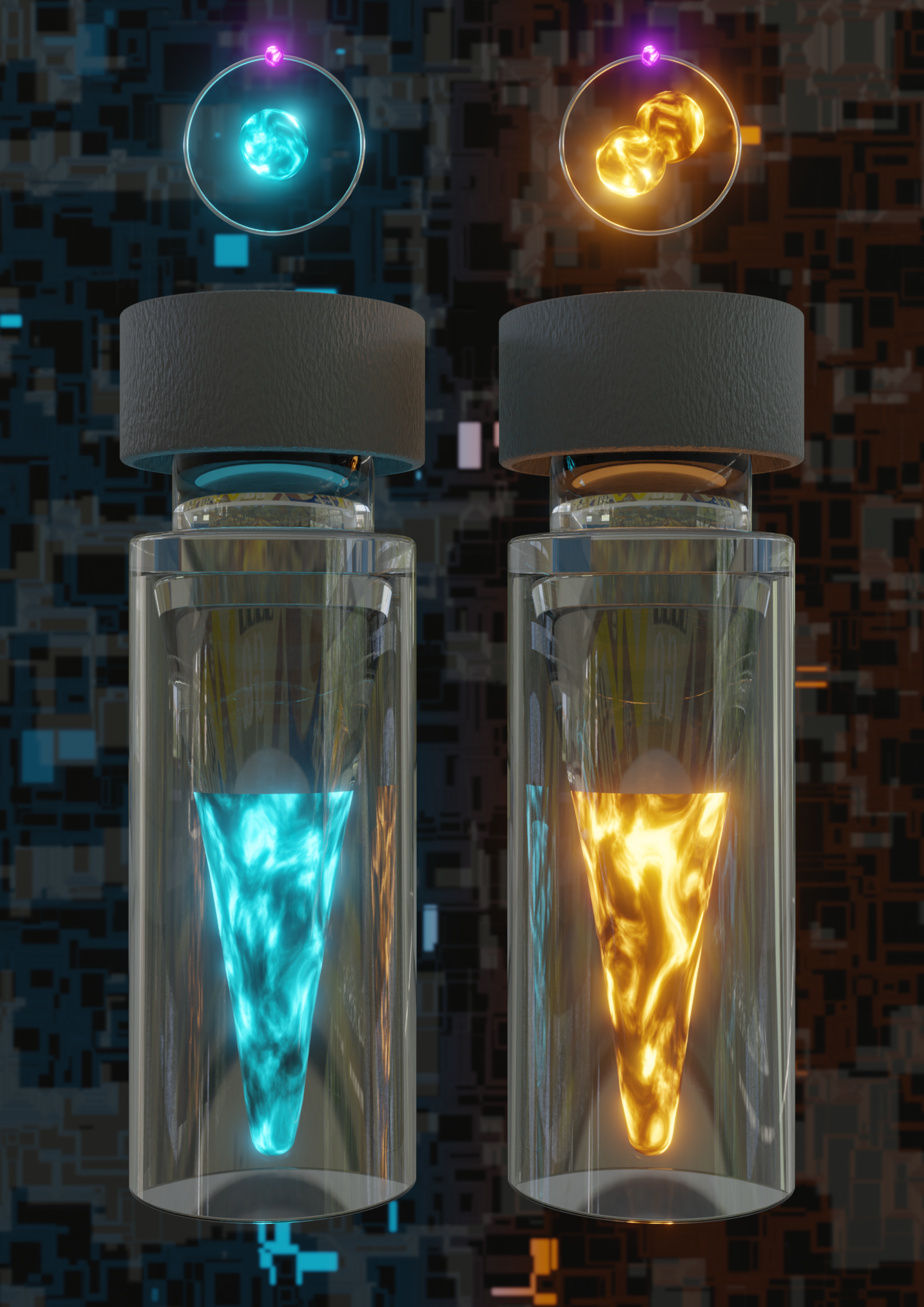
1.12.2 Using HDX-MS to investigate the human dUTPase:Stl interaction

Human dUTPase has for some time now been of interest to scientists because it acts as a survival factor for tumor cells and so has become a target for cancer chemotherapies [80]. In addition to small-molecule based approaches, biologics present an appealing way forward due to potential benefits such as increased selectivity and consequently a reduction in intracellular off-target effects. Therefore the discovery of Stl as a potent cross-species inhibitor of dUTPase has raised important questions about its potential viability as an anti-cancer drug. With the Vértessy group establishing that Stl could indeed bind to human dUTPase, our role in the collaboration became one of localisation: we ran standard HDX difference methodology to determine the regions responsible for the binding interaction, both in terms of human dUTPase and Stl. From this and a combination of additional biophysical techniques from other collaborators, such as size-exclusion chromatography in line with small-angle X-ray scattering (SEC-SAXS) restrained by HDX-MS data, a schematic model for human dUTPase:Stl complex assembly and Stl-DNA interaction was proposed. This model forms part of the bedrock of knowledge about this system that will be necessary if a biologic based on Stl is ever to be developed.

1.12.3 Using HDX-MS to investigate the functional plasticity of Stl

It has been found that Stl has the ability to inhibit enzymatic function of dUTPases from multiple species, even when those dUTPases share relatively little amounts of sequence and even structural homology, as explained previously. For example, Stl has been shown to inhibit representatives of both

distinct classes of dUTPases: the β-pleated homotrimeric and the all-α homodimeric dUTPases, despite these classes sharing essentially no structural similarity with only the ability to bind dUTP in common between them. The goal of this collaboration with the Vértessy group was to investigate the mechanism by which Stl can bind to and inhibit the function of these drastically dissimilar dUTPases by running HDX-MS difference experiments on a representative of each dUTPase class bound to Stl and then analyzing the data in order to accrue evidence for how this functional plasticity of Stl occurs. The proteins investigated in this study were: φ11 dUTPase from *S. aureus* phage representing the β-pleated homotrimeric dUTPases and φNM1 dUTPase from *S. aureus* phage representing the all-α homodimeric dUTPases, both in complex with Stl from *S. aureus*.



2 Methods

Unless otherwise stated, all reagents were purchased from Sigma-Aldrich or Thermo Fisher Scientific .

2.1 Production & purification of barnase and barstar

The plasmid of wild type barnase was provided by the Ikura group (Tokyo Medical and Dentistry University, Japan) in the pTZ416 vector under the control of the alkaline phosphatase promotor with a resistance to ampicillin. A plasmid for a barstar triple mutant (C40A, C82A & D39A) was also provided, however after numerous attempts to produce and purify a reasonable amount of protein from this plasmid resulted in failure (detailed in chapter 4.1), this plasmid was abandoned and a new plasmid was synthesised by Gene Universal (Newark, Delaware, USA). This plasmid (in pET-28a(+)) also contained the C40A & C82A mutations as well as an N-terminal 6His-tag with a TEV cleavage site to aid purification and was under the control of the lac promotor with a resistance to kanamycin. The C40A & C82A mutations are to eliminate a disulphide bond and the resultant structure is termed ”pseudo-Wild Type” due to its near-identical nature to the true Wild Type structure.

2.1.1 Barnase production & purification

Transformation of plasmid DNA into BL21-AI competent *E. coli* cells (Invitrogen, Inchinnan, UK) was carried out according to the “*E. coli* Competent Cells” protocol by Promega (Southampton, UK) (Appendix A). In order to amplify the amount of DNA in stock, single colonies were inoculated into 5 ml LB + 50 µg/ml ampicillin and incubated overnight at 37 °C with agitation at 220 rpm before being centrifuged at 4,400 rpm for 15 minutes. Cells were miniprepped according to the standard protocol for a Qiagen miniprep kit (Manchester, UK) (Appendix B). Resultant extracted DNA was flash frozen in liquid N_2 and stored at -20 °C. In order to make glycerol stocks out of the transformed cells for long term storage, single colonies were inoculated into 50 ml LB + 50 µg/ml ampicillin and incubated overnight at 37 °C with agitation at 220 rpm. Glycerol stocks were made by taking 500 µl of the overnight culture and adding it to 500 µl 50 % glycerol solution. The sample was mixed thoroughly, left to stand for 15 minutes and then flash frozen in liquid N_2 and stored at -70 °C.

For overexpression, the barnase glycerol stock was removed from the -70 °C freezer and placed on ice. The top was scrapped with an inoculating loop which was streaked onto an LB-agar plate containing 50 µg/ml ampicillin which was left in an incubator overnight at 37 °C. A single colony from the resultant grow was transferred via inoculating loop into 10 ml LB + 50 µg/ml ampicillin and left in an incubator overnight at 37 °C with agitation at 220 rpm. 6 ml of pre-culture was transferred into 1 L of minimal phosphate media (Appendix C) + 50 µg/ml ampicillin and incubated over night at 37 °C with agitation at 220 rpm. The expression culture was centrifuged at 4,200 rpm for 20 mins at 4 °C and the supernatant discarded. The cell pellet was re-suspended in 10 ml 50 mM Tris HCl pH 8 (final volume: 12.5 ml). 687 µl of glacial acetic acid was added and the solution left spinning in a rotator at 4 °C for 20 mins. The solution was then centrifuged at 4,000 rpm for 20 mins at 4 °C and the supernatant removed into a 3 kDa MWCO spin concentrator (Cytiva, Marlborough, MA, USA). The sample was centrifuged at 4,000 rpm at 4 °C until its volume reached approx. 5 ml. Sample removed from spin concentrator

and dialysed using a 3 kDa MWCO Slide-A-Lyzer dialysis cassette (Thermo Scientific, Waltham, MA, USA) vs. 500 ml 50 mM Tris HCl pH 8 at 4 °C with gentle stirring for 4 hours. The dialysis buffer was changed out for 500 ml fresh 50 mM Tris HCl pH 8 and left stirring overnight.

The sample was removed from the dialysis cassette and filtered through a 0.22 μm filter. Purification was undertaken on an ÄKTA Pure (Cytiva, Marlborough, MA, USA). A 5 ml HiTrap SP HP ion exchange column (Cytiva, Marlborough, MA, USA) was washed with 3 CVs dH_2O then 3 CVs 500 mM NaOH then 2 CVs 2 M NaCl then 3 CVs 50 mM Tris HCl pH 8. The sample was loaded into a 5ml loop and washed over the column with 3 CVs 50 mM Tris HCl pH 8 and a substantial peak at A280 nm was seen. The sample was eluted from the column using a linear gradient of 0-1 M NaCl over 10 mins at a flow rate of 1 ml/min and a fraction size of 1 ml. A very substantial peak at A280 nm was seen in fractions 8, 9 & 10, corresponding to a NaCl concentration of 0.8-1 M.

In order to test for purity, 7.5 μl of each fraction was combined with 2.5 μl gel loading dye and incubated in a dry bath for 10 mins at 70 °C. The samples were then loaded into a pre-cast NuPAGE 10 % Bis-Tris 1 mm gel (Invitrogen, Inchinnan, UK) and run at 200 V for 50 mins using 1 x MES SDS running buffer. The ladder used was PageRuler Plus (Thermo Scientific, Waltham, MA, USA). The gel was removed from the cassette and stained using InstantBlue (Sigma-Aldrich, Gillingham, UK) for approx. 1 hour before being imaged.

With sample purity confirmed, fractions 8-10 were combined and loaded into a 3 kDa MWCO Slide-A-Lyzer dialysis cassette and dialysed vs. 500 ml buffer E (5 mM potassium phosphate dibasic + 5 mM potassium phosphate monobasic, pH 7) overnight. The dialysis buffer was changed out for fresh 500 ml buffer E and left dialysing for 5 hours. The sample was extracted from the cassette and its concentration determined vs. a buffer E blank, following which its concentration was lowered to 40 μM by diluting with buffer E. Barnase sample aliquoted, flash frozen in liquid N_2 and stored at -70 °C.

In order to manufacture the barnase H102A mutant, forward and reverse primers were synthesised by Invitrogen (Inchinnan, UK) with the following sequences:
Forward – TGGCTGATTTACAAAACAACGGACGCTTATCAGACCTTTACAAAAATCAG
Reverse – CTGATTTTTGTAAAGGTCTGATAAGCGTCCGTTGTTTTGTAAATCAGCCA
Upon arrival, lyophilized DNA was solubilised in 1 ml H_2O, their concentration determined and then appropriately diluted with H_2O to reach approx. 125 ng/μl. Mutagenesis was carried out as stated in the "QuikChange II Site-Directed Mutagenesis Kit" protocol by Promega (Southampton, UK) (Appendix. D). Resultant BnH102A colonies were pre-cultured, miniprepped, transformed into BL21-AI, overexpressed, purified and stored as stated previously for the WT barnase. The only difference to the above is that only 100 ml of minimal phosphate media was used for the overexpression because the yield of BnH102A was found to be far greater than that of the WT.

2.1.2 Barstar production & purification

Transformation of plasmid DNA into BL21-DE3(pLysS) competent *E. coli* cells was carried out according to the "*E. coli* Competent Cells" protocol by Promega (Appendix A) (Southampton, UK). In

order to amplify the amount of DNA in stock, single colonies were inoculated into 5 ml LB + 50 µg/ml kanamycin & 34 µg/ml chloramphenicol and incubated overnight at 37 °C with agitation at 220 rpm before being centrifuged at 4,400 rpm for 15 minutes. Cells were miniprepped according to the standard protocol for a Qiagen miniprep kit (Appendix B) (Manchester, UK). Resultant extracted DNA was flash frozen in liquid N_2 and stored at -20 °C. In order to make glycerol stocks of the transformed cells for long term storage, single colonies were inoculated into 50 ml LB + 50 µg/ml kanamycin & 34 µg/ml chloramphenicol and incubated overnight at 37 °C with agitation at 220 rpm. Glycerol stocks were made by taking 500 µl of the overnight culture and adding it to 500 µl 50 % glycerol solution. The sample was mixed thoroughly, left to stand for 15 minutes and then flash frozen in liquid N_2 and stored at -70 °C.

For overexpression, the barstar glycerol stock was removed from the -70 °C freezer and placed on ice. The top was scrapped with an inoculating loop which was streaked onto an LB-agar plate containing 50 µg/ml kanamycin & 34 µg/ml chloramphenicol which was left in an incubator overnight at 37 °C. A single colony from the resultant grow was transferred via inoculating loop into 10 ml LB + 50 µg/ml kanamycin & 34 µg/ml chloramphenicol and left in an incubator overnight at 37 °C with agitation at 220 rpm. 200 µl of the pre-culture was inoculated into 200 ml 2xYT media + 50 µg/ml kanamycin & 34 µg/ml chloramphenicol and incubated at 37 °C with agitation at 220 rpm until the OD reached approx. 0.6. Protein expression was induced with 1 mM IPTG and the culture was incubated overnight at 37 °C with agitation at 220 rpm.

The overnight culture was decanted into centrifuge tubes and centrifuged at 4,000 rpm for 20 mins at 4 °C. The supernatant was discarded and the cell pellet re-suspended in 20 ml of lysis buffer (50 ml PBS 1x, 35 µl β-Mercaptoethanol (10 mM), 1 EDTA-free protease inhibitor tablet & 2 µl benzonase) after which the cells were lysed using a cell disruptor at 25 kPsi for 1 cycle. Cell lysate was clarified by centrifugation at 20,000 rpm for 30 mins at 4 °C. The supernatant was collected and spiked with 4 M NaCl and 1 M imidazole in order to bring their concentration in the supernatant up to 300 mM & 25 mM respectively. Purification was undertaken on an ÄKTA Pure. The supernatant was filtered through a 0.22 µm syringe filter and loaded into a 50 ml Superloop. A 1 ml HisTrap (Cytiva, Marlborough, MA, USA) column was equilibrated with 5 CVs Equilibration buffer (50 mM Tris-HCl pH 8, 300 mM NaCl & 25 mM imidazole) and the clarified cell lysate washed over it at 1 ml/min until the Superloop was empty. A substantial peak at A280 nm was seen at this time, representing the non-binding proteins. The barstar was eluted with Elution buffer (50 mM Tris-HCl pH 8, 300 mM NaCl & 300 mM imidazole) using a linear gradient from 0-100 % over 10 mins at 1 ml/min with a fraction size of 1 ml. A substantial peak was seen at A280 nm in fractions 6-10 (barstar and any other binding proteins). Fractions 6-10 were combined and concentrated down to 2.5 ml using a 3 kDa MWCO spin concentrator at 4,000 rpm at 4 °C. A PD-10 desalting column (Cytiva, Marlborough, MA, USA) was equilibrated with 25 ml 50 mM Tris-HCl pH 8, after which the combined purified fractions were added and centrifuged at 1,000 x g for 2 mins at 4 °C. Eluent removed into a fresh 3 kDa MWCO spin concentrator and centrifuged at 4,000 rpm at 4 °C until the total sample volume was approx. 750 µl. The His-tag cleavage reaction using AcTEV protease (Invitrogen, Inchinnan, UK) was set up as stated in Appendix E. The cleavage reaction was left at room temperature for 40 hours with gentle mixing on a rotator.

The cleavage reaction was recovered and diluted with Equilibration buffer to a total volume of 5 ml and filtered through a 0.22 µm syringe filter, after which it was loaded into a 5 ml loop. The 1 ml HisTrap column was re-equilibrated with 5 CVs Equilibration buffer and the sample washed over it with 15 ml Equilibration buffer at 1 ml/min with 1 ml fractions collected. A substantial peak at A280 nm was seen at this time (cleaved barstar and any other non-binding proteins). The remainder of un-cleaved barstar as well as AcTEV protease and any other binding proteins was eluted isocratically with 5 CVs Elution buffer. The flow through containing cleaved barstar and other non-binding proteins was concentrated down to 500 µl using a 3kDa MWCO spin concentrator at 4,000 rpm and 4 °C, whereupon it was filtered through a 0.22 µm syringe filter and loaded into a 500 µl loop. A Superdex 75/300 GL Size Exclusion (SEC) column (Cytiva, Marlborough, MA, USA) was equilibrated with 2 CVs 50 mM Tris-HCl pH 8 at a flow rate of 0.8 ml/min, after which the sample was run over the column with 1.5 CVs 50 mM Tris-HCl pH 8 at 0.8 ml/min with the eluent being collected in 1 ml fraction sizes. The first 0.3 CVs of eluent were not collected and a peak at A280 nm corresponding to barstar was seen in fraction 7. A gel to determine identity and purity of the barstar sample was carried out as described previously. Barstar's yield was determined as previously described with its concentration subsequently adjusted to 40 µM, aliquoted, snap frozen in liquid N_2 and stored at -70 °C.

Using this protocol, the majority of the barstar was un-cleaved by AcTEV and it was clear that a substantial amount of work would be needed in order to boost yields to acceptable levels. At this point time was starting to run out and so it was decided to contract out the remaining barstar production to Ruth Rose of the Protein Production Facility at Queen Mary University of London while we focused on collecting HDX data. The plasmid as well as the production and purification protocols we had developed up until this point would be used as a starting point for the Protein Production Facility to iterate upon until a large amount of cleaved barstar could be produced

2.2 HDX-MS experiments

2.2.1 Overview of generic experimental and analytical HDX-MS setup

HDX-MS experiments were all performed on a Synapt G2Si HDMS in tandem with an Acquity UPLC M-Class system with HDX and automation (Waters Corporation, Manchester, UK) and a LEAP PAL autosampler (Trajan Scientific Europe Ltd, Milton Keynes, UK) for sample management. The mass spectrometer was calibrated against NaI and sample data acquired with lock-mass correction using Leu-enkephalin every 30 seconds.
The basic experimental procedure is as follows:
5 µl of protein sample at 10-20 µM is mixed with 95 µl equilibration buffer/labelling buffer for reference files/labelling files respectively at 20 °C. After the appropriate incubation time has elapsed, 70 µl of sample is transferred into 70 µl quench at 1 °C to retard further deuteration. 50 µl of the sample is then digested on-line by a Waters Enzymate BEH pepsin column at 20 °C and the subsequent peptides immobilised on a Waters BEH C18 VanGuard pre-column for 3 minutes at a flow rate of 200 µl/min in buffer A (H_2O + 0.1 % formic acid, pH 2.5). Peptides are then eluted by use of a linear gradient of organic solvent, buffer B (acetonitrile + 0.1 % formic acid, pH 2.5), from 8-40 % over 6 minutes and then separated by UPLC using a Waters BEH C-18 analytical column before being transferred into the mass

spectrometer by ESI. All trapping and chromatography is performed at 0 °C to minimise back exchange of deuterium for hydrogen. Within the mass spectrometer, the sample first encounters a Step Wave ion guide to remove any neutral contaminants, a quadrupole to (in our experimental procedure) focus the ion beam and an IMS cell to gather drift time data. Lastly, the sample enters the ToF mass analyser for m/z separation and determination. Each sample is run with an 11 minute acquisition time with the majority of peptides eluting between 2 and 8 minutes. This experimental set up can be visualized in Figure 2.1.
Following on from each sample run is a clean blank which runs a saw-tooth gradient of buffer B from 8-85 % and back again over 4 minutes and repeated a second time in order to eliminate carry-over into the next sample run.

Peptides contained within the reference files were assigned using the ProteinLynx Global Server (PLGS) v3.0.2 (Waters Corporation, Manchester, UK) software in order to generate ion accounting files containing a list of all the peptides that could be found in the sample run. The deuterium uptake (ΔD) of each peptide at each time point in each state was subsequently determined with DynamX v3.0.0 (Waters Corporation, Manchester, UK) by calculating the difference between the centroids of the mass spectral envelopes of the labelled samples and the reference samples (Figure 2.2). Centroids were calculated using Equation 2.1:

$$centroid = \frac{\Sigma m_i I_i}{\Sigma I_i} \tag{2.1}$$

Where m_i denotes the m/z of peak i and I_i denotes the intensity of peak i.

In order to calculate the uptake difference between states, data for all time points in the bound state were summed and subtracted from the summed data of all the time points in the unbound state to form a peptide-level difference plot. This peptide-level data was further processed using a custom MATLAB script in order to display the total mass shift of each peptide plotted against the residue position to generate "Woods plots" showing the Δmass of each peptide on the residue scale comparing the bound to the unbound states. An overview of this methodology can be seen in Figure. 2.3.

Those residues which displayed a significant amount of deuterium uptake difference were determined by calculating Confidence Interval (CI) values. First, the Mean Squared Deviation (MSD) of the uptake was calculated across all peptides for each time point using Equation 2.2:

$$MSD_t = \langle SD_{i,t} \rangle^2 \tag{2.2}$$

Where MSD_t denotes the Mean Squared Deviation at time point t and $SD_{i,t}$ denotes the uptake Standard Deviation for peptide i at time point t. Then, the Standard Error of the Mean (SEM) was calculated from the MSD values using Equation 2.3:

$$SEM = \frac{\sqrt{\Sigma MSD_t}}{\sqrt{N}} \tag{2.3}$$

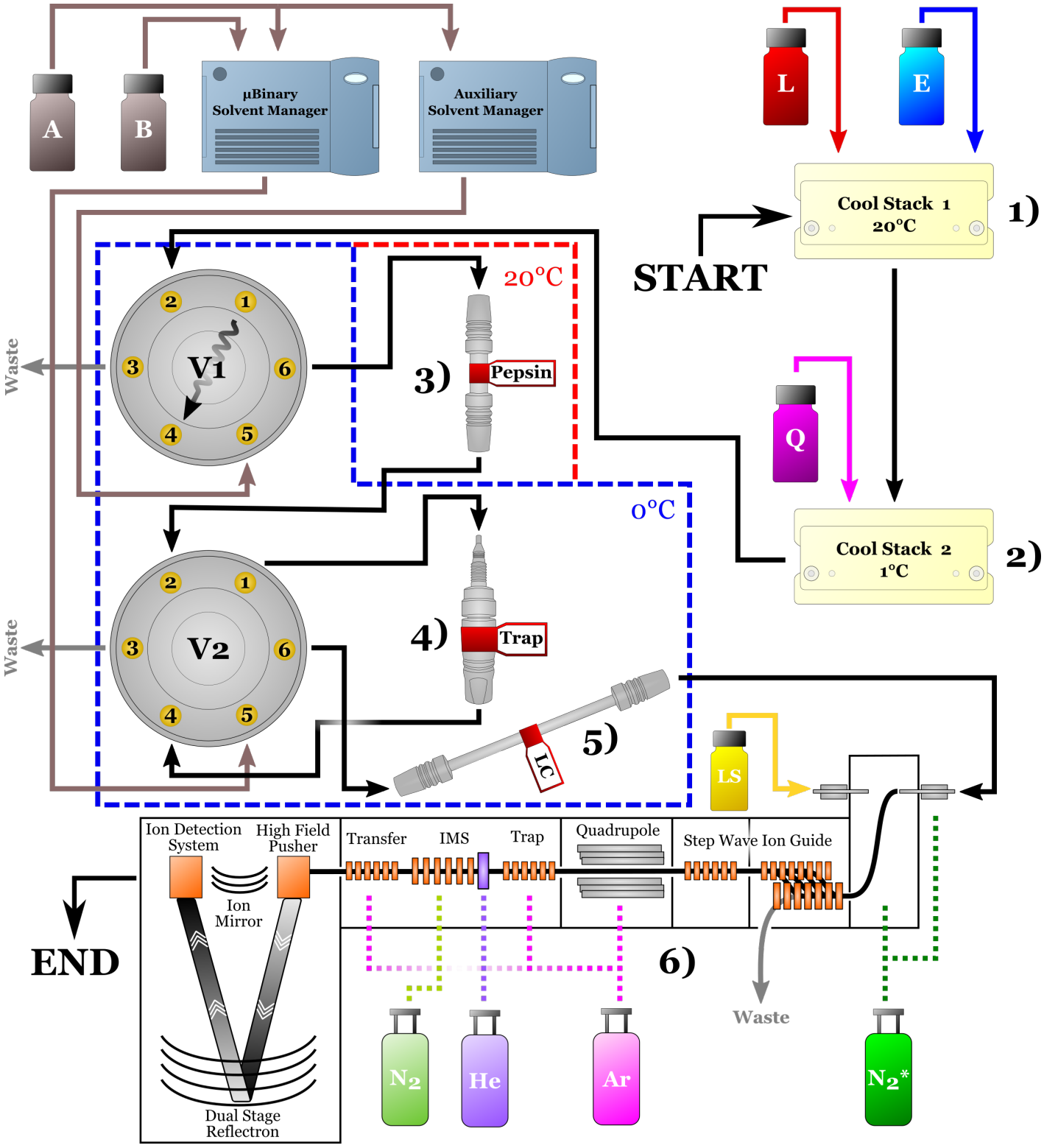

A
B
µBinary Solvent Manager
Auxiliary Solvent Manager
L
E
Cool Stack 1 20°C
1)
START
20°C
Waste
V1
3)
Pepsin
Q
0°C
Cool Stack 2 1°C
2)
V2
4)
Trap
5)
LC
LS
Ion Detection System
High Field Pusher
Ion Mirror
Dual Stage Reflectron
Transfer
IMS
Trap
Quadrupole
Step Wave Ion Guide
END
6)
Waste
N2
He
Ar
N2*

Figure 2.1: Overview of the experimental set up for the mass spectrometer. Flow path of the sample is represented as a solid black line; other solid coloured lines represent the flow of different solvents. Dotted lines represent the flow of gasses. Dashed lines encompass an environment with a particular controlled temperature (displayed as a number with the same colour). Solvents are represented as coloured vials, gasses as coloured gas cylinders. The mass spectrometer is represented as a cross section to show the flow path as well as constituent parts. All other parts of the experimental set up are displayed as a labelled diagrammatic facsimile of the real objects. **(1)** Samples are incubated with either buffer E (blue, E) or buffer L (red, L) as appropriate for a set amount of time. **(2)** Samples are then quenched with buffer Q (purple, Q) to retard further exchange. **(3)** Transfer into the pepsin column via valve 1 (V1) then occurs to digest the samples into peptides. **(4)** Peptides move through valve 2 (V2) and are immobilised on the trap column until they are eluted by an increasing percentage of organic solvent (brown, B) vs. inorganic solvent (brown, A). **(5)** Eluted peptides are separated by UPLC and enter into the mass spectrometer by ESI. **(6)** The ionized sample passes through a Step Wave ion guide to remove neutral contaminants, a quadrupole to (in our experimental procedure) focus the ion beam and an IMS cell to gather drift time data. m/z information for the ions is then determined by the ToF mass analyser. LockMass correction is obtained using LockSpray (yellow, LS).

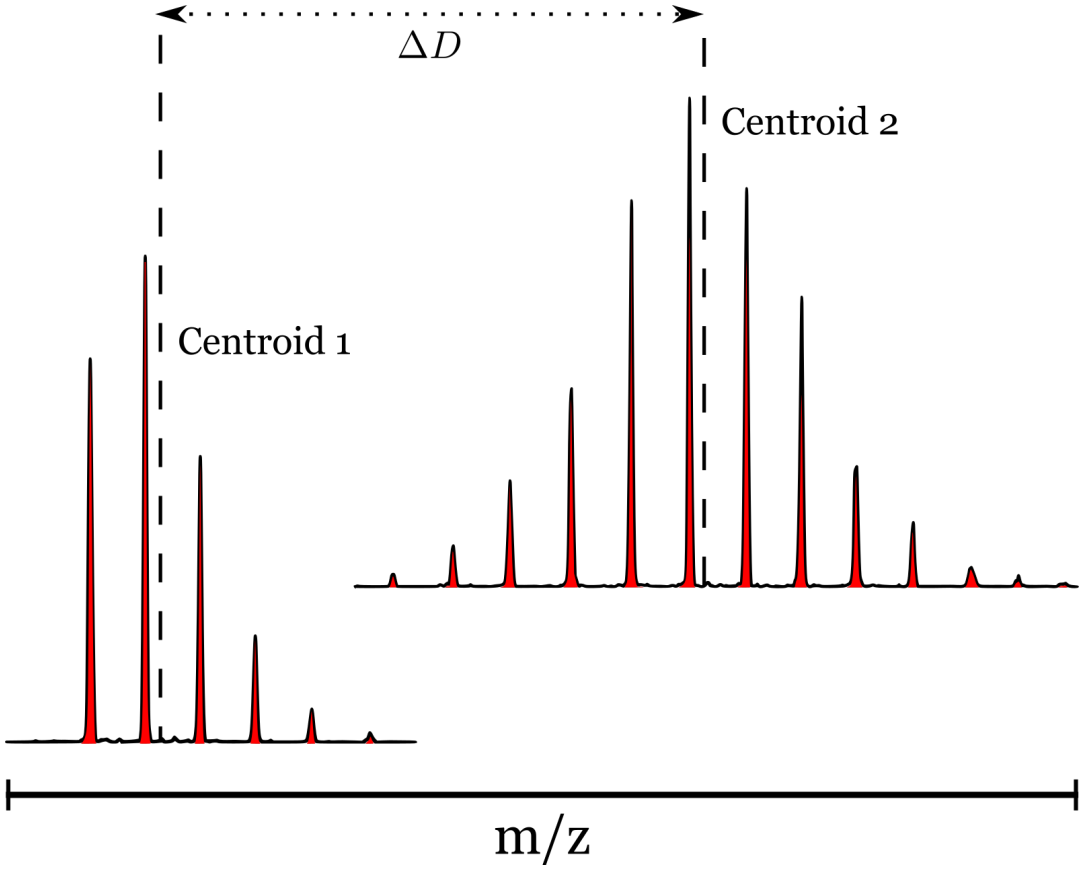


Figure 2.2: Calculation of deuterium uptake from centroids. Visual representation of how the change in deuterium uptake (ΔD) can be determined by calculating the difference between the centroids of two isotopic envelopes.

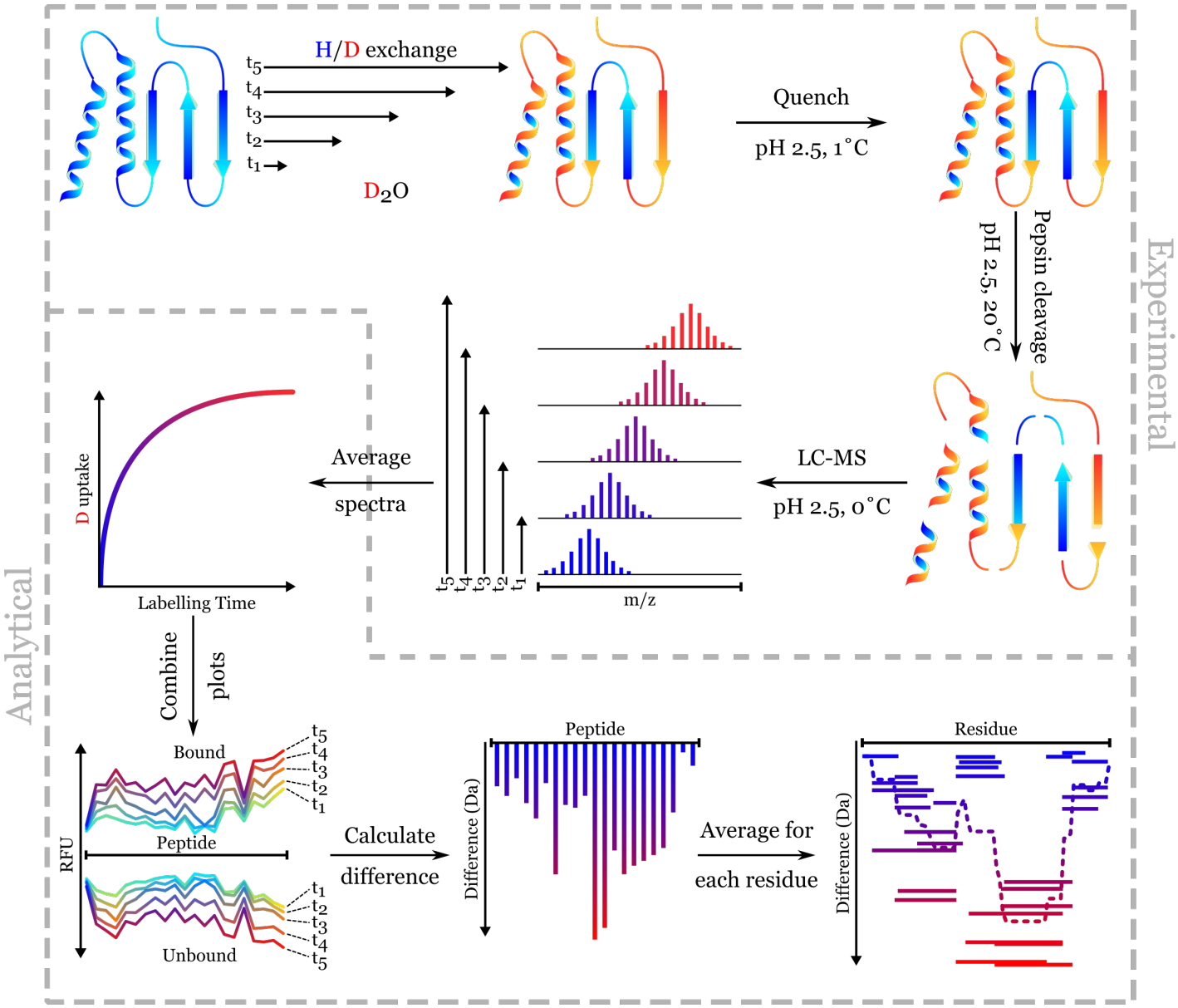


Figure 2.3: Overview of HDX-MS data collection & analysis pipeline. In all individual elements, blue represents a state of protonation and red represents a state of deuteration. In the experimental portion, a protein suspended in aqueous buffer is diluted for varying time points (t1-5) in deuterated buffer, causing deuteration to occur according to each residue's k_{HX}. The sample is quenched in ice-cold buffer Q to retard further in/back-exchange and then digested into peptides on-line with immobilised pepsin. Peptides are separated by LC and MS data is then collected for each peptide at each time point, typically causing the peptide's Gaussian distribution to shift to higher m/z values as deuteration time increases. In the analytical portion, each peptide's uptake "fingerprint" is determined by graphing its deuterium uptake as a function of labelling time. Plots for all peptides are then combined into a single "butterfly" plot graphing each peptide's RFU at each time point in both the bound and the unbound states. Bound RFU values are subtracted from unbound RFU values to construct a "difference" plot which is then averaged per residue to form a "Woods" plot. The advantage of Woods plots vs. difference plots is that they allow HDX differences to be seen on the residue level instead of just the peptide level and so allow for higher resolution data interpretation.

Where N denotes the number of replicates in the experimental data set. In the case of all our experiments, this number was 3.
Finally, the confidence interval values in Da were calculated by multiplying the SEM value by the appropriate value for the CI t-test at 2 degrees of freedom using Equation 2.4:

$$CI_p = SEM \cdot z_p \tag{2.4}$$

Where CI_p denotes the Confidence Interval at percent threshold p and z_p denotes the critical value at percent threshold p e.g. 9.925 for 99 % CI, 4.303 for 95 % CI etc. These calculations generate a threshold value in Da, everything above which (positive or negative) was thought of as "significant" uptake difference and everything below which was thought of as "not significant".

2.2.2 HDX-MS experiments on proteins involved in binary PPIs

Samples of GFP, GFP-nb & GFP-nb min were provided by Rebecca Beavil from the Randall Division of Cell and Molecular Biophysics, King's College London. Samples of barstar were provided by Ruth Rose from the School of Biological and Chemical Sciences, Queen Mary University of London. Barnase was produced in-house. All protein samples were diluted to 40 μM using the buffer they were provided in, aliquoted, snap frozen in liquid N_2 and stored at -70 °C. BEX were set up for each protein by loading a single aliquot (<1 ml) into a 3 kDa MWCO Slide-A-Lyzer dialysis cassette and then dialysing it against 100 ml buffer L (4.5 mM K_2HPO_4, 4.5 mM KH_2PO_4 in D_2O, pD 7) overnight at room temperature with gentle stirring. Samples were extracted from the cassette, filtered through a 0.22 μm syringe filter and incubated at 37 °C for between 2-3 weeks. Samples were removed from the incubator, aliquoted, snap frozen in liquid N_2 and stored at -70 °C.

Four unique binary PPIs were investigated in this thesis: BnWT:BspWT, BnH102A:BsY29F, GFP:GFP-nb & GFP:GFP-nbmin. Reference files with a 1:1 mixture of the two proteins were gathered for each interaction in sextuplicate along with six labelling time points: 15 seconds, 1 minute, 5 minutes, 25 minutes, 2 hours & 8 hours collected in triplicate for both the bound and unbound states. These specific time points were chosen so that uptake data could be gathered over an (approximately) logarithmic timescale. BEX and IEX were also gathered in triplicate. IEX are set up exactly the same as a reference file except the quench is made up in D_2O.

HDX-MS data was collected replicate-by-replicate i.e. an entire single data set including references, labelling time points and controls is collected before looping back to the start and collecting the next set of replicates. The experiments were set up like this in order to account for any changes which might occur during the 36 hours it took to run all the data sets as each reference, time point and control would have one of its replicates taken in the first 12 hours, one in the second 12 hours and one in the final 12 hours. In order to collect data for a single interaction, two of these 36 hour runs needed to be completed; one in which protein A was the unbound data set and one in which protein B was the unbound data set.

2.3 Native MS experiments

Native MS experiments were carried out on the Synapt G2Si HDMS calibrated against CsI with samples introduced via a NanoLockSpray Exact Mass Ionization Source (Waters Corporation, Manchester, UK). The instrument was run with positive polarity in sensitivity mode. Desalting of samples took place before native MS was carried out in order to reduce background noise. Two different desalting methods were used: Bio-spin desalting columns (Bio-Rad, Watford, UK) and Vivaspin desalting columns (Cytiva, Marlborough, MA, USA).
For Bio-spins, each were primed prior to an experiment by first centrifuging them for 2 mins at 1,000 x g to remove storage solution, followed by 5x 1 min centrifugations at 1,000 x g with 500 µl of appropriate buffer each time and discarding the flow through. Desalting was achieved by applying 50 µl of sample to the centre of the Bio-spin and centrifuging for 4 mins at 1,000 x g, after which the flow though was collected.
For Vivaspins, a column with an appropriate MWCO was chosen and then washed with 100 µl of appropriate buffer for 2 mins at 15,000 x g. 50 µl of sample was then added and the column topped up with 300 µl appropriate buffer before being centrifuged at 15,000 x g until approximately half of the initial volume remained. This step was repeated at least 4 more times. On the last desalting cycle, the column was centrifuged until the remaining volume was equal to that of the initial starting volume, at which point the sample was recovered.

Capillaries were made in-house with a Flaming/Brown P-97 micropipette puller (Sutter Instruments, Novato, CA, USA) and coated with Au:Pd (80:20) using a Quorum Q150RS sputter coater (Lewes, UK). Before use, each needle had its sealed tip snapped off using tweezers in order to allow sample flow through it. 3 µl sample was loaded into the needle and placed into the ion source with a high initial pusher gas flow in order to encourage initial sample flow, which was subsequently reduced as low as possible while still maintaining spectra intensity. Exact machine parameters used for data acquisition were tailored on a sample-by-sample basis depending on the behaviour of the spectra. Peak identity was confirmed using the webserver version of ESIprot (`https://www.bioprocess.org/esiprot/esiprotform.php`) [81].

2.4 Production of protein-protein complex poses

2.4.1 Molecular Dynamics simulations

Initial bound structures were downloaded from the Protein Data Bank (PDB) [82] (`http://www.rcsb.org/`) database with the following accession codes: BnWT:BspWT (1BRS), GFP:GFP-nb (3OGO), GFP:GFP-nbmin (3G9A). Missing atoms were modelled in using Prime [83, 84] and the protein preparation wizard interface in Maestro (Schrödinger LLC, New York, USA). Structures were then submitted as individual chains and as a complex to CHARMM-GUI's Solution Builder tool (`http://www.charmm-gui.org/?doc=input/solution`) in order to generate a series of input files for relaxational MD simulations of the proteins in both their bound and unbound conformations. Input files were generated using the following parameters:
Step 1 – Protein Solution System: using the structures downloaded and modelled previously.
Step 2 – Waterbox Size Options: fit waterbox size to protein size, waterbox type = rectangular, enter

edge distance = 10.0; Add Ions: 0.15 M KCl, ion placing method = Monte-Carlo.
Step 3 – Periodic Boundary Condition Options: generate grid information for PME FFT automatically.
Step 4 – Force Field Options: CHARMM36m, Input Generation Options: NAMD, Equilibration Input Generation Options: NVT ensemble, Dynamics Input Generation Options: NPT ensemble, temperature = 300 °K.

Input files for each structure to be simulated were prepared in the above manner and downloaded from CHARMM-GUI. Equilibration and Production runs were then prepared and executed from the input files using a CUDA-enabled NAMD v2.13 module installed on the research computing facility at King's College London, Rosalind (`https://rosalind.kcl.ac.uk`), utilising a Tesla V100 GPU (NVIDIA, Santa Clara, CA, USA).
Equilibration simulations were run with the following parameters:
Temperature = 300 °K;
Energy output = 125 steps;
Trajectory output = 1,000 steps;
Force Field Parameters – non-bonded exclusion policy = scaled1-4, 1-4scaling = 1.0, cutoff = 12.0 Å, switch distance = 10.0 Å, pair list distance = 16.0 Å, steps per cycle = 20, pair lists per cycle = 2;
Integrator Parameters – timestep = 2.0 fs, rigid bonds = all, non bonded frequency = 1, full electrostatic frequency = 1;
Constant Temperature Control – reassign frequency = 500 steps, reassign temperature = 300 °K;
Periodic Boundary Conditions – wrap water = on, wrap all = on, wrap nearest = off;
Particle Mesh Ewald – PME = yes, PME interpolation order = 6, PME grid spacing = 1.0;
Pressure and Volume Control – use group pressure = yes, use flexible cell = no, use constant ratio = no, langevin = on, langevin damping = 1.0, langevin temperature = 300 °K, langevin hydrogen = off;
Constant Pressure – langevin piston = on, langevin piston target = 1.01325, langevin piston period = 50.0, langevin piston decay = 25.0, langevin piston temperature = 300 °K;
Minimize = 10,000 steps; Run = 25,000 steps.
The system was Equilibrated for 50 ps to remove the effects of adding water, ions etc. around the protein.

Production simulations were run with the following parameters:
Temperature = 300 °K;
Energy output = 50,000 steps;
Trajectory output = 5,000 steps;
Force Field Parameters – non-bonded exclusion policy = scaled1-4, 1-4scaling = 1.0, cutoff = 12.0 Å, switch distance = 10.0 Å, pair list distance = 16.0 Å, steps per cycle = 20, pair lists per cycle = 2;
Integrator Parameters – timestep = 2.0 fs, rigid bonds = all; non bonded frequency = 1, full electrostatic frequency = 1;
Periodic Boundary Conditions – wrap water = on, wrap all = on, wrap nearest = off;
Particle Mesh Ewald – PME = yes, PME interpolation order = 6, PME grid spacing = 1.0;
Constant Pressure Control – use group pressure = yes, use flexible cell = no, use constant ratio = no, langevin piston = on, langevin piston target = 1.01325, langevin piston period = 50.0, langevin piston decay = 25.0, langevin piston temperature = 300 °K;

Constant Temperature Control – langevin = on, langevin damping = 1.0, langevin temperature = 300 °K, langevin hydrogen = off; Run = (5,000,000 – 25,000,000 steps).
The main Production step to produce a relaxed structure ran for between 10-50 ns, depending on how long it took before it was qualitatively judged that the system in question had reached equilibrium.

Upon completion of the Production run, trajectories were imported into VMD v1.9.3 [85] (Theoretical and Computational Biophysics Group at the Beckman Institute for Advanced Science and Technology of the University of Illinois at Urbana-Champaign) and the RMSDs of each frame (as compared to the first frame, backbone only) determined using the RMSD Trajectory Tool. Once it was qualitatively judged that the RMSDs had plateaued off, the system was considered to have reached equilibrium and the average frame (as determined by the RMSD Trajectory Tool, backbone only) was extracted to be used as the relaxed structure.

2.4.2 Protein-protein docking

Relaxed MD structures were used as the starting point for docking using the webserver version of HADDOCK. The Expert interface (`https://milou.science.uu.nl/services/HADDOCK2.2/haddockserver-expert.html`) was used as it provided a balance between customizability and ease of use.
Parameters selected for the docking procedures were as follows:
First Molecule – Structure Definition: relaxed chain A of complex; Restraint Definition: active residues = all those residues identified by HDX-MS to have a level of uptake decrease surpassing the 99 % CI, passive residues = define passive residues automatically around the active residues; Histidine Protonation States = automatically guess histidine protonation states using molprobity; Semi-flexible Segments = automatic.
Second Molecule – Structure Definition: relaxed chain B of complex; Restraint Definition: active residues = all those residues identified by HDX-MS to have a level of uptake decrease surpassing the 99 % CI, passive residues = define passive residues automatically around the active residues; Histidine Protonation States = automatically guess histidine protonation states using molprobity; Semi-flexible Segments = automatic.
Distance Restraints – radius (in Å) around active residues to automatically define passive residues = 6.5, remove non-polar hydrogens? = yes, randomly exclude a fraction of the ambiguous restraints (AIRs) = yes, number of partitions for random exclusion (%excluded=100/number of partitions) = 2.0.
Sampling Parameters – number of structures for rigid body docking = varied between 1,000-5,000, number of trials for rigid body minimisation = 5, sample 180 degrees rotated solutions during rigid body EM = yes, number of structures for semi-flexible refinement = varied between 200-1,000, solvent to use for the last iteration = water, number of structures for the explicit solvent refinement = varied between 200-1,000, epsilon constant for the electrostatic energy term = 10.0.
Parameters for Clustering – clustering method = RMSD, RMSD cut off = 7.5 Å, minimum cluster size = 4.

The RMSDs of the resultant docking poses vs. the crystal structure was then determined using a custom RMSD-calculating script based on the Kabsch algorithm method [86] that compared the atom-to-atom RMSDs of each pose vs. the crystal structure, allowing an RMSD to be assigned to each pose.

Given that the original crystal structures had undergone several processing steps in a variety of different programs in order to produce the docked complexes, the atom number and order no longer matched up exactly with the crystal structure and so several steps were needed to achieve parity again. First, the crystal structure and the top scoring docking pose were opened in UCSF Chimera [87] (Resource for Biocomputing, Visualization and Informatics, University of California, USA) and the models spilt into their respective individual chains. Next, the Matchmaker tool was used to align chain A of the docking pose with chain A of the crystal structure and chain B of the docking pose with chain B of the crystal structure. This then created a pseudo-crystal structure out of the top scoring docking pose with an almost identical fold to the crystal structure but with the atom ordering the same as the rest of the docking poses. The individual chains of the pseudo-crystal structure were then combined and exported as a .pdb file. Unfortunately, exporting .pdb files from UCSF Chimera causes a number of changes to the file which would need to be corrected or else the RMSD-calculating script would not work. These were accomplished using the tool PDBEditor (`http://www.bioinformatics.org/pdbeditor/wiki/`) which enabled us to reset the atom numbering from 1. One final edit that needed to be made was to delete any "TER" lines that had appeared as they contained duplicate atom numberings to the preceding atom. With these edits made, the pseudo-crystal structure now had the identical atom orderings and numberings as the rest of the docking poses. Poses with an RMSD value of $\leq$ 2.5 Å were considered to be native.

2.5 Obtaining residue-resolved lnPs using HDXmodeller

Experimental RFU data obtained by HDX-MS was used to calculate residue-resolved lnPs using HDXmodeller (`https://hdxsite.nms.kcl.ac.uk/Modeller`). First, we used the tool "k-intrinsic" (`https://hdxsite.nms.kcl.ac.uk/kintrinsic`) to calculate k_{int} values for each residue in the amino acid sequence of our proteins, using a temperature value of 293.15 K and a pD value of 7.0. Next, we calculated residue-resolved lnP values using HDXmodeller. Experimental RFUs determined by HDX-MS along with calculated k_{int} values were used as input files and ran using 50 replications, 1,000 iterations, an accuracy of $1e^{-6}$, a temperature of 293.15 K and a pD of 7.0 as input parameters. With residue-level lnPs calculated for all amino acids in the data set, we then used the tool "Occupier" (`https://hdxsite.nms.kcl.ac.uk/Occupier`) to calculate the occupancy of all peptides within our experimental HDX data sets in order to identify likely weakly constrained peptides. We then used this information to enable subsections within each data set to be created by deleting certain weakly constrained peptides, therefore creating distinct boarders between subsections of the proteins which had no bridging peptides and so no influence from other subsections. This information was combined with the residue-resolved lnPs to evaluate our confidence in the lnPs' accuracy using the tool "R-evaluator" (`https://hdxsite.nms.kcl.ac.uk/Revaluator`) to generate auto-validation R-matrix scores for the demarcated subsections as well as the protein as a whole. To generate a score for individual subsections, the start and end residue numbers of that subsection were used as input parameters.

2.6 Using HDXsimulator to classify protein structures

2.6.1 Generating protein decoy sets

Two different *ab initio* folding methods were used to generate decoys for use by HDXsimulator. The first was to use Rosetta's Abinitio protein folding program and the second was to use 3DRobot. For Rosetta, Robetta's Fragment Server (`http://old.robetta.org/fragmentsubmit.jsp`) was first used in order to generate 3-mer and 9-mer fragment files from the protein sequences. These were then combined with the original FASTA file and the native structure as input files upon which Rosetta's Abinitio program was run. 1,000 structures were generated by Rosetta, each of which had an RMS score comparing it to the original native structure. As this method of structure generation was *ab initio*, there were few native structures present and so the proportion of native structures needed to be enriched in order for HDXsimulator to work effectively. This was achieved using Rosetta's Relax application which allows for all-atom refinement of structures. The Abinitio structure with the lowest RMS score was selected and considered as the "pseudo-crystal" structure from this point onwards. This pseudo-crystal structure was run through Relax and 100 refinements generated, each differing from the pseudo-crystal structure by no more than 2 Å. These 100 structures were then ranked by using the same custom RMSD-calculating script described previously. Because this script compared atom-to-atom, the reference structure and the comparison structure had to have the exact same number of atoms in the exact same order, hence the need for "pseudo-crystal" structures as the actual crystal structures differed substantially in both respects and so would not work, even with substantial modification. The highest 20 structures by RMSD were selected for use as the native data set. Therefore, using these two methods, 1,000 mostly non-native and 20 native structures were generated.

In comparison to Rosetta, 3DRobot generates decoys from the input native structure itself. The web server version of 3DRobot (`https://zhanglab.ccmb.med.umich.edu/3DRobot/`) was used to generate 1,000 decoys with an RMSD cut-off of 20 Å. Again, the structure with the lowest RMSD score compared to the native structure was considered the "pseudo-crystal" structure for the rest of the pipeline. The primary difference between the decoys sets generated by these two methods is one of distribution: those generated by Rosetta have 20 near-native structures and then a large gap between them and the non-native structures, whereas those decoy sets generated by 3DRobot have a constant distribution between the pseudo-crystal and the least native structure. These decoys also had their RMSDs vs. the pseudo-crystal structure determined as previously described.

2.6.2 Exploring the boundaries of modelling protein conformation using HDXsimulator

The first implementation of the HDXsimulator pipeline was done mostly manually and, as we knew the process of testing and producing data sets would require lots of repetition, we would need a certain degree of automation. Therefore, in order to expedite the process of running multiple experiments in quick succession, a pipeline process was developed on a local Linux workstation involving the use of multiple Python programs and Bash scripts. The pipeline is described in depth here and illustrated in Figure 2.4. Select code is available in appendices. The first step was to calculate the RMSD of each decoy vs. the pseudo-crystal structure as described previously. Next, HDXsimulator was run on the pseudo-crystal structure, generating lnP values for each residue except prolines and the N-terminal

residue and RFU values for each peptide (as defined by an experimental peptide list). The residue-level pseudo-crystal lnP values were extracted and used as input for later programs. HDXsimulator was then run on all the decoys in the data set with the exception of the pseudo-crystal structure with the data stored for later use.

Next came the generation of synthetic lnP errors with which the efficacy of the HDXsimulator pipeline could be tested. This was done by using HDXsimulator itself to generate erroneous lnPs from the pseudo-crystal structure by varying the scaling factors βC and βH (Equation 1.10) to be different than the optimal values as determined by Best & Vendruscolo [30]. These scaling factors determined the cut-off distances for contacts and hydrogen bonds respectively to be considered for any given residue with its neighbouring residues. We generated these erroneous data sets by assigning a range within which the values of βC (default: 0.35) and βH (default: 2) could vary of 0.2-0.5 & 1-3 respectively and then used a pseudo-random number generator to assign a value to use for each repetition of the calculation. With these new varying scaling factors, HDXsimulator was run on the pseudo-crystal structure 1,000 times in order to produce a full suite of lnP error files for use in the pipeline. This code is available in Appendix F. After error generation, the RMSE as well as the R^2 of each synthetic error lnP replicate vs. the original pseudo-crystal lnP using default parameters was calculated and the values stored for subsequent analysis.

Each synthetic error lnP replicate was then used to calculate a theoretical k_{obs} value for each residue using Equation 2.5:

$$k_{obs} = \frac{k_{int}}{exp(lnP)} \tag{2.5}$$

Where k_{obs} is the theoretical observed rate constant for each residue, k_{int} is the intrinsic exchange rate for each residue as calculated by the tool k-intrinsic and lnP is the protection factor for each residue. This code is available in Appendix G. These k_{obs} values were then combined with the experimental peptide list and k_{int} values in order to calculate theoretical RFU values for each of the original synthetic error lnP replicates. This process was under the control of a Bash script which feeds each k_{obs} file in turn into the RFU-calculating program before running it and saving the output.

After all RFU values have been calculated, the last step of the pipeline was to generate ROC curves in order to calculate AUC values for each set of erroneous data, for both peptide-level RFU and residue-level lnP. This was accompanied by the automated construction of histograms & scatter plots in order to allow data analysis. The histograms enabled the spread of AUC values produced by a dataset to be visualised more easily, as well as the construction of scatter plots comparing AUC values with both RMSE and R^2 values on both the RFU as well as lnP levels.

2.6.3 Classifying native structures using HDXsimulator

Appropriate elements of the aforementioned pipeline were reused in order to automate the acquisition of experimental data as much as possible. All decoy sets used here were the same as those used previously. RMSD values of the decoys vs. the pseudo-crystal structure were calculated as previously

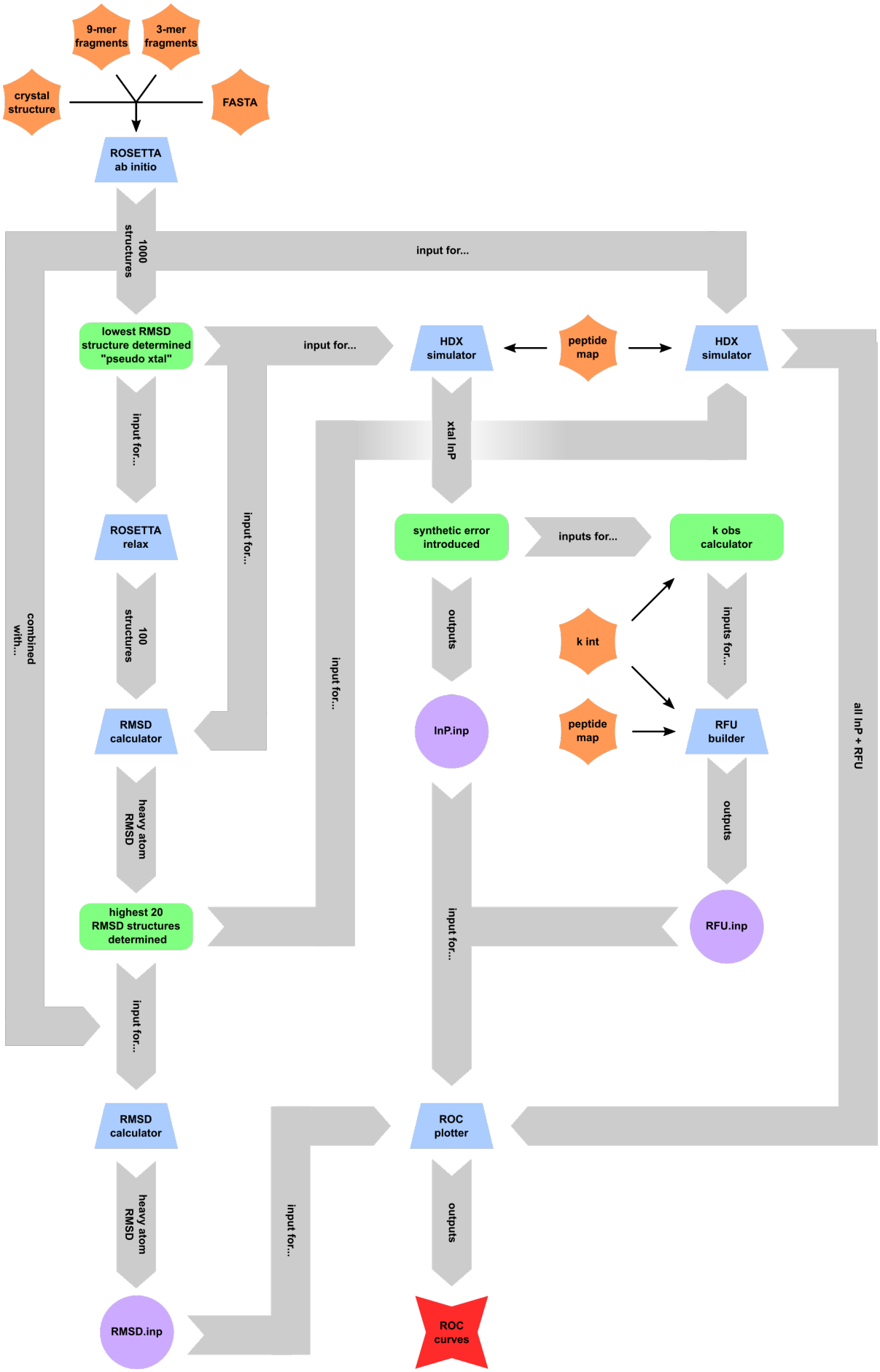

9-mer fragments
3-mer fragments
crystal structure
FASTA
ROSETTA ab initio
1000 structures
input for...
lowest RMSD structure determined "pseudo xtal"
input for...
HDX simulator
peptide map
HDX simulator
input for...
xtal InP
ROSETTA relax
input for...
synthetic error introduced
inputs for...
k obs calculator
100 structures
outputs
k int
inputs for...
combined with...
RMSD calculator
InP.inp
peptide map
RFU builder
all InP + RFU
heavy atom RMSD
input for...
outputs
highest 20 RMSD structures determined
input for...
input for...
RFU.inp
input for...
RMSD calculator
ROC plotter
heavy atom RMSD
input for...
outputs
RMSD.inp
ROC curves

Figure 2.4: Steps involved in the exploration of HDXsimulator. Schematic diagram detailing the various steps and calculations involved in the HDXsimulator exploration pipeline, as coded and controlled by Python and Bash scripts. Orange hexagons represent input files generated outside the pipeline, blue trapeziums represent steps developed by others, green rectangles represent steps developed by the author, purple circles represent major outputs of various branches of the pipeline that go on to serve as inputs for the final step, red star, the generation of ROC curves detailing the sensitivity of HDXsimulator to error. Grey arrows detail the movement of the pipeline from one stage to the next.

described and HDXsimulator was run on the decoy sets (using default parameters) in order to generate calculated lnP/RFU values for each residue/peptide of each decoy. These calculated values were then compared against experimental lnP from HDXmodeller and experimental RFU from our HDX-MS experiments in order to generate an RMSE metric on the residue/peptide-level for each decoy. These RMSE values were compared against the decoy's RMSD to generate a scatter plot. ROC plots were then generated based on these scatter plots, assuming a native cut off RMSD of ≤ 2.5 Å, in order to evaluate how effective RMSE was at correctly estimating native vs. non-native decoys. High AUC values indicated that structures were able to be accurately classified as either native or non-native. Conversely, low AUC values indicated structures were not able to be accurately classified as either native or non-native.

2.7 Experimental HDX-MS methods used for the dUTPase:Stl system

All dUTPase and Stl samples were obtained from the Vértessy group from the Budapest University of Technology and Economics. Samples were received frozen in dry ice and transferred to a -70 °C freezer for long term storage. Samples had the following stock concentrations: φ11 dUTPase (φ11DUT) – 277 µM, human dUTPase (hDUT) – 882 µM, φNM1 dUTPase (φNM1DUT) – 241 µM, Stl – 33 µM. All proteins were solubilised in a buffer containing 20 mM HEPES, 300 mM NaCl, 5 mM $MgCl_2$, pH 7.5.

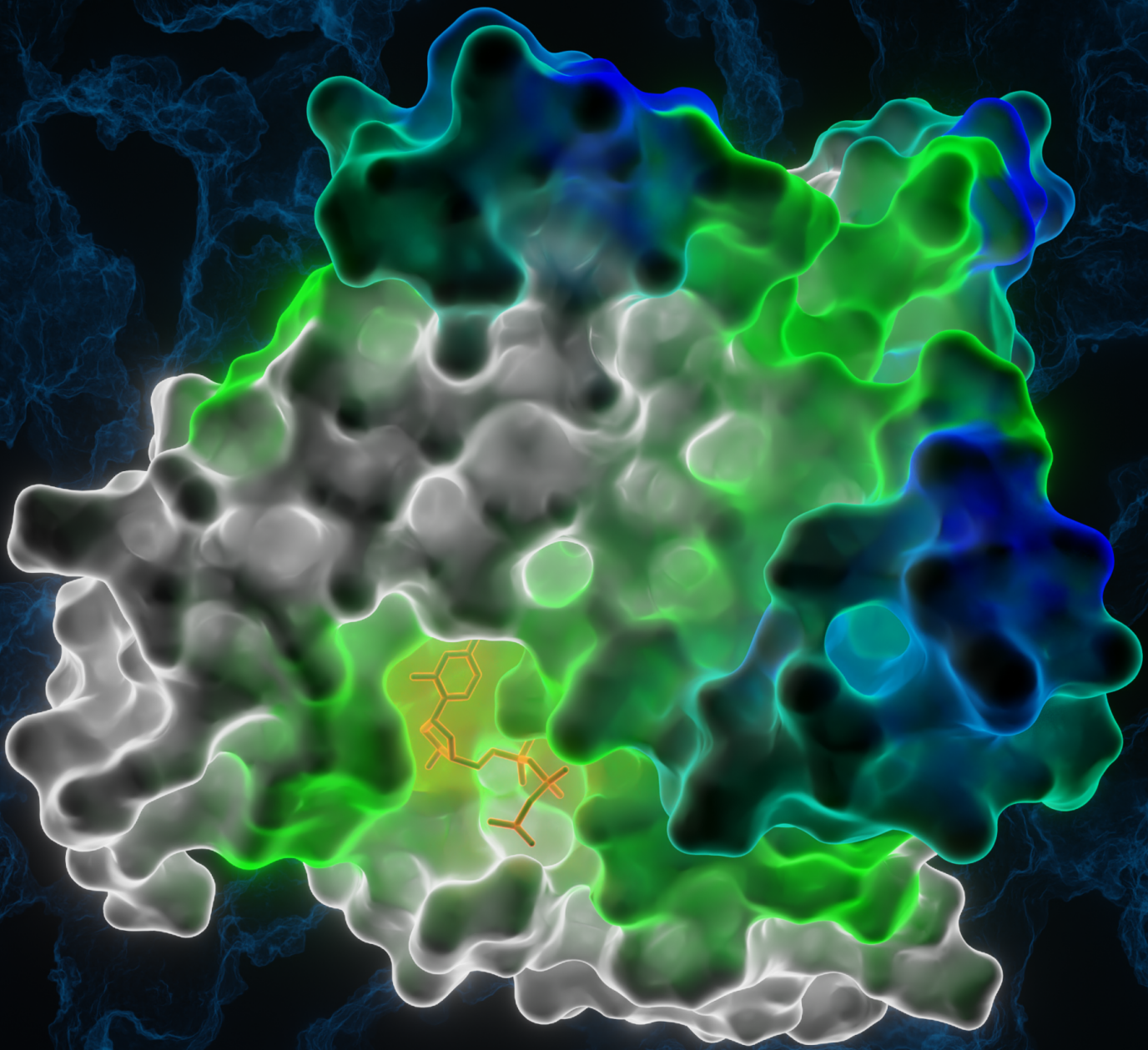
2.7.1 HDX-MS experimental set up for dUTPase:Stl

Each interaction of one dUTPase with Stl was broken down into two separate experimental runs. The first was to collect bound and unbound data for the dUTPase being studied, the second was to collect bound and unbound data for Stl. Each experimental run was set up identically, with un-deuterated reference files collected on a 1:1 mixture of the dUTPase and Stl in sextuplicate and 3 different labelling time points (1, 10 & 100 minutes) collected for both the bound and unbound protein states in triplicate. Bound data sets were collected with the relative concentration ratios of 1.2:1 in favour of the protein not currently being investigated. This was to ensure that the protein under investigation was fully saturated with substrate and so unbound species should be eliminated. Unbound data sets were collected on a 1:1 mixture of the protein currently being investigated and its buffer in order to make the concentration of the protein in the unbound data set approximately match that of the protein in the bound data set. Acquisitions were run consecutively in blocks i.e. all 6 reference files collected followed by all 3 unbound 1 minute files etc.

2.7.2 HDX-MS experimental and analytical procedure for dUTPase:Stl

HDX-MS experiments were performed and analysed as described in chapter 2.2. Buffer E contained 20 mM HEPES, 300 mM NaCl, 5 mM $MgCl_2$, pH 7.5 in H_2O, buffer L contained 20 mM HEPES, 300 mM NaCl, 5 mM $MgCl_2$, pD 7.5 in D_2O and the quench contained 2.4 % formic acid. Residues with mass difference values greater than the 99 % CI threshold (95 % CI for the hDUT data set) were marked and carried forward into the final analytical step: superimposing uptake difference data upon the three-dimensional structures of the proteins in order to elucidate binding and allosteric information. Data visualisation was achieved using the "Define Attributes" tool in UCSF Chimera with residues showing significant negative uptake difference values assigned a colour gradient from green (less significant) to blue (more significant). Significant positive uptake difference values were assigned a colour gradient

from orange (less significant) to red (more significant). Residues which showed no significance were represented as grey.



3 Results of the investigation into the interaction of dUTPase with Stl

In addition to utilising HDX-MS for the unorthodox purpose of determining protein structure, HDX-MS was also used during the course of this PhD for the far more traditional goal of informing on the location of a protein-ligand interaction in a biological system and so allowing inferences to be made about the system through analysis of the data. This system was the interaction of the proteinaceous inhibitor Stl with dUTPases from various different species and was a collaborative project between the Borysik group and the Vértessy group from the Budapest University of Technology and Economics. In the course of this project, we investigated two different aspects of the interaction: that of hDUT with Stl and the differences between trimeric and dimeric dUTPases from *S. aureus* with Stl. The goal of this work was, in the case of hDUT:Stl, to provide key novel structural insights that pave the way for further applications of the first discovered potent proteinaceous inhibitor of hDUT, which acts as a survival factor for tumour cells and is therefore a target for cancer chemotherapy. In the case of dimeric/trimeric dUTPase:Stl, the goal was to characterise the interactions between these proteins based on a range of biochemical and biophysical methods (particularly HDX) and shed light on the binding mechanism of the dimeric φNM1DUT and Stl.

3.1 The human dUTPase:Stl interaction

Our part in this collaborative effort to characterise the interaction of hDUT with Stl was to gather evidence on the location of the interaction. Sequence coverage for both proteins after digestion was high at 95.7 % & 94.0 % for hDUT and Stl respectively with an average peptide-per-amino acid redundancy of >2 in both cases. Peptide-level difference plots show a clear negative mass shift characteristic of a binding interface in both proteins, with a sum maxima of approx. -2 Da in the case of hDUT and approx. -14 Da in the case of Stl, localised to specific peptide regions. Woods plots were generated from this peptide-level data in which the location and mass shift of each peptide was presented on the primary sequence of the protein in question, enabling insights at the residue level to be made. In the case of hDUT, the largest negative Δmass can be found in the region of H34-L50 with other significant Δmass seen in the region of the C-terminus. However, these C-terminal mass differences converge more rapidly then those at the N-terminus, implying a weaker interaction. In addition, residues A89-G110 also show modest negative Δmass, suggesting a role in the interaction.

In the case of Stl, the difference plots show a very significant negative Δmass localised to a limited number of peptides with more minor changes seen across the whole of the protein. Peptides 21-24 display the most significant mass shifts with a clear delineation on either side. When mapped out over the amino acid sequence of Stl, these peptides correspond to residues Y98-Y113. In addition, minor differences in isotope uptake can be seen in most other peptides, indicating that Stl may undergo a global decrease in dynamics upon binding to hDUT. These results indicate that although binding in Stl is localised to a very specific region, the interaction is propagated across the entire rest of the protein. When these results are mapped onto the three-dimensional structures of hDUT and Stl, we can start to visualise the interaction much more easily. All these results can be seen in Figure 3.1.

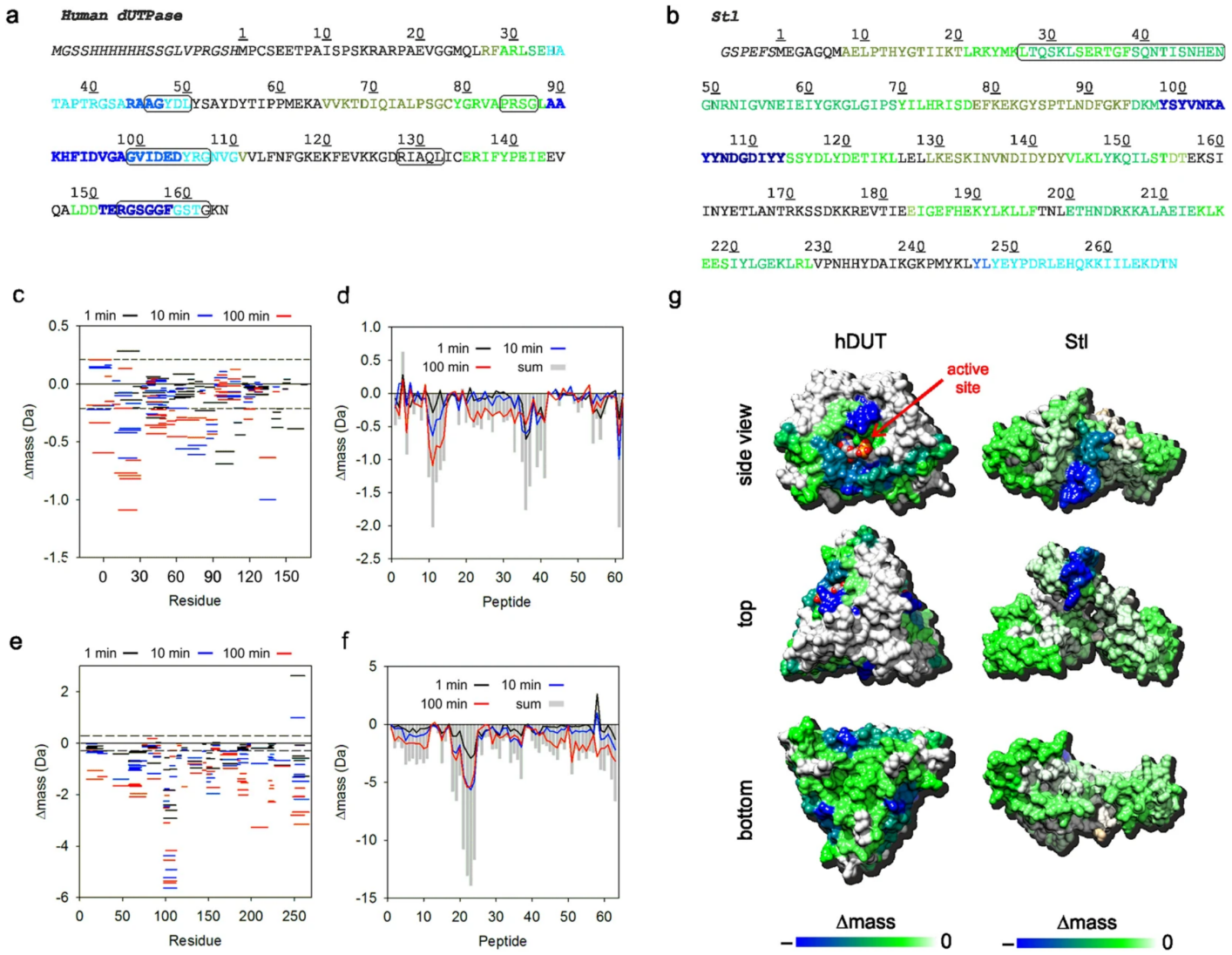


Figure 3.1: Representation of the hDUT:Stl HDX-MS results. (a & b) Sequence of hDUT and Stl proteins respectively. Numbering starts at the first residue of the Uniprot sequences of the proteins (Uniprot IDs: P33316-2 and Q9F0J8 respectively). Extension compared to Uniprot sequence is in italics. Active site residues in case of hDUT and the DNA binding motif of Stl are boxed. Sequence is coloured according to HDX data (Δmass accumulated across all labelling times) applying the colour-scheme displayed in (g). **(c–f)** HDX-MS difference data (c & e) and associate Woods plots (d & f) for hDUT (c & d) and Stl (e & f) showing the change in isotope uptake upon complexation of the proteins. Labelling time points are indicated by different colours and the dashed lines in (c) and (e) represent the 95 % confidence bands. **(g)** Representation of the HDX-MS difference data on the surface of the hDUT and Stl. In case of hDUT, an apo state structure is shown (PDB ID: 1Q5U), the C-terminal 13 residues are omitted from the representation since the position of these residues were not resolved in the crystal structure presumably because of flexibility. Position of the substrate analogue is shown based on the structural alignment of the apo and ligand-bound structures (3EHW). In case of Stl, a Phyre2 generated model is shown, which was compatible with synchrotron radiation circular dichroism and mutagenesis results obtained for the protein. Colouring is according to the scale at the bottom of the panel. Figure taken from [79].

In the course of this research, HDX-MS experiments were conducted on the individual proteins of hDUT and Stl as well as the two proteins mixed together. When combined with the other biophysical data presented in [79], our experiments provided additional and conclusive evidence for the formation of a complex between hDUT and Stl. Additionally, our data also reveals the location on each protein where the interaction takes place, information which is unique to our technique. For hDUT, the most significant isotope uptake differences were found in those peptides covering residues H34-L50 as well as A89-G110. These regions overlap with the first three of the five conserved motifs found within trimeric dUTPases, providing evidence for how Stl can bind to and inhibit different trimeric dUTPases which share comparatively little sequence homology. The C-terminal region of hDUT, which includes the fifth conserved motif, also shows a significant amount of deuterium uptake decrease upon complexation, however fluctuations of the HDX rates indicate the interactions involving this segment are weaker and more transient compared to those of other regions.

Trimeric dUTPases such as hDUT contain three distinct active sites for the hydrolysis of dUTP to dUMP, built up from residues of the five conserved motifs from all three subunits arranged in a specific pattern. Each singular active site is comprised of: motifs 1, 2 & 4 from one subunit along with motif 3 from another and the flexible motif 5 from the final subunit. Motifs 1, 2, 3 & 4 form the dUTP binding pocket between their two respective subunits while motif 5 acts as a lid, closing off the pocket upon substrate binding and locking the conserved residues in a catalytically competent arrangement [88–90]. We have shown that four of the five conserved motifs in hDUT are involved in the interaction with Stl which, in combination with the other biophysical data gathered by our collaborators, allows us to elucidate a potential method of inhibition.

It had been shown previously [74, 75] and reconfirmed in this study by the Vértessy group that the binding of dUTP and Stl to hDUT are mutually exclusive i.e. one can bind only if the other is not already present. A mechanistic model is therefore proposed in which Stl is only allowed to manoeuvre into the substrate binding pocket of hDUT if access is not hindered by either dUTP or the closed conformation of the flexible motif 5. The complex of hDUT and Stl may be further stabilised by motif 5 which is consistent with the decrease in Δmass observed in this motif upon complex formation. However, a caveat to this proposal is that, in accordance with our results, the stabilising effect is likely to be transient in nature. This proposal is supported by previously obtained results in which the Vértessy group found that a φ11 dUTPase mutant which lacked motif 5 did not see Stl binding perturbed [74].

When considering Stl, the HDX-MS experiments conducted in this study show that by far the largest amount of isotope uptake decrease occurred in the tyrosine-rich region from Y98-Y113 upon complexation to hDUT, indicating its involvement in the interaction. This finding is also consistent with previous results obtained by the Vértessy group in which they found that the removal of an N-terminal segment of Stl comprising residues 1-85 did not influence its ability to bind to and inhibit dUTPase [78]. We postulate that the global decrease in deuterium uptake seen in almost all peptides across the entire Stl sequence could be due to a decrease in flexibility that occurs upon complexation with hDUT. Previous experiments by the Vértessy group as well of those of collaborators in this study have led to the theory that the Stl dimer falls apart upon complexation with hDUT, however we see almost no positive

Δmass shifts in the Stl difference plot which would support this idea. Therefore we propose that the answer to this paradox is that the dimer interface of Stl might overlap with the Stl:hDUT interaction surface and hence any positive Δmass signals are supressed due to the presence of hDUT in the same location. This idea was previously theorised by the Vértessy group [78] and is given more credence by our data as well as a collaborator's SEC-SAXS data which utilised our HDX-MS results as restraints. SAXS is a technique with a variety of use cases such as enabling the low-resolution determination of the size and shape of macromolecules such as proteins. It is often used in place of traditional x-ray crystallography in situations where the protein under investigation cannot be crystallised, such as is the case for hDUT:Stl. With our results as a guide, SEC-SAXS has produced a model of the hDUT:Stl complex showing a stoichiometry of $hDUT_3Stl_3$ as well as $hDUT_3Stl_2$. A schematic model of hDUT:Stl complex assembly and Stl-DNA interaction is therefore proposed ([79] Figure 6 in paper) which explains the data this collaborative effort has gathered. It posits that while the complexation of hDUT to its substrate dUTP prevents Stl binding and inhibition, the complexation of Stl to its substrate DNA does not prevent hDUT binding, causing Stl dimer dissociation and the formation of $hDUT_3Stl_3$ as well as $hDUT_3Stl_2$ complexes.

3.2 The interaction of trimeric and dimeric dUTPases with Stl

Full results are detailed in Nyiri & Harris *et al.* 2019, which can be found in Appendix M. In this paper, we explored the functional plasticity of Stl and revealed new details for a better understanding of the different binding mechanisms of Stl for the two different classes of phage dUTPases. HDX-MS results point to the involvement of highly different Stl interaction surfaces when paired with a trimeric or a dimeric dUTPase. Based on these HDX results, as well as native MS results from previous studies, we put forward a schematic model which attempts to explain the mechanism of action of dUTPase inhibition by Stl, taking into account the differences between dUTPase classes. In this model (Scheme 1 in the paper), Stl dimerizes in solution and binds to DNA. Upon interaction with with either dUTPase, dimer dissociation (and therefore DNA binding dissociation) occurs and the Stl monomers instead bind to the dUTPase. For the trimeric dUTPases, a DUT_3Stl_2 or DUT_3Stl_3 complex forms, whereas for the dimeric dUTPases, the dUTPase dissociates as well and a DUT-Stl heterodimer forms. This leads to competitive inhibition of timeric dUTPases by Stl due to binding to the enzyme's active site and non-competitive inhibition of dimeric dUTPases by Stl due to perturbation of the enzyme's active site. Experimental HDX data acquisition and analysis as well as all related writings were done by the author.

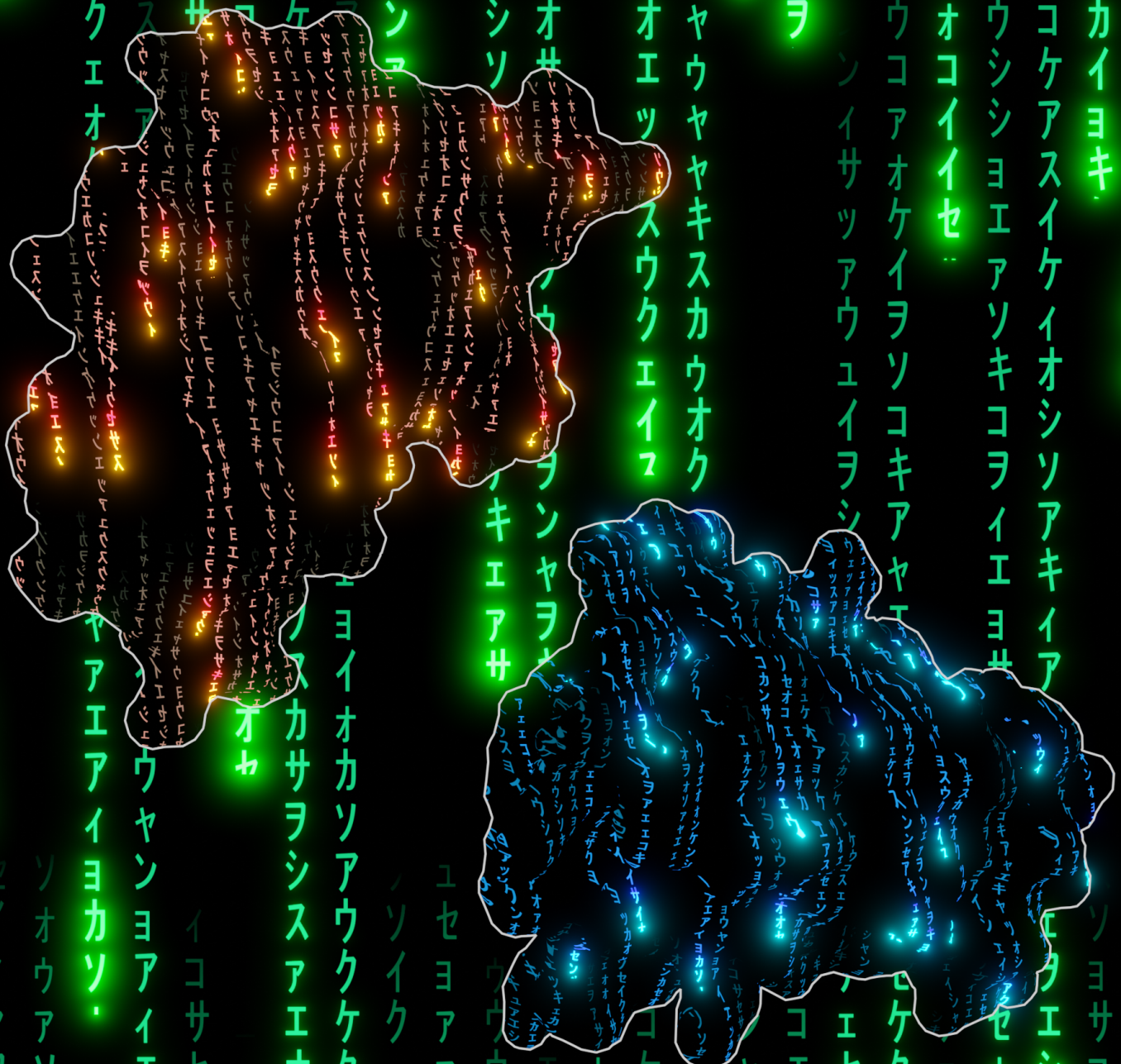
3.3 Summary

In this chapter, we investigated the interaction of the proteinaceous inhibitor Stl with several different dUTPases in order to advance our understanding of this novel interaction. We specifically looked at the binding of Stl with the timeric dUTPase hDUT in order to increase the breadth of our knowledge of this potentially therapeutic interaction, as well as the binding of Stl with the timeric dUTPase φ11DUT and the dimeric dUTPase φNM1DUT in order to try and better understand the functional plasticity of Stl.

hDUT has been designated as a target for onco-therapies due to its role in maintaining genome integrity via the removal of dUTP from the nucleotide pool, with several small molecule phase 1 trials

in progress [80]. In this study, we described the discovery of Stl as a novel potent proteinaceous inhibitor of hDUT. The hDUT:Stl interaction was characterised with a number of biophysical techniques, including HDX-MS, and a structural model of the interaction was obtained based on this data. Our HDX results were able to clearly delineate peptide segments around the hDUT active site that are involved in binding to Stl and this data is in line with the observed inhibition of the dUTPase enzymatic function and competition between Stl and dUTP for binding to dUTPase. Importantly, the entrance to the dUTP accommodating β-hairpin (i.e. the conserved Motif 3) as well as the interaction surface for the β-γ phosphate chain of the substrate (i.e. the conserved Motif 2) are both identified in the study as involved in Stl binding. Our evidence-based structural model offers insights into the mechanism of hDUT:Stl complexation in general and, additionally, the delineation of the peptide segments of Stl involved in interaction of hDUT, alluding to the possibility of development of peptide based inhibitors.

The previously unreported functional plasticity of Stl was also investigated by HDX-MS, which revealed details about the previously reported differing binding mechanisms of Stl for two different phage dUTPases [74, 76, 79], with HDX-MS experiments suggesting highly different interaction surfaces between Stl and the dimeric φNM1DUT and trimeric φ11DUT dUTPases. The study of this functional plasticity is important because *Staphylococcal* phages encode dUTPase representatives from both the trimeric and dimeric dUTPase families [74, 91] and so the investigation of Stl as a multi-purpose inhibitor has potential clinical relevance. Based on our HDX results, as well as native MS, a schematic model is proposed to explain this phenomena. Stl dimerises in solution and binds to DNA as dimers. Interaction of Stl monomers with dUTPases perturbs the dimerisation of the repressor, hence leading to the dissociation of the Stl-DNA complex. Trimeric φ11DUT dUTPase can form DUT_3Stl_2 or DUT_3Stl_3 complexes with Stl, while for the dimeric φNM1DUT, the complex is a DUT-Stl heterodimer. Based on our results, Stl binds directly to the active site of trimeric dUTPases and acts as a competitive inhibitor while for dimeric dUTPases it is a non-competitive inhibitor.



4 Results of experiments carried out for the purpose of optimisation

Numerous optimisation experiments were carried out over the course of this PhD in order to arrive at the final protocols stated in the Methods chapter. Those aspects of this work that required substantial amounts of optimisation are detailed below. The methods used for these optimisation experiments are included here rather than in the Methods chapter in order to aid understanding due to this chapter containing many small experiments, the details of which are necessary for understanding subsequent experiments. The aspects of the thesis which shall be covered in this chapter are: the experiments undertaken to improve the yields of both barstar and barnase, as neither initial protocol produced nearly enough for our needs. Optimising the hardware utilisation of our MD simulations in order to maximise the efficiency of their production. Improving native structure generation by protein-protein docking by varying the program/parameters used as well as the input HDX restraints data. Finally the exploration of the boundaries of modelling protein conformations using HDXsimulator in order to map and analyse the programs capabilities.

4.1 Improving barstar yield

Initial yields of BspWT using the protocol supplied by the Ikura group were too low to conduct meaningful experiments on. In order to remedy this, an extensive set of optimisation experiments were conducted in order to improve yields both at the initial production stage as well as the purification stage. The original protocol, as outlined by the Ikura group, can be found in Appendix H in order to allow a frame of reference.

4.1.1 Initial plasmid development

The impetus to start carrying out optimisation experiments on barstar came about due to our sudden inability to re-transform the DNA in our stocks into new competent cells. Despite nothing in our protocols changing and the DNA itself being stored at -70 °C, this sudden impotence forced us to seek an alternative source of barstar plasmid and thus prompted a more in-depth discussion about how to improve yields. An alternative source of BspWT DNA that closely matched what we already had was supplied by Addgene (Watertown, MA, USA) in the form of plasmid pMT643 (vector backbone: pUC19), a barstar plasmid with a resistance to ampicillin originally deposited by the Hartley group, that also contained the C40A & C82A mutations necessary for the elimination of disulphide bonds. Plasmid pMT643 was transformed into BL21(DE3)pLysS cells (Invitrogen, Inchinnan, UK) using the "*E. coli* Competent Cells" protocol by Promega (Southampton, UK) (Appendix A) with the resultant reaction spread on LB agar plates containing 50 µg/ml ampicillin & 34 µg/ml chloramphenicol and incubated overnight at 37 °C. A single colony was inoculated into 10 ml LB containing 50 µg/ml ampicillin & 34 µg/ml chloramphenicol and incubated overnight at 37 °C with agitation at 220 rpm.

The first parameter we decided to optimise was the induction time as the Ikura protocol merely states "4 h-o/n". 10 µl preculture was inoculated into each of 4x 200 ml 2xYT media containing 50 µg/ml ampicillin & 34 µg/ml chloramphenicol and incubated at 37 °C with agitation at 110 rpm until an OD of 0.6 was reached, whereupon expression was induced with 1 mM IPTG. Induced test expression cultures incubated for 4, 8, 19 and 24 hours at 37 °C with agitation at 110 rpm. After each induction time

had elapsed, the culture was removed from the incubator, centrifuged at 4,000 rpm for 10 minutes and the pellets frozen at -70 °C until they could be analysed. SDS-PAGE gel analysis carried out as stated in chapter 2.1.1. Results show similar levels of expression of BspWT between the 8, 19 and 24 hour time points with the 4 hour time point trailing behind (Figure 4.1 A). This result is important because it indicates that BspWT is not degraded over time which was a potential reason why the induction time as stated in the Ikura protocol was so variable, allowing us the convenience of an overnight induction.

Previous attempts at purification of BspWT with the original plasmid supplied by the Ikura group were inconsistent and resulted in a low protein yield. In order to test whether the BspWT from this new plasmid responded to the protocol any better, we carried out subsequent steps of cell lysis and purification as stated in Appendix H. Results indicated that initial ion exchange and subsequent SEC failed to purify the limited quantity on BspWT present in the sample, with contamination with several low yield proteins as well as a large amount of an unknown approx. 20 kDa protein (Figure 4.1 B). This unknown protein could in fact be a complex of barnase and barstar as the MW is approximately correct, however if this were the case it would require numerous further steps to unfold and re-purify the proteins individually and it was judged to not be worth the effort to attempt given that it might be a different protein altogether. Regardless, it was clear that the problems with purification had not been solved by changing to the new Addgene plasmid and, as BspWT production was still quite low, we decided to explore alternative options to try and solve both problems at the same time.

To do this, we contracted out with the company Gene Universal (Newark, Delaware, USA) to synthesise the BspWT gene *de novo* and insert it into the pET-28a vector with a C-terminal His-tag and kanamycin resistance instead of ampicillin. The resultant construct featured the BspWT gene seamlessly inserted downstream of the T7 promotor, lac operator and a ribosome binding site to aid expression, with no additional N-terminal residues. The gene then attaches seamlessly to a C-terminal 6x His-tag with no linker in between and a stop codon immediately after. This new BspWT construct was transformed into BL21(DE3)pLysS cells as stated previously with the resultant reaction spread on LB agar plates containing 50 µg/ml kanamycin & 34 µg/ml chloramphenicol and incubated overnight at 37 °C. A single colony was inoculated into 10 ml LB containing 50 µg/ml kanamycin & 34 µg/ml chloramphenicol and incubated overnight at 37 °C with agitation at 220 rpm. A test culture was set up in order to test the expression levels of this new construct. 1.2 ml of the overnight culture was inoculated into 200 ml 2xYT media and incubated at 37 °C with agitation at 110 rpm until an OD of 0.6 was reached, whereupon expression was induced with 1 mM IPTG. The expression culture was incubated overnight at 37 °C with agitation at 110 rpm. Gel analysis was then performed as previously described. The results show that this BspWT construct has far higher levels of protein expression than either the original construct or the Addgene construct using the same production conditions (Figure 4.1 C). The only problem was that the vast majority of the expressed protein is sequestered in inclusion bodies in the insoluble fraction which required some additional processing steps in order to access.

4.1.2 Initial purification tests

Inclusion body purification involves unfolding the protein using a chaotropic agent such as GnHCl in order to disrupt the aggregates, followed by refolding of the protein into its native conformation.

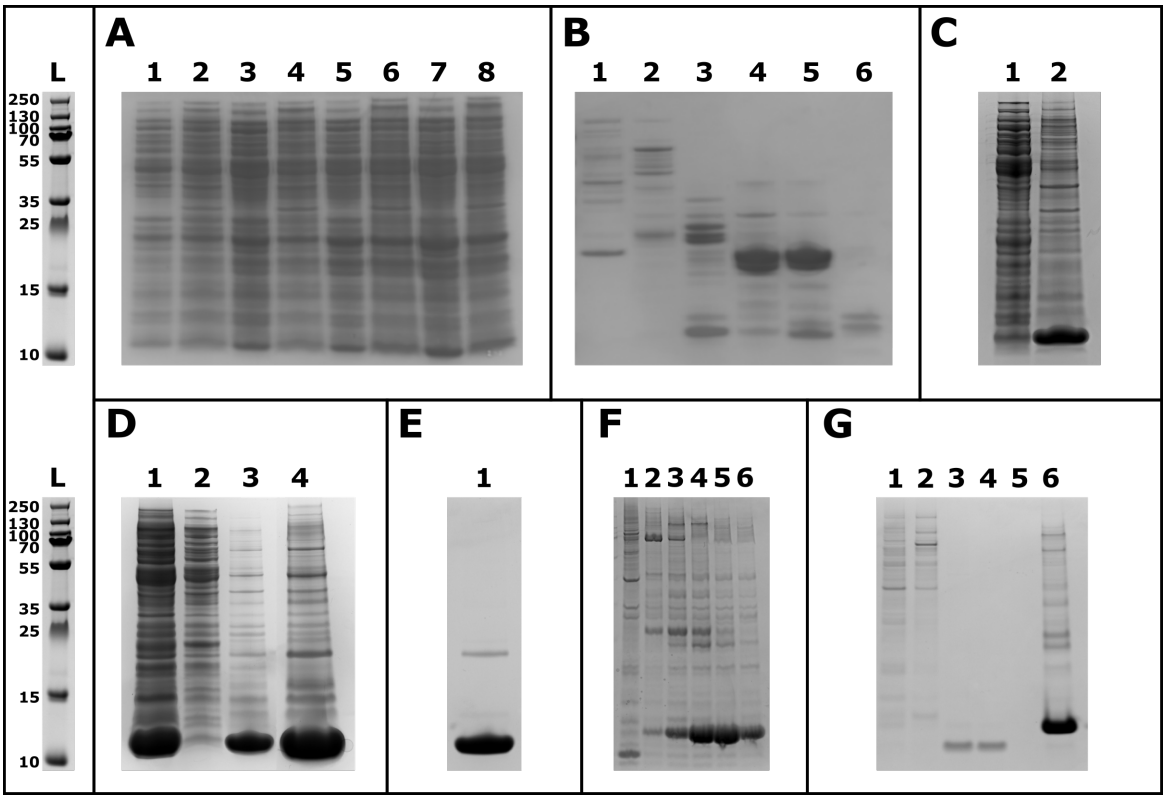


Figure 4.1: Barstar optimisation gels. A set of SDS-PAGE gels documenting the barstar optimisation process. The ladder (L) was PageRuler Plus Prestained Protein Ladder; the MWs (in kDa) of the marker proteins used are noted to the side. **(A)** 1 – 4 hour soluble, 2 – 4 hour insoluble, 3 – 8 hour soluble, 4 – 8 hour insoluble, 5 – 19 hour soluble, 6 – 19 hour insoluble, 7 – 24 hour soluble, 8 – 24 hour insoluble. **(B)** 1 – fraction 26, 2 – fraction 31, 3 – fraction 37, 4 – fraction 40, 5 – fraction 42, 6 – fraction 48. **(C)** 1 – soluble, 2 – insoluble. **(D)** 1 – whole cell lysate, 2 – soluble fraction, 3 – solubilised inclusion bodies, 4 – crashed pellet. **(E)** 1 – purified BspWT-His. **(F)** 1 – cleaved BspWT, 2 – fraction 5, 3 – fraction 6, 4 – fraction 7, 5 – fraction 8, 6 – fraction 9. **(G)** 1 – fraction 1, 2 – fraction 2, 3 – fraction 6, 4 – fraction 7, 5 – fraction 13, 6 – flow through (un-cleaved).

Refolding proteins correctly can be difficult, however we knew that theoretically barstar refolds easily from the literature as its initial interest to scientists was as a model system in folding studies. Cell pellet from the previous test culture was resuspended in lysis buffer containing 50 mM Tris HCl pH 8, 1 mM PMSF, 1 mM EDTA, 1 mM DTT & 1 % v/v Triton x-100. The solution was sonicated at 40 % amplitude for 6 mins, 5 seconds on, 10 seconds off to lyse the cells followed by centrifugation at 15,000 rpm for 20 mins at 4 °C. Resultant supernatant discarded and the pellet resuspended in more lysis buffer followed by another centrifugation at 15,000 rpm for 20 mins at 4 °C. Supernatant discarded and the pellet re-suspended in the same lysis buffer but without the Triton x-100 followed by another centrifugation at 15,000 rpm for 20 mins at 4 °C. Supernatant discarded and the remaining inclusion bodies resolubilised in the same lysis buffer but without Triton x-100 and with 6 M GnHCl. Sample left mixing overnight in a rotator at 4 °C. The sample was centrifuged at 15,000 rpm for 20 mins at 4 °C and the supernatant loaded into a 3.5 kDa MWCO Slide-A-Lyzer dialysis cassette and dialysed vs. 100x volume of buffer containing 50 mM Tris HCl pH 8, 300 mM NaCl, 5 % v/v glycerol, 25 mM imidazole & 1 mM DTT in order to refold the barstar by removing the chaotropic GnHCl. Dialysis buffer was changed once. During the second dialysis step, the sample in the cassette was observed to have gone cloudy and it was clear that the barstar had crashed out of solution when the concentration of GnHCl dropped low enough for aggregation to re-occur. The sample was extracted from the cassette, centrifuged at 15,000 rpm for 20 mins and samples of both the pellet and the supernatant taken for gel analysis. Results doubly show that barstar production using this new prep is far greater than any other prep tried before and also that the solubilisation of barstar aggregates is achieved using this method as can be seen by the very prominent band in the soluble inclusion bodies fraction (Figure 4.1 D). However, there is clearly a problem with refolding the protein after inclusion body solubilisation with 6 M GnHCl and this was the problem we tackled next.

The purification column we intended to use in order to take advantage of the His-tag on this new construct, a HisTrap HP, could tolerate 6 M GnHCl so it was decided to try and refold barstar on the column instead of using a dialysis cassette. This method would also have the advantage of allowing us to skip straight to purification, avoiding lengthy dialysis steps. Purification was carried out on an ÄKTA Pure. The pellet of the crashed protein was resolubilised in 4 ml equilibration buffer (50 mM Tris HCl pH 8, 300 mM NaCl, 25 mM imidazole & 6 M GnHCl) and loaded into a 5 ml loop. A 1 ml HisTrap HP column was equilibrated with 5 CVs of equilibration buffer and the solubilised sample washed over it using more equilibration buffer until the A280 nm baselined. The His-tagged barstar immobilised on the column was refolded by running a linear gradient of refolding buffer (50 mM Tris HCl pH 8, 300 mM NaCl & 25 mM imidazole) over the column at 0.5 ml/min for 30 ml to gradually remove the GnHCl. Pressure gauges on the ÄKTA Pure were closely monitored for signs of aggregation but no particular increases were seen. Refolded barstar was eluted isocratically using 10 CVs of elution buffer (50 mM Tris HCl pH 8, 300 mM NaCl & 300 mM imidazole) at 1 ml/min with fractions collected in 2 ml increments. A large single peak at A280 nm seen at this time, indicating successful purification. This was confirmed by gel analysis which showed that a large yield of BspWT-His had been purified with a much lower amount of a single contaminating protein also present that was suspected to be the barstar dimer (Figure 4.1 E). The identity of this contaminating protein was investigated by Native MS as described previously but proved inconclusive as no intense peaks could be found. The barstar-

containing fractions were dialysed vs. 100x volume 50 mM Tris HCl pH 8 in order to remove the high salt concentrations found in the elution buffer.

4.1.3 Final plasmid development

At this point, numerous HDX experiments were carried out on this purified BspWT sample in complex with BnWT. It was found that the results did not match up to a system that was fully bound and so native MS experiments as well as native gel analysis were carried out in order to see whether binding was fully occurring or not. We found that in fact a substantial amount of the complex sample was dissociated into the individual monomers which indicated that something was disrupting binding. This was because, with a K_d in the order of femtomolar, at experimental concentrations in the order of micomolar, full binding and no monomers should be seen. It was postulated that perhaps the un-cleaved His-tag was disrupting the binding site and so we contracted out with Gene Universal again to produce an identical version of the same BspWT plasmid with the exception of now having an N-terminal 6x His-tag with a TEV cleavage site between it and the protein.

This new BspWT plasmid was transformed and cultured using the same protocol as before, except that agitation during the expression stage was increased to 220 rpm in order to try and increase cell density. Gel analysis of a 200 ml test expression culture showed that on one hand this new plasmid seemed to produce substantially less protein than the previous plasmid did, however on the other hand the majority of the protein was now in the soluble fraction. It was suspected that the reason for the latter was because of the former. This meant that, while our yields would be lower, the production protocol would be simpler because we would not have to purify from inclusion bodies and this was judged to be a worthy trade-off.

Several improvements to the BspWT production protocol were made on the advice of the Booth group in an effort to make the whole process more efficient. The pellet of the test expression culture was resuspended in 50 ml PBS, 10 mM β-mercaptoethanol, 2 µl benzonase & 1 EDTA-free protease inhibitor tablet and incubated at room temperature for 10 mins before the cells were lysed using a cell disruptor at 25 kPsi for 2 cycles. The whole cell lysate was clarified by centrifugation at 20,000 rpm for 30 mins at 4 °C and the resulting supernatant removed and concentrated down to approx. 15 ml using 3 kDa MWCO spin concentrators at 4,000 rpm. A 1 ml HisTrap HP column was equilibrated with 5 CVs of equilibration buffer (50 mM Tris HCl pH 8, 300 mM NaCl, 25 mM imidazole) and the concentrated cell lysate spiked with 1125 µl 4 M NaCl & 400 µl 1 M imidazole in order to bring the concentrations of both up to approx. the same as the equilibration buffer. Cell lysate loaded into column via a 5 ml loop (in 3 batches) with equilibration buffer at 1 ml/min and then eluted using 10 CVs of elution buffer (50 mM Tris HCl pH 8, 300 mM NaCl & 300 mM imidazole) over a linear gradient at 1 ml/min with fraction sizes of 1 ml. A large peak at A280 nm was seen in fractions 4-10, corresponding to an imidazole concentration of 120-300 mM. These fractions were combined and dialysed vs. 100x volume of 50 mM Tris HCl pH 8 overnight at 4 °C using a 3 kDa MWCO Slide-A-Lyzer with gentle stirring.

4.1.4 His-tag cleavage

The goal now was to cleave off the His-tag using TEV protease, an enzyme that recognises and cleaves the sequence ENLYFQ/S where the "/" represents the cleaved peptide bond. AcTEV was acquired from Invitrogen which has the advantage of containing a His tag itself, allowing for easy purification after cleavage occurs (AcTEV and cleaved His tag will remain bound to the column whereas barstar will now flow straight through). The reaction was set up according to Appendix E. Cleavage reaction incubated at 16 °C for 20 hours. Reaction diluted with equilibration buffer to a total volume of 5 ml and loaded into a 5 ml loop before being washed over a 1 ml HisTrap HP column (equilibrated with 5 CVs equilibration buffer) with 10 ml equilibration buffer. The flow through was collected as this is the location of the cleaved BspWT protein. Proteins remaining on the column were eluted with a linear gradient of elution buffer for 10 CVs at 1 ml/min in 1 ml fractions. The flow though was dialysed vs. 100x volume of 50 mM Tris HCl pH 8 overnight at 4 °C using a 3 kDa MWCO Slide-A-Lyzer with gentle stirring. Gel analysis revealed the successful cleavage of BspWT, however with much less efficiency than hoped with the majority remaining un-cleaved (Figure 4.1 F). Additionally, purification had not been as successful as we had thought as a large number of contaminating proteins remained in the cleaved barstar fraction.

Additional SEC purification was carried out on the cleaved BspWT sample in order to clean it up. The sample was concentrated down to approx. 250 µl using a 3 kDa MWCO vivaspin concentrator and loaded into a Superdex 75 10-300 GL SEC column pre-equilibrated with 2 CVs 50 mM Tris HCl pH 8. Sample ran with 1.5 CVs 50 mM Tris HCl pH 8 at 1 ml/min with 1 ml fractions. Void volume of the column was discarded and small A280 nm peaks seen in fractions 2, 7, 10, 15, 22 & 38. Gel analysis indicates that the BspWT sample has been successfully purified with the majority of the high MW proteins visible in fraction 2 and the cleaved barstar in fraction 7. Several improvements to this protocol were put into effect for subsequent attempts:

- Resuspend the initial cell pellet in less PBS buffer in order to make concentration faster
- Incubate protein in AcTEV at a higher temperature in order and with mixing to increase cleavage efficiency
- Use desalting columns in order to remove imidazole and so avoid extra dialysis.

The barstar protocol was now optimised as far as we would take it and a final attempt was made using the protocol as described in chapter 2.1.2. Yield of cleaved protein was improved, however the vast majority remained un-cleaved (Figure 4.1 G) and it was clear that a substantial amount of work would be needed in order to boost yields to an acceptable level. Initial HDX tests on the small amount of cleaved protein we had been able to produce indicated that the interaction between barnase and barstar had been substantially improved by the new cleaved barstar and so we were sure this was the protein we wanted to pursue. However, because running this protocol took almost the entire week, there was very little time to be doing both HDX and protein production and so it was decided to contract out barstar production to Ruth Rose of the Protein Production Facility at Queen Mary University of London while we focused on HDX of other systems. The plasmid as well as the production and purification protocols we had developed up until this point would be used as a starting point for the

Protein Production Facility to iterate upon until a large amount of cleaved barstar could be produced.

To summarise, we developed a new, modern plasmid to replace the outdated (and eventually ineffective) original and, after several revisions, managed to successfully produce and purify protein using an entirely new protocol. The only problem that we left unfinished was that of His-tag cleavage which had proven to be far less efficient than we had hoped and so, in order to remedy this issue, we commissioned an outside source to help us finish off production using our developed protocols as a baseline.

4.2 Improving barnase yield

Initial yields of BnWT using the protocol supplied by the Ikura group were also too low to conduct meaningful experiments on. In order to remedy this, an extensive set of optimisation experiments were conducted in order to improve yields both at the initial production stage as well as the extraction and purification stages. The original protocols, as outlined by the Ikura group, can be found in Appendix I in order to allow a frame of reference.

4.2.1 Attempts to develop a new plasmid

Our first attempt to boost BnWT yields took its cues from our successful experiments on barstar where we had the gene synthesised in a modern, high-throughput plasmid. Like barstar, the plasmid we received barnase in from our collaborator was extremely old and lacked many of the optimised features that modern commercial plasmids contain. Therefore, we contracted with Gene Universal to synthesise the BnWT gene (as well as the accompanying barstar gene for barnase inactivation) in the pET-26b(+) containing a resistance gene for kanamycin, with a His-tag at the C-terminus. Desiccated DNA was resuspended in 50 µl ddH_2O and its concentration determined to be 300 ng/µl. A 50 ng/µl solution was prepared by diluting 10 µl of the stock with 50 µl ddH_2O. Diluted DNA transformed into BL21(DE3)pLysS cells using the "*E. coli* Competent Cells" protocol by Promega (Appendix A) with the resultant reaction spread on LB agar plates containing 50 µg/ml kanamycin & 34 µg/ml chloramphenicol and incubated overnight at 37 °C. A single colony was inoculated into 10 ml LB containing 50 µg/ml kanamycin & 34 µg/ml chloramphenicol and incubated overnight at 37 °C with agitation at 220 rpm. 10 µl preculture inoculated into 10 ml 2xYT media containing 50 µg/ml kanamycin & 34 µg/ml chloramphenicol and incubated at 37 °C until an OD of 0.6 was reached, whereupon expression was induced with 1 mM IPTG. A test expression culture was incubated overnight at 37 °C with agitation at 220 rpm.

550 µl acetic acid was then added to the test culture and left mixing in a rotator at 4 °C for 20 mins before being centrifuged at 13,000 rpm for 5 mins. Supernatant and pellet subsequently collected with the pellet resuspended in 50 µl 50 mM Tris HCl pH 8 and then sonicated at 50 % amplitude for 2x2 second bursts. Lysed pellet samples were centrifuged at 13,000 rpm for 10 mins and the supernatant and pellet collected. The pellet was then resuspended in 100 µl 50 mM Tris HCl pH 8 + 4 % SDS and centrifuged at 13,000 rpm for 10 mins. An SDS-PAGE gel was run using various samples from the above procedures using the protocol described previously. Gel analysis revealed little visible production of BnWT-His, theorised to be due to lack of barstar expression, leading to cell death. In order to test this, another expression culture was set up as described previously but on a larger scale (100 ml) with OD measurements taken every 30 mins for 7 hours post-inoculation of the starter culture. Aliquots for gel

analysis were taken 1, 2, 4 & 20 hours post induction. Results indicated that cell death was not occurring with steady growth seen throughout the experiment, capping out at and OD of approx. 6. Tellingly, no lag phase is seen post-induction and the gel confirms that that almost no production of BnWT-His took place (Figure 4.2 A).

We decided to switch our bacterial cells from BL21(DE3)pLysS to BL21-AI which are specifically designed for the expression of toxic proteins such as barnase. Transformation protocol is as previously described with the exception of a heat shock of 30 seconds and the LB agar plates lacking chloramphenicol as BL21-AI has no inherent resistance to this antibiotic. A 100 ml test expression culture was set up as before except for expression being induced with 1 mM IPTG + 0.1 % arabinose. OD measurements were taken every 30 mins for 7 hours with additional measurements taken the next day. Samples for gel analysis taken at 2, 4 & 20 hours post-induction. Results showed that in comparison to BL21(DE3)pLysS, BL21-AI cells did in fact see a substantial amount of cell death starting at approx. 5.5 hours post induction. From a peak measurement of approx. 3, OD values very rapidly decreased to a minimum of approx. 0.8 five hours after induction with only a slight recovery to approx. 1.2 after 22 hours post-induction. It was hoped that the cause of this cell death was the production of a large amount of barnase, however SDS-PAGE gel analysis (Figure 4.2 B) indicated that to not be the case. This gel was extremely busy and so in order to be sure that no BnWT-His was produced; a Western blot was carried out as follows:

An SDS-PAGE gel was run as previously described, however fixing stain was not added. Two squares of triple-stacked filter paper as well as the gel were soaked in transfer buffer (25 mM Tris, 190 mM glycine, 10 % v/v MeOH & 1 mM SDS). A square of PVDF was cut to be the same size as the gel and soaked briefly in methanol before rinsing in dH_2O and leaving to soak in transfer buffer. A stack consisting of filter paper followed by PVDF followed by the gel followed by filter paper was run on a transfer blotter at 45 mA for 1.5 hours. The PVDF membrane was removed and soaked overnight in 20 ml of 5 % milk powder in PBS-T (0.05 % TWEEN 20) at 4 °C on a rocker. The PVDF was rinsed thoroughly with PBS-T and before being soaked in 20 ml 5 % milk powder in PBS-T + 2 µl anti-His antibody and left on rocker for 1 hour. The membrane was once again rinsed with PBS-T and then blotted with horseradish peroxide before being imaged using chemiluminescence for 5 mins. The resultant image (Figure 4.2 C), while extremely unclear, does at least confirm our suspicions that no BnWT-His was produced as no strong bands are visible in any of the samples around the appropriate MW.

We believe this wholesale lack of barnase production to be due to the complex interaction between the barnase and barstar genes that both need to be included in the plasmid due to barnase's toxicity. While the plasmid we made did include both genes and was made by a professional plasmid-designing company, there was evidently some aspect of the original plasmid that we overlooked as our attempts resulted in cell death, no doubt due to uninhibited barnase production.

4.2.2 Production optimisation

With the failure of the BnWT-His construct to produce meaningful amounts of protein, our attention once again turned to the original protocol supplied to us by the Ikura group and how it might be

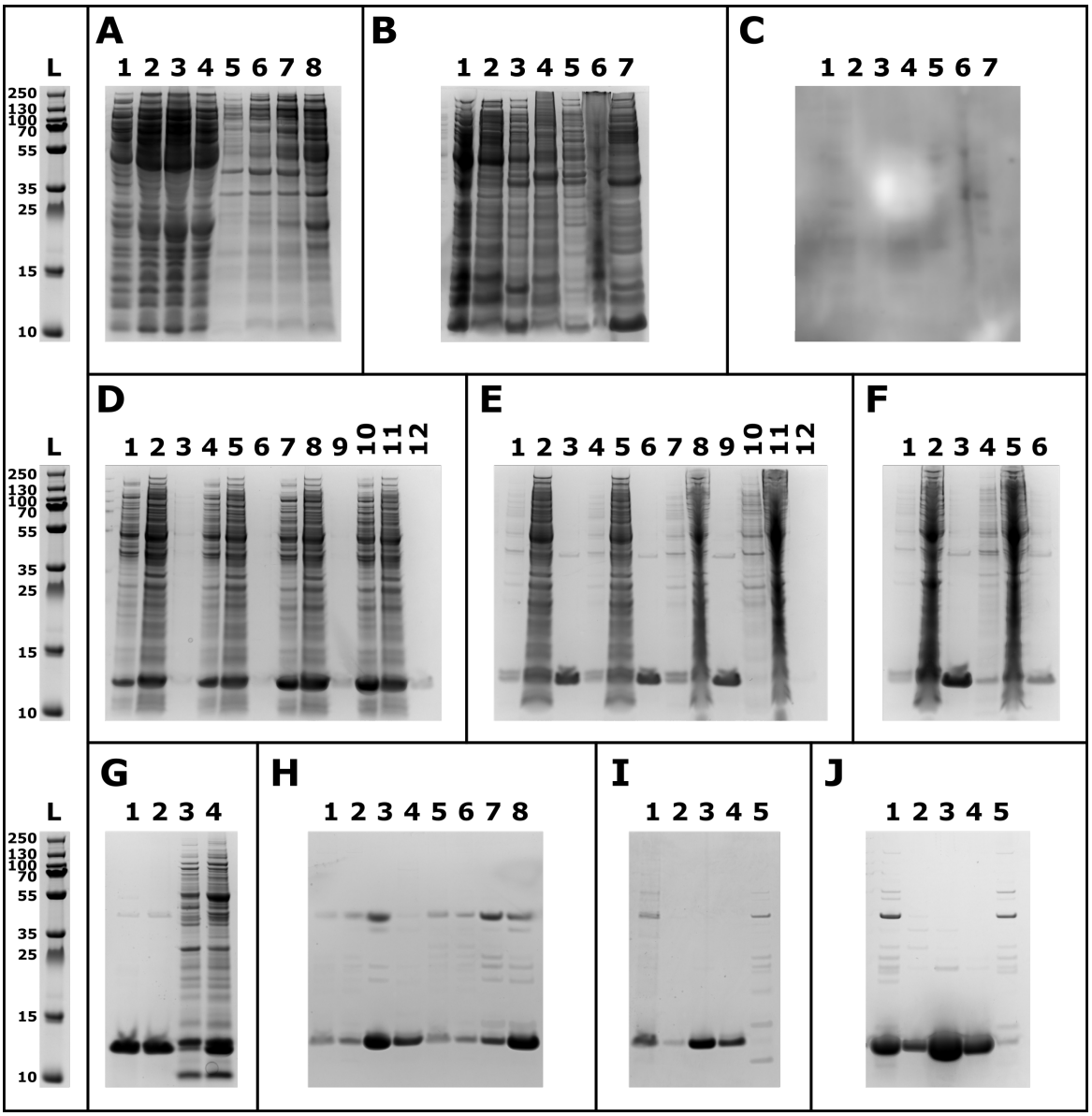

L
250
130
100
70
55
35
25
15
10
A
1 2 3 4 5 6 7 8
B
1 2 3 4 5 6 7
C
1 2 3 4 5 6 7
D
1 2 3 4 5 6 7 8 9 10 11 12
E
1 2 3 4 5 6 7 8 9 10 11 12
F
1 2 3 4 5 6
G
1 2 3 4
H
1 2 3 4 5 6 7 8
I
1 2 3 4 5
J
1 2 3 4 5

Figure 4.2: Barnase optimisation gels. A set of SDS-PAGE gels documenting the barnase optimisation process. The ladder (L) was PageRuler Plus Prestained Protein Ladder; the MWs (in kDa) of the marker proteins used are noted to the side. **(A)** 1 – 1 hour soluble, 2 – 2 hour soluble, 3 – 4 hour soluble, 4 – 20 hour soluble, 5 – 1 hour insoluble, 6 – 2 hour insoluble, 7 – 4 hour insoluble, 8 – 20 hour insoluble. **(B)** 1 – 2 hour soluble, 2 – 2 hour insoluble, 3 – 4 hour soluble, 4 – 4 hour insoluble, 5 – 20 hour soluble, 6 – 20 hour insoluble, 7 – 20 hour media. **(C)** 1 – 2 hour soluble, 2 – 2 hour insoluble, 3 – 4 hour soluble, 4 – 4 hour insoluble, 5 – 20 hour soluble, 6 – 20 hour insoluble, 7 – 20 hour media. **(D)** 1 – 30 °C 110 rpm soluble, 2 – 30 °C 110 rpm insoluble, 3 – 30 °C 110 rpm acetic acid, 4 – 30 °C 220 rpm soluble, 5 – 30 °C 220 rpm insoluble, 6 – 30 °C 220 rpm acetic acid, 7 – 37 °C 110 rpm soluble, 8 – 37 °C 110 rpm insoluble, 9 – 37 °C 110 rpm acetic acid, 10 – 37 °C 220 rpm soluble, 11 – 37 °C 220 rpm insoluble, 12 – 37 °C 220 rpm acetic acid. **(E)** 1 – 0 mM phosphate soluble, 2 – 0 mM phosphate insoluble, 3 – 0 mM phosphate supernatant, 4 – 0.01 mM phosphate soluble, 5 – 0.01 mM phosphate insoluble, 6 – 0.01 mM phosphate supernatant, 7 – 0.1 mM phosphate soluble, 8 – 0.1 mM phosphate insoluble, 9 – 0.1 mM phosphate supernatant, 10 – 1 mM phosphate soluble, 11 – 1 mM phosphate insoluble, 12 – 1 mM phosphate supernatant. **(F)** 1 – BL1-AI soluble, 2 – BL21-AI insoluble, 3 – BL21-AI media, 4 – BL21-pLysS soluble, 5 – BL21-pLysS insoluble, 6 – BL21-pLysS media. **(G)** 1 – acetic acid (media), 2 – acetic acid (pellet, supernatant), 3 – cold osmotic shock (pellet), 4 – chloroform (pellet). **(H)** 1 – Tris HCl post acetic acid, 2 – Tris HCl post dialysis, 3 – Tris HCl fraction 5, 4 – Tris HCl fraction 6, 5 – sodium acetate post acetic acid, 6 – sodium acetate post dialysis, 7 – sodium acetate fraction 5 , 8 – sodium acetate fraction 6. **(I)** 1 – post acetic acid, 2 – fraction 8, 3 – fraction 9, 4 – fraction 10, 5 – flow through. **(J)** 1 – post dialysis, 2 – fraction 8, 3 – fraction 9, 4 – fraction 10, 5 – flow through.

improved. In the first stage of optimisation, we focused on adjusting the temperature and agitation rate of the expression culture to maximise production as qualitatively judged by an SDS-PAGE gel. 600 µl preculture was inoculated into 100 ml baffled flasks of minimal phosphate media (Appendix C) + 50 µg/ml ampicillin, whereupon they were incubated overnight in 1 of 4 different test conditions: 30 °C, 110 rpm agitation (original stated conditions); 30 °C, 220 rpm agitation; 37 °C, 110 rpm agitation & 37 °C, 220 rpm agitation. 15 ml of each culture was removed and 825 µl acetic acid added to release barnase from the periplasm. Samples for gel analysis were taken at each stage and processed as stated previously. Qualitative results indicate that, perhaps unsurprisingly, the 37 °C & 220 rpm agitation sample produced the most BnWT, followed by 37 °C, 110 rpm agitation then 30 °C, 220 rpm agitation and finally the original condition of 30 °C, 110 rpm agitation (Figure 4.2 D). From these results we can see that temperature has a larger impact on production than agitation rate, presumably due to higher cell growth.

The next step was to vary the amount of phosphate contained within the minimal media to see if it would influence the degree of BnWT expression (as the plasmid is under the control of the alkaline phosphatase promotor). This was done by varying the levels of Na_2HPO_4 & NaH_2PO_4 in the neutral phosphate buffer (Appendix C). The original stated value of 0.1 mM phosphate (0.05 mM Na_2HPO_4 & 0.05 mM NaH_2PO_4) was tested as well as zero phosphate, 0.01 mM phosphate (0.0025 mM Na_2HPO_4 & 0.0025 mM NaH_2PO_4) and 1 mM phosphate (0.5 mM Na_2HPO_4 & 0.5 mM NaH_2PO_4). 600 µl preculture inoculated into 100 ml minimal media + 50 µg/ml ampicillin and left incubating overnight at 37 °C with agitation at 220 rpm. 1 ml samples taken from each culture and treated with 55 µl acetic acid to release the barnase. Samples for gel analysis were taken at each stage and processed as stated previously. Qualitative results indicate that varying phosphate concentration seems to have little effect on production of barnase until the concentration approaches 1 mM where production completely stops (Figure 4.2 E). This is probably due to there being too much phosphate to induce production via the alkaline phosphatase promotor. Given that almost no noticeable difference could be seen in this gel, we decided to stay with the original stated value of 0.1 mM phosphate throughout subsequent experiments.

After this we decided to test whether transforming the original BnWT plasmid into BL21-AI cells would increase yields considering this cell line is specialised for toxic proteins. Transformation into BL21-AI carried out as previously described. Two 100 ml expression cultures were set up, the first containing BnWT plasmid in the original BL21(DE3)pLysS cells and the second containing the plasmid in BL21-AI cells. Both expression cultures were set up using the best conditions as optimised thus far. After growth, samples from the cultures were taken and processed as previously described and analysed via gel. The gel showed that BnWT production was far higher in the BL21-AI cells, however BL21(DE3)pLysS production did appear to be lower than it had been in previous experiments, although this was possibly due to the gel imager auto-adjusting contrast against a much more intense band in the BL21-AI lane (Figure 4.2 F). Therefore, we decided to switch over to BL21-AI cells as they certainly did not decrease yields and had the potential to produce even more than the BL21(DE3)pLysS with further optimisation.

With a new cell line established, we moved on to attempting to optimise the release of the protein

from the periplasm as previous results had suggested that not all protein produced made it into the media following the osmotic shock protocol we had used thus far. Three alternative techniques we tested vs. the original: acetic acid on a pelleted and resuspended sample, cold osmotic shock using a sucrose solution & chloroform extraction. Test expression cultures were set up as previously described. After growth, 1 ml samples were taken from each of the 4 cultures in order to test the different release methods. One sample was treated as normal in order to serve as a baseline for comparison. For the other sample treated with acetic acid, the culture was centrifuged at 13,000 rpm for 5 mins and the supernatant discarded, after which the pellet was resuspended in 100 µl 50 mM Tris HCl pH 8 to which 5.5 µl acetic acid was added (proportionally the same amount as the larger sample). From this point on, the sample was treated the same as the normal acetic acid sample. In order to test cold osmotic shock, the culture was centrifuged as above and the pellet resuspended in 500 µl 20 % sucrose, 1 mm EDTA, 30 mM Tris HCl pH 8 and subjected to gentle mixing for 10 mins at room temperature. The sample was then centrifuged at 13,000 rpm for 10 mins at 4 °C and the supernatant discarded. The pellet was then rapidly resuspended in 500 µl ice cold dH_2O and subjected to gentle mixing at 4 °C for 10 mins. Finally, the sample was centrifuged at 13,000 rpm for 10 mins and the supernatant collected. Lastly, in order to test chloroform release, a sample was pelleted as described previously, the supernatant discarded and then the pellet resuspended in some residual supernatant. 10 µl chloroform was added and the cells briefly vortexed before being incubated at room temperature for 15 mins. 100 µl 10 mM Tris HCl pH 8 was added to the cells followed by centrifugation at 6,000 x g for 20 mins, after which the supernatant was collected. All supernatant samples were filtered through a 0.22 µm filter and concentrated to a total volume of 50 µl using a 5 kDa MWCO vivaspin at 15,000 x g. Samples were then run and analysed on a gel as previously described.

Both the acetic acid samples showed similar high levels of expression with high purity, however the samples treated with the sucrose solution as well as chloroform showed similar levels of protein release but at the cost of much lower purity (Figure 4.2 G). This is presumably because these two methods actually lysed the cells despite only being intended to release protein from the periplasm. Interestingly, for the first time we can see a band matching the MW of barstar in these two samples which has never been visible before in any fraction of a barnase gel. This experiment was important because it showed that barnase yield could be maintained even if the expression culture was pelleted and resuspended, against the guidance of the original protocol which called for acetic acid to be added directly to the media. Up until now, barnase purification had involved a very time-consuming step where the entire expression culture was filtered and run over an ion exchange column, a process which could take more than a day if several litres of culture were involved. Now with this result in hand, we could cut the amount of liquid requiring processing to a fraction of what it was before.

4.2.3 Purification optimisation

With the optimisation of the production and release of BnWT brought to acceptable levels, our attention turned to purification. This had always been a particularly weak part of the original protocol with results varying substantially between batches both in terms of purity of sample as well as location of sample (e.g. sometimes BnWT could be found within normal fractions but sometimes could be found in flow-through etc.). The first stage of purification tests involved testing two different buffers:

50 mM Tris HCl pH 8 and 50 mM sodium acetate pH 5. A 100 ml test expression culture was grown using the optimised conditions as stated thus far and split in two after growth had concluded. After centrifugation, the two pellets were resuspended, one with 5 ml 50 mM Tris HCl pH 8 and one with 5 ml 50 mM sodium acetate pH 5, to both of which was added 275 µl acetic acid. Samples were left mixing for 20 mins at 4 °C before being centrifuged at 4,000 rpm for 15 mins at 4 °C and the pellets discarded. Samples were dialysed vs. 100x volume of 50 mM Tris HCl pH 8 or 50 mM sodium acetate respectively using 3 kDa MWCO Side-A-Lyzer cassettes in order to remove the acetic acid and return the samples to their buffered pH values. Dialysis occurred for 2 hours at 4 °C with gentle stirring with one round being sufficient to increase the pH of the sodium acetate sample from 3.0 to 4.9, however the Tris HCl samples required two rounds in order to bring its pH from 2.4 to 7.8.

Test purifications were carried out on an ÄKTA Pure using a 5 ml HiTrap SP HP cation column. The column was washed with 3 CVs dH_2O followed by 3 CVs 0.5 M NaOH followed by 2 CVs 2 M NaCl followed by 3 CVs 50 mM Tris HCl pH 8 or 50 mM sodium acetate pH 5 respectively. Samples were filtered with a 0.22 µm syringe filter and loaded into a 5 ml loop before being washed over the column with 3 CVs of the respective sample buffer. For this first purification test, samples were eluted isocratically using 3 CVs 1 M NaCl in 50 mM Tris HCl pH 8 for the Tris HCl sample and 0.5 M NaCl in 50 mM sodium acetate pH 5 for the sodium acetate sample with fraction sizes of 3 ml. Substantial peaks at A280 nm were seen in fractions 5 + 6 in both samples and these fractions were analysed by SDS-PAGE gel as previously described. The gel showed that on one hand, BnWT is now seen in the correct place (the fractions instead of the flow-through), however it also shows that isocratic elution is not sufficient to completely purify the samples (Figure 4.2 H). Fractions 5 + 6 of the Tris HCl sample appeared to be slightly cleaner than the sodium acetate sample, therefore we decided to take the Tris HCl method forward to further purification trials using gradient elution.

Another test expression culture was set up and purification carried out as stated previously, however instead of isocratic elution, a linear gradient of 0-1 M NaCl in 50 mM Tris HCl pH 8 at 1 ml/min was used with 1 ml fraction sizes. Peaks at A280 nm were seen in fractions 8, 9 & 10, corresponding to a NaCl concentration of 0.8-1 M. Gel analysis was carried out as stated previously and showed that BnWT had indeed been purified using gradient elution with strong bands matching BnWT in fractions 8, 9 & 10 with almost no contaminating bands visible (Figure 4.2 I). With the success of this optimisation, a 1 L expression culture was grown in order to make sure these procedures worked on upscaled reagents. Procedure used is as stated in chapter 2.1.1. Gel analysis on these far more concentrated samples showed that some faint contaminating bands were now visible which were too low concentration to be picked up by the previous gel, however it was determined that the concentration of these contaminants was so low compared to BnWT production as to not be worth further efforts at purification (Figure 4.2 J).

With all these optimisations steps combined, the amount of BnWT that could be successfully purified from a 1 L culture was approx. 3.08 mg/L, in comparison to approx. 0.5 mg/L using the original protocol, a greater than 6-fold increase. In addition, changes to the pipeline protocols allowed for much simpler and more consistent production and purification, a benefit unto its own. To summarise,

we originally attempted to ameliorate the yield problems we were having by replacing the original plasmid with one of more modern design, however these experiments proved unsuccessful. Therefore we returned to and successfully optimised the conditions for the originally supplied plasmid, eventually producing and purifying yields more than sufficient for our needs.

4.3 Improving Molecular Dynamics performance

MD simulations were performed using NAMD in order to relax the crystal structures of the proteins so that their exact bound conformations did not influence subsequent docking simulations. MD is famously a very computationally expensive technique and so a substantial amount of time was spent optimising our simulations so they could be run as quickly as possible. This was done primarily by comparing how quickly simulations would run on various different CPU/GPU configurations and whether it was more efficient to devote resources to complete one simulation quickly or multiple simulations at once but more slowly.

In order to properly benchmark each configuration, a number of metrics were calculated. First, as the size of the system heavily influences the time taken to run MD, the size of the simulation in terms of the number of atoms was determined. Next, the time taken to run the simulation was established, as reported by NAMD, and divided by the number of nanoseconds the simulation was run for in order to determine the time taken to calculate 1 ns worth of trajectories (Equation 4.1):

$$T = \frac{t/l}{3.6 \times 10^3} \tag{4.1}$$

Where T is the time per nanosecond in hours, t is the time taken to run the simulation in seconds and l is the length of the simulation in nanoseconds. The size of the simulation was then incorporated into the calculation in order to determine the time per nanosecond per atom (Equation 4.2):

$$T_a = \frac{T}{s} \cdot 3.6 \times 10^6 \tag{4.2}$$

Where T_a is the time per nanosecond per atom in milliseconds and s is the size of the system in atoms. Equation 4.2 helps us to benchmark how long it takes to run a given MD simulations regardless of its size, however there is one more parameter we must take into account and that is the number of simulations which can be run simultaneously on any given piece of computer hardware with a certain amount of resources allocated to each simulation. This is simply determined by dividing the result of Equation 4.2 by the number of parallel simulations that can be run (Equation 4.3):

$$E = \frac{T_a}{p} \tag{4.3}$$

Where E is the efficiency of the method in milliseconds and p is the number of parallel simulations that can be run with a given resource allocation.

Our initial MD simulations were run on a local Linux workstation with 32 CPU cores available.

NAMD is known to scale well with multiple CPU cores and so we wanted to test and see if it would be more efficient to run: 1 simulation with all 32 cores, 2 simulations in parallel each with 16 cores or 4 simulations in parallel each with 8 cores. The efficiency metric described earlier would help us evaluate this by providing a score by which these three configurations could be judged, with the lower the score the better. Additionally, the other metrics would allow us insight into how the size of the simulation factored into this efficiency score. Several simulations of varying size were run at each core count and the data averaged to give a single figure for each. Results (Figure 4.3) indicate that running all 32 cores on one simulation is the least efficient method with a value of 463 ms, while 16 cores on 2 simulations and 8 cores on 4 simulations have similar efficiency scores at an average of 308 & 299 ms respectively. While NAMD is known for its scalability, this is not in evidence here, likely because our systems are too small to fully take advantage of the extra cores and so compute power is going to waste. 16 and 8 cores had very similar efficiency scores, indicating that our systems do scale at least up until 16 cores. This result also gave us flexibility in terms of whether we wanted to run 2 or 4 simulations depending on if speed or throughput was our priority at that moment as little time would be lost regardless of which method we used.

Simulations were run for a time using this set up, however there were a large number to do and we knew that we would likely have to repeat many of them as new information emerged. Therefore, we endeavoured to increase efficiency even more by taking advantage of the King's College London HPC cluster Rosalind and the number of NVIDIA TESLA V100 GPUs they had installed. By making use of these specialist data science GPUs, we were able to dramatically increase our efficiency scores, with an average of 35 ms being achieved over a large number of systems, despite the fact that due to university limitations we could only use 1 GPU at a time (with 1 system running). This massive efficiency increase of almost 10x clearly demonstrates the power of dedicated GPU acceleration over more traditional CPU approaches and therefore all simulations were run in this way going forward.

4.4 Improving protein-protein docking

Optimisation experiments have so far been carried out on the BnWT:BspWT interaction in order to increase the number of native poses (RMSD $\leq$ 2.5 Å from the crystal structure) generated. Initial docking experiments conducted as part of a different project had utilised local installs of PatchDock and FibreDock and found them to be extremely user unfriendly; therefore the first thing we did was to find an alternative docking program to use.

We first tried using a local install of RosettaDock as the docking module features a constraint flag that we thought could be used to factor in our HDX data into the docking process. However, we soon found that RosettaDock is not optimised to run with a large amount of constraints and that above a certain threshold number (well below the amount we needed) it would become unstable and error out. We could have still used RosettaDock for the un-constrained docking we planned to do alongside the HDX-marked docking for comparison, however we decided that it would be better if these two different docking modes were done using the same program to allow the results they generated to be more comparable.

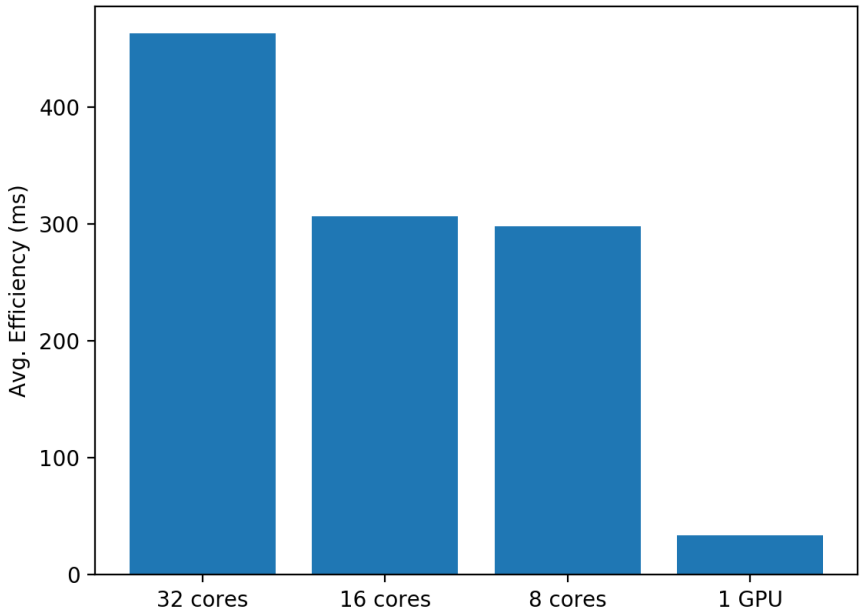


Figure 4.3: Optimisation of MD simulation efficiency. Graph showing how the average efficiency of running MD simulations varies with different hardware configurations; lower is better. CPU configurations run on a local Linux workstation, GPU configuration run on an HPC cluster.

Therefore, we needed to find a docking program that was explicitly set up with the use of constraints in mind. HADDOCK seemed the obvious choice here as the encoding of information from identified protein interfaces into AIRs is its main selling feature. HADDOCK also conveniently offers a webserver-based version of its docking program which would simplify its use on our end immensely. Therefore, with our new docking program selected, we turned to optimising the conditions under which the program was run. There were 3 different conditions we focused on during this phase: the CI cut off for the data that would be used to generate the AIRs, the number of structures that would be generated during initial rigid body docking and lastly the number of those initial structures that would undergo flexible refinement and subsequent refinement in explicit solvent.

Starting with the CI threshold, we varied using those residues surpassing the 99 %, 99.9 % & 99.95 % CI data in order to either broaden or narrow the number of residues that became AIRs, with the 99 % CI including most residues in the data set and 99.9 % CI and 99.95 % CI subsequently including only those residues with the strongest uptake difference. All data sets were generated using 1,000 rigid body structures with the top 200 by score selected for subsequent refinement. Interestingly, we found an inverse relationship between the strictness of the CI cut off and the number of native poses generated. Figure 4.4 A showed that the 99.95 % CI generated no native poses, with a large gap in RMSD between two poses at 2.9 & 3.1 Å and the rest of the data set. The 99.9 % CI (Figure 4.4 B) generated 1 native pose with a distinct gap in RMSD between a more native-like cluster ending at 3.6 Å and the rest of the data set. The 99 % CI (Figure 4.4 C) performed by far the best with 6 native poses with a much smaller gap in terms of RMSD between the more native-like cluster ending at 3.1 Å and the rest of the data set. We believe that the reason for this inverse relationship between CI cut off and number of native structures is that the stricter cut offs exclude certain residues that, despite having lower uptake, are nevertheless important for the interaction and therefore HADDOCK is less able to correctly dock the proteins in the correct orientation.

With a CI cut off of 99 % settled on, we next went about optimising the number of rigid body and refined structures to be generated. All data sets so far had used a 1,000/200 split between these two parameters, so our first attempt involved increasing the number of structures that were refined from only the top 200 to all 1,000. This would enable us to test the efficacy of HADDOCK's score function by seeing if any poses out of the top 200 were in fact native. Figure 4.4 D showed us that the score function appears to be quite accurate as again, only 6 native poses are found. With increasing the number of refined structures not generating any additional native poses, we moved on to increasing the number of rigid body structures that are initially generated in order to increase the amount of sampling of the conformation space of the complex. For this data set we used a 5,000/200 split and found that the actual number of native poses generated went down (Figure 4.4 E) to only 3. This shows that there is quite a bit of variety in the poses that are generated by HADDOCK, even when using HDX data as AIRs and so we may have to potentially run several replicates in order to get a true idea of the efficacy of a certain set of conditions.

Lastly, we attempted to boost the number of native poses by going as high in terms of sampling as was practical for a single data set with a 5,000/1,000 split (Figure 4.4 F). In this data set we saw our best

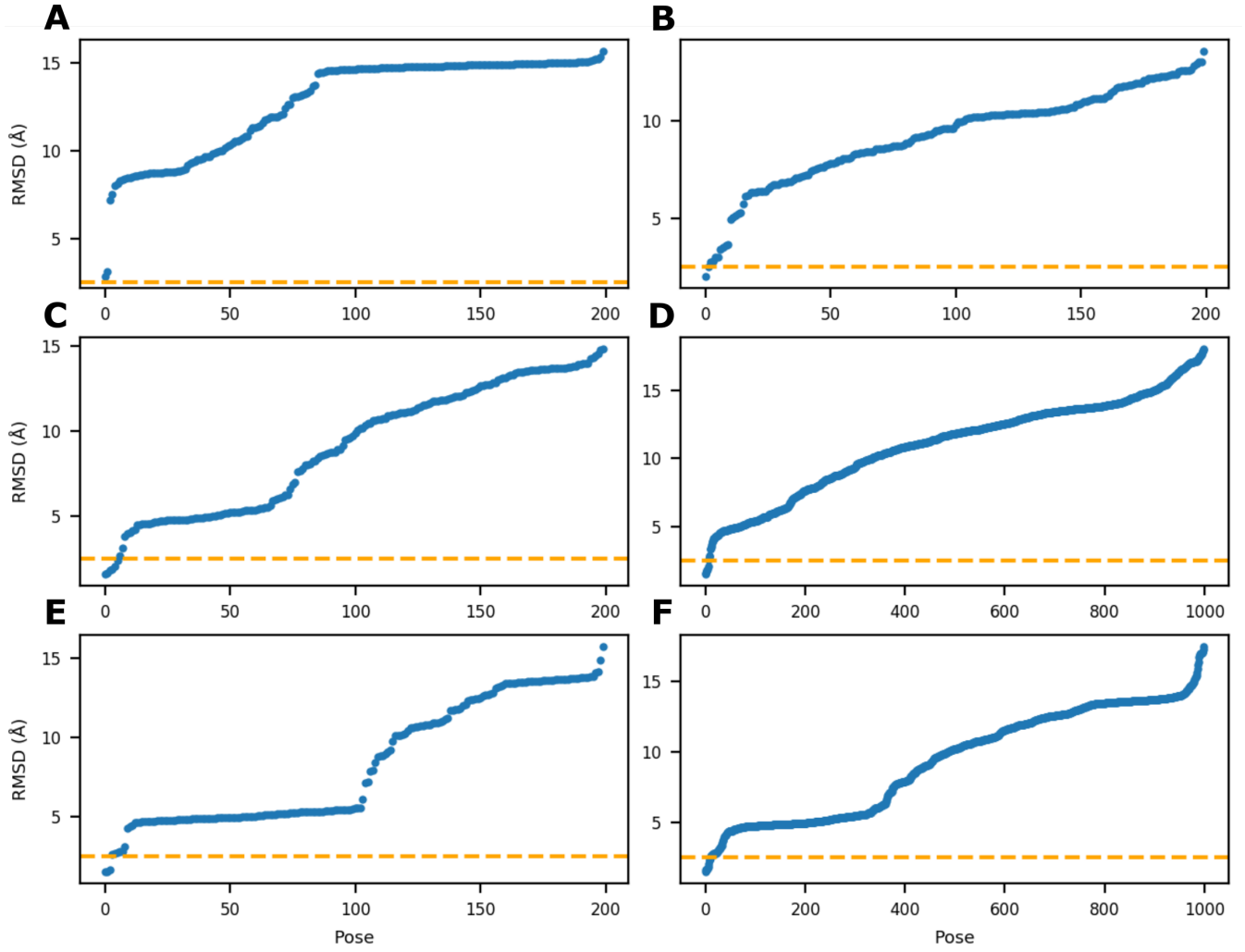


Figure 4.4: HADDOCK docking optimisation. Plots showing the effect of various different parameters on the ability of HADDOCK to successfully dock the BnWT:BspWT interaction. Dashed orange lines show the 2.5 Å cut off for a pose to be considered native. **(A)** AIRs constructed using all residues exceeding the 99.95 % CI, 1,000 rigid body poses and 200 refined poses. **(B)** AIRs constructed using all residues exceeding the 99.9 % CI, 1,000 rigid body poses and 200 refined poses. **(C)** AIRs constructed using all residues exceeding the 99 % CI, 1,000 rigid body poses and 200 refined poses. **(D)** AIRs constructed using all residues exceeding the 99 % CI, 1,000 rigid body poses and 1,000 refined poses. **(E)** AIRs constructed using all residues exceeding the 99 % CI, 5,000 rigid body poses and 200 refined poses. **(F)** AIRs constructed using all residues exceeding the 99 % CI, 5,000 rigid body poses and 1,000 refined poses.

results yet with 10 native poses. It is perhaps not surprising that the number of native poses increased with greater sampling, however why we saw this with the 5,000/1,000 split and not with the 5,000/200 split when previous results indicated that HADDOCK's scoring function could accurately reflect the top rigid body poses is not fully understood. The most likely explanation is simply that of the semi-random nature of the docking, even with AIRs, and that were the data sets to be repeated we would likely get slightly different results. The reason this was not done at this time was one of resource allocation: we deemed it would be more beneficial to start optimising and producing data sets for other interactions rather than try and optimise BnWT:BspWT above what has already been presented. When looking at the plots together, we can see a distinct double sigmoidal pattern emerging in most of the RMSD distributions. This indicates that, at least for this data set, HADDOCK has a preference for certain RMSD ranges which may be the result of the AIRs used (e.g. we can see the first plateau around 5 Å and the second plateau around 13 Å in the three data sets using the 99 % CI AIRs). It is likely that these preferred ranges correspond to two stable conformers that are therefore more populated than other less sterically favourable ones.

4.5 Exploring the boundaries of modelling protein conformation using HDXsimulator

A large variety of different methodologies were tested in order to map and analyse HDXsimulator's capabilities and limitations as to how it responded to lnP values that were artificially error-laden. This was done in order to determine the relationship between the deviation of the erroneous lnPs and subsequent RFU values calculated for each decoy in a data set from their original values and the eventual AUC score assigned to that particular data set. Our goal in doing this was to help us improve future versions of HDXsimulator by having a clear and logical understanding of the factors that influence the relationship between AUC and the RMSE/R^2 of the original reference compared to the error. Our expectations were to see AUC values decrease as error between original and erroneous lnPs/RFUs increased, as represented by an RMSE/R^2 metric. This would make sense as, if HDXsimulator was calculating lnPs correctly, deviation from the originals should result in the ROC curves being less able to accurately classify the structures the lnPs belong to as being native or not. Initially, we thought this would be a relatively simple affair, however we soon discovered that linking AUC to RMSE/R^2 in the manner we envisioned was not quite so straight forward.

4.5.1 Error generation through a Gaussian distribution

Our first method of error generation was to vary the true lnP values (as calculated by HDXsimulator using optimised scaling factors from the pseudo-crystal structure) using a Gaussian distribution around a certain level of standard deviation. The standard deviations chosen were 1, 2, 3, 4, 5 & 6 and three replicates at each standard deviation were calculated. This code is available in Appendix J. The pipeline, as described in chapter 2.6.2, was run on 7 different data sets: BnWT, BspWT, GFP, GFP-nb & GFP-nbmin using Rosetta-generated decoys and BnWT & GFP using 3DRobot-generated decoys. We found that lnP error generation using this method was not particularly effective in producing a wide range of AUC values, with almost all standard deviations in the Rosetta data sets producing AUC values of approx. 1 on both the lnP and RFU levels despite a good distribution of R^2 values from approx. 0.3-0.9 (Figure 4.5 A). AUC values for the two 3DRobot data sets were lower, however they also did

not see any substantial variation with standard deviation, with all data points producing similar AUC values regardless of R^2 value on both the lnP and RFU levels (Figure 4.5 B).

If standard deviations of 1-6 were not sufficient to cause errors large enough to distinguish between native and non-native conformations, our next step was to see if increasing those errors even further might be. We therefore upped the standard deviations about which the error was calculated to include 6, 8, 10, 12, 14 & 16 and re-ran all of the previous data sets. While the frequency of AUC values substantially below 1 did increase using this method, particularly with the RFU-level data, there was no discernible correlation between higher standard deviations (therefore lower R^2 scores) and lower AUC on either the RFU or lnP level (Figure 4.5 C). Those replicates that did display low AUC values appeared to be almost random. It was postulated that a potential reason for this maybe that, because the value of the errors are intrinsically related to the true values (as they are varied around them), the method may retain a kind of "memory" of the true values and therefore be mostly unable to distinguish native from non-native structures. This idea was given more credence when, upon testing the method using completely random values, we saw AUC values of approx. 0.5, exactly what we would expect (data not shown).

4.5.2 Error generation through shuffling

In order to test this theory, we decided to use a different method of error generation, one that would not retain as much "memory" of the true values. In this method, rather than changing the values themselves, we instead shuffled the true values within certain limits. This system was set up much like the Gaussian error generation, except instead of standard deviations determining the severity of the error, it was the number of places within which the value could be shuffled. E.g. equivalent to a standard deviation of 1, using this new methodology a value could be shuffled up or down the list of lnPs by 1 position, equivalent to bestowing the lnP of a residue upon its neighbour. This would lead to files where the true values were almost in the correct position but not quite. On the other end of the spectrum, a value could be shuffled up to 12 positions from its original place, leading to files with values that were in very different places from the original. This code is available in Appendix K.

To evaluate this method, we tested two data sets, shuffling 1-6 positions as well as 2-12 positions (in 2 position increments) on the BspWT_Rosetta data set. In addition to R^2, we also started measuring the RMSE of the error values vs. the true values for all subsequent data sets, in order to see if any differences between these two popular methods could be determined. The 1-6 position shuffle data set was the first to see AUC values substantially below 1 and correlated to decreasing R^2 (and increasing RMSE), but only for data on the lnP-level. A line of best fit showed a general decrease in AUC as the degree of shuffling increased on the lnP-level with AUC values decreasing to approx. 0.85 in those files that had been shuffled 6 times (Figure 4.5 D). Data on the RFU-level showed almost no changes at all with increasing error. Interestingly, the 2-12 position shuffle data set (Figure 4.5 E), while showing similar performance in terms RFU/lnP-level data, did not show this same trend despite even greater amounts of deviation from the original. Because the maximum values for R^2 and RMSE were very similar between both these data sets despite one being shuffled to a far greater extent than the other, it was clear that once a certain amount of shuffling had occurred (e.g. approx. 6 places), the data set was

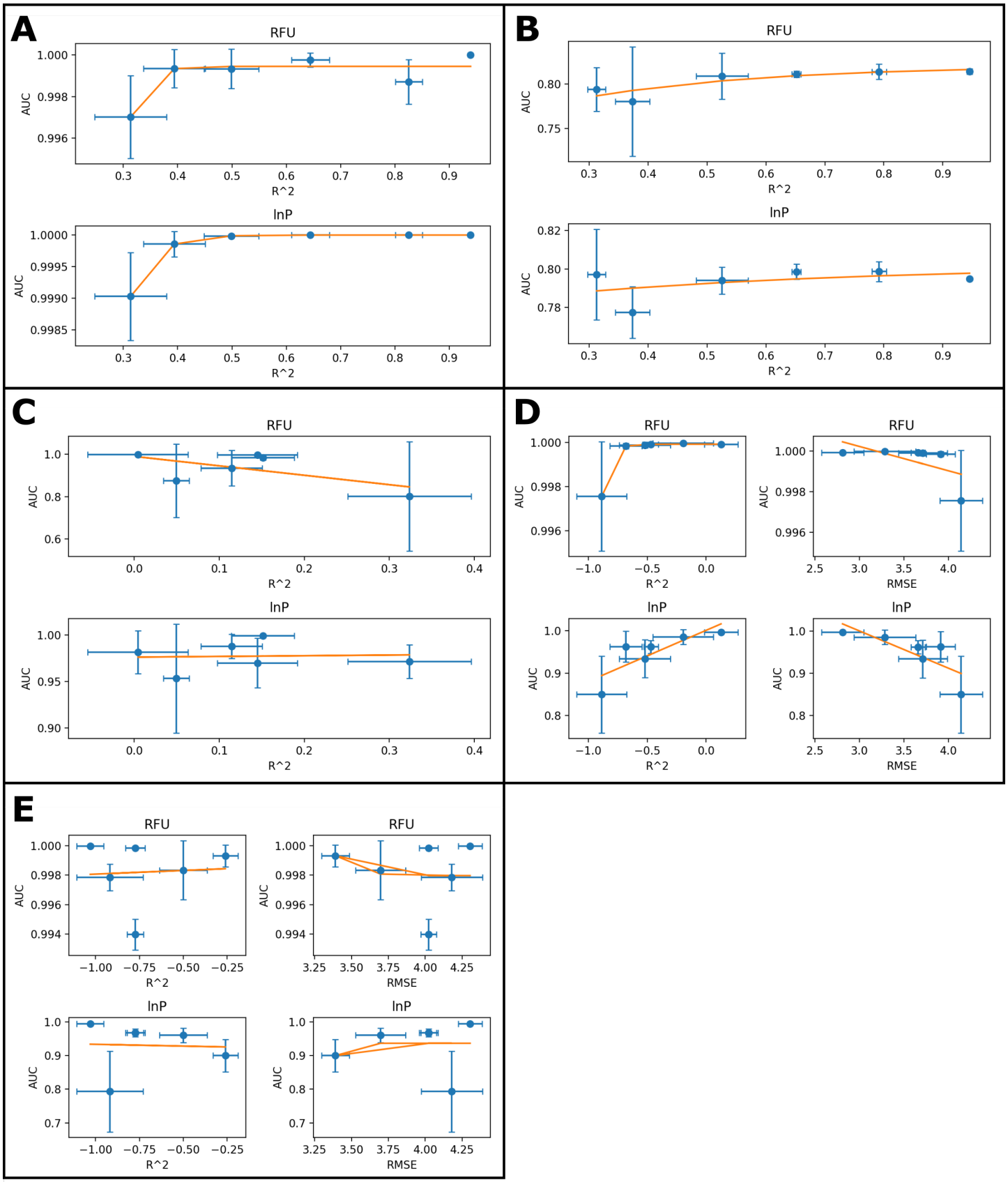

A
RFU
AUC
R^2
lnP
B
RFU
AUC
R^2
lnP
C
RFU
AUC
R^2
lnP
D
RFU
AUC
R^2
RMSE
lnP
E
RFU
AUC
R^2
RMSE
lnP

Figure 4.5: Exploring the capabilities of HDXsimulator part 1. Representative scatter plots demonstrating the effect of different methods of lnP error generation on the ability of HDXsimulator to differentiate between native and non-native structures. Each data point represents the average value of 3 replicates, with error bars to show the standard deviation for both AUC and R^2/RMSE values. The orange line represents the line of best fit. **(A)** Effect of errors generated by a Gaussian distribution about the true values with standard deviations of 1-6 on a decoy set generated by Rosetta. **(B)** Effect of errors generated by a Gaussian distribution about the true values with standard deviations of 1-6 on a decoy set generated by 3DRobot. **(C)** Effect of errors generated by a Gaussian distribution about the true values with standard deviations of 6, 8, 10, 12, 14 & 16. **(D)** Effect of errors generated by shuffling the true values 1-6 positions from their original locations. **(E)** Effect of errors generated by shuffling the true values 2, 4, 6, 8, 10 & 12 positions from their original locations.

already as different from the original as it could get and further shuffling did nothing, mathematically speaking. We therefore concluded that the differences between the data sets came down to the random chance of the shuffling and that in order for us to perceive trends, we would need to increase our sample size substantially.

We decided to make 2 changes in order to try and achieve this. The first was to switch from semi-random shuffling within a range to completely random shuffling in order to enable the full distribution of possible errors. This was done in conjunction with the second change which was to increase the number of replicates we generated. Previous data sets included only 18 data points (3 replicates of 6 different standard deviations/shuffling), whereas now each data set would include 1,000 data points. It was thought that these modifications would enable us to sample the full spectrum of possible lnP errors and so allow a greater range of AUC values to be generated. Results indicated that this had been achieved as for the first time we saw not only a considerable amount of AUC values substantially below 1, but also a correlation between AUC and both R^2 and RMSE, but only on the lnP-level (Figure 4.6 A). RFU data displayed some values considerably below 1 but no correlation between AUC and R^2/RMSE. The histogram showed that on the lnP scale, while the majority of the AUC results remain high (approx. 650 are above 0.8), the rest tend towards 0.5, indicating that for these error files HDXsimulator was not able to distinguish native from non-native structures reliably, a first for this methodology. Furthermore, in the scatter plot, we also for the first time saw a correlation between the AUC value calculated for each lnP error file and its divergence from the true lnP values, as measured by both R^2 and RMSE.

Now that we had had some success with error generation, we decided to try and increase the resolution of the data we were generating for each protein data set. This would take a form similar to how higher resolution data was produced by HDXmodeller by splitting the protein data set into different domains and running the pipeline on those domains individually rather than the protein as a whole. This was done because we wanted to have a better understanding of what parts contributed towards producing a superior data set and which parts contributed to producing an inferior data set. Therefore, we modified the code of the pipeline to consider only residues within a certain range when calculating both AUC and R^2/RMSE e.g. residues 1-20 or 20-30 instead of the whole protein. Data sets were then generated as before with each being focused on only one domain of the protein under investigation, resulting in a much higher resolution view. In the case of BspWT_Rosetta, we split the protein up into 10 amino acid domains, from 1-10, 11-20 etc. and the results confirmed that different domains of the protein did indeed react differently to the pipeline with termini seeing relatively little variance in AUC values (Figure 4.6 B) and the central domains showing high variance (Figure 4.6 C). Interestingly, whereas in previous data sets both R^2 and RMSE metrics showed correlation to AUC on the lnP level, for these domain data sets R^2 was seemingly unable to differentiate between AUC values. However, this is likely because a few results have very low R^2 values (-2.5 and lower), which is pulling the graphs to the left and causing the data points to appear to not have correlation. This can be seen by comparing the x axis scales for the R^2 values in Figure 4.6 B & 4.6 C.

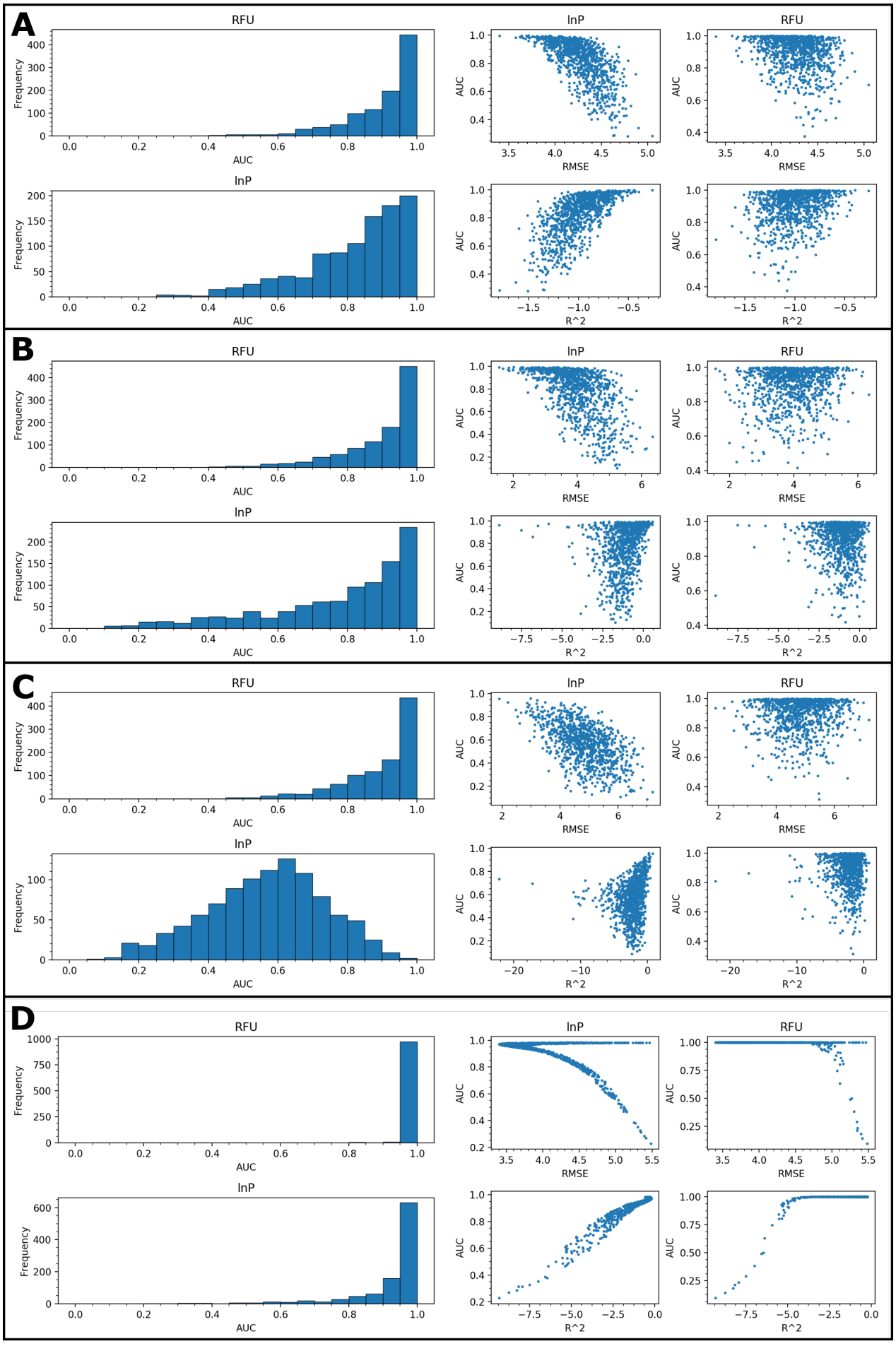

A
RFU
Frequency
AUC
lnP
RMSE
R^2
B
C
D

Figure 4.6: Exploring the capabilities of HDXsimulator part 2. Representative histograms and scatter plots demonstrating the effect of different methods of lnP error generation on the ability of HDXsimulator to differentiate between native and non-native structures. **(A)** Effect of errors generated by random shuffling of the true values. **(B)** Effect of errors generated by random shuffling of the true values. Data shown for a terminal region of the BspWT_Rosetta data set. **(C)** Effect of errors generated by random shuffling of the true values. Data shown for a central region of the BspWT_Rosetta data set. **(D)** Effect of errors generated by varying the βC & βH scaling factors within HDXsimulator to produce "incorrect" data sets.

4.5.3 Error generation through HDXsimulator

We next decided to try and improve the degree of correlation seen between the AUC and R^2/RMSE values. In all data sets thus far, while some correlation was present, there was still a wide range of potential R^2/RMSE values that could give rise to any particular AUC value and we wanted to narrow this as much as possible. In order to accomplish this, we attempted a very different form of error generation than anything we had tried before. Instead of using HDXsimulator to generate a "true" lnP data set and then attempt to introduce errors into it, instead we modified two scaling factors within HDXsimulator and then used the program itself to generate "incorrect" data sets for comparison against data generated using the default parameters determined by Best & Vendruscolo [30] to be optimal. This was done as stated in chapter 2.6.2. Using this methodology, we could finally see a much stronger correlation on the lnP-level between AUC and both R^2 and RMSE with much less variation at any given value then had been seen before (Figure 4.6 D). An odd trend we noticed however was the tendency for RMSE values to have a "plateau" or "hook" at AUC values close to 1 that remained entirely separate from the rest of the trend. Fortunately, the R^2 values do not exhibit this character to nearly the same degree and so we leaned towards using R^2 as the metric of choice in the future.

Of all the error generation methods we tried, only using HDXsimulator itself produced the kind of strong correlation between AUC and R^2 that we were looking for. We believe that this method of error generation succeeded where the others failed because it inherently altered the actual method of lnP calculation whereas the others simply modified true values that had already been calculated. To explain, the two scaling factors within HDXsimulator we varied to introduce synthetic error into the lnP values calculated for any given residue were βC, the distance in Å that contacts with other residues were considered, and βH, the distance in Å that hydrogen bonds with other residues were considered. By varying these values away from the defaults that Best & Vendruscolo calculated to be optimal, we produced errors that nevertheless fundamentally considered the tertiary structure of the decoy from which they were produced. As the tertiary structure of the protein is primarily responsible for determining any given residues lnP, the errors we produced using this method can be considered an authentic depiction of the decoy's lnPs, only calculated using a suboptimal expression. Therefore it is not surprising that such a method would be able to show a strong correlation between the AUC value and R^2 whereas other methods of error generation, which had no regard for the tertiary structure of the decoys, could not.

To summarise, we intended to determine the relationship between the deviation of erroneous lnPs/-subsequent RFU values and the eventual AUC score calculated for a particular data set, with our goal being to chart HDXsimulator's capabilities and limitations to enable us to improve future versions of the program. We expected to quickly see a relationship between increasing RMSE/R^2 and decreasing AUC, however observing this relationship proved more challenging than originally anticipated. Our initial plan was to use Gaussian error to vary the true lnP values, however we quickly found that no matter how much error we introduced, the method retained a "memory" of the true values and would still be able to distinguish between native and non-native structures. If judged by these results alone, it would indicate that the lnPs generated by HDXsimulator were irrelevant to the overall success of the method which couldn't be true. Hence, we set about an extensive optimisation program in order

to try and find a method of error generation that would show a difference in outcome based on the disparity of the lnPs from the true values. After much trial and error, we found that optimal method for error generation was using HDXsimulator itself utilising suboptimal scaling factors which produced the correlated AUC and R^2 values that we were looking for. The results presented here tell us that the relationship between AUC and RMSE/R^2 is much more complicated than we had initially believed, with hidden factors challenging our understanding of the boundaries of modelling protein conformation using HDXsimulator. With a suitable method for correlating AUC and error thus determined, we proceeded with one final test to see how this method of error generation affected AUC on the subdomain level. This work is detailed in chapter 5.6.

4.6 Summary

Over the course of this thesis, numerous different aspects of the project required extensive optimisation in order to bring them up to a standard that could report meaningful data. Those aspects, detailed above were: improving the yields of barstar and barnase, benchmarking MD simulation performance on different types of hardware, improving native structure output of protein-protein docking simulations and finally exploring the boundaries of modelling protein conformation using HDXsimulator.

Barstar and barnase were important for this project because the large amount of literature data available for their interaction (and that of mutants) enabled us to gather two sets of data using the same production/HDX protocols, thereby saving time. Unfortunately, neither of these proteins are commercially available, necessitating their production and purification in-house. While at one time very popular to study [92–97], these two proteins have since fallen by the wayside of academic interest and hence what protocols exist are optimised for production methods of many years ago. When attempting to replicate the protocol of the Ikura group (complete with gifted plasmids), we found it impossible to produce comparable yields to what the protocol claimed. Hence we had to conduct extensive optimisation experiments in the case of barnase and outright abandon the provided plasmid in favour of another in the case of barstar in order to produce enough protein for our needs.

The optimisation steps carried out for barnase that led to an increase in yield included: increasing the incubation temperature to 37 °C and agitation rate to 220 rpm, re-transforming the original plasmid into BL21-AI cells (specialised for toxic proteins), pelleting the culture and treating with acetic acid in order to release the protein from the periplasm and finally purification using a linear gradient of 0-1 M NaCl in 50 mM Tris HCl pH 8 over a 5 ml HiTrap SP HP cation column. These optimisations increased our yields from 0.5 mg/L to 3.08 mg/L, a 6-fold increase.

For barstar, the optimisation steps carried out that led to an increase in yield included: *de novo* synthesis of the barstar gene and insertion into a pET-28a vector with a N-terimal cleavable His-tag, increasing incubation temperature to 37 °C and agitation to 220 rpm, cell lysis by cell disruptor and purification using a linear gradient of 50mM Tris HCl pH 8, 300 mM NaCl & 300 mM imidazole. Our attempts to produce our own barstar ended with the optimisation of the cleavage of the His-tag using TEV protease, in which we made good progress but not enough produce significant amounts of cleaved protein in the time we had remaining. Therefore we commissioned the protein production facility at Queen Mary's University of London to finish optimisation/production of cleaved barstar using the protocol we had developed thus far as a basis.

Numerous MD simulations needed to be run during this thesis in order to relax bound PPI structures for subsequent docking. Therefore we decided to optimise our hardware usage in order to maximise the efficiency of production in terms of simulation time per nanosecond per atom and so produce the required simulations as quickly as possible. We found that, given the hardware we had access to, the most efficient means of production was to run simulations one by one using a NVIDIA TESLA V100 GPU. This produced an Efficiency (Equation. 4.3) of 35 ms per nanosecond per atom, which was almost 10x greater then the Efficiencies of using various different CPU configurations.

For the docking itself, optimisations were carried out in order to enrich the number of native poses (RMSD $\leq$ 2.5 Å) generated by the simulations. Being able to produce a certain percentage of native poses (approx. 2 % of total) was a requirement as otherwise we would not be able to identify whether the method could correctly classify poses as being native or not. Initial attempts were made to use RosettaDock, however we found that its constraints flag feature was not suited for use with the large number of constraints identified using HDX and so we moved on to a program more suited for use with constraints: HADDOCK. We found our best results using this program came when we used a CI cut off for the HDX data incorporated into the AIRs of 99 %, initial rigid body sampling of 5,000 poses and subsequent flexible refinement of the top scoring 1,000 structures. Using these parameters we were able to generate 10 native structures out of those 1,000 refined poses, a rate of 1 %. Clearly more optimisations will be required however this is a good start.

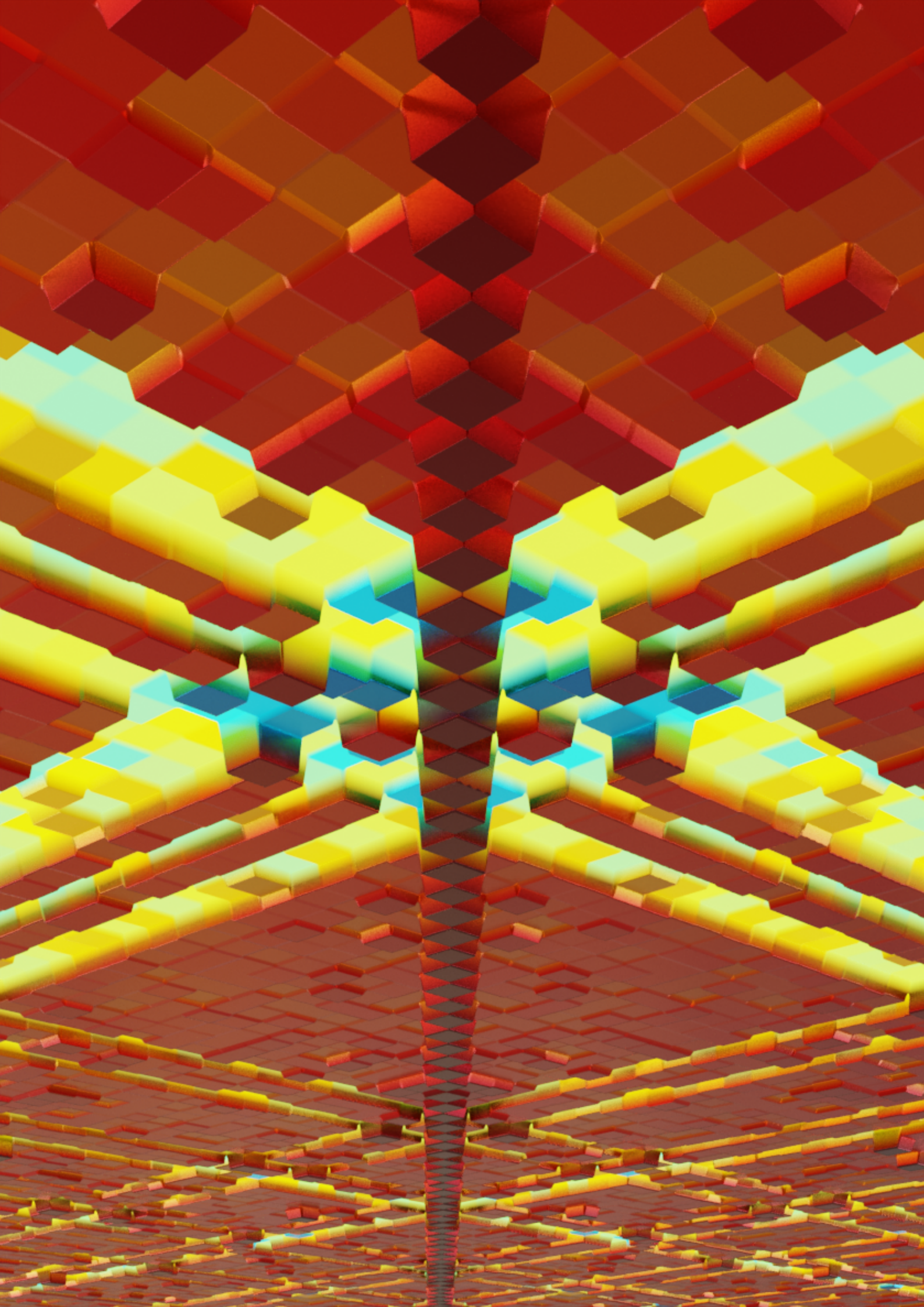
Finally, we explored the boundaries of modelling protein conformation using HDXsimulator. This was done to map and analyse how HDXsimulator responded to lnP values that were artificially error-laden, in order to determine the relationship between the deviation of the erroneous lnPs and subsequent RFU and eventual AUC values calculated for each decoy. Our goal in doing this was to help us improve future versions of HDXsimulator by having a clear and logical understanding of the factors that influence the relationship between AUC and the RMSE/R^2 of the original reference compared to the error. We attempted a number of different methods in order to see the expected link between RMSE/R^2 and AUC:

The first was varying the lnPs about a Gaussian distribution within defined standard deviation cut offs. This method did not produce a meaningful decrease in AUC regardless of the amount of error introduced (measured by R^2 compared to the original values), counter to our expectations, especially as when the experiment was repeated with completely random values the AUC was approx. 0.5 as expected.

Next we tried not modifying the values themselves but instead shuffling the true values randomly a set number of places from their original location. This produced the first correlated decrease between RMSE/R^2 and AUC with those lnPs shuffled 6 places showing eventual AUC scores of 0.85. A further change was implemented to this shuffling methodology by changing to completely random shuffles (i.e. no place limit) but also upping the sampling substantially from 18 repeats to 1,000. This was done in order to sample the full spectrum of possible lnP errors and so allow a greater range of AUC values to be generated. This change had a substantial effect on our results with for the first time the generation of a large amount of AUC results below 0.8. It was also at this time that we modified our code to be able to

calculate values for individual protein subsections as opposed to just the whole protein as before. This technique allowed us to increase the resolution of our results, allowing for more detailed inferences.

Finally, we attempted to improve the correlation between the RMSE/R^2 and AUC values by overhauling our error generation methodology completely by switching to errors generated by HDXsimulator itself rather than by post-processing its data. This was accomplished by modifying two scaling factors used by HDXsimulator from values previously found to be optimal, so that the results calculated deviated while still fundamentally taking into account the tertiary structure of the decoy from which they were produced. This method produced the strongest correlation between RMSE/R^2 and AUC of all those tested, with much less variation at any given value then had been seen before. Therefore this was the method of error generation that was brought forward into the next stage of the thesis.



5 Results of experiments performed to enable the classification of protein structures

In this chapter, we will be examining the results of those aspects of the thesis which directly relate to our stated goal of ascertaining HDXsimulator's ability to classify protein structures as either native or non-native. Our aim is to show how these results build upon each other in order to build up to a final metric allowing us to quantify our technique's ability. We shall cover: the yields of barnase and the native MS validation experiments undertaken to demonstrate the character of this recombinant protein and the complex between it and barnase. The HDX-MS experiments conducted on all the binary PPIs investigated in this work and how their results are a valid representation of their interaction together for the purpose of bringing the data forward for use by our method. How we used the newly developed HDXmodeller tool to model residue-resolved lnP values for each interaction, enabling us to compare modelled lnPs with lnPs calculated from structures. The relaxed outputs of the MD simulations that will be used in future work to extend the methodology presented in this thesis on single proteins to that of binary PPIs. The docked protein-protein structures that take the relaxed MD structures as input and provide the eventual native structures against which decoy structures will be compared when this methodology is later extended. The results of our investigation into the boundaries of modelling protein conformation using HDXsimulator and finally the results of our inquiry into HDXsimulator's ability to distinguish between native and non-native structures, the primary goal of this thesis.

5.1 Production of barnase and validation of barnase and barstar

The improvements detailed in chapter 4.2 lead to an increase in BnWT yield to 3.08 mg/L culture. BnH102A was also produced using the exact same procedure as the WT protein and had a yield of 34.2 mg/L culture.

Native MS experiments were carried out as described in chapter 2.3 in order to check for the correct mass and therefore confirm the identity of the proteins we produced, as well as the formation of the complex. For BnWT (theoretical mass: 12,383 Da), intense peaks were seen in the native spectrum at 1,769.84 & 2,064.64 m/z, corresponding to an experimentally determined mass of 12,384 $\pm$ 0.55 Da. For BnH102A (theoretical mass: 12,317 Da), intense peaks were seen in the native spectrum at 1,760.4 & 2,053.63 m/z, corresponding to an experimentally determined mass of 12,323.45 $\pm$ 5.44 Da (Figure 5.1 A). For the BnWT:BspWT complex (theoretical mass: 22,587 Da), intense peaks were seen in the native spectrum at 2,510 & 2,824 m/z, corresponding to an experimentally determined mass of 22,587 $\pm$ 0.02 Da (Figure 5.1 B). The identity and binding of proteins not produced during this study were confirmed during the course of HDX-MS experiments.

With these experiments, we demonstrated the ability of barnase and barstar to fully bind to each other using native MS. This was in contrast to previous experiments on different constructs which indicated incomplete binding. Both barnases and the barnase:barstar complex were found to have experimental MWs that were almost identical to their theoretical MWs and complex displayed very intensive binding peaks with little visible unbound protein remaining. At this stage we considered the protein production stage of the project to have been completed successfully and could finally move on to the

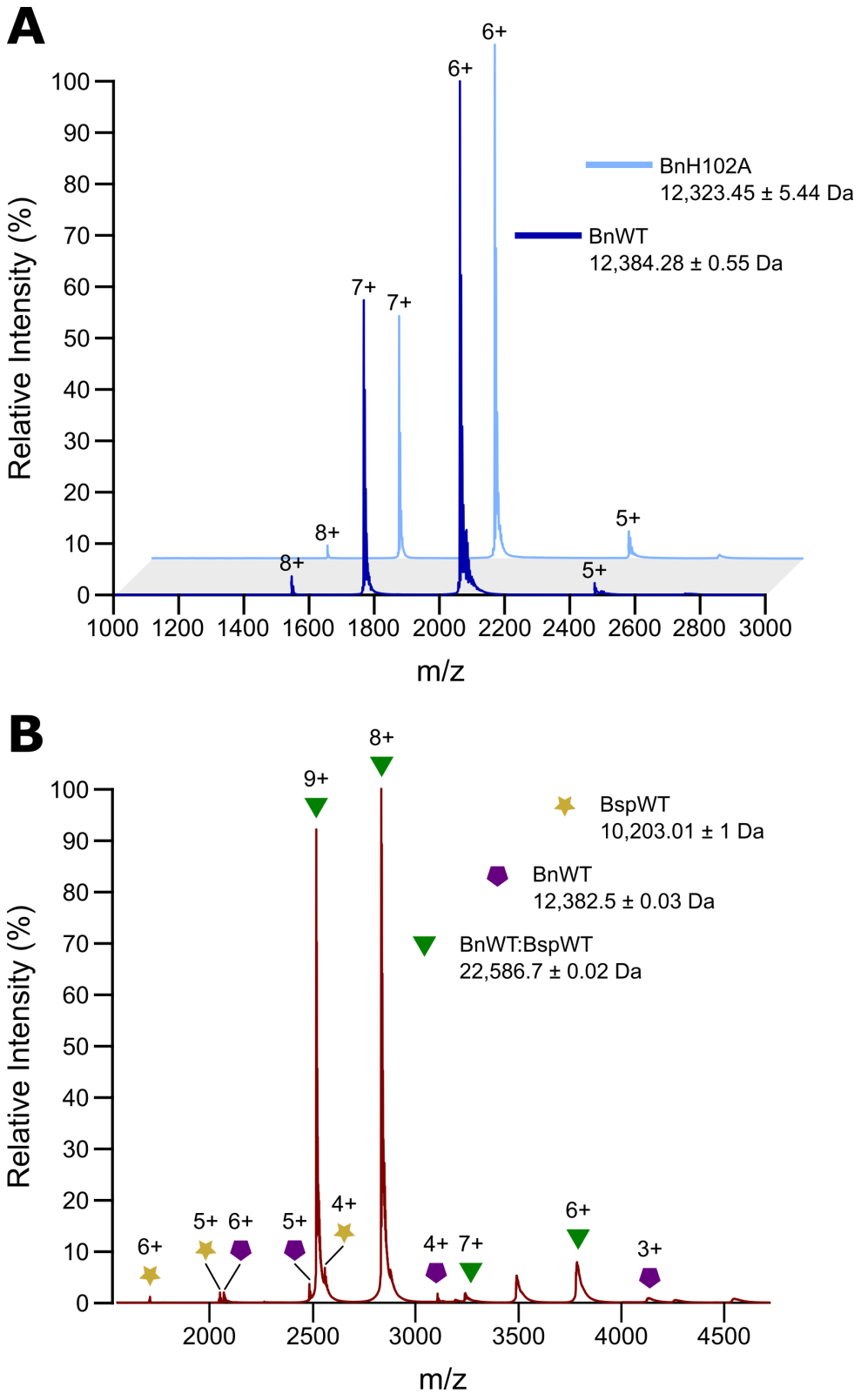


Figure 5.1: Native MS spectra of barnase and the barnase:barstar complex. (A) Native spectra for BnWT (dark blue) and BnH102A (light blue) displaying their experimental masses. Charge states annotated next to relevant peaks. **(B)** Native spectra for the BnWT:BspWT complex displaying constituent peaks for BnWT:BspWT (green triangle), BnWT (purple pentagon) and BspWT (yellow star). Experimental masses and charge states are annotated.

collection of novel data.

5.2 HDX-MS experiments on binary PPI protein complexes

The HDX-MS experiments we carried out over the course of this thesis were a means to an end rather than the end goal itself. All of the systems we used were, necessarily, already well characterised with extensive empirical data including crystal structures and therefore the point of us conducting these experiments was not to develop any new insight into the proteins or their binding interactions. Rather our HDX data was used in order to provide an experimental "reality" that the calculated data produced by HDXsimulator could be compared against in order to judge the veracity of the technique. Therefore, while our experimental data did not need to offer new insight, it did need to portray an accurate picture of the individual proteins and their interactions as understood by the scientific community that study them. If this standard was not met then any conclusions we drew about HDXsimulator, our main goal, would be inherently flawed. In analysing our HDX data, our primary concern was therefore comparison to the zeitgeist of the protein/system in general as well as our own knowledge of HDX outputs and how they correspond to each other. Statistics relating to the analysis of each individual protein in DynamX are available in Table 5.1.

5.2.1 The BnWT : BspWT interaction

5.2.1.1 BnWT

In DynamX, BnWT displayed extremely high levels of digestion and data quality was extremely good across the majority of the peptides with only a few that made it through the filtering process needing to be excluded from the final data set. On the peptide level (Figure 5.2 A) we could see that the first 7 N-terminal residues displayed high levels of RFU, characteristic of termini which are typically unprotected, however residues F7-Y13 saw a substantial drop in RFU which we attributed to being part of an α-helix and therefore having increased protection. Peptides including residues L14-L42 show behaviour typical of a well-protected region with the RFU values well spread out from the short to the long time points and little variation in values across the whole region. Peptides including residues A43-W94 display behaviour that is by comparison much more indicative of a less protected domain, with a much narrower spread of RFU values from the short to the long time points. Peptides including residues L95-K108 interestingly displayed the spread out RFU values characteristic of a protected region despite being a terminus found on the outside of the structure.

The exceptional coverage and redundancy with data quality to match of this data set lead us to be extremely confident that our results were accurate. This data set showed no symptoms of any unfolding and, when corrected for back exchange, most of the peptides displayed a good spread of RFU values with little overlapping at different time points. However, even when corrected for back exchange, we did not see RFU values approaching 1 in the majority of peptides, despite this being a relatively small protein with little internal volume and our longest time point being 8 hours. This was a common observance that we saw in the majority of the data sets collected for this thesis and did seem to have a correlation with protein size (i.e. larger proteins having lower average corrected RFU). Therefore, our explanation is that 8 hours was probably not sufficient in order to fully saturate the protein due

	Peptides	*Coverage (%)*	*Redundancy*	*BEX avg. (RFU)*	*Back exchange (RFU)*
BnWT	98	99.1	12.58	0.6	0.4
BnH102A	95	100	13.62	0.55	0.45
BspWT	49	88.6	7.26	0.65	0.35
BsY29F	47	90.9	6.67	0.6	0.4
GFP	151	94.9	8.77	0.45	0.55
GFP-nb	47	97.6	5.12	0.6	0.4
GFP-nbmin	57	99.3	6.53	0.6	0.4

Table 5.1: Statistics of protein data sets after analysis in DynamX. Table displaying various statistics related to the 7 individual protein data sets after the completion of analysis in the program DynamX. ”Peptides” describes the total number of peptides in the whole data set, ”Coverage” describes the percentage of the amino acid sequence covered by those peptides, ”Redundancy” describes the average number of peptides covering each amino acid, ”BEX avg.” describes the approximate average RFU of the back exchange control across all the peptides & ”Back exchange” is derived from 1 minus the BEX avg. and describes the approximate amount of RFU lost to back exchange across all peptides.

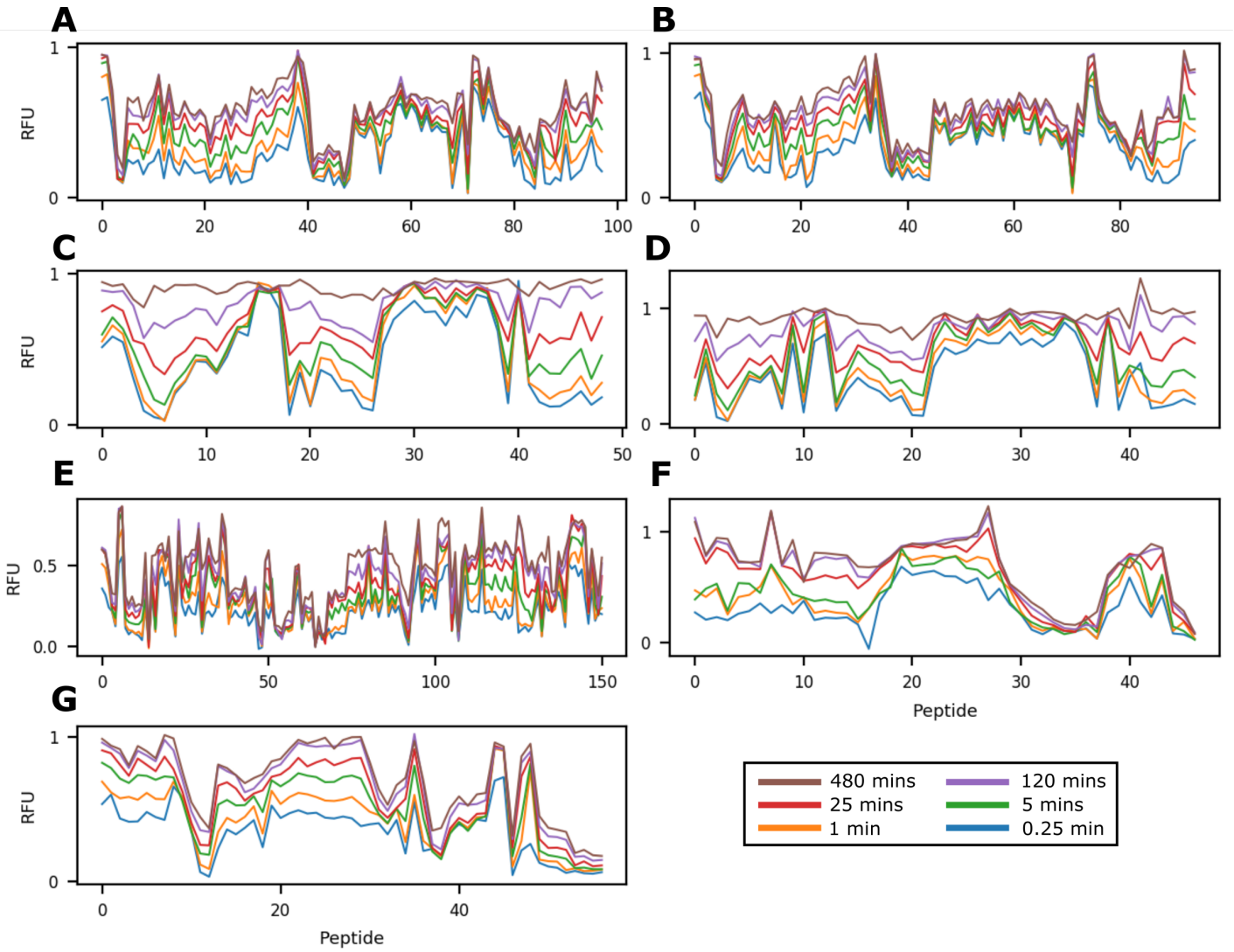


Figure 5.2: RFU of individual proteins after HDX analysis. Plots describing how the degree of deuterium uptake in individual peptides of a protein changes over varying deuteration times. RFU values shown for proteins in the unbound state with results corrected for extraneous exchange. **(A)** BnWT, **(B)** BnH102A, **(C)** BspWT, **(D)** BsY29F, **(E)** GFP, **(F)** GFP-nb, **(G)** GFP-nbmin. Deuteration times displayed in the legend.

to protection and that had a longer time point been used we would have likely seen saturation in the majority of peptides. The reason this was not done at the time is one of practicality; each experimental run already took around 40 hours to complete assuming no machine faults (which were common) and so the idea of having more and longer time points did not make practical sense.

From the perspective of BnWT, the interaction with BspWT brought about very significant changes across most of the protein (Figure 5.3 A). Woods plots revealed that the N-terminal region between residues A1-D12 was the only region to see no difference at all between the states. Then starting from residue Y13, the sum difference between the bound and unbound states began to increase to a local maxima of -7.27 Da around residue G40 before seeing a slight drop off to -4.89 Da around residue V45. From here, the sum uptake difference again began to increase to a maximum value of -11.23 Da and, staying put near this value until in between residues N58-D75. Uptake difference then fell back sharply to below -2 Da in the region around residues N84-D93 before seeing another equally sharp rise to between -6 to -10 Da at the C-terminus. From these results we can see that binding to BspWT causes very significant differences across almost the entire length of the BnWT protein, though especially concentrated in the central domain and the C-terminus. Due to the very high amount of uptake difference, two different significance thresholds were calculated as described in chapter 2.2.1 in order assess the interaction. For the 99 % CI, a value of 1.74 Da was found which included almost every residue in the protein. Therefore in addition, the 99.9 % CI was also calculated and a value of 5.55 Da was found which eliminated all but the central and C-terminal domains.

5.2.1.2 BspWT

In DynamX, BspWT displayed good levels of digestion and data quality was fairly good across most of the peptides, however with a large number needing to be excluded from the final data set. On the peptide level (Figure 5.2 C) we could see that up until approx. residue E52, BspWT displayed a wide spread of RFU values which then became suddenly narrower approx. between residues E52-L71 before spreading out again. This region of low RFU variance had been an issue in earlier HDX tests on the barstar construct with an un-cleaved N-terminal His-tag. In those experiments, we noted that the peptides in this domain had unusually high RFU values even at low time points and that the difference between values at 15 seconds and 8 hours was virtually non-existent. This type of behaviour is characteristic of an unstructured random coil and was part of the reason we decided to abandon the original barstar construct in favour of the current cleaved version. While the current cleaved construct does display elements of this previous behaviour, it is to a much lesser degree as in this data set there is a clear distinction between the RFU values of the various time points, despite them being close together. We therefore decided that it was unlikely this region was unfolded and so this construct was acceptable to use.

From the perspective of BspWT, the interaction with BnWT again saw very significant changes across most of the protein (Figure 5.3 C). No coverage was available for the first 7 residues and Woods plots showed only a comparatively small amount of uptake difference was seen in the remainder of the N-terminal residues from E8-D15 of around -2 Da. After this point however, BspWT sees a sudden increase in uptake difference in the region from Q18-L34 with an average value of approx. -9 Da and

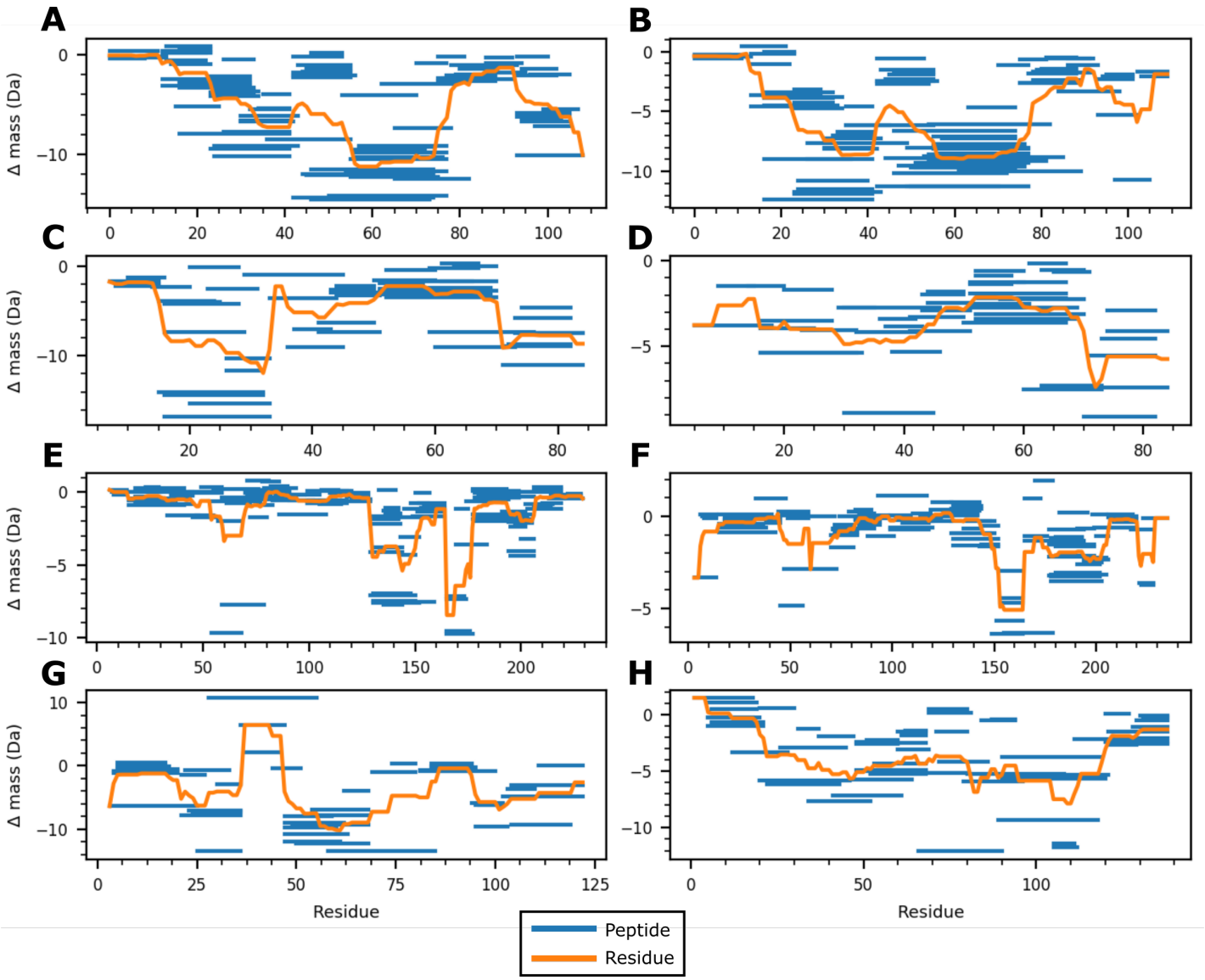


Figure 5.3: Δmass of individual proteins upon complexation with their binding partner. Woods plots describing how the mass of individual residues of proteins changes when in the presence of their binding partner in a binary PPI. Results are shown on the peptide level (experimental data, blue) where every bar is a discrete peptide, and on the residue level (calculated data, orange) where each data point is the average value of every peptide that covers that residue. **(A)** BnWT (BnWT:BspWT), **(B)** BnH102A (BnH102A:BsY29F), **(C)** BspWT (BnWT:BspWT), **(D)** BsY29F (BnH102A:BsY29F), **(E)** GFP (GFP:GFP-nb), **(F)** GFP (GFP:GFP-nbmin), **(G)** GFP-nb (GFP:GFP-nb), **(H)** GFP-nbmin (GFP:GFP-nbmin).

a maximum of -11.92 Da seen in residue N33. Uptake difference then falls to sharply to an average of approx. -4 Da in the region of residues D35-L71 before rising again in the C-terminal region of residues Q72-T85 to an average of approx. -8 Da. Like with BnWT, these results show that binding causes highly significant differences across almost the entire protein though especially concentrated in the region of Q18-L34 and the C-terminus. Like for BnWT, because the uptake difference values were so high, two different significance thresholds were calculated as described in chapter 2.2.1 in order assess the interaction. For the 99 % CI, a value of 1.71 Da was found which included every residue in the protein for which we had data for. Therefore in addition, the 99.9 % CI was also calculated and a value of 5.43 Da was found which eliminated the N-terminus and the central domain.

BspWT was one of only two data sets (the other being BsY29F) which saw deuterium saturation in the corrected RFU plots. This makes sense considering it is the smallest protein we investigated in this study at only 89 residues and therefore would have the least overall protection. While the data in terms of redundancy and quality was not as strong as barnase, we nevertheless consider this to be a perfectly viable data set, especially when compared to some of the problems we had faced with barstar prior to acquiring it. As described previously, the original barstar construct with an un-cleaved C-teriminal His-tag that we collected HDX data for displayed unusual RFU values around the region of V50-V70 that were consistent with that region of the protein being unfolded. In these data sets, peptides covering these residues had RFU vales that were very high and very close together at all time points which is characteristic of a random coil and not a folded protein. However no other region of the protein displayed this behaviour and circular dichroism analysis (data not shown) also indicated that the protein was folded correctly. On the basis of these results we were prepared to let this inconsistency slide, however upon running this barstar construct with barnase, we started to see problems indicative of a lack of binding between the two in certain data sets. Considering the barnase-barstar interaction is one of the strongest ever discovered (including most of the mutant interactions) [61], we knew that something must be wrong and so we endeavoured to replace the original construct with the cleaved N-terminal His-tag construct used in the final data sets. These data sets do not see the same problems around V50-V70 that the previous construct did (they are still quite high but the time points are more spread out) and upon complexation with barnase, the interactions were completely rescued, indicating that something had indeed been wrong with the previous construct which had now been fixed. Whether this problem was indeed a matter of unfolding or potentially something to do with the His-tag blocking the interaction interface (unlikely considering the tag was on the reverse face of the protein from the interface and should not have been able to get in the way) is unknown but regardless the issue has been fixed and so we were confident that the BspWT data set would be useable for computational analysis with HDXsimulator.

5.2.1.3 Visualisation of the interaction

Deuterium uptake differences exceeding the 99 % CI cut off for both proteins were mapped onto the structures of the proteins in their bound conformation (Figure 5.4 A). Data visualisation was achieved using the "Define Attributes" tool in UCSF Chimera with residues showing significant negative uptake difference values assigned a colour gradient from green (less significant) to blue (more significant). Significant positive uptake difference values (if any) were assigned a colour gradient from orange (less

significant) to red (more significant). Residues which showed no significance were represented as grey and those for which there was no coverage were represented in white. These results broadly match up with the crystal structure of the complex within the scope of peptide-resolution data, with those surfaces close to the interface between the proteins displaying higher levels of uptake difference whereas those surfaces on the opposite side of the proteins from the interaction showing comparatively much reduced uptake difference. Due to the small size of the proteins involved (especially barstar), it is not surprising to see significant levels of difference in areas not directly associated with the partner protein as it is common for differences to propagate to nearby areas. Based on these results and of the individual RFU plots, we propose that the HDX data collected for these proteins is a legitimate representation of the individual proteins and the complex as a whole and therefore is appropriate to take forward into the computational stages of this project.

5.2.2 The BnH102A : BsY29F interaction

5.2.2.1 BnH102A

In DynamX, BnH102A displayed excellent levels of digestion and data quality was very good across the majority of the peptides with only a small number needing to be excluded from the final data set. On the peptide level (Figure 5.2 B) we could see, unsurprisingly, a very similar picture to that painted for BnWT, including over the peptides covering the A102 mutated residue.

Like with the WT barnase, for BnH102A we had exceptional coverage and redundancy with data quality to match, leading us to be extremely confident that our results are accurate. This data set also showed no symptoms of any unfolding and, when corrected for back exchange, most of the peptides displayed a good spread of RFU values with little overlapping at different time points. Likewise, even when corrected for back exchange, we did not see RFU values approaching 1 in the majority of peptides.

From the perspective of BnH102A, the interaction with BsY29F displayed significant deuterium uptake difference across most of the protein (Figure 5.3 B). Woods plots revealed within this data set 3 distinct regions that could be demarcated from each other. The first could be found between residues Q15-A46 with an average uptake difference of approx. -6 Da and a maximum uptake difference of -8.63 Da. The second domain was between residues P47-S91 with an average uptake difference of approx. -6 Da and a maximum difference of -8.93 Da. The final region was found between residues S92-T107 with an average uptake difference of approx. -3.5 Da and a maximum uptake difference of -5.89 Da. Like for the previous data set, because the uptake difference values were so high, two different significance thresholds were calculated as described previously in order assess the interaction. For the 99 % CI, a value of 1.74 Da was found which included almost every residue in the protein except for the N-terminus and 3 residues near the C-terminus. Therefore in addition, the 99.9 % CI was also calculated and a value of 5.54 Da was found which eliminated most of the residues at the termini.

5.2.2.2 BsY29F

In DynamX, BsY29F displayed good levels of digestion and data quality was fairly good across most of the peptides, however with a large number needing to be excluded from the final data set. On the

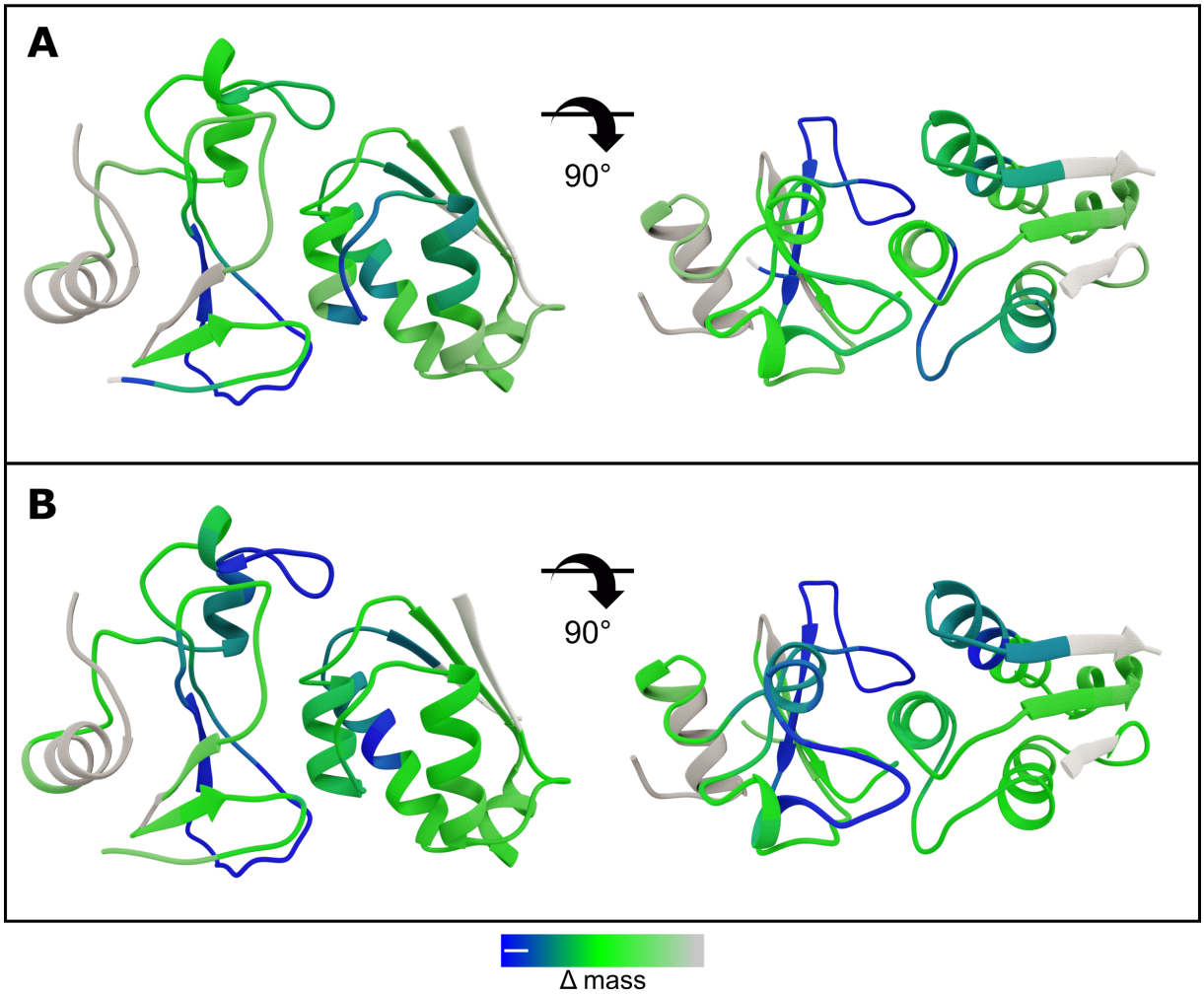


Figure 5.4: Barnase & barstar HDX-MS data mapped onto structures. Uptake difference data surpassing the 99 % CI of the BnWT:BspWT and BnH102A:BsY29F interactions assigned to a colour spectrum and mapped onto their respective three-dimensional structures. The difference values represented by the colour gradient are assigned relative to the highest value in each individual protein's data set, hence colours are not necessarily comparable between structures. **(A)** The BnWT:BspWT interaction from the front (left) and the top (right). **(B)** The BnH102A:BsY29F interaction from the front (left) and the top (right).

peptide level (Figure 5.2 D) we could see that again this mutant showed very similar levels of uptake as the pWT protein did, including over the F29 mutated residue.

From the perspective of BsY29F, the interaction with BnH102A displayed significant uptake difference across the entire surface of the protein for which we had data for (Figure 5.3 D). Demarcating different regions was difficult for this data set, however Woods plots suggested 3 potential separate domains. The first could be found between residues N6-D15 with an average uptake difference of approx. -3 Da and a maximum uptake difference of -3.76 Da. The second domain was between residues L16-W53 with an average uptake difference of approx. -3.5 Da and a maximum difference of -4.87 Da. The final region was found between residues E68-T85 with an average uptake difference of approx. -5 Da and a maximum uptake difference of -7.37 Da. Like for the previous data set, because the uptake difference values were so high, two different significance thresholds were calculated as described previously in order assess the interaction. For the 99 % CI, a value of 1.46 Da was found which included every residue in the protein which we had data for. Therefore in addition, the 99.9 % CI was also calculated and a value of 4.66 Da was found which eliminated all but two regions in the centre and at the C-terminus.

BsY29F was the only other data set (along with BspWT) which saw deuterium saturation in the corrected RFU plots. This makes sense considering it is the smallest protein we investigated in this study at only 89 residues and therefore would have the least overall protection. While the data in terms of redundancy and quality was not as strong as barnase, we nevertheless consider this to be a perfectly viable data set, especially when compared to some of the problems we had faced with barstar prior to acquiring it. Like for BspWT, initial constructs of BsY29F displayed symptoms of unfolding, however like with the new construct of BspWT, this new version of BsY29F does not see the same problems around V50-V70 that the previous construct did. Likewise, upon complexation with barnase, the interactions were completely rescued, indicating that something had indeed been wrong with the previous construct which had now been fixed.

5.2.2.3 Visualisation of the interaction

Deuterium uptake differences exceeding the 99 % CI cut off for both proteins were mapped onto the structures of the proteins in their bound conformation (Figure 5.4 B) as described previously. These results had a similar distribution to the WT:pWT interaction in terms of regions of significance and regions of insignificance. In terms of BnH102A, we do not see any changes around the H102A region compared to the BnWT data set as a result of the interaction with BsY29F, however significant differences can be seen around residues Y24-A43 which do show comparatively more uptake difference in this data set compared to BnWT. In terms of BsY29F, we did in fact see a change compared to BspWT around the Y29F region, with these residues having substantially reduced uptake difference whereas most of the other regions stayed largely the same. Like with the WT-pWT data set, it is difficult to make detailed inferences due to the small size of the proteins combined with the large surface area covered by the interaction interface. However based on these results and the individual RFU plots, we propose that the HDX data collected for these proteins is a legitimate representation of the individual proteins and the complex as a whole and therefore is appropriate to take forward into the computational stages of this project.

5.2.3 The GFP : GFP-nb interaction

5.2.3.1 GFP

In DynamX, GFP displayed good levels of digestion and data quality was very good across most of the peptides with only a relatively small number needing to be excluded from the final data set. The BEX was surprisingly inconsistent, with considerable variation seen as well as a very low average RFU across all peptides. On the peptide level (Figure 5.2 E) we could see that, similar to the BEX, RFU values were quite variable and in general very low with a maximum value across the whole data set of $<$ 0.45. Generally speaking, the N-terminal half of the protein displays a lower maximum RFU with time points being more condensed together in comparison to the C-terminal half of the protein which shows a slightly higher maximum RFU and an expanded set of time points.

From the perspective of GFP, the interaction with GFP-nb saw strong, localised uptake differences in a number of domains (Figure 5.3 E). Woods plots revealed that the N-terminal region of the protein showed no significant uptake differences until residues in the range of V55-F71 which saw a comparatively small uptake difference of between approx. -2 to -3 Da. The next area of the protein to see a significant uptake difference were residues between K131-A154 which displayed an average of approx. -4.5 Da with a maximum of -5.4 Da. This region was followed by a brief dip down to lower difference values which then increased very sharply between residues K166-L178 to see the largest values in the data set with an average of approx. -6.5 Da and a maximum of -8.46 Da. Only one more region showing any significance was found in the protein between residues L195-L207 with an average of -1.5 Da and a maximum of -2 Da. Due to the far more localised nature of this data set when compared to barnase and barstar, only the 99 % CI was calculated and a value of 1.82 Da was determined which included the 4 prominent regions outlined above.

5.2.3.2 GFP-nb

In DynamX, GFP-nb displayed acceptable levels of digestion and data quality was moderate across most of the peptides with quite a large number needing to be excluded from the final data set. On the peptide level (Figure 5.2 F) we could see that the first 47 N-terminal residues displayed a very widespread RFU distribution between 0.1-0.5 which then narrowed substantially between approx. residues W48-F69 to 0.3-0.55. Maximum RFU even at the longest time points then dropped considerably between approx. residues T70-Y95 to around 0.2 despite BEX values for this region remaining high, implying substantial protection in this region. C-terminal residues then return to a more spread out distribution as seen in the N-terminus.

From the perspective of GFP-nb, the interaction with GFP saw more widespread uptake differences similar to those seen in barnase and barstar (Figure 5.3 G). This is likely due to the protein's comparative small size and so the majority of the surface being involved in some way with the interaction. No coverage was available for the first 3 residues and residue Q4 showed a strong uptake difference of -6.37 Da. However, looking at the peptide map we believe this to be an erroneous data point as only one fairly long peptide covers this residue and this peptide also extends into a region of high uptake difference, hence its high value. The first legitimate region showing a significant amount of uptake

difference is in between residues R20-W37 with an average uptake difference of approx. -5 Da and a maximum difference of -6.26 Da. This is followed immediately by the only region of significant positive uptake difference (i.e. deuterium uptake increases upon binding) seen in all of the data sets recorded for the binary PPIs, between residues Y38-E47 with an average uptake difference of approx. 5 Da and a maximum of 6.45 Da. After this, two regions of significant negative uptake difference once again dominate the data set, with the first being between residues W48-S86 with an average uptake difference of approx. -6 Da and a maximum of -10.15 Da and the second being between residues Y96-H123 with an average difference of approx. -4 Da and a maximum of -6.9 Da. Like with the previous barnase and barstar data sets, because a large amount of the proteins sequence showed an uptake difference, two CI values were calculated as described previously, a 99 % CI of 2.01 Da and a 99.9 % CI of 6.41 Da. However, because the value of the 99.9 % CI was so high, it excluded several regions with nevertheless high uptake completely and so it was decided to only use the 99 % CI for this data set.

The GFP-nb data set was of lesser quality compared to the other proteins studied (with the possible exception of GFP-nbmin & barstar), though still perfectly usable. Similarities can be seen to barnase and barstar in terms of the majority of the protein being involved in the interaction in some way, although this is not surprising considering they are of similar size. Despite this, certain degrees of localisation could be seen with the reverse faces of the nanobody showing no significant uptake and regions of the highest uptake difference being located at the interface. GFP-nb was particularly interesting as it was the only data set in this entire thesis (with the exception of the φNM1 data set, see chapter 3.2) that displayed a region of significant positive uptake difference. This region, Y38-E47, is located at the binding interface and is indeed, according to the literature [66], directly involved in the interaction. This is highly unusual as regions directly involved in binding almost always see a negative uptake difference as the act of binding renders them more protected from hydrogen exchange thanks to a combination of greater hydrogen bond interactions and lower solvent accessibility among others. Regions of positive uptake difference are normally seen away from the binding interface and are usually allosteric changes brought about by binding, such as the reverse face of a transmembrane transporter protein involved in the rocker-switch mechanism of active transport [98]. This region is covered by three separate peptides, each of which display this behaviour so we are confident in its validity, however quite how a region manages to be involved in binding and have a positive uptake difference is somewhat a mystery.

5.2.3.3 Visualisation of the interaction

Deuterium uptake differences exceeding the 99 % CI cut off for both proteins were mapped onto the structures of the proteins in their bound conformation (Figure 5.5 A) as described previously. These results match up very well with what we would expect from the crystal structure, especially in the case of GFP. In comparison to the smaller proteins, GFP displays good localisation of uptake differences to those peptides at the interface, with almost all other regions showing no significant uptake difference. In terms of GFP-nb, the interaction is a little less clear cut, with uptake difference seen over most of the protein's surface, although with notably higher intensity at the interface compared to on the other side of the protein, with the exception of the poly-His tag. Interestingly, GFP-nb gives us the only example of significant positive uptake difference seen in any data set, in the region between Y38-E47. Overall, this data matches up well with what we would expect given the crystal structure, especially for GFP,

and so we propose that the HDX data collected for these proteins is a legitimate representation of the individual proteins and the complex as a whole and therefore is appropriate to take forward into the computational stages of this project.

5.2.4 The GFP : GFP-nbmin interaction

5.2.4.1 GFP

From the perspective of GFP, the interaction with GFP-nbmin again saw strong, localised uptake differences in a number of domains (Figure 5.3 F). Woods plots revealed that the N-terminal residues between G4-L7 showed a degree of significance like that which was seen in GFP-nb, however like for the nanobody we believe that these data points may be erroneous as they are the result of only a single peptide that does not match up with either of the other 3 peptides covering that region. We believe this to also be the reason for the single residue V61 showing a significant uptake difference. Therefore we believe that the first region showing significant uptake difference is between residues V150-N170 with an average difference of approx. -3.5 Da and a maximum value of -5.07 Da. This is followed by a region containing residues between S175-Q204 with an average uptake difference of approx. -2 Da with a maximum of -2.43 Da and lastly a small region between residues E222-I229 with an average uptake of approx. -2.5 Da and a maximum of -2.69 Da. As with the previous GFP data set, only the 99 % CI was needed due to its far more localised nature compared to what had been found in the smaller proteins. A value of 1.64 Da was calculated which included the 3 prominent regions outlined above.

GFP data sets for both its interactions with GFP-nb & GFP-nbmin showed good data quality and redundancy as well as the highest degree of specificity of all the interactions studied in this thesis with results being highly localized to the interface with very few exceptions. This allowed us to clearly demarcate between GFP:GFP-nb and GFP:GFP-nbmin despite the fact that both nanobodies bind on the same side of GFP and have a considerable amount of overlapping residues. GFP-nb binds relatively higher up GFP and in a horizontal orientation, with those regions of GFP showing the greatest uptake difference being located higher up the protein's side. In comparison, GFP-nbmin binds relatively lower down and in a more vertical orientation and this change in position is represented on GFP with those regions showing the greatest uptake difference being located lower down the protein's side. GFP also showed the lowest levels of corrected RFU, which fits with our theory of 8 hours not being a long enough time point to see deuterium saturation, but still with very limited overlapping of time points. While unrelated to the strict implementation of these data sets being used in order to evaluate HDXsimulator, the fact that the residues comprising the chromophore (a tripeptide consisting of S65, Y66 & G67) displayed a significant amount of uptake difference despite not being involved in the interface (at the 99 % CI for GFP:GFP-nb and at the 95 % CI for GFP:GFP-nbmin) is worthy of note and consideration. In both data sets, the chromophore residues show an uptake decrease, despite the fact that only GFP-nbmin causes a change in the fluorescence characteristics of GFP. It would be interesting to test the GFP-enhancer nanobody [68], which causes the fluorescence of GFP to increase, to see if a positive uptake difference could be seen in the chromophore or if all changes to the chromophore result in an uptake decrease. Overall, we have very strong evidence that both of the GFP data sets are a legitimate representation of the interaction and therefore appropriate for use in testing HDXsimulator.

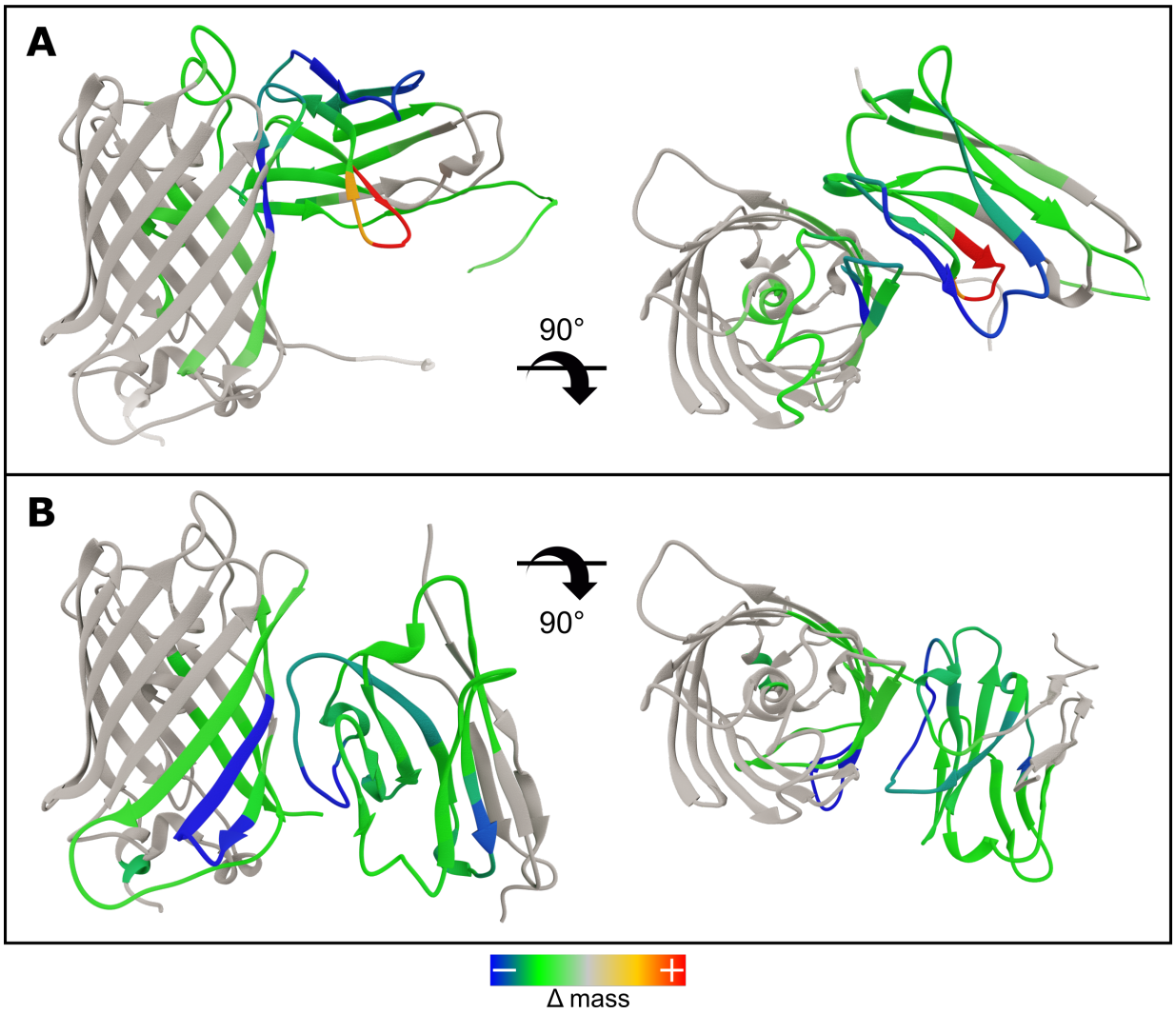


Figure 5.5: GFP & GFP nanobodies HDX-MS data mapped onto structures. Uptake difference data surpassing the 99 % CI of the GFP:GFP-nb and GFP:GFP-nbmin interactions assigned to a colour spectrum and mapped onto their respective three-dimensional structures. The difference values represented by the colour gradient are assigned relative to the highest value in each individual protein's data set, hence colours are not necessarily comparable between structures. **(A)** The GFP:GFP-nb interaction from the front (left) and the top (right). **(B)** The GFP:GFP-nbmin interaction from the front (left) and the top (right).

5.2.4.2 GFP-nbmin

In DynamX, GFP-nbmin displayed good levels of digestion and data quality was moderate across most of the peptides with quite a large number needing to be excluded from the final data set. On the peptide level (Figure 5.2 G) we could see that residues A2-F69 display a wide RFU distribution which then tightens up considerably between residues T70-G112. The C-terminal residues open the RFU distribution up again but interestingly have a much lower maximum RFU compared to the N-terminus, implying a far greater level of protection.

From the perspective of GFP-nbmin, the interaction with GFP saw a significant amount of uptake difference across almost the entire primary sequence, as we have come to expect from proteins of small size (Figure 5.3 H). Woods plots show that all of the residues between S23-Y122 display notable negative uptake difference with the only variation being in the magnitude. Demarcating distinct regions is difficult but it is possible to distinguish 3 from the background. The first stretches between residues S23-T79 with an average uptake value of approx. -4.5 Da and a maximum value of -5.64 Da. The second includes residues V80-Y95 with an average uptake difference of approx. -5 Da and a maximum value of -6.82 Da and the last includes residues Y96-Y122 with an average of approx. -6 Da and a maximum of -7.84 Da. Only residues at the N & C-termini show no significant uptake difference. Like with previous data sets, because a large amount of the proteins sequence showed an uptake difference, two CI values were calculated as described previously, a 99 % CI of 2.21 Da and a 99.9 % CI of 7.05 Da. However, because the value of the 99.9 % CI was so high, it excluded almost every amino acid and so it was decided to only use the 99 % CI for this data set as well.

Like GFP-nb, the GFP-nbmin data set was of lesser quality compared to the other proteins studied, though still perfectly usable. Also like GFP-nb, the majority of the protein is involved in the interaction in some way, although this is not surprising considering they are of similar size. Despite this, a certain degree of localisation could be seen with the reverse faces of this nanobody showing no significant uptake and regions of the highest uptake difference being located at the interface.

5.2.4.3 Visualisation of the interaction

Deuterium uptake differences exceeding the 99 % CI cut off for both proteins were mapped onto the structures of the proteins in their bound conformation (Figure 5.5 B) as described previously. These results align very closely with what we would expect to see from the conformation of the complex crystal structure, with GFP again seeing uptake differences being highly localised to the interface and little difference being seen in the rest of the protein. Unlike with the GFP:GFP-nb data set, the region around the chromophore does not display much significant uptake difference, with the exception of V61, despite GFP-nbmin actually having a visual effect on it. However, when looking at this region in the Woods plots, a distinctive protrusion can be seen in the region of the chromophore but at a value below the threshold of the 99 % CI. If the CI threshold were lowered to 95 %, this region becomes significant, therefore it is debatable as to whether HDX-MS can detect the effect of GFP-nbmin on the chromophore of GFP or not depending on where one sets the CI cut-off. GFP-nbmin also saw more localised differences compared to its counterpart nanobody, with both N & C-terminal regions located on the reverse side of the protein from the interface seeing no significant uptake differences, while

those regions at the interface saw significant uptake difference. Interestingly there was no sign of the positive difference seen in the GFP-nb data set anywhere in this data set, although this may not be surprising considering that the two nanobodies share relatively little sequence homology. Based on these results and of the individual RFU plots, we propose that the HDX data collected for these proteins is a legitimate representation of the individual proteins and the complex as a whole and therefore is appropriate to take forward into the computational stages of this project.

5.3 Obtaining residue-resolved lnPs using HDXmodeller

In our previous work on classifying structures [31], we could only make use of calculated vs. experimental RFU data as there is no way to obtain experimental lnPs from HDX data in order to allow comparison with calculated lnPs. HDXmodeller was developed by Ramin Salmas of the Borysik group in order to fix this problem as it allows residue-resolved lnP values to be modelled from experimental peptide-level HDX-MS data. An essential part of the program are auto-validation matrices that compare various replicates against each other in a pair-wise manner in order to allow calculation of an R-matrix value for each pair. This value indicates the program's confidence in the lnPs it has modelled and these values can be averaged across all of the pairs to give a single confidence value for the entire data set. R-matrix scores can be calculated for a protein as a whole as well as for specific subsections of the protein, assuming no bridging peptides, in order to enable insight into which domains of the protein contribute to greater modelling confidence and which contribute to lesser modelling confidence. These scores can be used to inform users about the likely accuracy of the lnP values produced. Therefore, rather than analyse the lnP outputs themselves, we shall analyse the R-matrix scores for each protein data set as they take into account lots of different factors related to the lnP values. Each protein was run as a whole and also split into subsections to see which parts of the proteins contributed towards high confidence (high R-matrix scores) and which contributed towards low confidence (low R-matrix scores).

Taking the whole protein into account (Figure 5.6), BnWT had an R-matrix score of 0.775 when averaged over all 50 replicates (A), BnH102A had the lowest score of 0.640 (B), BspWT had 0.726 (C), BsY29F had 0.655 (D), GFP had 0.671 (E), GFP-nb had 0.645 (F) and GFP-nbmin had the highest score of 0.795 (G). Therefore, of the whole protein data sets, BnWT, BspWT and GFP-nbmin fall into the "high" (≥ 0.7) bin with regard to accuracy and BnH102A, BsY29F, GFP and GFP-nb fall into the "fair" (0.5-0.69) bin with regard to data accuracy. None of these data sets are considered to have "low" (0-0.49) data accuracy. Hence, we can judge that most of the modelled lnPs for these data sets are likely to be quite accurate as a substantial number of the replicates converge to a small range of agreed upon values, especially those of BnWT, BspWT & GFP-nbmin. With the broad success of the whole protein data sets, we wanted to obtain a higher resolution view of each protein to see if subsections were fairly consistent across the protein or if there were differences between them which were being smoothed out in the whole-protein score. These results are described below and are also summarised in Table 5.2.

On the level of demarcated domains (Figure 5.7), BnWT had subsections from residues A1-Y13, Y13-A43, A43-F56, F56-D93 & W94-I109. Due to the N-terminal residue of each subsection not being considered in the calculation, there was therefore no overlap in this or in any following data sets. The A1-Y13 subsection had an R-matrix score of 0.924 (A1), the highest of any subsection tested, the Y13-

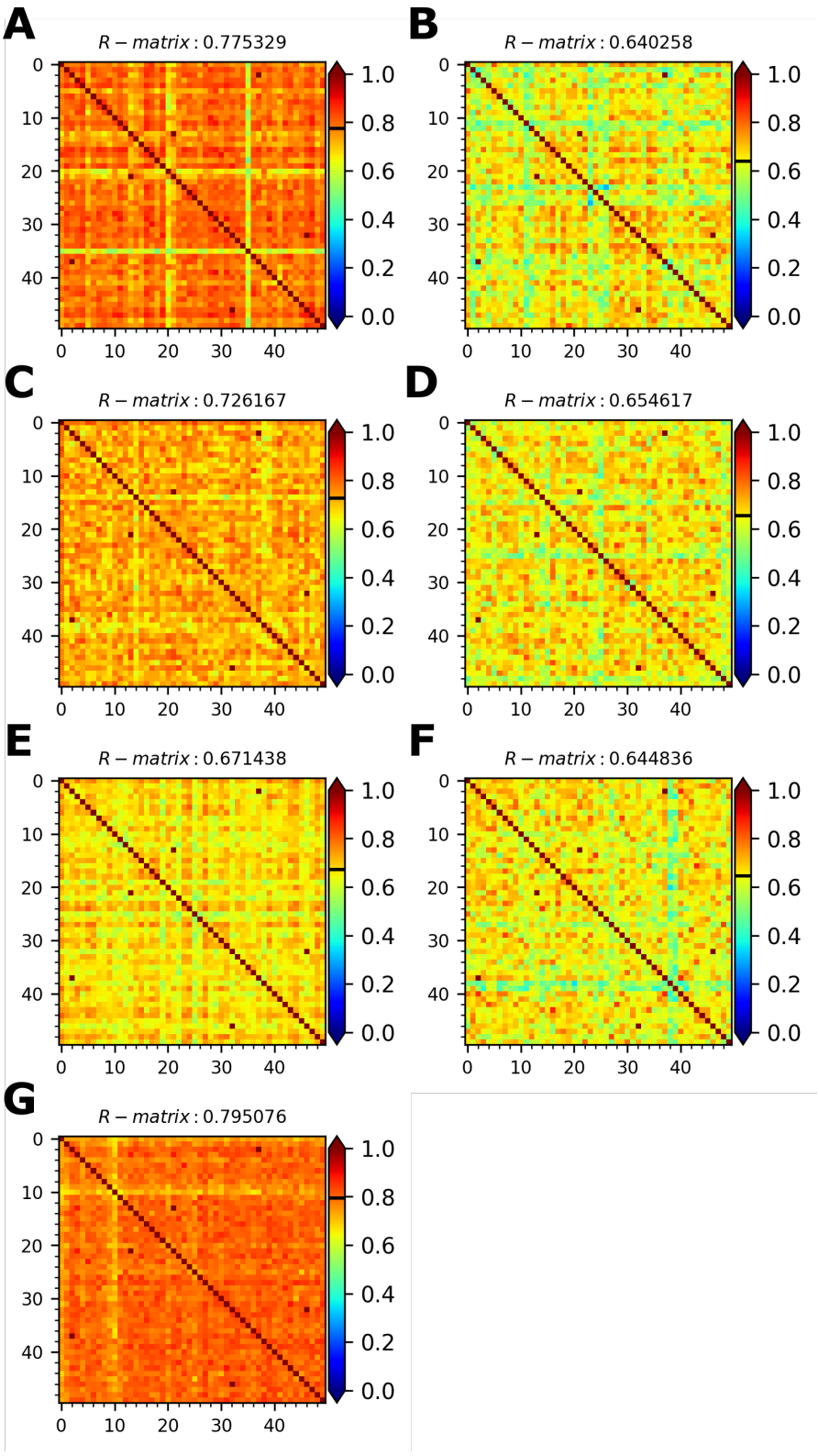


Figure 5.6: Confidence in the residue-level lnPs calculated for whole proteins. Auto-validation matrices displaying the average R-matrix score of the 50 pair-wise replicates generated to determine the residue-level lnP values for entire proteins. Score of each individual pair-wise replicate represented by a coloured square within each plot where the colour relates to the R-matrix score according to the colour bar to the side. **(A)** BnWT, **(B)** BnH102A, **(C)** BspWT, **(D)** BsY29F, **(E)** GFP, **(F)** GFP-nb, **(G)** GFP-nbmin.

	Subsection 1	*Subsection 2*	*Subsection 3*	*Subsection 4*	*Subsection 5*	*Subsection 6*
BnWT	A1-Y13	Y13-A43	A43-F56	F56-D93	W94-I109	
	0.924	0.699	0.421	0.891	0.635	
BnH102A	A1-L14	L14-A43	A43-Y78	Y78-W94	W94-R110	
	0.921	0.510	0.665	0.423	0.782	
BspWT	E8-L16	L16-L34	L34-E52	E52-L71	Q72-T85	
	0.362	0.892	0.594	0.360	0.306	
BsY29F	D15-N33	N33-E52	E52-L71	Q72-T85		
	0.552	0.463	0.763	0.099		
GFP	L7-F46	F46-F99	F100-F130	F130-F165	K166-L207	L207-T230
	0.760	0.594	0.588	0.702	0.661	0.606
GFP-nb	L5-L21	S22-W37	E48-F69	L82-Y95	Y95-F103	
	0.832	0.272	0.560	0.218	0.722	
GFP-nbmin	A2-L22	S23-E48	E48-T70	T70-C97	D121-H139	
	0.521	0.871	0.838	0.707	0.745	

Table 5.2: R-matrix values of individual protein subsections. Table displaying R-matrix values of the subsections of the 7 individual protein data sets after analysis by HDXmodeller. Residues covered by each protein's subsections are indicated along with the R-matrix value determined for that subsection below.

A43 subsection had 0.699 (A2), the A43-F56 subsection had 0.421 (A3), the F56-D93 subsection had 0.891 (A4) and the W94-I109 subsection had 0.635 (A5). Therefore, subsections A1-Y13 & F56-D93 fall into the "high" data accuracy bin, subsections Y13-A43 and W94-I109 fall into the "fair" data accuracy bin and subsection A43-F56 falls into the "low" data accuracy bin.
BnH102A had subsections from residues A1-L14, L14-A43, A43-Y78, Y78-W94 & W94-R110. The A1-L14 subsection had an R-matrix score of 0.921 (B1), the L14-A43 subsection had 0.510 (B2), the A43-Y78 subsection had 0.665 (B3), the Y78-W94 subsection had 0.423 (B4) and the W94-R110 subsection had 0.782 (B5). Therefore, subsections A1-L14 and W94-R110 fall into the "high" data accuracy bin, subsections L14-A43 and A43-Y78 fall into the "fair" data accuracy bin and subsection Y78-W94 falls into the "low" data accuracy bin.
BspWT had subsections from residues E8-L16, L16-L34, L34-E52, E52-L71 & Q72-T85. The E8-L16 subsection had an R-matrix score of 0.362 (C1), the L16-L34 subsection had 0.892 (C2), the L34-E52 subsection had 0.594 (C3), the E52-L71 subsection had 0.360 (C4) and the Q72-T85 subsection had 0.306 (C5). Therefore, subsection L16-L34 falls into the "high" data accuracy bin, subsection L34-E52 falls into the "fair" data accuracy bin and subsections E8-L16, E52-L71 & Q72-T85 fall into the "low" data accuracy bin.
BsY29F had subsections from residues D15-N33, N33-E52, E52-L71 & Q72-T85. The D15-N33 subsection had an R-matrix score of 0.552 (D1), the N33-E52 subsection had 0.463 (D2), the E52-L71 subsection had 0.763 (D3) and the Q72-T85 subsection had 0.099 (D4), the lowest of any subsection tested. Therefore, subsection E52-L71 falls into the "high" data accuracy bin, subsection D15-N33 falls into the "fair" data accuracy bin and subsections N33-E52 & Q72-T85 fall into the "low" data accuracy bin.

On the level of demarcated domains (Figure 5.8), GFP had subsections from residues L7-F46, F46-F99, F100-F130, F130-F165, K166-L207 & L207-T230. The L7-F46 subsection had an R-matrix score of 0.760 (A1), the F46-F99 subsection had 0.594 (A2), the F100-F130 subsection had 0.588 (A3), the F130-F165 subsection had 0.702 (A4), the K166-L207 subsection had 0.661 (A5) and the L207-T230 subsection had 0.606 (A6). Therefore, subsections L7-F46 & F130-F165 fall into the "high" data accuracy bin and subsections F46-F99, F100-F130, K166-L207 & L207-T230 fall into the "fair" data accuracy bin. None of the subsections fall into the "low" data accuracy bin.
GFP-nb had subsections from residues L5-L21, S22-W37, E48-F69, L82-Y95 & Y95-F103. The L5-L21 subsection had an R-matrix score of 0.832 (B1), the S22-W37 subsection had 0.272 (B2), the E48-F69 subsection had 0.560 (B3), the L82-Y95 subsection had 0.218 (B4) and the Y95-F103 subsection had 0.722 (B5). Therefore, subsections L5-L21 & Y95-F103 fall into the "high" data accuracy bin, subsection E48-F69 falls into the "fair" data accuracy bin and subsections S22-W37 & L82-Y95 fall into the "low" data accuracy bin.
GFP-nbmin had subsections from residues A2-L22, S23-E48, E48-T70, T70-C97 & D121-H139. The A2-L22 subsection had an R-matrix score of 0.521 (C1), the S23-E48 subsection had 0.871 (C2), the E48-T70 subsection had 0.838 (C3), the T70-C97 subsection had 0.707 (C4) and the D121-H139 subsection had 0.745 (C5). Therefore, subsections S23-E48, E48-T70, T70-C97 & D121-H139 fall into the "high" data accuracy bin and subsection A2-L22 falls into the "fair" data accuracy bin. None of the subsections fall into the "low" data accuracy bin.

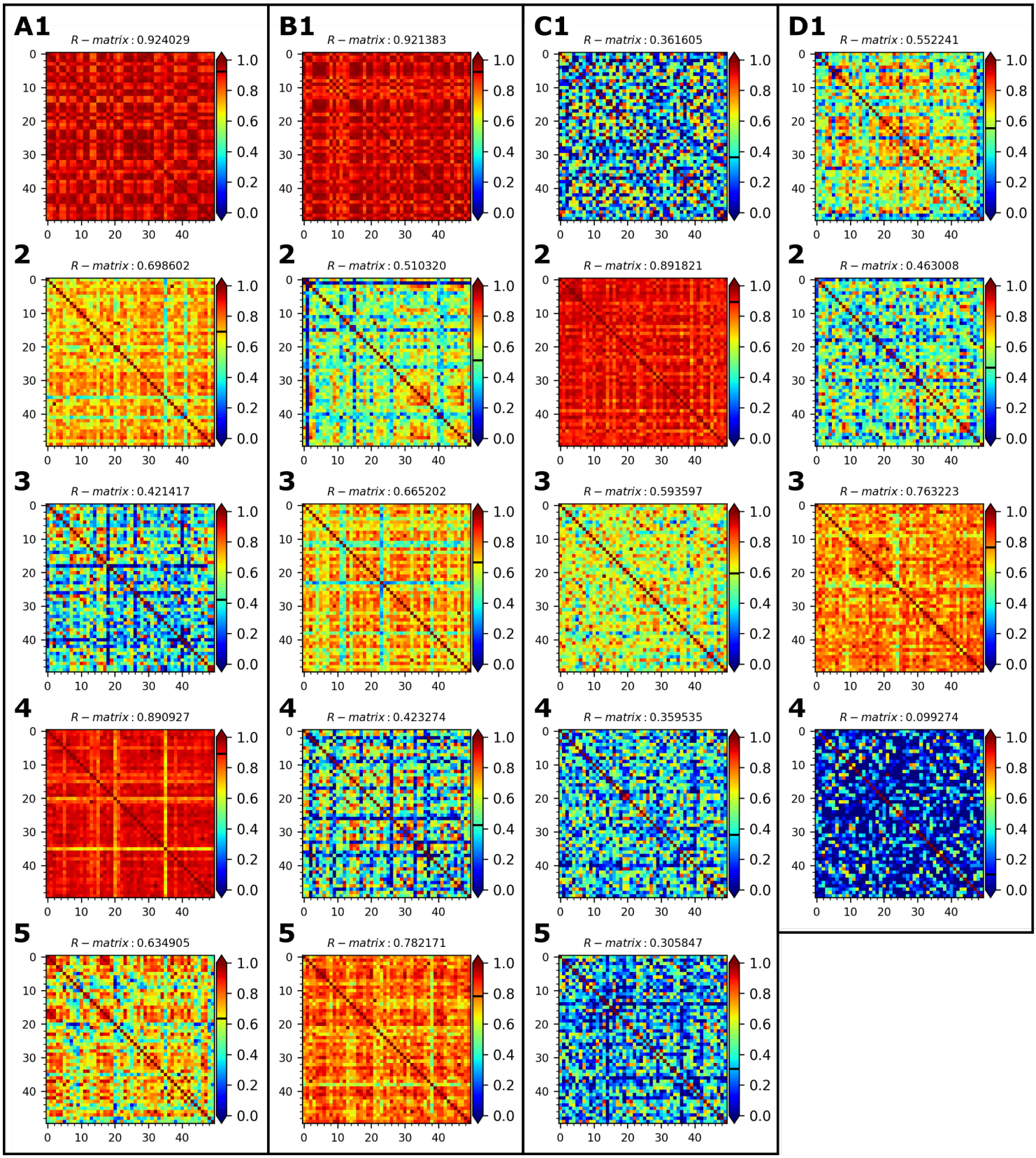

A1
R − matrix : 0.924029
B1
R − matrix : 0.921383
C1
R − matrix : 0.361605
D1
R − matrix : 0.552241
2
R − matrix : 0.698602
R − matrix : 0.510320
R − matrix : 0.891821
R − matrix : 0.463008
3
R − matrix : 0.421417
R − matrix : 0.665202
R − matrix : 0.593597
R − matrix : 0.763223
4
R − matrix : 0.890927
R − matrix : 0.423274
R − matrix : 0.359535
R − matrix : 0.099274
5
R − matrix : 0.634905
R − matrix : 0.782171
R − matrix : 0.305847

Figure 5.7: Confidence in the residue-level lnPs calculated for barnase-barstar subsections. Auto-validation matrices displaying the average R-matrix score of the 50 pair-wise replicates generated to determine the residue-level lnP values for subsections of the barnase and barstar proteins. Score of each individual pair-wise replicate represented by a coloured square within each plot where the colour relates to the R-matrix score according to the colour bar to the side. **(A1)** BnWT subsection A1-Y13, **(A2)** BnWT subsection Y13-A43, **(A3)** BnWT subsection A43-F56, **(A4)** BnWT subsection F56-D93, **(A5)** BnWT subsection W94-I109. **(B1)** BnH102A subsection A1-L14, **(B2)** BnH102A subsection L14-A43, **(B3)** BnH102A subsection A43-Y78, **(B4)** BnH102A subsection Y78-W94, **(B5)** BnH102A subsection W94-R110. **(C1)** BspWT subsection E8-L16, **(C2)** BspWT subsection L16-L34, **(C3)** BspWT subsection L34-E52, **(C4)** BspWT subsection E52-L71, **(C5)** BspWT subsection Q72-T85. **(D1)** BsY29F subsection D15-N33, **(D2)** BsY29F subsection N33-E52, **(D3)** BsY29F subsection E52-L71, **(D4)** BSY29F subsection Q72-T85.

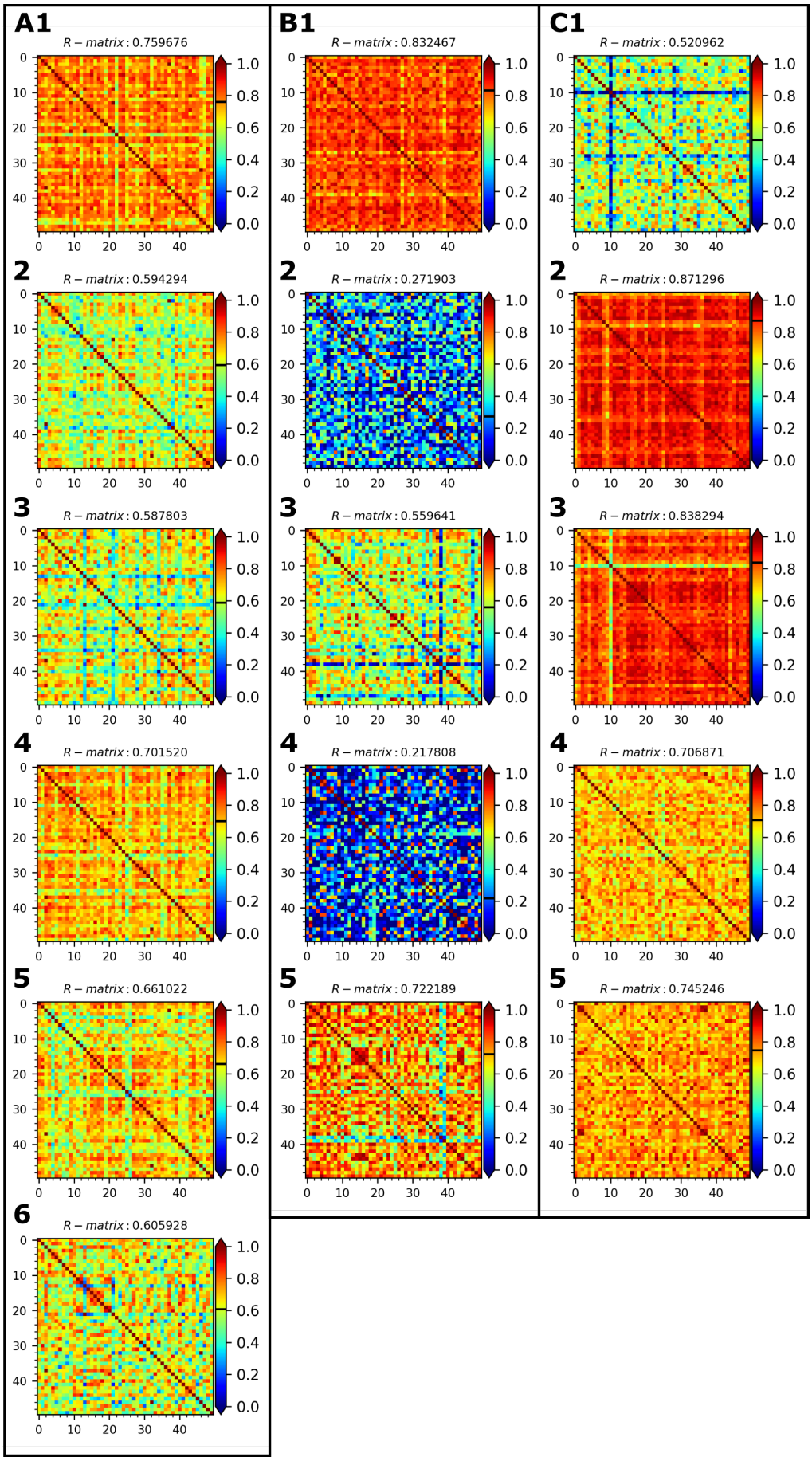

A1
R − matrix : 0.759676
2
R − matrix : 0.594294
3
R − matrix : 0.587803
4
R − matrix : 0.701520
5
R − matrix : 0.661022
6
R − matrix : 0.605928
B1
R − matrix : 0.832467
2
R − matrix : 0.271903
3
R − matrix : 0.559641
4
R − matrix : 0.217808
5
R − matrix : 0.722189
C1
R − matrix : 0.520962
2
R − matrix : 0.871296
3
R − matrix : 0.838294
4
R − matrix : 0.706871
5
R − matrix : 0.745246

Figure 5.8: Confidence in the residue-level lnPs calculated for GFP-GFPnbs subsections. Auto-validation matrices displaying the average R-matrix score of the 50 pair-wise replicates generated to determine the residue-level lnP values for subsections of the GFP, GFP-nb & GFP-nbmin proteins. Score of each individual pair-wise replicate represented by a coloured square within each plot where the colour relates to the R-matrix score according to the colour bar to the side. **(A1)** GFP subsection L7-F46, **(A2)** GFP subsection F46-F99, **(A3)** GFP subsection F100-F130, **(A4)** GFP subsection F130-F165, **(A5)** GFP subsection K166-L207, **(A6)** GFP subsection L207-T230. **(B1)** GFP-nb subsection L5-L21, **(B2)** GFP-nb subsection S22-W37, **(B3)** GFP-nb subsection E48-F69, **(B4)** GFP-nb subsection L82-Y95, **(B5)** GFP-nb subsection Y95-F103. **(C1)** GFP-nbmin subsection A2-L22, **(C2)** GFP-nbmin subsection S23-E48, **(C3)** GFP-nbmin subsection E48-T70, **(C4)** GFP-nbmin subsection T70-C97, **(C5)** GFP-nbmin subsection D121-H139.

On the level of individual domains, we found that there were indeed considerable differences between subsections in all of the protein data sets, with some domains having extremely high R-matrix scores and some with extremely low scores. In some proteins such as GFP these were fairly minor, however with some such as GFP-nb they were extreme with certain subsections having excellent R-matrix scores and others with terrible ones. It is also important to point out that the score assigned to the protein as a whole is not a simple average of all the differing subsections. This can be seen quite clearly in the BsY29F data set where it has a whole-protein R-matrix score of 0.655 but the individual subsections have scores of 0.552, 0.463, 0.763 & 0.099; a mean average of 0.469. This seeming paradox is explained by remembering that changing the precise shape of the peptide maps can have profound changes on the R-matrix score, and therefore the residues comprising a specific subsection may perform differently when taken as a whole, with all the other peptides present, compared to when taken individually. Therefore there is a certain degree of disconnect between the two different resolutions used by HDXmodeller which must be born in mind when evaluating data.

5.4 Molecular Dynamics simulations

With HDX data sets now generated for each protein and modelled residue-level lnPs determined, we moved on to generating structure libraries which could be used to benchmark HDXsimulator. In order to use HDXsimulator to classify the structures of binary complexes as well as individual proteins, we needed to undertake some additional steps in order to prepare suitable structures for use as training data sets. Decoy libraries of individual proteins could be generated immediately using a program such as Rosetta or 3DRobot, however the generation of binary complexes was a little more involved. The first step in this process was to relax our input crystal structures using MD simulations. Relaxed starting structures would be important for docking later on because we needed to ensure that the method was capable of producing native complexes from HDX restraints and flexible refinement alone. There is little challenge to docking structures that are already in the perfect conformation and when this method is tested out for real, it will be of course be on unbound structures. Therefore, as the crystal structures we had were of bound complexes, we needed to relax them with MD in order to more accurately represent an unbound conformation. NAMD was the obvious choice of software for this work because members of the group had previous experience with it, allowing for faster learning and swifter error resolution.

Two different types of simulations were carried out: “unbound” in which the individual proteins were relaxed in a simulation without their binding partner and “bound” in which the relaxation was carried out with the binding partner present. The goal of having these two different forms of simulation was that it would enable us to see what effects, if any, the presence or absence of the binding partner had on subsequent protein-protein docking. Unbound MD simulations were carried out on BnWT, BnH102A, BspWT, BsY29F, GFP, GFP-nb & GFP-nbmin. Bound MD simulations were carried out on BnWT:BspWT & GFP:GFP-nb. As the purpose of these MD simulations was to simply relax the structures instead of make mechanistic insights, simulations were comparatively short and only run until the RMSD of the later frames vs. the starting frame stabilised. MD trajectories were analysed using VMDs RMSD Trajectory Tool (RMSDTT) and calculations were made using backbone atoms only.

BnWT (Figure 5.9 A) was run for a total of 10 ns with the RMSD stabilising relatively quickly after

1 or 2 ns and the average RMSD across every frame being 1.09 Å. The average structure was calculated using the RMSDTT and the frame closest to this average structure determined to be frame 479. Therefore this structure was carried forward into the docking stage as the representative structure of relaxed BnWT.
BnH102A (Figure 5.9 B) was run for a total of 10 ns with the RMSD stabilising very quickly after about 1 ns and the average RMSD across every frame being 0.99 Å. The average structure was calculated using the RMSDTT and the frame closest to this average structure determined to be frame 197. Therefore this structure was carried forward into the docking stage as the representative structure of relaxed BnH102A.
BspWT (Figure 5.9 C) was run for a total of 10 ns with the RMSD stabilising relatively quickly after 1 or 2 ns and the average RMSD across every frame being 0.98 Å. The average structure was calculated using the RMSDTT and the frame closest to this average structure determined to be frame 431. Therefore this structure was carried forward into the docking stage as the representative structure of relaxed BspWT.
BsY29F (Figure 5.9 D) was run for a total of 10 ns with the RMSD stabilising almost immediately after less than 1 ns and the average RMSD across every frame being 0.8 Å. The average structure was calculated using the RMSDTT and the frame closest to this average structure determined to be frame 477. Therefore this structure was carried forward into the docking stage as the representative structure of relaxed BsY29F.
GFP (Figure 5.9 E) was run for a total of 20 ns with the RMSD taking substantially longer to stabilise compared to the other proteins studied, around 10 ns, likely because of its larger size. The average RMSD across every frame was 2.5 Å. The average structure was calculated using the RMSDTT and the frame closest to this average structure determined to be frame 505. Therefore this structure was carried forward into the docking stage as the representative structure of relaxed GFP.
GFP-nb (Figure 5.9 F) was run for a total of 10 ns with the RMSD showing a distinctive wave pattern that repeats every 3-4 ns or so and is likely the result of the poly-His tag flailing about. The average RMSD across every frame was 2.73 Å. The average structure was calculated using the RMSDTT and the frame closest to this average structure determined to be frame 944. Therefore this structure was carried forward into the docking stage as the representative structure of relaxed GFP-nb.
GFP-nbmin (Figure 5.9 G) was run for a total of 10 ns with the RMSD taking quite a long time to stabilise. This appears to be because the poly-His tag takes some time to extend out from its initial position whereupon it stays in an extended state. The average RMSD across every frame was 5.22 Å. The average structure was calculated using the RMSDTT and the frame closest to this average structure determined to be frame 64. N.B. due to an input error, this trajectory only has 100 frames covering 10 ns instead of the usual 1,000, hence the data for this system will be slightly less precise. Nevertheless, frame 64 was carried forward into the docking stage as the representative structure of relaxed GFP-nb.

BnWT:BspWT (Figure 5.9 H) was run for a total of 10 ns with the RMSD stabilising very quickly after about 1 ns and the average RMSD across every frame being 1.31 Å. The average structure was calculated using the RMSDTT and the frame closest to this average structure determined to be frame 111. Therefore this structure was carried forward into the docking stage as the representative structure of the relaxed BnWT:BspWT complex.
GFP:GFP-nb (Figure 5.9 I) was run for a total of 20 ns due to the size of the complex with the RMSD

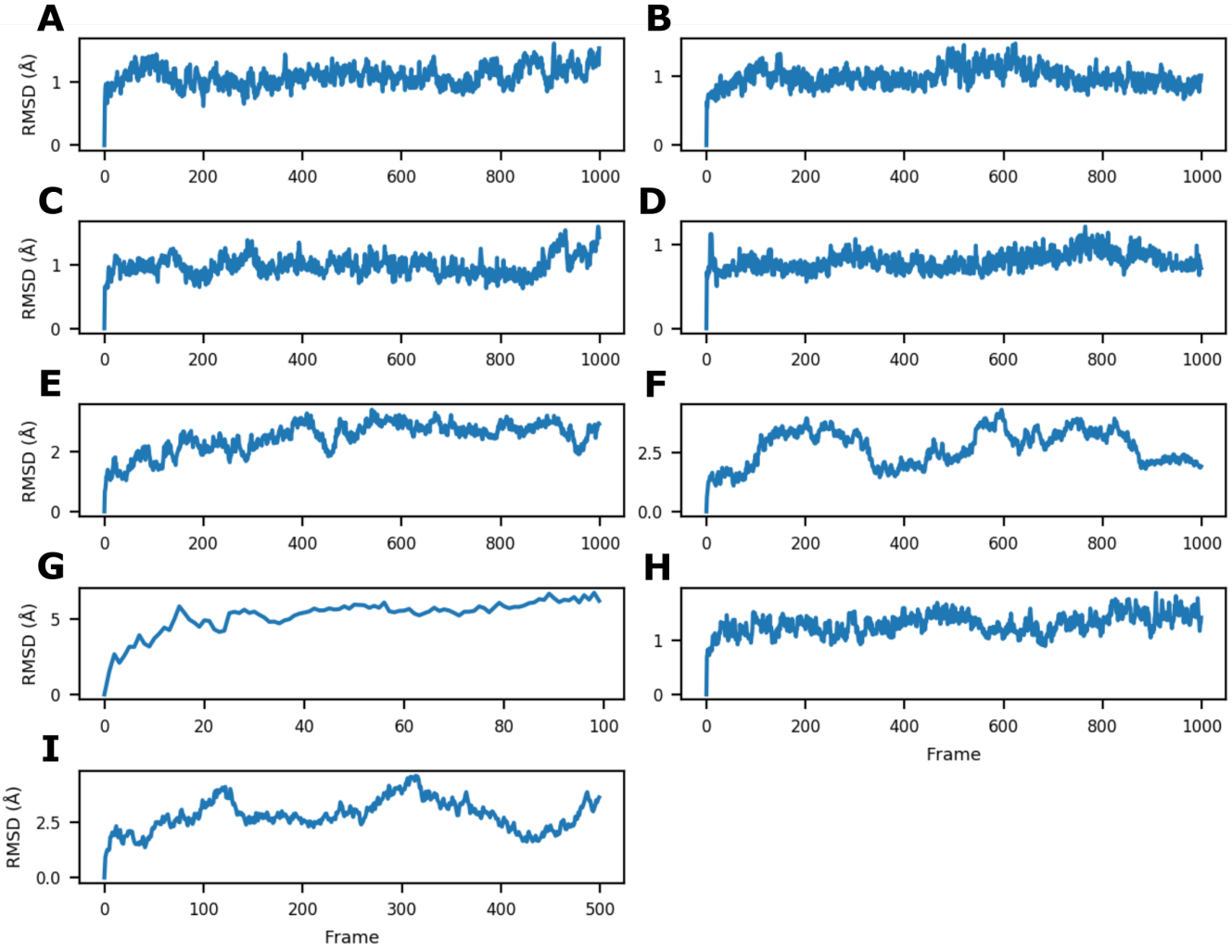


Figure 5.9: Change in RMSD over the course of a relaxational MD simulation. Plots describing how a structure's RMSD changes vs. the initial starting frame over the course of a relaxational MD simulation. Relaxation deemed to have fully occurred upon seeing a qualitative stabilization of RMSD value over a prolonged period of frames. **(A)** BnWT, **(B)** BspWT, **(C)** BnH102A, **(D)** BsY29F, **(E)** GFP, **(F)** GFP-nb, **(G)** GFP-nbmin, **(H)** BnWT:BspWT, **(I)** GFP:GFP-nb. MD simulations A-D & F-H were run for 10 ns, MD simulations E & I were run for 20 ns.

stabilizing after about 5 ns with the average RMSD across every frame being 2.82 Å. The average structure was calculated using the RMSDTT and the frame closest to this average structure determined to be frame 218. Therefore this structure was carried forward into the docking stage as the representative structure of the relaxed GFP:GFP-nb complex.

As we were just relaxing the various structures and not looking for any mechanistic insight, the results of the MD simulations were not too surprising. Most of the structures stabilised very quickly, within a few nanoseconds, which makes sense considering most of the proteins used in this study are quite small. In the same vein, it also makes sense that GFP took the longest to stabilise considering it is more than twice the size of every other protein. The most interesting results were those for GFP-nb and the GFP:GFP-nb complex which saw a wave pattern in the RMSD trace over time. This pattern in the GFP-nb plot can be explained by the poly-His tag which is unstructured and so moves about randomly as time progresses. As GFP-nb is small and otherwise very stable, the His tag makes up a substantial amount of the deviation seen in this system and so its movement is represented prominently in the RMSD plot. However this would not explain why this wave pattern is also seen in the GFP:GFP-nb plot as now the His tag is only a very small part of the whole system. A potential answer to this question lies in GFP itself which also has an unstructured N-terminal region (not a His tag) which sees similar waving motions over time. Unlike with the nanobody, this is not as visible in the plot of GFP on its own because of how much larger the structured region of GFP is compared to GFP-nb. However, when the two proteins are combined in a complex, the cumulative deviation caused by two disordered regions is now large enough to be visible in the RMSD plot.

5.5 Protein-protein docking

With most of the structures now relaxed, both in the unbound and bound conformations, the next step was to generate libraries of protein-protein docking poses for eventual use in benchmarking HDXsimulator's ability to classify bound structures. The process of docking relaxed MD structures into complexed poses was chronologically the last part of the project we undertook before the commencement of the writing-up process. Therefore, while we have established the methods that will be used to take this project into its final stages (determining complex structure from HDX data), we did not have time to complete all the necessary work and this will be something for future group members to finalise. Nevertheless, we have discovered some interesting aspects of those systems we did have a chance to analyse. As of writing, only two data sets have been attempted: BnWT:BspWT & GFP:GFP-nb, both of which were produced using the unbound conformers of the constituent proteins. Data sets were produced using the webserver version of HADDOCK and the RMSD of the produced poses compared to the crystal structure as described previously.

The number of "native" structures produced by a docking method depends on where the RMSD cut off is set. For BnWT:BspWT, if the cut off is set to 2.5 Å, 10 structures out of the 1,000 selected to undergo flexible refinement and subsequent refinement in explicit solvent were found to be native. This number represents 1 % of the total refined structures and is about half of what we were ideally looking for, based on our results for the Rosetta decoy sets where data sets of 1,000 decoys where enriched with 20 native structures. However, if the cut off is increased just slightly to 3 Å, 26 structures out of

the 1,000 selected to undergo flexible refinement and subsequent refinement in explicit solvent were found to be native. This number represents 2.6 % of the total refined structures. Therefore we can see that different results can be obtained depending on where the cut off is placed which will be a point of optimisation in the future.

Figure 5.10 A displays all 1,000 docking poses superimposed on top of each other, using BnWT as the anchor point. From this we can see that there is quite a large variety of different orientations, which makes sense considering that the majority of both proteins' residues were marked as being significant for the purposes of constructing the AIRs. Additionally, the AIRs are a binary on/off, meaning that the complex shape of the interaction's difference plot is lost. However despite this, poses are mostly limited to a horizontal plane through the true binding site with no poses found with BspWT on "top" or on "bottom" of BnWT. Figure 5.10 B displays the exact same orientation of the poses but this time only those poses deemed to be native (according to a cut off of 2.5 Å) are shown. From this we can see the location of the true binding site and, when comparing it to Figure 5.10 A, we can also see that the majority of the 1,000 poses fall approximately within this location.

In comparison to BnWT:BspWT, the docked poses for GFP:GFP-nb showed no native structures, with the closest pose having an RMSD of 7.2 Å. This may be partially due to the lower sampling used in this data set, only 1,000 rigid body structures followed by the top 200 selected to undergo flexible refinement and subsequent refinement in explicit solvent; but this would not explain the complete lack of anything approaching a native structure. When looking at the structures superimposed upon each other using GFP as the anchor (Figure 5.10 C), we can see that, while mostly on the correct side, GFP-nb poses are rotated at all sorts of angles compared to the crystal structure, giving rise to the high RMSD values. This is in comparison to BnWT:BspWT where the correct orientation was seen in almost all the poses. We believe this to be caused by two factors. The first is that the GFP:GFP-nb complex lacks the "lock and key" geometry seen in the BnWT:BspWT complex. Instead, the binding surfaces are relatively flat which no doubt caused HADDOCK to struggle as, like most docking programs, it relies to a certain extent on cavities and protrusions in order to fit proteins together. The second factor is that, while the AIRs for GFP were highly localised, the AIRs for GFP-nb were not. This means that HADDOCK likely struggled to assign a proper orientation for GFP-nb, hence why it is found at so many different angles.

While we are satisfied with the BnWT:BspWT data set, the GFP:GFP-nb data set yet requires substantial optimisation before it can be brought forward to serve as a test for HDXsimulator. There are several parameters that could be optimised, like with BnWT:BspWT, including the CI threshold to use as AIRs, the number of initial structures for rigid body docking as well as the number that are then taken forward for subsequent refinement. There is also the possibility that 2.5 Å maybe an unnecessarily strict cut off.

With the data from these two simulations in mind, we assert that HDX data has the potential to be very capable of guiding protein-protein docking simulations, however that capacity is diminished the greater a protein's surface is covered by the AIRs. In cases such as these, the docking program can only rely on steric characteristics which explains why BnWT:BspWT produced a good number

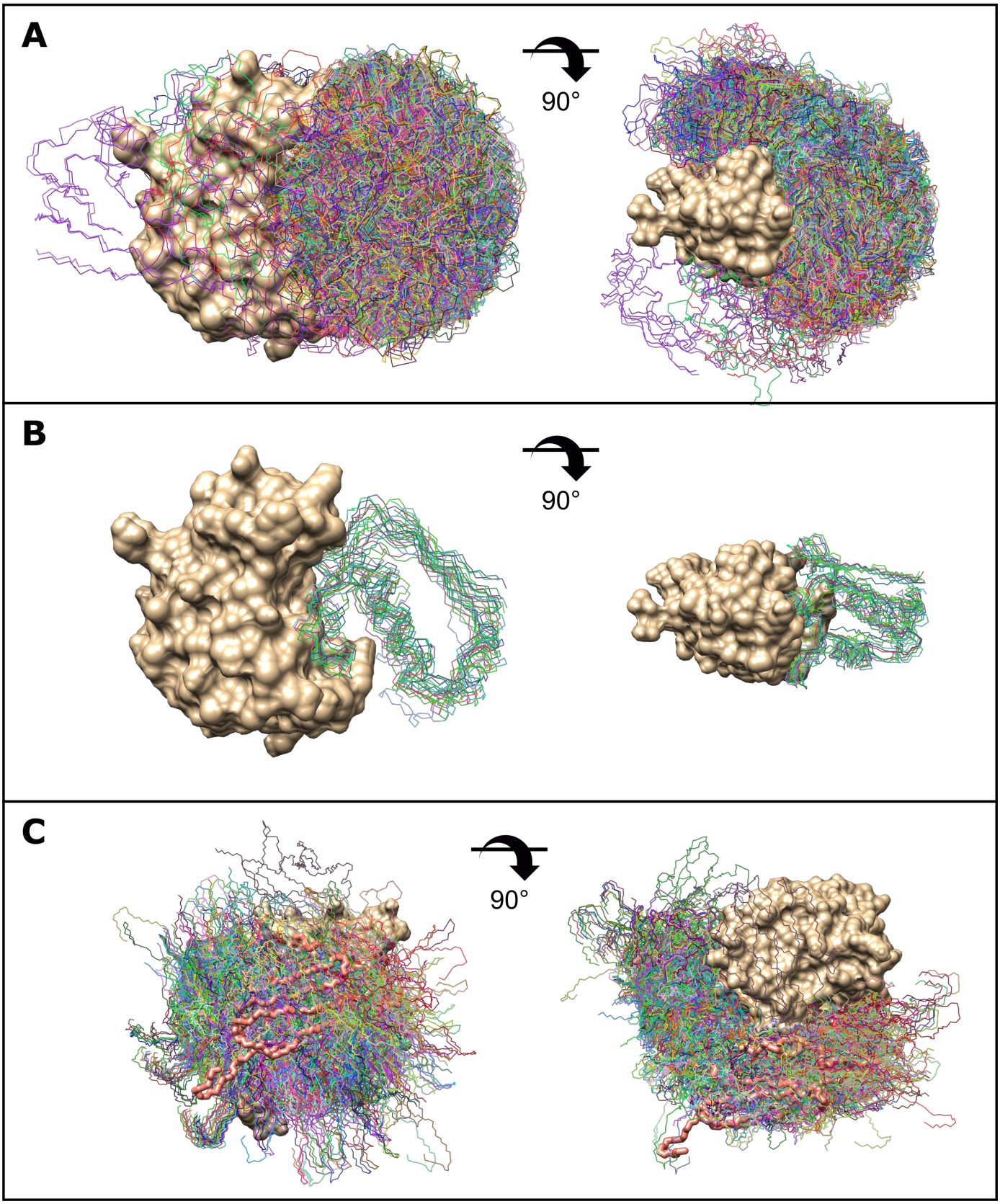

A
90°
B
90°
C
90°

Figure 5.10: Docking poses simulated for the BnWT:BspWT & GFP:GFP-nb interactions. HADDOCK-generated docking poses showing the spread of structures obtained when using AIRs constructed from residues with uptake difference values surpassing the 99 % CI. Barnase and GFP (tan, surface view) were used as the anchor points in their respective sub-figures with the partner protein (multicoloured, wire view) rotated around it according to the docking pose. **(A)** BnWT:BspWT; all 1,000 refined poses viewed from the front (left) and the top (right). **(B)** BnWT:BspWT; 10 native (RMSD $\leq$ 2.5 Å) refined poses viewed from the front (left) and the top (right). **(C)** GFP:GFP-nb; all 200 refined poses viewed from the front (left) and the top (right). The native pose of GFP-nb is represented in red, thick wire view.

of native poses (easily discernible shape complementarity), whereas GFP:GFP-nb did not (poor shape complementarity between the structures). Therefore we propose that the best docking simulations will likely be gained from those interactions which have both specific HDX-generated AIRs and also show good shape complementarity, although of course more data sets will be required to confirm this.

5.6 Exploring the boundaries of modelling protein conformation using HDXsimulator

Simultaneously to our work generating binary complex pose libraries, we began the process of testing HDXsimulator using libraries of individual protein decoy structures we had previously produced. HDXsimulator is a program developed by Ramin Salmas from the Borysik group to enable the calculation of theoretical protection factors and RFU values from three-dimensional coordinates on the residue level. These calculated lnP/RFU values can then be compared to modelled/experimentally determined lnP/RFU values determined by HDXmodeller/experimental HDX-MS in order to generate an RMSE metric which serves to rank the structures by how close to experimental values they are. In order to test the program's ability to distinguish between native and non-native structures, ROC curves were constructed with the curve's AUC being indicative of its efficacy. In chapter 4.5, we began the process of mapping the capabilities and limitations of HDXsimulator by introducing synthetic error into lnP data sets and seeing how the AUC score of the data set responded. Our initial attempts could not elucidate a relationship between AUC and the RMSE/R^2 of the synthetic error data, however after much trial and error, we found that optimal method for error generation was using HDXsimulator itself utilising suboptimal scaling factors which produced the correlated AUC and R^2 values that we were looking for. With our method of choice identified, we then proceeded to generate a full data suite from the BspWT_Rosetta data set, split up into 10 residue domains, in order to see how AUC scores responded to synthetic error on the level of individual subsections. These findings are displayed in Figure 5.11.

From this data we can see that there is a strong linear correlation between AUC and R^2 in almost all the different domains of the BspWT_Rosetta data set. Domains representing the central section of the protein (C-G) display the strongest correlation with very tight grouping between AUC and R^2 values, while domains closer to the termini (A, B, H, I) display weaker correlation but nevertheless far greater than any of the other error generation methods attempted previously. Almost all domains display a degree of ambivalence at the higher R^2 values were the correlation between AUC and R^2 is substantially worse. This "plateau" was seen far more prevalently in test RFU-level data sets and those using RMSE as the comparison metric but is seen to a lesser extent here as well. Plateauing is especially extensive in data representing residues G31-A40 (D) and residues Q61-V70 (G), however it does fall away and the correlation becomes linear again after R^2 values drop slightly. The range of the R^2 and AUC values also saw a considerable amount of variation over the differing domains of the data set. Some domains, such as residues G31-A40 (D), saw a relatively small range from 0 to -2 while others, such as residues L71-E80 (H), saw a much larger range from 0 to -20. These differing ranges of R^2 values indicate that varying the scaling factors of HDXsimulator has a greater effect on some domains compared to others, however this does not seem to translate into the pipeline being more or less capable of differentiating between them as there is no obvious correlation between the range of R^2 values and the correlation to AUC. Similarly, the range of AUC values differed substan-

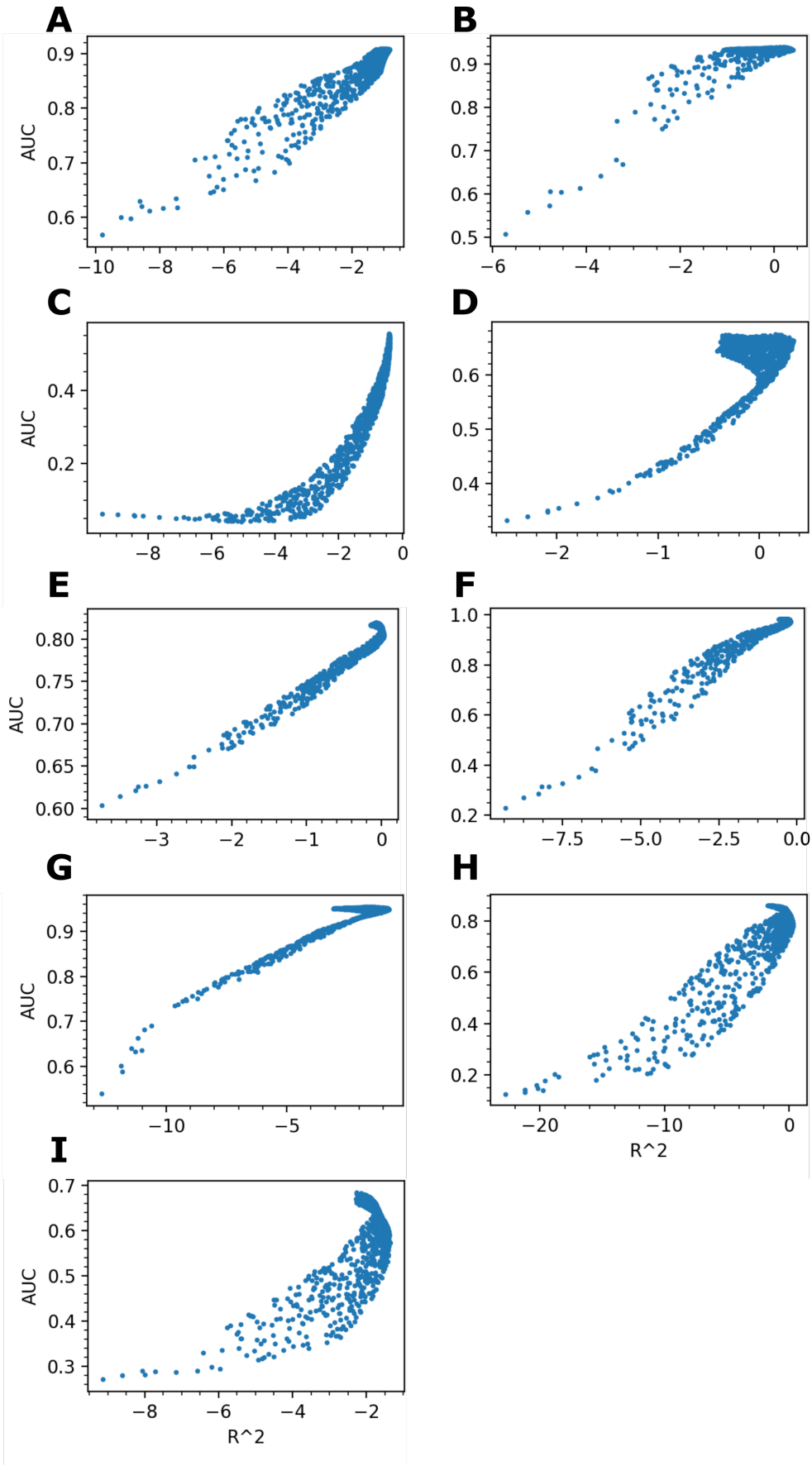

A
B
C
D
E
F
G
H
I
AUC
R^2

Figure 5.11: Testing HDXsimulator's ability to distinguish between erroneous lnPs. Plots showing how the AUC value (HDXsimulator's ability to distinguish between native and non-native structures) changes depending on how different erroneous input lnPs are from the "true" lnP values, as measured by their R^2 value. Plots generated for the BspWT_Rosetta data set in 10 residue increments. **(A)** Residues K1-I10, **(B)** residues R11-L20, **(C)** residues K21-Y30, **(D)** residues G31-A40, **(E)** residues L41-V50, **(F)** residues L51-K60, **(G)** residues Q61-V70, **(H)** residues L71-E80, **(I)** residues G81-S89.

tially between domains, such as residues K1-I10 (A) with a range of 0.9-0.6 and residues K21-Y30 (C) with a range of 0.5-0.1. This indicates that HDXsimulator's capacity to distinguish between native and non-native structures is influenced down to the level of a few residues and that as a consequence the AUC value for any overall protein is likely somewhat of an average of the values of individual domains.

The last point to consider when looking at the BspWT_Rosetta data set is the characteristic "hook" seen at the higher AUC/R^2 values in most of the domains. This could perhaps be the result of the inverse of the effect discussed above wherein similar structures, having similar AUC values, nevertheless have slightly different R^2 values because of small differences in lnP as caused by the varying scaling factors used in its creation. The sharp demarcation between the two clusters of AUC values a given R^2 can give rise to at the high end with no intermediary values could be due to specific conformers being sterically favoured and so the resultant structures and AUCs being one or the other rather than a range in between.

To summarise, we intended to determine the relationship between the deviation of erroneous lnPs/-subsequent RFU values and the eventual AUC score calculated for a particular data set, with our goal being to chart HDXsimulator's capabilities and limitations to enable us to improve future versions of the program. After much trial and error, we found that optimal method for error generation was using HDXsimulator itself utilising suboptimal scaling factors which produced the correlated AUC and R^2 values that we were looking for. The results of our optimisation work as well as the experiments presented here tell us that the relationship between AUC and RMSE/R^2 is much more complicated than we had initially thought, with hidden factors challenging our understanding of the boundaries of modelling protein conformation using HDXsimulator. These factors affect not only how data calculated from protein decoys reacts to AUC on the whole protein level, but also on the level of distinct subsections as our experiments with the BspWT_Rosetta data set demonstrate. The subsections of this protein have quite different profiles from each other, showing that, like with HDXmodeller, data generated for a protein as a whole is not necessarily representative of its individual constituent parts. This work is still in its infancy and at present there is much we do not understand, including what these hidden factors are and it is clear that more work will have to be undertaken if we are to fully understand this relationship. However, the experiments presented here are a solid foundation upon which subsequent research can be based.

5.7 Differentiating between native and non-native structures using HDXsimulator

The HDXsimulator pipeline was developed as an extension of our earlier work [31] for the purpose of enabling residue-level lnP data to be used, in addition to peptide-level RFU data, as a scoring metric in the evaluation of native vs. non-native structures. In order to appraise the program's ability to make this distinction, we tested it on 7 unique protein decoy sets with its ability to distinguish native from non-native determined by the construction of ROC curves and subsequent AUC values. Our previous work had informed us that HDXsimulator reacted differently to different proteins and it therefore followed that it would also likely react differently to decoy sets produced by contrasting algorithms. In our research, we found only two distinct methods for decoy generation that fit our criteria: Rosetta's Abinitio/Relax applications and 3DRobot, therefore we initially endeavoured to have a Rosetta and 3DRobot

decoy set for each protein. However, we unfortunately only managed to obtain two 3DRobot decoy sets (for BnWT & GFP) before the webserver version of the program closed for the CASP14 season and we did not have time to make use of the local install version before the writing of this thesis. Despite this, the two data sets we did acquire have proven very informative to our overall conclusions and we believe that similar results would have been seen had the we managed to obtain the others. The data sets tested were: BnWT_Rosetta, BspWT_Rosetta, GFP_Rosetta, GFP-nb_Rosetta, GFP-nbmin_Rosetta, BnWT_3DR & GFP_3DR.

5.7.1 Whole-protein data sets

We first looked at the AUC values of the 7 data sets taken as a whole using the RMSE of the calculated RFU compared to experimental HDX RFU as the scoring metric, similarly to our previous work. In terms of RFU, the BnWT_Rosetta data set produced an AUC score of 1, BspWT_Rosetta had 0.0055, GFP_Rosetta had 0.3022, GFP-nb_Rosetta had 0.6674, GFP-nbmin_Rosetta had 0.7525, BnWT_3DR had 0.6006 and GFP_3DR had 0.8044 (Figure 5.12).
These data sets displayed a large amount of variety in terms of AUC score depending on the protein as well as the decoy set. For example, the BnWT_Rosetta decoy set produced an AUC of 1, indicating that HDXsimulator was able to perfectly distinguish between native and non-native structures, however the BnWT_3DR decoy set produced an AUC of 0.6006, despite being based on the exact same protein. A similar result in the opposite direction could be seen for GFP where the Rosetta decoy set produced an AUC of 0.3022 whereas the 3DRobot decoy set produced an AUC of 0.8044. These kinds of results were found quite often in other data sets as we shall see, indicating a certain degree of decoy-dependence on the part of HDXsimulator.

We next compared the whole protein data sets using the RMSE of the calculated lnP vs. the lnP modelled by HDXmodeller as the scoring metric in order to see what, if any, changes could be identified between using RFU or lnP. In terms of lnP, the BnWT_Rosetta data set produced an AUC score of 0.9989, BspWT_Rosetta had 0.0071, GFP_Rosetta had 0.2898, GFP-nb_Rosetta had 0.9862, GFP-nbmin_Rosetta had 0.2463, BnWT_3DR had 0.5222 and GFP_3DR had 0.8264 (Figure 5.13).
For the most part, we found that lnP AUC scores were relatively consistent with those produced using RFU, with 4 of the 7 being within 0.025 AUC of each other. This is not altogether surprising considering that modelled lnP is derived from experimental RFU and calculated RFU is derived from calculated lnP. The BnWT_3DR data set saw a little more variance with an RFU AUC of 0.6006 and a lnP AUC of 0.5222, however the most interesting results are those for GFP-nb and GFP-nbmin which both saw large differences between their RFU and lnP AUCs. For GFP-nb, using lnP as the scoring metric increased its AUC from 0.6674 to 0.9862, whereas for GFP-nbmin the exact opposite is true, with using lnP decreasing its AUC from 0.7525 to 0.2463. These two results indicate that while RFU and lnP are mostly consistent, there is the potential for large deviations to occur, possibly due to inconsistencies in experimental HDX data that lead to diverging experimental RFUs vs. modelled lnPs.

A final point to mention is the RMSE scale of the GFP-nb RFU-level data set, which has an RMSE variance between 8.1-8.8, compared to all the other RFU data sets, which all have RMSEs from 0-0.3. This indicates that HDXsimulator was much less able to estimate RFUs for GFP-nb compared to all the other

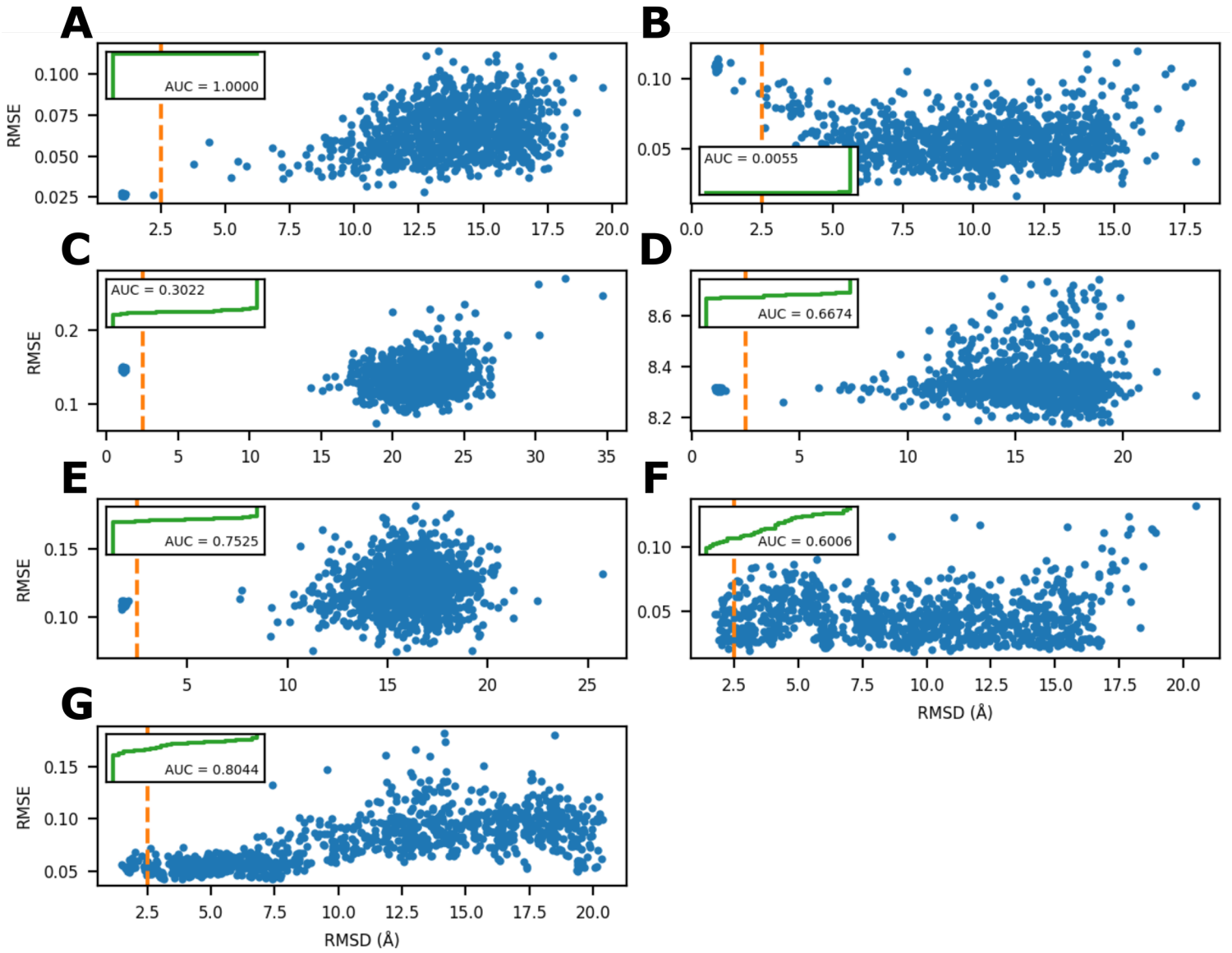


Figure 5.12: ROC plots for whole protein data sets – RFU. Scatter plots (blue points) comparing the decoys' RMSD and RMSE values for whole protein data sets using RFU as the comparison metric from which can be derived ROC curves (inset, green line) and subsequent AUC values (indicated within inset plot). Orange dashed line represents that native cut off of 2.5 Å. **(A)** BnWT_Rosetta, **(B)** BspWT_Rosetta, **(C)** GFP_Rosetta, **(D)** GFP-nb_Rosetta, **(E)** GFP-nbmin_Rosetta, **(F)** BnWT_3DR, **(G)** GFP_3DR.

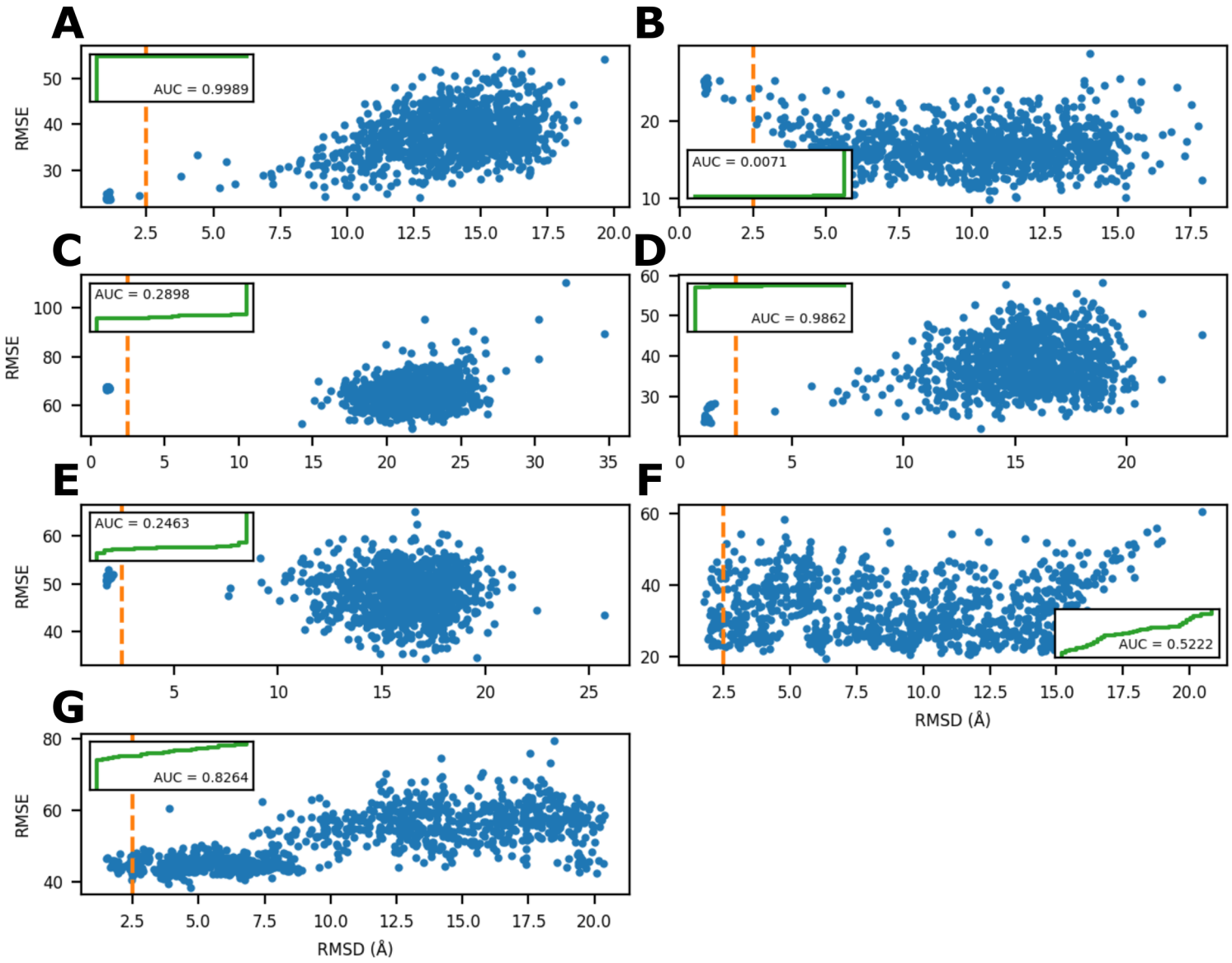


Figure 5.13: ROC plots for whole protein data sets – lnP. Scatter plots (blue points) comparing the decoys' RMSD and RMSE values for whole protein data sets using lnP as the comparison metric from which can be derived ROC curves (inset, green line) and subsequent AUC values (indicated within inset plot). Orange dashed line represents that native cut off of 2.5 Å. **(A)** BnWT_Rosetta, **(B)** BspWT_Rosetta, **(C)** GFP_Rosetta, **(D)** GFP-nb_Rosetta, **(E)** GFP-nbmin_Rosetta, **(F)** BnWT_3DR, **(G)** GFP_3DR.

protein data sets, but only on the RFU-level as this scale discrepancy is not seen for the lnP-level data. This could be due to aforementioned inconsistencies in experimental HDX data, however this would not explain why the GFP-nbmin data, which also saw deviation between RFU and lnP AUC values, does not show this same scale difference in the RFU-level data. The intricacies of unusual HDXsimulator outputs such as these will no doubt require further thorough examination if such details are to be fully understood.

5.7.2 Subsection data sets

In addition to allowing comparison to RFU-level data for whole proteins, having the ability to compare calculated and modelled lnPs enabled us to, for the first time, investigate each protein data set in higher resolution by breaking them down into individual subsections. With this higher resolution view, we could see, for each whole protein data set, which subsections contributed towards higher AUC and which contributed towards lower AUC. For this work, we chose the same subsections as were used for HDXmodeller to see if any comparisons with R-matrix values could be seen at this higher resolution view, however the method is not limited to these specific subsections and any sequential number of residues could be analysed in this way.

For BnWT_Rosetta (Figure 5.14), the A1-Y13 data set produced an AUC of 0.9501, Y13-A43 had 0.8183, A43-F56 had 0.9719, F56-D93 had 0.9678 and W94-I109 had 0.8503.
For BspWT_Rosetta (Figure 5.15), the E8-L16 data set produced an AUC of 0.3099, L16-L34 had 0.3540, L34-E52 had 0.2204, E52-L71 had 0.0016 and Q72-T85 had 0.8079.
For GFP_Rosetta (Figure 5.16),the L7-F46 data set produced an AUC of 0.2727, F46-F99 had 0.7285, F100-F130 had 0.8488, F130-F165 had 0.0809, K166-L207 had 0.5613 and L207-T230 had 0.3822.
For GFP-nb_Rosetta (Figure 5.17), the L5-L21 data set produced an AUC of 0.5239, S22-W37 had 0.7816, E48-F69 had 0.7216, L82-Y95 had 0.9783 and Y95-F103 had 0.9507.
For GFP-nbmin_Rosetta (Figure 5.18), the A2-L22 data set produced an AUC of 0.3292, S23-E48 had 0.0291, E48-T70 had 0.9596, T70-C97 had 0.0802 and D121-H139 had 0.7494.
For BnWT_3DR (Figure 5.19), the A1-Y13 data set produced an AUC of 0.5107, Y13-A43 had 0.3618, A43-F56 had 0.5886, F56-D93 had 0.5338 and W94-I109 had 0.5683.
For GFP_3DR (Figure 5.20), the L7-F46 data set produced an AUC of 0.7943, F46-F99 had 0.5801, F100-F130 had 0.7713, F130-F165 had 0.7805, K166-L207 had 0.6682 and L207-T230 had 0.4807.

We found that the protein data sets used in this work had an interesting mix of results in terms of their subsections, with some having fairly consistent AUC values and some having vastly different AUC values. Those that were quite consistent were: BnWT_Rosetta, GFP-nb_Rosetta and BnWT_3DR; whereas those that saw large variations were: BspWT_Rosetta, GFP_Rosetta, GFP-nbmin_Rosetta and GFP_3DR. For the majority of the data sets therefore, the whole protein AUC value was in fact taking into account differing subsections that had quite disparate AUC results and smoothing across them. This was similar to the results we saw with HDXmodeller where most data sets were a combination of higher and lower scoring subsections that were being amalgamated into a single R-matrix score. Such discrepancies highlight the need for the residue-level lnP method we have developed in this work as this information is lost when only considering whole protein peptide-level RFU data.

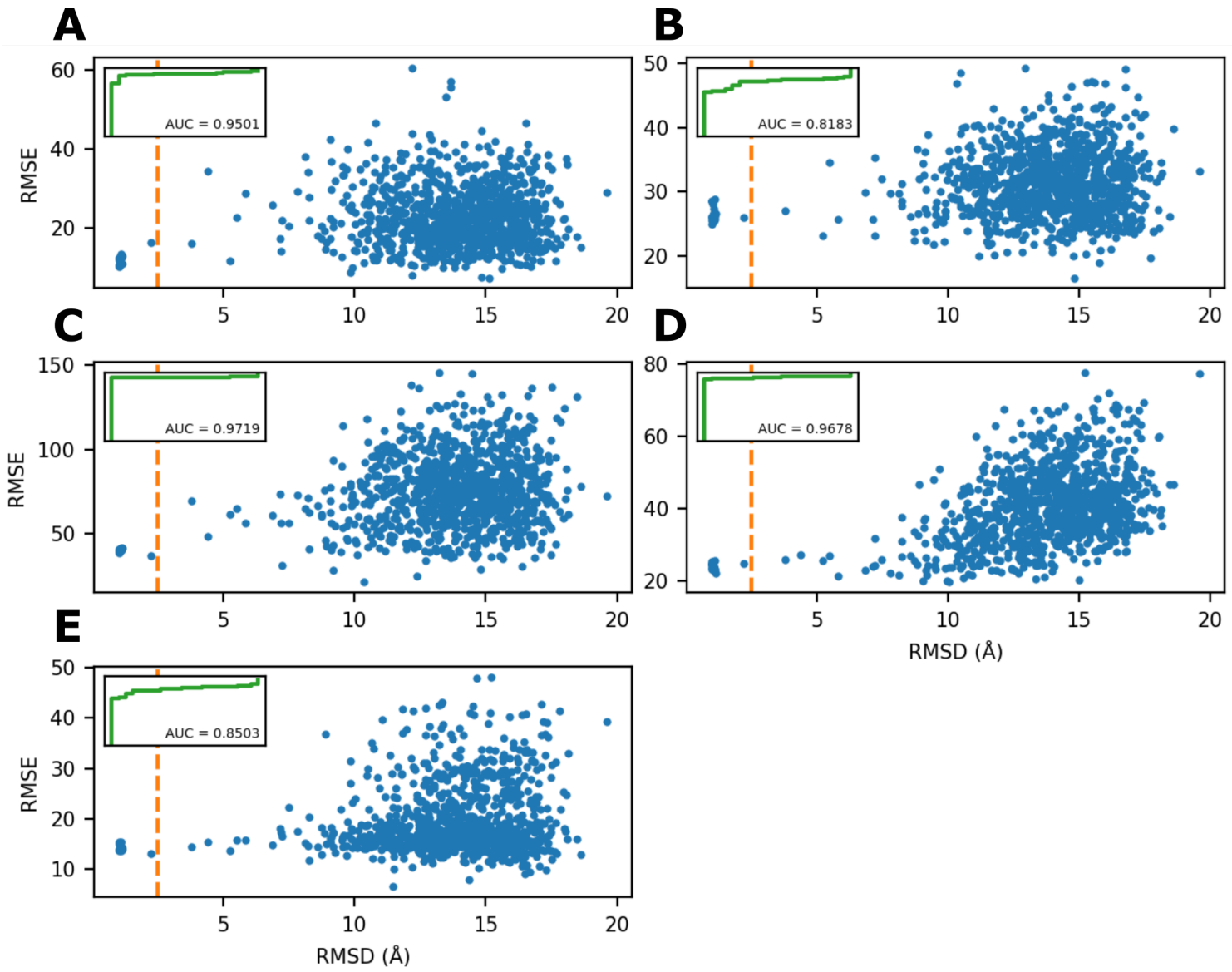


Figure 5.14: ROC plots for BnWT_Rosetta subsections – lnP. Scatter plots (blue points) comparing the decoys' RMSD and RMSE values for BnWT_Rosetta subsections using lnP as the comparison metric from which can be derived ROC curves (inset, green line) and subsequent AUC values (indicated within inset plot). Orange dashed line represents that native cut off of 2.5 Å. **(A)** A1-Y13, **(B)** Y13-A43, **(C)** A43-F56, **(D)** F56-D93, **(E)** W94-I109.

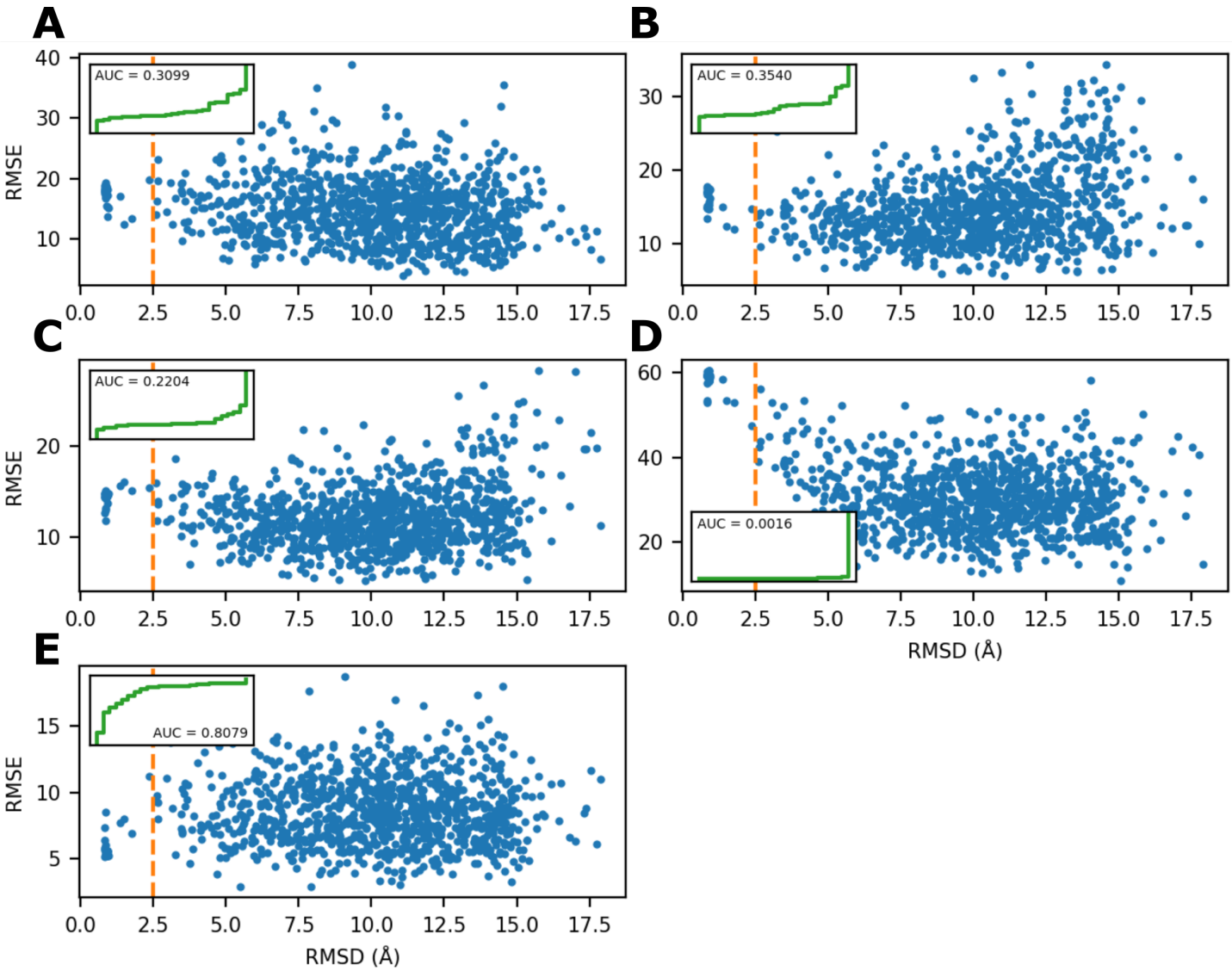


Figure 5.15: ROC plots for BsWT_Rosetta subsections – lnP. Scatter plots (blue points) comparing the decoys' RMSD and RMSE values for BsWT_Rosetta subsections using lnP as the comparison metric from which can be derived ROC curves (inset, green line) and subsequent AUC values (indicated within inset plot). Orange dashed line represents that native cut off of 2.5 Å. **(A)** E8-L16, **(B)** L16-L34, **(C)** L34-E52, **(D)** E52-L71, **(E)** Q72-T85.

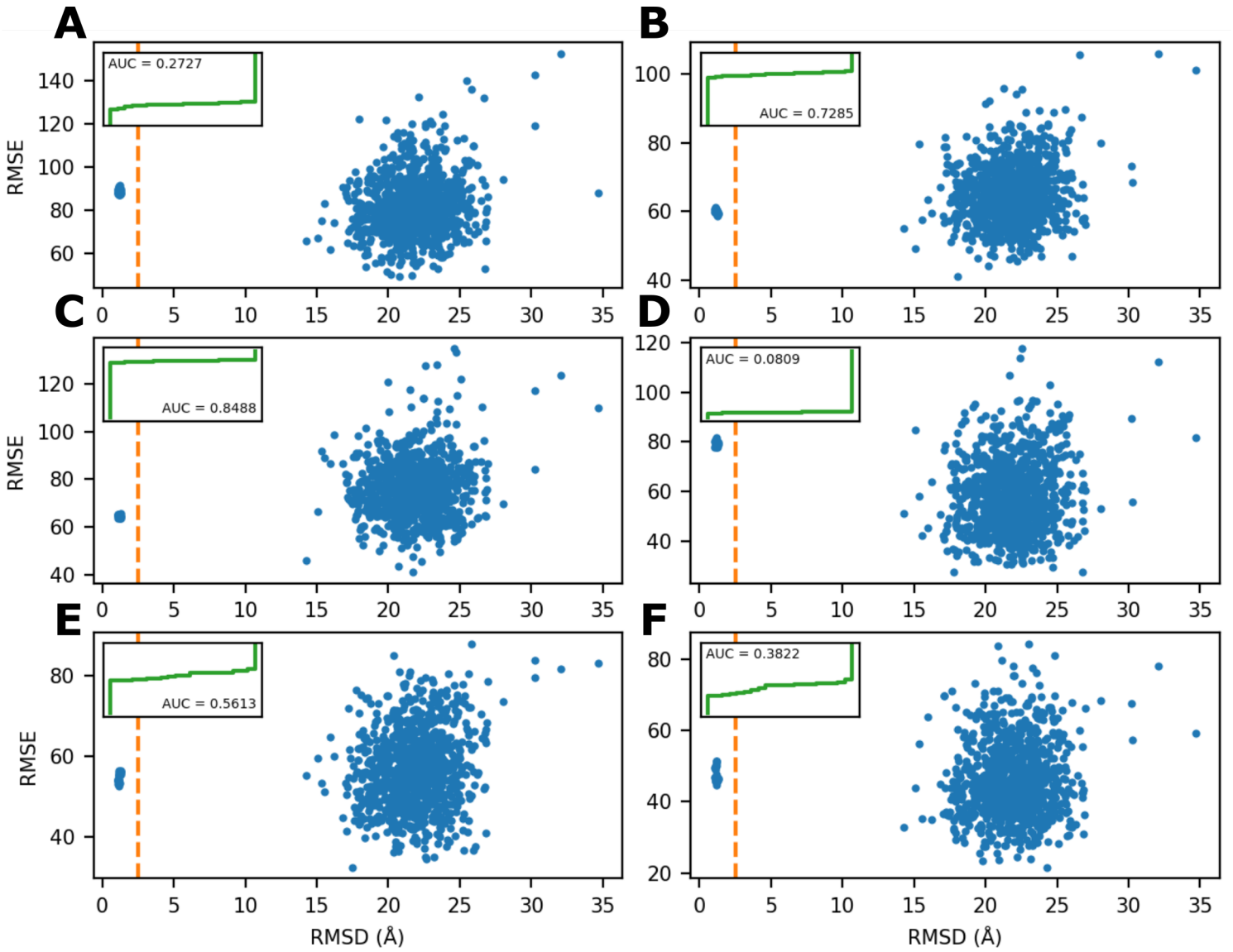


Figure 5.16: ROC plots for GFP_Rosetta subsections – lnP. Scatter plots (blue points) comparing the decoys' RMSD and RMSE values for GFP_Rosetta subsections using lnP as the comparison metric from which can be derived ROC curves (inset, green line) and subsequent AUC values (indicated within inset plot). Orange dashed line represents that native cut off of 2.5 Å. **(A)** L7-F46, **(B)** F46-F99, **(C)** F100-F130, **(D)** F130-F165, **(E)** K166-L207, **(F)** L207-T230.

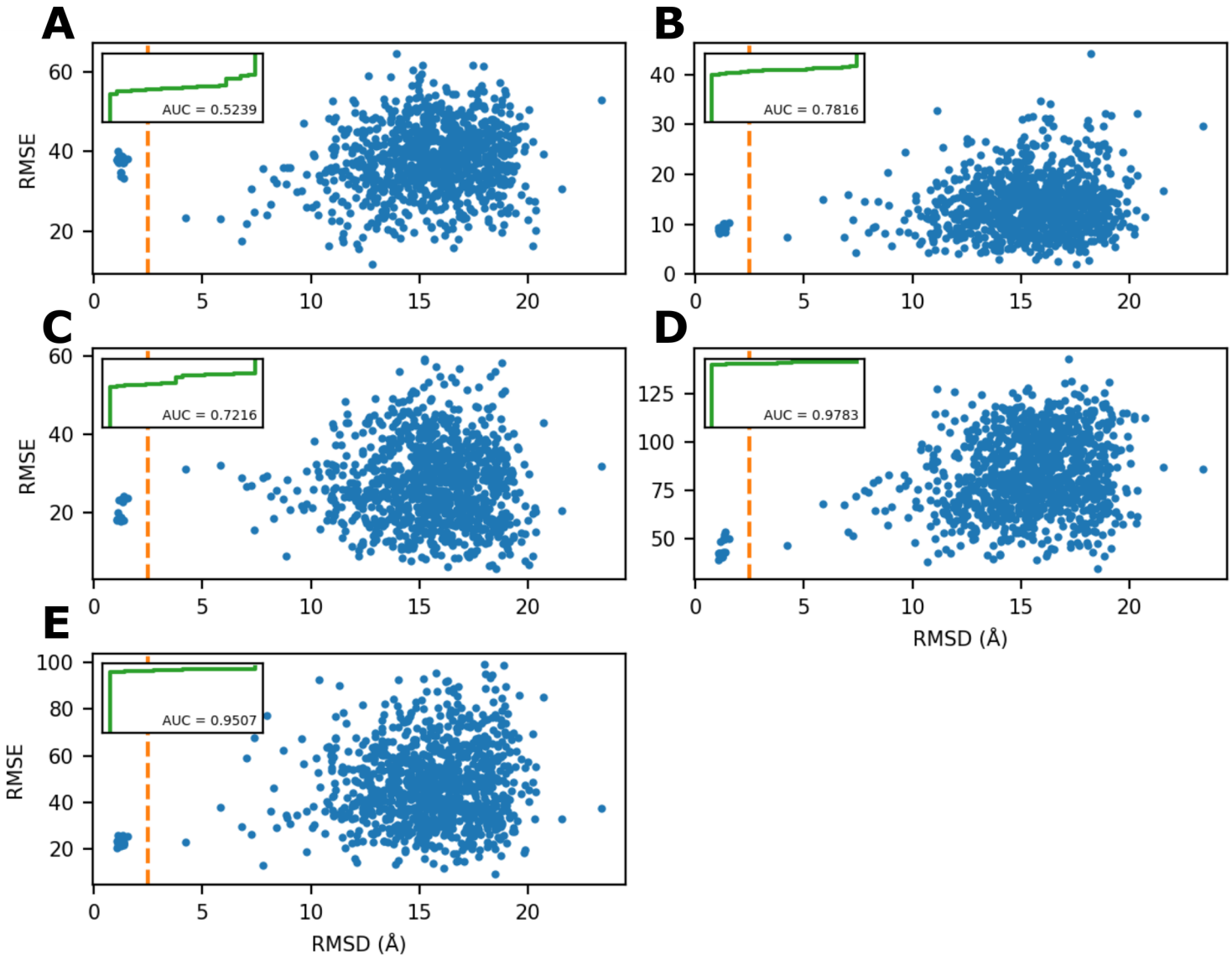


Figure 5.17: ROC plots for GFP-nb_Rosetta subsections – lnP. Scatter plots (blue points) comparing the decoys' RMSD and RMSE values for GFP-nb_Rosetta subsections using lnP as the comparison metric from which can be derived ROC curves (inset, green line) and subsequent AUC values (indicated within inset plot). Orange dashed line represents that native cut off of 2.5 Å. **(A)** L5-L21, **(B)** S22-W37, **(C)** E48-F69, **(D)** L82-Y95, **(E)** Y95-F103.

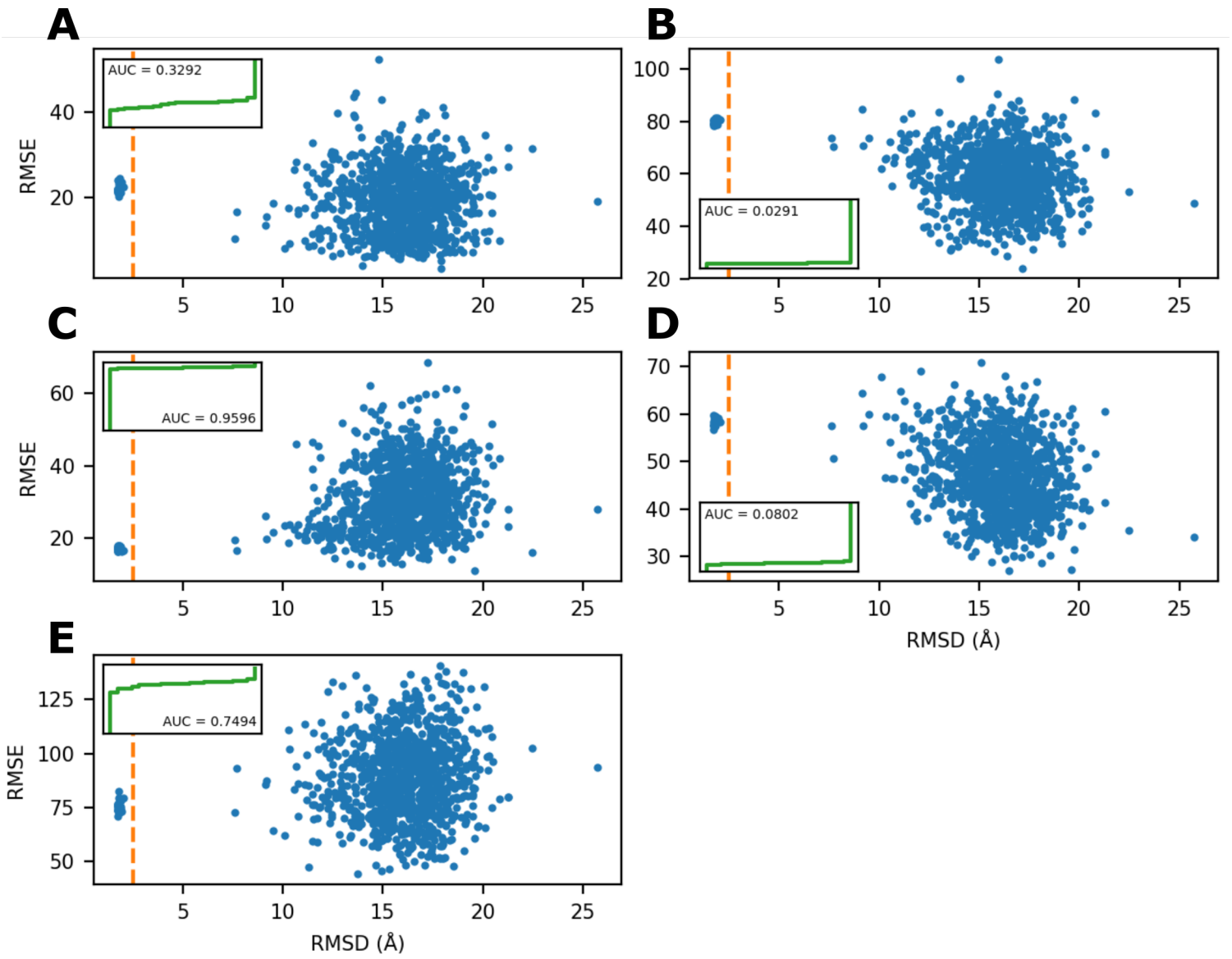


Figure 5.18: ROC plots for GFP-nbmin_Rosetta subsections – lnP. Scatter plots (blue points) comparing the decoys' RMSD and RMSE values for GFP-nbmin_Rosetta subsections using lnP as the comparison metric from which can be derived ROC curves (inset, green line) and subsequent AUC values (indicated within inset plot). Orange dashed line represents that native cut off of 2.5 Å. **(A)** A2-L22, **(B)** S23-E48, **(C)** E48-T70, **(D)** T70-C97, **(E)** D121-H139.

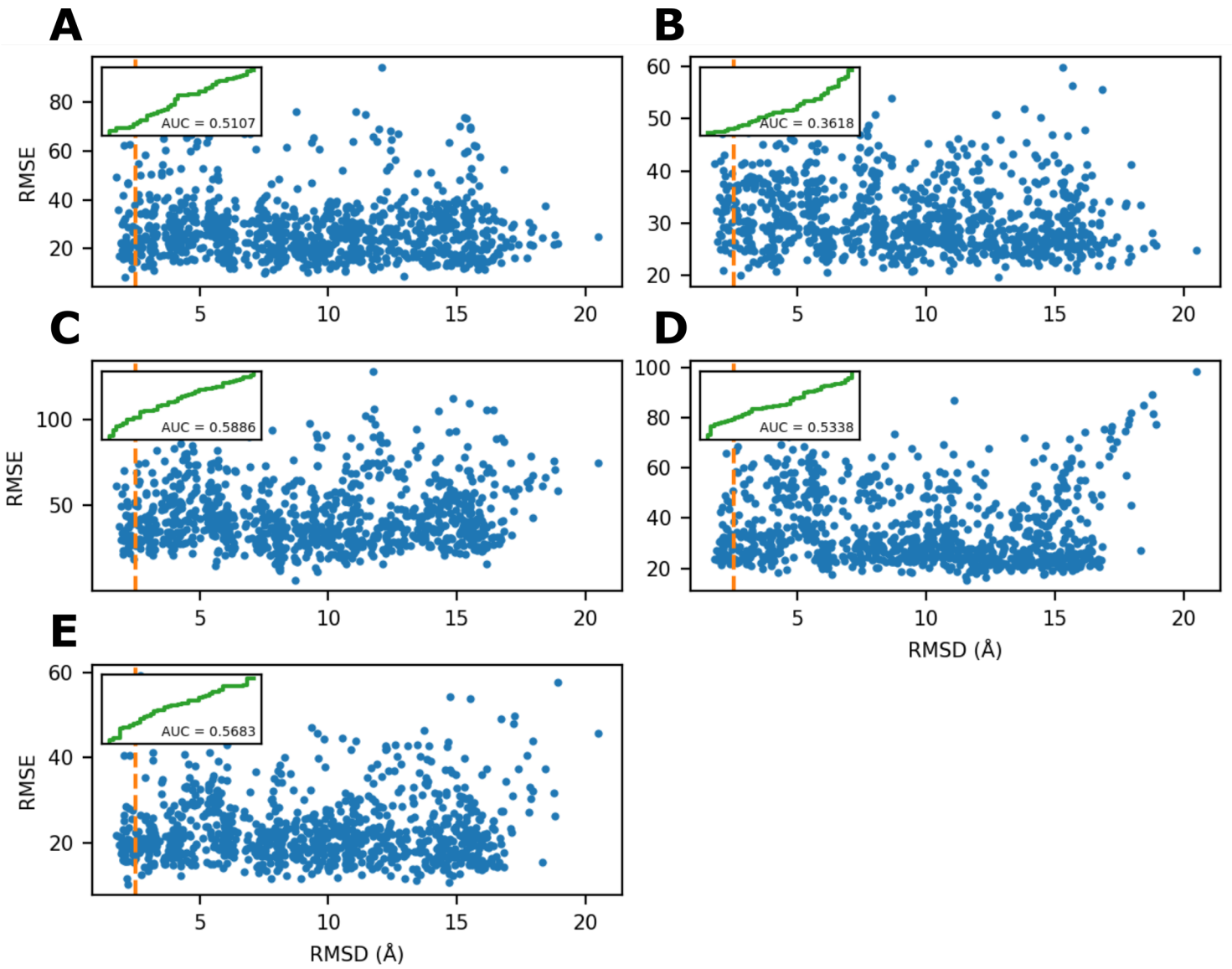


Figure 5.19: ROC plots for BnWT_3DR subsections – lnP. Scatter plots (blue points) comparing the decoys' RMSD and RMSE values for BnWT_3DR subsections using lnP as the comparison metric from which can be derived ROC curves (inset, green line) and subsequent AUC values (indicated within inset plot). Orange dashed line represents that native cut off of 2.5 Å. **(A)** A1-Y13, **(B)** Y13-A43, **(C)** A43-F56, **(D)** F56-D93, **(E)** W94-I109.

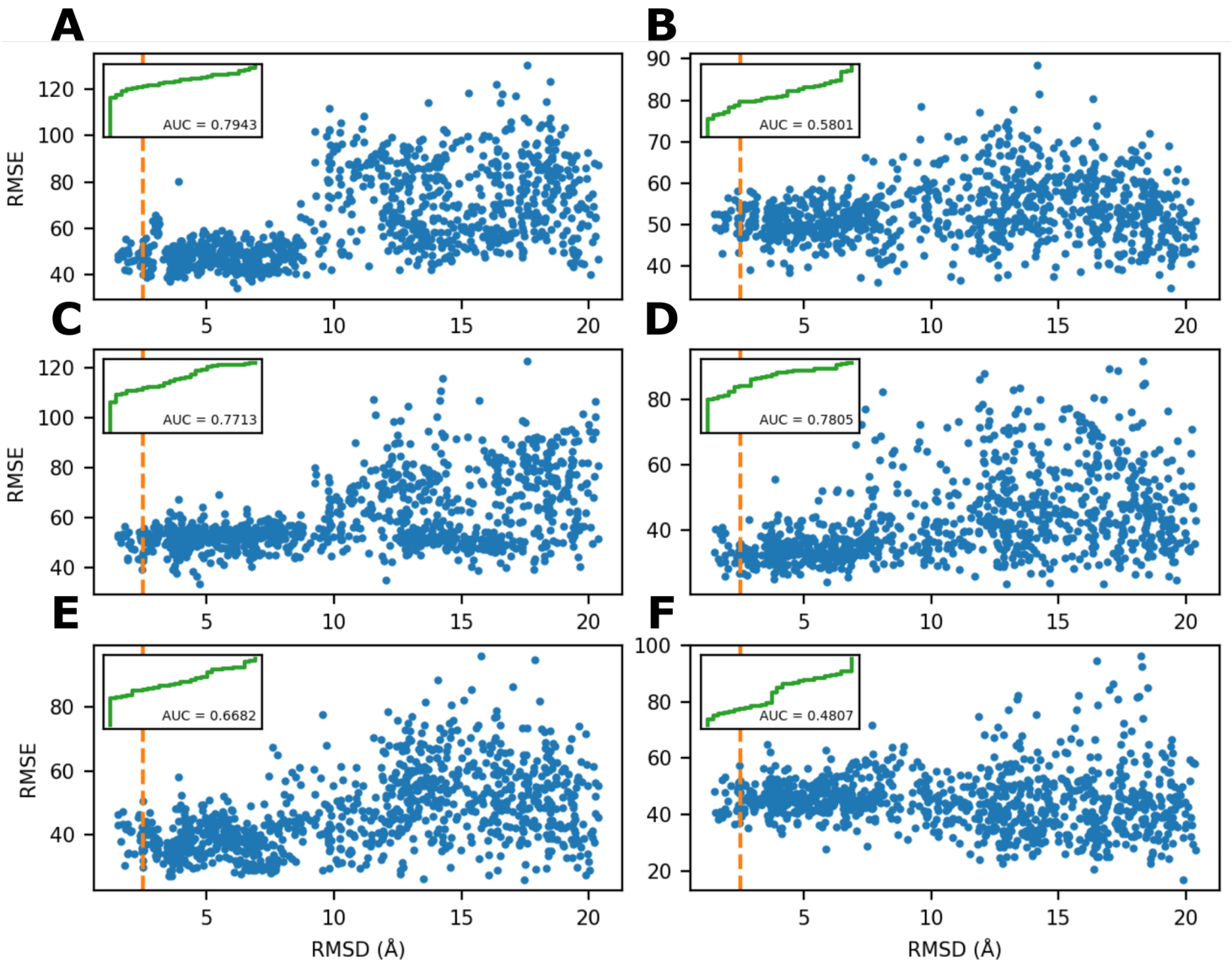


Figure 5.20: ROC plots for GFP_3DR subsections – lnP. Scatter plots (blue points) comparing the decoys' RMSD and RMSE values for GFP_3DR subsections using lnP as the comparison metric from which can be derived ROC curves (inset, green line) and subsequent AUC values (indicated within inset plot). Orange dashed line represents that native cut off of 2.5 Å. **(A)** L7-F46, **(B)** F46-F99, **(C)** F100-F130, **(D)** F130-F165, **(E)** K166-L207, **(F)** L207-T230.

5.7.3 Combined data sets

With all these AUC scores produced for both the protein data sets as a whole as well as individual subsections, we wanted to get an idea of the AUC scores we would obtain for the data sets all combined together. This would give us a statistical score that summarised HDXsimulator's ability to determine native structures, taking into account all the individual data sets we used in this work. We produced three combined data sets: one taking into account all the whole protein's scores on the RFU level, one taking into account all the whole protein's scores on the lnP level and one taking into account all the subsection's scores on the lnP level (Figure 5.21). To do this, we combined all the RMSD and RMSE values comprising individual data sets into one large data set and then determined the AUC score for this combined set. The combined whole proteins set on the RFU level produced and AUC score of 0.5923, the combined whole proteins set on the lnP level produced an AUC of 0.5866 and the combined subsections set on the lnP level produced and AUC of 0.5554.

We found that our average AUC values were very similar between the whole protein RFU and lnP data, differing by 0.0057, which tells us that the metrics are virtually identical in their ability to correctly classify native vs. non-native decoys. In terms of comparing whole protein to subsection lnP, we can again see that the AUC values are very similar, differing by 0.0312. This tell us that subsection level data is an accurate representation of the individual parts that make up the whole protein data. If instead we had seen substantial deviation of the subsection data AUC from the whole protein data AUC, it would indicate that either one or the other is being calculated incorrectly, or possibly that we are seeing a similar effect to that which was observed with HDXmodeller where altering the data set can fundamentally change the programs ability to calculate lnP.

The AUC values themselves, at approx. 0.6, indicate that, taking into account all the data sets, the methodology has an approx. 60 % chance to correctly identify any given decoy as either native or non-native. This is a positive result but is of course far below the accuracy that would be required for researchers to trust that the structures identified are in fact native or not for use in their own research. With this in mind, there are several points that we have identified as being potentially part of the cause of these lower overall AUC values. These possible areas of future improvement shall be considered in the Discussion.

5.8 Summary

In this chapter, we have seen results presented that contribute to the overall stated aim of this thesis, that is, the ability of HDXsimulator to correctly classify a protein structure as being native or non-native. These results cover: the final yields of barnase as well as native MS validation of the interaction between it and barstar, HDX-MS experiments on all the binary PPIs investigated in this work, how HDXmodeller modelled residue-resolved lnP values for each interaction, the results of relaxational MD simulations as well as protein docking for use in future work on binary PPI structure determination, the results of the investigation into the boundaries of modelling protein conformations with HDXsimulator and finally, the primary goal of this thesis, the results of our inquiry into HDXsimulator's ability to distinguish between native and non-native structures.

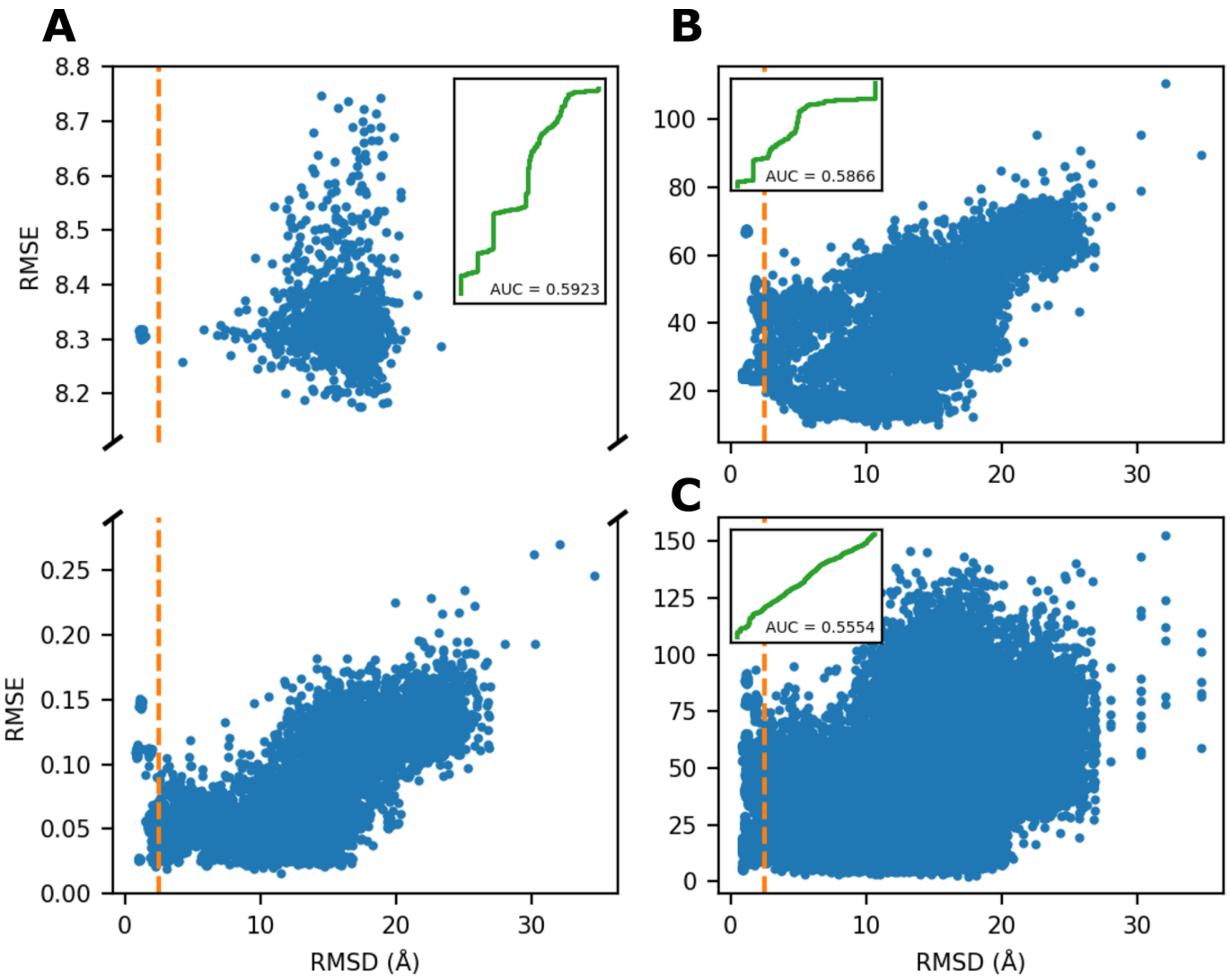


Figure 5.21: Combined ROC plots for all whole protein and subsection data sets – RFU/lnP. Scatter plots (blue points) comparing the decoys' RMSD and RMSE for whole protein and subsection data sets using RFU/lnP as the comparison metric from which can be derived ROC curves (inset, green line) and subsequent AUC values (indicated within inset plot). Orange dashed line represents that native cut off of 2.5 Å. **(A)** Combined whole protein data sets – RFU. Spilt axes necessary for viewing due to GFP-nb_Rosetta data set having much higher RMSE values compared to the others. **(B)** Combined whole protein data sets – lnP. **(C)** Combined subsection data sets – lnP.

The final yields of BnWT and BnH102A after the improvements detailed in chapter 4.2 resulted in 3.08 mg/L and 34.2 mg/L pure protein being produced respectively. Native MS experiments carried out on BnWT, BnH102A and BnWT:BspWT confirmed the correct MWs and binding of these proteins. The identity and binding of proteins not produced during this study were confirmed during the course of HDX-MS experiments. The acquisition of these proteins was the bedrock on which all our subsequent experiments were based as without them there would be no experimental data with which to compare our simulated results.

HDX-MS experiments were carried out in order to provide an experimental reality against which HDXsimulator's calculations could be compared. In addition, to ensure the validity of said experimental reality, a comparison was made of the locational data elucidated by the HDX profiles to the scientific community's understanding of the binding interaction, as found in the literature. We found that, in all cases, our results for the interactions when mapped onto the protein structures of the binary PPI matched with available crystal structures of the bound proteins. Therefore we are confident in saying that our HDX results accurately represent the interactions of the binary PPIs and so are valid to take forward and use for comparison with calculated data.

HDXmodeller enables residue-level lnPs to be calculated from peptide-level HDX data, with the reliability of the modelled values being represented by a novel auto-validation matrix. Such a method of HDX modelling has long been a goal of practitioners in the field [99–104], however it has proved challenging to accurately determine all the underlying variables. Additional restraints encoded by the peptide ion envelopes provide a potential way forward as they contain clues regarding the distribution of isotope along a peptide [105–107]. We used HDXmodeller as a way to obtain modelled residue-level lnPs in order to enable comparison with the residue-level lnPs calculated by HDXsimulator. We found that the auto-validation R-matrix scores varied significantly across our data sets with differences seen not only across proteins but across protein subsections as well. Whole-protein data sets for BnWT, BspWT and GFP-nbmin displayed R-matrix scores in the "high" data bin (≥ 0.7) while BnH102A, BsY29F, GFP and GFP-nb displayed R-matrix scores in the "fair" data bin (0.5-0.69). Subsection-level data sets varied even more, including subsections within the same protein, indicating that local factors play a significant role in the protein's overall amenity to modelling (data available in Table 5.2).

In order to enable future use cases of our method on binary complexes, we carried out MD simulations in order to generate relaxed structures of our proteins that could then subsequently be used in docking simulations. We used NAMD to produce the relaxed structures, for which two different types of simulations were run: bound (with the binding partner present) and unbound (with the binding partner missing). For each type, the simulation was run for between 10-20 ns, until the RMSD of subsequent frames stabilised vs. the initial frame. For most proteins, this occurred within a few nanoseconds, however larger proteins such as GFP took longer. The average structures were calculated using VMD's RMSDTT and the frames closest to the average carried forward into the docking stage of the project.

Protein-protein docking was carried out on two of the binary PPIs: BnWT:BspWT & GFP:GFP-nb using HADDOCK on the unbound conformers of the constituent proteins. With a native cut off of 2.5

Å, the BnWT:BspWT interaction produced 10 native structures out of 1,000 selected to undergo flexible refinement. If the cut off was raised to 3 Å, this number increased to 26 out of 1,000. When the poses were visualised in Chimera, we could see that the location on BspWT relative to BnWT was always on a horizontal plane through the true binding site with no poses found with BspWT on "top" or on "bottom" of BnWT. In comparison the GFP:GFP-nb interaction produced no native poses with the closest being 7.2 Å from the crystal structure. When looking at the poses in Chimera, we can see that while the nanobody is mostly on the correct side of GFP, they are rotated at all sorts of angles compared to the crystal structure, giving rise to the high RMSD values. We attributed this to the lack of "lock and key" geometry for this complex as well as a lack of localisation of the AIRs for GFP-nanobody. Therefore we conclude that further optimisation is required before this docking technique can be brought forward for use with HDXsimulator, however we are confident that HDX data has the potential to be capable of accurately guiding protein-protein docking simulations.

When exploring the boundaries of modelling protein conformation using HDXsimulator, we intended to chart HDXsimulator's capabilities and limitations to enable us to improve future versions of the program. After much trial and error, we found that the optimal method for error generation was using HDXsimulator itself with suboptimal scaling factors which produced the correlated AUC and R ^2 values that we were looking for. The results of our optimisation work tell us that the relationship between AUC and RMSE/R^2 is much more complicated than we had initially thought, with hidden factors challenging our understanding. Protein subsections were shown to have quite different profiles from each other, showing that, like with HDXmodeller, data generated for a protein as a whole is not necessarily representative of its individual constituent parts. It is clear that more work will have to be undertaken if we are to fully understand this relationship.

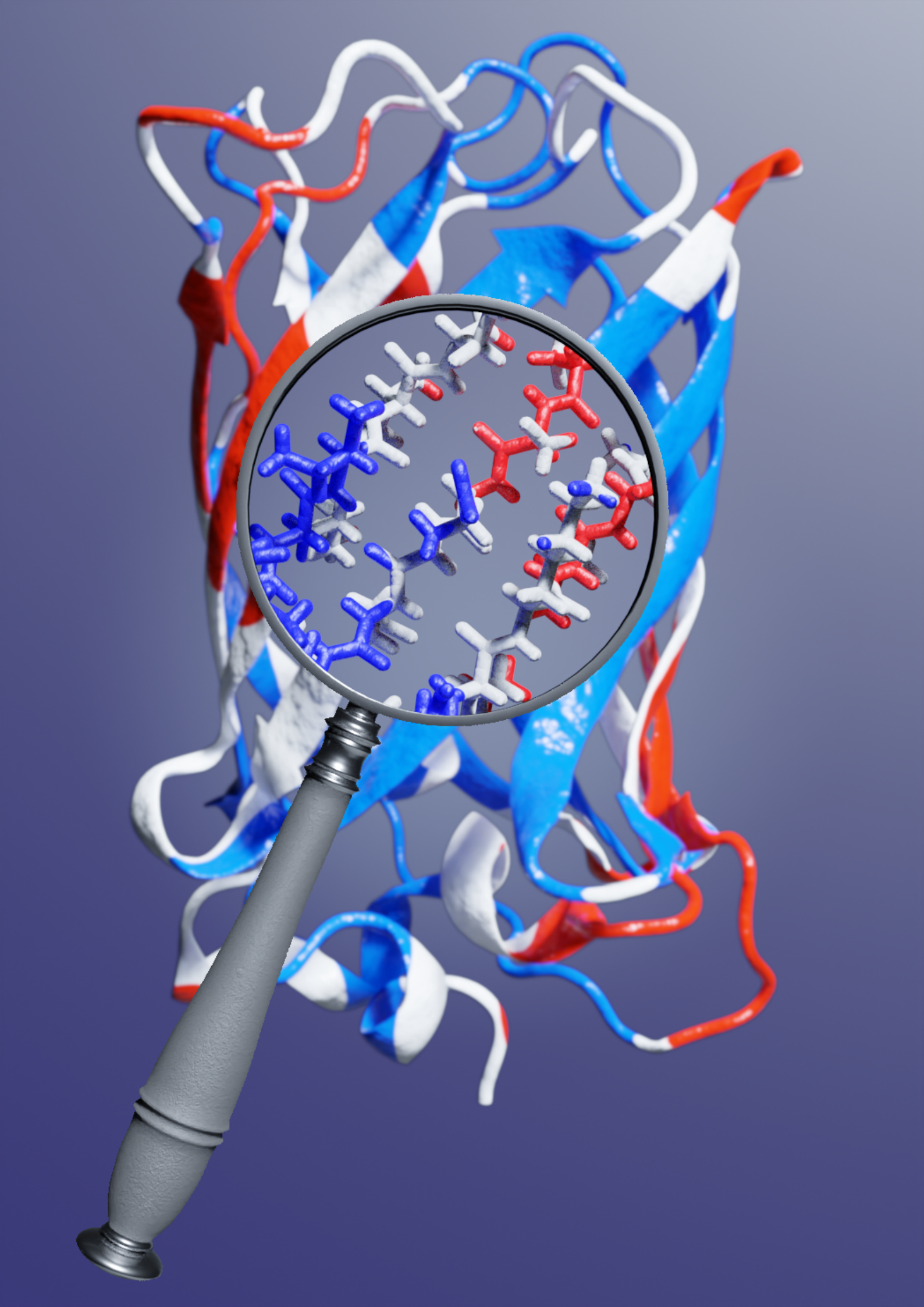
The final part of our work was to test HDXsimulator's ability to determine a protein's *in silico* structure, generated by Rosetta or 3DRobot, as being native or not. This is accomplished by deriving the in silico protection factor from 3D structures, a process which has long been of interest to researchers [29, 30] and has become increasingly popular of late with several different methods being attempted [108–112]. For HDXsimulator, we tested both residue-level lnP data on the level of whole proteins and subsections as well as peptide-level RFU data on the level of whole proteins in order to evaluate which metrics worked the best.

For whole protein data sets, we found a great variety in terms of the AUC scores that measure HDXsimulator's ability to tell native from non-native. Proteins such as BnWT's Rosetta data set produced an AUC of 1, meaning that HDXsimulator could accurately classify structures in almost every case; while proteins such as BspWT's Rosetta data set produced an AUC of 0.0055, meaning that almost every structure was misclassified. We found that those AUCs derived from lnP data closely matched those derived from RFU data in almost all cases.

A similar picture was seen when looking at lnP comparison data for individual protein subsections, with certain ones being highly conducive to accurate structure classification and others not, even within the same protein. These results highlight the need for the residue-level lnP method we have developed in this work as this information is lost when only considering whole protein peptide-level RFU data.

Finally, we combined all the AUC scores produced into three data sets: one charting all our RFU

data for whole proteins, one for lnP data for whole proteins and one for lnP data for protein subsections. When taken together, the combined whole proteins set on the RFU level produced and AUC score of 0.5923, the combined whole proteins set on the lnP level produced an AUC of 0.5866 and the combined subsections set on the lnP level produced and AUC of 0.5554. This indicates that, taking into account all the data sets, the methodology as it currently stands has an approx. 60 % chance to correctly identify any given decoy as either native or non-native. At greater then simple random chance, this result is a positive first step towards the accurate determination of *in silico* protein structure and so backs our stated goal of leveraging HDX-MS to determine native protein structures.



6 Discussion

6.1 The dUTPase : Stl interaction

Mapping the interaction of various dUTPases with their proteinaceous inhibitor Stl was a collaborative project undertaken between the Borysik group and the Vértessy group from the Budapest University of Technology and Economics. This project utilised the far more traditional use-case of HDX-MS: locating the site of interaction between a protein and a ligand, in this case another protein. Our overarching goal was the investigation of the mechanism by which Stl can inhibit different dUTPases that share relatively little structural or sequence homology, with a key answer to this question being the location of binding between the two proteins. The project was split into two sub-projects: the first being the investigation of the interaction between hDUT and Stl and the second being the investigation of the interaction of a trimeric and a dimeric dUTPase with Stl.

For the investigation into the hDUT:Stl interaction, HDX-MS was one of a number of analytical techniques brought to bear, the other major one being SEC-SAXS, to try and elucidate the mechanism of inhibition. We ran HDX-MS experiments at three different time points: 1 min, 10 mins & 100 mins on hDUT and Stl complexed with each other as well as the two proteins on their own in order to generate bound minus unbound difference plots to allow quantitative appraisal of the data. Difference data was also mapped onto the structures of the individual proteins to allow qualitative appraisal of the data.

In terms of hDUT, the largest negative Δmass were found in the region of H34-L50 with other significant Δmass seen in the region of the C-terminus. However, these C-terminal mass differences converged more rapidly then those at the N-terminus, implying a weaker interaction. In addition, residues A89-G110 also showed modest negative Δmass, suggesting a role in the interaction. In terms of Stl, a very significant negative Δmass could be seen localised to a limited number of residues (Y98-Y113) with more minor changes seen across the whole of the protein, indicating that Stl may undergo a global decrease in dynamics upon binding to hDUT. These results indicate that although binding in Stl is localised to a very specific region, the interaction is propagated across the entire rest of the protein.

When combined with the other biophysical data presented in [79], our experiments provided additional and conclusive evidence for the formation of a complex between hDUT and Stl as well as the location of the interaction. The regions of significant uptake decrease observed in hDUT overlap with the first three of the five conserved motifs found within trimeric dUTPases, providing evidence for how Stl can bind to and inhibit different trimeric dUTPases which share comparatively little sequence homology. The C-terminal region of hDUT, which includes the fifth conserved motif, also shows a significant amount of deuterium uptake decrease upon complexation, however fluctuations of the HDX rates indicate the interactions involving this segment are weaker and more transient compared to those of other regions.

Previous work conducted on hDUT had shown that the binding of dUTP and Stl to hDUT are mutually exclusive [74,75]. This combined with our new locational data enabled a mechanistic model to be proposed in which Stl is only allowed to manoeuvre into the substrate binding pocket of hDUT if access

is not hindered by either dUTP or the closed conformation of the flexible motif 5. The complex of hDUT and Stl may be further stabilised by motif 5 which is consistent with the decrease in Δmass observed in this motif upon complex formation. However, a caveat to this proposal is that, in accordance with our results, the stabilising effect is likely to be transient in nature.

Previous experiments by the Vértessy group as well of those of other collaborators have led to the theory that the Stl dimer falls apart upon complexation with hDUT, however we saw almost no positive Δmass shifts in the Stl difference plot which would support this idea. Therefore we propose that the dimer interface of Stl might overlap with the Stl:hDUT interaction surface and hence any positive Δmass signals are suppressed due to the presence of hDUT in the same location. This idea was previously theorised by the Vértessy group [78] and is given more credence by our data as well as a collaborator's SEC-SAXS data which utilised our HDX-MS results as restraints.

With our results as a guide, SEC-SAXS has produced a model of the hDUT:Stl complex showing a stoichiometry of $hDUT_3Stl_3$ as well as $hDUT_3Stl_2$. A schematic model of hDUT:Stl complex assembly and Stl-DNA interaction is therefore proposed ([79] Figure 6 in paper) which explains the data this collaborative effort has gathered. It posits that while the complexation of hDUT to its substrate dUTP prevents Stl binding and inhibition, the complexation of Stl to its substrate DNA does not prevent hDUT binding, causing Stl dimer dissociation and the formation of $hDUT_3Stl_3$ as well as $hDUT_3Stl_2$ complexes.

In conclusion, hDUT has a prominent role in maintaining genome integrity via the conversion of dUTP to dUMP, thus sanitising the nucleotide pool and preventing the misincorporation of uracil into DNA. This essential task has led to the enzyme being considered as a potential target for oncotherapies and accordingly several small molecular drug targets have been developed to inhibit hDUT function [80]. In this collaborative study, we set out to understand and document the molecular mechanism for the inhibition of hDUT by Stl through the formation of a complex between the two proteins using a range of biophysical techniques, including HDX-MS. The present results on the interaction between hDUT and Stl, including the acquired SEC-SAXS model based on HDX data restraints provide plausible explanations for the observed mutual inhibition of Stl and hDUT's physiological function in their complex. Our HDX results enabled a clear delineation of the peptide segments around the hDUT active site that are also involved in the binding surface to Stl and this data is consistent with the observed inhibition of hDUT's enzymatic function, as well as the mutually exclusive contest between dUTP and Stl to bind to hDUT. Importantly, both conserved motifs 2 and 3 which are have previously been shown to be involved in dUTP substrate binding are identified by HDX-MS in the current study as being involved in the interface for Stl binding, validating a hypothesis that had previously been based only on computational models.

With these results in mind, we argue that proteinaceous inhibition of hDUT by Stl has the potential to be a powerful tool in the arsenal of researchers investigating the function of dUTPases and suggest further exploitation of Stl as a specific inhibitor of dUTPases for therapeutic-based applications. It should also be noted that as our HDX results elucidate clear Stl peptide segments responsible for the inhibition of hDUT, the possibility of developing peptide-based inhibitors cannot be underestimated.

Discussion of the results relating to the interaction of trimeric and dimeric dUTPases with Stl can be found in the paper in Appendix M.

6.2 Production of barnase and barstar

We decided to produce two of the proteins we would use for the benchmarking of HDXsimulator in house. Barnase and barstar have long been known to the scientific community in the study of protein folding and as such there is a wealth of literature regarding their structures and kinetics, including many mutants. We collaborated with the Ikura group from the Tokyo Medical and Dentistry University who supplied us with plasmids for both barnase and barstar as well as protocols for their production and purification. We had hoped that this portion of the project would therefore be quite straight forward, however this proved not to be the case and a substantial amount of optimisation was required to produce and purify enough of each protein to meet our needs.

We believe that the reasons for most the difficulties we faced were twofold: the first was the relative lack of protein production experience on our part compared to the Ikura group, especially experience relating to these specific proteins, and the second was the difference in equipment used in our labs compared to theirs. As any experienced researcher can attest to, over the course of working on a system, one unconsciously learns and carries out many small steps which do not necessarily make it into a final protocol, but which nevertheless have an impact on the final result. Therefore, we can easily see how a protocol that worked perfectly for one group may be a failure for another. This is what we believe to be the primary reason why we had to develop entirely new protocols for both proteins, as detailed in chapter 4, although of course it is also possible that the initial plasmids may have been faulty or any number of other possible causes of error.

Eventually we developed the barnase protocol to produce more than enough protein for our needs and also substantially optimised the barstar protocol, however in the interests of time we contracted out the final part, His tag cleavage, to an outside source. With both proteins in hand, we confirmed their identity and successful binding to each other by native MS, in which BnWT (theoretical mass: 12,383 Da) had an experimentally determined mass of 12,384 $\pm$ 0.55 Da; BnH102A (theoretical mass: 12,317 Da) had an experimentally determined mass of 12,323.45 $\pm$ 5.44 Da and the BnWT:BspWT complex (theoretical mass: 22,587 Da) had an experimentally determined mass of 22,587 $\pm$ 0.02 Da. These experiments are evidence for the correct character of the barnase WT and H102 mutant as well as the successful binding of the two WT proteins. Native MS was used for these proteins/interaction and not all proteins/interactions for the sake of expediency: with the conformation of the identity of both in house proteins, as well as the pWT barstar and the BnWT:BspWT complex, it was deemed likely that the other proteins/interactions would be of the correct character as well. Given that the identity of the other proteins could also be confirmed during the course of our HDX-MS experiments, it was deemed a better use of time to take this route rather then use additional native MS as this way we could kill two birds with one stone. At this stage we considered the protein production stage of the project to have been completed successfully and could finally move on to the collection of novel data.

6.3 HDX-MS experiments on binary PPI protein complexes

With the requisite proteins produced or otherwise acquired, we moved on to the collection of novel HDX data which would enable the subsequent benchmarking of HDXsimulator. To that end, for each binary PPI, we collected HDX profiles for both proteins in the unbound state as well as the bound state at 0.25, 1, 5, 25, 120 & 480 minute time points. BEX and IEX were collected for each data set in order to correct RFU values for extraneous exchange. Corrected RFU profiles for the individual proteins would go on to be used to develop HDXsimulator's ability to classify individual decoy structures, whereas calculated difference plots will see use in the future in developing HDXsimulator's ability to classify bound docking poses when that method has been fully developed.

The corrected peptide-level RFU plots showed that all of the proteins displayed behaviour indicative of a well-folded protein, with the various time points showing a good spread of RFU values across most peptides. This was the single most important aspect we needed to confirm in order to allow subsequent comparison against simulated data. The region from approx. peptides 25-35 in the barstar profiles that we had previously identified as being potentially unfolded still displayed some of the same signs: having both high RFU values and the time points being relatively close together, however to a much lesser degree than what we had observed previously. Therefore we were satisfied that this protein was now completely folded, an attribute further confirmed by our native MS data of the BnWT:BspWT complex showing almost no free protein peaks.

The other salient point to mention with these RFU plots was the low levels of total uptake seen in some of the data sets. Smaller proteins such as barstar saw the higher time points approaching an RFU on 1, however in the larger proteins such as GFP, we could see that even the 480 minute time point didn't come close to achieving deuterium saturation in the majority of the peptides. Given that this is only seen in the larger proteins, it is mostly likely that 480 minutes was simply not long enough to achieve complete deuterium saturation and that longer time points would have seen RFU values approaching 1.

Residue-level quantitative difference plots and qualitative coloured structures were constructed in order to allow us to ascertain the veracity of the interaction between both members of each binary PPI, the former of which will be used in the future for the classification of docking poses as either native or non-native. Most data sets show strong uptake differences that are spread over a large amount of the protein's sequence which makes sense considering that most of the proteins are quite small and so the interaction surface takes up a large proportion of the total sequence. This combined with the comparative low-resolution nature of peptide HDX is why all proteins except GFP see the majority of their residues above the 99 % CI threshold. This data is given more support when mapped onto crystal structures as it can be qualitatively observed that, for each binary PPI, those regions on both proteins with the highest uptake difference are juxtaposed opposite each other in three-dimensional space.

Overall, we are confident in the accuracy of our data as a true representation of the interactions between the proteins involved in the binary PPIs presented in this work. The RFU plots for each individual protein do not display any serious irregularities indicative of a fundamental flaw with the proteins and the combination of native MS data, highly significant difference plots and mapped structures all pro-

vide strong evidence that our assessment of the interactions matches up with the scientific consensus. Therefore we were happy to take all these data sets forward to use as benchmarks for the development of HDXsimulator.

6.4 Obtaining residue-resolved lnPs using HDXmodeller

Unlike NMR, HDX-MS cannot typically be used to calculate experimental lnPs due to the low-resolution nature of its data. However thanks to recent advances is methodology as presented in (Salmas and Borysik, 2020. Accepted for publication), we now have the ability to calculate modelled lnPs from experimental HDX-MS data using the program HDXmodeller. The standout feature of the program is an auto-validation function that takes into consideration the quality of the entire optimisation process through the use of a covariance matrix over different replicates. Replicates are compared against each other in a pair-wise manner and their degree of correlation quantitatively assessed through the calculation of an R-matrix score, with high scores indicating high modelling confidence and low scores indicating low modelling confidence. There are several factors which are known to affect the R-matrix value and, although attempts have been made, we have not been able to devise a scoring metric which can take them all into account and describe how likely any particular data set is to have high confidence. Some of these factors include the occupancy of any particular amino acid as well as precise shape of the peptide map, taking into account specific overlaps etc., in addition to the RFU values of the peptides themselves and how closely they match up. For these reasons, it was not possible to predict how HDXmodeller would react to any particular data set and rules which seemed to apply to one would not apply for another and so a substantial amount of tweaking was done to each individual data set to try and get the best results out of it.

We used HDXmodeller to calculate modelled lnPs for our 7 protein data sets: BnWT, BnH102A, BspWT, BsY29F, GFP, GFP-nb & GFP-nbmin because these values would be needed for use as a comparison data set for our residue-level work with HDXsimulator down the line. Modelled lnPs were calculated for these proteins as a whole as well as for specific demarcated subsections in order to see if any differences were visible at a higher resolution view. On the whole protein level, we saw that all of the data sets had an R-matrix score classifying them as either ”high” (≥ 0.7) or ”fair” (0.50-0.69), however when we moved to the subsection level we saw considerable differences start to emerge that were not visible in the lower resolution view. Almost every data set contained subsections with R-matrix scores in the ”low” (< 0.5) category as well as the fair and high categories, indicating that the program’s confidence in its modelled lnPs fluctuated dramatically across the sequence of the various proteins.

Having the ability to demarcate regions which broadly contribute to high scores and those which broadly contribute to low is a useful analysis tool which we made use of during the HDXsimulator section of the project. For this, a high resolution view was important because it would enable us to attempt to correlate the R-matrix score of HDXmodeller with the AUC score of HDXsimulator for the same subsections down the line and so allow us to see what, if any, the affect high scoring lnPs had on eventual AUC values. We could therefore make more informed analyses based only on certain parts of the proteins, whereas we could only speak generally about the protein as a whole if we did not have

this high-resolution view.

6.5 Molecular Dynamics simulations & protein-protein docking

Before a library of docking poses with which to test HDXsimulator could be generated, we needed to carry out MD simulations on the bound crystal structures of the proteins utilised in this study. As there is little challenge in docking proteins that are already in the perfect orientation, MD was carried out in order to relax the protein's structures and so provide a realistic approximation of what docking unbound crystal structures or even, later down the line unbound *in silico* structures determined to be native by HDXsimulator, would be like.

We used CHARMM-GUI to prepare structures for MD and then ran equilibration and production runs using NAMD on the individual constituent proteins of each binary PPI as well as the complexes as a whole. Relaxation was qualitatively judged to have been successfully completed when the RMSD of subsequent frames plateaued off when compared to the first frame of the simulation. For most of the proteins this happened very quickly, within a few nanoseconds, however some, most notably the larger simulations like the bound complexes took substantially longer due to the greater number of atoms involved.

After relaxation, we began to carry out docking simulations using the webserver version of HADDOCK with AIRs generated using the results of our HDX experiments. Those residues which passed the 99 % CI threshold were marked as significant and defined as active restraints with the passive restraints automatically defined in order to guide the simulations toward the native complex structure. Due to this work taking place chronologically at the end of the project, we did not have time to complete all of the docking simulations but nevertheless we have learned much from those few we have carried out. Using a native cut off of $\leq$ 2.5 Å, the docking of BnWT:BspWT produced 10 native poses whereas the the docking of GFP:GFP-nb produced 0 native poses.

When looking at our initial HDX data sets for the BnWT:BspWT interaction, we were worried because almost every residue exceeded the 99 % CI threshold and so when applying this data to the AIRs of HADDOCK, it would be the equivalent of not having restraints at all (when everything is significant, nothing is). This highlights a key weakness of docking programs such as HADDOCK and PatchDock-/FiberDock which allow the marking of residues: it is a binary on/off switch that completely ignores the complex shape of the HDX data that gave rise to it. Under such a system, a residue that only just passed the CI threshold is given the exact same weight as a residue that exceeds the CI several times over, despite the latter clearly being of greater significance to the interaction than the former. With this problem in mind, we sought to ameliorate the issue by instead upping our CI threshold to extremely strict levels and therefore only include those residues which played the greatest part in the interaction in the AIRs. This idea was trialled using CI thresholds of 99.9 & 99.95 %. However we found that, contrary to what we predicted, these stricter CI thresholds actually made the docking simulations worse compared to the 99 % CI threshold, not better. We suspect that this may have been due to these higher thresholds omitting residues with relatively lower uptake difference that are nevertheless important in the binding interaction and therefore the program is struggling to correctly align the structures. This

would certainly explain why the stricter the CI threshold, the worse our results become. These results hint at a potential limitation of using HDX data to mark small proteins such as BnWT & BspWT. Because the interaction affects virtually the whole protein, we are effectively not using any restraints and so are mostly relying on the docking algorithm to correctly align the proteins based on intrinsic factors of the structures themselves. Therefore the efficacy of HDX data for use as AIRs in circumstances such as these should be treated with caution.

GFP:GFP-nb displayed interesting behaviour in as much that it while it produced no native poses at all, we think the reasons why can be explained by our previous observations of the BnWT:BspWT docking simulation. The fact that the only element causing non-native RMSD values was the rotation of GFP-nb about a single axis shows that when highly specific HDX data is used as AIRs, as is the case with GFP, a very accurate location for the binding partner can be found by HADDOCK. The problem lies in the fact that because the data for GFP-nb was not as specific, HADDOCK could not decide on the proper orientation and so it found generating native structures very difficult. It would have been beneficial to our analysis of this problem had one of our binary PPIs been between two larger proteins of comparable size to GFP with equally specific interaction profiles. We believe that such an interaction would have produced many more native poses than the GFP:GFP-nb interaction as such specific restraints would likely have been able to overcome the lack of shape complimentarity that appears to have so hindered our docking of GFP:GFP-nb. However, it should be noted that we did not have nearly as much time to optimise the the docking of GFP and GFP-nb compared to BnWT and BspWT, including replicate runs and so it is very possible that with more time we would have been able to produce a library of GFP:GFP-nb poses with a comparative number of native poses.

With the data from these two simulations in mind, we assert that HDX data has the potential to be very capable of guiding protein-protein docking simulations, however there are a few caveats that need to be taken into account. The capacity to accurately dock structures is diminished the greater a protein's surface is covered by the AIRs as each subsequent active/passive marked residue reduces the comparative weight of the ones that came before until, eventually, the AIRs become meaningless. In cases such as these, the docking program can only rely on steric characteristics such as protrusions and hollows which explains why BnWT:BspWT produced a good number of native poses (easily discernible shape complementarity), whereas GFP:GFP-nb did not (poor shape complementarity between the structures). Based on the evidence we have accumulated from the two distinct docking simulations attempted thus far, we propose that the best docking simulations will likely be gained from those interactions which have both specific HDX-generated AIRs and also show good shape complementarity, although of course more data sets will be required to confirm this. Future researchers using our methodology will therefore likely be able to predict the efficacy of attempts to generate native complex structures by analysing HDX profiles of the proteins in the bound configuration and comparing that to the unbound structures (either crystal or generated using our decoy method) to see if one or both of the conditions we set out for successful docking is met.

6.6 Exploring the boundaries of modelling protein conformation using HDXsimulator

HDXsimulator is a program developed by Ramin Salmas of the Borysik group for the purposes of calculating residue-level lnP values from three-dimensional structures. Based on the methods of Vendruscolo & Paci *et al.* and Best & Vendruscolo [29, 30], HDXsimulator accepts as input a number of three-dimensional models such as decoys and returns outputs of calculated lnP and RFU values for each structure in the decoy set. Calculated lnP is then compared with modelled lnP generated by HDXmodeller and calculated RFU values with experimentally determined RFUs in order to generate RMSE metrics for each model. For the purposes of testing, HDXsimulator also compares the three-dimensional coordinates of the models themselves with a (pseudo) native structure to generate an RMSD metric for each model. This allows for quantitative analysis via the construction of a ROC plot, which tests the programs classification efficacy i.e. whether a decoy is native or non-native.

We started by mapping the boundaries of modelling protein conformation by HDX-MS using HDXsimulator in order to determine its capabilities and limitations. This was done by comparing lnP values calculated using scaling factors (βH & βC) previously determined [30] to be optimal against lnPs that were deliberately error-laden in order to see the relationship between the AUC score and the RMSE/R^2 of the reference data vs. the error data. Our goal in doing this was to help us improve future versions of the program by having a clear and logical understanding of the factors that influence the relationship between these two factors. Our expectations were to see AUC values decrease as error between original and erroneous lnPs/RFUs increased, as represented by the RMSE/R^2 metric, however this proved more challenging than we had originally thought. In order to expedite the process of producing data sets, we developed a pipeline process in order to automate the majority of the data handling using a combination of Python and Bash scripts. This allowed us to greatly reduce the time taken to produce individual data sets as well as reducing the likelihood of user error causing invalid results.

Our concerns about efficiency proved well founded as we had to trial multiple different error generation methods until we found one which produced the expected relationship between AUC and RMSE/R^2. In addition to Gaussian distributions, we also attempted error introduction via shuffling of the true lnP values within the data set, initially within certain limits and then completely randomly, as well as using HDXsimulator itself to produce the errors. We also enabled the code to consider only certain parts of the protein data set as we had learned from HDXmodeller that different domains of the protein often behave differently. Therefore we reasoned that it would be beneficial for us to have a higher resolution view that we could eventually correspond to the same experimental domains we had identified as being important in our work with HDXmodeller.

Of these methods, only using HDXsimulator itself with suboptimal scaling factors produced the expected relationship between the values. We believe this method succeeded where the others failed because it was the only one that inherently altered the actual method of lnP calculation whereas the others simply modified true values that had already been calculated. By varying the scaling factors away from the values previously determined to be optimal, errors were produced that fundamentally considered the tertiary structure of the decoy which gave rise to them. As the tertiary structure of the

protein is primarily responsible for determining any given residues lnP, the errors we produced using this method can be considered an authentic depiction of the decoy's lnPs, only calculated using a sub-optimal expression. Therefore it is not surprising that such a method would be able to show a strong correlation between the AUC value and R^2 whereas other methods of error generation, which had no regard for the tertiary structure of the decoys, could not.

The results of using this method of error generation on different domains of the BspWT_Rosetta data set were interesting because they, like those of HDXmodeller, highlight how distinct tertiary structures can react differently to the same methodology. In our results, we could see that the central domains of BspWT_Rosetta have a comparatively tight correlation between the R^2 and AUC values, whereas at the termini, the correlation is noticeably weaker with a far greater range of potential AUC values that could be generated from any given R^2 value. The reason why one particular R^2 value can give rise to multiple different AUCs can be deduced when one considers what exactly R^2 represents: a single number that attempts to summarise how a whole protein's worth of synthetic lnPs compares to a set of reference values. Therefore we can see that there could be multiple different combinations of synthetic lnPs that all differ from each other quite substantially but nevertheless have the same R^2 vs. the reference data set. These disparate synthetic lnPs could therefore lead to very different AUC values as some of the structures they represent might be very close to the native while others may be quite dissimilar. This is a mathematical limitation of using a metric such as R^2 as it cannot take such differences into account and so reports them as a single value. With this in mind, it makes sense why we see this behaviour more at the termini compared to the central domains: the decoys have greater variation at the termini due to less steric clashes, hence the lnPs calculated across these domains will be more varied than those in the central regions where possible tertiary structure is more restricted.

6.7 Differentiating between native and non-native structures using HDXsimulator

6.7.1 Overview

With the our now greater understanding of the capabilities and limitations of HDXsimulator, we moved on to the most important part of this project: testing the program's ability to correctly classify structures from the various decoys sets prepared earlier as being native or non-native. We tested this classification ability for the 7 whole protein data sets on both the peptide RFU level (as we had done previously in [31]) as well as on the residue lnP level that had been enabled by the development of HDXmodeller. Calculated RFU/lnP was compared against experimental RFU/modelled lnP to generate an RMSE metric which was then compared against the RMSD of the decoys vs. the pseudo-crystal structure in order to generate an AUC score which described HDXsimulator's ability to accurately classify the structures. The development of modelled lnPs for comparison against calculated lnPs on the residue-level also enabled us to acquire a much higher resolution view of each protein data set than was possible before by assessing individual subsections independently of the rest of the protein. This is not possible to do with peptide-level data as any peptides bridging between the subsections can carry influence from those other subsections into the one you are looking at, making it impossible to determine which part of the protein is responsible for those RFU values. This is not a problem on the level of residue-resolved lnPs because on the scale of single amino-acids, no bridging between subsections can occur.

On the whole protein level, we found that our AUC results for RFU and lnP data were for the most part fairly congruent with each other for the same protein, which makes sense considering that that modelled lnP is derived from experimental RFU and calculated RFU is derived from calculated lnP. However AUC values varied dramatically between different proteins with some data sets such as BnWT_Rosetta having an almost perfect score of 0.9989 whereas others such as BspWT_Rosetta had completely the opposite score of 0.0071.
On the subsection level, we found that, similar to our results with HDXmodeller, there were often considerable variations in terms of AUC score within the different subsections of the protein data sets, with some subsections that were able to be accurately classified and others that were not.
When all our results were pooled together to obtain single AUC scores representing all the data sets on the level of whole protein RFU, whole protein lnP and subsection lnP, we found that these scores came to 0.5923, 0.5866 & 0.5554 respectively.

When taken all together, these results are indicative of a method that is still in relative infancy and as such there are many aspects that will require further investigation. When using a binary classification method such as ROC curves, AUC values typically range from 0.5, meaning no better than guessing, to 1.0, meaning that classification is always correct. However in our results, there are numerous instances of AUC values falling substantially below 0.5. This means that for these data sets, the correct classification is being actively selected against, for example in the case of results for BspWT_Rosetta where all the native poses are in fact being classified by the ROC as non-native. We believe this unusual result to be a consequence of the manner in which the Rosetta decoy sets were produced as it is only in these data sets that values so substantially below 0.5 are seen. In these decoy sets, the native structures are a single cluster with often times a large gap in terms of RMSD between them and the rest of the non-native decoys. Therefore these native decoys tend to move as a group and so if the calculated RFU/lnP tends to designate one of them as non-native according to the ROC curve, they are likely to all be considered non native, hence the extremely low AUC values exhibited by certain data sets. While some values below 0.5 are seen in decoy sets produced by 3DRobot, in which this clustering of the native decoys does not occur, it is no where near as prominent as that seen for Rosetta. Thus it is more likely to be simply the result of the imbalance between the small number of native decoys and the large number of non-native decoys and therefore the disproportionate effect of one incorrectly classified native decoy has on the AUC score compared to one incorrectly classified non-native decoy. This class imbalance affects the Rosetta decoy sets too of course which likely exacerbates the problems brought about by clustering.

In addition to comparing different scoring metrics at varying resolutions, we also wanted to see if there was any correlation between the R-matrix score, i.e confidence, of the modelled lnPs and the AUC value obtained for lnP-level data. Theoretically there should be a degree of agreement between these two metrics because if inaccurate (low R-matrix) lnPs are used as the "true" values against which calculated lnP for the native decoys are compared, the RMSE score between them should be quite high, leading to low AUC values. However, our previous work indicated that having highly accurate lnP data was not a prerequisite to obtaining high-scoring AUCs (hence our use of the minimalist equation of Vendruscolo & Paci *et al.*) and our work here further confirms this. We found that, on both the whole

protein level and the subsection level, there was almost no discernible correlation between high or low AUC values and high or low R-matrix values. For whole proteins, the R-matrix values were relatively consistent, ranging between 0.645-0.795, yet the AUC values obtained from these same data sets ranged from 0.0071-0.9989. Individual subsections presented a similar story, for example in the GFP_Rosetta data set, subsection F100-F130 had an R-matrix score of 0.588 and an AUC of 0.8488 whereas subsection K166-L207 had a relatively similar R-matrix score of 0.661 but a much lower AUC of 0.5613. This trend of disconnected R-matrix and AUC values can be seen in the majority of the data sets and, while the reason why is not particularly clear, we believe it might be due to partially inaccurate lnP modelling/calculation reporting compounding on top of each other to produce results that do not match. An auto-validation metric for HDXsimulator, similar to that for HDXmodeller, might help alleviate this problem.

A final point to consider when evaluating this data is the nature of the pseudo-crystal structures we used. For each decoy set generated, the structure with the lowest RMS compared to the crystal structure was taken and termed the pseudo-crystal structure for purposes of comparison. This was necessary because the numbering and exact order of the atoms in the decoys vs. the crystal structure was completely different, preventing their use with our RMSD calculating script and HDXsimulator. These differences are so extreme as to not be easily solvable by .pdb editing programs. Therefore our only recourse was to use one of the generated structures as a stand in for the real crystal structure for purposes of analysis, however the veracity of this technique is dependant on how close to the real crystal structure the pseudo-crystal structure is. We found that, at least for the Rosetta data sets, some pseudo-crystal structures were very close to the real crystal structure (e.g. BspWT) while others (e.g. GFP-nb) were substantially further away, as determined by their RMS score as reported by Abinitio. It is difficult to quantify what effect, if any, the variance of the pseudo-crystal structure from the real crystal structure had on our AUC results. For example, the BspWT pseudo structure was almost identical to the real crystal structure and produced terrible AUC results while the GFP-nb pseudo structure was quite different from the real crystal structure and produced very good AUC results. Regardless, having these structures differ is something to be avoided if at all possible as at best it complicates our analysis and at worst probably causes a certain amount of mis-classified decoys. This will no doubt be another problem to overcome in the continued development of HDXsimulator, one that could probably be ameloriated through increased *ab initio* sampling and/or data-driven decoy generation.

With the above observations in mind, we believe there are a number of areas of possible improvement which may help to improve HDXsimulator's ability to accurately classify decoy structures as being native or non-native. These are detailed in the next section.

6.7.2 Recommendations for improving the classification of structures

The first is that, as it currently stands, the methodology is highly decoy-dependent as can seen by the fact that the decoys produced by Rosetta and 3DRobot were not necessarily better or worse than each other, only different. This is evident when comparing the two proteins for which we had two decoy sets as, for BnWT, Rosetta produced superior AUC values while for GFP, 3DRobot produced superior AUC values. With only a sample size of two different methods, it is not possible to draw definitive

conclusions about what exactly makes for a good decoy set, however we suspect that a likely area of weakness in both these methods is their *ab initio* nature leading to a massive over-balance of non-native structures compared to native ones. Such discrepancies between the numbers of structures that fall into each class is likely having an effect on the AUC. Therefore a method which incorporates experimental data to guide structure generation, such as docking using HADDOCK (as will be discussed later), would produce proportionally more native structures and have less class imbalance and so improve the performance of the ROC curves.

Another point related to the generation of decoys is the number of structures actually produced for each protein data set. In our current methodology, approx. 1,000 decoys are evaluated for each data set, but this number is almost certainly too small to fully map the conformational landscape of any of the proteins we investigated. This number was chosen for our work for reasons of economy: we had many such decoy sets to produce and it was unknown at the time what an appropriate number for each would be, hence we settled on 1,000 as a compromise between production time and thoroughness. We predict that had a higher number of structures been chosen, for example 10,000 or even 100,000, we would see improved results as a greater proportion of the possible conformers would have been sampled. Such numbers would not have been practical for the work conducted here but may be possible for researchers investigating a single protein.

HDXsimulator itself is of course an area of possible improvement. The method used in this work to estimate protection factors from decoys is based on the same method proposed by Vendruscolo & Paci *et al.* in 2003 and as such makes use of a simple interpretation of the various factors that contribute to any given residue's protection. Such an equation is very useful for high throughput methods such as ours, however because only two factors, contacts and hydrogen bonds, are considered, it leaves open the possibility of other potentially important elements being ignored. With this in mind, we are in the exploratory stage of developing a new approach to estimate protection factors from structures based on thermodynamic terms. However, our preliminary experiments indicated that this method currently performs worse than that of Vendruscolo & Paci *et al.* in terms of the RMSE comparison to modelled lnPs (data not shown) and it is clear that a considerable amount of work will need to be done before we can start to seriously consider switching over to this new method.

The proteins themselves are another important factor when considering the efficacy of the technique as presented in this thesis. As can be seen from our results, different proteins (and within the same protein, different decoy sets) produce very different results from one another, some excellent e.g. BnWT_Rosetta, and some terrible e.g. BspWT_Rosetta, and our combined score is of course an amalgamation of all these results together. As we were only able to sample 7 data sets during the course of this work, one bad data set can severely affect the score, even though it may be an exception and not a rule. We could see this clearly when we removed the two worst data sets from the overall calculation (BspWT_Rosetta & GFP_Rosetta) and our final score of the whole protein lnP data set improved from 0.5866 to 0.6716 (data not shown). The solution to this problem is of course to increase the number of data sets analysed to the point where the true trend of AUC values can be seen and it is not influenced by a few erroneous data sets. This was not practical for us to do during this work as 10s if not 100s of

data sets from different proteins would be required; however our long-term plan is to turn data acquisition into a community project where researchers from all over the world can submit their HDX data. This will allow us to acquire many more data sets then could ever be obtained in-house and will likely be an integral step to improving the HDXmodeller/HDXsimulator methodologies.

Another source of possible problems which may affect the accuracy of HDXsimulator is the experimental HDX data itself. Such data is considered by this methodology to be "true" data and therefore any deviation away from it is considered to be error and thus the mark of an increasingly non-native structure. However, as with all experimental data regardless of method or application, this is not necessarily the case and while every effort was taken to try and ensure that the experimental HDX acquired for these proteins was as accurate as possible (replicates, extraneous exchange controls etc.), there is always the possibility that some parts of some data sets may in fact be "false" and this would of course affect the resultant AUC value/s for those proteins. Biological replicates and replicates run by other labs using different hardware would be a way of identifying if this had occurred to a substantial degree. However, from a practical perspective, it took 15+ weeks just to run one set of HDX experiments on all these proteins and, after the aforementioned years spent producing barnase/barstar, we simply did not have the required time to dedicate to producing biological replicates of every protein data set for comparison. Furthermore, because HDX's primary interest to the scientific community is as a difference method, corrected RFU values are rarely reported and so it was not possible to compare our data to literature values. All of this leaves the possibility that some of our experimental HDX results may have been partially mistaken, however within the bounds of this project, this was not possible to definitively prove or deny.

Finally, the way in which native/non-native decoys are classified could itself be altered. For example, 2.5 Å is commonly used as a cut off for native structures, hence why we used this value in our experiments, however this value could be moved depending on how rigid we wanted to be about the definition of the structures. For instance, if we did not need a strictly "native" structure, we could increase the RMSD cut off to something higher like 5 Å. This would give the ROC curve considerably more leeway in terms of class assignment and so result in a higher AUC, at the cost of structures potentially being less native than before. Another, more substantial change we could look into would be the method of classification itself. The current implementation uses ROC analysis, however other classification methods exist which might potentially be superior for our data sets. For example, Precision-Recall curves may present a better solution than ROC curves because they work better on imbalanced data sets such as ours. Precision-Recall curves summarize the trade-off between the recall or true positive rate (the same as ROC curves) and the precision or positive predictive value which is the fraction of positive results that are true positives as opposed to the false positive rate used by ROC curves. Precision-Recall curves tend to perform better than ROC curves on imbalanced data sets because of the use of true negatives in the ROC curve's false positive rate which is avoided by Precision-Recall curves.

Overall, we believe that the results presented here for HDXsimulator represent a positive first step upon which a reliable and effective method for predicting protein structure can be built. This method has not been in development long, with the first diagnostic data sets only being recently produced,

hence why the results are not quite up to the standard where native and non-native decoys can reliably discriminated. From the data sets produced in this work, we have identified numerous avenues for improvement that can be explored in the future which have the potential to produce a method that will be of great benefit to structural biologists and analytical chemists alike.

6.8 Summary & conclusions

The work undertaken in this thesis is extensive in its breadth and variety, incorporating a diverse range of different techniques spanning multiple scientific disciplines. From classical biochemical approaches such as the manipulation of DNA, cell culture and the production of proteins to analytical chemistry in the form of HDX and native mass spectrometry and even computational chemistry and computer science techniques such as MD, protein docking and Python programming. This thesis illustrates perfectly the modern trend of science as a multi-disciplinary undertaking, incorporating multiple different elements from across the spectrum of scientific understanding in order to answer ever more complicated questions.

Before the main project of structure classification began, we first worked on determining the location of binding between various different dUTPases and Stl for which there were no complex crystal structures available. Samples of three different dUTPases: human, φ11 & φNM1 as well as Stl were received from the Vértessy group and the location of binding determined by HDX-MS. This information was combined with data from other structural techniques in order to develop a model for the mechanism of inhibition of dUTPase by Stl. We found that in the case of the hDUT:Stl interaction, the complexation of hDUT to its substrate dUTP prevents Stl binding and inhibition, while the complexation of Stl to its substrate DNA does not prevent hDUT binding. This causes Stl dimer dissociation and the formation of $hDUT_3Stl_3$ as well as $hDUT_3Stl_2$ complexes. In the case of the interaction of timeric and dimeric dUTPases with Stl, we found that Stl's inhibitory plasticity is as a result of different binding surfaces interacting with timeric and dimeric dUTPases. Upon complexation, the timeric dUTPases form a $dUTPase_3Stl_3$ complex and the dimeric dUTPases form a $dUTPase_1Stl_1$ complex.

The main question we set out to answer with the work undertaken in this thesis was: can an analytical technique such as HDX-MS be combined with computational chemistry in order to allow the deduction of native structures *ex nihilo*? Initial work carried out [4] set the stage by proving that peptide-level RFU data calculated from three-dimensional coordinates using the methodology of Vendruscolo & Paci*et al.* could be successful in selecting for native structures from a background of non-native ones. This was expanded upon in [31] where we quantified this ability to discriminate between native and non-native structures using calculated vs. experimental RFU data by using a binary classification system in the form of ROC curves. These initial works identified a number of areas in which the technique could be improved, and these became the main thrust of the work carried out in this PhD. Our first goal was to expand the list of experimental data sets available for analysis by acquiring HDX data on more protein systems. Binary PPIs were chosen for this because they would allow collected data sets to be used both for the individual proteins themselves as well as for the complexes. We leveraged past experience with protein production in order to produce two of the proteins, barnase and to a certain extent barstar, in-house and contracted out to specialists to produce the rest. With proteins in hand,

we carried out extensive characterisation of each system by HDX-MS, looking not only at difference data but also at absolute RFU values corrected for extraneous exchange.

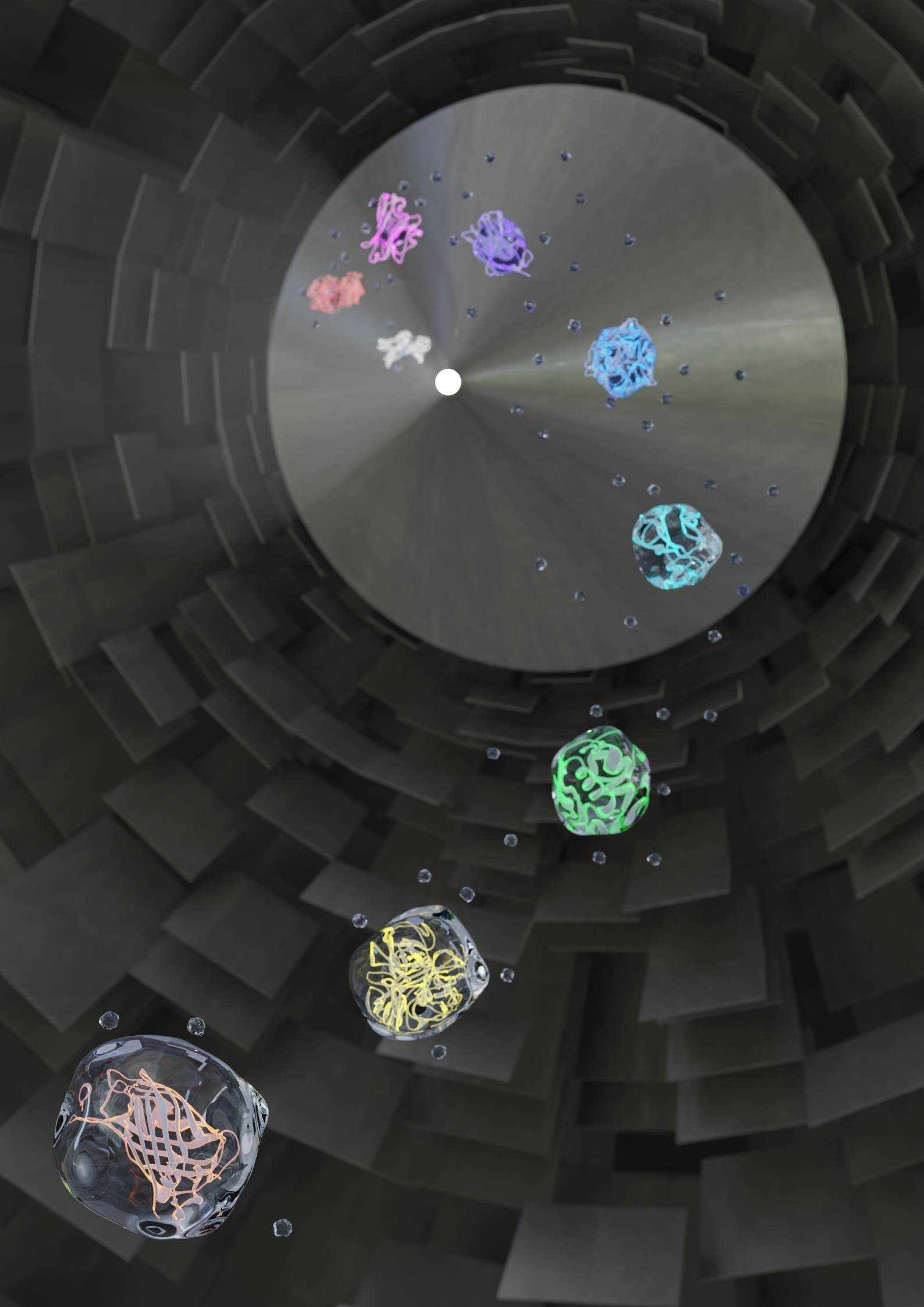
Simultaneously to this acquisition of data, Ramin Salmas developed HDXmodeller to enable the calculation of residue-level lnP data from peptide-level RFU outputs and HDXsimulator to take advantage of this modelled lnP data and allow structures to be discriminated using residue-level lnP data in addition to peptide-level RFU data. This development also enabled higher resolution data acquisition by allowing protein data sets to be broken down into subsections of any length. An extensive process to map the boundaries of modelling protein conformation using HDXsimulator was then undertaken in order determine its capabilities and limitations, followed by production runs of the individual protein's data sets in both the original whole protein RFU form as well as the new whole protein lnP and subsection lnP forms. When data sets were combined, we obtained decent results as determined by ROC curves of approx. 0.6, meaning that our current methodology can classify protein structures with 60 % accuracy. However importantly, this figure takes into account several very bad data sets and if these are excluded, our accuracy increases to approx. 70 %. Using these results, we were able to suggest several ways in which the methodology could be improved so that future classification accuracy could approach the level needed for users to be confident in its results. Concurrently, we started to develop the method to work not only on single proteins but also on binary PPIs. We began by conducting MD experiments on the crystal structures of the binary complexes in order to relax them and so not bias subsequent docking steps. With this completed, protein-protein docking was carried out on two of the four binary complexes we would be studying, at which point work on this PhD ceased.

In conclusion, the work conducted in this thesis has laid the groundwork for the eventual development of an effective technique for determining protein structures from HDX-MS outputs. While the AUC results for our initial combined data set were lower than we had hoped for, this is not altogether surprising considering this is a brand-new project that we have had to develop from the ground up. These results are therefore encouraging and leave plenty of opportunity for subsequent development by others to progress the technique to the point where protein structures can be reliably classified into native and non-native by their HDX data alone. Furthermore, we have begun the developmental steps that will be required to enable this methodology to discriminate between native and non-native complexes as well; further expanding its use cases to include the extremely important class of binary PPIs. We believe therefore that we have successfully answered the primary research question of this PhD project, that an analytical technique such as HDX-MS can indeed be leveraged to determine native protein structures. We anticipate that further improvement of the methodology will be a community endeavour with data contributions from HDX groups all over the world and that, when sufficiently developed, this technique will be an invaluable tool for scientists studying proteins for which no experimental structures exist.

6.9 Future work

As a methodology near the start of its development, there are numerous paths which could be taken to improve upon what is presented here. Most of these are discussed at length in chapter 6.7.2 and involve the procurement of additional decoy sets from different methods, increasing the number

of decoys per data set, modifying HDXsimulator so it can predict lnPs with greater accuracy, acquiring and generating experimental data for additional proteins, producing replicates of HDX data to ensure validity and, finally, modifying the method by which class assignment is determined for potentially better accuracy. Beyond making improvements to HDXsimulator, which may take an indeterminate amount of time, there are also other areas that could be worked on, such as continuing the development of the binary PPI branch of this thesis so that it can interface with HDXsimulator. There is also a lot of work to be done from a community relations point of view. If we are to acquire the 10s or 100s of data sets we will need to properly validate the methodology, we will need to make the use of the tools as simple as possible in order to encourage others to use them. Related to this, we will need to increase community awareness of our tools, not just through papers but also through conferences.



References

[1] J C Kendrew, G Bodo, H M Dintzis, R G Parrish, H Wyckoff, and D C Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181:662–666, 1958.

[2] H Deng, Y Jia, and Y Zhang. Protein structure prediction. *International Journal of Modern Physics B*, 32(18), 2018.

[3] Koyel Mitra, Anne Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732, 2013.

[4] Antoni J. Borysik. Simulated Isotope Exchange Patterns Enable Protein Structure Determination. *Angewandte Chemie - International Edition*, 56(32):9396–9399, 2017.

[5] Eugen Goldstein. Ueber eine noch nicht untersuchte Strahlungsform an der Kathode inducirter Entladungen. *Annalen der Physik*, 300(1):38–48, 1898.

[6] J. J. Thomson. Bakerian Lecture:—Rays of Positive Electricity. *Proceedings of the Royal Society A*, 89(607), 1913.

[7] Richard Herzog. Ionen- und elektronenoptische Zylinderlinsen und Prismen. I. *Zeitschrift für Physik*, 89:447–473, 1934.

[8] J. Mattauch and R. Herzog. Über einen neuen Massenspektrographen. *Zeitschrift für Physik*, 89:786–795, 1934.

[9] Jeffery Dempster. New Methods in Mass Spectroscopy. *Proceedings of the American Philosophical Society*, 75(8):755–767, 1935.

[10] Kenneth T. Bainbridge and Edward B. Jordan. Mass spectrum analysis 1. the mass spectrograph. 2. the existence of isobars of adjacent elements. *Physical Review*, 50(4):282–296, 1936.

[11] M. M. Wolff and W. E. Stephens. A pulsed mass spectrometer with time dispersion. *Review of Scientific Instruments*, 24(8):616–617, 1953.

[12] B. Mamyrin, V. Karataev, D. Shmikk, and V. Zagulin. The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution. *Soviet Journal of Experimental and Theoretical Physics*, 64:82–89, 1973.

[13] Malcolm Dole, L. L. Mack, R. L. Hines, R. C. Mobley, L. D. Ferguson, and M. B. Alice. Molecular beams of macroions. *The Journal of Chemical Physics*, 49(5):2240–2249, 1968.

[14] Masamichi Yamashita and John B. Fenn. Electrospray ion source. Another variation on the free-jet theme. *Journal of Physical Chemistry*, 88(20):4451–4459, 1984.

[15] John B Fenn, Matrhias Mann, Chin Kai Meng, Shek Fu Wong, and Craig M Whitehouse. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science*, 246(4926):64–71, 1989.

[16] C. S. Ho, C. W. K. Lam, M. H. M. Chan, R. C. K. Cheung, L. K. Law, L. C. W. Lit, K. F. Ng, M. W. M. Suen, and H. L. Tai. Electrospray Ionisation Mass Spectrometry: Principles and Clinical Applications. *Clinical Biochemist Reviews*, 24(1):3–12, 2003.

[17] James J Pitt. Principles and Applications of Liquid Chromatography- Mass Spectrometry in Clinical Biochemistry. *Clinical Biochemist Reviews*, 30:19–34, 2009.

[18] Viswanatham Katta and Brian T. Chait. Conformational Changes in Proteins Probed by Hydrogen-exchange Electrospray-ionization Mass Spectrometry. *Rapid Communications in Mass Spectrometry*, 5:214–217, 1991.

[19] Aase Hvidt and K. Linderstrom-Lang. Exchange of hydrogen atoms in insulin with deuterium atoms in aqueous solutions. *Biochimica et Biophysica Acta*, 14(4):574–575, 1954.

[20] S. J. Leach and Julie Hill. Studies on Ribonuclease Conformation and Racemization Using Tritium-Hydrogen Exchange and Optical Rotatory Dispersion. *Biochemistry*, 2(4):807–813, 1963.

[21] David Weis. *Hydrogen Exchange Mass Spectrometry of Proteins: Fundamentals, Methods, and Applications*. 2016.

[22] David Nguyen, Leland Mayne, Michael C. Phillips, and S. Walter Englander. Reference Parameters for Protein Hydrogen Exchange Rates. *Journal of the American Society for Mass Spectrometry*, 29(9):1936–1939, 2018.

[23] Yawen Bai, John S Milne, Leland Mayne, and S Walter Englander. Primary Structure Effects on Peptide Group Hydrogen Exchange. *Proteins*, 17(1):75–86, 1993.

[24] Walter Englander, Nancy W Downer, and Harry Teitelbaum. Hydrogen Exchange. *Annual Review of Biochemistry*, 41:903–924, 1972.

[25] Jane Clarke, Laura S. Itzhaki, and Alan R. Fersht. Hydrogen exchange at equilibrium: A short cut for analysing pathways? *Trends in Biochemical Sciences*, 22(8):284–287, 1997.

[26] Lars Konermann, Xin Tong, and Yan Pan. Protein structure and dynamics studied by mass spectrometry: H/D exchange, hydroxyl radical labeling, and related approaches. *Journal of mass spectrometry*, 43:1021–1036, 2008.

[27] Zhongqi Zhang and David L Smith. Determination of amide hydrogen exchange by mass spectrometry: a new tool for protein structure elucidation. *Protein Science*, 2(4):522–531, 1993.

[28] Yoshitomo Hamuro, Stephen J. Coales, Kathleen S. Molnar, Steven J. Tuske, and Jeffery A. Morrow. Specificity of immobilized porcine pepsin in H/D exchange compatible conditions. *Rapid Communications in Mass Spectrometry*, 22:1041–1046, 2008.

[29] Michele Vendruscolo, Emanuele Paci, Christopher M. Dobson, and Martin Karplus. Rare Fluctuations of Native Proteins Sampled by Equilibrium Hydrogen Exchange. *Journal of the American Chemical Society*, 125(51):15686–15687, 2003.

[30] Robert B. Best and Michele Vendruscolo. Structural interpretation of hydrogen exchange protection factors in proteins: characterization of the native state fluctuations of CI2. *Structure*, 14(1):97–106, 2006.

[31] Matthew J. Harris, Deepika Raghavan, and Antoni J. Borysik. Quantitative Evaluation of Native Protein Folds and Assemblies by Hydrogen Deuterium Exchange Mass Spectrometry (HDX-MS). *Journal of The American Society for Mass Spectrometry*, 30(1):58–66, 2018.

[32] Ramin Ekhteiari Salmas and Antoni James Borysik. Hdxmodeller an online webserver for high-resolution hdx-ms with auto-validation. *Nature Communications Biology*, 4, 2021.

[33] Ramin Ekhteiari Salmas and Antoni James Borysik. Characterization and management of noise in hdx-ms data modeling. *Analytical Chemistry*, 93:7323–7331, 5 2021.

[34] Jayawant N. Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, 2010.

[35] Kim T. Simons, Charles Kooperberg, Enoch Huang, and David Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, 268(1):209–225, 1997.

[36] Kim T. Simons, Ingo Ruczinski, Charles Kooperberg, Brian A. Fox, Chris Bystroff, and David Baker. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins: Structure, Function and Genetics*, 34(1):82–95, 1999.

[37] Richard Bonneau, Jerry Tsai, Ingo Ruczinski, Dylan Chivian, Carol Rohl, Charlie E.M. Strauss, and David Baker. Rosetta in CASP4: Progress in ab initio protein structure prediction. *Proteins: Structure, Function and Genetics*, 45(SUPPL. 5):119–126, 2001.

[38] Richard Bonneau, Charlie E.M. Strauss, Carol A. Rohl, Dylan Chivian, Phillip Bradley, Lars Malmström, Tim Robertson, and David Baker. De novo prediction of three-dimensional structures for major protein families. *Journal of Molecular Biology*, 322(1):65–78, 2002.

[39] Philip Bradley, Kira M.S. Misura, and David Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871, 2005.

[40] Srivatsan Raman, Robert Vernon, James Thompson, Michael Tyka, Ruslan Sadreyev, Jimin Pei, David Kim, Elizabeth Kellogg, Frank DiMaio, Oliver Lange, Lisa Kinch, Will Sheffler, Bong-hyun Kim, Rhiju Das, Nick V Grishin, and David Baker. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins*, 77(9):89–99, 2009.

[41] Yifan Song, Frank DiMaio, Ray Yu-Ruei Wang, David Kim, Chris Miles, TJ Brunette, James Thompson, and David Baker. High resolution comparative modeling with RosettaCM. *Structure*, 21(10):1735–1742, 2013.

[42] Haiyou Deng, Ya Jia, and Yang Zhang. 3DRobot: Automated generation of diverse and well-packed protein structure decoys. *Bioinformatics*, 32(3):378–387, 2016.

[43] Martin Karplus and J. Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9(9):646–652, 2002.

[44] Scott A. Hollingsworth and Ron O. Dror. Molecular Dynamics Simulation for All. *Neuron*, 99(6):1129–1143, 2018.

[45] B. J. Alder and T. E. Wainwright. Phase transition for a hard sphere system. *The Journal of Chemical Physics*, 27(5):1208–1209, 1957.

[46] J Andrew McCammon, Bruce R Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267:585–590, 1977.

[47] Kresten Lindorff-Larsen, Paul Maragakis, Stefano Piana, Michael P. Eastwood, Ron O. Dror, and David E. Shaw. Systematic validation of protein force fields against experimental data. *PLoS ONE*, 7(2), 2012.

[48] Sunhwan Jo, Taehoon Kim, Vidy Ashankara G. Iyer, and Wonpil Im. CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM. *Journal of Computational Chemistry*, 29(11):1859–1865, 2008.

[49] Jumin Lee, Xi Cheng, Jason M. Swails, Min Sun Yeom, Peter K. Eastman, Justin A. Lemkul, Shuai Wei, Joshua Buckner, Jong Cheol Jeong, Yifei Qi, Sunhwan Jo, Vijay S. Pande, David A. Case, Charles L. Brooks, Alexander D. MacKerell, Jeffery B. Klauda, and Wonpil Im. CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *Journal of Chemical Theory and Computation*, 12(1):405–413, 2016.

[50] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[51] Ilya A. Vakser. Protein-protein docking: From interaction to interactome. *Biophysical Journal*, 107(8):1785–1793, 2014.

[52] Dina Schneidman-Duhovny, Yuval Inbar, Ruth Nussinov, and Haim J. Wolfson. PatchDock and SymmDock: Servers for rigid and symmetric docking. *Nucleic Acids Research*, 33:363–367, 2005.

[53] Efrat Mashiach, Ruth Nussinov, and Haim J. Wolfson. FiberDock: Flexible induced-fit backbone refinement in molecular docking. *Proteins*, 78(6):1503–1519, 2010.

[54] Efrat Mashiach, Ruth Nussinov, and Haim J Wolfson. FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking. *Nucleic Acids Research*, 38:457–461, 2010.

[55] Jeffrey J. Gray, Stewart Moughon, Chu Wang, Ora Schueler-Furman, Brian Kuhlman, Carol A. Rohl, and David Baker. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology*, 331(1):281–299, 2003.

[56] Sergey Lyskov and Jeffrey J. Gray. The RosettaDock server for local protein-protein docking. *Nucleic acids research*, 36:233–238, 2008.

[57] Cyril Dominguez, Rolf Boelens, and Alexandre M.J.J. Bonvin. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125:1731–1737, 2003.

[58] G. C.P. van Zundert, J. P.G.L.M. Rodrigues, M. Trellet, C. Schmitz, P. L. Kastritis, E. Karaca, A. S.J. Melquiond, M. Van Dijk, S. J. De Vries, and A. M.J.J. Bonvin. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology*, 428:720–725, 2016.

[59] Joël Janin, Kim Henrick, John Moult, Lynn Ten Eyck, Michael J.E. Sternberg, Sandor Vajda, Ilya Vakser, and Shoshana J. Wodak. CAPRI: A critical assessment of PRedicted interactions. *Proteins: Structure, Function and Genetics*, 52(1):2–9, 2003.

[60] Robert W. Hartley. Barnase and barstar: two small proteins to fold and fit together. *Trends in Biochemical Sciences*, 14(11):450–454, 1989.

[61] G Schreiber and A R Fersht. Energetics of protein protein interactions-analysis of the barnase barstar interface by single mutations and double mutant cycles. *Journal of Molecular Biology*, 248(2):478–486, 1995.

[62] R. W. Hartley and J. R. Smeaton. On the reaction between the extracellular ribonuclease of Bacillus amyloliquefaciens (Barnase) and its intracellular inhibitor (Barstar). *Journal of Biological Chemistry*, 248(16):5624–5626, 1973.

[63] Rafael Yuste. Fluorescence microscopy today. *Nature Methods*, 2(12):902–904, 2005.

[64] K. H.S. Arun, C. L. Kaul, and P. Ramarao. Green fluorescent proteins in receptor research: An emerging tool for drug discovery. *Journal of Pharmacological and Toxicological Methods*, 51:1–23, 2005.

[65] Osamu Shimomura, Frank H. Johnson, and Yo Saiga. Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, Aequorea. *Journal of Cellular and Comparative Physiology*, 59:223–239, 1962.

[66] Marta H Kubala, Oleksiy Kovtun, Kirill Alexandrov, and Brett M Collins. Structural and thermodynamic analysis of the GFP : GFP-nanobody complex. *Protein Science*, 19:2389–2401, 2010.

[67] S. Muyldermans, T. N. Baral, V. Cortez Retamozzo, P. De Baetselier, E. De Genst, J. Kinne, H. Leonhardt, S. Magez, V. K. Nguyen, H. Revets, U. Rothbauer, B. Stijlemans, S. Tillib, U. Wernery, L. Wyns, Gh Hassanzadeh-Ghassabeh, and D. Saerens. Camelid immunoglobulins and nanobody technology. *Veterinary Immunology and Immunopathology*, 128:178–183, 2009.

[68] Axel Kirchhofer, Jonas Helma, Katrin Schmidthals, Carina Frauer, Sheng Cui, Annette Karcher, Mireille Pellis, Serge Muyldermans, Corella S. Casas-Delucchi, M. Cristina Cardoso, Heinrich Leonhardt, Karl Peter Hopfner, and Ulrich Rothbauer. Modulation of protein properties in living cells using nanobodies. *Nature Structural and Molecular Biology*, 17(1):133–139, 2010.

[69] Tomas Lindahl. Instability and decay of the primary structure of DNA. *Nature*, 362(6422):709–715, 1993.

[70] B G Vértessy and J Toth. Keeping uracil out of DNA: physiological role, structure and catalytic mechanism of dUTPases. *Accounts of Chemical Research*, 42(1):97–106, 2009.

[71] Ildiko Pecsi, Rita Hirmondo, Amanda C. Brown, Anna Lopata, Tanya Parish, Beata G. Vertessy, and Judit Tóth. The dutpase enzyme is essential in Mycobacterium smegmatis. *PLoS ONE*, 7(5), 2012.

[72] Carles Úbeda, Elisa Maiques, Peter Barry, Avery Matthews, María Ángeles Tormo, Íñigo Lasa, Richard P. Novick, and José R. Penadés. SaPI mutations affecting replication and transfer and enabling autonomous replication in the absence of helper phage. *Molecular Microbiology*, 67(3):493–503, 2008.

[73] Rosanne L.L. Hill and Terje Dockland. The type 2 dUTPase of bacteriophage NM1 initiates mobilization of Staphylococcus aureus bovine pathogenicity island 1. *Journal of Molecular Biology*, 428(1):142–152, 2016.

[74] Judit E. Szabó, Veronika Németh, Veronika Papp-Kádár, Kinga Nyíri, Ibolya Leveles, Ábris Bendes, Imre Zagyva, Gergely Róna, Hajnalka L. Pálinkás, Balázs Besztercei, Olivér Ozohanics, Károly Vékey, Károly Liliom, Judit Tóth, and Beáta G. Vértessy. Highly potent dUTPase inhibition by a bacterial repressor protein reveals a novel mechanism for gene expression control. *Nucleic Acids Research*, 42(19):11912–11920, 2014.

[75] Rita Hirmondó, Judit E. Szabó, Kinga Nyíri, Szilvia Tarjányi, Paula Dobrotka, Judit Tóth, and Beáta G. Vértessy. Cross-species inhibition of dUTPase via the Staphylococcal Stl protein perturbs dNTP pool and colony formation in Mycobacterium. *DNA Repair*, 30:21–27, 2015.

[76] András Benedek, István Pölöskei, Olivér Ozohanics, Károly Vékey, and Beáta G. Vértessy. The Stl repressor from Staphylococcus aureus is an efficient inhibitor of the eukaryotic fruitfly dUTPase. *FEBS Open Bio*, 8(2):158–167, 2017.

[77] Kinga Nyíri, Matthew J. Harris, Judit Matejka, Olivér Ozohanics, Károly Vékey, Antoni J. Borysik, and Beáta G. Vértessy. HDX and Native Mass Spectrometry Reveals the Different Structural Basis for Interaction of the Staphylococcal Pathogenicity Island Repressor Stl with Dimeric and Trimeric Phage dUTPases. *Biomolecules*, 9(488), 2019.

[78] Kinga Nyíri, Veronika Papp-Kádár, Judit E. Szabó, Veronika Németh, and Beáta G. Vértessy. Exploring the role of the phage-specific insert of bacteriophage Φ11 dUTPase. *Structural Chemistry*, 26:1425–1432, 2015.

[79] Kinga Nyíri, Haydyn D. T. Mertens, Borbála Tihanyi, Gergely N. Nagy, Bianka Kőhegyi, Judit Matejka, Matthew J. Harris, Judit E. Szabó, Veronika Papp-Kádár, Veronika Németh-Pongrácz, Olivér Ozohanics, Károly Vékey, Dmitri I. Svergun, Antoni J. Borysik, and Beáta G. Vértessy. Structural model of human dUTPase in complex with a novel proteinaceous inhibitor. *Scientific Reports*, 8(4326), 2018.

[80] Kaku Saito, Hirotaka Nagashima, Kazuharu Noguchi, Kunihiro Yoshisue, Tatsushi Yokogawa, Eiji Matsushima, Takeshi Tahara, and Shigeru Takagi. First-in-human, phase I dose-escalation study

of single and multiple doses of a first-in-class enhancer of fluoropyrimidines, a dUTPase inhibitor (TAS-114) in healthy male volunteers. *Cancer Chemotherapy and Pharmacology*, 73(3):577–583, 2014.

[81] Robert Winkler. ESIprot: a universal tool for charge state determination and molecular weight calculation of proteins from electrospray ionization mass spectrometry data. *Rapid Communications in Mass Spectrometry*, 24(3):285–294, 2010.

[82] Helen M. Berman, John D. Westbrook, Zukang Feng, Gary L. Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[83] Matthew P. Jacobson, Richard A. Friesner, Zhexin Xiang, and Barry Honig. On the role of the crystal environment in determining protein side-chain conformations. *Journal of Molecular Biology*, 320(3):597–608, 2002.

[84] Matthew P. Jacobson, David L. Pincus, Chaya S. Rapp, Tyler J.F. Day, Barry Honig, David E. Shaw, and Richard A. Friesner. A Hierarchical Approach to All-Atom Protein Loop Prediction. *PROTEINS: Structure, Function, and Bioinformatics*, 55(2):351–367, 2004.

[85] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[86] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, A32(5):922, 1976.

[87] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, 2004.

[88] Ildikó Pécsi, Judit E. Szabó, Scott D. Adams, István Simon, James R. Sellers, Beáta G. Vértessy, and Judit Tóth. Nucleotide pyrophosphatase employs a P-loop-like motif to enhance catalytic power and NDP/NTP discrimination. *Proceedings of the National Academy of Sciences of the United States of America*, 108(35):14437–14442, 2011.

[89] Gergely N. Nagy, Reynier Suardíaz, Anna Lopata, Olivér Ozohanics, Károly Vékey, Bernard R. Brooks, Ibolya Leveles, Judit Tóth, Beata G. Vértessy, and Edina Rosta. Structural Characterization of Arginine Fingers: Identification of an Arginine Finger for the Pyrophosphatase dUTPases. *Journal of the American Chemical Society*, 138(45):15035–15045, 2016.

[90] Ildiko Pecsi, Ibolya Leveles, Veronika Harmat, Beata G. Vertessy, and Judit Toth. Aromatic stacking between nucleobase and enzyme promotes phosphate ester hydrolysis in dUTPase. *Nucleic Acids Research*, 38(20):7179–7186, 2010.

[91] Janine Bowring, Maan M Neamah, Jorge Donderis, Ignacio Mir-Sanchis, Christian Alite, J Rafael Ciges-Tomas, Elisa Maiques, Iltyar Medmedov, Alberto Marina, and José R Penadé. Pirating conserved phage mechanisms promotes promiscuous staphylococcal pathogenicity island transfer. *eLife*, 6:e26487, 2017.

[92] Luis Serrano, Andreas Matouschek, and Alan R Fersht. The folding of an enzyme iii. structure of the transition state for unfolding of barnase analysed by a protein engineering procedure. *Journal of Molecular Biology*, 224:805–818, 1992.

[93] Martine Prevost, Shoshana J Wodak, Bruce Tidor, and Martin Karplus. Contribution of the hydrophobic effect to protein stability: Analysis based on simulations of the ile-96-ala mutation in barnase. *Proceedings of the National Academy of Sciences of the United States of America*, 88:10880–10884, 1991.

[94] C Nick Pace, Douglas V Laurents, and Rick E Erickson. Urea denaturation of barnase: ph dependence and characterization of the unfolded state. *Proceedings of the National Academy of Sciences of the United States of America*, 31:2728–2734, 1992.

[95] Gideon Schreiber and Alan R Fersht. The refolding of cis-and trans-peptidylprolyl isomers of barstar. *Biochemistry*, 32:11195–11203, 1993.

[96] M C R Shastry, Vishwas R Agashe, and Jayant B Udgaonkar. Quantitative analysis of the kinetics of denaturation and renaturation of barstar in the folding transition zone. *Protein Science*, 3:1409–1417, 1994.

[97] V Guillet, A Lapthornl, R W Hartley, and Y Mauguen. Recognition between a bacterial ribonuclease, barnase, and its natural inhibitor, barstar. *Structure*, 1:165–176, 1993.

[98] Chloe Martens, Mrinal Shekhar, Antoni J. Borysik, Andy M. Lau, Eamonn Reading, Emad Tajkhorshid, Paula J. Booth, and Argyris Politis. Direct protein-lipid interactions shape the conformational landscape of secondary transporters. *Nature Communications*, 9(4151), 2018.

[99] Ernst Althaus, Stefan Canzar, Carsten Ehrler, Mark R Emmett, Andreas Karrenbauer, Alan G Marshall, Anke Meyer-Bäse, Jeremiah D Tipton, and Hui-Min Zhang. Computing h/d-exchange rates of single residues from data of proteolytic fragments. *BMC Bioinfornatics*, 11, 2010.

[100] Zhongqi Zhang, Aming Zhang, and Gang Xiao. Improved protein hydrogen/deuterium exchange mass spectrometry platform with fully automated data processing. *Analytical Chemistry*, 84:4942–4949, 6 2012.

[101] Daniel J. Saltzberg, Howard B. Broughton, Riccardo Pellarin, Michael J. Chalmers, Alfonso Espada, Jeffrey A. Dodge, Bruce D. Pascal, Patrick R. Griffin, Christine Humblet, and Andrej Sali. A residue-resolved bayesian approach to quantitative interpretation of hydrogen-deuterium exchange from mass spectrometry: Application to characterizing protein-ligand interactions. *Journal of Physical Chemistry B*, 121:3493–3501, 4 2017.

[102] Chris Gessner, Wieland Steinchen, Sabrina Bédard, John J. Skinner, Virgil L. Woods, Thomas J. Walsh, Gert Bange, and DIonysios P. Pantazatos. Computational method allowing hydrogen-deuterium exchange mass spectrometry at single amide resolution. *Scientific Reports*, 7, 12 2017.

[103] Simon P. Skinner, Gael Radou, Roman Tuma, Jeanine J. Houwing-Duistermaat, and Emanuele Paci. Estimating constraints for protection factors from hdx-ms data. *Biophysical Journal*, 116:1194–1203, 4 2019.

[104] Benjamin T Walters. Empirical Method To Accurately Determine Peptide-Averaged Protection Factors from Hydrogen Exchange MS Data. *Analytical Chemistry*, 89(2):1049–1053, 2017.

[105] Zhong Yuan Kan, Benjamin T. Walters, Leland Mayne, and S. Walter Englander. Protein hydrogen exchange at residue resolution by proteolytic fragmentation mass spectrometry analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 110:16438–16443, 10 2013.

[106] Darko Babić, Saša Kazazić, and David M. Smith. Resolution of protein hydrogen/deuterium exchange by fitting amide exchange probabilities to the peptide isotopic envelopes. *Rapid Communications in Mass Spectrometry*, 33:1248–1257, 8 2019.

[107] Zhongqi Zhang. Complete extraction of protein dynamics information in hydrogen/deuterium exchange mass spectrometry data. *Analytical Chemistry*, 92:6486–6494, 5 2020.

[108] Jürgen Claesen and Argyris Politis. Poppet: a new method to predict the protection factor of backbone amide hydrogens. *Journal of the American Society for Mass Spectrometry*, 30:67–76, 1 2018.

[109] Filip Persson and Bertil Halle. How amide hydrogens exchange in native proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 112:10383–10388, 8 2015.

[110] Vincent J Hilser and Ernesto Freire. Structure-based calculation of the equilibrium folding pathway of proteins. correlation with hydrogen exchange protection factors. *Journal of Molecular Biology*, 262:756–772, 1996.

[111] Hossein Mohammadiarani, Vincent S. Shaw, Richard R. Neubig, and Harish Vashisth. Interpreting hydrogen-deuterium exchange events in proteins using atomistic simulations: Case studies on regulators of g-protein signaling proteins. *Journal of Physical Chemistry B*, 122:9314–9323, 10 2018.

[112] Robert G Mcallister and Lars Konermann. Challenges in the Interpretation of Protein H/D Exchange Data: A Molecular Dynamics Simulation Perspective. *Biochemistry*, 54(16):2683–2692, 2015.

Appendices

A *E. coli* Competent Cells protocol by Promega

Single-Use Competent Cells Standard Transformation Protocol

1. Remove competent cells from −70 °C and place on ice for 5 minutes or until just thawed.
2. Add 1–50 ng of DNA (in a volume not greater than 5 µl) to the Competent Cells. Move the pipette tip through the cells while dispensing. Quickly flick the tube several times. Do not vortex.
3. Immediately return the tubes to ice for 5–30 minutes.
4. Heat-shock cells for 15–20 seconds in a water bath at exactly 42 °C. Do not shake.
5. Immediately place the tubes on ice for 2 minutes.
6. Add 450 µl of room-temperature SOC medium to each transformation reaction, and incubate for 60 minutes at 37 °C with shaking (approximately 225 rpm). For best transformation efficiency, lay the tubes on their sides and tape them to the platform.
7. For each transformation reaction, we recommend plating 100 µl of undiluted cells and 1:10 and 1:100 cell dilutions on antibiotic plates. Incubate the plates at 37 °C overnight.

B Miniprep protocol by Qiagen

Protocol: Plasmid DNA Purification using the QIAprep Spin Miniprep Kit and a Microcentrifuge

1. Resuspend pelleted bacterial cells in 250 µl Buffer P1 and transfer to a microcentrifuge tube. Ensure that RNase A has been added to Buffer P1. No cell clumps should be visible after resuspension of the pellet. If LyseBlue reagent has been added to Buffer P1, vigorously shake the buffer bottle to ensure LyseBlue particles are completely dissolved. The bacteria should be resuspended completely by vortexing or pipetting up and down until no cell clumps remain.
2. Add 250 µl Buffer P2 and mix thoroughly by inverting the tube 4–6 times. Mix gently by inverting the tube. Do not vortex, because this will result in shearing of genomic DNA. If necessary, continue inverting the tube until the solution becomes viscous and slightly clear. Do not allow the lysis reaction to proceed for more than 5 min. If LyseBlue has been added to Buffer P1, the cell suspension will turn blue after addition of Buffer P2. Mixing should result in a homogeneously colored suspension. If the suspension contains localized colorless regions, or if brownish cell clumps are still visible, continue mixing the solution until a homogeneously colored suspension is achieved.
3. Add 350 µl Buffer N3. Mix immediately and thoroughly by inverting the tube 4–6 times. To avoid localized precipitation, mix the solution thoroughly, immediately after addition of Buffer N3. Large culture volumes (e.g., $\geq$5 ml) may require inverting up to 10 times. The solution should become cloudy. If LyseBlue reagent has been used, the suspension should be mixed until all trace

of blue has gone and the suspension is colorless. A homogeneous colorless suspension indicates that the SDS has been effectively precipitated.

4. Centrifuge for 10 min at 13,000 rpm ($\sim$ 17,900 x g) in a table-top microcentrifuge. A compact white pellet will form.

5. Apply 800 µl of the supernatant from step 4 to the QIAprep 2.0 spin column by pipetting.

6. Centrifuge for 30–60 s. Discard the flow-through.

7. Recommended: Wash the QIAprep 2.0 spin column by adding 0.5 ml Buffer PB and centrifuging for 30–60 s. Discard the flow-through. This step is necessary to remove trace nuclease activity when using endA+ strains, such as the JM series, HB101 and its derivatives, or any wild-type strain, which have high levels of nuclease activity or high carbohydrate content. Host strains, such as XL-1 Blue and DH5α , do not require this additional wash step.

8. Wash QIAprep 2.0 spin column by adding 0.75 ml Buffer PE and centrifuging for 30–60 s.

9. Discard the flow-through, and centrifuge at full speed for an additional 1 min to remove residual wash buffer. Important: Residual wash buffer will not be completely removed unless the flow-through is discarded before this additional centrifugation. Residual ethanol from Buffer PE may inhibit subsequent enzymatic reactions.

10. Place the QIAprep 2.0 column in a clean 1.5 ml microcentrifuge tube. To elute DNA, add 50 µl Buffer EB (10 mM TrisHCl, pH 8.5) or water to the center of each QIAprep 2.0 spin column, let stand for 1 min and centrifuge for 1 min.

C Minimal phosphate media protocol

For 1 L low phosphate media (0.1 mM phosphate), add:

900 ml H_2O + 0.4 g casamino acids $\rightarrow$ autoclave
100 ml 10 x MOPS
10 ml 20 % glucose
0.1 ml 1 M neutral phosphate buffer
1 ml 20 mg/ml adenine
50 µl 10 mg/ml thiamine
1 ml 50 mg/ml ampicillin
1 ml 34 mg/ml chloramphenicol

10 x MOPS:

MOPS 0.4 M
Tricine 42 mM
NH_4Cl 95 mM
K_2SO_4 2.8 mM
$MgCl_2$ 5.3 mM
NaCl 0.5 M

$CaCl_2$ 5 mM
$FeSO_4$ 0.1 M
Adjust to pH 7.4 with NaOH
Filter sterilize and store at 4 °C
For 1 L 10 x MOPS, add 10 µl micronutrients before use

Micronutrients for 10 x MOPS:
Ammonium molybdate 3 mM
Cobalt chloride 64 mM
Manganese chloride 80 mM
Boric acid 0.4 M
Copper sulphate 16 mM
Zinc sulphate 11 mM
Filter sterilize and store at 4 °C

1 M neutral phosphate buffer:
Na_2HPO_4 0.5 M
NaH_2PO_4 0.5 M
Filter sterilize or autoclave

D QuikChange II Site-Directed Mutagenesis protocol by Promega

Mutant Strand Synthesis Reaction (Thermal Cycling):

1. Synthesize two complimentary oligonucleotides containing the desired mutation, flanked by unmodified nucleotide sequence. Purify these oligonucleotide primers prior to use in the following steps (see Mutagenic Primer Design).

2. Prepare the control reaction as indicated below:

   - 5 µl of 10× reaction buffer
   - 2 µl (10 ng) of pWhitescript 4.5-kb control plasmid (5 ng/µl)
   - 1.25 µl (125 ng) of oligonucleotide control primer #1 [34-mer (100 ng/µl)]
   - 1.25 µl (125 ng) of oligonucleotide control primer #2 [34-mer (100 ng/µl)]
   - 1 µl of dNTP mix
   - 38.5 µl ddH_2O (to bring the final reaction volume to 50 µl)

   Then add:

   - 1 µl of PfuUltra HF DNA polymerase (2.5 U/µl)

3. Prepare the sample reaction(s) as indicated below:

   - 5 µl of 10× reaction buffer
   - x µl (5–50 ng) of dsDNA template

- x µl (125 ng) of oligonucleotide primer #1
- x µl (125 ng) of oligonucleotide primer #2
- 1 µl of dNTP mix
- ddH_2O to a final volume of 50 µl

Then add:

- 1 µl of PfuUltra HF DNA polymerase (2.5 U/µl)

4. Cycle each reaction using the cycling parameters outlined in the following table.

Segment	*Cycles*	*Temperature*	*Time*
1	1	95 °C	30 seconds
2	12-18	95 °C	30 seconds
		55 °C	1 minute
		68 °C	1 minute/kb of plasmid length

5. Adjust segment 2 of the cycling parameters according to the type of mutation desired (see the following table):

Type of mutation desired	*Number of cycles*
Point mutations	12
Single amino acid changes	16
Multiple amino acid changes	18

6. Following temperature cycling, place the reaction on ice for 2 minutes to cool the reaction to $\leq$ 37 °C.

Dpn I Digestion of the Amplification Products:

1. Add 1 µl of the Dpn I restriction enzyme (10 U/µl) directly to each amplification reaction.

2. Gently and thoroughly mix each reaction mixture by pipetting the solution up and down several times. Spin down the reaction mixtures in a microcentrifuge for 1 minute and immediately incubate each reaction at 37 °C for 1 hour to digest the parental (ie., the nonmutated) supercoiled dsDNA.

Transformation of XL1-Blue Supercompetent Cells:

1. Gently thaw the XL1-Blue supercompetent cells on ice. For each control and sample reaction to be transformed, aliquot 50 µl of the supercompetent cells to a prechilled 14-ml BD Falcon polypropylene round-bottom tube.

2. Transfer 1 µl of the Dpn I-treated DNA from each control and sample reaction to separate aliquots of the supercompetent cells. As an optional control, verify the transformation efficiency of the XL1-Blue supercompetent cells by adding 1 µl of the pUC18 control plasmid (0.1 ng/µl) to a 50-µl aliquot of the supercompetent cells. Swirl the transformation reactions gently to mix and incubate the reactions on ice for 30 minutes.

3. Heat pulse the transformation reactions for 45 seconds at 42 °C and then place the reactions on ice for 2 minutes. Note: This heat pulse has been optimized for transformation in 14-ml BD Falcon polypropylene round-bottom tubes.

4. Add 0.5 ml of NZY+ broth preheated to 42 °C and incubate the transformation reactions at 37 °C for 1 hour with shaking at 225–250 rpm.

5. Plate the appropriate volume of each transformation reaction, as indicated in the table below, on agar plates containing the appropriate antibiotic for the plasmid vector.

Reaction type	*Volume to plate*
pWhitescript mutagenesis control	250 µl
pUC18 transformation control	5 µl (in 200 µl of NZY+ broth)
Sample mutagenesis	250 µl on each of two plates (entire transformation reaction)

6. Incubate the transformation plates at 37 °C for >16 hours.

E His-tag cleavage reaction protocol by Invitrogen

Recommended Conditions for Cleavage of a Fusion Protein:

1. Add the following to a microcentrifuge tube:
   - Fusion Protein 20 µg
   - 20X TEV Buffer 7.5 µl
   - 0.1 M DTT 1.5 µl
   - AcTEV Protease, (10 units) 1.0 µl
   - Water to 150 µl
2. Incubate at 30 °C. Remove 30 µl aliquots at 1, 2, 4, and 6 hours.
3. Add 30 µl 2X SDS sample buffer (125 mM Tris-HCl, pH 6.8; 4 % SDS; 1.4 M β-mercaptoethanol; 20 % (v/v) glycerol; 0.01 % bromophenol blue). Keep samples at -20 °C until experiment is complete.

F Code for synthetic lnP error generation using HDXsimulator

```
import random
import os

count = 1
while count ¡= 1000:
  betac = random.uniform(0.2, 0.5)
  betah = random.uniform(1, 3)
  os.system("sed -i s/Betac=.*/Betac=" + str(betac) + "/ simulator.py")
  os.system("sed -i s/Betah=.*/Betah=" + str(betah) + "/ simulator.py")
  os.system("python3 HDXsimulator.py")
```

```
  os.system("awk '-print $6 ' path/to/output/files* ¿ path/to/new/directory/lnPkobs
                                                -SD-" + str(count))
  print(count)
  count = count + 1
```

G Code for the calculation of k_{obs} values

```
import math
import os
import glob
import shutil

inputlnP = "path/to/input/file/lnPkobs-*"
inputkint = "path/to/input/file/kint.inp"

listoffiles = glob.glob(inputlnP)
for filename in listoffiles:
  lnP = []
  kint = []
  with open(filename, "r") as f:
    lines = f.readlines()
    for line in lines:
      if line != "0“n":
        lnP.append(line)
  with open(inputkint, "r") as ff:
    lines = ff.readlines()
    for line in lines:
      kint.append(line)
  with open(filename + "@out", "w") as fff:
      fff.write("0“n")
  for l, k in zip(lnP, kint):
    value = math.exp(float(l))
    kobs = float(k)/value
    with open(filename + "@out", "a") as fff:
      fff.write(str(kobs) + "“n")

destdir = "path/to/output/files"
for file in glob.glob("path/to/input/files/*@out"):
  shutil.copy(file, destdir)
  try:
    os.remove(file)
  except:
    pass

for file in glob.glob("path/to/output/files/*@out"):
  p1, p2, p3 = file.split("/")
  par1, par2 = p3.split(" ")
  part1, part2 = par2.split("@")
  os.rename(file, "path/to/output/files/" + part1)
```

H Original barstar protocol by the Ikura group

1. Transformation of *E. coli* BL32(DE3)/pLysE:
    - Add DNA (<50 ng) to each competent cells (∼ 25 µl) on ice
    - Mix the contents gently
    - Store on ice for 30 mins
    - Transfer into a bath preheated at 42 °C
    - Store for 45 sec
    - Rapidly transfer to an ice bath
    - Keep on ice for 2 min
    - Add 0.5 ml of LB preheated at 37 °C
    - Incubate at 37 °C for 1 h with 200 rpm
    - Transfer 100 µl onto agar-plate containing 50 µg/ml Amp & 34 µg/ml Cam
    - Incubate at 37 °C o/n
2. Preculture:
    - Inoculate a single colony into 6 ml of LB containing 50 µg/ml Amp & 34 µg/ml Cam & 1 % glucose
    - Incubate at 37 °C o/n with 200 rpm
3. Large scale culture:
    - Inoculate 6 ml of the o/n culture into 1 L of 2xYT containing 50 µg/ml Amp & 34 µg/ml Cam
    - Incubate at 37 °C with 110 rpm
    - Add 1 ml of 1M IPTG after OD600nm reaches 0.5 (2.5-3.5 h)
    - Incubate at 37 °C with 110 rpm with 110 rpm for 4h-o/n
    - Harvest by centrifugation (4000 rpm and 10 min)
    - Store at -20 °C
4. Lysis & salting out:
    - Resuspend the cells with 50 ml/2L culture lysis buffer
    - Homogenize by homogenizer
    - Sonicate with 190 W and 15 min
    - Centrifuge for 15 min at 15000 rpm and collect the supernatant
    - Add ammonium sulphate up to 40 % at 4 °C and mix well
    - Centrifuge for 15 min at 15000 rpm and collect the supernatant
    - Add ammonium sulphate up to 80 % at 4 °C and mix well
    - Centrifuge for 15 min at 15000 rpm and collect the pellet

- Dissolve with 20 ml/2L culture dialysis buffer
- Dialyze o/n in 2 L of dialysis buffer

5. Purification – Ion exchange (Q sepharose prep):
    - Equilibrate with 3 vol. of starting buffer at 4 ml/min (150 ml
    - Load the sample after filtration
    - Wash with 3 vol. of starting buffer at 4 ml/min (150 ml
    - Run with a linear gradient from 0 to 1 M NaCl at 4 ml/min (250 ml
    - Collect the elution with 4 ml of fraction size
    - Dialysis o/n in 2 L dialysis buffer
6. Purification – gel filtrate (superdex G-75 prep):
    - Equilibrate with 2 vol. of dialysis buffer at 3 ml/min
    - Load the sample after filtration
    - Run with more than 1 vol. of the buffer at 3 ml/min (350 ml)
    - Collect the elution with 4 ml of fraction size

I Original barnase protocol by the Ikura group

1. Transformation of *E. coli* BL32(DE3)/pLysS:
    - Add DNA (<50 ng) to each competent cells ($\sim$ 25 µl) on ice
    - Mix the contents gently
    - Store on ice for 30 mins
    - Transfer into a bath preheated at 42 °C
    - Store for 45 sec
    - Rapidly transfer to an ice bath
    - Keep on ice for 2 min
    - Add 0.5 ml of LB preheated at 37 °C
    - Incubate at 37 °C for 1 h with 200 rpm
    - Transfer 100 µl onto agar-plate containing 50 µg/ml Amp & 34 µg/ml Cam
    - Incubate at 37 °C o/n
2. Preculture:
    - Inoculate a single colony into 6 ml of LB containing 50 µg/ml Amp & 34 µg/ml Cam & 1 % glucose
    - Incubate at 37 °C o/n with 200 rpm
3. Large scale culture

- Inoculate 6 ml of the o/n culture into 1 L of low phosphate media containing 50 µg/ml Amp & 34 µg/ml Cam
- Incubate at 30 °C with 110 rpm

4. Osmotic shock & ion exchange I (SP sepharose):
    - Add 55 ml of acetic acid into 1 L culture
    - Stir for 15-30 min at 4 °C
    - Centrifuge for 15 min at 7500 rpm and collect the supernatant
    - Load on SP sepharose column (gel vol. = ∼50 ml) at 3 ml/min
    - Wash with 3 vol. of sodium acetate (pH 5) at 4 ml/min (150 ml)
    - Elute with high-salt buffer at 4 ml/min and collect the elution with 1000 drops/tube
    - Dialyze o/n in 2 L dialysis buffer
5. Purification – ion exchange II (SP sepharose prep):
    - Load sample after filtration
    - Wash with 3 vol. of starting buffer at 4 ml/min (150 ml)
    - Run with a linear gradient from 0 to 0.5 M NaCl at 3 ml/min (250 ml)
    - Collect the elution with 4 ml of fraction size

J Code for the calculation of Gaussian error lnP values

```
import numpy as np

stdev = [1, 2, 3, 4, 5, 6]
repeats = 3
inputfile = "path/to/input/file/lnP.txt"
outputfile = "path/to/output/files/lnPkobs"

for i in stdev:
  data = []
  count = 1
  while count $¡$= repeats:
    with open(inputfile, "r") as f:
      lines = f.readlines()
      for line in lines:
        myarray = np.fromstring(line, dtype=float, sep=',')
        syntherror = np.random.normal(size = 1, loc = myarray, scale = i)
        data.append(syntherror)
    with open(outputfile + "-SD%s-%s.txt" % (i, count), "w") as ff:
      ff.write("0“n")
    with open(outputfile + "-SD%s-%s.txt" % (i, count), "a") as ff:
      np.savetxt(ff, data, newline="“n")
    del data[:]
    count = count + 1
```

K Code for the semi-random shuffling of lnP values

```
import random

shuffle = [1, 2, 3, 4, 5, 6]
repeats = 3

inputfile = "path/to/input/file/lnP.txt"
outputfile = "path/tp/output/files/lnPkobs"
lst = []

with open(inputfile, "r") as f:
  lines = f.readlines()
  for line in lines:
    lst.append(line.strip())

for maxdistance in shuffle:
  count = 1
  while count ¡= repeats:
    initialarray = lst
    elementtaken = [(i ¡ maxdistance) or i ¿ (len(initialarray) + maxdistance -
                                              1) for i in range(len(initialarray) +
                                               maxdistance * 2)]
    result = [0] * len(initialarray)
    for i in range(len(initialarray)):
        if not elementtaken[i]:
            elementtaken[i] = True
            result[i] = initialarray[i - maxdistance]
            continue
        possiblepositions = [position for position in range(i - maxdistance, i +
                                                  maxdistance + 1) if not
                                                  elementtaken[maxdistance +
                                                  position]]
        position = random.choice(possiblepositions)
        result[i] = initialarray[position]
        elementtaken[maxdistance + position] = True
    with open(outputfile + "-SD%s-%s.txt" % (maxdistance, count), "w") as ff:
      ff.write("0“n")
    with open(outputfile + "-SD%s-%s.txt" % (maxdistance, count), "a") as ff:
      for i in result:
        ff.write(str(i) + "“n")
    count = count + 1
```

L PAPER: Quantitative Evaluation of Native Protein Folds and Assemblies by Hydrogen Deuterium Exchange Mass Spectrometry (HDX-MS).

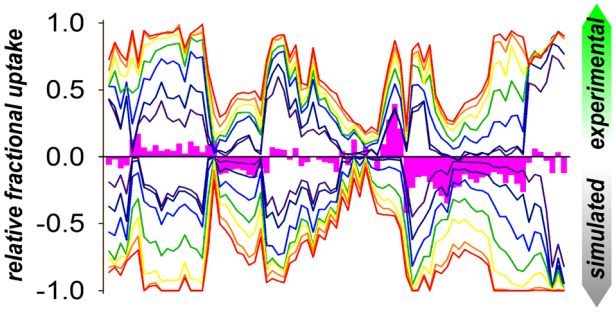
J. Am. Soc. Mass Spectrom. (2018)
DOI: 10.1007/s13361-018-2070-3

FOCUS: HONORING CAROL V. ROBINSON'S ELECTION TO THE NATIONAL ACADEMY OF SCIENCES: RESEARCH ARTICLE

Quantitative Evaluation of Native Protein Folds and Assemblies by Hydrogen Deuterium Exchange Mass Spectrometry (HDX-MS)

Matthew J. Harris, Deepika Raghavan, Antoni J. Borysik

Department of Chemistry, King's College London, Britannia House, London, SE1 1DB, UK



Abstract. Hydrogen deuterium exchange mass spectrometry (HDX-MS) has significant potential for protein structure initiatives but its relationship with protein conformations is unclear. We report on the efficacy of HDX-MS to distinguish between native and non-native proteins using a popular approach to calculate HDX protection factors (PFs) from protein structures. The ability of HDX-MS to identify native protein conformations is quantified by binary structural classification such that merits of the approach for protein modelling can be quantified and better understood. We show that highly accurate PF calculations are not a prerequisite for HDX-MS simulations that are capable of effectively discriminating between native and non-native protein folds. The simulations can also be performed directly on unique structures facilitating high-throughput evaluation of many alternate conformations. The ability of HDX-MS to classify the conformations of homo-protein assemblies is also investigated. In contrast to protein monomers, we show a significant lack of correspondence between the simulated and experimental HDX-MS data for these systems with a subsequent decrease in the ability of HDX-MS to identify native states. However, we demonstrate surprisingly high diagnostic ability of the simulated data for assemblies in which a significant proportion of the individual chains occupy protein-protein interfaces. We relate this to the number of peptides that can sample alternate subunit orientations and discuss these observations within the larger context of applying HDX-MS to evaluate protein structures.

Keywords: Hydrogen deuterium exchange mass spectrometry, Protein structure

Received: 15 March 2018/Revised: 14 September 2018/Accepted: 14 September 2018

Introduction

Hydrogen deuterium exchange mass spectrometry (HDX-MS) reports on time-dependent changes in the deuterium uptake of a protein in D_2O solvent with a structural probe at virtually every amino acid along the protein backbone [1–3]. Despite many advantages of HDX-MS including speed and sensitivity, the method is normally limited to providing qualitative insight into protein conformations. Protein structures are typically required to inform on experimental outputs but the use of HDX-MS to determine protein structures is something of a novelty. We recently demonstrated the potential for simulating the HDX-MS patterns of proteins to elucidate the structures of hetero-protein assemblies [4]. Here, HDX protection factors (PFs) were estimated from atomic coordinates and then used to modify the chemical exchange rates of residues to calculate the isotope uptake of each peptide. The approach facilitated the high-throughput ranking of docking poses based on pairwise comparisons with experimental data. Importantly, it permitted the quantitative discrimination of different poses without the need for additional processing or user interpretation.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13361-018-2070-3) contains supplementary material, which is available to authorized users.

Correspondence to: Antoni Borysik; *e-mail:* antoni.borysik@kcl.ac.uk

Published online: 02 October 2018

The potential for determining native protein folds by HDX-MS is another exciting application of the technique. Accurately predicting protein exchange rates remains a significant challenge although the ability of predictive tools to discriminate between native and non-native folds by HDX-MS has not been previously investigated or quantified [5–8]. Here, we extend our previous work on HDX-MS protein modelling to investigate the performance of these methods to identify native protein folds and the conformations of homomeric protein assemblies. We show that the HDX-MS patterns of proteins simulated directly from their atomic structures are sufficiently accurate to discriminate between native and non-native protein folds. In contrast, the simulated HDX-MS profiles of homo-protein complexes are shown to correspond poorly with their respective experimental outputs. Surprisingly, the capacity to discriminate between native and non-native quaternary structures of protein complexes is high for protein assemblies in which each subunit has multiple interchain contacts. We relate this to an increase in the number of peptides that can sample alternate chain orientations in these systems. Taken together, these data add to our understanding of the use of HDX-MS for structural evaluation and provide an important foundation on which future developments in the area can be built.

Methods

Mass Spectrometry

HDX-MS experiments were performed on a Synapt G2Si HDMS coupled to an Acquity UPLC M-Class system with HDX and automation (Waters Corporation, Manchester, UK). Human alpha lactalbumin (Athens Research and Technology Inc., Athens, USA), enolase from baker's yeast (Sigma-Aldrich Ltd., Dorset, UK) and serum amyloid P component (SAP) from human serum (Merck Chemicals Ltd., Nottingham, UK) were purchased as lyophilised powder, and barnase was prepared in-house. The isotope uptake of each protein was determined using a continuous labelling workflow at 20 °C. Each protein was dissolved in buffer E (10 mM potassium phosphate pH 7.0) to a final concentration of 5–10 μM. Isotope labelling was initiated by diluting 5 μl of each protein into 95 μl of buffer L (10 mM potassium phosphate in D_2O pD 6.6) for various time points. Aliquots of each reaction were taken and quenched by diluting in equal volumes of ice-cold 2% formic acid. Human alpha lactalbumin was quenched in an equal volume of 10 mM phosphate buffer containing 0.4 M tris(2-carboxyethyl)phosphine hydrochloride (Bertin Pharma, Bretonneux, France) and 1.5% HCl to promote pepsin digestion by reduction of disulphide bonds and barnase quench solutions contained 4 M urea. Proteins were digested online with a Waters Enzymate BEH pepsin column at 20 °C. The coverage and redundancy of alpha lactalbumin and barnase digestion were enhanced by increasing the column pressure to 7000 psi with the aid of a back pressure regulator (Waters Corporation). Peptides were trapped on a Waters BEH C18 VanGuard pre-column for 3 min at a flow rate of 200 μl/min in buffer A (0.1% formic acid ~ pH 2.5) before being applied to a Waters BEH C-18 analytical column. Peptides were eluted with a linear gradient of buffer B (0.1% formic acid in acetonitrile ~ pH 2.5) at a flow rate of 40 μl/min. All trapping and chromatography were performed at 0.5 °C to minimise back exchange. MS data were acquired using an MS^E workflow in HD mode with extended range enabled to reduce detector saturation and maintain peak shapes and all labelling time points were obtained in triplicate. The MS was calibrated separately against NaI and the MS data were obtained with lock mass correction using Leu-enkephalin. Peptides were assigned with the ProteinLynx Global Server (PLGS, Waters Corporation, Manchester, UK) software and the isotope uptake of each peptide determined with DynamX v3.0. The isotope uptake of each peptide was corrected for back/in exchange according to methods outlined by Zhang [1]. Fully deuterated protein samples were prepared by dissolving lyophilised samples in buffer L; each sample was then sterilised using a 0.2-μm syringe filter prior to incubation at 37 °C for at least 3 weeks. The isotope uptake of each peptide is reported as the relative fractional uptake (RFU) which is the observed mass shift of a peptide normalised to the maximum possible change in mass.

Simulating Protein HDX-MS Patterns

HDX protection factors (PFs) were estimated according to near-contacts criteria and hydrogen bonding as previously described where the protection of residue i ($\ln P_i^{sim}$) is expressed as the number of heavy atoms (N_i^C) and hydrogen bond acceptors (N_i^H) within defined distance cutoffs from the backbone amide each weighted by an empirically determined scaling term (β) (Eq. 1) [4, 5]:

$$\ln P_i^{sim} = N_i^C \beta_C + N_i^H \beta_H \quad (1)$$

When compared to experimental data previously obtained by NMR, Eq. 1 significantly overestimates the PFs of backbone amides [9]. To account for this discrepancy, a separate exclusion parameter (excl) was introduced that allowed the outputs to be rescaled by omitting the contribution of all heavy atoms from the contact calculations of user-defined residues: where excl = 0 reports all heavy atoms for PF calculations of residue i; excl = 1 omits the atoms of residue i; excl = 2 omits the atoms of residue i and immediately adjacent residues and so on. In addition to this, a smoothing function was also introduced for atom counting within the cutoff distance, where dist(h, O) and dist(n, heavyAtom) are the linear distances relating to the respective hydrogen bond and contact calculations and hcut and heavycut are the respective cutoff distances of 2.4 and 6.5 Å (Supporting Information, Fig. S1, Eq. 2) [10]:

$$\ln P_i^{sim} = \frac{\beta_H}{1 + e^{10\mathrm{dist}(h,O)-\mathrm{hcut}}} + \frac{\beta_C}{1 + e^{5\mathrm{dist}(n,\mathrm{heavyAtom})-\mathrm{heavycut}}} \quad (2)$$

PFs were simulated directly from the corresponding crystal structures (1A4V, 1A2P, 1SAC and 3ENL) with missing structure built using Modeller [11–15]. In the case of alpha

lactalbumin, PFs were also calculated from a protein ensemble generated by molecular dynamics (MS) simulations of 1A4V in explicit water. MD simulations were performed using the OPLS/AA force field implemented within GROMACS 4.6.7 [16]. Production MD simulations were carried out at 300 K for 100 ns following energy minimisation and extensive solvent equilibration. One hundred structures were taken along the 100-ns trajectory and protection factors expressed as the average values taken across all conformations. Alpha lactalbumin and barnase decoy sets were prepared using 3DRobot with the output set to 1000 structures [17]. A range of enolase and SAP decoys were prepared using a local installation of SymmDock V1.0 without constraints yielding ca. 10,000 and 5000 transformants for enolase and SAP respectively [18]. Transformants were then refined on a local installation of SymmRef V1.2 using the recommended settings to remove steric clashes and allow for backbone and sidechain flexibility [19].

The simulated PFs were used to generate HDX-MS patterns of each protein using an in-house script implemented within MATLAB. In the case of enolase and SAP, the PFs of each residue were taken as the average across all protein chains. The code takes as input the protein sequence, experimental peptide list of a protein and the start and end positions of each peptide along with the experimental temperature and pD. It then calculates the intrinsic chemical exchange rates (k_{int}) of each backbone amide proton according to previously defined near-neighbour effects using the modified exchange factors for acidic residues [20, 21]. The intrinsic exchange rates and PFs are then used to determine the observed exchange rates (k_{obs}) for each residue according to Eq. 3. The isotope uptake of each peptide is then calculated from the following polyexponential function, where D_t is the total number of deuterium atoms incorporated into the peptide at time t, N is the total number of exchangeable positions and k_i is the observed hydrogen exchange rate constant of residue i (Eq. 4):

$$k_{obs} = \frac{k_{int}}{PF} \tag{3}$$

$$D_t = N - \sum_{i-1}^{N} \exp(-k_i t) \tag{4}$$

Proline residues were discounted along with amino-terminal groups to ensure that the simulated RFU calculations were in line with experimental outputs processed by DynamX.

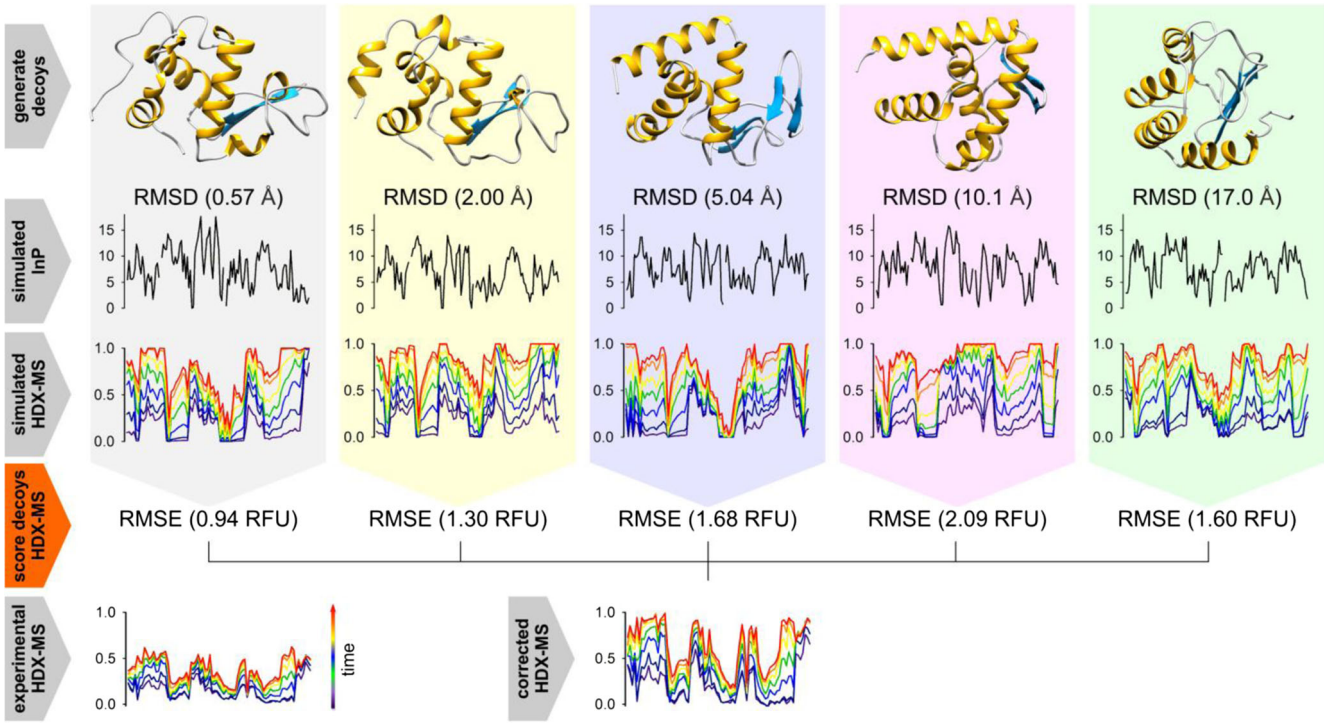


Figure 1. Outline of the HDX-MS simulation workflow and analysis: A set of decoys were first prepared for each protein and the RMSD of each decoy determined by alignment with the native structure. **(top row)** Five example decoys are shown for alpha lactalbumin along with their corresponding RMSD. **(second row)** PFs simulated directly for each decoy according to Eq. 1. **(third row)** The PFs were used to modify the chemical exchange rates and the isotope uptake of each residue determined and projected onto an experimental peptide list to generate a library of simulated HDX-MS profiles. **(fourth row)** The library of HDX-MS simulations was then compared to that of experimental HDX-MS data to obtain the RMSE of each simulation as shown. **(bottom row)** Prior to alignment with the simulated HDX-MS data, all experimental outputs were first corrected for extraneous exchange. Following this process, the simulated HDX-MS profiles were then ranked according to their RMSE with the experimental outputs and their ability to identify native structures evaluated based on their performance in binary structural classification

Expression and Purification of Barnase

Unless stated otherwise, all chemicals were purchased from Fluorochem Ltd., Derbyshire, UK, Sigma-Aldrich Ltd., Dorset, UK, or VWR International Ltd., Leicestershire, UK. Overexpression of wild-type barnase (*Bacillus amyloliquefaciens* ribonuclease) was directed from the plasmid pTZ416 under the control of the alkaline phosphatase promotor and was kindly provided by Prof Teikichi Ikura (Tokyo Medical and Dentistry University, Japan) [22]. The plasmid was transformed into BL21(DE3)pLysS cells and plated onto LB agar plates containing ampicillin (50 mg/ml) and chloramphenicol (34 mg/ml). A single colony was used to inoculate 50 ml LB containing ampicillin and chloramphenicol and incubated overnight at 37 °C with agitation at 220 rpm; 1.2 ml of the pre-culture was then used to inoculate 200 ml low-phosphate media containing ampicillin and chloramphenicol and incubated overnight at 30 °C with agitation at 110 rpm. The low-phosphate media was prepared as follows. For 1 l low-phosphate media, 0.4 g casamino acids was added to 900 ml H_2O and autoclaved. To this, 100 ml 10 × concentrate filter sterilised MOPS (3-(*N*-morpholino)propanesulfonic acid) was added containing 10 ml 20% glucose, 0.1 ml 1 M neutral phosphate buffer, 1 ml of 20 mg/ml adenine, 50 µl 10 mg/ml thiamine, 1 ml 50 mg/ml ampicillin and 1 ml 34 mg/ml chloramphenicol. The concentrated MOPS buffer contained 0.4 M MOPS, 42 mM tricine, 95 mM NH_4Cl, 2.8 mM K_2SO_4, 5.3 mM $MgCl_2$, 0.5 M NaCl, 5 mM $CaCl_2$ and 0.1 M $FeSO_4$ adjusted to pH 7.4 with NaOH which was then filter sterilised. Immediately prior to use, 10 µl micronutrients was added to the MOPS buffer which contained 3 mM ammonium molybdate, 64 mM cobalt chloride, 80 mM manganese chloride, 0.4 M boric acid, 16 mM copper sulphate and 11 mM zinc sulphate sterilised by filtration. The 1 M neutral phosphate buffer contained 0.5 M Na_2HPO_4 and 0.5 M NaH_2PO_4 which was then autoclaved. After overnight incubation, 11 ml acetic acid was added to the cell culture and left mixing for 20 min at 4 °C to promote the release of barnase into the media by osmotic shock. The cells were then centrifuged at 7500 rpm for 15 min and the supernatant retained for purification following vacuum filtration through a 0.22-µm filter. Barnase was then equilibrated against two column volumes of dialysis buffer of 50 mM TrisHCl (tris(hydroxymethyl)aminomethane hydrochloride) pH 8.0 before purification by size exclusion chromatography on a Superdex 75 10/300 GL column (GE Healthcare Life Sciences, Little Chalfont, UK). The purification and identity of barnase were confirmed by SDS/PAGE electrophoresis and mass spectrometry.

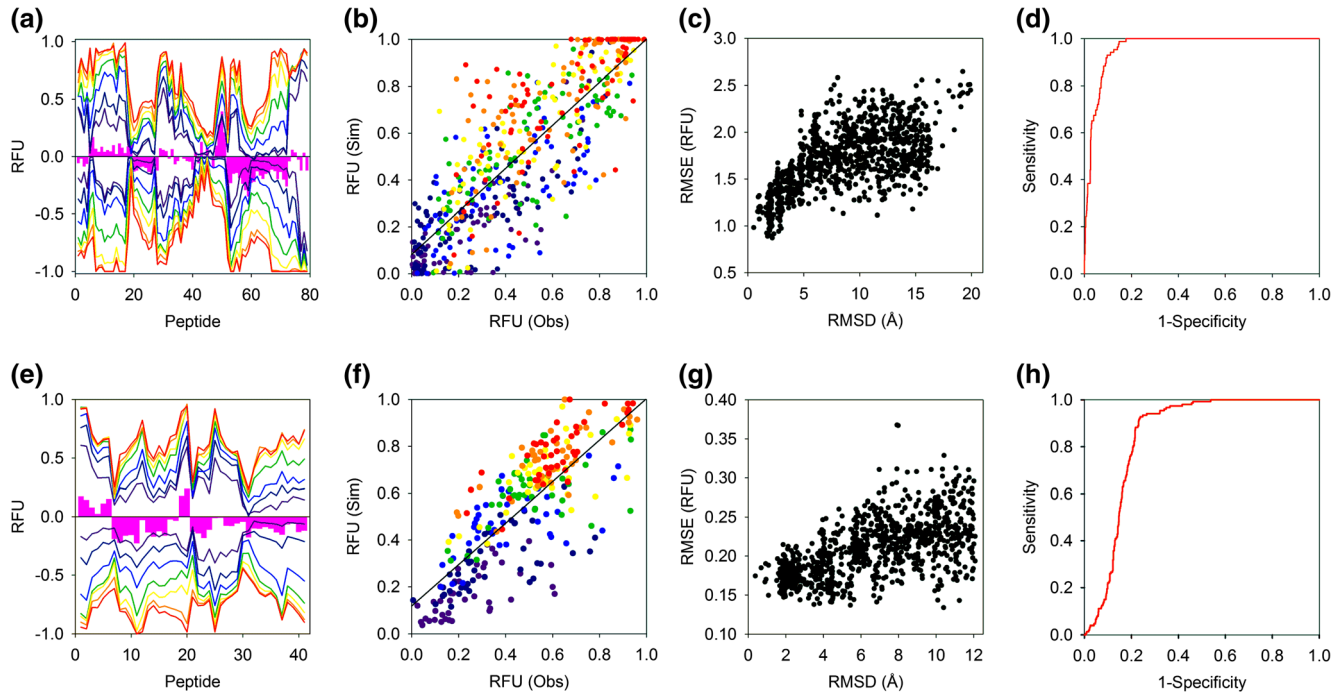


Figure 2. Native folds of alpha lactalbumin and barnase investigated by HDX-MS: (**a**, **e**) Mirror plots comparing experimental (positive) and simulated (negative) HDX-MS outputs. Experimental data were acquired at 0.25, 1, 5, 20, 60, 240 and 480 min at 293.15 K (coloured dark blue through red respectively). The pink bars denote the time-averaged difference in RFU between the experimental and simulated data and are shown to highlight areas of significant change. (**b**, **f**) Scatterplot comparing observed and simulated HDX-MS data of all RFU time points with different labelling times coloured as in (**a**). (**c**, **g**) The relationship between the RMSE and RMSD of 1000 decoys. The RMSE was calculated by pairwise comparison of the simulated and experimental HDX-MS data and the RMSD determined by alignment with the crystal structure. (**d**–**h**) ROC plots demonstrating the ability of the HDX-MS simulations to classify protein structures. Decoys with an RMSD ≤ 2.5 Å with the crystal structure were classified as native. Alpha lactalbumin and barnase data are shown in the upper and lower four figures, respectively

Evaluation of HDX-MS Simulations to Identify Native Structures

The ability of the HDX-MS simulations to discriminate between native and non-native protein structures was quantified from the associated receiver operator characteristic (ROC) plots of a binary classification test. The RMSE of each HDX-MS simulation was obtained by pairwise comparison with the associated experimental outputs across all peptides and labelling time points. The RMSD of each decoy was determined by alignment with the relevant native crystal structure using the McLachlan algorithm implemented on a locally installed copy of ProFit v3.1 with decoys having an RMSD ≤ 2.5 Å classified as native [23, 24]. A ROC plot was then generated for each dataset using SigmaPlot 13.0 (Systat Software Inc., London, UK) and the ability of the HDX-MS simulations to identify native structures determined from the area under the curve (AUC) where values > 0.9 were considered excellent, > 0.8 good, 0.6–0.8 poor to fair and below 0.6 failed.

Results and Discussion

Many different methods have been developed to estimate the HDX behaviour of proteins but the capacity of these approaches to discriminate between native and non-native states by HDX-MS has not been previously tested or quantified. The ability of HDX-MS to identify native protein folds was evaluated with alpha lactalbumin and barnase with the PFs of these proteins simulated according to Eq. 1 after minor optimisation (Fig. S1, "Methods") [5]. The PFs were used to modify the chemical exchange rates of these proteins from which the isotope uptake of each residue was determined and projected onto experimental peptide lists to simulate HDX-MS outputs ("Methods"). The ability of the HDX-MS simulations to discriminate between native and non-native folds was evaluated using decoy sets of 1000 different protein conformations. HDX-MS data was simulated for each decoy generating a library of HDX-MS profiles which were ranked according to their correspondence with experimental data obtained in-house (Fig. 1, "Methods"). A binary classification test was then performed to evaluate the efficacy to which the HDX-MS simulations could discriminate between native and non-native protein folds. The diagnostic ability of the simulated HDX-MS profiles was quantified from the area under the curve (AUC) of the associated ROC plots which is a measure of the success rate of correctly classifying structures selected at random ("Methods").

HDX-MS data simulated for the native states of alpha lactalbumin and barnase correlated surprisingly well with experimental outputs of the proteins. For alpha lactalbumin, the experimental and simulated outputs are practically identical over the first ~ 45 peptides with the accuracy of the simulation only breaking down marginally toward the C-terminal end of the protein. The correspondence between the experimental and simulated data of alpha lactalbumin and barnase is comparable with respective RMSE of 0.174 and 0.165 RFU (Fig. 2(a, e)). The simulated RFU of all labelling time points and peptides agrees well with the experimental data with no significant discrepancies in the gradient of the fit between these data (Fig. 2(b, f)). While the native state HDX-MS simulations of both proteins compare equally well with their respective experimental outputs, there are significant differences in their overall diagnostic ability. For a set of 1000 protein decoys, there are many native (low RMSD) alpha lactalbumin structures that also yield HDX-MS simulations that align closely with the experimental outputs (low RMSE). This contrasts with the barnase decoy set where the clustering around native structures that also generates accurate HDX-MS simulations is qualitatively less apparent (Fig. 2(c, g)). Differences in the ability of HDX-MS to discriminate between native and non-native protein folds of these proteins were confirmed from the

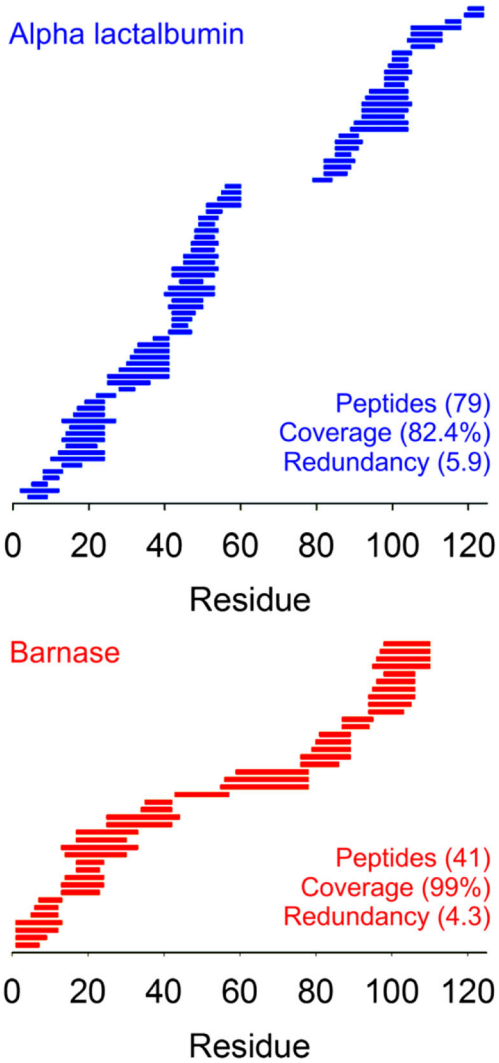


Figure 3. peptide maps of alpha lactalbumin and barnase: The peptide maps of alpha lactalbumin (blue) and barnase (red) that comprise the HDX-MS data of these proteins are shown along with the respective number of peptides, coverage and redundancies. The ~ 20 residue region missing from the alpha lactalbumin data spans two of the four disulphide bonds of the protein

associated ROC plots. The alpha lactalbumin and barnase data have respective AUC values of 0.96 and 0.85 indicating that the HDX-MS simulations of alpha lactalbumin are > 3-fold more likely to correctly identify native and non-native structures than those of barnase (Fig. 2(d, h)). Differences in the diagnostic ability of the HDX-MS of these proteins could reflect variations in the number of peptides that comprise each dataset. While both proteins have similar chain lengths, the barnase HDX-MS profile is comprised of around 50% fewer peptides. Despite a significant region of missing peptides around two of the disulphide bonds of alpha lactalbumin, the peptide redundancy is significantly higher for this protein. High redundancy may enhance the ability of the alpha lactalbumin HDX-MS data to discriminate between different folds resulting in the exceptionally high AUC (Fig. 3).

The accuracy of the HDX-MS simulations of these proteins is remarkable given that the underlying PF estimates correlate poorly with previously determined experimental values (Fig. S1). The HDX-MS data were also simulated directly from crystal structures of the proteins which neglect the ensemble property of HDX and the understanding that exchange is driven by protein motion. The coefficients β_C, β_H (Eq. 1) were previously found by fitting experimental PFs from a limited number of proteins to structural ensembles generated by molecular dynamics (MD) simulations [5]. Surprisingly, however, we found that PFs simulated from the ensemble average of alpha lactalbumin corresponded less well with the experimental PFs of this protein. HDX-MS data simulated from the ensemble average also compared less well with experimental outputs (Fig. S2). Overall, PFs simulated from an MD ensemble of alpha lactalbumin reduced the accuracy of the HDX-MS simulations. While these results are somewhat unexpected, they agree with recent observations showing that data simulated from single structures can improve the correlation with experimental HDX data [25].

We then applied the same approach to characterise the structures of the homo-protein assemblies enolase and SAP. Here, we assume the native fold of the proteins and investigate the ability of the HDX-MS simulations to identify the native chain organisation. In contrast to the HDX-MS simulations of the protein monomers, those obtained for the native protein complexes are characterised by an overall lack of correspondence with their respective experimental outputs (Fig. 4(a, e)). The HDX-MS simulations fail to broadly capture the experimental data with RMSE for the respective HDX-MS simulations of enolase and SAP of 0.219 and 0.212 RFU. The correspondence between all peptides and time points is also asymmetrical with the RFU of the simulations either under or overestimating the experimental values (Fig. 4(b, f)). Despite the poor accuracy of the HDX-MS simulations of both protein

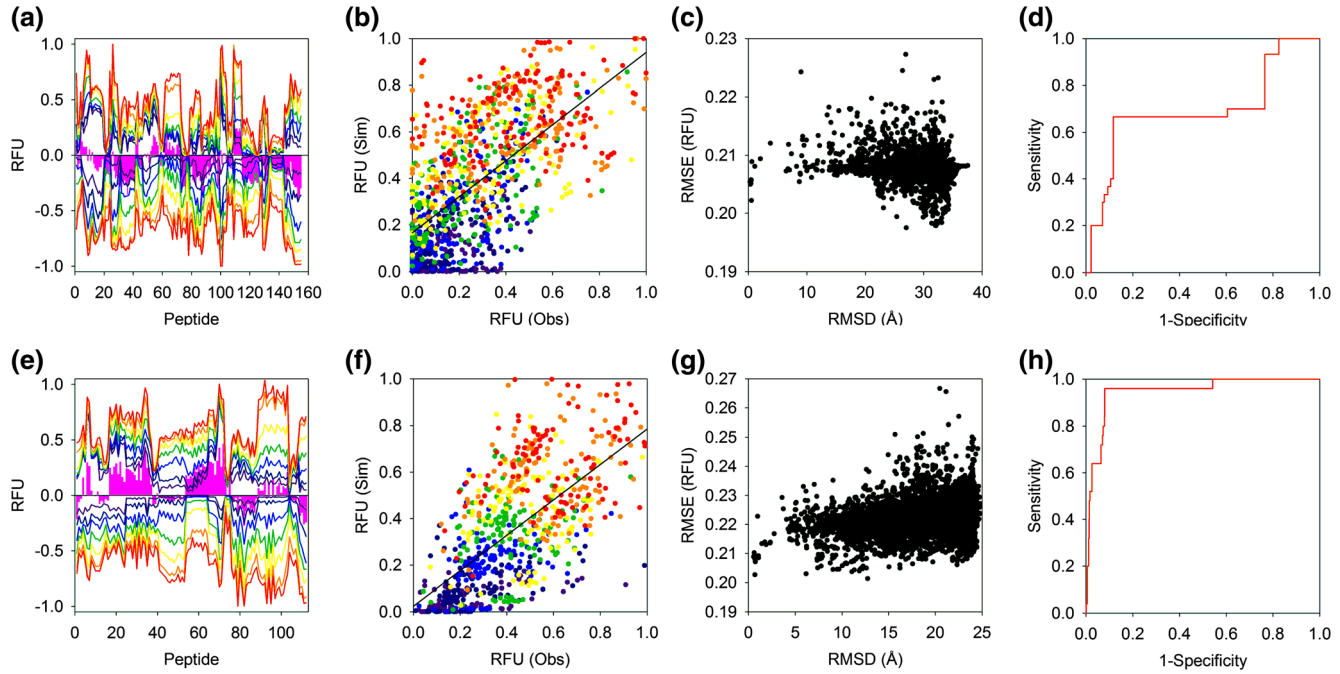


Figure 4. Native structures of enolase and SAP investigated by HDX-MS: (**a**, **e**) Mirror plots comparing experimental (positive) and simulated (negative) HDX-MS outputs. Experimental data were acquired at 0.25, 1, 5, 20, 60, 240, and 480 min at 293.15 K (coloured dark blue through red respectively). The pink bars denote the time-averaged difference in RFU between the experimental and simulated data and are shown to highlight areas of significant change. (**b**, **f**) Scatterplot comparing observed and simulated HDX-MS data of all RFU time points with different labelling times coloured as in (**a**). (**c**, **g**) The relationship between the RMSE and RMSD for a range of decoys. The RMSE was calculated by pairwise comparison of the simulated and experimental HDX-MS data and the RMSD determined by alignment with the crystal structure. (**d**–**h**) ROC plots demonstrating the ability of the HDX-MS simulations to classify protein structures. Decoys with an RMSD ≤ 2.5 Å with the crystal structure were classified as native. Enolase and SAP data are shown in the upper and lower four figures, respectively

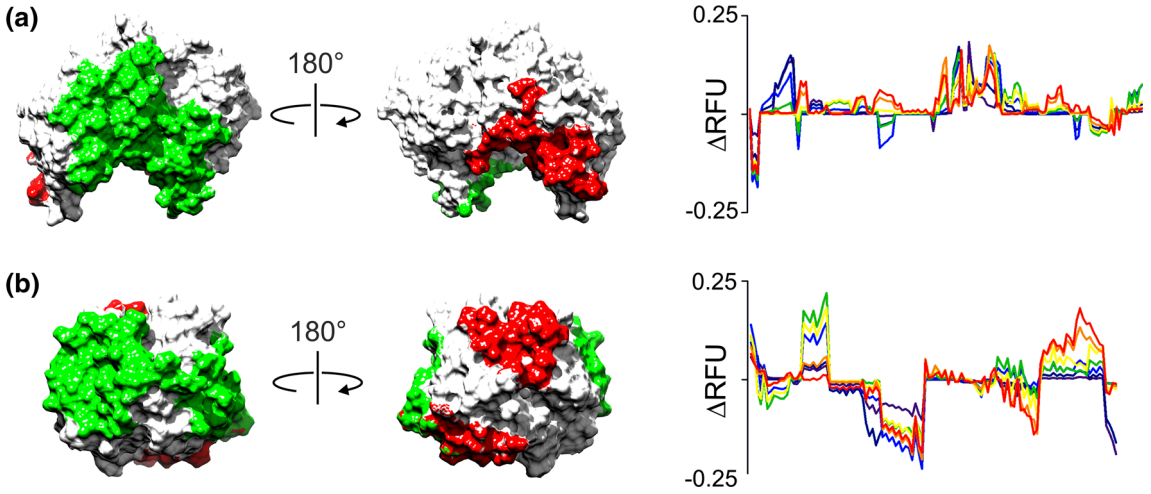


Figure 5. ΔRFU for different chain orientations of enolase and SAP: (**a**) native (green) and non-native (red) protein-protein interfaces shown on a single enolase protein chain. Interfacial regions were defined using a 6.5-Å distance cutoff as used in Eq. 1. The plot shows the ΔRFU between the native and non-native assembly for all peptides. (**b**) as per (**a**) but shown for SAP the ΔRFU between the native and non-native SAP assemblies for all peptides is also shown. Data in the ΔRFU plots reflect the seven different labelling times from 15 s to 8 h, coloured dark blue to red respectively. Non-native interfaces for both proteins represent assemblies with the highest RMSD after alignment with the native complex

complexes, there are significant differences in their ability to discriminate between native and non-native structures. The ability of the enolase simulations to identify native structures is poor with the associated ROC plot indicating failure with an AUC of 0.69 (Fig. 4(c, d)). In contrast, however, the ability of the SAP HDX-MS simulations to correctly classify structures is extremely high with the AUC if the associated ROC plot indicating a success rate of 95% (Fig. 4(g, h)).

Given the inaccuracy of the HDX-MS simulations of both enolase and SAP, the high diagnostic ability of the SAP simulations is unexpected. This is likely attributed to differences in the number of interchain contacts in these proteins. Whereas a significant proportion of each SAP monomer is buried in subunit interfaces of the pentameric complex, the buried regions of each enolase chain are limited to a single dimeric interface. Accordingly, the likelihood of peptides probing protein-protein interfaces is much higher in SAP such that the HDX-MS outputs of this complex can more effectively differentiate between different chain orientations. To highlight this, HDX-MS data were simulated for both enolase and SAP showing the change in RFU (ΔRFU) between the native and a non-native protein complex. As expected, the proportion of each protein chain buried in protein-protein interfaces is significantly higher in SAP with the consequence that many more SAP peptides exhibit large changes in their RFU for the different subunit poses and the ΔRFU of the SAP peptides is more widespread and pronounced (Fig. 5). We suggest that the increased number of interchain contacts in SAP enhances the ability of the HDX-MS simulations of this protein to discriminate between different assembly structures. High numbers of interchain contacts must therefore be particularly important for the modelling of homo-protein complexes by HDX-MS and may in some cases overcome limitations in the accuracy of the simulated data.

Conclusion

The aim of this work was to quantify the ability of HDX-MS to discriminate between native and non-native protein conformations based on a popular approach to estimate PFs from protein structures. The efficacy of the method was evaluated on the peptide level using the PF estimates to calculate HDX-MS outputs of proteins and their assemblies and then comparing these simulations to experimental data obtained in-house. The ability of HDX-MS to identify native structures was quantified based on their performance in binary structural classification to provide insight into the use of HDX-MS for protein modelling.

We show that HDX-MS data simulated directly from protein atomic structures can be highly diagnostic for native protein folds, even when the underlying PFs of these data are poorly defined. For alpha lactalbumin, PF calculations (lnP) with an RMSE of only 2.86 over 44 residues were sufficient to generate HDX-MS outputs capable of discriminating between native and non-native states with a success rate of > 95% (Fig. S1). Our data suggest that high-peptide redundancy may be more important than overall coverage in the ability of HDX-MS to differentiate between native and non-native structures. The alpha lactalbumin HDX-MS data significantly outperformed that of barnase in binary structural classification despite having a peptide coverage of only 82% compared with 99% for barnase. Although the native state HDX-MS simulations of both these proteins agreed equally well with their respective experimental profiles, the peptide redundancy of the alpha lactalbumin data is significantly higher. We propose that the high-peptide redundancy of the alpha lactalbumin HDX-MS outputs enhances the capacity of these data to differentiate between different folds resulting in the exceptionally high AUC. Remarkably, protein ensembles were not required

for these calculations and even reduced the accuracy of the simulated protection factors. While this observation contradicts accepted relationships between protein motions and exchange behaviour, the capacity to generate accurate HDX-MS data from unique states is appealing because of the associated benefits with regard to throughput.

HDX-MS data simulated for homo-protein assemblies compared significantly less well with experimental outputs. This could be due to significant differences in the HDX behaviour of protein complexes and the fact that Eq. 1 was never optimised for use with large multi-chain proteins. To better understand the scope of Eq. 1, HDX-MS data were simulated over a range of different β_C, β_H weighting values and the outputs compared the experimental data. While the expression could be marginally optimised to improve the correspondence between the simulated and experimental profiles, this did not improve the ability of the simulations to correctly classify the quaternary conformations of protein assemblies (Fig. S3). The inability of Eq. 1 to describe the HDX behaviour of protein assemblies may originate from more pronounced EX1 exchange in these assemblies which is not defined by the current approach. However, no significant EX1 signatures were visible in the experimental isotope patterns of these proteins suggesting that equilibrium exchange (EX2) dominates the isotope uptake of these proteins (data not shown). Interestingly, the HDX-MS simulations of the pentameric protein assembly SAP were shown to be highly diagnostic of the native complex in spite of their poor correspondence with experimental data. We suggest that this stems from a greater number of protein-protein interfaces in this complex with an associated increase in the number of peptides available to sample native and non-native chain orientations. However, this observation also points to a limitation in the characterisation of homo-protein complexes in that knowledge of peptide redundancy and coverage in the native interface can only be had with the aid of a high-resolution structure. This is not a challenge for hetero-proteins however, as the degree of peptide sampling in the native interface can be inferred directly from associated HDX-MS difference data without the need for any structural reference. Indeed, the ability of HDX-MS to provide detailed footprinting information on the protein-protein interfaces of hetero-protein complexes in the absence of any structural information is one of the major strengths of the technique.

We have demonstrated that a simple expression used to calculate protein exchange behaviour is sufficient to simulate HDX-MS data that can effectively differentiate between native and non-native protein folds. While these data are limited to a few selected protein structures and further work is required to understand the scope of these expressions, they do provide an important window in the use of HDX-MS for protein modelling. Peptide redundancy appears to be more important than overall coverage for these approaches and a high degree of interchain contacts is essential for HDX-MS guided modelling of protein complexes. Future work to characterise and develop improved expressions for calculating the PFs of proteins from their atomic structures may unlock previously untapped potential of HDX-MS in areas such as ab initio protein folding and high-throughput structure determination. This will require a greater understanding of the relationship between protein structure and HDX for which the present work represents a useful platform.

Open Access

References

1. Zhang, Z., Smith, D.L.: Determination of amide hydrogen exchange by mass spectrometry: a new tool for protein structure elucidation. Protein science : a publication of the Protein Society. **2**, 522–531 (1993)
2. Robinson, C.V., Gross, M., Eyles, S.J., Ewbank, J.J., Mayhew, M., Hartl, F.U., Dobson, C.M., Radford, S.E.: Conformation of GroEL-bound alpha-lactalbumin probed by mass spectrometry. Nature. **372**, 646–651 (1994)
3. Marciano, D.P., Dharmarajan, V., Griffin, P.R.: HDX-MS guided drug discovery: small molecules and biopharmaceuticals. Curr. Opin. Struct. Biol. **28**, 105–111 (2014)
4. Borysik, A.J.: Simulated isotope exchange patterns enable protein structure determination. Angew. Chem. **56**, 9396–9399 (2017)
5. Best, R.B., Vendruscolo, M.: Structural interpretation of hydrogen exchange protection factors in proteins: characterization of the native state fluctuations of CI2. Structure. **14**, 97–106 (2006)
6. Craig, P.O., Latzer, J., Weinkam, P., Hoffman, R.M., Ferreiro, D.U., Komives, E.A., Wolynes, P.G.: Prediction of native-state hydrogen exchange from perfectly funneled energy landscapes. J. Am. Chem. Soc. **133**, 17463–17472 (2011)
7. Liu, T., Pantazatos, D., Li, S., Hamuro, Y., Hilser, V.J., Jr. Woods, V.L.: Quantitative assessment of protein structural models by comparison of H/D exchange MS data with exchange behavior accurately predicted by DXCOREX. J. Am. Soc. Mass Spectrom. **23**, 43–56 (2012)
8. Park, I.H., Venable, J.D., Steckler, C., Cellitti, S.E., Lesley, S.A., Spraggon, G., Brock, A.: Estimation of hydrogen-exchange protection factors from MD simulation based on amide hydrogen bonding analysis. J. Chem. Inf. Model. **55**, 1914–1925 (2015)
9. Schulman, B.A., Redfield, C., Peng, Z.Y., Dobson, C.M., Kim, P.S.: Different subdomains are most protected from hydrogen exchange in the molten globule and native states of human alpha-lactalbumin. J. Mol. Biol. **253**, 651–657 (1995)
10. Best, R.B.: personnal communication. (2016)
11. Chandra, N., Brew, K., Acharya, K.R.: Structural evidence for the presence of a secondary calcium binding site in human alpha-lactalbumin. Biochemistry. **37**, 4767–4772 (1998)
12. Martin, C., Richard, V., Salem, M., Hartley, R., Mauguen, Y.: Refinement and structural analysis of barnase at 1.5 A resolution. Acta Crystallogr. D Biol. Crystallogr. **55**, 386–398 (1999)
13. Emsley, J., White, H.E., O'Hara, B.P., Oliva, G., Srinivasan, N., Tickle, I.J., Blundell, T.L., Pepys, M.B., Wood, S.P.: Structure of pentameric human serum amyloid P component. Nature. **367**, 338–345 (1994)
14. Stec, B., Lebioda, L.: Refined structure of yeast apo-enolase at 2.25 A resolution. J. Mol. Biol. **211**, 235–248 (1990)
15. Sali, A., Blundell, T.L.: Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. **234**, 779–815 (1993)
16. Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., Berendsen, H.J.: GROMACS: fast, flexible, and free. J. Comput. Chem. **26**, 1701–1718 (2005)

17. Deng, H., Jia, Y., Zhang, Y.: 3DRobot: automated generation of diverse and well-packed protein structure decoys. Bioinformatics. **32**, 378–387 (2016)
18. Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., Wolfson, H.J.: PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res. **33**, 363–367 (2005)
19. Mashiach-Farkash, E., Nussinov, R., Wolfson, H.J.: SymmRef: a flexible refinement method for symmetric multimers. Proteins. **79**, 2607–2623 (2011)
20. Bai, Y., Milne, J.S., Mayne, L., Englander, S.W.: Primary structure effects on peptide group hydrogen exchange. Proteins. **17**, 75–86 (1993)
21. Mori, S., van Zijl, P.C., Shortle, D.: Measurement of water-amide proton exchange rates in the denatured state of staphylococcal nuclease by a magnetization transfer technique. Proteins. **28**, 325–332 (1997)
22. Urakubo, Y., Ikura, T., Ito, N.: Crystal structural analysis of protein-protein interactions drastically destabilized by a single mutation. Protein science : a publication of the Protein Society. **17**, 1055–1065 (2008)
23. McLachlan, A.D.: Rapid comparison of protein structures. Acta Cryst. **A38**, 871–873 (1982)
24. http://www.bioinf.org.uk/software/profit/
25. Devaurs, D., Antunes, D.A., Papanastasiou, M., Moll, M., Ricklin, D., Lambris, J.D., Kavraki, L.E.: Coarse-grained conformational sampling of protein structure improves the fit to experimental hydrogen-exchange data. Front. Mol. Biosci. **4**(13), (2017)

M PAPER: HDX and Native Mass Spectrometry Reveals the Different Structural Basis for Interaction of the Staphylococcal Pathogenicity Island Repressor Stl with Dimeric and Trimeric Phage dUTPases.

biomolecules

MDPI

Article

HDX and Native Mass Spectrometry Reveals the Different Structural Basis for Interaction of the Staphylococcal Pathogenicity Island Repressor Stl with Dimeric and Trimeric Phage dUTPases

Kinga Nyíri [1,2,*,†], Matthew J. Harris [3,†], Judit Matejka [1,2], Olivér Ozohanics [4,5], Károly Vékey [4], Antoni J. Borysik [3,*] and Beáta G. Vértessy [1,2,*]

1 Department of Applied Biotechnology and Food Sciences, Budapest University of Technology and Economics, 1111 Budapest, Hungary
2 Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, 1117 Budapest, Hungary
3 Department of Chemistry, King's College London, Britannia House, London SE1 1DB, UK
4 Institute of Organic Chemistry, Research Centre for Natural Sciences, Hungarian Academy of Sciences, 1117 Budapest, Hungary
5 Department of Medical Biochemistry, Semmelweis University, 1085 Budapest, Hungary
* Correspondence: knyiri@mail.bme.hu (K.N.); antoni.borysik@kcl.ac.uk (A.J.B.); vertessy@mail.bme.hu (B.G.V.)
† These authors contributed equally to this paper.

Received: 13 July 2019; Accepted: 11 September 2019; Published: 14 September 2019

Abstract: The dUTPase enzyme family plays an essential role in maintaining the genome integrity and are represented by two distinct classes of proteins; the β-pleated homotrimeric and the all-α homodimeric dUTPases. Representatives of both trimeric and dimeric dUTPases are encoded by *Staphylococcus aureus* phage genomes and have been shown to interact with the Stl repressor protein of *S. aureus* pathogenicity island SaPIbov1. In the present work we set out to characterize the interactions between these proteins based on a range of biochemical and biophysical methods and shed light on the binding mechanism of the dimeric φNM1 phage dUTPase and Stl. Using hydrogen deuterium exchange mass spectrometry, we also characterize the protein regions involved in the dUTPase:Stl interactions. Based on these results we provide reasonable explanation for the enzyme inhibitory effect of Stl observed in both types of complexes. Our experiments reveal that Stl employs different peptide segments and stoichiometry for the two different phage dUTPases which allows us to propose a functional plasticity of Stl. The malleable character of Stl serves as a basis for the inhibition of both dimeric and trimeric dUTPases.

Keywords: dUTPase; inhibition; interaction surface; Stl staphylococcal repressor

1. Introduction

Infections caused by *Staphylococcus aureus* are hazardous for both humans and livestock especially since *S. aureus* strains develop resistance and adapt to the new hosts rapidly via horizontal gene transfer (HGT)[1,2]. Highly mobile *S. aureus* pathogenicity islands (SaPI) play a key role in this process since they frequently carry genes encoding toxic shock syndrome toxin, staphylococcal enterotoxin B, and other superantigens [3]. The spread of SaPIs is mediated by the so-called helper phages through a unique mechanism in which SaPIs residing in the staphylococcal genome replicate autonomously upon helper phage invasion or prophage activation. Thereafter a specific derepressor protein of the phage relieve the repression of the genes responsible for SaPI excision, replication, and packaging [4].

It has been shown that in case of SaPIbov1 pathogenicity island homotrimeric dUTPase enzymes of specific phages are responsible for the SaPI induction through direct interaction with Stl; the master repressor protein of the SaPI lifecycle [5,6]. Homotrimeric phage dUTPases share a conserved core and also frequently contain an approximately 30–40-residue-long, diverse phage-specific insertion. It has been hypothesized that this phage-specific insert plays an important role in the SaPI induction since the pH15 phage dUTPase, which lacks this insertion region, cannot function as a SaPI-derepressor [5,7]. However, it has been shown that a mutant φ11 phage dUTPase lacking this phage-specific insert also interacts with the Stl, although it disrupts the Stl-DNA interaction less effectively than the wild-type protein and has a reduced capacity to induce SaPIbov1 [6–8]. Moreover, it was also reported that mycobacterial dUTPase lacking the phage specific insert can also bind to Stl in vitro and in vivo [9]. On the other hand, φSaov3 and φB2 phage dUTPases that contain the same sequence of the insert as the SaPI inducing dUTPases of phage φ11 and 80α, respectively, are not capable of derepression [10]. In addition, it has also been established that neither φ11 nor 80α phage dUTPases can bind to the substrate dUTP and Stl at the same time, suggesting the involvement of the dUTPase active site in the dUTPase:Stl complex formation [6,7]. The active site of trimeric dUTPases is built up as the following: A substrate binding pocket is formed by conserved motifs 1,2,4 from one of the protomers and motif 3 of a second protomer, motif 5 from the third protomer closes the active site upon substrate binding [11,12]. As dUTP hinders dUTPase:Stl complex formation, it was suggested that Stl binds to dUTPase in an open, substrate-free conformation, while it is unable to bind to dUTPase when the binding pocket is in closed conformation [6]. It has also been shown that motif 5 has negligible contribution to the protein–protein interaction in case of φ11 phage dUTPase while it has somewhat more pronounced role in case of 80α phage dUTPase [6,7]. Taking these data together it was appealing to hypothesize that the substrate binding pocket is directly involved in Stl-dUTPase interaction. Although dUTPase activity per se is not essential for SaPI mobilization there is evidence that certain mutations of specific residues in motif 4 and motif 3 influence the SaPI induction capability of 80α phage dUTPase, which argues that the dUTPase active site has key role in the complex formation with Stl [7,13].

In parallel to these studies it has also been revealed that not only the homotrimeric phage dUTPases but a homodimeric dUTPase from φNM1 phage is also capable to interact with the Stl of SaPIbov1 [14,15]. Hill et al. provided clear evidence also for the direct interaction of the φNM1 phage dUTPase and Stl [14]. This finding is surprising since homodimeric and homotrimeric dUTPases share no structural similarity: dimeric dUTPases are all-α helical proteins while trimeric dUTPases have a β-pleated 3D fold (Figure 1) [12,16,17]. The two active sites of dimeric dUTPases are built up symmetrically on the interface of the dimer by 5 motifs as one protomer provides motif 1,2,4,5 and the other donates motif 3. Although it has turned out that in case of φNM1 phage dUTPase the enzymatic activity is not essential for SaPI mobilization [15], the two structurally highly different dUTPase families have only the dUTP binding ability in common, so it is suggestive to speculate on the role of this region in Stl binding.

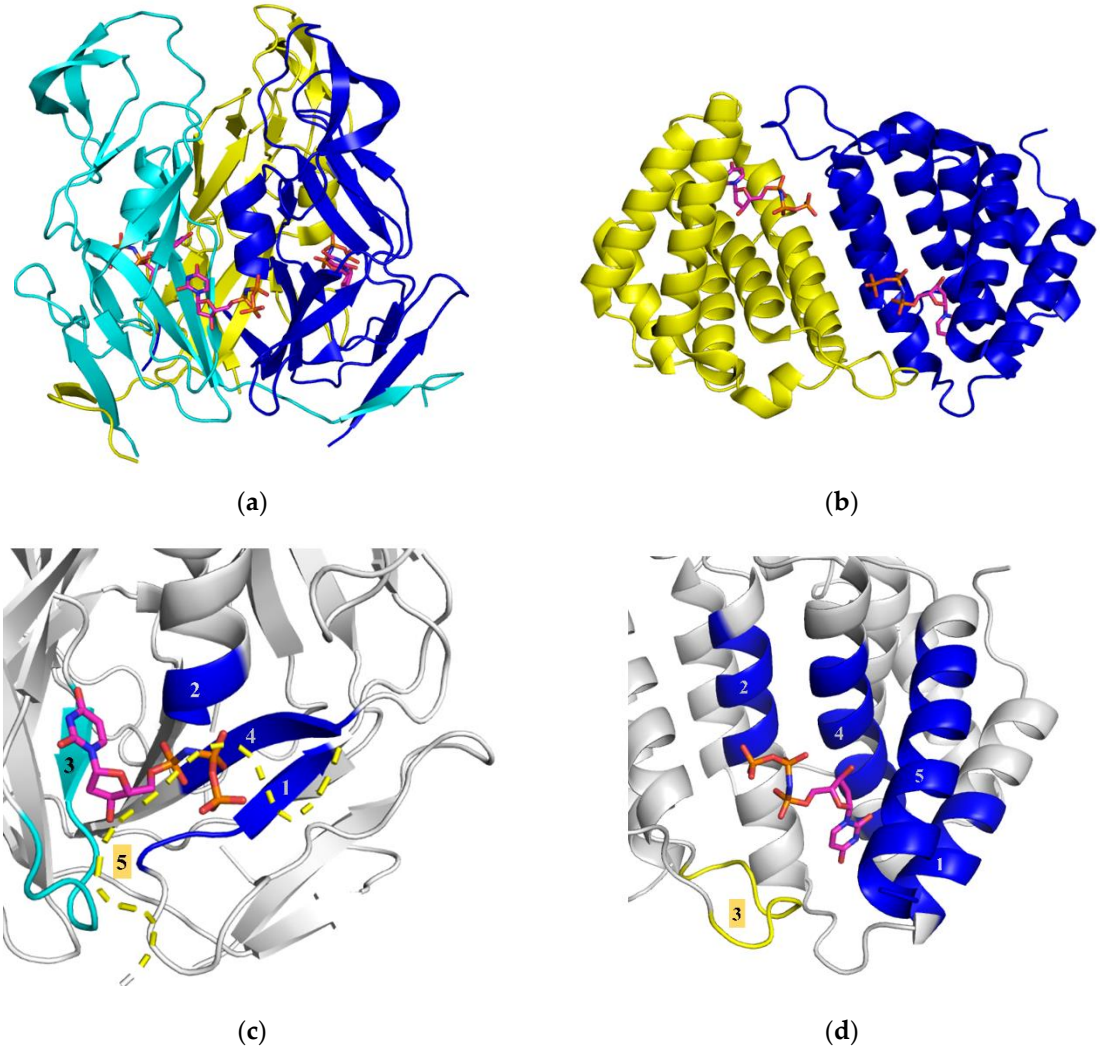


Figure 1. Comparison of homotrimeric and homodimeric phage dUTPases. (**a**) Structure of the β-pleated trimeric φ11 phage dUTPase (PDB ID 4GV8) protein is represented as cartoon, substrate analogue dUPNPP shown as sticks with atomic coloring as carbon: magenta, oxygen: red, nitrogen: blue, phosphorus: orange. (**b**) An all-α-helical dimeric dUTPase protein of φDI Staphylococcal phage (PDB ID 5MYD, [17]) is represented as cartoon, dUPNPP shown as sticks with atomic coloring as carbon: magenta, oxygen: red, nitrogen: blue, phosphorus: orange. (**c**) Close-up representing the architecture of the trimeric φ11 phage dUTPase active site (PDB ID 4GV8 [18], motif 5 is not localized in the φ11 phage dUTPase electron density map, so that is modeled based on the 80α phage dUTPase structure [PDB ID: 3ZEZ]). Conserved motifs are colored by subunits, substrate analogue dUPNPP shown as sticks with atomic coloring according to (a). The substrate binding pocket is constituted at the interface of two subunits by conserved motifs 1, 2, and 4 of one subunit and motif 3 of the other subunit. Upon dUTP hydrolysis the pocket is closed by motif 5 of the third subunit. (**d**) Close-up representing the architecture of the all-α-helical dimeric dUTPase protein of φDI Staphylococcal phage (PDB ID 5MYD, [17]). Conserved motifs are colored by subunits, substrate analogue dUPNPP shown as sticks with atomic coloring according to (a). The substrate binding pocket is located at the dimer interface.

No detailed study is yet available which investigates the peptide segments involved in the formation of the binding surface of Stl with phage dUTPases of different folds. Furthermore, the large structural differences between homotrimeric and homodimeric dUTPases poses the question whether Stl promiscuity is orchestrated by different binding peptide segments of the Stl protein or if dUTPases present the same binding surface to Stl (potentially comprising the active site where the substrate

dUTP is accommodated) despite their different folds. This question was also investigated by Bowring et al. in their study published in 2017 [19] using two truncated Stl constructs: Stl-1-175 and Stl-87-267. This latter construct with two additional residues (Stl-85-276) was first characterized and termed as Stl-C-terminal domain in our previous work published in 2015 [20]. Bowring et al. concluded that these two constructs interact differently with the trimeric φ11 phage dUTPase and the dimeric φO11 phage dUTPase. We have previously shown that the Stl C-terminal domain binds to the trimeric φ11 phage dUTPase with high affinity compared to the full-length Stl, and also strongly inhibit the enzymatic activity of the enzyme [20]. The results reported by Bowring et al. [19] were in disagreement with our previous data [20]; however, the reason for this discrepancy was not addressed in [19].

Herein we set out to explore the binding mechanism of the interaction between homodimeric and homotrimeric dUTPases and Stl. We also aim to clarify the controversy between our previous data [20] and the study of Bowring et al. [19]. Based on enzyme inhibition assays, we here show that homodimeric φNM1 dUTPase has similar affinity to Stl as the homotrimeric φ11 dUTPase. We also show that the binding of Stl to the homodimeric φNM1 dUTPase results in dissociation of the homodimer and the formation of heterodimeric Stl:dUTPase assemblies. These events may be important for dUTPase inhibition given that the active sites of this protein are located in the dUTPase dimer interface. This is markedly different from the trimeric dUTPases which interact with Stl without the change of oligomeric state. In order to provide exclusive insight to the structural details of complex formation, we performed hydrogen deuterium exchange mass spectrometry measurements. This pioneering technique can reveal information such as the change in H/D exchange rate upon complex formation [21–23]. If a decrease is observed in a specific area, that region is suggested to be directly involved in the protein–protein interaction [24,25]. Based on our hydrogen deuterium exchange mass spectrometry (HDX-MS) results, we identify regions of both Stl and homotrimeric and homodimeric dUTPase proteins which are involved for complex formation.

2. Materials and Methods

2.1. Cloning, Expression, and Purification of Proteins

The φNM1 phage dUTPase (DUTφNM1, Uniprot ID: A0EWK2, residues 2-178) was amplified from the pET21A vector provided by the courtesy of Dokland laboratory [14] using 5′-TATTGGATCCATGG CTAGCACTAACACATTAACA-3′ forward and 5′-GGTCCTCGAGTTACACGTATCCTTTTCCTGCG-3′ reverse primers and cloned to a pGEX-4T-1 vector in frame with the thrombin cleavable amino-terminal GST tag by using BAMHI and XhoI restriction sites. The resulting construct was validated by sequencing (Eurofins MWG Operon). DUTφ11 was expressed from a pET-15b plasmid created by cloning of the codon-optimized cDNA of DUTφ11 that was cloned into the vector from Novagen with *NdeI* and *XhoI* restriction sites using the services of Eurofins MWG Operon. A truncated mutant of the φ11 dUTPase lacking the phage specific insert DUTφ11$^{\Delta insert}$ and Stl were expressed from constructs designed earlier [18,26]. Sequences of the proteins are shown in Supplementary Table S1.

Proteins were expressed in *E. coli* strain BL21 Rosetta (DE3) propagated on Luria–Bertrani broth till OD_{600} = 0.6 and then induced with 5 mM isopropyl-β-D-1-thiogalactopyranoside (IPTG) for 4 h at 30 °C in case of Stl and DUTφNM1 and 37 °C for DUTφ11 and DUTφ11$^{\Delta insert}$. The cells were then harvested by centrifugation (30 min 16000*g*) and stored at −80 °C.

Purification of GST-tagged Stl and DUTφNM1 proteins were carried out as described earlier for the case of Stl [20]. Briefly, cell pellets were resuspended by Potter–Elvehjem homogenizer in 30 mL buffer A (50 mL HEPES (pH = 7.5), 200 mM NaCl) supplemented with 2 mM dithiothreitol (DTT), ca. 2 µg/mL RNase and DNase, and an EDTA-free complete ULTRA protease inhibitor tablet (Roche). The cell suspension was sonicated (4 × 60 s), and centrifuged (16000*g*, 30 min). The supernatant was loaded on a pre-equilibrated benchtop glutathione-agarose affinity-chromatography column (GE Healthcare) and then the column was washed with ten volumes of buffer A. The GST tag was removed by overnight on-column cleavage of the fusion-protein by of 80 unit thrombin (GE Healthcare) in

4 mL buffer A at 20 °C. Pure proteins (>95% as verified by SDS gel electrophoresis) were obtained in the flow-through.

Purification of DUTφ11 was carried out by NiNTA affinity chromatography and a subsequent gel filtration as the following: protein was solubilized in 50 mL lysis buffer (50 mM TRIS·HCl, pH = 8.0, 300 mM NaCl, 0.5 mM EDTA, 0.1% Triton X-100, 10 mM 2-mercaptoethanol, 5 mM benzamidine, 1 mM PMSF; ca. 2 μg/mL RNase and DNase and an EDTA-free complete ULTRA protease inhibitor tablet (Roche)). Following 4 × 60 s sonication, the supernatant from centrifugation (16,000*g*, 30 min) was applied onto a Ni-NTA column (Novagen) pre-equilibrated with lysis buffer containing 15 mM imidazole. After removing the contaminants by washing the column with ten bed volumes of low salt and high salt buffers (50 mM HEPES pH = 7.5, supplemented with 30 mM KCl or 300 mM KCl, respectively), DUTφ11 was eluted with 500 mM imidazole dissolved in low salt buffer. After elution DUTφ11 was dialyzed against buffer B (50 mM HEPES, pH = 7.5, 300 mM NaCl, 5 mM $MgCl_2$) and then gel-filtrated in buffer B on a GE Healthcare S200 Increase 10/300 (24 mL) column.Purity of the obtained protein preparation was above 95% based on SDS gel electrophoresis results.

Purification of DUTφ11$^{\Delta insert}$ was carried out as described earlier [6]. Shortly, protein was solubilized in the same way as DUTφ11 in low salt buffer (20 mM HEPES (pH = 7.5), 100 mM NaCl, 5 mM $MgCl_2$, 10 mM 2-mercaptoethanol) supplemented with 2 μg/mL RNase and DNase and one tablet of Complete ULTRA EDTA-free protease inhibitor. Supernatants were directly loaded on a Q-Sepharose column (5 mL) equilibrated with low salt buffer and eluted by applying 25 mL of a linear gradient up to 1000 mM NaCl; dUTPase appeared at 0.3–0.5 M NaCl. The second purification step was gel-filtration performed as in the case of DUTφ11. The purified DUTφ11$^{\Delta insert}$ appeared as single bands of at least 95% purity on SDS-PAGE.

All protein preparations were either used freshly or frozen in liquid nitrogen, and stored at −80 °C in small aliquots. Concentration of the proteins was determined based on the absorbance value measured at 280 nm by NanoDrop 2000 UV-Vis spectrophotometer using the extinction coefficients calculated based on amino acid composition (http://web.expasy.org/protparam) (Supplementary Table S1).

2.2. dUTPase Enzyme Activity Assay

Proton release during the transformation of dUTP into dUMP and PPi was followed using a Jasco V550 spectrophotometer at 559 nm and 293 K. Reaction mixtures contained DUTφNM1 dUTPase enzyme and Stl protein at different concentrations of 0–300 nM in 1 mM HEPES–HCl (pH = 7.5) buffer containing 5 mM $MgCl_2$, 150 mM KCl, and 40 mM phenol red pH indicator. The reaction was initialized with 30 mM dUTP after pre-incubation of proteins for 5 min. The initial velocity was determined from the slope of the first 10% of the progress curve. Quadratic binding equation was fitted to the data.

2.3. Native Gel Electrophoresis

Native gel electrophoresis was performed in 12% polyacrylamide gel. After 1-h pre-electrophoresis with constant voltages of 100 V in Tris-HCl buffer (pH = 8.7), 15 μL of the premixed samples was loaded onto the gel and electrophoresed for 1.5 h on 150 V. In order to avoid protein denaturation the apparatus was cooled on ice during procedure. Gels were stained by Page Blue protein staining solution (Thermo Fisher). Species and concentrations of monomers are indicated on Figure 2b.

2.4. Chemical Crosslinking

Stl and DUTφNM1 samples of 20 μM concentration and the dUTPase-Stl mixtures of 1:1 molar ratio (40 μM total protein concentration) were prepared and incubated for 5 min at 20 °C, then 20 mM disuccinimidyl suberate (DSS) was added to the samples, followed by a further incubation at 20 °C for 1 h. Quenching of the crosslinking reaction was performed by the addition of 5 μL 100 mM (pH = 7.5) Tris buffer to 40 μL of samples and were analyzed by SDS-PAGE on a 12% gel using Page

Ruler prestained protein ladder as a molecular weight marker. Gels were stained by Page Blue protein staining solution (Thermo Fisher).

2.5. Native Mass Spectrometry

For the mass spectrometry measurements of DUTφNM1 and the DUTφNM1:Stl complex, a commercial Waters QTOF Premier instrument equipped with an electrospray ionization source was used in positive ion mode. The mass spectra were recorded under native conditions, and the mixtures contained the proteins at concentration of 40 μM in 5 mM NH_4HCO_3 buffer solution (pH = 8.0). These conditions allow transfer of the native protein complexes to the gas phase. The capillary voltage was 2600 V, the sampling cone voltage was 128 V, and the temperature of the source was kept at 363 K. Mass spectra were obtained in the mass range of 1500–6000 *m*/*z*.

2.6. HDX-MS

HDX-MS acquisitions were performed on a Synapt G2Si HDMS coupled to an Acquity UPLC M-Class system with HDX and automation (Waters Corporation, Manchester, UK). The deuterium uptake of the DUTφ11, DUTφNM1, and Stl proteins was determined using a continuous workflow with labelling taking place at 20 °C. Each protein was solubilized in Buffer S (20 mM HEPES, 300 mM NaCl, 5 mM $MgCl_2$, pH = 7.5) to a working concentration of 10–20 μM. Deuterium labelling was initiated by diluting 5 μL of each protein sample into 95 μL of Buffer L (20 mM HEPES, 300 mM NaCl, 5 mM $MgCl_2$ in D_2O, pD = 7.1). After various incubation times, samples were quenched in Buffer Q (2.4% formic acid) at 1 °C to retard further deuteration or back-exchange and were then digested on-line with a Waters Enzymate BEH pepsin column at 20 °C. Trapping of the peptides occurred on a Waters BEH C18 VanGuard pre-column for 3 min at a flow rate of 200 μL/min in 0.1% formic acid (pH = 2.5) before being applied to a Waters BEH C-18 analytical column. Elution of the peptides was achieved using a linear gradient of Buffer E (0.1% formic acid in acetonitrile, pH = 2.5) at a flow rate of 40 μL/min. To minimize back-exchange all trapping and chromatography stages of the experiment are performed at 0.5 °C. Determination of the bound HDX profile of each protein was carried out by pre-mixing the proteins at approximately equimolar concentrations. MS data were acquired using an MSE workflow in HD mode with extended range enabled to reduce the detector saturation and maintain peak shapes. Undeuterated reference acquisitions were obtained in sextuplicate for each protein along with labelling acquisitions of 1, 10, and 100 min, which were obtained in triplicate. The MS was calibrated using NaI and MS data were obtained with lock mass correction using Leu-enkephalin.

Peptides were assigned with the ProteinLynx Global Server (PLGS) (Waters Corporation, Manchester, UK) software package, with the deuterium uptake of each assigned peptide being determined with DynamX v3.0 (Waters Corporation, Manchester, UK). Evaluation of the data fitting as well as determining the error of each dataset were performed as previously described [27]. The total Δmass of each peptide was then plotted against the residue position, allowing the generation of "Woods plots" which describe the Δmass of each peptide in the bound state [28]. The average Δmass across all peptides at each residue was then calculated. Residues with values exceeding the 99% confidence bands are noted and defined as part of the interaction surface of Stl and phage dUTPases. In all cases sequence coverage was above 90% and redundancy was above 3.

2.7. Homology Models

In case of Stl the formerly generated and validated Phyre2 model was used [20,26], 3D homology model of DUTφNM1 was created also with Phyre2 based on the crystal structure of φDI and φO11 phage dUTPases (PDB ID: 5MYD, 5MIL) [17,19,29].

3. Results and Discussion

3.1. Stl Inhibits the Enzymatic Activity of Homodimeric and Homotrimeric dUTPases with Comparable Inhibition Constant

It has been shown that the homodimeric DUTφNM1 induces the replication of SaPIbov1 through complex formation with Stl, and this complex formation also results in the decrease of enzymatic activity of the DUTφNM1 enzyme [14]. In the present study we quantitatively analyzed the inhibition of DUTφNM1 by Stl (Figure 2a). Based on steady-state enzymatic activity measurements of DUTφNM1 performed in the presence of Stl of different concentrations, we found that the maximal inhibition was about 40%, thus half of the original enzymatic activity was retained even at relatively high concentration of Stl. This markedly differs from the complete loss of dUTPase enzymatic activity observed upon homotrimeric DUTφ11-Stl complex formation within the same steady-state assay conditions [6]. This result on its own does not necessarily implicate weaker binding per se, as for example in case of competitive inhibition observed for the DUTφ11-Stl system [6], the extent of inhibitory effect on enzymatic activity is determined by the dissociation and association kinetics of both the inhibitor and the ligand. Indeed, the apparent inhibitory constant found in case of the homotrimeric DUTφ11 ($K_{i,\ app}$ = 27 ± 5 nM, c.f. [6]), is comparable to the data we obtained here for the homodimeric DUTφNM1, $K_{i,\ app}$ = 34 ± 14 nM (Figure 2a). The exact mechanism of inhibition can only be revealed by detailed transient kinetic and thermodynamic characterization of the dimerization and substrate binding of the DUTφNM1 in the presence and absence of Stl, which was beyond the scope of this study.

3.2. Mechanism of Interaction with Stl is Markedly Different between Homodimeric and Homotrimeric Phage dUTPases

The stoichiometry of the DUTφNM1-Stl complex was then investigated using various biochemical and biophysical assays to critically evaluate and expand the suggestion for the existence of a heterodimer based on the chemical crosslinking by Hill and Dokland [14]. The proteins were first characterized by native gel electrophoresis on their own or premixed (Figure 2b). Mixtures of the Stl and DUTφNM1 proteins represented different ratios of the proteins in the samples (see numbers of ratios and concentrations of monomers above the specific lanes on Figure 2b). Lanes containing the individual proteins (either DUTφNM1 or Stl on their own) show bands corresponding to the homodimeric assemblies as previously described [6,14]. Upon mixing the two proteins, a new third band clearly emerged, that was not present in the samples of the individual proteins (highlighted with an arrow on the figure). The presence of this new band argues for the formation of a DUTφNM1-Stl complex.

To confirm the complex formation and investigate its stoichiometry, the assemblies were then investigated by chemical crosslinking (Figure 2c). Previous cross linking experiments have reported a 1:1 DUTφNM1-Stl heterodimer [14], although the short spacers (ca. 5 Å) used in these experiments may have resulted in overrepresentation of these assemblies. We performed the crosslinking of the proteins by using disuccinimidyl suberate (DSS) which possesses a ca. 11 Å linker distance in order to identify any higher order complex of DUTφNM1 and Stl. SDS-PAGE analysis of the individual proteins after crosslinking resulted in the presence of two bands for each protein with molecular weights corresponding to those expected for the monomeric and dimeric proteins (Figure 2c). In the premixed samples containing a mixture of DUTφNM1 and Stl in 1:1 molar ratio a unique band is present with a molecular weight (ca. 55 kDa) consistent with that expected for a DUTφNM1:Stl heterodimer in accordance with the native gel electrophoresis experiments. As we have not found any other assembly of higher molecular weight we also concluded that the DUTφNM1-Stl complex is likely to consist of one monomer of Stl and one monomer of DUTφNM1.

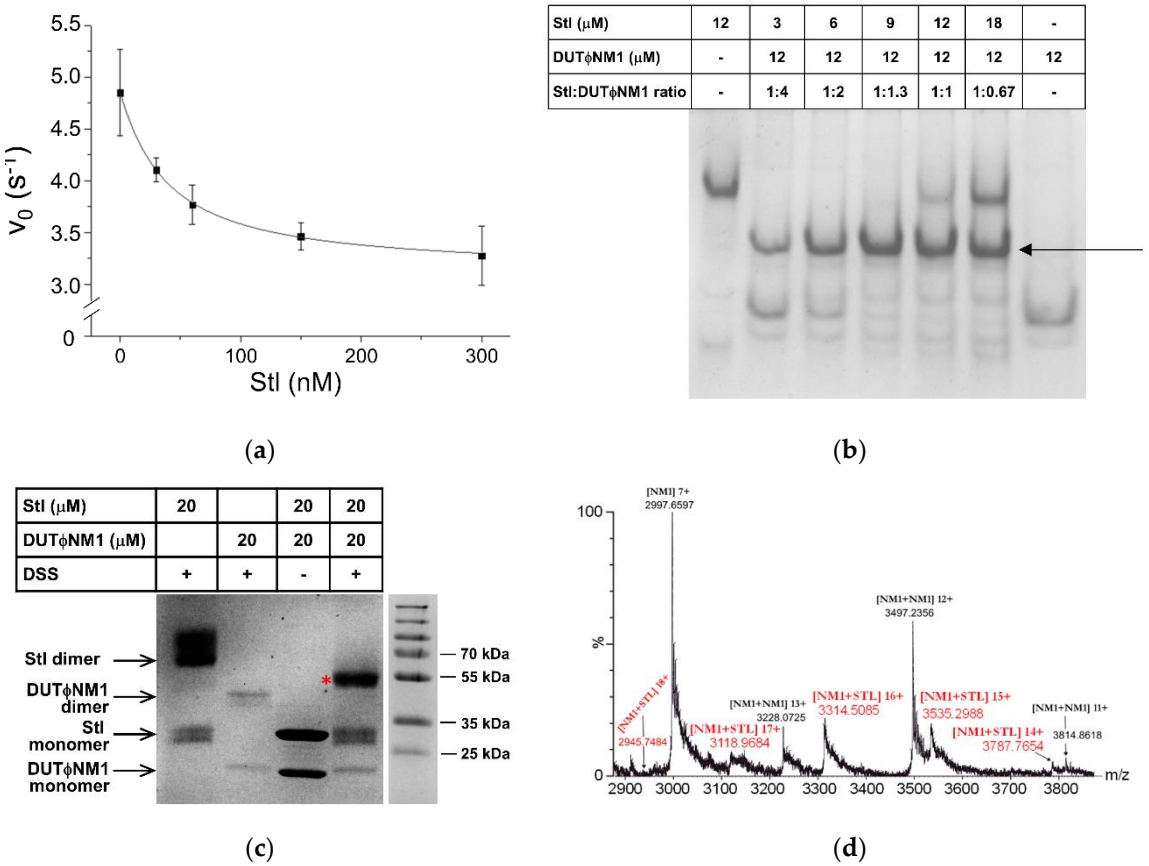


Figure 2. Biophysical characterization of φNM1 phage dUTPase-Stl interaction. (**a**) Enzyme activity of homodimeric DUTφNM1 in the presence and absence of Stl (steady-state conditions). The maximal extent of inhibition was about 40%, the binding is characterized by the apparent inhibitory constant of $K_{i,\,app}$ = 34 ± 14 nM. (**b**) Native gel electrophoresis of Stl, DUTφNM1 and the mixture of the two proteins of various molar ratios, species, and concentrations of monomers are indicated in the figure. Note the emerging band (indicated by arrow) in the samples containing the mixture of the two proteins that suggests the formation of an Stl-DUTφNM1 heterodimer, situated between the bands of the individual dimeric proteins. (**c**) SDS-PAGE analysis of Stl, DUTφNM1, and their mixture after chemical crosslinking induced by the reagent disuccinimidyl suberate (DSS). The mixture of the untreated proteins was also loaded on the gel as a control. The band corresponding to the molecular mass of the DUTφNM1-Stl heterodimer is denoted with a red star. (**d**) The native mass spectrum of the DUTφNM1-Stl mixture. Peak series with *m*/*z* values of 2946 (18+), 3119 (17+), 3315 (16+), 3535 (15+), 3788 (14+) (highlighted in red) indicate the presence of an assembly associated with the molar mass of 53020 ± 7 Da, which corresponds to a 1:1 complex of Stl (32.0 kDa) and DUTφNM1 (21.0 kDa), constituting the DUTφNM1-Stl heterodimer (designated as "NM1+Stl" on the figure). Peaks corresponding to DUTφNM1 monomer (NM1) and homodimer (NM1+NM1) are also present in the spectrum.

The 1:1 composition of the DUTφNM1–Stl complex was also confirmed by native mass-spectrometry measurements (Figure 2d, Figure S1). In the spectrum, the monomer form of the DUTφNM1 protein was the most abundant showing a Gaussian-like distribution of the *m*/*z* values 1907.934 (+11), 2098.611 (+10), 2331.718 (+9), 2623.025 (+8), 2997.660 (+7), 3497.052 (+6), 4196.261 (+5), 5245.075 (+4) (Figure S1a). The mass calculated from these MS peaks is 20976 Da, which corresponds to the molar mass of 20976 Da calculated based on amino acid composition of the protein (http://web.expasy.org/protparam) (Supplementary Table S1). Although less abundant, still the peaks corresponding to a dimer of DUTφNM1 were also observable in the spectrum as series of peaks with *m*/*z* 3228.073 (+13), 3497.235 (+12), 3814.862 (+11) associated with the molar mass of 41950 Da, well agreeing with the expected 41952 Da for a homodimer of DUTφNM1 (Figure S1a). New peaks of *m*/*z* 2945.7484 (+18), 3118.9684 (+17),

3314.5295 (+16), 3533.9619 (+15), 3787.103 (+14), 4078.3411 (+13), 4418.1189 (+12), 4819.6745 (+11) were also emerging in the spectra of the mixture of the DUTφNM1 and Stl. These peaks are associated with the molar mass of 53020 Da. This clearly shows that the complex consists of one monomer Stl (32016 Da) and one monomer DUTφNM1 (20976 Da), i.e., 1:1 stoichiometry is observed.

In contrast to this, the stoichiometry of the complex resulting from the interaction of the homotrimeric DUTφ11 with Stl was found to be 3:2 or 3:3 dUTPase: Stl [6,8], suggesting that DUTφ11 remains in homotrimeric oligomeric state upon complex formation. Small-angle X-ray scattering studies suggested that in the trimeric human dUTPase-Stl complex, Stl is present in monomers [26]. Here we observe that in the interaction of the DUTφNM1 protein with Stl, both proteins dissociate into monomers and form a heterodimer. These results can provide a possible explanation on the mechanism of enzymatic inhibition of the DUTφNM1 protein, since in case of DUTφNM1, the active site is located at the dimer interface of the protein (cf. Figure 1), which is likely affected by the Stl binding. It has been suggested that Stl as other similar repressors binds to its cognate DNA site as a dimer. Complex formation with either DUTφNM1 of DUTφ11 involves monomers of Stl (cf. [6,7,26] and present work), which provides a potential model for the perturbation of the Stl-DNA complex through dissociation of the repressor dimers to interact with the derepressor in both types of complexes.

3.3. dUTPase Active Sites are Directly Involved in the Complex Formation with Stl

The protein surfaces responsible for the interaction between DUTφNM1:Stl and DUTφ11:Stl complexes were investigated using hydrogen deuterium exchange mass spectrometry (HDX-MS), similarly as in a previous study [26]. This method reports on a protonated protein's time-dependent uptake of deuterium when dissolved in a fully deuterated solvent, in which changes can be localized to peptide units across the protein backbone. In binding assays, HDX-MS outputs are typically reported by changes in the rate of isotope uptake (Δmass) between unbound and bound protein complexes, yielding characteristic difference plots that provide unique insight into protein–protein interfaces [23]. In the present study, we determined HDX-MS data for all proteins either in isolation or in premixed samples and then difference plots were prepared by subtraction of the uptake data obtained for the bound protein complexes from those obtained for the proteins in their unbound conformations (cf. Methods).

The HDX-MS difference plots of both dimeric and trimeric dUTPases exhibited large changes in isotope uptake in the presence of Stl consistent with the binding of the inhibitor to these proteins (Figures 3 and 4). In the case of DUTφ11, the most conspicuous Δmass data were observed for peptides that spanned most of the active site segments as well as the phage specific insert of the protein (Figure 3, peptide numbering is shown on Figure S2). This indicates the direct involvement of DUTφ11 active site in the complex formation with Stl (cf. also Supplementary Figure S3) and is consistent with previous work showing that the dUTP substrate and the Stl compete for the same binding site [6]. Similar conclusions were drawn for the trimeric human dUTPase by HDX-MS [26] and for the trimeric dUTPase from *E. coli* by mutational analysis [30]. So that it seems that Stl may have a uniform binding mode to the trimeric dUTPases. As it has been shown that the phage specific insert is not essential for the binding of DUTφ11 to Stl [8], possibly the interaction of the residues of the insert with Stl is the consequence of binding of the inhibitor protein to the active site of DUTφ11. This hypothesis has been reinforced by the HDX-MS data obtained for a truncated mutant, which lacks the phage specific insert, DUTφ11$^{\Delta insert}$ (Figure S2 and S4–S6). In the experiments with this mutant protein and Stl, the large Δmass data for the dUTPase active site segments were clearly preserved (cf. Figure S5b).

It is important to note that no significant Δmass was observed for the DUTφ11 conserved motif 5. This finding is also in agreement with the previous results indicating that the truncated mutant of the DUTφ11 protein lacking motif 5 was capable of binding to Stl with similar affinity as that of the full length, wild-type protein [6,7].

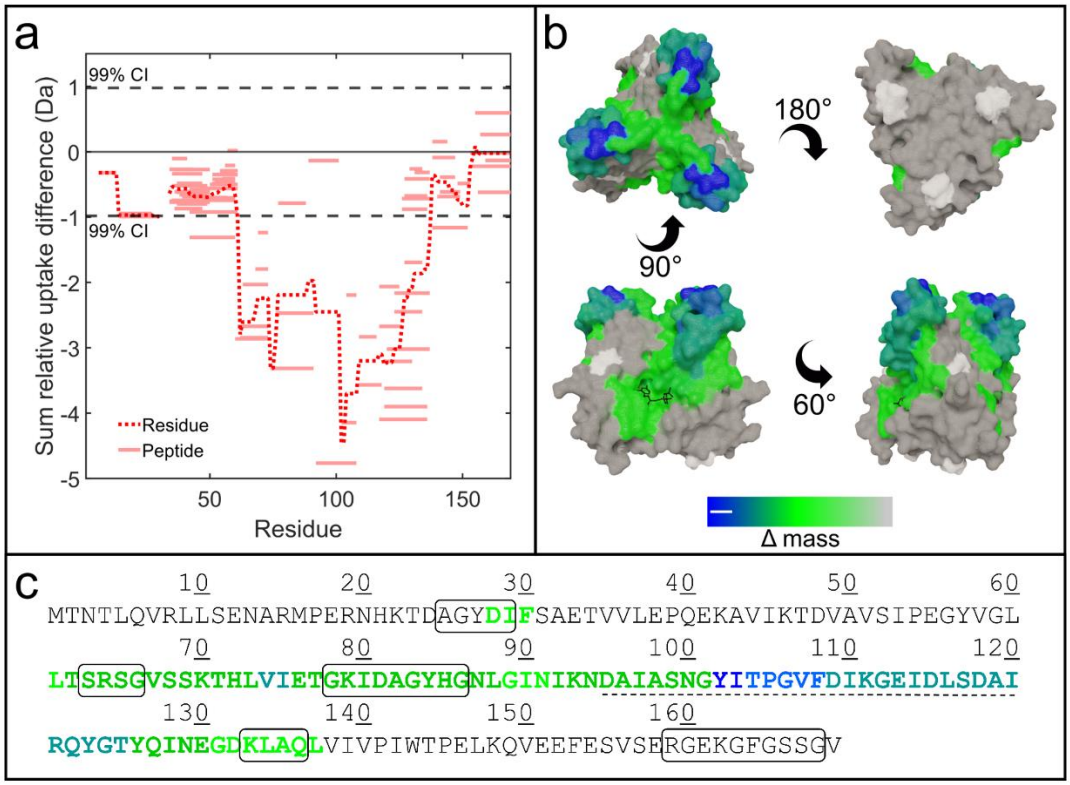


Figure 3. Hydrogen deuterium exchange mass spectrometry (HDX-MS) results for DUTφ11 upon mixing with Stl. (**a**) HDX difference plot showing the Δmass of each peptide (solid lines) as well as the average amount per residue (dotted line). The dashed lines represent the 99% confidence bands evaluated over the whole dataset. (**b**) X-ray crystal structure (PDB ID 4GV8 [23]) of DUTφ11 colored according to HDX-MS difference data following the color gradient shown at the bottom of the panel; substrate analogue dUPNPP is also shown as black sticks in order to ease visualization of the active sites (views: top, bottom, sides). Regions which could not be probed by HDX-MS are shown in white. (**c**) Sequence of the DUTφ11 is shown, where the conserved active site building motifs are boxed and the phage specific insert is underlined with a dashed line. Letter coloring is according to HDX difference data.

The HDX-MS results for the DUTφNM1–Stl complexes revealed significantly more complex binding interaction as compared with the outputs of DUTφ11-Stl (Figure 4, peptide numbering is shown on Figure S6). Peptides spanning residues 15–36 and 155–171 of DUTφNM1 show significant decreases in Δmass consistent with the binding of Stl and occlusion of these sites from isotope exchange. These regions contain active site residues of potential key-importance in enzymatic activity including Q17, D21 (residues of motif 1 responsible for uracil binding), K159 and R166 (residues of motif 5 responsible for phosphate binding) as determined by sequence alignment against a dimeric dUTPase with detailed study on the active site [31], and dimeric phage dUTPase structures [17,19] (Figure 4b). This suggests that part of the active site is directly involved in the protein–protein complex formation, as these residues of the protein become less accessible to the solvent upon complex formation (cf. also Supplementary Figure S2). However, we also note here that specific mutation of K159 to alanine did not abolish DUTφNM1:Stl complex formation either in vitro or in vivo, thus this residue is not an essential factor in the protein–protein interaction [15].

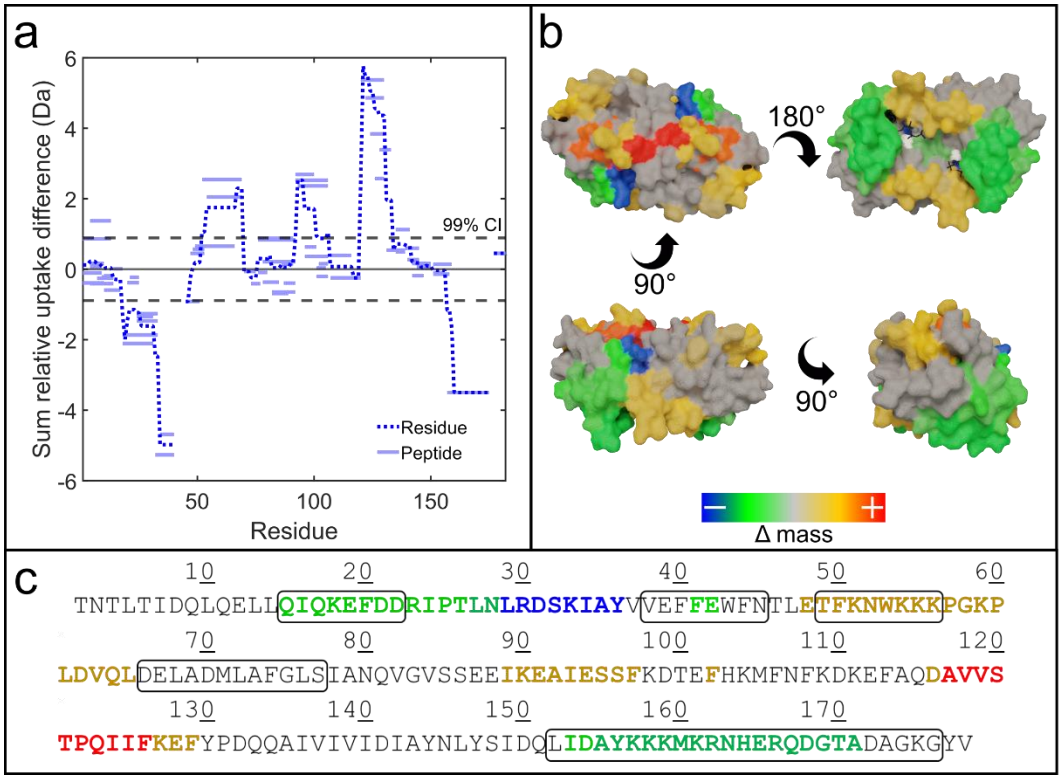


Figure 4. HDX-MS results for DUTφNM1 upon mixing with Stl. (**a**) HDX difference plot showing the Δmass of each peptide (solid lines) as well as the average amount per residue (dotted line). The dashed lines represent the 99% confidence bands evaluated over the whole dataset. (**b**) Homology model of DUTφNM1 colored according to HDX-MS difference data following the color gradient shown at the bottom of the panel; substrate analogue dUPNPP is also shown as black sticks in order to ease the visualization of the active sites (views: top, bottom, sides). Regions which could not be probed by HDX-MS are shown in white. (**c**) Sequence of the DUTφNM1 protein. Conserved active site building motifs are boxed, letter coloring is according to HDX difference data.

In the presence of Stl, the DUTφNM1 protein also shows extended regions of positive Δmass including peptides covering residues 48–65, 89–97, and 116–129. The results potentially indicate that these protein regions may become more solvent accessible in the presence of Stl presumably as a consequence of dimer dissociation as observed in the crosslinking assay, or undergo other conformational changes. To better understand the significance of these results, the HDX-MS outputs were mapped onto a 3D homology model of DUTφNM1 dimer (Figure 4c). According to the analysis of residue–residue interactions across the dimer interface of the in silico 3D model performed with DIMPLOT [32], residues 45–58 and 111–124 are located on the dimer interface. Thus, the increase in H/D exchange rate detected in these regions might correspond to the increased solvent accessibility of these segments because of the dissociation of DUTφNM1 homodimer upon complex formation with Stl. All in all, the native gel electrophoresis, chemical crosslinking, enzyme activity, native MS, and HDX-MS experiments all support the formation of a heterodimer complex constituting DUTφNM1 and Stl.

It has been formerly proposed that a short segment with common sequence (GVSS) of the DUTφ11 and DUTφNM1 might have a role in complex formation [15] (cf. also sequence alignment on Supplementary Figure S9). Although this segment of DUTφ11 (residues 66–69) showed significant decrease in H/D exchange rate, the same segment of DUTφNM1 (residues 83–86) did not present significant HDX signal upon complex formation, arguing against this hypothesis. Further analysis is required to decide on the potential role of this segment in complex formation.

3.4. Different Regions of Stl Mediate the Promiscuity of this Protein for dUTPase Binding

We next compared the different plots of Stl in the presence of the dUTPases representing the homodimeric and homotrimeric families in order to better understand the interesting capability of Stl to interact with both types of dUTPases (cf. Figure 5, peptide numbering is shown on Figures S4 and S8). These experiments were also expected to shed light on the contradiction between the data published by Nyiri et al. in 2015 [20] and Bowring et al. in 2017 [19] for interaction of Stl-C-terminal domain (Stl-85-267) with DUTφ11. The HDX-MS outputs revealed dramatically different Δmass profiles for Stl, depending on which dUTPase is added. In the presence of DUTφ11, Stl exhibits significant negative mass shifts across the protein backbone. This suggests that the binding of Stl to DUTφ11 induces a global conformational tightening of the protein and also implies that Stl protein has a larger conformational space in the absence of dUTPase. In addition to these global changes in protein conformation, Stl also displays a dramatically pronounced negative mass shift localized to the protein region of residues ca. 98Y – 113Y (Stl-98-113). This suggests that this region plays a major role in the interaction of Stl with DUTφ11. Since this tyrosine-rich region is part of the Stl-85-267 (termed as Stl-C-terminal domain (or Stl-87-267, termed as $Stl^{\Delta HTH}$) truncated constructs, these HDX-MS data are consistent with our previous results which showed that this truncated Stl construct lacking the N-terminal 84 residues is fully capable of binding to and inhibiting the enzymatic activity of the DUTφ11 enzyme [20]. It is also of interest to point out that the very same peptide segment was also shown to be similarly involved in the interaction of Stl with human dUTPase, another trimeric dUTPase [26], strengthening the role of this segment in interactions with representatives from this family of dUTPases.

As we also showed earlier, the N-terminal 84 residues contains the DNA-binding helix-turn-helix motif of the repressor protein [20]. So the results presented in this study reinforce the suggestion that DNA-binding and protein–binding functions of Stl can be associated with different segments of the protein.

In case of the DUTφNM1-Stl complex, peptides covering the region of the N-terminal 200 residues of Stl were identified with no significant HDX change, however peptides from the 60 residue-long segment situated at the very C-terminal part of the Stl sequence showed pronounced negative signal. Within the region, the segment of Stl-227-247 shows the largest shifts. On the one hand, these results are consistent with the former finding that truncation of the N-terminal 84 residues did not perturb the complex formation between Stl and DUTφNM1 [15] or Stl and another dimeric phage dUTPase from phage φO11 [19]. On the other hand, these results also clearly delineate the different segments of the Stl-C-terminal domain that are used by the repressor protein for complex formation with DUTφ11 and human dUTPase [26] (homotrimeric dUTPase family) or DUTφNM1 (homodimeric dUTPase family).

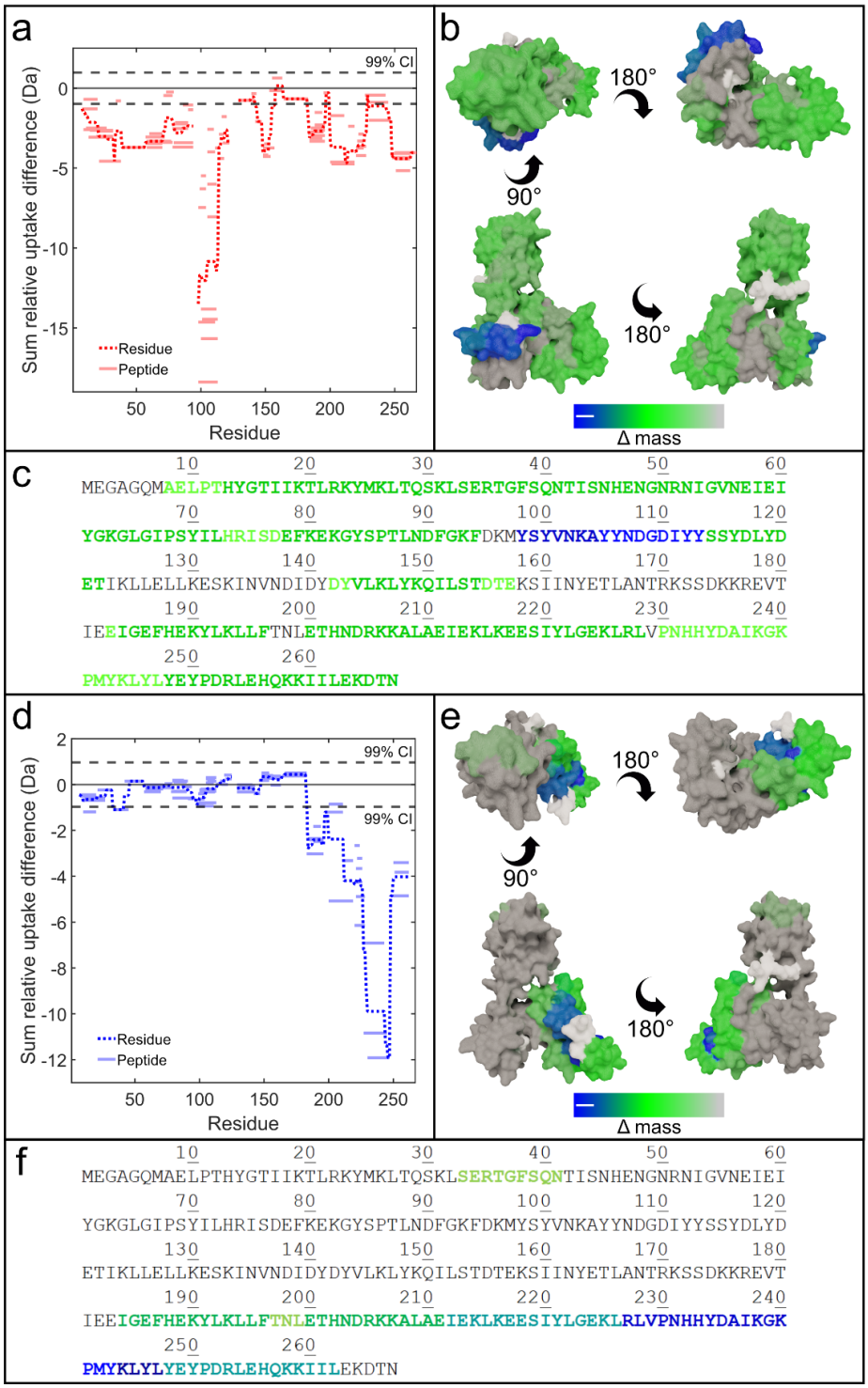


Figure 5. Experimental results and schematic models showing how the Stl repressor protein of SaPIbov1 interacts with drastically different families of phage dUTPases. (**a**,**d**) HDX difference plots showing the Δmass of each peptide (solid lines) as well as the average amount per residue (dotted lines) of Stl upon mixing with DUTφ11 (a) and DUTφNM1 (d). The dashed lines represent the 99% confidence bands evaluated over the whole dataset. (**b**,**e**) Homology model of Stl colored according to the HDX-MS difference data obtained upon binding to DUTφ11 (b) and DUTφNM1 (e), following the color gradient shown at the bottom of the panel (views: top, bottom, sides). Regions which could not be probed by HDX-MS are shown in white. (**c**,**f**) Sequence of Stl is shown with letter coloring according to HDX difference data when bound to DUTφ11 (c) and DUTφNM1 (f).

4. Conclusions

Altogether, our kinetic, cross-linking, native mass spectrometry, and HDX-MS experiments suggest previously unreported functional plasticity of *S. aureus* pathogenicity island repressor protein Stl and revealed new details for a better understanding of the different binding mechanisms of Stl for the two different phage dUTPases. The HDX-MS experiments shown here suggested highly different interaction surface of Stl with the dimeric DUTφNM1 and trimeric DUTφ11 dUTPases. Native mass spectrometry data here and in earlier papers [6,26,33] provided direct experimental data for the distinct Stl-binding mechanisms of the two dUTPase families. The two families of dUTPases have evolved separately and constitute drastically different protein folds and active site architecture, as reviewed in [12,34]. It is of interest to consider that Staphylococcal phages encode dUTPase representatives from both families (cf. [6]and [19]) such that the presence of a dUTPase is a conserved character of these phages.

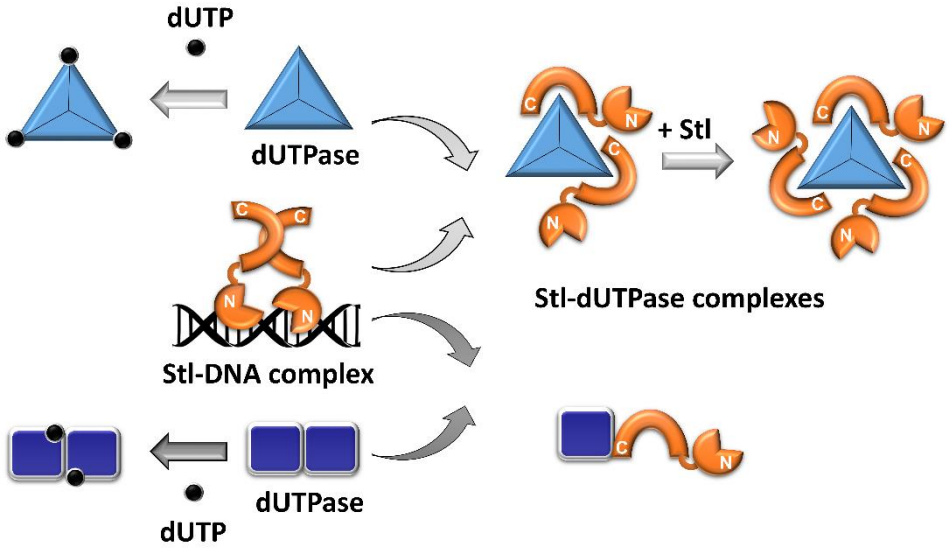
Why do phages encode their own dUTPases? A reason for this may arise from the interesting lack of endogenous dUTPase from *S. aureus* strains [35]. Since dUTPase is important for preventive DNA repair [11], phages may increase their chances by encoding their own copy of this important enzyme, either from the trimeric or from the dimeric dUTPase family. SaPIs on the other hand have evolved to rely on phages for their life cycle. The Stl repressor of SaPIbov1 has been adapted to interact with both types of dUTPases that will function as derepressors of Stl function, allowing induction of the pathogenicity island mobile genetic elements (SaPIs). In this case, the conserved presence of dUTPases within the phages is also profitable for SaPIbov1. The fact that HDX-MS data showed that different segments of this Stl repressor target phages encoding different dUTPases underlines the suggestion that mobile genetic elements may gain benefit from conditions of low dUTP levels to ensure uracil-free DNA environment. One possible advantage is that the low dUTP level, provided by the dUTPase enzymatic action, enhances the fidelity of SaPI replication via diminution of the mutation rate [36], while reducing the potential for the selective evolution of phages to escape SaPI interference in parallel.

It is especially interesting that even if integrated prophages contain the dUTPase gene, the expression of the protein is most likely being repressed [6]. In some of the *S. aureus* strains a specific inhibitor protein of the uracil DNA glycosylase, namely SaUGI presumably moderates uracil excision, while the survival of other strains is yet unexplained [35,37,38]. It has also been demonstrated that the prophage free *S. aureus* RN450 strain, which does not contain the dUTPase gene possesses elevated genomic uracil content compared to *dut*+ *ung*+ bacteria [35]. It seems likely that mobile genetic elements may encode either dUTPase or SaUGI to escape the damaging uracil-DNA repair, which can impair their horizontal gene transfer [9]. In addition to this, uracil content of a mobile genetic element might also prevent its integration into the genome of the new host, as it was recently demonstrated in the case of human immunodeficiency virus [39,40]. It is also tempting to hypothesize that some SaPIs recognize phages through dUTPase-Stl interaction, in order to ensure their uracil-free replication [6]. This could have a dual advantage: i) Replicated SaPI genomes are not fragmented by the host repair mechanism, ii) mutation rate of phages is not elevated by the damaging base excision repair (BER), which hinder their ability to escape the SaPI interference.

Based on our results Scheme 1 shows a schematic model, which describes the interaction of Stl with dimeric and trimeric dUTPases.

Stl protein dimerizes in solution and based on the similarity with other repressors and the symmetry of the specific binding site of the protein within the SaPI DNA, it is assumed that Stl binds to DNA as dimers (Scheme 1) [26,41]. Interaction of Stl monomers with dUTPases perturbs the dimerization of the repressor, hence it leads to the dissociation of the Stl-DNA complex. Trimeric DUTφ11 dUTPase can form DUT_3Stl_2 and DUT_3Stl_3 complexes with Stl, while in the case of dimeric enzyme DUTφNM1 the complex is a DUT-Stl heterodimer (Scheme 1) [6,9,26,33]. Based on our results Stl binds directly to the active site of trimeric dUTPases and it acts as a competitive inhibitor of these enzymes [6]. As also presented herein, Stl also reduces the enzymatic activity of dimeric dUTPases, although via a mechanism somewhat different from that observed for the trimeric enzymes. We show

direct evidence from native mass spectrometry that inhibition of dimeric dUTPases by Stl during complex formation between the two proteins results from perturbation of the active site architecture, which resides at the dimer interface of the enzyme.



Scheme 1. Interaction of the Stl protein with dimeric and trimeric dUTPases. Stl (orange, N-terminus is denoted with letter N, C-terminal with letter C) forms dimers in solution and may bind to DNA (black) as dimers. Perturbation of Stl dimerization by dUTPases leads to the dissociation of the Stl-DNA complex. Stl inhibits both the trimeric (light blue triangles) and dimeric dUTPases (dark blue rectangles). The inhibition is based on competition between Stl and the substrate, dUTP (black dots) in case of the trimeric dUTPases. Stl monomers and dUTPase trimers form DUT_3Stl_2 and DUT_3Stl_3 complexes. We found that the region of residues ca. 98Y – 113Y of Stl protein has a major contribution in the interaction of Stl with trimeric dUTPases (cf. also [26]). The substrate binding site of dimeric dUTPases, which is located at the dimerization surface of the enzyme, is impaired upon formation of a heterodimeric complex of the dUTPase with Stl. This explains the reduction of the enzymatic activity of dimeric dUTPases in the presence of Stl. We found that peptides from the 60 residue-long segment situated at the very C-terminal part of the Stl sequence play a key role in the heterodimer formation.

Supplementary Materials: The following are available online at http://www.mdpi.com/2218-273X/9/9/488/s1, Figure S1: Annotated mass spectra of the mixture of the DUTφNM1 (NM1) with Stl (STL) protein (a) and Stl protein (b) measured under native electrospray conditions. Figure S2: Coverage map of HDX-MS difference plots for DUTφ11 and DUTφ11$^{\Delta insert}$ upon complex formation with Stl. Figure S3: Active sites of DUTφ11 and DUTφNM1 colored according to HDX-MS difference data obtained upon complex formation of those with Stl protein. Figure S4: Coverage map of HDX-MS difference plots for Stl upon complex formation with DUTφ11 and DUTφ11$^{\Delta insert}$. Figure S5: HDX-MS difference data obtained for DUTφ11 and DUTφ11$^{\Delta insert}$ upon complex formation with Stl. Figure S6: HDX-MS difference data obtained for Stl upon complex formation with DUTφ11 and DUTφ11$^{\Delta insert}$. Figure S7: Coverage map of HDX-MS difference plots for DUTφNM1 upon complex formation with Stl. Figure S8: Coverage map of HDX-MS difference plots for Stl upon complex formation with DUTφNM1. Figure S9: Sequence alignment of φ11 phage trimeric dUTPase and φNM1 phage dimeric dUTPase generated by NPS@ server. Table S1: Sequences of protein constructs used.

Author Contributions: Conceptualization: K.N., B.G.V.; data curation: K.N., M.J.H., J.M., O.O.; formal analysis: K.N.; funding acquisition: A.J.B., K.V. and B.G.V.; investigation: K.N., M.J.H., J.M., O.O.; methodology: K.N., M.J.H., O.O., A.J.B., B.G.V.; project administration: K.N. and B.G.V.; resources: A.J.B., K.V., and B.G.V.; software: K.N. and M.J.H.; supervision: A.J.B. and B.G.V.; validation: K.N., M.J.H., O.O., A.J.B., B.G.V.; visualization: K.N., M.J.H., O.O., A.J.B., B.G.V.; writing—original draft, K.N., M.J.H., A.J.B., B.G.V.; writing—review and editing: K.N., M.J.H., A.J.B., B.G.V.

Note Added in Proof: After submission of our manuscript a relevant paper has been published (R. Ciges-Tomas et al. *Nat. Commun.* 2019, *10*, 3676.), where using different independent methods similar results were reached reinforcing the conclusions of both studies.

Funding: This work was supported by the National Research, Development and Innovation Office of Hungary (K119993 to KV, K119493, NVKP_16-1-2016-0020, 2017-1.3.1-VKE-2017-00002, 2017-1.3.1-VKE-2017-00013, VEKOP-2.3.2-16-2017-00013 NKP-2018-1.2.1-NKP-2018-00005 to BGV,), and the BME-Biotechnology FIKP grant of EMMI (BME FIKP-BIO). The work of KN was supported through the New National Excellence Program of the Ministry of Human Capacities [ÚNKP-16-3_VBK-038]. A. Borysik acknowledges the financial support from the Royal Society sponsor reference RG150222. M. Harris is a Biotechnology and Biological Sciences Research Council – industrial Collaborative Awards in Science and Engineering (BBSCR/iCASE) funded postgraduate student with industrial support from Waters Corporation.

Acknowledgments: The authors thank Terje Dokland and Rosanne L. L. Hill for their generous help by providing a plasmid encoding the φNM1 dUTPase.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Fitzgerald, J.R. Human origin for livestock-associated methicillin-resistant *Staphylococcus aureus*. *MBio* **2012**, *3*, e00082-12. [CrossRef] [PubMed]
2. Juhas, M. Horizontal gene transfer in human pathogens. *Crit. Rev. Microbiol.* **2013**, *7828*, 101–108. [CrossRef] [PubMed]
3. Novick, R.P.; Christie, G.E.; Penadés, J.R. The phage-related chromosomal islands of Gram-positive bacteria. *Nat. Rev. Microbiol.* **2010**, *8*, 541–551. [CrossRef] [PubMed]
4. Lindsay, J.A.; Ruzin, A.; Ross, H.F.; Kurepina, N.; Novick, R.P. The gene for toxic shock toxin is carried by a family of mobile pathogenicity islands in *Staphylococcus aureus*. *Mol. Microbiol.* **1998**, *29*, 527–543. [CrossRef] [PubMed]
5. Tormo-Más, M.A.; Mir, I.; Shrestha, A.; Tallent, S.M.; Campoy, S.; Lasa, I.; Barbé, J.; Novick, R.P.; Christie, G.E.; Penadés, J.R. Moonlighting bacteriophage proteins derepress staphylococcal pathogenicity islands. *Nature* **2010**, *465*, 779–782. [CrossRef] [PubMed]
6. Szabó, J.E.; Németh, V.; Papp-Kádár, V.; Nyíri, K.; Leveles, I.; Bendes, Á.Á.; Zagyva, I.; Róna, G.; Pálinkás, H.L.; Besztercei, B.; et al. Highly potent dUTPase inhibition by a bacterial repressor protein reveals a novel mechanism for gene expression control. *Nucleic Acids Res.* **2014**, *42*, 11912–11920. [CrossRef] [PubMed]
7. Maiques, E.; Quiles-Puchalt, N.; Donderis, J.; Ciges-Tomas, J.R.; Alite, C.; Bowring, J.Z.; Humphrey, S.; Penadés, J.R.; Marina, A. Another look at the mechanism involving trimeric dUTPases in *Staphylococcus aureus* pathogenicity island induction involves novel players in the party. *Nucleic Acids Res.* **2016**, *44*, 5457–5469. [CrossRef]
8. Nyíri, K.; Papp-Kádár, V.; Szabó, J.E.; Németh, V.; Vértessy, B.G. Exploring the role of the phage-specific insert of bacteriophage Φ11 dUTPase. *Struct. Chem.* **2015**, *26*, 1425–1432. [CrossRef]
9. Hirmondó, R.; Szabó, J.E.; Nyíri, K.; Tarjányi, S.; Dobrotka, P.; Tóth, J.; Vértessy, B.G. Cross-species inhibition of dUTPase via the Staphylococcal Stl protein perturbs dNTP pool and colony formation in Mycobacterium. *DNA Repair* **2015**, *30*, 21–27. [CrossRef]
10. Frígols, B.; Quiles-Puchalt, N.; Mir-Sanchis, I.; Donderis, J.; Elena, S.F.; Buckling, A.; Novick, R.P.; Marina, A.; Penadés, J.R. Virus satellites drive viral evolution and ecology. *PLoS Genet.* **2015**, *11*, e1005609. [CrossRef]
11. Vértessy, B.G.; Tóth, J. Keeping uracil out of DNA: Physiological role, structure and catalytic mechanism of dUTPases. *Acc. Chem. Res.* **2009**, *42*, 97–106. [CrossRef] [PubMed]
12. Nagy, G.N.; Leveles, I.; Vértessy, B.G. Preventive DNA repair by sanitizing the cellular (deoxy)nucleoside triphosphate pool. *FEBS J.* **2014**, *281*, 4207–4223. [CrossRef] [PubMed]
13. Tormo-Más, M.Á.; Donderis, J.; García-Caballer, M.; Alt, A.; Mir-Sanchis, I.; Marina, A.; Penadés, J.R. Phage dUTPases control transfer of virulence genes by a proto-oncogenic G protein-like mechanism. *Mol. Cell* **2013**, *49*, 947–958. [CrossRef] [PubMed]
14. Hill, R.L.L.; Dokland, T. The Type 2 dUTPase of bacteriophage φNM1 initiates mobilization of *Staphylococcus aureus* bovine pathogenicity island 1. *J. Mol. Biol.* **2016**, *428*, 142–152. [CrossRef] [PubMed]
15. Hill, R.L.L.; Vlach, J.; Parker, L.K.; Christie, G.E.; Saad, J.S.; Dokland, T. Derepression of SaPIbov1 is independent of φNM1 type 2 dUTPase activity and is inhibited by dUTP and dUMP. *J. Mol. Biol.* **2017**, *429*, 1570–1580. [CrossRef] [PubMed]

16. Nyíri, K.; Vértessy, B.G. Perturbation of genome integrity to fight pathogenic microorganisms. *Biochim. Biophys. Acta Gen. Subj.* **2017**, *1861*, 3593–3612. [CrossRef]
17. Donderis, J.; Bowring, J.; Maiques, E.; Ciges-Tomas, J.R.; Alite, C.; Mehmedov, I.; Tormo-Mas, M.A.; Penadés, J.R.; Marina, A. Convergent evolution involving dimeric and trimeric dUTPases in pathogenicity island mobilization. *PLoS Pathog.* **2017**, *13*, e1006581. [CrossRef]
18. Leveles, I.; Németh, V.; Szabó, J.E.; Harmat, V.; Nyíri, K.; Bendes, A.A.; Papp-Kádár, V.; Zagyva, I.; Róna, G.; Ozohanics, O.; et al. Structure and enzymatic mechanism of a moonlighting dUTPase. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2013**, *69*, 2298–2308. [CrossRef]
19. Bowring, J.; Neamah, M.M.; Donderis, J.; Mir-Sanchis, I.; Alite, C.; Ciges-Tomas, J.R.; Maiques, E.; Medmedov, I.; Marina, A.; Penadés, J.R. Pirating conserved phage mechanisms promotes promiscuous staphylococcal pathogenicity island transfer. *Elife* **2017**, *6*, e26487. [CrossRef]
20. Nyíri, K.; Kőhegyi, B.; Micsonai, A.; Kardos, J.; Vértessy, B.G. Evidence-based structural model of the Staphylococcal repressor protein: Separation of functions into different domains. *PLoS ONE* **2015**, *10*, e0139086. [CrossRef]
21. Zhou, B.; Zhang, Z.-Y. Application of hydrogen/deuterium exchange mass spectrometry to study protein tyrosine phosphatase dynamics, ligand binding, and substrate specificity. *Methods* **2007**, *42*, 227–233. [CrossRef] [PubMed]
22. Ling, J.M.L.; Silva, L.; Schriemer, D.C.; Schryvers, A.B. Hydrogen–deuterium exchange coupled to mass spectrometry to investigate ligand–receptor interactions. In *Neisseria Meningitidis—Advanced Methods and Protocols*; Humana Press: Totowa, NJ, USA, 2012; pp. 237–252.
23. Mistarz, U.H.; Brown, J.M.; Haselmann, K.F.; Rand, K.D. Probing the binding interfaces of protein complexes using gas-phase H/D exchange mass spectrometry. *Structure* **2016**, *24*, 310–318. [CrossRef] [PubMed]
24. Wales, T.E.; Engen, J.R. Hydrogen exchange mass spectrometry for the analysis of protein dynamics. *Mass Spectrom. Rev.* **2006**, *25*, 158–170. [CrossRef] [PubMed]
25. Engen, J.R. Analysis of protein conformation and dynamics by hydrogen/deuterium exchange MS. *Anal. Chem.* **2009**, *81*, 7870–7875. [CrossRef] [PubMed]
26. Nyíri, K.; Mertens, H.D.T.; Tihanyi, B.; Nagy, G.N.; Kőhegyi, B.; Matejka, J.; Harris, M.J.; Szabó, J.E.; Papp-Kádár, V.; Németh-Pongrácz, V.; et al. Structural model of human dUTPase in complex with a novel proteinaceous inhibitor. *Sci. Rep.* **2018**, *8*, 4326. [CrossRef] [PubMed]
27. Houde, D.; Berkowitz, S.A.; Engen, J.R.; Fadgen, K.E.; Brown, J.; Engen, J.R.; Lee, C.T.; Steen, J.A.; Steen, H.; Mayer, M.P.; et al. The utility of hydrogen/deuterium exchange mass spectrometry in biopharmaceutical comparability studies. *J. Pharm. Sci.* **2011**, *100*, 2071–2086. [CrossRef]
28. Roberts, V.A.; Pique, M.E.; Hsu, S.; Li, S.; Slupphaug, G.; Rambo, R.P.; Jamison, J.W.; Liu, T.; Lee, J.H.; Tainer, J.A.; et al. Combining H/D exchange mass spectroscopy and computational docking reveals extended DNA-binding surface on uracil-DNA glycosylase. *Nucleic Acids Res.* **2012**, *40*, 6070–6081. [CrossRef] [PubMed]
29. Kelly, L.A.; Mezulis, S.; Yates, C.; Wass, M.; Sternberg, M. The Phyre2 web portal for protein modelling, prediction, and analysis. *Nat. Protoc.* **2015**, *10*, 845–858. [CrossRef] [PubMed]
30. Benedek, A.; Temesváry-Kis, F.; Khatanbaatar, T.; Leveles, I.; Surányi, É.V.; Szabó, J.E.; Wunderlich, L.; Vértessy, B.G. The role of a key amino acid position in species-specific proteinaceous dUTPase inhibition. *Biomolecules* **2019**, *9*, 221. [CrossRef]
31. Moroz, O.V.; Harkiolaki, M.; Galperin, M.Y.; Vagin, A.A.; González-Pacanowska, D.; Wilson, K.S. The crystal structure of a complex of Campylobacter jejuni dUTPase with substrate analogue sheds light on the mechanism and suggests the "basic module" for dimeric d(C/U)TPases. *J. Mol. Biol.* **2004**, *342*, 1583–1597. [CrossRef]
32. Laskowski, R.A.; Swindells, M.B. LigPlot+: Multiple ligand–protein interaction diagrams for drug discovery. *J. Chem. Inf. Model.* **2011**, *51*, 2778–2786. [CrossRef] [PubMed]
33. Benedek, A.; Pölöskei, I.; Ozohanics, O.; Vékey, K.; Vértessy, B.G. The Stl repressor from *Staphylococcus aureus* is an efficient inhibitor of the eukaryotic fruitfly dUTPase. *FEBS Open Biol.* **2018**, *8*, 158–167. [CrossRef] [PubMed]
34. Galperin, M.Y.; Moroz, O.V.; Wilson, K.S.; Murzin, A.G. House cleaning, a part of good housekeeping. *Mol. Microbiol.* **2006**, *59*, 5–19. [CrossRef] [PubMed]

35. Kerepesi, C.; Szabó, J.E.; Papp-Kádár, V.; Dobay, O.; Szabó, D.; Grolmusz, V.; Vértessy, B.G. Life without dUTPase. *Front. Microbiol.* **2016**, *7*, 1768. [CrossRef] [PubMed]
36. Hirmondo, R.; Lopata, A.; Suranyi, E.V.; Vertessy, B.G.; Toth, J. Differential control of dNTP biosynthesis and genome integrity maintenance by dUTPases. *Sci. Rep.* **2017**, *7*, 6043. [CrossRef] [PubMed]
37. Wang, H.C.; Hsu, K.C.; Yang, J.M.; Wu, M.L.; Ko, T.P.; Lin, S.R.; Wang, A.H.J. Staphylococcus aureus protein SAUGI acts as a uracil-DNA glycosylase inhibitor. *Nucleic Acids Res.* **2014**, *42*, 1354–1364. [CrossRef] [PubMed]
38. Mir-Sanchis, I.; Roman, C.A.; Misiura, A.; Pigli, Y.Z.; Boyle-Vavra, S.; Rice, P.A. Staphylococcal SCCmec elements encode an active MCM-like helicase and thus may be replicative. *Nat. Struct. Mol. Biol.* **2016**, *23*, 891–898. [CrossRef]
39. Yan, N.; O'Day, E.; Wheeler, L.A.; Engelman, A.; Lieberman, J. HIV DNA is heavily uracilated, which protects it from autointegration. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 9244–9249. [CrossRef]
40. Weil, A.F.; Ghosh, D.; Zhou, Y.; Seiple, L.; McMahon, M.A.; Spivak, A.M.; Siliciano, R.F.; Stivers, J.T. Uracil DNA glycosylase initiates degradation of HIV-1 cDNA containing misincorporated dUTP and prevents viral integration. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, E448–E457. [CrossRef]
41. Papp-Kádár, V.; Szabó, J.E.; Nyíri, K.; Vertessy, B.G. In vitro analysis of predicted DNA-binding sites for the Stl repressor of the *Staphylococcus aureus* SaPIBov1 pathogenicity island. *PLoS ONE* **2016**, *11*, e0158793. [CrossRef]