The Application of Nuclear Magnetic Resonance to Investigate Biological Processes

Charles Buchanan

Pembroke College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2022

Abstract

To understand disease, we need to understand the behaviour of individual biomolecules, how they interact, and how this complex 'dance' gives rise to organism level phenotypes. Proteins are an important class of molecules on which to focus, the 'machinery' of the cell. It is clear that to understand how they perform their various functions, we need a detailed description of not only their 'static' or average structure, but also the dynamical motions that govern how they move. Biological NMR, as a solution state, atomic resolution technique, is very well suited to this characterisation. Despite their power, biological NMR techniques are not widely applied, mostly because the barrier to entry is very high when compared to alternative techniques such as cryo-EM. In this thesis, a deconvolution algorithm, UnidecNMR, is developed to automate 'peak picking', an aspect identified as a rate limiting step in all NMR analysis. This step is manual on all but the simplest targets. UnidecNMR is demonstrated to be either as good, or better, than a human for real experimental data (chapter 2). It is then embedded into multiple pipelines to address experiment-specific problems, and by doing so, we develop novel methodology. Initially, application to 3D triple resonance data leads to near-automatic backbone and side chain assignment, together with a visualisation that lets a user clearly follow how the data maps onto the assignment (chapter 3). An application to 4D Methyl-TROSY NOE spectra and pseudo-3D CPMG relaxation dispersion data from a dynamical 396kDa oligomer unveils previously undocumented dynamics likely important for function (chapter 4). To further improve the sensitivity of triple resonance NMR data, 'Seedless' band selective GRAPE pulses are developed, where we experimentally note that specific improvements of B1 field homogeneity compensation during NMR experiments can provide dramatic sensitivity enhancements. For example, the sensitivity of an NHSQC at 950 MHz is increased by 45%, and an HNCACO at 600 MHz by 55% (chapter 5). Finally, it formed the core of a new saturation transfer analysis pipeline for studying protein - ligand interactions. This led to atomic resolution, kinetic and structural information about a very relevant interaction between the SARS-CoV-2 spike protein and N-linked glycans found on cell surfaces (6). These applications are rendered feasible through the capabilities of UnidecNMR. By providing this software to the community, and demonstrating what can be achieved using these approaches, I hope this thesis will lower the barrier for entry to bring biological research into the field to enable a greater understanding of protein dynamics and hence deepen our understanding of disease.

The Application of Nuclear Magnetic Resonance to Investigate Biological Processes



Charles Buchanan

Pembroke College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2022

This thesis is dedicated to Mum, Dad and my brother George with all my love.

Acknowledgements

Scientific

I have been very fortunate during my DPhil to collaborate with excellent scientists. My thanks to: Gogulan Karunanithy and Virginia-Casablancàs Antras for their expression of isotopically labelled ABP1P for the GRAPE project; Michael Barber and Reid Alderson who had previous made samples of Hsp16.5; Olga Tkachenko who expressed αB-Crystallin, which was used as part of the UniDecNMR benchmark; Gaurav Bhole and Prof Jonathan Jones, for the Quantum Information Theory underlying the Seedless project; Lachlan Deimel, who is leading the immmunology side of the Glycan-antibody project; Aziz Khan and Andrew Giltrap for both making the uSTA project so enjoyable; Prof. Ben Davis for teaching me the importance of deadlines.

Personal

My time in the Baldwin group has been the most rewarding of my life. I arrived disillusioned with education and desperate to leave. In the early days, Andy allowed me (sometimes begrudgingly...!) to spend an inordinate amount of time rowing. I learnt much from Sean Bowden, surpassed all my expectations, competed in the Isis-Goldie race twice and the Boat Race once and was selected for the cancelled 2020 Boat Race. Meanwhile I somehow became obsessed with chemistry. Andy showed me uncompromising standards, gave me freedom to explore and a drive to succeed. Thank you to both Andy and Sean, I look forward to many years of learning ahead.

Thank you to Mum for her eagle-eyed proofreading.

To Rob, thank you for your love, support, and for putting up with my intensity.

Abstract

To understand disease, we need to understand the behaviour of individual biomolecules, how they interact, and how this complex 'dance' gives rise to organism level phenotypes. Proteins are an important class of molecules on which to focus, the 'machinery' of the cell. It is clear that to understand how they perform their various functions, we need a detailed description of not only their 'static' or average structure, but also the dynamical motions that govern how they move. Biological NMR, as a solution state, atomic resolution technique, is very well suited to this characterisation. Despite their power, biological NMR techniques are not widely applied, mostly because the barrier to entry is very high when compared to alternative techniques such as cryo-EM. In this thesis, a deconvolution algorithm, UnidecNMR, is developed to automate 'peak picking', an aspect identified as a rate limiting step in all NMR analysis. This step is manual on all but the simplest targets. UnidecNMR is demonstrated to be either as good, or better, than a human for real experimental data (chapter 2). It is then embedded into multiple pipelines to address experiment-specific problems, and by doing so, we develop novel methodology. Initially, application to 3D triple resonance data leads to near-automatic backbone and side chain assignment, together with a visualisation that lets a user clearly follow how the data maps onto the assignment (chapter 3). An application to 4D Methyl-TROSY NOE spectra and pseudo-3D CPMG relaxation dispersion data from a dynamical 396kDa oligomer unveils previously undocumented dynamics likely important for function (chapter 4). To further improve the sensitivity of triple resonance NMR data, 'Seedless' band selective GRAPE pulses are developed, where we experimentally note that specific improvements of B1 field homogeneity compensation during NMR experiments can provide dramatic sensitivity enhancements. For example, the sensitivity of an NHSQC at 950 MHz is increased by 45%, and an HNCACO at 600 MHz by 55% (chapter 5). Finally, it formed the core of a new saturation transfer analysis pipeline for studying protein - ligand interactions. This led to atomic resolution, kinetic and structural information about a very relevant interaction between the SARS-CoV-2 spike protein and N-linked glycans found on cell surfaces (6). These applications are rendered feasible through the capabilities of UnidecNMR. By providing this software to the community, and demonstrating what can be achieved using these

approaches, I hope this thesis will lower the barrier for entry to bring biological research into the field to enable a greater understanding of protein dynamics and hence deepen our understanding of disease.

Contents

List of Figures **xiii**

List of Abbreviations **xvii**

1 Introduction **1**
1.1 Tackling Disease . . . 1
1.2 Nuclear Magnetic Resonance . . . 2
1.2.1 NMR observables . . . 3

2 UnidecNMR **7**
2.1 Acknowledgements . . . 7
2.2 Introduction . . . 8
2.3 UnidecNMR Theory . . . 8
2.4 Application to simulated data . . . 9
2.5 Application to real data . . . 13
2.5.1 Application to heavily overlapped 1D sugar spectra . . . 13
2.5.2 Application to 2D NHSQC data . . . 14
2.5.3 Application to 3D backbone spectra . . . 15
2.5.4 Applying UnidecNMR to 3D backbone spectra . . . 16
2.5.5 Designing a GUI for 2- and 3-D UnidecNMR data . . . 18
2.5.6 Applying UnidecNMR to 3D backbone spectra . . . 19
2.6 Application to 3 and 4D NOESY spectra . . . 21
2.7 Conclusions . . . 23
2.8 Materials and Methods . . . 25

3 Automated Assignment Strategy **27**
3.1 Designing the Assignment strategy . . . 27
3.1.1 Data structure . . . 28
3.2 Initial Assignment algorithm . . . 28
3.3 Presenting ambiguous data . . . 30
3.4 Conclusions . . . 34

4 Studying the dynamics of a 396kDa sHSP oligomer at atomic resolution 37
4.1 Acknowledgements . . . 38
4.2 Introduction . . . 38
4.2.1 Small heat shock proteins . . . 38
4.2.2 Methyl-TROSY NMR . . . 41
4.3 Spectra . . . 45
4.3.1 Initial temperature ramp . . . 45
4.3.2 Assignment of MILV labelled full length protein . . . 45
4.3.3 UnidecNMR application to 3D NOESY spectra . . . 49
4.4 Dynamics experiments . . . 51
4.4.1 The fitting problem . . . 53
4.5 CPMG data . . . 54
4.5.1 Comparison with alphaB Crystallin . . . 55
4.5.2 Domain switching dynamics . . . 56
4.5.3 N-terminal domain . . . 57
4.5.4 Spin-state selective CPMG . . . 65
4.5.5 α-lactalbumin . . . 66
4.6 Alpha-crystallin domain . . . 67
4.6.1 Comparison with full length protein . . . 68
4.6.2 Assignment of ACD . . . 71
4.6.3 Temperature Ramp . . . 71
4.7 Conclusions . . . 74
4.8 Future work . . . 75
4.9 Materials and Methods . . . 76

5 Optimised pulsing during NMR experiments: Seedless 77
5.1 Acknowledgements . . . 78
5.2 Introduction . . . 78
5.2.1 Pulses and the Bloch equation . . . 79
5.3 Seedless pulses . . . 82
5.3.1 Pulse optimisation . . . 82
5.3.2 GRAPE pulses . . . 82
5.3.3 Point-to-point vs Unitary . . . 85
5.3.4 B1 inhomogeneity . . . 86
5.4 Using Flourine NMR as a Seedless proof of concept . . . 86
5.5 A critical component of heteronuclear NMR: Insensitive Nuclei Enhanced by Polarisation Transfer . . . 88
5.6 Triple resonance experiments . . . 89

5.6.1 The requirement for carbon selectivity 91
5.6.2 Alpha-Beta carbon decoupling 92
5.6.3 Pulse sequence analysis 93
5.6.4 The benefits of Seedless pulses 94
5.7 Triple resonance results 96
5.7.1 HNCO 97
5.7.2 B1 inhomogeneity 100
5.7.3 Introducing Hydrogen and Nitrogen pulses 102
5.7.4 Full triple resonance results 103
5.8 NHSQC results 105
5.9 Conclusions 105
5.10 Materials and methods 107

6 Universal Saturation Transfer Analysis 109
6.1 Acknowledgements 109
6.2 Introduction 110
6.3 Development of uSTA 112
6.4 BSA-Tryptophan application 114
6.5 Haemaglutinnin influenza application 115
6.6 Cryptic SARS-CoV-2 spike sialoside binding 117
6.7 Cryo-EM pinpoints the sialoside attachment site 121
6.8 Discussion and conclusions 126
6.9 Future Work 128
6.10 Materials and Methods 128

7 BAR-1, a sugar binding antibody 129
7.1 Acknowledgements 129
7.2 Introduction 130
7.3 Glycan binding antibody design 131
7.4 uSTA application 132
7.4.1 Comparison with X-Ray data 132
7.4.2 Kd calculations 134
7.4.3 Mutational study 134
7.5 Conclusions 135

Appendices

A uSTA Supplementary figures 139

B uSTA Theory **161**
B.1 uSTA theory 161
B.1.1 Kinetics 164
B.1.2 Combining kinetics and spin evolution 165
B.1.3 Relaxation treatment 167
B.1.4 Using the matrix 168
B.2 Simulating 170

References **171**

List of Figures

2.1 UnidecNMR method 9
2.2 The results of UnidecNMR are relatively tolerant to the choice of peak shape 10
2.3 Simulation of NMR data with two difficulty axes 11
2.4 Importance of the squash function 12
2.5 Results of 1D simulated spectra 12
2.6 Results of 2D simulated spectra 14
2.7 Application of UnidecNMR to 1D proton spectra of highly overlapped sugars. 15
2.8 Results of 2 and 3D real data 16
2.9 Diagrammatic representation of assignment 17
2.10 UnidecNMR Processing GUI 19
2.11 UnidecNMR 2d peak list alignment GUI 20
2.12 UnidecNMR peak fitting GUI 21
2.13 UnidecNMR 2D viewer GUI 22
2.14 Automatic picking of 3 and 4D methyl NOE spectra. 24

3.1 Logical walk GUI 31
3.2 Spin Frame GUI 32
3.3 Projection tab GUI 34
3.4 Preliminary assignment reporting figure 35

4.1 HSP16.5 structure 40
4.2 The interior of HSP16.5 displays dynamical behaviour 43
4.3 Preliminary HSP16.5 spectra 44
4.4 NOESY bonding 46
4.5 4D Ile-only NOESY 47
4.6 Resimulated NOESY spectrum 48
4.7 Assignment of Methionine 43 49
4.8 Full assignment of the MILV Hsp16.5 52
4.9 UnidecNMR lineshape fitting GUI 54
4.10 CPMG dispersion of ACD-loop methyl groups 56

4.11 CPMG fits of full length HSP16.5 58
4.12 The N-terminus of HSP16.5 59
4.13 Alphafold ensemble of Alphafold dimer 64
4.14 Alphafold ensemble of Alphafold dimer 64
4.15 NTD methionine ACD NOEs 65
4.16 CPMG fits of full length HSP16.5 methionine residue 66
4.17 Spin-state selective CPMG fitting 67
4.18 Addition of aggregating α-Lac to Hsp16.5 68
4.19 The aggregation of model protein α-lactalbumin in the presence of the ACD of HSP16.5 69
4.20 HSP16.5 ACD vs oxidised HSP27 ACD 69
4.21 Full-length ACD spectra comparison 70
4.22 ACD construct assignment spectra 71
4.23 Assigned Temperature ramp of the ACD construct 73
4.24 ACD construct CPMG temperature dependence 74

5.1 Amplitude and phase of pulses 81
5.2 Hard pulse profile 81
5.3 Amplitude and phase of pulses 83
5.4 The trajectory of a Seedless Unitary pulse 85
5.5 Hard pulse profile 87
5.6 INEPT and Indirect acquisition blocks 89
5.7 BMRB carbon chemical shift statistics 92
5.8 Pulse sequence analysis 94
5.9 Imperfect CACB decoupling using Seedless pulses 95
5.10 HNCO and HNCA pulse sequnces 97
5.11 CO vs CA dependency of Seedless HNCO improvements 98
5.12 Individual HNCO carbon pulse replacement 99
5.13 2D and 3D experiments show similar patterns 101
5.14 Replacing Hard with Seedless one by one 103
5.15 Full triple resonance results 104
5.16 Full triple resonance results 106

6.1 uSTA identifies a sugar- binding site and pose in the unusual NTD of early-pandemic SARS-CoV-2 spike. 110
6.2 Development of the uSTA method 116
6.2 Continued from previous page 117
6.3 uSTA allows mapping of a designed cryptic sugar-binding site in H1N1 influenza hemagglutinin (HA) 118
6.4 uSTA reveals interaction of sialosides with SARS-CoV-2 spike protein 120

6.4 Continued from previous page . . . 121
6.5 Quantitative uSTA analyses allow comparison and predictive restraints for protein ligand-binding prediction . . . 122
6.5 Continued from previous page . . . 123
6.6 Comparison of SARS-CoV-2 glycan attachment mechanisms and variant evolution via uSTA suggests binding away from the RBD that is lost. . . . 124
6.7 Isolation of Sialic-acid binding site . . . 125
6.7 Continued from previous page . . . 126

7.1 Initial application of uSTA to BAR-1 ligands . . . 133
7.2 Crystallographic data for BAR-1 and amidine-lysine linked 2,3 sialyllactose . . . 133
7.3 Bloch-McConnell simulation of BAR-1 and amidine-lysine linked 2,3 sialyllactose . . . 134
7.4 BAR-1 mutants . . . 135

A.1 Homology Analyses . . . 140
A.2 Schematic Representation of the Classical STD Experiment and the Deconvolution Approach. . . . 141
A.2 Continued from previous page . . . 142
A.3 A Summary of the Workflow in a Classical NMR STD Experiment and Application of uSTA to Spike • Sugar-hybrid 5 and TreR • Tre systems. . . . 143
A.3 Continued from previous page . . . 144
A.4 Modelling of Current STD Dependencies/Methods. . . . 145
A.4 Continued from previous page . . . 146
A.5 The uSTA Concept . . . 146
A.6 Workflow of uSTA . . . 147
A.7 Advantages of the uSTA Pipeline, Emphasising Specific Advantages Over A More Conventional Analysis . . . 148
A.8 The Importance and Reliability of Ligand Subtraction when Calculating uSTA Surfaces . . . 149
A.9 Comparison of SARS-CoV-2 glycan attachment mechanisms and variant evolution via uSTA suggests binding away from the RBD that is lost. . . . 150
A.10 Comparison of SARS-CoV-2 glycan attachment mechanisms and variant evolution via uSTA suggests binding away from the RBD that is lost. . . . 151
A.10 Continued from previous page . . . 152

A.11 Combination of uSTA with HADDOCK Allowed Ranking of Docked Model Ensembles 153
A.11 Continued from previous page 154
A.12 uSTA Analysis of Spike from Variants of Concern 155
A.13 Effects Upon Sialoside Binding of the RBD-Blocking Neutralizing C5 Antibody 156
A.14 Cryo-EM Coulombic maps for sialic acid bound to spike 157
A.15 Cryo-EM Coulombic maps for ligands bound to spike 158
A.16 Cryo-EM Coulombic maps confirm an addition binding mode for aromatic ligands in a distinct spike region 159

List of Abbreviations

ABP1P . . . . Actin Binding Protein 1

ACD . . . . . Alpha Crystallin Domain

ACE-2 . . . . . Angiotensin Converting Enzyme-2

ATCase . . . . Aspartate Transcarbamoylase

ATP . . . . . Adenosine Triphosphate

BCR . . . . . B-cell Receptor

B1 . . . . . Radiofrequency pulse amplitude

B0 . . . . . Magnetic field strength

BSA . . . . . Bovine Serum Albumin

BMRB . . . . Biological Magnetic Resonance Data Bank

CATIA . . . . CPMG, Anti-Trosy and Trosy Intelligent Analysis

CEST . . . . . Chemical Exchange Saturation Transfer

CPMG . . . . Carr-Purcell-Meiboom-Gill

CTD . . . . . C-terminal Domain

DDX4N . . . . DEAD-Box RNA Helicase 4 N-terminal domain

EIN . . . . . E-coli Enzyme I N-terminal Domain

EM . . . . . Electron Microscopy

Fab . . . . . Fragment Antigen-Binding region

FID . . . . . Free Induction decay

FT . . . . . Fourier transform

GPCR . . . . . G-Protein Coupled Receptors

GRAPE . . . . Gradient Ascent Pulse Engineering

GUI . . . . . Graphical User interface

HADDOCK . High Ambiguity Driven biomolecular DOCKing

HA . . . . . Haemaglutinnin (Influenza surface protein)

HEL . . . . . Hen Egg Lysozyme

HIV . . . . . Human Immunodeficiency Virus

HSQC . . . . . Heteronuclear Single Quantum Coherence

HMQC . . . . Heteronuclear Multiple-Quantum Correlation

IDP . . . . . Intrinsically Disordered Protein

Ig . . . . . Immunoglobulin

INEPT . . . . Insensitive Nuclei Enhanced by Polarisation Transfer

ITC . . . . . Isothermal Calorimetry

MAGMA . . . Methyl Assignment by Graph Matching

MERS . . . . . Middle Eastern Respiratory Syndrome

NAc . . . . . N-Acetyl group (corresponding to the Sialic acid moiety)

Neu5Ac . . . . N-Acetylneuraminic Acid

NHSQC . . . . Nitrogen Hydrogen Single Quantum Coherence

NMR . . . . . Nuclear Magnetic Resonance

NOE . . . . . Nuclear Overhauser Effect

NOESY . . . . Nuclear Overhauser Effect SpectroscopY

NTD . . . . . N-terminal Domain

P2P . . . . . Point-to-Point

PDB . . . . . Protein Databank

PRE . . . . . Paramagnetic Relaxation Enhancement

ppm . . . . . parts per million

PBS . . . . . Phosphate Buffered Saline

QCIF . . . . . Quad-resonance Inverse Cryoprobe with Fluorine

RF . . . . . Radiofrequency

SARS-CoV-2 . Severe Acute Respiratory Syndrome Coronavirus 2

SHM . . . . . Somatic Hypermutation

sHSP . . . . . Small heat shock proteins

SiaLac 2,3-sialyllactose

SPR . . . . . Surface Plasmon Resonance

STD . . . . . Saturation Transfer Difference

T1 . . . . . longitudinal Relaxation time

T2 . . . . . Transverse Relaxation time

TCI . . . . . Triple-resonance Inverse Cryoprobe

TOCSY . . . . Total Correlation Spectroscopy

TROSY . . . . Transverse Optimised Spectroscopy

uSTA . . . . . Universal Saturation Transfer Analysis

1 Introduction

Contents

1.1 Tackling Disease 1
1.2 Nuclear Magnetic Resonance 2
1.2.1 NMR observables 3

1.1 Tackling Disease

Since the beginning of the 20th century, there has been remarkable progress in the understanding and treatment of disease. A small subset includes: the discovery and enhancement of antibiotics [1]; life-saving vaccines including polio[2], smallpox [3] and SARS-CoV-2 [4]; cancer treatments[5]; near 100% effective prophylaxis of Human Immunodeficiency Virus (HIV)[6]. These all result from a steady drive to understand smaller and smaller components of life, spawning a scientific community both vast and complex, from multinational corporations to passionate individuals. One of the major challenges of this type of study is the distillation of extreme biological complexity into a set of independent and dependent variables that accurately demonstrate the efficacy of a newly presented model. Different approaches have resulted in a wide range of tools, from X-Ray crystallography

to the encyclopedia of cellular biology. The accumulated development of these approaches underpins human understanding of disease.

Biophysics sits at the interface of many disciplines and applies physical approaches to these complex biological systems. This often involves isolating a particular protein or proteinaceous complex to unpick thermodynamic and kinetic parameters. Experimental methods that span the range of physical science have been used to probe aspects of biological complexity, including X-ray crystallography, electron microscopy, mass spectrometry and Nuclear Magnetic Resonance (NMR) spectroscopy. The results can then be related to structural behaviour *in vitro* and biological function, with an assumption that behaviour is similar *in vivo*. For example, NMR and Mass Spectrometry were used to link the kinetic assembly/disassembly of oligomeric species to the dynamical regions of a small heat shock protein[7]. There is a continual drive for development at the forefront of these techniques to broaden their scope [8][9]. This work is usually undertaken by researchers with highly technical knowledge and, without translation, these important developments are often only accessible to a small group of specialists. Software design can guide a broad range of researchers through these techniques, enabling them to access the forefront without specialised knowledge[10]. Whilst software is not itself a scientific discovery, its importance lies in enabling the study of prohibitively time-consuming systems.

Most of this thesis will detail the study of diverse biological systems using NMR, facilitated by the distillation of technical knowledge into computational tools that demonstrate their utility through the resulting biological conclusions.

1.2 Nuclear Magnetic Resonance

Specific applications of Biological NMR (BioNMR) techniques will be discussed at the beginning of each chapter, but broadly it can probe the properties of materials through the intrinsic spin of atomic nuclei. It is used in a diverse set of applications, for example: organic chemists characterise small molecules, biochemists probe 'invisible' excited states of proteins and physicists apply it to quantum computation.

Whilst vast quantities of data have been produced on static protein structures in the last few decades, protein dynamics are comparatively poorly understood[11]. NMR can study molecules at atomic resolution in the solution state, meaning the observables sample the full statistical ensemble of spin states produced by a magnetic field. This ensemble can be manipulated using Radiofrequency pulses and its time-dependent evolution within the magnetic field recorded. This evolution is extremely sensitive to the local environment of the spin ensemble. Time-resolved information about the local environment enables observation of interchange between species in the ensemble at various timescales [12]. NMR is in a unique position, therefore, to study the dynamical behaviour of biomolecules. For example, intrinsically disordered proteins (IDPs) make up roughly 30% [13] of the human proteome, but have no clear structured state. They interconvert rapidly between many conformations as they tumble. While other structural techniques struggle with IDPs, solution state NMR probes the entire statistical ensemble, enabling study of this enormous class of proteins at atomic resolution. Work has been done by pioneering NMR groups across the world to develop experiments and mathematical models to probe protein dynamics at many timescales. This behaviour has often been linked to function[7][14], for example enzyme activity[14]. However, much of this work remains the preserve of highly specialised groups, and to maximise the impact of this in the biological world, tools are needed to 'lower the bar' to make these methods more accessible to a wider audience of scientists.

1.2.1 NMR observables

In order to understand the power of Biomolecular NMR, it is important to consider the observables. Using sequences of Radiofrequency pulses and delays (called 'Pulse Sequences'), Nuclear spin magnetisation can be made to precess around the bulk magnetic field and recorded as a Free Induction Decay. This data can then be Fourier transformed to produce an N-dimensional frequency-domain spectrum consisting of a number of resonances, or peaks. Encoded within the position and magnitude of these peaks are physical properties of the sample. Therefore, quantifying the position and

magnitude of these resonances is a crucial step towards understanding the underlying systems. Despite the historic development of many algorithms [15][16][17][18], none has reached widespread adoption and reliable quantification of biological spectra remains an extremely skilled process, requiring months to years of training to produce publication worthy fidelity. To address this, we developed a new approach, designed to be automated, but as effective as a skilled user, with zero tolerance to error. While many computational methods are happy to get high accuracy on large datasets, to the experimentalist studying a single problem, the only tolerable error rate in peak picking is zero. The method described in Chapter 2 of this thesis, UnidecNMR, appears to accomplish this. To demonstrate the effectiveness of this method, we apply it to several cases, covered in the following chapters:

1. Chapter 3 outlines the development of software to enhance triple resonance assignment experiments using UnidecNMR outputs, which was fundamental to achieving the results in the previous two chapters.

2. Chapter 4 details the implementation of UnidecNMR within tools to study the multiple timescale dynamics of an oligomeric heat shock protein complex, hsp16.5, at unprecedented resolution, revealing intriguing large scale fluctuations on the interior of the 396kDa protein assembly.

3. Chapter 5 discusses the optimisation of Radiofrequency pulses for application during NMR experiments, resulting in substantial increases in sensitivity for widely used experiments.

4. Chapter 6 describes the incorporation of UnidecNMR within uSTA, a novel approach to analysis of Saturation Transfer Difference (STD) NMR experiments to extract protein-ligand interaction information. This includes application to pathogen-sugar interactions, including SARS-CoV-2, from which multiple biological conclusions are drawn.

5. Chapter 7 further demonstrates the utility of the uSTA pipeline by characterising a challenging glycan-antibody interaction.

Taken as a whole, this thesis is focused on creating tools to unlock the breadth and depth of information accessible with NMR in order to enhance the resulting biological conclusions in even the most challenging systems.

2
UnidecNMR

Contents

2.1 Acknowledgements 7
2.2 Introduction 8
2.3 UnidecNMR Theory 8
2.4 Application to simulated data 9
2.5 Application to real data 13
2.5.1 Application to heavily overlapped 1D sugar spectra . . . 13
2.5.2 Application to 2D NHSQC data 14
2.5.3 Application to 3D backbone spectra 15
2.5.4 Applying UnidecNMR to 3D backbone spectra 16
2.5.5 Designing a GUI for 2- and 3-D UnidecNMR data . . . 18
2.5.6 Applying UnidecNMR to 3D backbone spectra 19
2.6 Application to 3 and 4D NOESY spectra 21
2.7 Conclusions 23
2.8 Materials and Methods 25

2.1 Acknowledgements

This work has been previously written into paper format, has been submitted and is undergoing the peer-review process. This has been rephrased in my own words. All figures were made by me, with minor editing by Prof. Baldwin. Some figures and captions have come from the paper draft, which were all written by me with minor editing by Prof. Baldwin.

2.2 Introduction

Deconvolution is a mathematical approach that attempts to extract the location and intensity from each spectrum. A peak-shape function and noise threshold are provided by the user and the algorithm will then iteratively converge on a resulting function. The convolution of this function with the peak-shape faithfully represents the data within the spectrum.

2.3 UnidecNMR Theory

Previous deconvolution algorithms have been developed [19][20]. UniDec (Universal deconvolution) was originally devised as an extension of the Richardson-Lucy algorithm to deconvolve mass spectrometry data and separate the mass and charge dimensions [19][20][21]. Our intention was to apply a similar deconvolution approach to NMR data. The previous mass spectrometry data shares many features with NMR data analysis. The method relies on the assumption that a spectrum of intensities, I, can be reasonably expressed as a convolution of a peak shape function, g, and an array of delta functions or sources, f, each of which spans the same set of spectral frequencies, i. The algorithm aims to perform the deconvolution that removes the peak shape function from the data, providing a user with a list of sources, f, that dictates both the peak positions and intensities. The kernel iterates (t) on the intensities of each spectral element:

$$f_i^{t+1} = f_i^t \frac{I_i}{(f^t * g)_i} \tag{2.1}$$

When the back-calculated spectrum, $(f * g)$, has the same intensity as the data I at frequency i, then the ratio is equal to 1, and the algorithm has converged [22]. It is worth noting here that we provide considerable tools for the user to visually compare the back-calculated spectrum with the raw data. This is our solution to a common pitfall with Bayesian approaches, whereby the initial assumptions (e.g. noise threshold and peak shape function) cause valid data points to be excluded as solutions. The action of the algorithm is to suppress sources that are

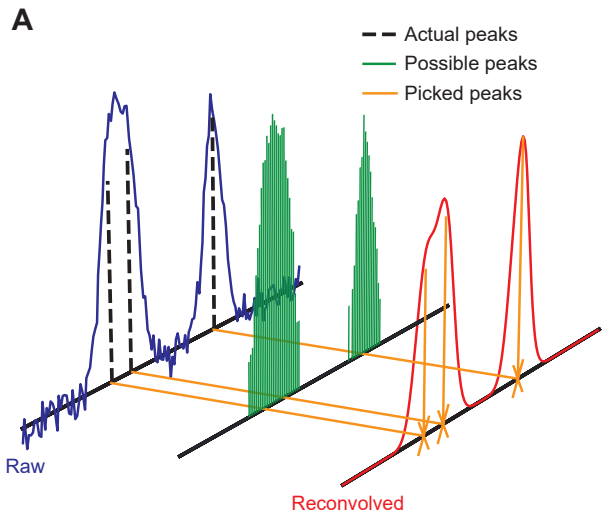


Figure 2.1: A schematic of the UnidecNMR method: all locations within a raw spectrum (Blue) above a noise threshold are initially possible peak locations (green) before UnidecNMR iteratively filters them down to final locations (orange). The reconvolved spectrum convolves these final peak positions with the inputted peak shape function to give data that can be compared against the raw data (Blue)

adjacent to the true 'centre' of a resonance. A user supplies a noise threshold and a peak shape function g (Figure 2.1a) and the intensities of possible peak positions are iteratively adjusted according to equation until convergence, where the final values in f provide the central locations and intensities of the picked peaks. The selected peak shape g should be a reasonable match for the average resonance in a spectrum, although the final results are reasonably tolerant to this parameter being incorrectly set (Figure 2.2), as expected in the case in experimental data where there is a wide range of peak shapes. The algorithm can apply in 1, 2, 3 or 4 dimensions (as shown in later Figures).

2.4 Application to simulated data

To test a naive implementation of UnidecNMR, we simulated spectra. A Gaussian function was convolved with a peak location function that had two non-zero points at a given peak separation. Gaussian noise was added at a specified signal:noise ratio to produce each simulated spectrum. A database was created using the 2 difficulty axes: signal-to-noise and separation. 9 reasonable values were chosen to

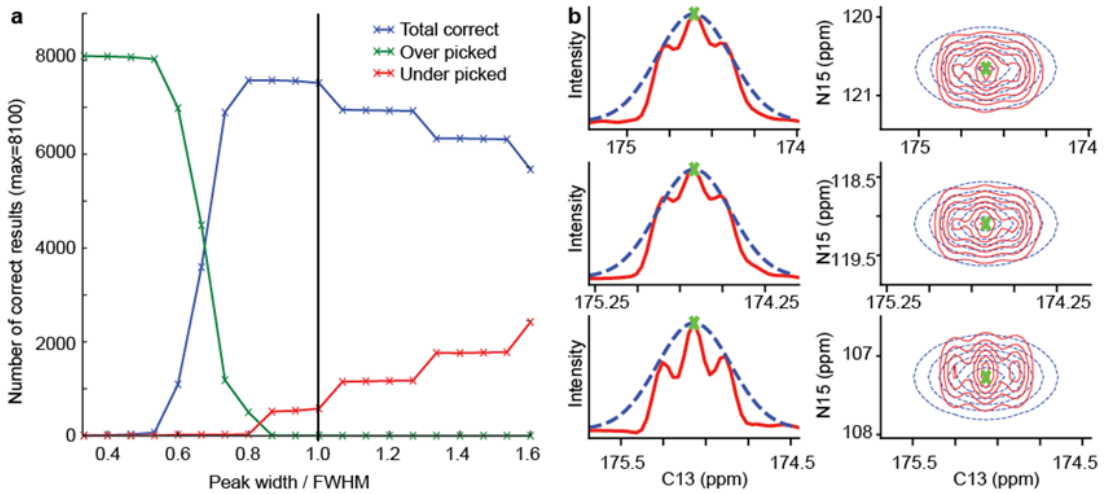


Figure 2.2: The accuracy of 2D peak picking is shown against the full simulated dataset of variable signal-to-noise and separation values (a), with different peak widths shown as a function of Full-Width-Half-Maximum. Maximum accuracy was achieved if the multiple of the peak width used for calculation was between 0.8 and 1 of the 'actual' peak width. When the peak width is set to be lower than this, the algorithm will over-pick the spectrum. When the peak width is set to be wider than this, the algorithm tends to under-pick, with the success rates falling off gradually. The discrete form of the plot is due to the 'squash' window being specified in the program as an integer of points. In cases where peak widths are substantially wider than this, UnidecNMR can be run twice with different settings for the peak shape. (b) shows excerpts from the HNCO spectrum from Hsp16.5, where the decoupler was mis-set and the peak shape in the 13C dimension resembles a triplet. While undesirable, this spectrum can be accurately analysed using UnidecNMR by setting the peak shape to be broad enough to encompass the triplet.

create each axis and produce 500 spectra (each with a different sample of noise from a Gaussian distribution) for each combination, giving 40,500 test spectra. The aim was to maximise the number of test spectra from which the peak location function could be accurately extracted. We accepted any solution that was within two points of the initially simulated peak location.

The naive implementation of UnidecNMR gave undesirable distributions of peak locations. This was rational, as the convolution of one Gaussian function with a second Gaussian function results in a third Gaussian function [23]. However, distributions of potential peak locations are not the desired output (Figure 2.4A). Therefore, a new algorithm was devised to impose the desired output without distributions of line shape. We have called this the 'squash function'. It is a simple clustering algorithm which identifies all maxima in the deconvolution output and

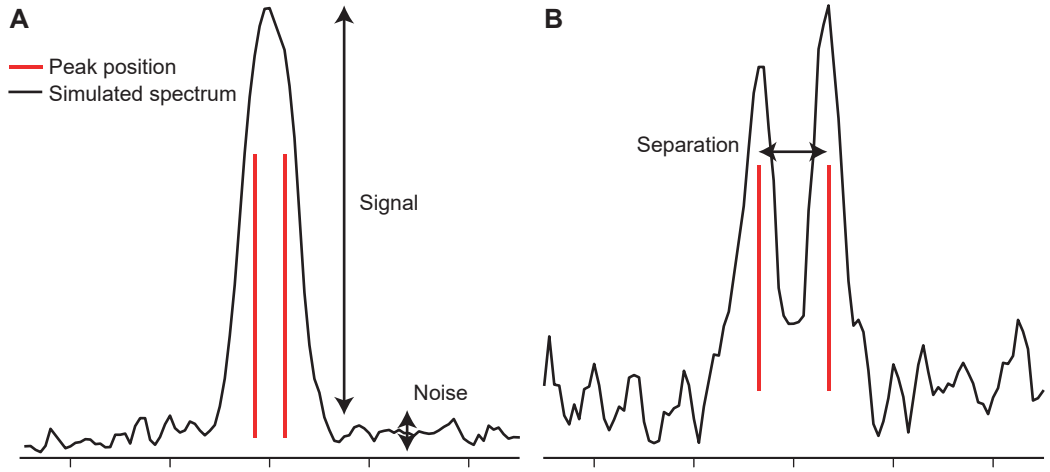


Figure 2.3: This shows our simulated NMR data with two difficulty axes: signal-to-noise ratio, (A) and peak separation (B). The initial peak location is indicated.

clusters the resulting intensity into a singular point. We defined the clustering width in a parameter called 'Squash' and optimised as a function of the peak full width half maximum (Figure 2.5E, F). The clustered output is then used as the next input for another run of the deconvolution algorithm, which optimises the height of the results. As any position that is set to zero in the deconvolution input is not enabled as a deconvolution output (similar to the initial noise threshold), this prevents the distributed peak position output. The combination of this algorithm with a corresponding protocol gave good results qualitatively (Figure 2.4B), but to quantify this it had to be applied to the full dataset of simulated 1D spectra.

The naive implementation of UnidecNMR gave similarly poor results on the dataset (Figure 2.5C). Examples again showed a distribution of peak positions (Figure 2.5A). Application of the full protocol reliably gave near perfect scores except at very low signal-to-noise or separation, where an experienced spectroscopist would also struggle. The optimisation of two of the protocol parameters is shown to be very important (Figure 2.5E,F). Both parameters are expressed as a multiple of the peak shape width, meaning they do not need spectroscopist input. They were optimised by taking a sub-sample of the most challenging parts of the dataset and iteratively varying the two parameters and summing the results to look for changes.

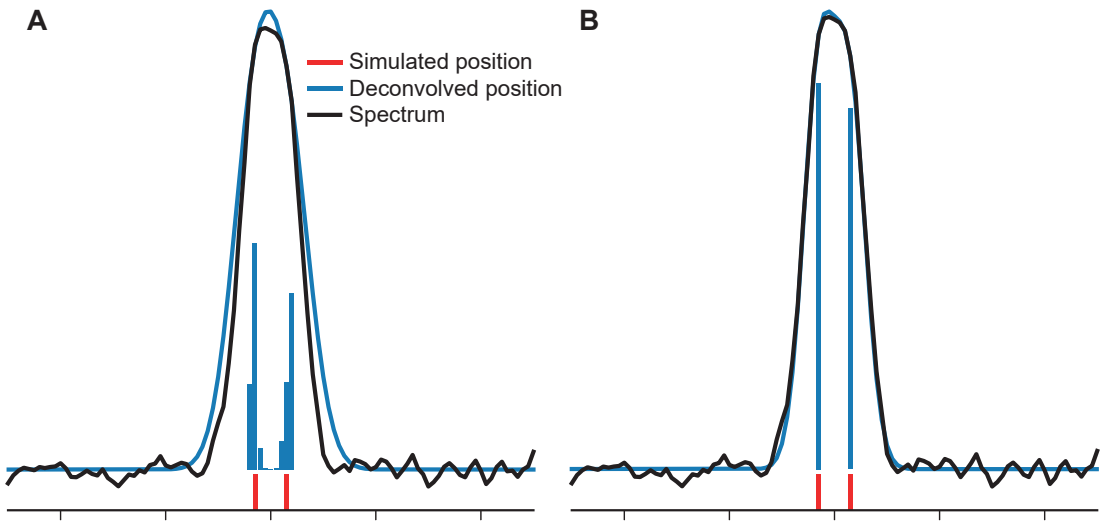


Figure 2.4: A distribution of outputs is given by the naive UnidecNMR implementation (A), while the desired delta function condition is imposed by the squash function (B)

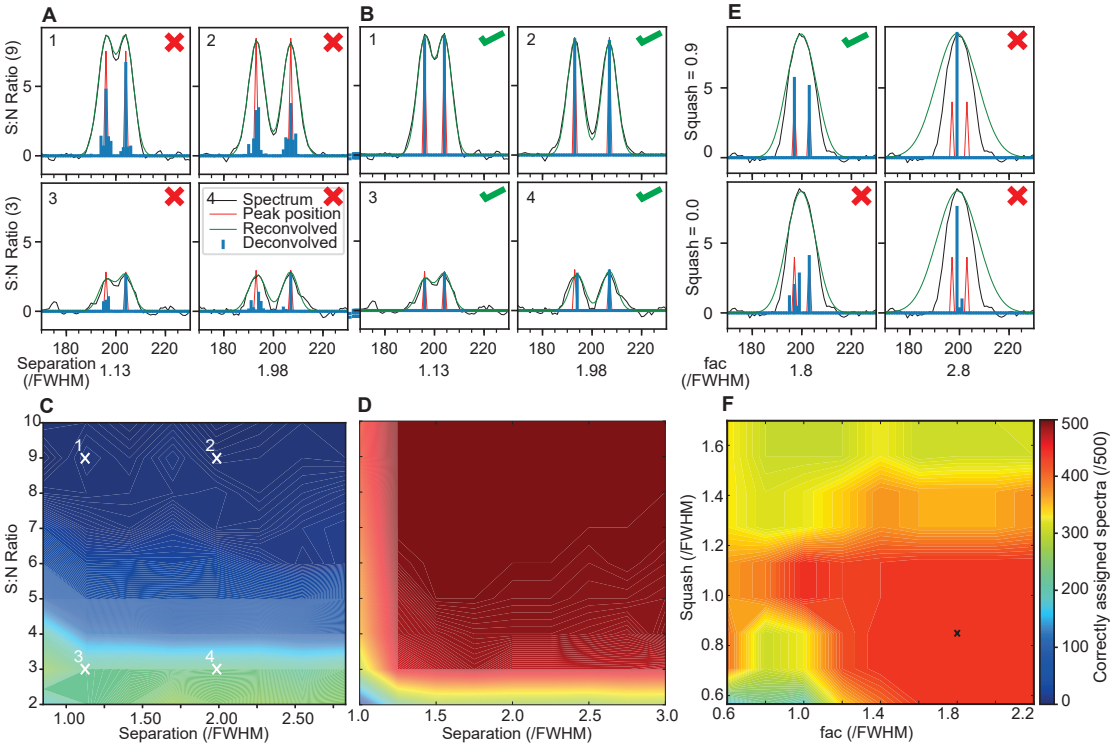


Figure 2.5: Poor results on the 1D dataset (C) are shown from naive deconvolution while examples (A) demonstrate that a distribution of peak positions is found rather than the desired delta function. The maximum signal-to-noise for the spectra in A, B and E are indicated in brackets on the y-axis, taken from the full dataset at the numbered points indicated in C. Post squash function and protocol optimisation results (D) are far better, with examples (B) showing that a delta function condition has been imposed. Taking one of the more difficult regions from D, we demonstrate the importance of tuning two of these parameters (E,F)

These results inspired confidence, and the algorithm was applied to 2D spectra, which are more relevant to Biomolecular NMR. We extended the optimised 1D protocol into a second dimension and applied it to a 2D simulated dataset (Figure 2.6). Using the same simulated dataset, we compared the performance to the following set of previously published algorithms:

1. NMRNet employs a convolutional neural network based on machine learning
2. PICKY relies on a singular value decomposition of spectra
3. WaVPeak takes advantage of a wavelet smoothing and clustering technique

NMRNet [17] is a Neural network based technique which was not optimised for simulated data and performed much better on real data. PICKY [15] performed well at low difficulty, but at low signal-to-noise or separation, the results were substantially worse. WaVPeak [16] gave slightly less predictable results and again struggled at lower signal-to-noise or separation. The results demonstrated that UnidecNMR far outperformed the competitors on this simulated dataset.

2.5 Application to real data

2.5.1 Application to heavily overlapped 1D sugar spectra

Having optimised for simulated data with promising results, the algorithm was applied to real data. Sugar molecules contain multiple homonuclear coupled, chemically similar environments with low chemical shift dispersion. This combination means the coupling constants are similar to the chemical shift dispersion, causing extra complication via 'strong' coupling. UnidecNMR was applied to some trisaccharide data, and was able to identify many multiplet components in heavily overlapped regions (Figure 2.7). Applications of this to Saturation transfer difference NMR spectra will be discussed in Chapter 6.

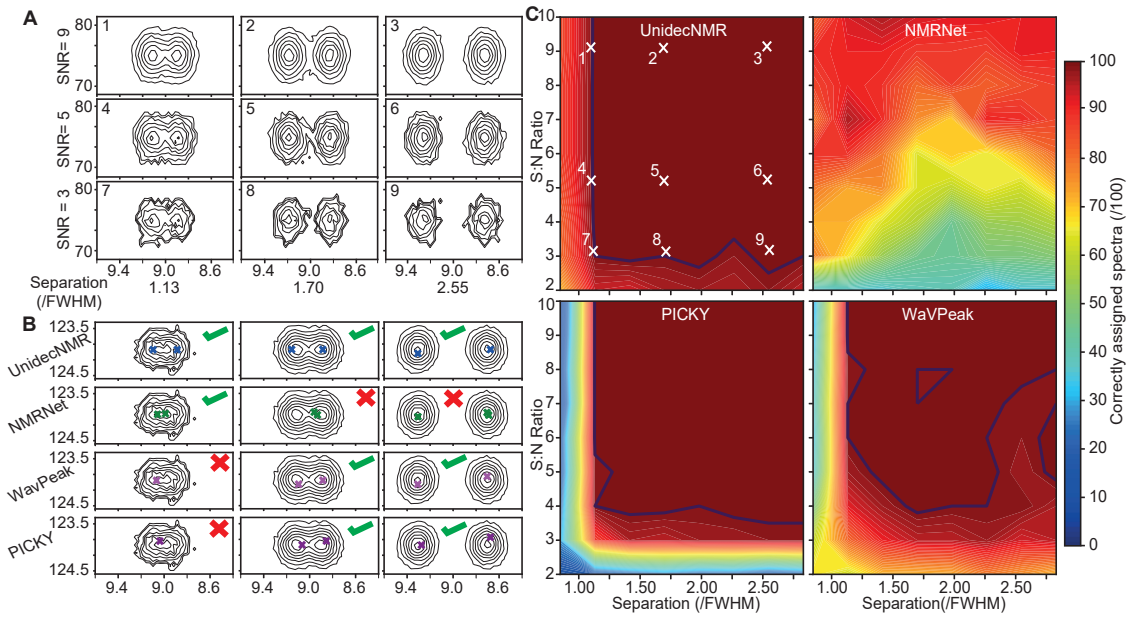


Figure 2.6: Performance of UnidecNMR on 2D simulated data (7200 spectra). (A) shows Samples of simulated data illustrating a range in signal-to-noise on the y-axis, and separation on the x-axis. (B) Illustrative examples of three spectra that provide a representative discrimination of the various algorithms tested. Only NMRNet [17] and UnidecNMR were able to resolve the peak with the closest separation whereas only WaVPeak 14, PICKY 10 and UnidecNMR were able to distinguish the peaks with the widest separation. (C) Overall performance on the entire simulated 2D dataset. Only UnidecNMR, PICKY 10 and WaVPeak 14 achieve 100% accuracy at the relatively trivial high separation/high signal-to-noise limit. The lowest combination of separation and signal-to-noise at which the algorithms have a 100% success rate can be compared. UnidecNMR outperforms the other algorithms on these synthetic datasets.

2.5.2 Application to 2D NHSQC data

The 1H-15N Heteronuclear Single Quantum Coherence (NHSQC) was chosen as the 2D test spectrum because it forms the backbone of much biomolecular NMR. The NHSQC dataset included Ubiquitin, the core domain of HSP27 [25], the core domain of HSP16.5 (see Chapter 4) and the core domain of αB-crystallin [26]. The latter three are complicated by significant chemical exchange. UnidecNMR demonstrated superior performance to the other algorithms and produced results as good as a skilled spectroscopist (Figure 2.8).

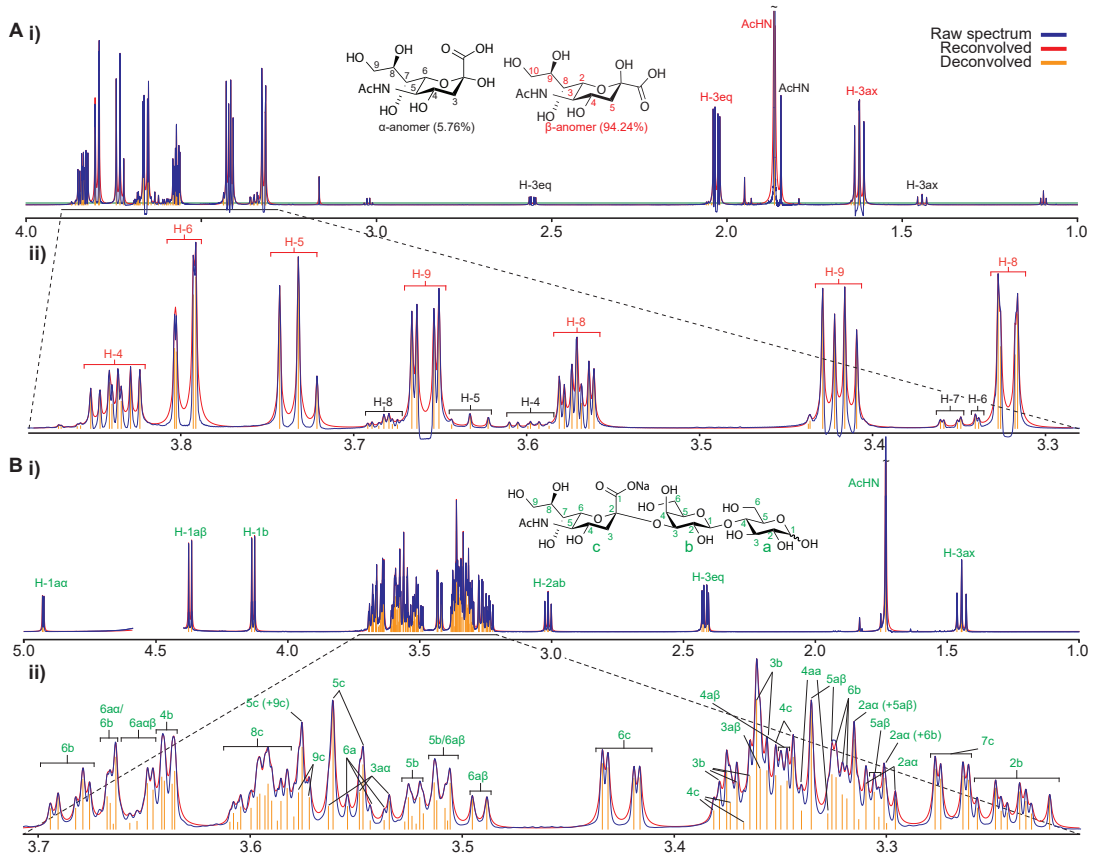


Figure 2.7: (A) spectrum of N-Acetylneuraminic acid is shown (i) containing interconverting α and β anomers both present in the spectrum. Assignments of the deconvolution outputs to α and β anomers are indicated in black and red respectively. (A ii) zoom-in of the sugar region (3.3-3.9ppm). (b) spectrum of 2,3 alpha-sialyllactose (i) with assignments indicated in green. (B ii) zoom in of the sugar region (3.2-2.7ppm) showing the precise deconvolution achieved by UnidecNMR. Assignments were made using selective 1D TOCSY and HSQC spectra, as shown in the uSTA manuscript [24].

2.5.3 Application to 3D backbone spectra

Introduction to assignment experiments

NHSQC spectra provide a structural map of smaller proteins [27] (up to 70kDa with deuteration [28]). However, the peaks need to be assigned to particular residues within each protein. As we expect roughly 1 N-H per residue (excluding side chains and prolines), once this assignment has been performed, the spectrum becomes a residue specific fingerprint of the protein. This atomic resolution can lead to very relevant structural and dynamical biological insight [25]. There are a number of protocols to assign NHSQC spectra, but the most popular involves extending

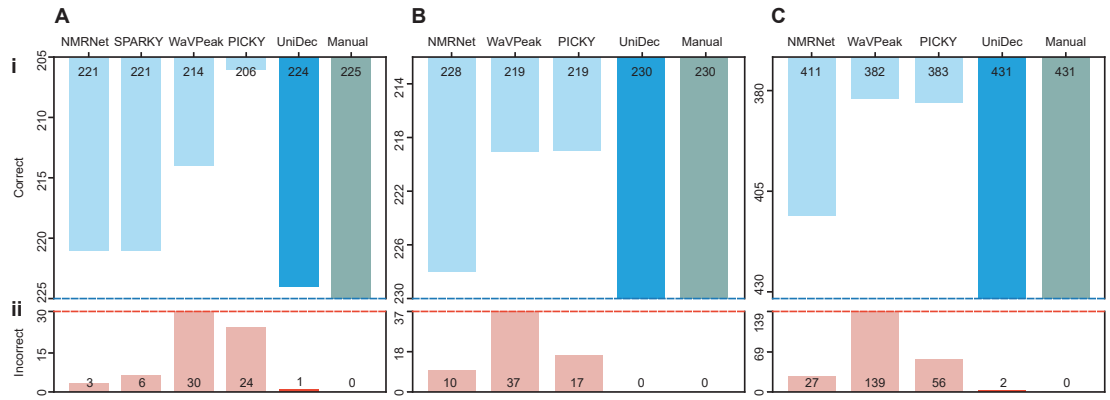


Figure 2.8: Relative performance of UnidecNMR versus a series of peak-picking algorithms. Correctly picked (i) and incorrect picks (ii) are shown against a set of 4 HSQC (a), HNCO (b) and HNCA (c) spectra. The performance of UnidecNMR is substantially improved when analysing the HNCA spectra when a 2D peak list was provided ('boring' mode).

the NHSQC into a third, carbon dimension. This requires uniform ^{13}C and ^{15}N labelling of the protein. 3D experiments of this type, for example the HNCO, HNCA, HNCACO and HNCOCA share Nitrogen and Hydrogen dimensions, with different carbon frequencies recorded in the third dimension[27][29]. The logic of these spectra is explored in Figure 2.9. Projecting these 3D spectra onto the 2D N-H plane, therefore, produces very similar NHSQC like spectra in all of them.

2.5.4 Applying UnidecNMR to 3D backbone spectra

The UnidecNMR approach is very flexible, as potential deconvolution solutions can be allowed/disallowed at particular frequencies based on any prior information. Deconvolved 2D N-H peak locations can therefore be used as a prior restraint for the 3D spectra, meaning deconvolution solutions are only enabled at the N-H peak locations. This limits the search from a naive 3D to a set of 1D carbon spectra, called 'bores', greatly reducing the potential deconvolution solutions. As the same NHSQC peak list can be applied to all 3D spectra, this approach creates clean data structures (all carbon frequencies are organised by N-H peak), whilst also filtering out most of the 3D data. This logic mirrors what human spectroscopists do during a technique called strip-plots [18]. This logic was purposely encoded into the pulse sequence and can be thought of as encoding Bayesian logic within the spectrum.

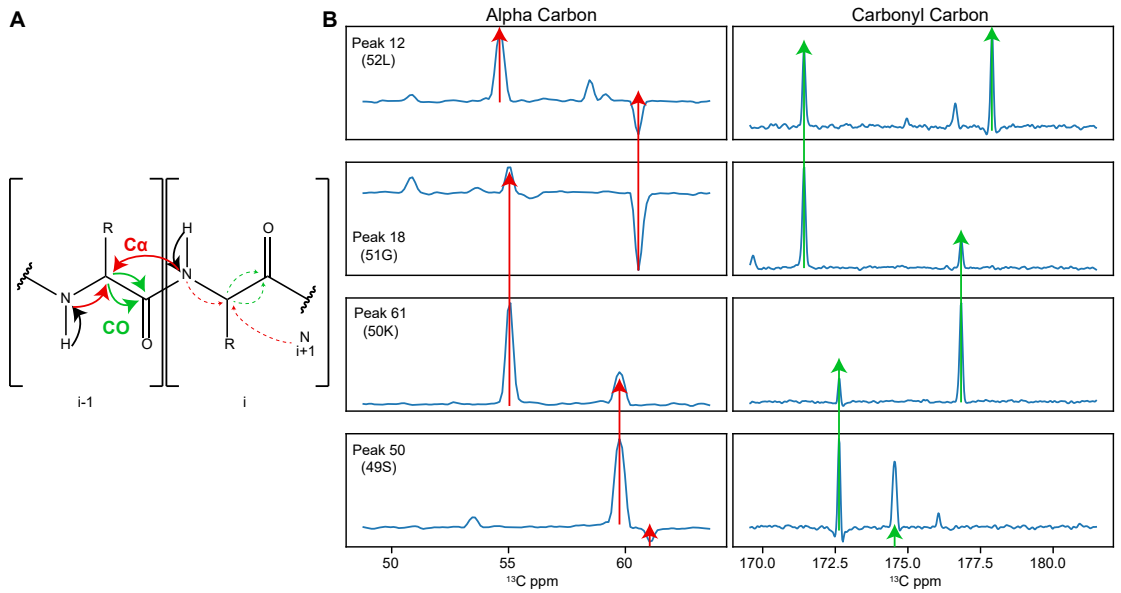


Figure 2.9: The path of magnetisation during the HNCA (black then red arrows in A) and HNCACO (black then red then green arrows in A) experiments is shown (A). Both experiments start with magnetisation on H, before transferring to N and then transferring to Cα. The standard range of coupling constants between ith N and the ith Cα is -8 to -13, while N and the i-1 Cα is -4 to -11. This similarity means that transferring from N to Cα causes a split in magnetisation transfer to both the ith and i-1 Cα. When the carbon dimension is acquired, two (ith and i-1) resonant frequencies are seen. The HNCA indirectly records the Larmor precession of the Cα spins, while the HNCACO does not, instead transferring magnetisation to the adjacent Carbonyl carbon before subsequently recording. This maintains the split in magnetisation, resulting in two (again ith and i-1) carbonyl peaks. Plots of specific N-H peak 'bores' for both the alpha (HNCA) and carbonyl (HNCACO) are shown (B). When the NH bore corresponding to amino acid i is lined up next to the NH bore corresponding to amino acid i-1, the i-1 carbon peaks from the ith bore should line up with the i carbon peak from the i-1 bore. This enables chains of amino acids to be linked and assigned to specific amino acids.

To the best of our knowledge, traditional peak picking algorithms do not employ this logic when analysing a spectrum. We have found this to both greatly expedite data analysis and also exclude significant portions of noise where peaks are unlikely to be present. In the case that this approach misses intensity, this is easily visible by overlaying the resimulated spectrum on the actual data.

In more complex systems, producing an accurate 2D peak-list from the NHSQC alone can be challenging because of peak overlap. The 3D spectra are less likely to have this problem, and the HNCO was used as an additional step to create a more accurate 2D peak list due to its high signal:noise ratio. We designed

the following workflow:

1. Use UnidecNMR to deconvolve the NHSQC
2. Restrain the HNCO using this peak list and run UnidecNMR
3. Adjust the 2D peak list based on the HNCO results

Step 2 identifies any peaks overlapped in the N-H dimensions by spreading them out into the carbon dimension. Step 3 includes these peaks in a final 2D peak-list, which results in a more accurate, master peak list to restrain the rest of the spectra. In the case that there is overlap in all 3 HNCO dimensions, further analysis of the other experiments will show this, and the master list can be readjusted.

2.5.5 Designing a GUI for 2- and 3-D UnidecNMR data

One problem with previous peak picking algorithms is a 'black box' approach, whereby results must be confirmed by manually searching the spectrum, which involves a similar workload to the original peak-picking problem. To resolve this problem, we designed a graphical user interface (GUI) for UnidecNMR. This enables easy implementation of the above workflow, giving greater control over input parameters and visualisation tools for the outputted data. It is organised into tabs, with the first presenting the user with options for a standard NMR analysis workflow: spectral processing, processed data reading, 2D peak list alignment, UnidecNMR input parameters (including peak shape fitting) and a report panel for review.

Spectral processing

The raw free induction decay (FID) from the spectrometer needs processing to produce a spectrum. Spectrometers produce FID data in a variety of formats, the most prevalent being Bruker and Varian. This provides a barrier to entry for a new NMR spectroscopist. Within the GUI, options are provided to let users process 1-4 dimensional spectra easily. It enables phasing (Figure 2.10), the application of different window functions, zero-filling, baselining, linear prediction and processing of non-uniformly sampled data.

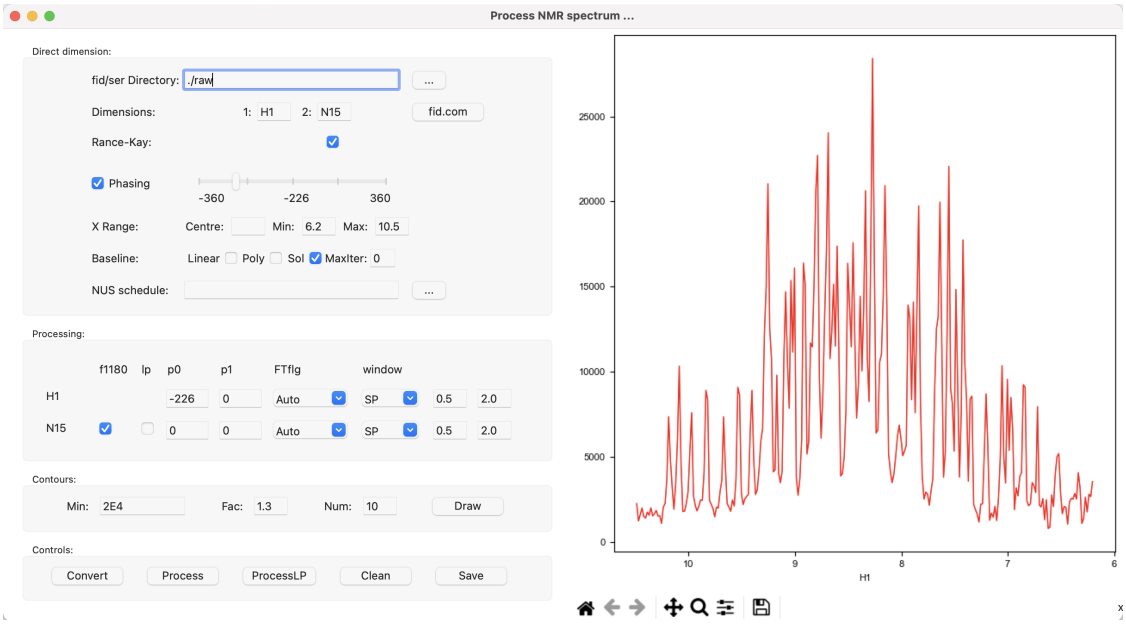


Figure 2.10: GUI input for processing 1-4 dimensional spectra. In this example we are phasing the directly detected (in this case ^{1}H) dimension of a 2 dimensional NHSQC spectrum using the slider.

2.5.6 Applying UnidecNMR to 3D backbone spectra

2D peak list alignment

Sometimes there is movement in the N-H dimension due to incorrect parameter setting or slight heating of the sample (e.g. in any sequences with TOCSY elements that deposit high amounts of radiofrequency into the sample). The ability is provided to adjust a 2D peak list for each spectrum, ensuring that the 1D deconvolutions were at the correct N-H coordinates (Figure 2.11). The GUI includes maximise and move functions that greatly expedite this task. It also enables interactive boring into the 2D N-H projection to visualise the carbon dimension, bringing the data much closer to the user.

UnidecNMR input parameters

UnidecNMR requires a noise threshold and peak-shape function. Without a GUI, defining a peak shape function needs arduous trial and error, involving iterative deconvolution runs. We provide a pop-up window enabling users to adjust the peak-shape dimensions (Figure 2.12). The noise threshold can also be easily

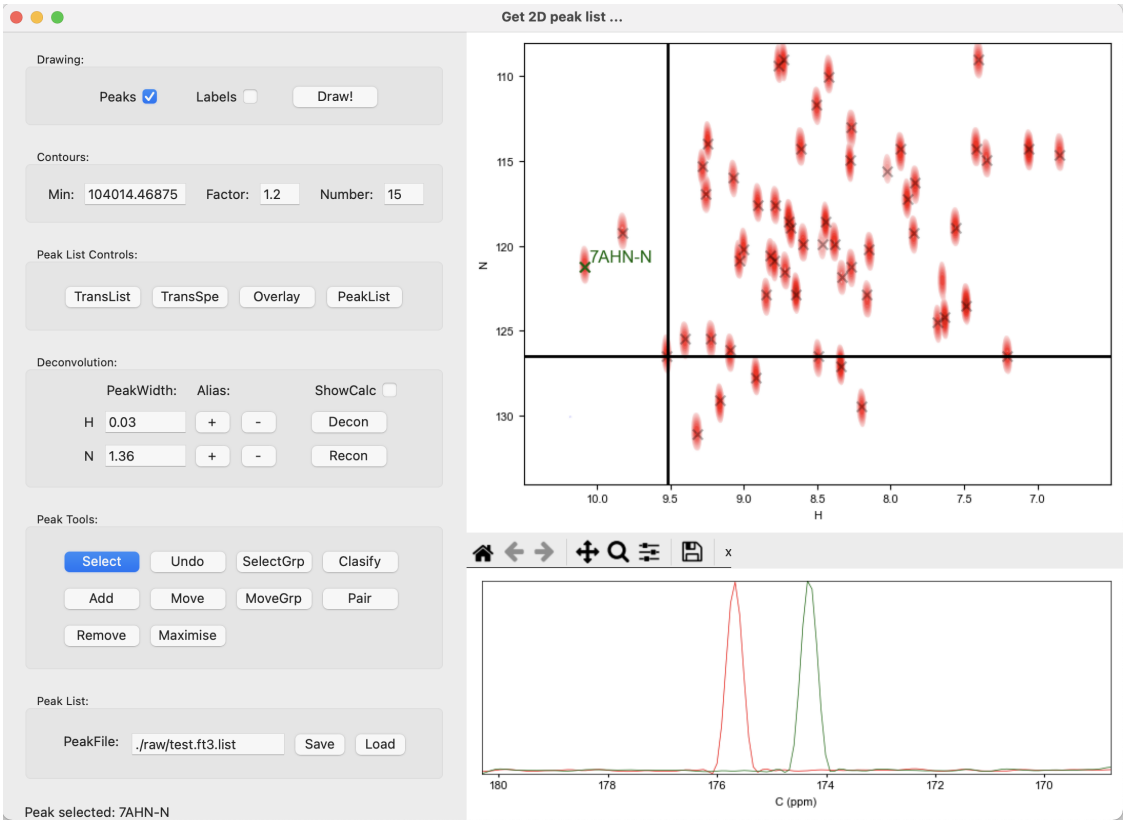


Figure 2.11: GUI input for Applying a 2 dimensional peak list to a 3 dimensional backbone assignment spectrum. The 2D N-H projection is seen in the top right of the figure, while the Carbon dimension corresponding to the crosshair position is seen in red below the projection. The selected peak (green) is also shown in both views. The projection element of the GUI has been blitted to increase performance: browsing the spectrum is now very fluid.

adjusted in the data viewer. Once these have been defined, UnidecNMR can be run and the results analysed.

Data visualisation

Once UnidecNMR results have been loaded, the data can be viewed in a few different ways. A 2D spectrum is shown as a contour plot, on which the identified peak positions and the re-simulated spectrum can be overlaid using toggles (Figure 2.13). A 3D spectrum can be viewed in the following ways: in its entirety; as 3x2D projections; or each bore can be viewed as either 1D or limited view 2D. The equivalent re-simulated spectrum and identified peak positions can be overlaid for easy analysis of all these representations. The limited view 2D bore mode presents

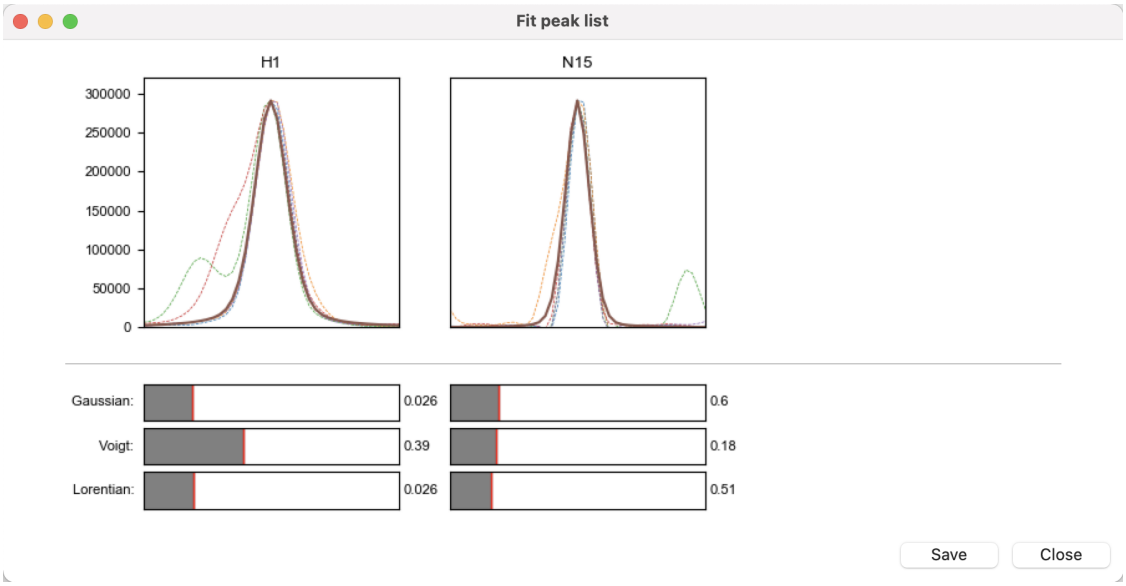


Figure 2.12: GUI input for peak fitting in 1-4 dimensions. Three controls are given for the peak-shape in each dimension: a Gaussian component, lorentzian component and the pseudovoigt parameter which governs the contribution of each. These sliders enable rapid matching of peak-shape to the data and much easier use of the UnidecNMR algorithm.

2 spectra enabling three possible views: 2 separate bores; raw and resimulated data from a single bore; normal and orthogonal 2D views from a single bore.

2.6 Application to 3 and 4D NOESY spectra

An introduction to the logic of Methyl-TROSY NOESY spectra

Protocols involving uniformly deuterated proteins with selective side chain $^{13}CH_3$ labelling have been devised to enable NMR studies of larger proteins (up to 1MDa)[30] [31]. They have been used to show conformational changes of biologically relevant systems, giving unparalleled insight into the importance of dynamical behaviour. Crucial to this process, like the backbone experiments, is a correct assignment of the spectral resonances. This is often performed using time-consuming and expensive mutagenesis, in which labelled residues are sequentially mutated out. A potentially quicker and more cost-effective approach is to graph match 3D or 4D NOESY spectra with distances from a predetermined crystal structure [32]. A critical step in this approach is peak-picking the spectra, which is currently extremely time-consuming. UnidecNMR can solve this problem by first deconvolving

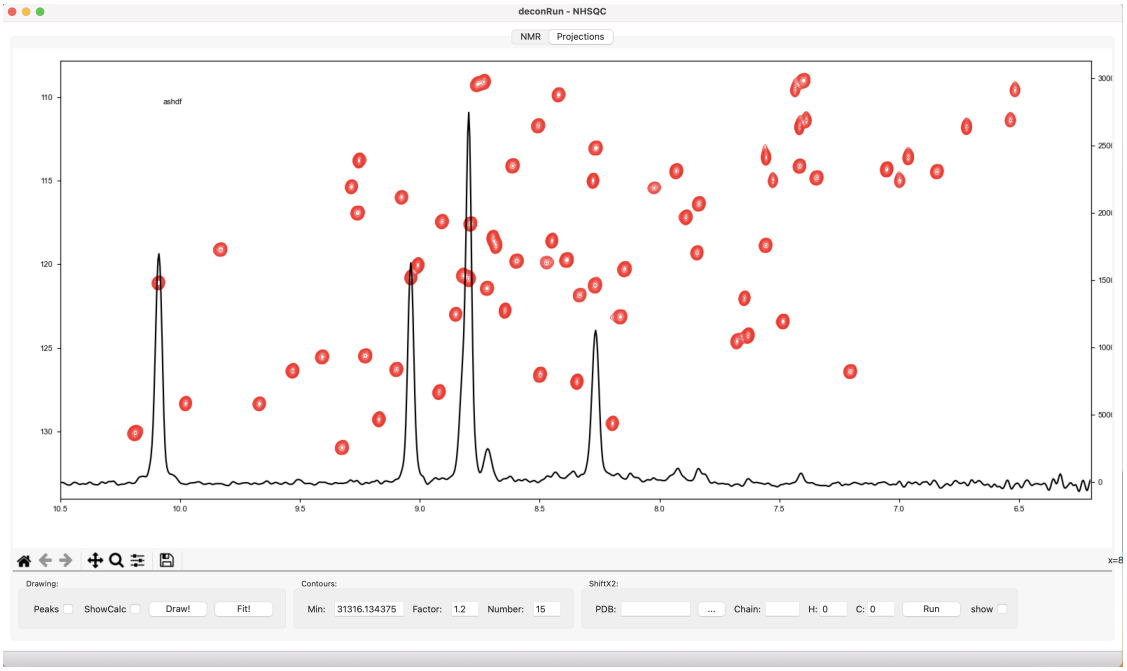


Figure 2.13: 2D data visualisation is very easy with UnidecNMR GUI: the data can be scrolled through horizontally (black) and vertically, and peaks and resimulated spectra can be overlaid.

a 2D C-H peak list from a high-resolution HMQC spectrum, which can then be used to restrain the 3D or 4D deconvolution (similar to the backbone N-H peak list). This limits the search in 3 and 4D spectra to 1D and 2D bores respectively. Within those bores, peaks are only possible at certain locations, also defined by the HMQC peak list, which further demonstrates the value of the flexible UnidecNMR approach. This is discussed and demonstrated in more detail during an application in Chapter 4.

4D UnidecNMR visualisation

Producing a 4D peak list without visualisation tools would make the results difficult to validate, so the GUI includes modes specific to this problem. 4x2D projections are viewable on one tab, alongside identified peak positions and the deconvolved spectrum. It is relatively easy to gain confidence in the UnidecNMR results through comparison of the resimulated spectrum and the raw data. We also created a viewer that enables iteration through the 2D bores, with options to compare the resimulated spectrum with the original, and view different sub-spectra side by side.

Methyl-TROSY NOE Results

We were able to input the UnidecNMR results straight into the Methyl Assignment by Graph Matching (MAGMA) algorithm and successfully assign the data (Figure 2.14[33]). The 3D data were provided for Aspartate transcarbamoylase (ATCase) by Prof. Lewis Kay's lab [8], while the 4D data of E-coli Enzyme I (EIN) were provided by Prof. Marius Clore's lab[34]. Astonishingly, for the 3D spectrum provided by Lewis Kay's lab and the 4D spectrum of EIN, UnidecNMR produced even more reasonable NOE restraints than an experienced spectroscopist, highlighting its utility. It should be noted that while the 14Ådistance could be considered a 'long-distance' NOE for protonated samples, ATCase are uniformly deuterated except specific methyl groups. Hydrogen-Deuterium cross-relaxation is much less efficient meaning pathways for cross-relaxation are substantially reduced [35]. This enables long distance NOEs, and the particular case of ATCase is further explored in the MAGMA paper [33]. This assignment strategy also relies on crystal structures accurately resembling solution conditions. Long-distance NOEs are physically feasible in this case, although it could be indicative of crystallography artefacts.

2.7 Conclusions

The authors of the original 'triple resonance' 13C/ 15N /1H backbone assignment experiment, HNCO, noted that "because of the low level of resonance overlap in the 3D spectra, much of the peak picking can be done in a fully automated manner." [27]. Despite the widespread adoption of the techniques, fully automated analysis has not yet been achieved.

UnidecNMR provides a significant step forwards in automating the data analysis of 1-4D NMR spectra. A user must supply two parameters: a noise threshold and a peak shape to back-calculate the spectrum. The determination of both is assisted by the software, but manual iteration is also feasible. The raw data and resimulated result use the nmrPipe [10] format. Deconvolved peak lists and resimulated spectra can therefore be viewed in either our GUI or other nmrPipe compatible software.

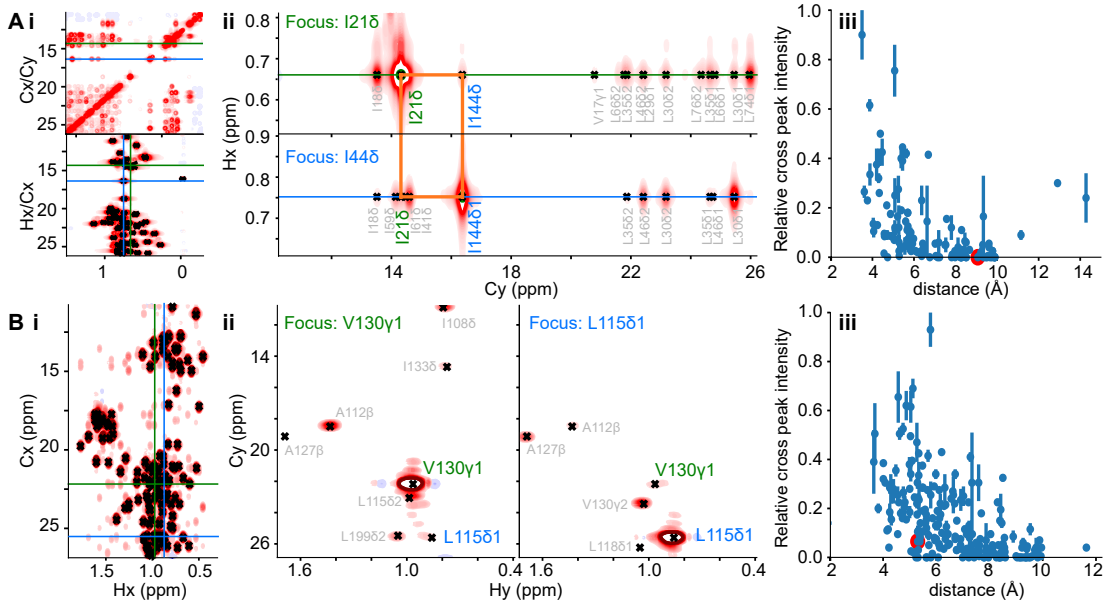


Figure 2.14: Results from a 3D methyl NOESY spectrum from ATCase (a) and a 4D methyl NOESY spectrum of EIN (b). These figures were generated from outputs within our GUI. i) The location of two selected slices are indicated, with respect to the relevant projections. It is desirable when analysing 4D spectra to take 2D slices that have lower resolution than the reference planes that include the direct dimension, as shown for EIN. ii) The corresponding slices focusing on identified NOEs from a pair of resonances, and a cross peak between them. The specific cross peak feature is indicated (orange). Orthogonal views can be selected within the GUI allowing a user to verify that all resonances are centred on the selected plane. The deconvolved version of the spectrum can also be shown side-by-side with the raw data. iii) The cross peak signal intensity, shown as a mean and standard deviation of the two reciprocal resonances, shown versus the expected C-C distance from the corresponding structure. 460 pairs of cross peaks were identified for ATCase and 660 for EIN, 60% more than obtained from a manual analysis. The picked peaks fall within a sensible range of distances indicating that the NOEs are consistent with the expected structures.

The performance of UnidecNMR is substantially better than competing algorithms for both the synthetic and real datasets tested in this work.

The ability to allow/disallow deconvolution solutions enables the logic underlying the spectra to be used as restraints, taking a different approach to competing algorithms and mirroring the techniques of a manual spectroscopist. 2D peak lists limit the solutions of 3 and 4D applications substantially, and the symmetry requirement can be imposed on NOESY spectra. The resulting peak lists are in a format which enables assignment [32] and structural calculations.

The design of a Graphical User interface prevents the technique from being a

'black box', and enables spectroscopists to iterate through, and edit, results. We employ the nmrGlue [36] and Matplotlib [37] python packages in this software. The deconvolution algorithm is written in C++ for speed, and can be executed from the command line to enable separate automated workflows. The package has been tested in Linux, Mac and Windows.

Overall, against the experimental data tested here, UnidecNMR either matches, or significantly outperforms an experienced spectroscopist. The GUI enables users to process, deconvolve and visualise spectra within a few clicks. UnidecNMR can also be included in a fully-automatic pipeline, although manual inspection of the results is recommended. UnidecNMR substantially reduces the time required to quantify spectra, and can hopefully accelerate the work of both new and experienced spectroscopists who are using biomolecular NMR to address a research problem. The software and datasets will be free for academic use and available for download.

Having developed an accurate peak picker, the next step was to start implementing this into experiments that would benefit from accelerated workflows.

2.8 Materials and Methods

The core domain αB Crystallin sample was expressed by Olya Tkachenko during her DPhil and her thesis details the protocol including sample preparation and NMR experiments [38]. The core domain of hsp27 was expressed by Reid Alderson during his DPhil and the protocol including sample preparation and NMR experiments are published [39]. Hsp16.5 was expressed by Michael Barber during his DPhil and his thesis details the protocol [40].

3

Automated Assignment Strategy

3.1 Designing the Assignment strategy

As discussed in the previous chapter, assignment is a critical step to probing protein structures and dynamics with NMR. In the original triple resonance assignment paper, Ad Bax noted, "Because of the low level of resonance overlap in the 3D spectra, much of the 3D peak picking can be done in a fully automated manner" [27]. Currently, the slowest part of the process is accurately peak picking the various spectra. On model proteins, this is easy and automatable, but many biological samples are complicated by chemical exchange and low-solubility. Spectral assignment, therefore, remains a largely manual effort. We hypothesised that intelligent application of UnidecNMR could fix this problem.

Having designed the UnidecNMR algorithm and GUI to provide well organised data suitable for triple resonance assignment, the next step was to create a tool to collate that data. There were four goals for this software.

1. Confidently assign as many peaks automatically as the data allow and collate ambiguous options.

2. Present ambiguous options to facilitate user intervention

3. Create data visualisation structures within the GUI to enable confidence in the results

4. Automatically produce a self-contained set of figures for inclusion in supplementary information of publications

It was rationalised that this approach would both create an extremely useful tool to complete the assignment and raise the level for data transparency within the field. This chapter describes the development and design of this algorithm around a simple test case: the ABP1P SH3 domain required in Chapter 5. The software was also used in Chapter 4 to assign the very challenging exchange-broadened hsp16.5 ACD construct.

3.1.1 Data structure

The UnidecNMR algorithm was at the core of the design. As described in Chapter 2, the use of restricted 'boring' mode within UnidecNMR produces a clean data structure organised by self-consistent peak labels. The first step is to define a master peak list using the protocol described in Chapter 2. This could then be applied to the rest of the triple resonance spectra. Subtle changes to the 2D peak list can be made within the UnidecNMR GUI to create spectrum specific versions of the master peak list. Once the UnidecNMR algorithm has been run on all available spectra, the data can be loaded into the assignment GUI.

3.2 Initial Assignment algorithm

Automatically upon loading, the algorithm goes through a few standard steps (CO chemical shifts are used as an example):

1. Carbon axis spectral pairs are aligned: for example the largest peak of each bore of the HNCO is extracted (which should only contain one peak, the i-1) and the smallest peak from each bore of the HNCACO (which should be the i-1 peak if it is visible above the noise). The difference for each bore is then

calculated and the median taken as a reference number. All peaks from the more complicated experiment (i.e. the HNCACO in this case) are moved by this reference value.

2. In the case of the F1180 trick [41], where a peak is negative, the unfolded peak position is calculated.

3. The referenced (and unfolded) peaks are preliminarily classified according to the logic of the experiment. For example, the HNCO should only have one peak per bore (i-1), while the smaller peak in the HNCACO will be classified as the i-1 peak. Some error catching logic is used here: If there are two peaks in the HNCACO, but the larger peak is in the same location as the HNCO peak, the larger HNCACO peak will be redefined as the i-1 peak and the smaller the i-th peak. If there is only one peak in the HNCACO, and that does not match the HNCO position, a 'virtual peak' is added to the HNCACO spectrum at the HNCO position. The visible HNCACO peak is then classified as the ith peak.

4. Potential connectivities between peaks are identified. The designated ith peak of the HNCACO should match up with an HNCO/HNCACO peak of another bore. All possible connectivities within a particular chemical shift tolerance are identified and saved.

5. We check for potential forward/backward connectivities of the peaks. For each given N-H bore we have two peaks per carbon: the ith and the i-1 peak. The ith peak should be visible in the i+1 bore, while the i-1 should be visible in the i-1 bore. We check that the peak is visible in both directions.

6. The BMRB chemical shift statistics have been embedded into the code. Using the dataset of carbon, nitrogen and hydrogen chemical shifts for each peak, the amino acid probabilities are calculated for the ith and i-1 peaks. Using some gentle exclusion criteria, possible amino acids are included alongside connection information to make a graph of connectivities. Alternatively, if

there are SHIFTX2 [42] data, they can be read here to make a separate graph for MAGMA.

7. The submission script for the MAGMA graph matching algorithm is written. The target graph is the sequence broken by N-H invisible prolines, while the matching graph provides possible connectivities with some excluded due to chemical shift probabilities.

While the Carbonyl chemical shifts from the HNCO and HNCACO have been used as an example, the software can also take the equivalent Cα and Cβ experiments, HNCA, HNCOCA, HNCACB and HNCOCACB respectively. In the case of an IDP with reduced carbon chemical shift dispersion, the option for the HNCANH and HNCOCAHN experiments has been added. A similar approach was taken despite slightly different logic for these experiments. These steps assemble first inferences from the UnidecNMR outputs and automatically generate scripts in the correct format for downstream analysis. This completes the first design criterion.

3.3 Presenting ambiguous data

The user is presented with the data in the most logical format to enable editing. Both confident and ambiguous connections can be iterated through. There are often missing peaks and logical ambiguities present in complex systems, requiring manual intervention. The walk (Figure 3.1) and spins (Figure 3.2) frames provide this ability. The walk frame presents the C, Carbonyl and C beta data (and NNH data from the IDP) of three N-H bores, the spins frame presents data for one N-H bore, including all the chemical shift data and the corresponding possible connections. These connections are displayed as a connectivity graph. Iterating through N-H peaks using the spins frame will automatically update the walk frame, enabling manual confirmation and intervention. The spins frame will present all options that meet the criteria from points 1 to 5 above, with a user-defined threshold for differences in chemical shift between ith and i-1 peaks. For example, we often present all carbon peaks that are within 0.3ppm when trying to match the ith peak

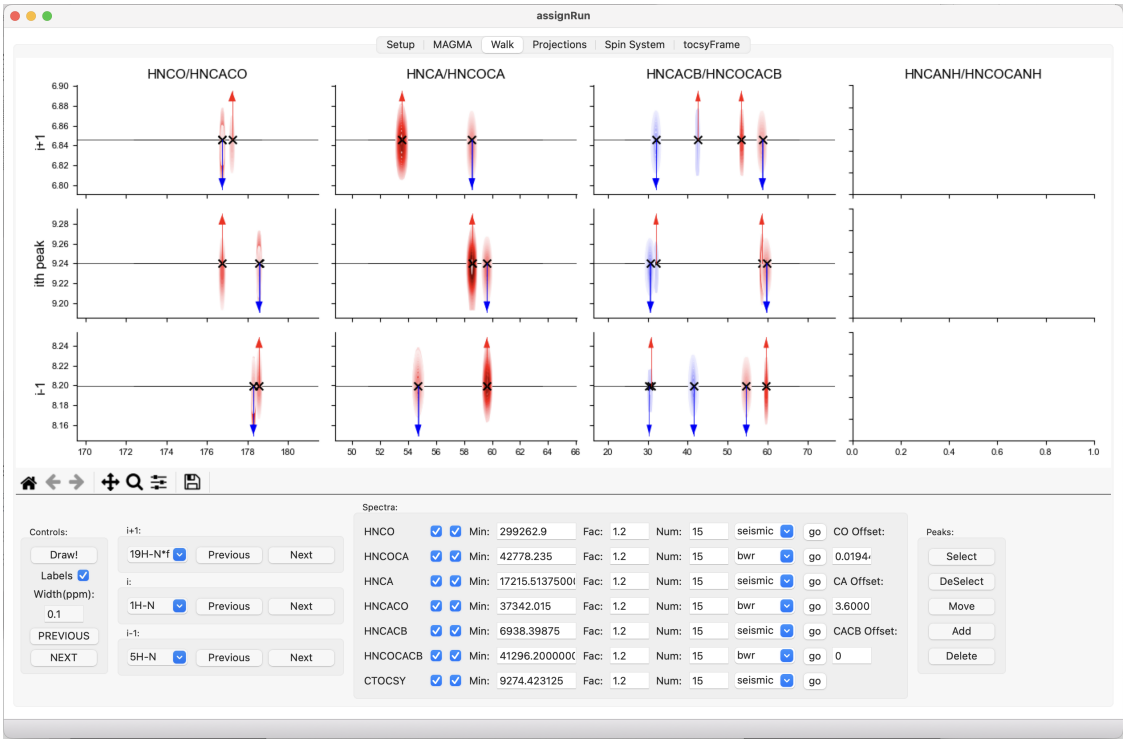


Figure 3.1: The GUI window for walking through the assignment, showing alpha, carbonyl and beta spectra overlaid. The flow of the peaks is easy to identify due to the arrows, which are designed as a visual cue. Each of the i-1, ith and i+1 peaks can be changed in the GUI controls to compare overlap. Erroneous assignments are easy to detect as they do not line up well in all spectra. There are display options, including noise thresholds, offset controls (initial values are calculated automatically), and peaks can be added, moved and removed.

from the HNCACO with a corresponding i-1 peak from the HNCO. Combining this with the data from Cα and Cβ excludes peaks that are close by chance. The closest match preferred and only impossible links removed. For example, if there is good overlap in the carbonyl peaks but no overlap in the Cα peaks, this option is removed. Conversely, if there is good overlap in the carbonyl peaks and missing data for the Cα peaks (perhaps due to exchange broadening), this option will not be removed. We designed this to be as similar as possible to what we would do as manual users. The spins frame tabulates the dataset of chemical shift information for each bore, showing Hydrogen, Nitrogen and Carbon chemical shift information. The graph plot can be switched to chemical shift, which shows the probability of a particular amino acid given the set of chemical shifts. There is the option to load in and/or calculate SHIFTX2 data from a crystal structure to further direct assignment. This

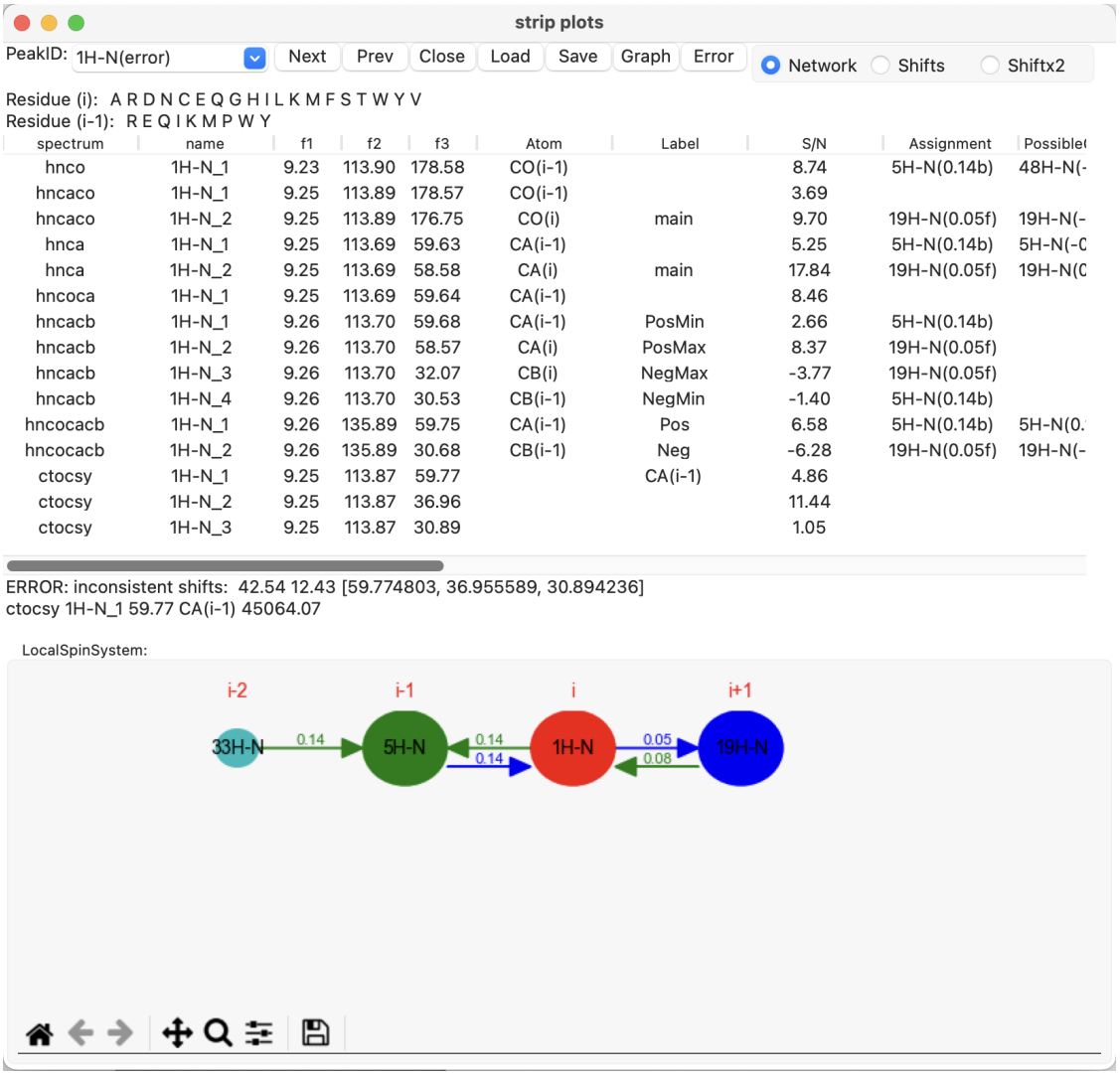


Figure 3.2: The spins frame shows the database of chemical shift information and the connectivity information.

produces a plot of probabilities of each residue in the chain, based on the similarity of the experimental chemical shifts to the SHIFTX2 calculated shifts.

In the case that N-H peaks are missing from the graph, the user can look for missed peaks in the projection pane, which aligns all N-H projections and displays the bores, enabling the user to add/delete peaks from the master peak list (Figure 3.3). This feeds down to the particular peaklist for each spectrum in the dataset. This data browsing enables the user to identify whether a potential peak is a piece of noise or significant intensity. If a new peak is added to the master 2D peak list, UnidecNMR instances can be started from the GUI to rerun the peak

picking algorithm iteratively over all spectra. This data visualisation and ability to re-run the deconvolution algorithm helps to address the risk that the Bayesian approach results in missing peaks and data.

Once spin systems have been created, the user can use two tools to try and assign the systems to the sequence. The first is the MAGMA graph matching algorithm, which has been updated to include a 'backbone mode'. This mode includes the directionality of the graph, which is not relevant to the previous, methyl-NOESY, use case. During the algorithm above, we continually update the NMR graph to include the logic that we have deduced. This graph is then fed to the MAGMA algorithm, which exhaustively searches all possible combinations and arrangements of the subgraph until it finds one that matches. This algorithm therefore greatly benefits from extra connectivities being found: breaking down just 2 subgraphs into smaller pieces dramatically increases the calculation time. The second mode is more interactive, which can provide assignment guesses to the user and enables manually assignment of the spin systems. Where the chemical shift data don't agree the interactive mode will alert the user. The interactive mode can be used to restrict and greatly expedite the MAGMA search.

Once the assignment has been completed by using the interactive or MAGMA modes, the spin-system pane has the option to read in the assignment and produce standardised PDF outputs that can be inserted into any supplementary (Figure 3.4). Chemical shift and connectivity information is included in these generated figures, enabling the reader to gain quick insight and confidence in the assignment. It is hoped that this presents a more rigorous data presentation than the currently standard peaklists.

It was easy to apply this algorithm to the spectra of a test protein, the SH3 domain of ABP1P. The algorithm unambiguously assigned all the residues with naive UnidecNMR inputs and gave us a chemical shift database for the whole protein. This process took a mildly experienced master's student around 30 minutes.

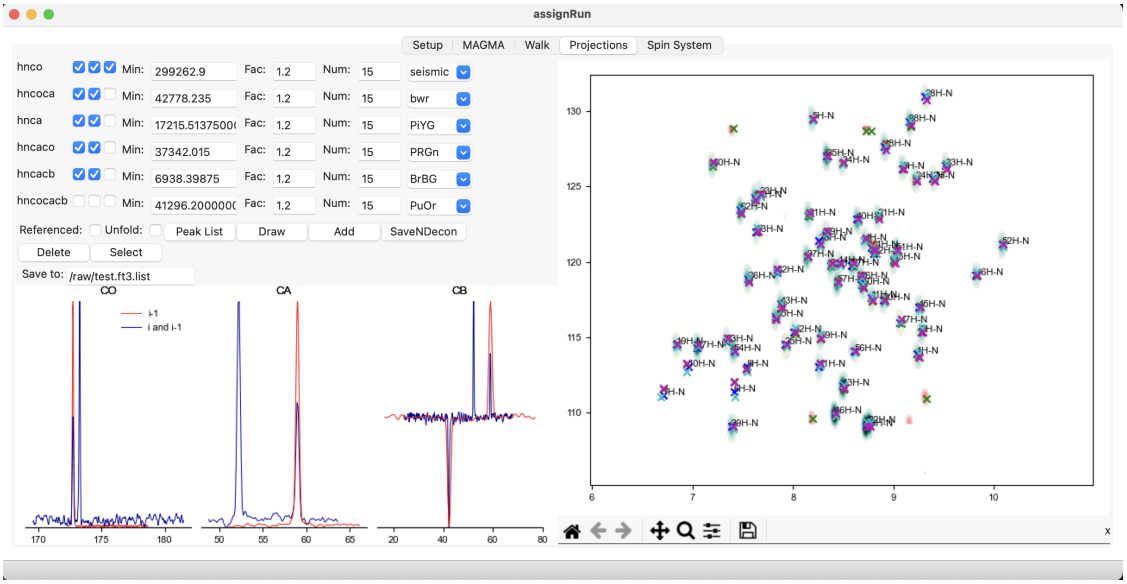


Figure 3.3: The GUI tab for Boring through the multiple assignment datasets simultaneously to make sure the data make sense. We can add and remove peaks here and spawn new iterations of UnidecNMR to re-run the full algorithmic workflow. There are options to view some/all 2D projections and their 2D peak lists alongside noise thresholds and contour colours. Bores for each of the Carbon axes are also shown, with i-1 and i/i-1 spectra shown. These are rapidly scrollable with effectively no lag, enabling millions of data points to be navigated at will. We can fold and unfold peaks in the Nitrogen dimension according to their sweep widths, and reference the peaks according to the nitrogen referencing offsets that are manually calculated.

3.4 Conclusions

The design of this assignment strategy around the UnidecNMR algorithm appears to be highly effective in both generating an assignment, and providing a visualisation for both student and PI to inspect the results. This will be made available to the NMR community where we hope assignment on challenging and interesting biological systems can be performed by non-specialist research groups.

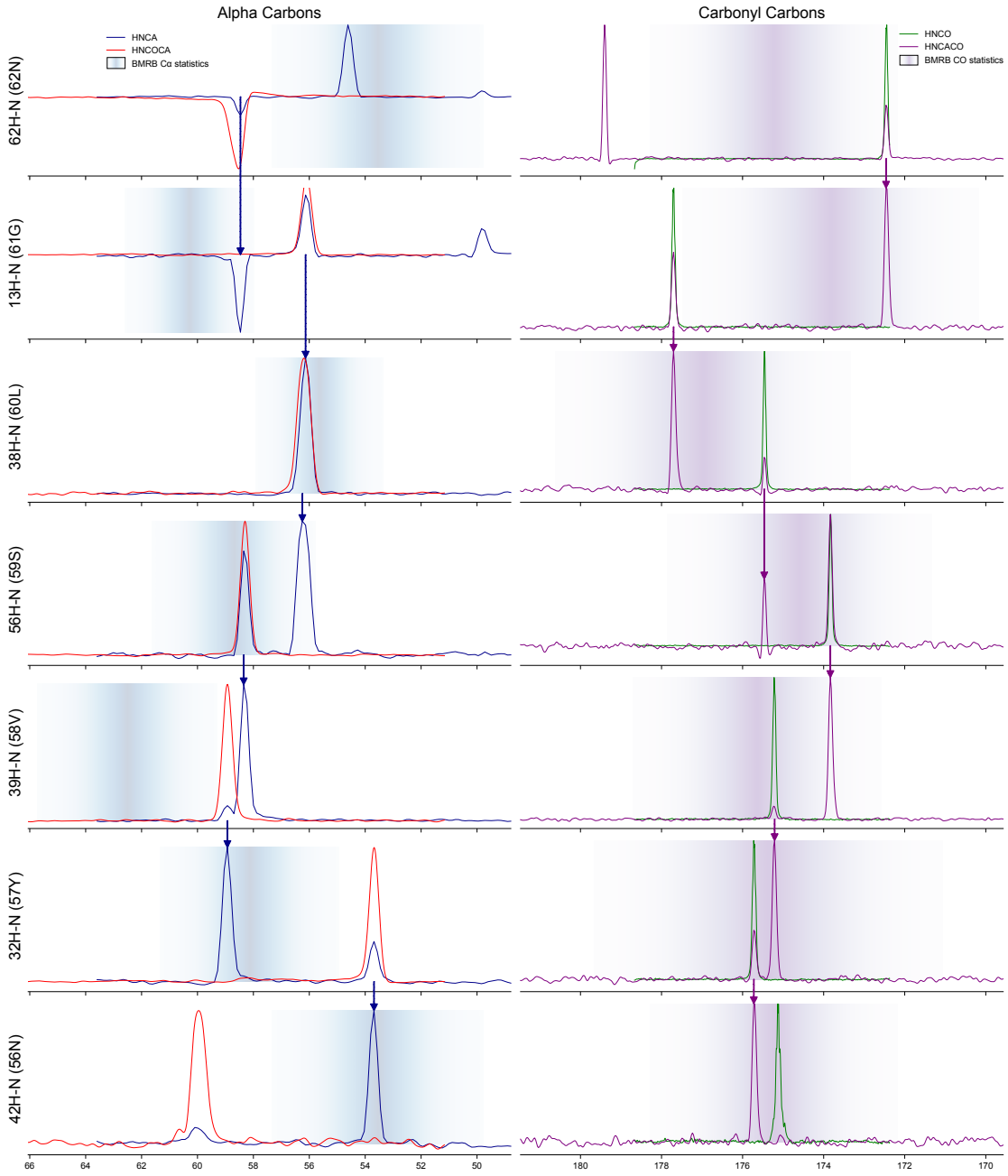


Figure 3.4: A section of the standardised assignment figure. Peak numbers and assigned residue numbers are shown on the y-axis. The logical links between ith and i-1 residues are shown with arrows for both the alpha carbon and carbonyl experiments. If there is a significant mismatch, it would be obvious. The average chemical shifts and their standard deviations, derived from the BMRB statistics, are shown for the ith peak of the assigned residue. Where there are outliers, it is obvious and can be examined further.

4

Studying the dynamics of a 396kDa sHSP oligomer at atomic resolution

Contents

4.1 Acknowledgements 38
4.2 Introduction 38
4.2.1 Small heat shock proteins 38
4.2.2 Methyl-TROSY NMR 41
4.3 Spectra 45
4.3.1 Initial temperature ramp 45
4.3.2 Assignment of MILV labelled full length protein 45
4.3.3 UnidecNMR application to 3D NOESY spectra 49
4.4 Dynamics experiments 51
4.4.1 The fitting problem 53
4.5 CPMG data 54
4.5.1 Comparison with alphaB Crystallin 55
4.5.2 Domain switching dynamics 56
4.5.3 N-terminal domain 57
4.5.4 Spin-state selective CPMG 65
4.5.5 α-lactalbumin 66
4.6 Alpha-crystallin domain 67
4.6.1 Comparison with full length protein 68
4.6.2 Assignment of ACD 71
4.6.3 Temperature Ramp 71
4.7 Conclusions 74
4.8 Future work 75
4.9 Materials and Methods 76

4.1 Acknowledgements

My focus in this study was the data analysis of historically recorded datasets. Prof Andrew Baldwin created many of the samples and recorded almost all of the data starting while he was in Lewis Kay's lab in Toronto and continuing when he moved to Oxford. Michael Barber and Reid Alderson produced the alpha crystallin domain (discussed later) samples, while Reid also produced a few of the Methyl TROSY labelled samples.

4.2 Introduction

The next step was to design a UnidecNMR approach to the analysis of high molecular weight structural and dynamical studies. Generally, this involves psuedo3D relaxation data, thermal titrations and NOE spectra which all need to be peak picked and the information synthesised to understand the behaviour of the system in solution. The focus of this study was Hsp16.5 rather than tool creation, but the tools greatly expedited the work. Some tools in this chapter enable the measurement of the local environment of the excited state of the protein. Enabling biochemists to access this information could greatly increase understanding of protein dynamics.

4.2.1 Small heat shock proteins

The delicate 3D protein structures that underpin our cellular machinery are thermodynamically less favourable than large, toxic aggregates [43]. If aggregation occurs, it is often irreversible. Fortunately, this process is hindered by kinetic barriers; typically proteins must misfold into a higher energy state before aggregation. Despite this, cells still require defence mechanisms against protein aggregation. One of these, chaperone proteins, can assist other proteins to fold during expression, refold after denaturation and stabilise off-pathway unfolded states to prevent conversion to aggregation-prone states. While other heat shock proteins (e.g. HSP70 and GROEL)

require ATP to refold proteins actively, small heat shock proteins (sHSPs) are ATP-independent and bind to misfolding proteins early in the chaperoning process. They have been referred to as the "paramedics of the cell" due to their role in regulating protein mis-folding and degredation[44]. However, it has also been shown that over-expression of small heat shock proteins has deleterious effects, as they stabilise toxic aggregates in cells and are prone to aggregation themselves [45] [46]. They have been related to many diseases[47] including neurodegenerative amyloidosis [48]. Significant effort has already gone into understanding both mammalian and plant small heat shock proteins [49] [50] [51].

They have proven challenging targets for atomic resolution study. Mammalian sHSPs generally form polydisperse oligmers, resulting in a range of structures and isoforms. Conversely, plant and archaea sHSPs are generally monodisperse. This has been hypothesised to be a regulation mechanism, whereby the oligomeric states display less chaperone activity than the monomer/dimer subunits, preventing the deletrious effects of over-expression [45] [46]. While there is some crystallographic data on the monodisperse small heat shock proteins, information about their inherent dynamic assembly and function is sparse. More generally, the role of dynamical behaviour of proteins is poorly understood in comparison to the vast array of crystallographic structures in the protein data bank (PDB) [52].

The monomeric proteins are made up of three domains: N-terminal (NTD), alpha-crystallin (ACD) and C-terminal (CTD). The ACD is highly conserved between paralogs, while the NTD and CTD are more variable [49]. The ACDs have generally been shown to be structured and include a beta sandwich region, while the C- and N-termini have proven more difficult to study. Each domain appears to play an important role in oligomerisation. The ACD homodimerises to create a building block for the oligomers[26]; mutating the N-terminus has been shown to shift the polydispersity distribution[53] and the C-terminus contains an IXI motif (or IPV) that, through mutational work and NMR study, has been shown to regulate polydispersity[7].

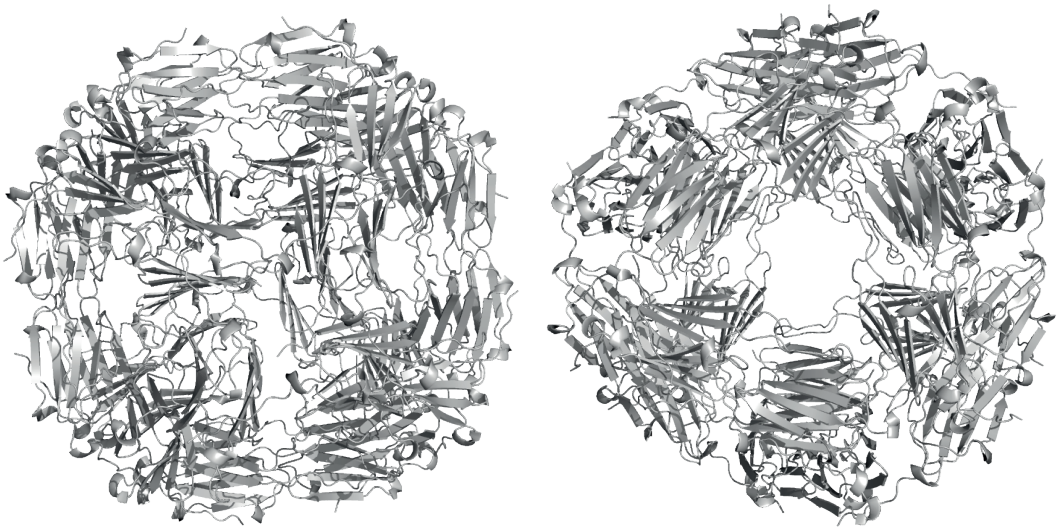


Figure 4.1: HSP16.5 X-ray Crystallography-derived structure from PDB 1SHS showing the 24-mer assembly of the protein. All monomer environments are symmetric and form dimers to create the edges of a 'hollow spherical complex with octahedral symmetry' [54]. The first 33 residues of the N-terminus are unresolved.

MjHSP16.5

Methanocaldococcus jannaschii (Mj) heat shock protein 16.5 (Hsp16.5) forms a monodisperse 24-mer[54] (Figure 4.1). It was the first small heat shock protein to be crystallised, showing it adopts octahedral symmetry with 12 copies of the dimeric environment per 24-mer [54] producing a sphere. These monomers are shown to be linked through two interfaces. The monomer-dimer interface is defined by an unstructured loop of one monomer making contact with the base of a beta sandwich (B·S) in another. This interaction is symmetric and known as a 'domain swap', where two monomers have exchanged structural elements to create a dimer[55]. The second interface is defined by the final 6 C-terimnal residues of one monomer embedding into a groove at the top of the ACD beta sandwich of another. These are arranged in the sphere so that the CTD-ACD interaction is on the exterior of the shpere, while the loop-B·S interaction is on the interior. The first 32 residues of the NTD are not resolved in the crystal structure, which was reasoned to be indicative of disorder. Based on the location of the 33rd residue, however, they are assumed to be contained within the central cavity of the sphere. While the crystal structure

paints an initial picture of the oligomeric assembly, we wanted to use NMR to probe the 'slow' (ms) and 'fast' (ns/ps) dynamics of the protein to understand how these might relate to oligomerisation and chaperone ability. As Mj is found in volcanic environments, the protein is stable beyond 80 degrees, which, when combined with the symmetry, makes Hsp16.5 an excellent target to study by NMR.

4.2.2 Methyl-TROSY NMR

Classical techniques to dissect protein structures and dynamics with NMR generally centre around backbone U-[^{15}N]-^{1}H HSQC based spectra [27]. As the size of the protein increases, however, the tumbling of the molecule slows down. This affects the correlation time of spins within the protein, meaning that two spins have a higher chance of being static with respect to each other. This is required for transverse cross-relaxation pathways to become efficient, so as the size of the protein increases, the initially coherent signal spreads out in the transverse plane more quickly, disabling the ability to record the oscillating signal for a long time. This causes peaks to broaden substantially within the fourier transformed spectrum, and resolution and intensity are lost. The first solution to this came from Kurt Wüthrich's lab, which devised the idea of ^{15}N-^{1}H Transverse Relaxation Optimised SpectroscopY (TROSY) NMR [56]. This relies on fully deuterating the protein before exchanging the amide protons with solvent. Heteronuclear ^{1}H-^{2}D cross-relaxation is much less efficient than homonuclear ^{1}H-^{1}H, as it scales with gyromagnetic ratio. Therefore the cross-relaxation pathways are minimised. The cross-relaxation rate is also dependent on the strength of the dipolar interaction. In the case of ^{15}N-^{1}H, this can be minimised by exploiting the TROSY effect, which balances chemical shift anisotropy against dipole-dipole coupling. In large proteins, this causes some spin states to relax more slowly and correspondingly produces much higher resolution spectra.

The principle of transverse relaxation optimisation was extended by Lewis Kay's lab to side chain ^{13}C labelled methyl groups, including alanines, valines, leucines, isoleucines, threonines and methionines [30]. This again requires a fully deuterated protein excluding $^{13}CH_3$ groups on side chains. Therefore, isotopically

labelled amino acid precursors were devised to be efficiently incorporated into proteins through *E. Coli* metabolic pathways [57]. Once the proteins are purified, this enables high-resolution spectra of the methyl region to be recorded, generally using experiments based on the heteronuclear multiple quantum coherence transfer (HMQC). The methyl groups benefit from the more mobile environment at the end of side chains, which slows relaxation; they also have three protons in the same magnetic environment, increasing their sensitivity to NMR. Methyl-containing amino acids are generally well distributed within proteins [58]. The HMQC experiments have the minimal number of pulses that perform, by definition, the mixing of states. Part of the benefit of TROSY is that spin states within the methyl group have different relaxation properties, some faster and some slower. As the protein becomes larger, these differences become more pronounced and the slowly relaxing states produce substantially better spectra. Pulsing spins causes interconversion between states, meaning an initially defined, slowly relaxing state will also spend some time in faster relaxing states. As an example, the inclusion of an extra 360 degree pulse during a TROSY pulse sequence would mix the fast- and slow-relaxing states, causing a significantly lower amplitude FID during acquisition. When exploiting the TROSY effect, therefore, pulses must be carefully considered. Carbon decoupling is switched off during acquisition as it will continually mix states causing much faster relaxation.

There are a few highly cited examples of successful dynamical studies of large proteins using NMR [59] [60] [14] [7] [61]. It remains, however, the work of highly specialised BioNMR groups. This is partially because there are few tools to expedite the data analysis, despite significant work to develop isotopic labelling pathways and precisely apply spin physics to pulse sequences.

In this chapter, I probe the oligomeric Hsp16.5 assembly at atomic resolution using Methyl-TROSY (Figure 4.2A) and discover the dynamic nature of the interior of the sphere (Figure 4.2B). This millisecond timescale behaviour extends from the monomer-dimer interface (loop and Beta 'sandwich' interface, henceforth abbreviated to 'loop-B·S') to the unresolved N-terminal domain (Figure 4.2B). Unlike a previous study of the polydisperse αB-crystallin, the C-terminal domain,

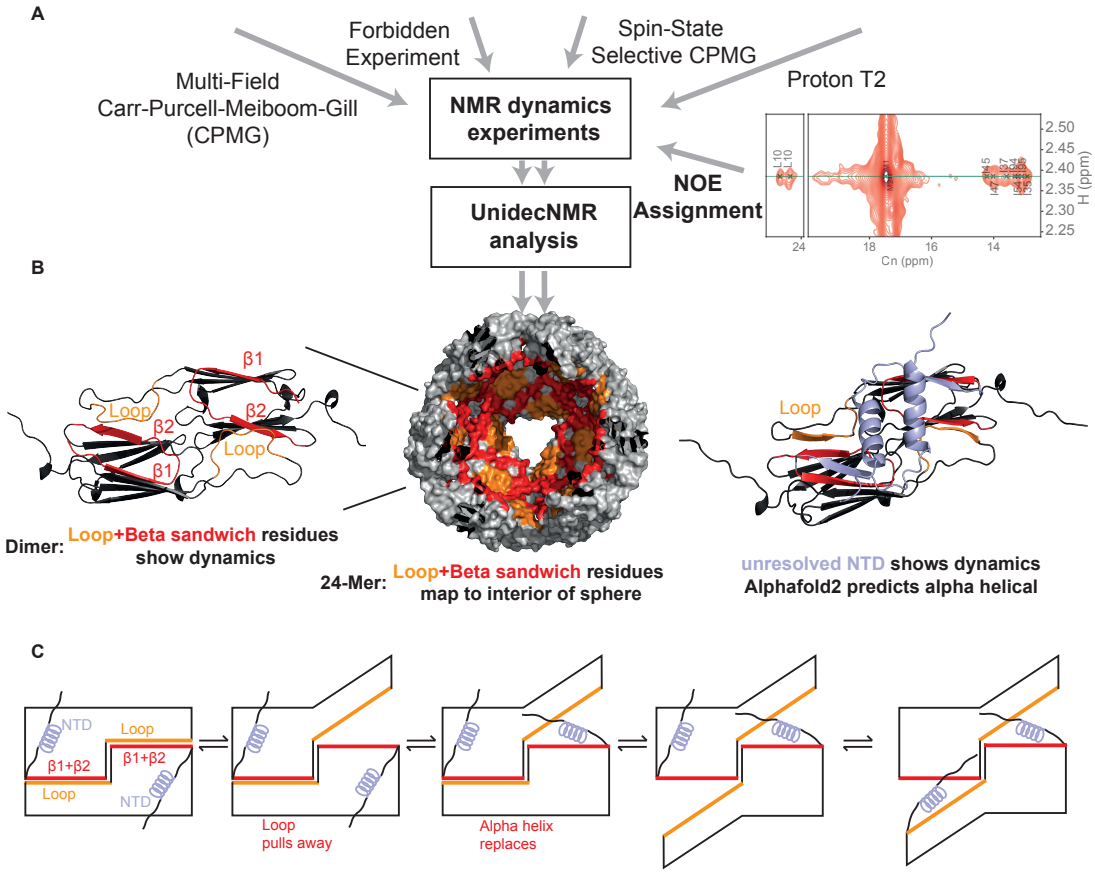


Figure 4.2: The NMR experiments that were used in this work to probe multiple timescales of chemical exchange are shown (A).Structural information from a combination of the X-Ray structure (PDB: 1SHS) and predictions from Alphafold are shown (B). The millisecond dynamics established from experiments in A (particularly CPMG) are mapped on, demonstrating that these residues cluster on the interior surface of the sphere. This may indicate a strucutral re-arrangement on the interior of the sphere. The Alphafold dimer prediction of an alpha-helical NTD and its position close to the residues found on the interior of the sphere are also shown. Finally a model is proposed where the millisecond dynamics on the $\beta1$+$\beta2$ and loop residues may be in concert with the millisecond NTD dynamics also observed in the CPMG experiments (C). This will require more work to prove or disprove.

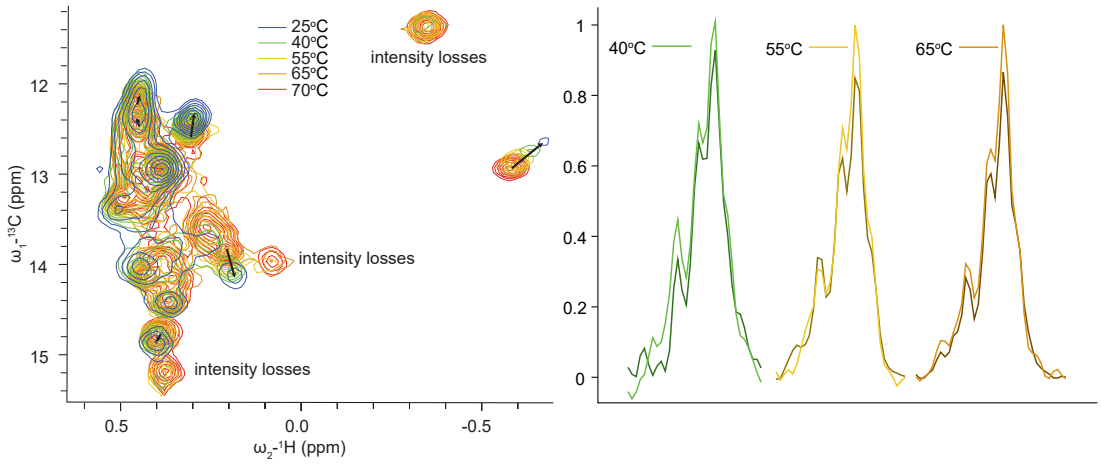


Figure 4.3: HMQC spectra at multiple temperatures shows residue specific degredation of signal at low temperature. Differences in intensity between low and high frequency CPMG spectra are indicative of increasing chemical exchange at higher temperatures and therefore dynamics. All of these spectra were recorded on a 500Mhz Varian spectrometer in Toronto with Room temperature probe.

located on the outside surface of the sphere, does not show evidence of ms dynamics. A significant region of the NTD is predicted to be highly mobile by fast dynamics NMR experiments, alpha-helical by combining alphafold and NOE data, and showing millisecond dynamics by CPMG experiments. To the best of our knowledge, this provides the first broad-coverage atomic resolution study of an oligomeric small heat shock protein assembly, and provides unique insight into the challenging N-Terminal Domain. Therefore, Hsp16.5 is presented as model system for understanding chaperone oligomeric assemblies and their client interactions more broadly.

I will start by demonstrating our assignment strategy, before detailing the analysis of the dynamics experiments and resulting conclusion of interior dynamics. Finally, initial comparisons to a smaller, ACD-only construct are made, indicating similar loop-$B \cdot S$ interactions. Throughout this chapter, analysis tools were embedded into the UnidecNMR platform to facilitate future studies.

4.3 Spectra

4.3.1 Initial temperature ramp

Acquiring preliminary HMQC and Carr-Purcell-Meiboom-Gill (CPMG) spectra at many temperatures is a usual first step when investigating a new methyl-labelled protein. This reveals whether the spectra are reasonable, alongside any evidence of chemical exchange indicative of an 'invisible' excited state [11]. To investigate Hsp16.5, Prof. Baldwin prepared a U-[^2D] $^{13}C^1H_3$ Isoleucine sample when he was a post-doctoral student in Prof. Lewis Kay's lab in Toronto. The resulting HMQC spectra demonstrate well resolved peaks at higher temperatures, with atom specific intensity losses at lower temperatures, while the CPMG experiments show significant relaxation dispersion at higher temperatures, indicative of protein dynamics (Figure 4.3). This is unusual compared to the full-length polydisperse small heat shock proteins, which generally produce poor, subunit-exchange broadened spectra [7]. The well-resolved monodisperse Hsp16.5 therefore provided an opportunity to study the dynamical behaviour of an oligomeric sHSP assembly. As a result, Prof. Baldwin made a Methionine, Isoleucine, Leucine and Valine (MILV) labelled protein to provide better coverage of the full protein sequence. The initial HMQC spectra on this sample gave substantially more peaks, as expected. Pseudo-3D CPMG spectra of this protein indicated significant atom-specific relaxation dispersion. These data confirmed the protein as a highly suitable target for NMR dynamical study.

4.3.2 Assignment of MILV labelled full length protein

Introduction to assignment

The first step of the study is to assign the spectral peaks to particular methyl groups, unlocking the atomic resolution data. Methyl-TROSY labelling is generally not a technique for ab-initio structure calculation, so a crystal structure is required. There are generally two approaches to the assignment problem: brute-force point mutagenesis [14], and NOESY graph matching [32]. Point mutagenesis requires the time-consuming expression of multiple expensive samples. NOESY data from

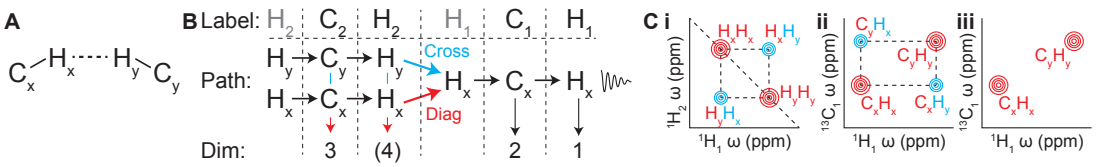


Figure 4.4: The potential pathways of the HCCH NOESY [62] are shown, starting from what a simplified chemical system (A), and showing the magnetisation pathways (B) for both diagonal and cross peaks. The nuclei are labelled x and y with respect to the chemical system, and 1 and 2 according to whether they are acquired before or after the NOESY element during the pulse sequence. The two possible pathways have been laid out looking backwards from the direct dimension. One of these pathways comes from the 'diagonal' (Red), where there has been no transfer of magnetisation according to the NOESY cross-relaxation pathways. The other, cross peak, pathway (blue) is dependent on the strength of the dipolar coupling between spins Hx and Hy. The final row indicates how the acquisition of these chemical shifts (except 1 these are indirectly acquired) adds up to make 3 and 4d spectra. The spectral outcome for this system, assuming reasonable dipolar coupling, is shown (C) in three perspectives, one where we are looking at the first and second recorded hydrogens (Ci) which show both diagonal peaks and cross peaks, as H1 and H2 are on different sides of the NOESY transfer block. Secondly the second recorded hydrogen and the first Carbon correlations are shown, which could either have originated from a cross peak pathway (Blue), or a diagonal pathway (Red). Finally the second hydrogen and second carbon correlations are shown (Ciii), which only shows scalar coupled spins, as there is no NOESY element between them in the pathway (B).

a single sample, however, can be expressed as a distance graph and matched to a separate graph produced from a crystal structure [8]. This is the basis of the MAGMA graph matching algorithm [32]. However, MAGMA must have a NOESY distance graph without false positive connections to produce a confident assignment. Equally, without enough connections between nodes, the potential combinations create unreasonable calculation times. Therefore we needed to create the most accurate distance graph possible based on the NOESY data.

Traditionally, experienced spectroscopists have created this graph by manually sifting through the data to find NOESY cross peaks. Some enhancement has been achieved by using paramagnetic relaxation enhancement (PRE) probes [34]. However, there are millions of data points in a typical 4D spectrum, making manual analysis very laborious. Accessible tools are needed to make the technique more attractive to biochemists needing dynamical information. We used UnidecNMR to construct this.

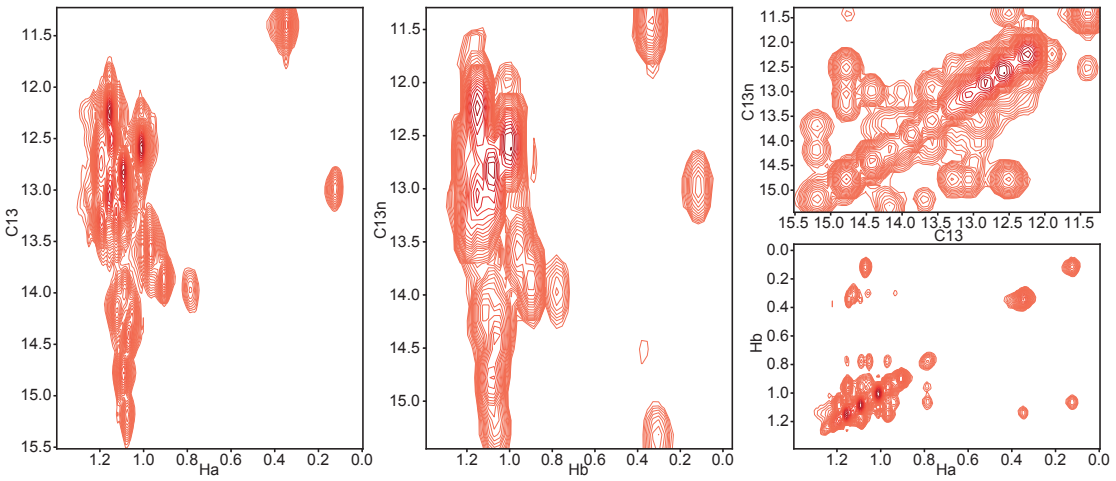


Figure 4.5: The Isoleucine regions of 4D NOE spectral projections are shown. This spectrum was recorded at 60 degrees celsius on a 600Mhz Varian Spectrometer with a room temperature probe.

NOESY spectra

The first consideration was the logical structure of the data, which provided rules to simplify spectral analysis. Experimentally, to minimise relaxation, the NOESY spectra are ^{13}C-^{1}H HMQC-NOESY-HMQC [62]. The logic of the experiment is shown in Figure 4.4. Where spins are dipolar coupled, the NOESY element enables transfer of magnetisation. Recording the chemical shifts of H atoms before and after this element, for example, enables identification of dipolar coupled (and therefore spatially close) H spins. The carbon dimensions before and after the NOESY element can also be recorded to extend the spectra to three or four dimensions. In 3 dimensions, C_1 and H_1 will be recorded (which together represent a normal HMQC) alongide the third, C_2 dimension. When we 'bore' into each peak in the C_1:H_1 HMQC plane, a 1D C_2 strip is produced that shows both 'diagonal' and 'cross' peaks, corresponding to carbons that are spatially close[1]. This is demonstrated in Figure 4.6.

Extending this principle to a 4th dimension, the H_2 chemical shift is also recorded to produce a hypercube. Whilst 4D data is harder to visualise, it becomes easier

[1]While correlating the C_2 and H_1 dimensions, the idea of a 'Diagonal' peak does not exist, as they are distinct elements. However, 'Diagonal' and 'cross' peak nomenclature will be used here to signify peaks that arise from the 'diag' and 'cross' pathways in Figure 4.4 respsectively.

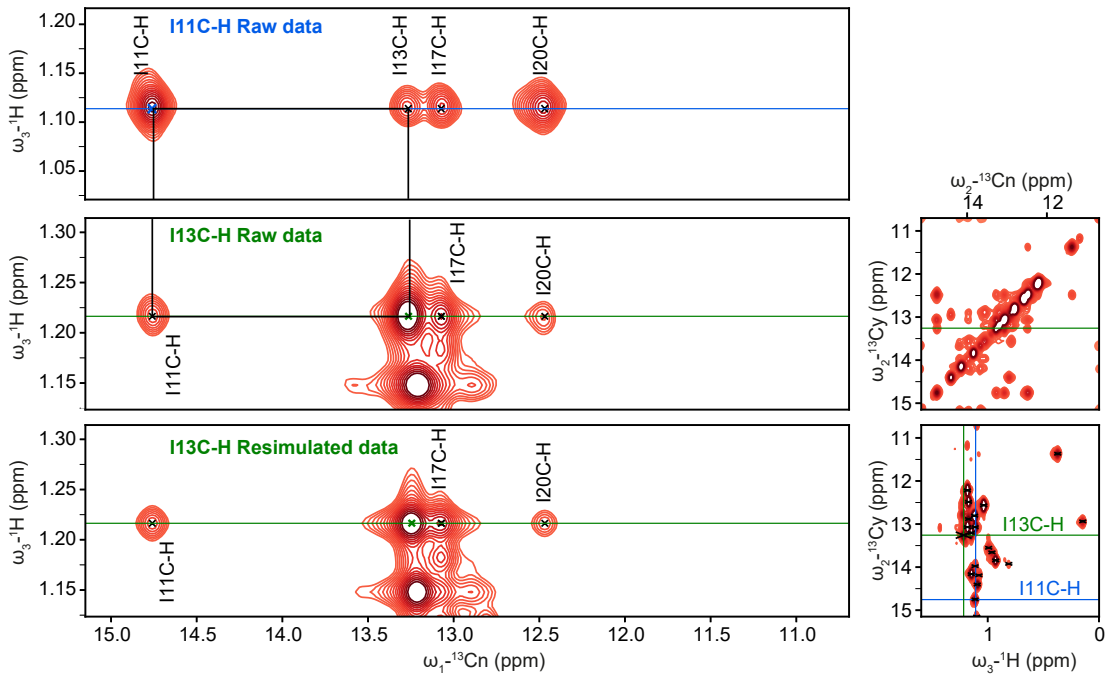


Figure 4.6: An example of data visualisation within the GUI is shown: Raw data is presented for two peaks that share a cross peak, I11C-H and I13C-H, with black lines indicating the location of the cross peaks. In the case where there is a question mark over the cross peak, this visualisation can be useful to confirm alignment of the peak. Resimulated data from UnidecNMR deconvolution outputs are shown for comparison. These show good agreement and make it simple to gain confidence in the UnidecNMR results.

when the dimensions are related to the correlated spins. For example, considering the pathways in 4.4B, peaks only arise from two possibilities, either cross peaks or diagonal peaks. A diagonal peak will have coordinates $[H_1, C_1, H_1, C_1]$ while a cross peak $[H_2, C_2, H_1, C_1]$. There are only a limited set of potential peaks in the spectrum. These can be picked by looking at the $H_1 : C_1$ projection, which will only contain scalar coupled peaks (Figure 4.5). Therefore if we peak pick that projection, we can limit our search in the third and fourth dimensions to those locations. Indeed, taking that projection, we are still able to apply the same logic of 'boreing' that is essential to the triple resonance assignment project (Chapter 2 and Chapter 5). If we 'bore' from the peak locations in the $H_1 : C_1$ projection, we will get a plane of the $H_2 : C_2$ dimensions that contains both diagonal and cross peaks. In this sense we are looking at the potential starting methyls ($H_2 : C_2$ in Figure 4.4)

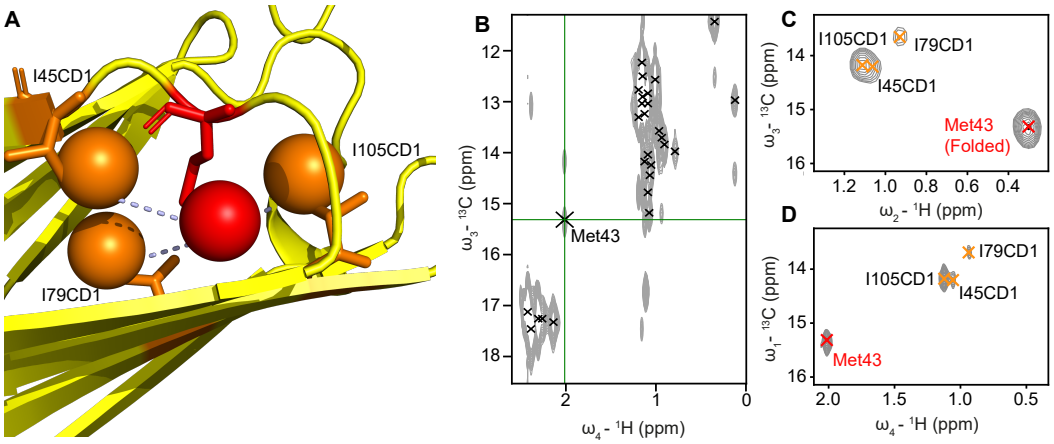


Figure 4.7: A zoom-in of the crystal structure showing the proximity of Met43 to three surrounding methyl groups (A). The peak corresponding to Met43 is shown on one of the 2D projections of the 4D spectrum (B). Two 2D planes of the 4D are shown: $\omega_2 - {}^1H$ vs $\omega_3 - {}^{13}C$ (C) and $\omega_4 - {}^1H$ and $\omega_1 - {}^{13}C$ (D). Cross-peaks are visible to all three of the spatially-close residues. Before assignment is completed, these become peaks that are known to be spatially close, which we can form into a graph for matching with the crystal structure.

of the end methyl peak ($H_1 : C_1$ in Figure 4.4) that we have bored. Equally, we can rotate the hypercube and to look at the $H_2 : C_2$ projection, or starting methyl group, and bore down to get all the finishing metyhl groups. Functionally each 'bore' produces 2 2D spectra that contain a subset of HMQC peaks. In order for it to be a valid cross-peak we would expect it to be visible from both directions, i.e. the starting methyl sees finishing methyl and vice-versa. The knowledge that we should only be able to see a subset greatly helps us: we employ the restraint that we only expect peaks at subset positions: anything outside this cannot be a solution to the deconvolution. This also creates a very clean data structure for NOE data, where the cross peaks are effectively links with intensities between peak indexed CH_3 groups. This can be thought of as a graph, whose nodes are the CH_3 groups and the edges are the NOE links.

4.3.3 UnidecNMR application to 3D NOESY spectra

It was important to build tools that would solve the HSP16.5 problem but also provide a more general solution to the difficult problem of Methyl-TROSY NOESY anal-

ysis.

To simplify the data analysis, we started by taking the 3D MILV spectrum and extracting the Isoleucine region in order to see only the Ile-Ile NOE peaks. There are 21 Isoleucine residues in each monomer of Hsp16.5, 20 of which are visible in the crystal structure. I27 is in the unresolved N-terminal region. Analysis of the 60 degree HMQC spectrum shows 21 clear peaks and some broader underlying peaks. Diffusion data performed previously by Andrew Baldwin (not shown) had found that the clear peaks diffuse in line with an object that would have a similar size to a 24-mer state. It should therefore be possible to assign all 21 peaks to the 24-mer state crystal structure (I27 by a process of elimination). Clear NOE cross peaks are visible in the data and needed accurate identification. Applying previously discussed logic to UnidecNMR enabled us to peak pick the spectra accurately. Following this, the UnidecNMR GUI makes it easy to iterate through the HMQC bores to examine identified cross-peaks. The raw and resimulated data can also be compared (Figure 4.6). Two bores are provided simultaneously and can be linked with symmetry lines to help identify cross-peaks.

The UnidecNMR output is automatically reformulated into a distance graph to be used as an input to the MAGMA [32] graph matching algorithm. This reformulation imposes a requirement for cross-peak symmetry on the UnidecNMR outputs to remove non-symmetric peaks from the distance graph. A striking feature of the Ile spectrum was the 2 ring shifted Isoleucine peaks that appear at lower H shift. Searching the crystal structure [54], there are only two residues that are close enough to be shifted by rings: Ile 99 and Ile 68. Therefore, we can use this as prior information before conducting the graph match; the ring shifted peaks can only be one of those two residues. Once the corresponding crystal structure graph is loaded in automatically by MAGMA the graph matching process begins. We used a mixed approach for this work; MAGMA provided us with some solutions, which we compared manually with the pymol structure to confirm the NOESY pattern. Starting with the ring-shifted methyls, it became relatively easy to work through the Isoleucines. Having assigned the Iles, we moved onto the full MILV spectrum,

using the isoleucines as a scaffold to complete the remainder of the assignments. There are 6 methionines in the protein, of which 4 (Met1, Met14, Met23, Met28) are in the unresolved N-terminus, which was hypothesised to be disordered. Met87 and Met43 are resolved in the ACD of the crystal structure. Met87 is on a loop that forms a significant part of the monomer-dimer interface, while Met43 sits in the bottom loop of the beta sandwich. Using the Ile framework, it is easy to identify Met43, which is surrounded by other residues and has a different chemical shift to the other methionines, which is indicative of structure. The confirmation of the assignment is shown in Figure 4.7. Although Met87 is on a dynamic loop, some weak NOE peaks are visible to nearby Ile residues also on the loop. While the NTD residues are unresolved in the crystal structure, we will later use an Alphafold2 structure to confirm the assignment of Met14.

Due to the isotopic labelling pathway, the methyl groups of valine are randomly labelled either at position CD1 or CD2 and Leucine at position CG1 or CG2. This means each peak is half as intense as the M and I peaks and they are more crowded. We assigned the majority of these with the help of a sample that had both methyls labelled for each L/V residue. A shorter mixing-time NOE experiment was recorded, which only showed peaks for geminal methyls, enabling identification of L/V residue methyl pairs. We were often uncertain about CD1/CD2 assignments, leaving them ambiguous. The lower peak intensity meant noisier spectra and more difficult analysis of the dynamics experiments. They were therefore investigated less during this work. The final assignment is shown (Figure 4.8).

4.4 Dynamics experiments

We used previously devised experiments to probe dynamics of our protein at multiple timescales. R_2 measurements (millisecond) [63], CPMG (millisecond) [31], spin state selective CPMG (millisecond) [9] and an experiment which probes a forbidden transition indcative of pico- to nano-second time-scale dynamics [64].

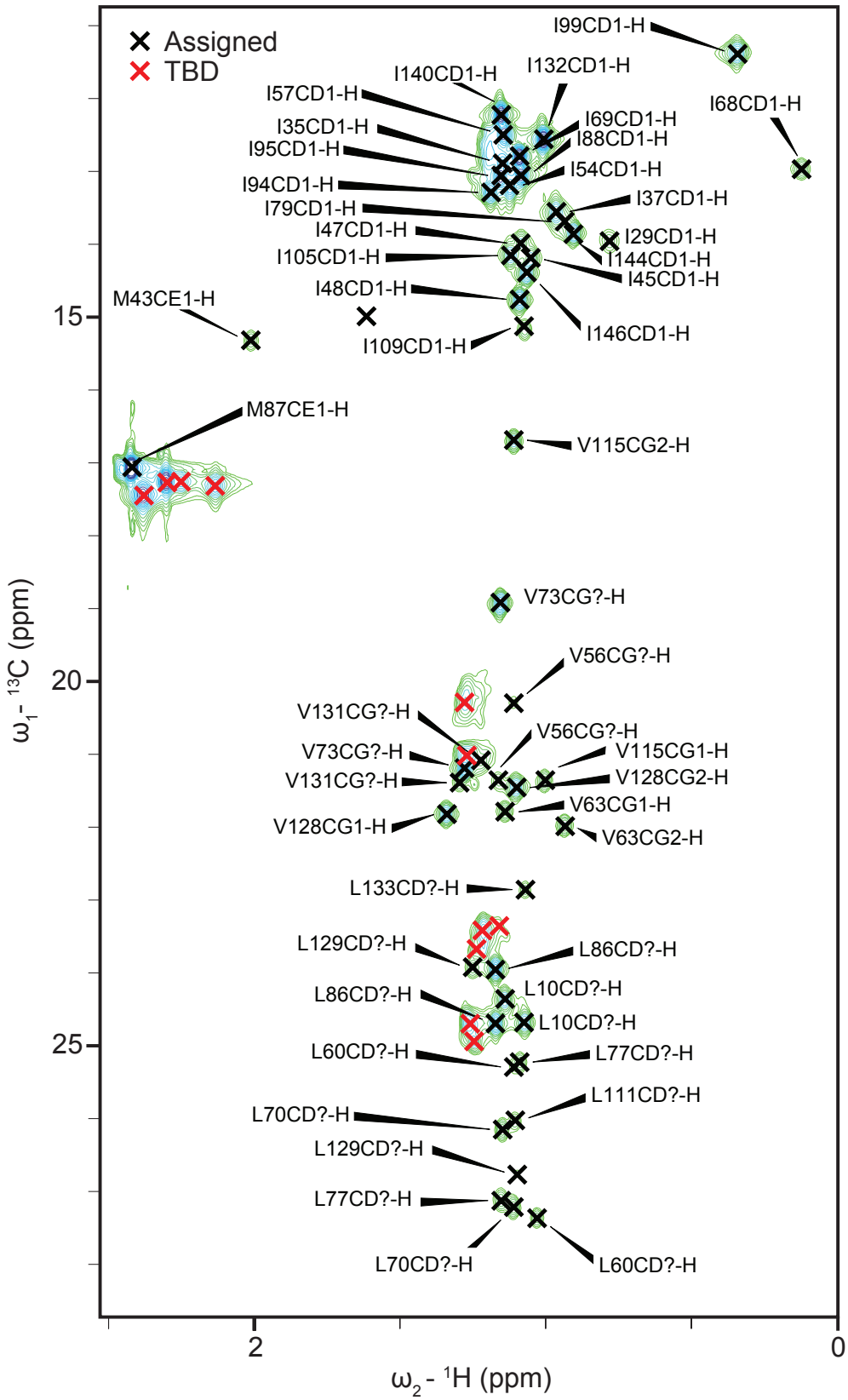


Figure 4.8: The full assignment of the MILV labelled Hsp16.5 is shown. Where there are question marks for the Leucine and Valine residues it was not clear from the Graph Matching procedure whether these belong to the $\delta 1/\delta 2$ or $\gamma 1/\gamma 2$ carbons respectively. This spectrum was recorded on a 700Mhz spectrometer in Toronto with a Room Temperature probe.

4.4.1 The fitting problem

Prof. Baldwin recorded high quality CPMG, T_2 and 'Forbidden' experiment data. These experiments are encoded by intensity on a Pseudo-3D axis, so functionally there are many HMQC planes that require the peak volume to be correctly fitted. This is currently very laborious. Typically FID from NMR experiments decay exponentially, producing spectra with lorentzian lineshapes. However, while the signal decays, noise is even throughout the FID. To increase the signal to noise ratio, the FID is usually multiplied by a 'window function', which weights earlier points (with more signal) higher than later (noisier) points. The resulting peakshape is roughly estimated by the fourier transform of the window function. For example, the fourier transform of a gaussian decay function produces a gaussian peakshape. Applying a gaussian window function to the exponentially decaying FID, however, will produce a mixed gaussian-lorentzian peakshape that can be described by a Psuedo-Voigt function. This function is generally parametrised by three variables: the FWHM of the lorentzian, the sigma of the gaussian and the 'voigt' parameter, which describes the contribution of the the gaussian and lorentzian components. Accurately fitting this function in noisy data is challenging, as there is substantial cross-over between these variables. From the perspective of the fitting problem, a purely gaussian peakshape is very desirable. To obtain this, the standard route is to multiply the data by an exponential function to negate the exponential decay, before applying a gaussian decay function. Finding the correct parameters for this window function requires some manual fiddling, but it generally makes the fitting problem far easier. With this in mind there is a potentially excellent solution as an extension to UnidecNMR, which provides a well-fitted peakshape and peaklist. If it were possible to build a peak fitting algorithm, UnidecNMR would provide excellent starting conditions and enable the data to be fitted accurately. This is currently a work in progress; accurately fitting processed NMR data in a general sense is a very difficult problem. Currently the fitter cannot fit mixed gauss-lorentz functions reliably. The GUI window for this is shown in Figure 4.9. When this is working effectively we are hopeful that it will add full dynamical studies to the capabilities

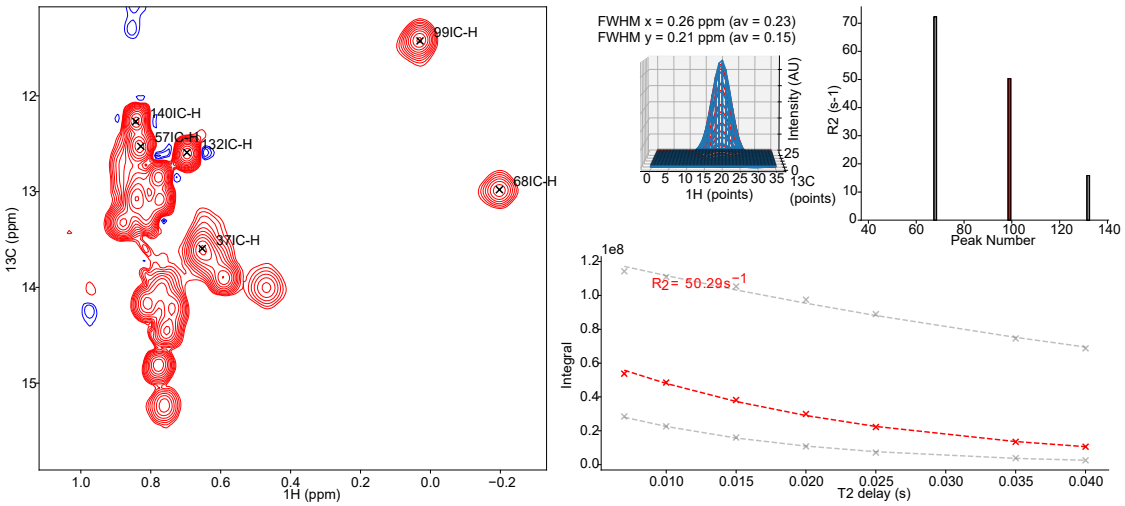


Figure 4.9: The main display for fitting Pseudo 3D data is shown, including the currently selected slice spectrum slice of a T2 experiment including the peaks we are fitting, the 3D intensity plot of the data with fitted in red and raw data in blue. The decay intensity, fits and resulting R2 are shown in the bottom right and the results per residue are shown in the top right.

of the UnidecNMR GUI. Until this is fully functioning, we must use FuDa from Flemming Hansen [65]. This is a very reliable peak fitting algorithm.

4.5 CPMG data

The CPMG experiment is extremely useful for probing chemical exchange of proteins. Chemical exchange occurs when time-dependent changes in spin environment change the NMR spectra. The details of this are very well described elsewhere [66] [67], however it is worth noting the two numbers that govern chemical exchange: k_{ex} and $\Delta\omega$, meaning the exchange rate [2] and difference in chemical shift between the two states, respectively. The underlying physics of the experiment are well described elsewhere [66]. Here, it is enough to know that if there is a dependence of peak integral on CPMG frequency, the experiment is revealing exchange behaviour.

[2] $k_{ex} = k_+ + k_-$ where k_+ is the forward and k_- the backward rate of a two state interconversion reaction

4.5.1 Comparison with alphaB Crystallin

αB Crystallin is a human small heat shock protein that has been previously studied using NMR methyl-TROSY techniques. Unlike HSP16.5, αB crystallin forms polydisperse oligomers that make it very difficult to study. Previous work showed that a model with oligomer-size-independent kinetic parameters explains Mass spectrometry data well, meaning the oligomeric distribution becomes an equilibrium of the on-rate (which is truly independent of oligomer size) and the off-rate (which scales linearly with oligomer size as there are more monomers that can dissociate in larger oligomers). The link is therefore made between monomer-oligomer kinetics and size distribution. This work was furthered with atomic resolution CPMG NMR of the IXI motif in the αB-crystallin C-terminus to give kinetic parameters that correlated well with whole protein mass spectrometry data. Overall, the dynamics of the CTD has been shown to regulate polydispersity [7]. Using the forbidden experiment [64], they also demonstrated that the -IXI- residues were disordered in the ground state, with an $S^2\tau_c$ parameter of around 3.5ns. This produced a model of a disordered C-terminus making hydrophobic contacts with other monomers at a rate that enables polydisperse oligomer distributions.

For Hsp16.5, residues I144 and I146 are the equivalent IXI motif on the C-terminus. From CPMG measurement, there is no relaxation dispersion on these residues. The forbidden experiment also demonstrates that the residues are much more structured, with an $S^2\tau_c$ of 35ns (compared with 3.5ns for αB [7]). It is not possible for this to be following the same model as αB Crystallin as it is not polydisperse and therefore cannot have oligomer-size independent kinetics. Perhaps there is some relationship between the lack of dynamical behaviour observed in the IXI motif and the lack of polydispersity, as this structural rigidity may have a role to play in stabilising mono- vs poly- dispersity. Further work would be required to establish this link, perhaps performing mutagenesis to introduce polydispersity [49] before probing millisecond dynamics of the IXI motif with CPMG.

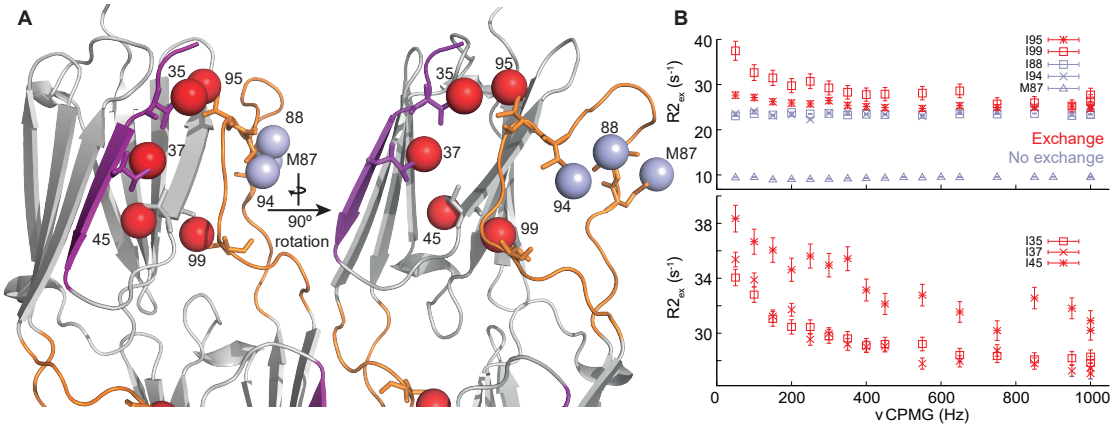


Figure 4.10: Excerpts from the crystal structure demonstrate the location of residues that are (red) and are not (lilac) showing CPMG relaxation dispersion. This demonstrates that residues at the interface between the Beta sandwich (35, 37 and 45) and flexible loop (95 and 99) show evidence of a significant exchange contribution (B), while methyls not at the interface but still on the loop (87,88 and 94) demonstrate no exchange. As kinetic parameters are not residue specific, we can infer that the larger dispersions on the red residues indicates their larger Delta Omega (or chemical shift change) values, indicative of more involvement in any structural rearrangement between exchanging species.

4.5.2 Domain switching dynamics

Initial results on other parts of the protein indicated residue specific dynamical behaviour. This is shown in Figure 4.10, which isolates the dispersions to one interface of the protein: the loop and N-terminal side of the beta sandwich. The residues marked in red on this figure are showing dispersions, while those that aren't are lilac. Residues 95 and 99 from the loop are engaging the B·S and show dispersions, while 87, 88 and 94 are pointing in the opposite direction to the B·S and show no dispersions. The chemical environments (and therefore $\Delta\omega$) of the red residues experience the largest change while the lilac experience the smallest change. A model consistent with this finds the loop residues marked in red (95 and 99) embedded within the hydrophobic B·S in the ground state and solvent exposed in the excited state. Conversely, the lilac residues are likely solvent exposed in both ground and excited state, with a much smaller $\Delta\omega$.

In order to quantify this, the experiment must be performed at multiple fields, as the $\Delta\omega$ will increase with field strength[3]. The resulting multi-field data can be

[3]While the ppms of excited and ground state will stay the same, the frequency difference in Hz

modelled to quantify residue specific $\Delta\omega$ and global k_{ex} and population parameters. Prof. Baldwin had repeated this experiment at 600Mhz, 750Mhz and 950Mhz, all at 60^oC. The spectra were processed using nmrPipe [10] and fitted using the FuDa algorithm. The resulting relaxation dispersion curves were fitted using CATIA [68] (another software package developed by Flemming Hansen) that fits the Bloch-McConnell equations to experimental data. Functionally, this is very similar to the Bloch-McConnell [67] equation fitter in appendix B, but rather than a saturation pulse before the pulse sequence, there is a transverse CPMG block during the pulse sequence.

The data and their corresponding fits are shown in Figure 4.11. Including the exchanging methionines from Figure 4.12A both worsens the fit of the isoleucines and does not explain the methionine data. This indicated the model might need to be expanded to a 3-state system. To confirm this, we fitted the N-terminal methionines in isolation, which again did not explain the data well.

4.5.3 N-terminal domain

The N-terminus of HSP16.5 is not reported in the crystal structure. More generally, structural studies of small heat shock proteins have centred around dimer ACD constructs that lack both a CTD and NTD. This simplifies the system by preventing oligomerisation, and the CTD can be expressed as a separate peptide, enabling CTD-ACD binding studies. Information about the N-terminal domain, however, has remained elusive. It has been shown to be important for chaperone activity [69] [70]. There has been some more recent Cryo-EM work on a R107G point mutant of Hsp16.5 that shows differences from the wild-type [53]. The wild-type shows density in the middle of the interior cavity while the mutant shows density at the edges. Work was then done to quantify interactions with an aggregating, substrate protein using a Thioflavin-T assay. This showed that the point mutant, with the different N-terminal conformation interacted more with the substrate than

will grow linearly with field strength

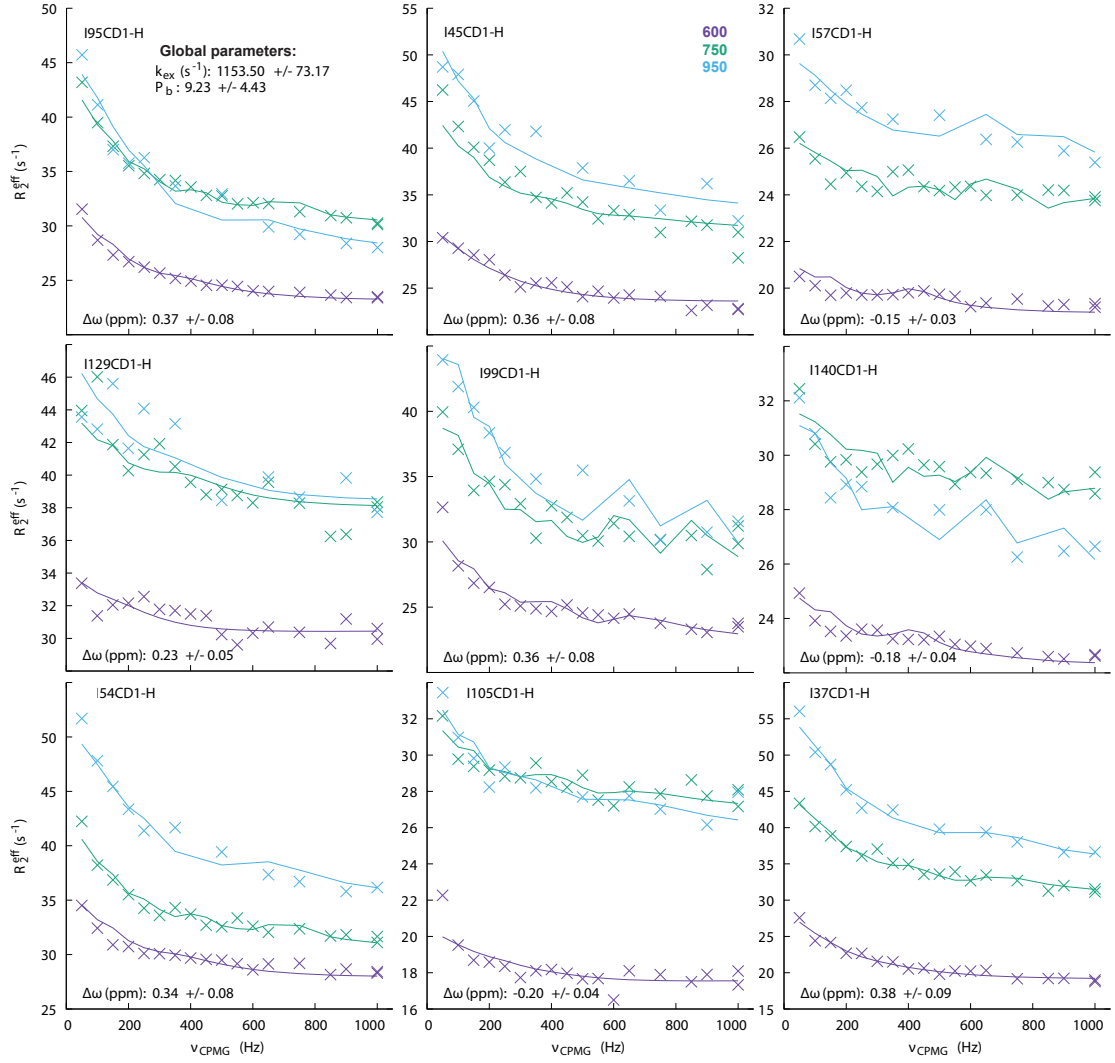


Figure 4.11: CPMG data of Isoleucine residues showing their respective fits are shown. Globally fitted p_b and k_{ex} (population of the excited state and rate of exchange respectively) are shown while residue specific delta omegas are also shown. The z flip rates (CzHz-CzHzHz rates) were measured and used as restraints, together with off resonance effects. In the absence of these, the fitted curves were smooth. The jagged appearance in the simulations would be removed if during fitting, a range of B1 fields were used in the analysis. These data were all collected at 60 Degrees Celsius on Varian 600Mhz, General Electric 600, 750 and 950 spectrometers all with Room temperature probes.

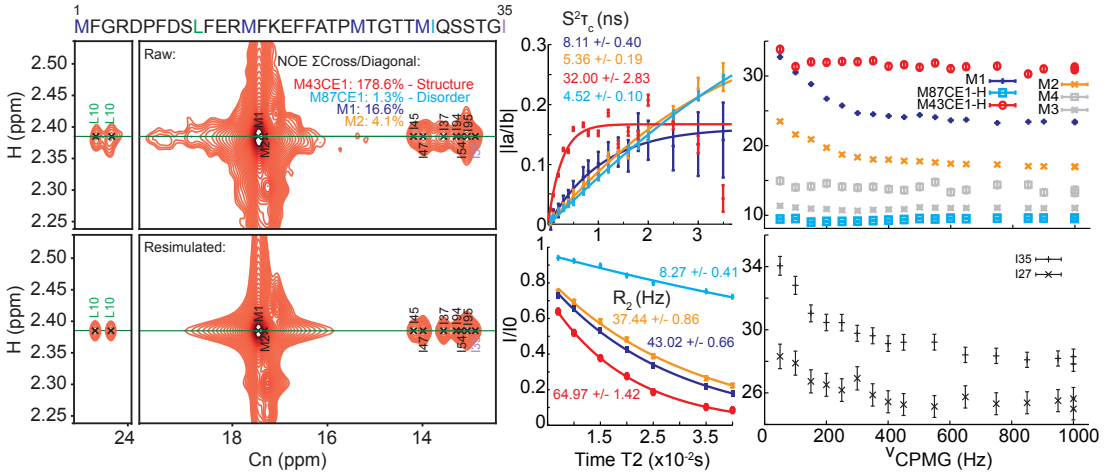


Figure 4.12: The sequence of the N-terminus is shown alongside raw and UnidecNMR resimulated NOE data for one of the exchanging N-terminal methionines indicating some long-range interactions. The sum of the cross-peak intensity divided by the intensity of the diagonal is also shown. M43CE1 is shown in the crystal structure to be in a structured region, while M87CE1 is on a flexible loop and appears to be disordered. Neither show signs of dispersion, so they are used throughout the Figure as comparative benchmarks. The results of the forbidden and T2 relaxation experiments are shown with fits and corresponding numbers. CPMG plots are also shown at one field, indicating dispersion in two of the N-terminal methionines (M1 and M2) while two do not have dispersions. The two Isoleucine residues from the N-term are also shown to have dispersions. The highest quality data was taken to demonstrate this: in this case the Isoleucine data came from a cryoprobe equipped 800Mhz spectrometer at 70^{o}C, while the Methionines were a 600Mhz spectrometer at 60^{o}C.

the wild type. The conclusion was that conformational changes in the N-terminus were important for substrate binding.

Environment of the NTD

The crystal structure shows the diameter of the interior of the sphere to be 65Å. From this we can calculate the concentration of the protein and water in the interior, enabling us to make some simple conclusions about the environment within the sphere.

Firstly we can assume a sphere to calculate the volume:

$$V(m^{-3}) = \frac{4}{3}\pi r^3 = 1.43 \times 10^{-25} m^3 \tag{4.1}$$

Then we can use the density of water to Figure out how many atoms of water there should be in the sphere.

$$N_A(mol^{-1}) \cdot \rho(g \cdot m^{-3})/M_W(g \cdot mol^{-1}) \cdot V(m^{-3}) = n \tag{4.2}$$

$$6.022 \times 10^{23} \cdot 997000/18.02 \cdot 1.43 \times 10^{-25} m^3 = 4773 \tag{4.3}$$

Therefore in the volume of the interior there would be 4773 water molecules assuming it has the same properties as bulk water. We know that there are 24 copies of the N-terminus in the interior, which means the concentration of the N-terminal peptide in the interior is:

$$N(mol) = 24/6.022 \times 10^{23} = 3.985 \times 10^{-23} \tag{4.4}$$

$$N(mol)/V(dm^3) = 3.985 \times 10^{-23}/1.43 \times 10^{-22} = 0.278(M \cdot dm^3) \tag{4.5}$$

If we now consider the mass balance of water vs. N-term (mass of residues 1-34 = 3912.48):

$$24 \times 3912.48/N_A = 1.55934 \times 10^{-19}(g) \tag{4.6}$$

Assuming the protein has a density of $1.35 \times 10^6 gm^{-3}$, the volume it occupies is $1.155 \times 10^{-25}(m^3)$, which leaves $2.75^{-26}(m^3)$ for water. Using the formula above we can calculate the number of molecules of water assuming bulk density:

$$6.022 \times 10^{23}(n \cdot mol^{-1}) * 997000(g \cdot m^{-3})/18.02(g \cdot mol^{-1}) * 2.75 \times 10^{-26} m^3 = 916 \tag{4.7}$$

The mass of which is:

$$916 * 18.02/N_A = 2.74 \times 10^{-20}(g) \tag{4.8}$$

These calculations estimate the mass inside the sphere at 85% protein and 15% solvent. This leaves little room for any (particularly large or aggregating)

proteinaceous substrates inside the sphere. While the chaperone may still act on unfolded chains as the full 24-mer, this may point to Hsp16.5 chaperoning proteins using the exterior of the sphere or disassembling into smaller components that themselves chaperone. In terms of concentrations, this highly proteinaceous interior is not dissimilar to the phenomenon of phase separation. Performing the same calculation on phase separated DDX4N, which is 400 mg/ml in the condensed phase [71], the mass percentage protein comes out at around 53%. Therefore, the interior of the sphere is more exclusionary of solvent than phase separated protein. We can repeat this calculation on the structured part of the protein in isolation (excluding the 65 Åinterior diameter and but including the outer shell of structured residues up to 120 Å) to give around 50% protein using the same assumptions. We cannot rule out the N-terminus extruding from the sphere or a solution state expansion of the sphere away from the cyrstallographic structure.

Two of the methionine residues have been confidently assigned from the crystal structure: residue 87 is on the dimer loop region which gives an $S^2\tau_c$ of 4.5+/-0.1ns, indicating significant mobility and disorder. Residue 43 is also offset in ^1H and ^{13}C chemical shift in the middle of the ACD Immunoglobulin (Ig) domain and gives an $S^2\tau_c$ of 32.0+/-2.8ns, which indicates much more restricted motion and therefore structure. These results agree with the crystal structure. Neither residue shows any dispersion in the CPMG experiment. We decided to use these residues as a benchmark for structure/disorder parameters for the N-terminus. The 4 N-terminal residues cluster in chemical shift close to the disordered Met87. $S^2\tau_c$ values could be extracted for 3 of these, giving results of 4.1+/-0.1ns, 5.4+/0.2ns and 8.1+/-0.4ns. These are indicative of a locally mobile ground state. In terms of chemical exchange, defining the 'state' of any protein can be misleading: the high local mobility measurements of the NTD residues means the ground state must fluctuate on the 'fast' (ns/ps) timescale, likely interchanging between many chemically similar states. High local mobility is generally considered a measure of 'disorder', which in bulk solution usually implies solvent exposure. In this case, however, the high protein concentration indicates the ground state is likely to be

dominated by transient protein-protein interactions, again similar to liquid-liquid phase separation. This could be proven with isotopic dilution studies looking for inter vs intramolecular NOEs.

In the disordered case $k_{ex} \gg \Delta\omega$ and therefore the spins are in the fast exchange regime with a population weighted average. As k_{ex} decreases, the exchange regime moves towards intermediate. In this we can consider the Arrhenius equation, which relates the rates of reaction with the energy of the transition state:

$$k = Ae^{\frac{-E_a}{RT}} \tag{4.9}$$

Where k is the rate of reaction and E_a is the activation energy, or energy level difference to the transition state. At equilibrium for an isomerisation, $k_{ex} = k_+ + k_-$, meaning k_{ex} is related to the forwards and backwards activation energies. Equally, at equilibrium:

$$K_D = \frac{k_+}{k_-} = \frac{[B]}{[A]} \tag{4.10}$$

In order for the CPMG experiment to show dispersions, generally there must be around >0.5% excited state, meaning the K_D must be 0.005 or larger and the k_{ex} is on the millisecond timescale [66]. From these simple equations we can the following: the ground $\rightarrow$ excited transition has to be 1-2 orders of magnitude slower than the excited->ground transition. Assuming that the transition state for the forwards and backwards reactions is the same, we can infer that the excited state is thermodynamically less stable. Therefore, the fact we are seeing well resolved peaks and dispersions puts significant restraints on how the ground and excited states can be related.

It is notable that some NOE cross-peaks are observable from the N-terminal methionine peaks. In order for the cross-relaxation pathway enabled by dipolar coupling to occur, the spins must be stationary with respect to each other for a duration. In disordered proteins and peptides, NOE transfer is very weak/non-existent beyond neighbouring residues because the timescale of motion is faster than the required stationary time. As the local environment becomes more

structured/ordered, correlated motion becomes more probable, as does an NOE cross-peak. This may suggest that we are observing a partially structured and locally dynamic ground state. A model that is consistent with this would be some secondary structure (i.e. alpha-helix or beta sheet) without tertiary structure.

There are significant NOE cross peaks between both methyls of the N-terminal Leu10 (assigned through elimination) and one of the NTD methionines. To understand the NOEs we turned to protein structure predictions from Alphafold2 [72], which has emerged in recent years with much promise of accurate protein structure prediction. Looking in the Alphafold database of structures [73], the monomer HSP16.5 ACD and CTD match the crystal structure almost exactly. Intriguingly, the NTD is predicted to be alpha-helical. To expand this, we created and ran our own monomeric Alphafold calculation using colabfold [74]. This gave very similar results to the database for the monomer. However, the monomer is not an observed species, so we built up the oligomeric complexity. The dimer again shows an alpha-helical structure for residues 6-20 of the NTDs, which group together at the bottom of the Ig domain (Figure 4.13). Alphafold gives 5 predictions with associated residue-specific confidence scores (Figure 4.13). All 5 predicted structures suggest an alpha-helical structure for the NTD. Overlaying these shows only small variations in the prediction of the alpha-helical position. However, this is not a stable construct, so we moved to larger oligomeric species. We rented time using the vast.ai platform [75] and developed scripts to install and run colabfold on their superior GPU systems. Unfortunately, Alphafold predicts dissociated dimers for the 24-mer construct which is not correct. This is an ongoing part of the project and in the future, I would like to rule out incorrect setup. Conversely, it is reasonable to suggest that Alphafold does not have a large dataset of this sort of locally mobile, partially structured and exchanging domains on which to base its prediction.

The Alphafold dimer predicts an alpha-helix between NTD residues 6 and 20, which positions Leu10 and Met14 very close to each other (Figure 4.14 A), meaning the methionine with a strong NOE to Leu10 can be assigned as Met14 (Figure 4.14 B+C). Further, the ^{1}H chemical shift of this peak is moved signficantly

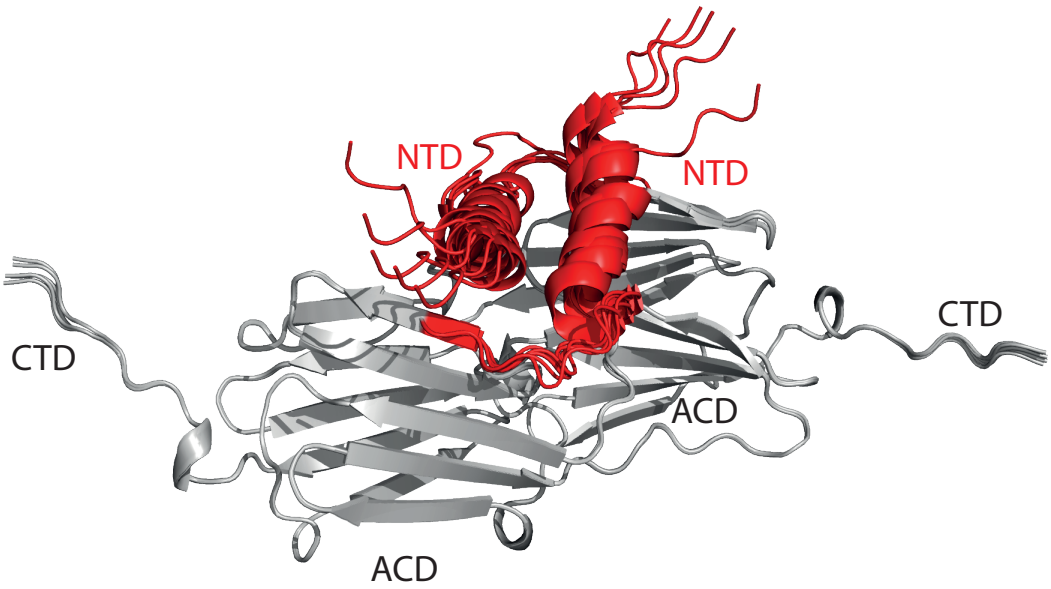


Figure 4.13: The resulting dimers of HSP16.5 are shown with the N-terminal domain highlighted in red. The predicted alpha helicity is not present in the crystal structure.

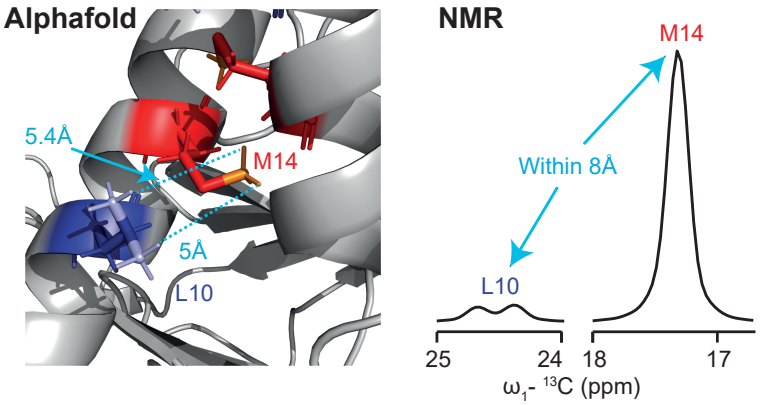


Figure 4.14: The Alphafold prediction of the NTD is an alpha-helix which positions L10 and M14 close (A), and certainly within the possibility of an NOE (5Å). The chemical shift of residue M14 is offset in hydrogen, similar to the structured M43 (B), and the 3D NOE map shows significant intensity for between Leu10 and M14.

away from the others, indicative of structure. The combination of these results improve the evidence base for alpha helicity of this region. To confirm this, we are planning to do some Circular Dichroism measurements to look for evidence of alpha-helicity, as we anticipate this to be the only alpha-helix in the structure. This has been speculated in previous work [53].

One other NTD Met peak has some weak but symmetric NOE peaks to the

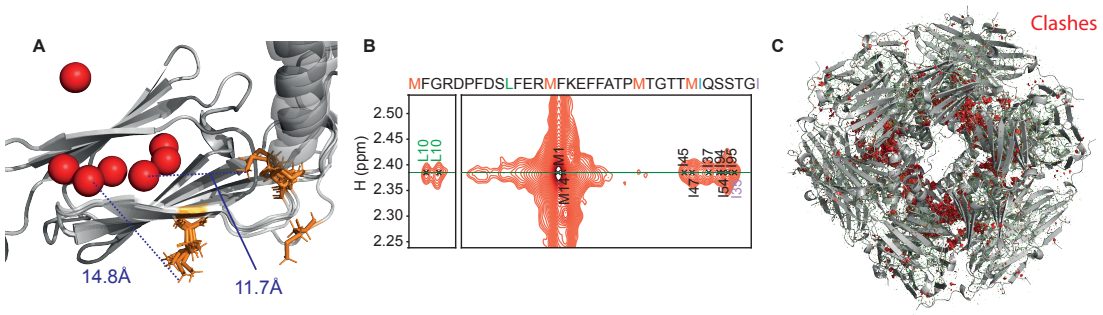


Figure 4.15: NOEs between one of the methionines and the ACD (B) cannot be explained by either methionine in the Alphafold model (A). This is rationalised when the Alphafold models are aligned with the 1SHS crystal structure via the very similar ACDs (C). This shows substantial clashing between NTD alpha helices (red)

ACD. When mapped onto the Alphafold predicted dimer they cluster to a particular region (the B·S loop interface), but none of the Alphafold methionines are close enough (within 10Å[4] Figure 4.15). Using the similarity of the ACD, we were able to map the NTD alpha helices onto the full 24-mer structure, but this showed signifiant clashes between NTDs (Fig 4.15C). We know that the dimer construct that we have folded is not the stable form in vitro, and so the utility of this model is limited. While there is evidence of alpha-helical structure in the NOE measurements that correlates with the predicted alpha-helix from Alphafold, the orientation of the predicted dimer alpha-helix cannot be extrapolated to the full 24-mer NOE data.

4.5.4 Spin-state selective CPMG

Attempting to fit the same single quantum coherence data did not produce very good fits by eye (Figure 4.16), with significant deviations between the data and the simulation. Therefore we turned to the more elaborate spin-state selective CPMG experiment to probe the N-terminus further[9]. Unlike the standard CPMG experiment, the single quantum coherence coupling evolution is not refocused during indirect acquisition, resulting in a quartet in the indirect dimension. Linear combinations of phase-cycled spectra 'select' for each multiplet component to produce 4 pseudo-3D spectra containing one spin-state each. These spin-states have their own relaxation properties, and can each be fitted globally to produce

[4]No NOE peaks are found in the structure at a distance greater than 10Å

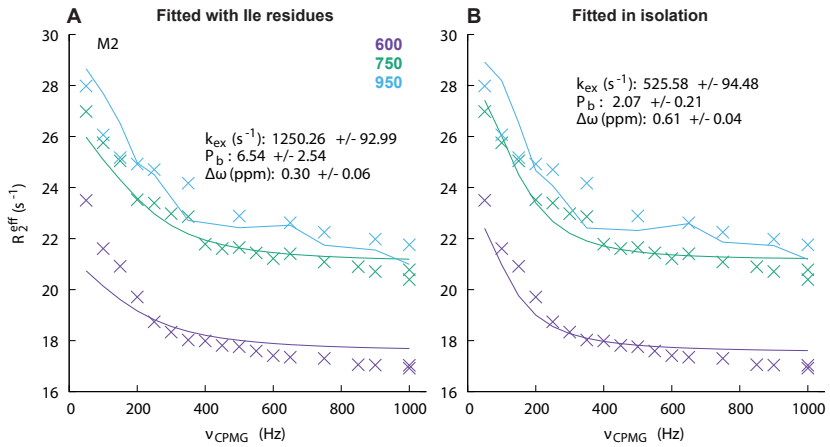


Figure 4.16: 2 attempts to fit the methionine data are shown. Neither a global fit including Isoleucine residues from Figure 4.11 could explain the data well nor a methionine only fit. This was assessed by eye, with the 600 fit showing significant deviations from the data. The z flip rates (CzHz-CzHzHz rates) were measured and used as restraints, together with off resonance effects. In the absence of these, the fitted curves were smooth. The jagged appearance in the simulations would be removed if, during fitting, a range of B1 fields were used in the analysis.

individual CPMG dispersion profiles. Each multiplet component has a different relaxation rate dependent on the combination of the TROSY effect and global tumbling time of the molecule. These data can be fit to extract the difference between the ground and excited state R_2 (transverse relaxation) and $S^2\tau_c$ (ps/ns motional parameter) values. Unfortunately this experiment requires extremely high signal-to-noise and clean spectra, as the inner lines of the multiplet contain 1/3rd the intensity of the outer lines. Only one of the NTD methionine residues provided good fits. Fitting the multiple field single quantum data with the spin-selective data produced fits (Figure 4.17) that indicated an disorder to order transition through positive ΔR_2 and $\Delta S^2\tau_c$ parameters. The ground state $S^2\tau_c$ parameter is also fitted and is in good agreement with the result from the separate forbidden experiment.

4.5.5 α-lactalbumin

To probe spectral differences on addition of a client protein, unlabelled α-lactalbumin was added to the mixture under aggregating conditions [76]. There were some significant peak appearances, but it has not yet been possible to assign or understand

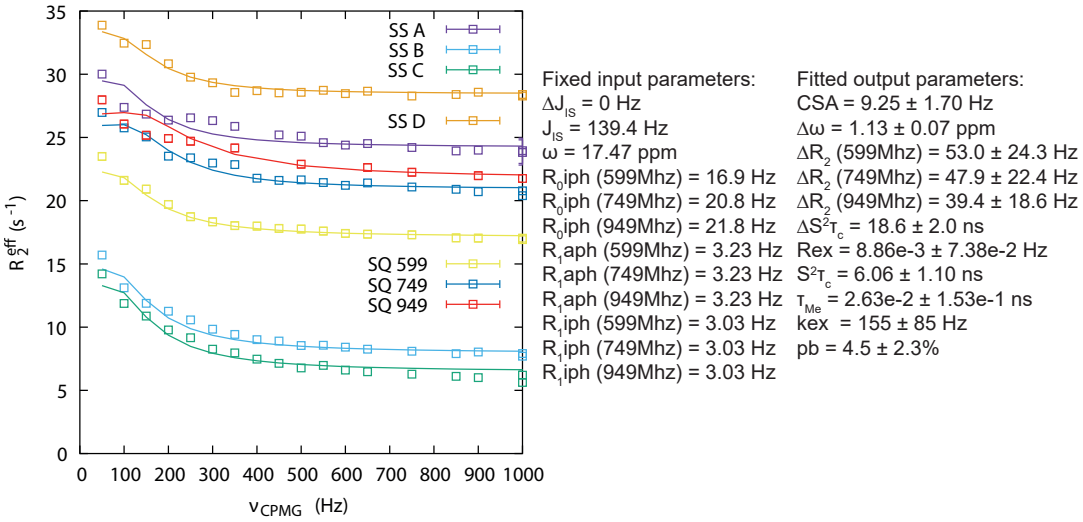


Figure 4.17: The fits for both the spin state selective and single quantum coherence dispersion data at multiple fields are shown for one of the methionines in the N-terminus. This was the only residue with high enough quality data to perform this fitting. The extracted parameters are all reasonable and indicate a strong positive ΔR_2 within error. The delta $S^2\tau_c$ is also positive. All experiments were measured at 60 degrees celsius. The ground state $S^2\tau_c$ agrees well with the recorded forbidden experiment data (5.36 ± 0.19ns). R_1 inphase and antiphase numbers were measured experimentally at 600Mhz and assumed to be field independent.

these changes (Figure 4.18). This may be indicative of substantial structural changes related to chaperone activity. The α-Lac was unlabelled, and at these concentrations we would not expect natural abundance spectra to produce peaks with this intensity (circa 10% of the maximum intensity and broader than the sharpest peaks).

4.6 Alpha-crystallin domain

Removing the NTD and CTD is an often used trick to prevent polydispersity in small heat shock proteins [39]. It produces an α-crystallin only dimer construct, although a subpopulation of monomers has been shown for some monomers [39] [77]. This construct is much smaller and easier to study using NMR. Other members of the Baldwin Group produced a U[^{15}N −^{13}C] ACD sample that could be assigned using triple resonance experiments [27]. Light scattering experiments demonstrated that this construct was an excellent chaperone (Figure 4.19). There are 113 residues, so subtracting 6 Prolines and the N-terminus gives 106 potential peaks. At 25

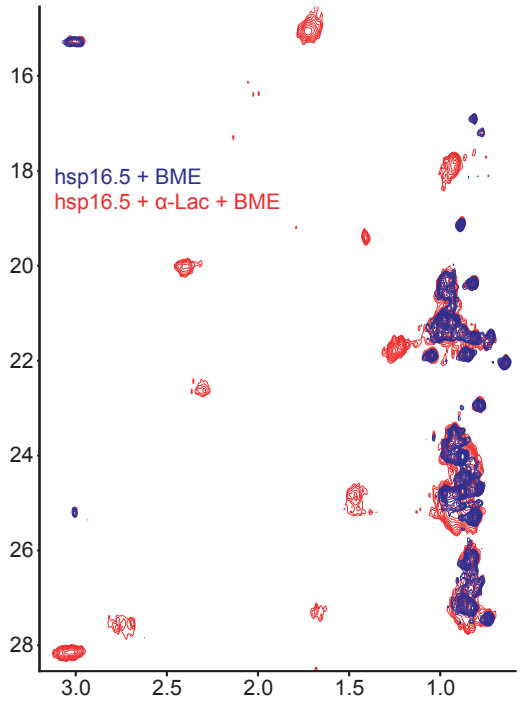


Figure 4.18: α-Lac was added to Hsp16.5 under aggregating conditions (the addition of beta mercaptoethanol (BME)). There were some substantial peak appearances and general peak broadening of the Hsp16.5 peaks, but it has not yet been possible to understand these changes.

degrees the spectral quality was significantly worse than expected for the small protein size (24.7kDa as a dimer) (Figure 4.20). An NHSQC temperature ramp was recorded, showing substantially better spectra at 50 degrees. Therefore, assignment spectra were recorded at this temperature. There was some limited, unquantifiable evidence of relaxation dispersion at 25 but not 50 degrees.

4.6.1 Comparison with full length protein

The ACD construct is particularly useful if it can inform on interactions of the full length protein. To experimentally determine the similarity, the similar HMQC experiments were recorded (Figure 4.21). The TROSY spectrum is not decoupled during acquisition, but the multiplet only has one sharp component, meaning the peak position is shifted by -0.5J. J varies slightly between individual methyl groups so the overlay will never be perfect. Homonuclear ^{13}C coupling can also evolve in the uniformly labelled construct, producing doublets in the indirect dimension. The

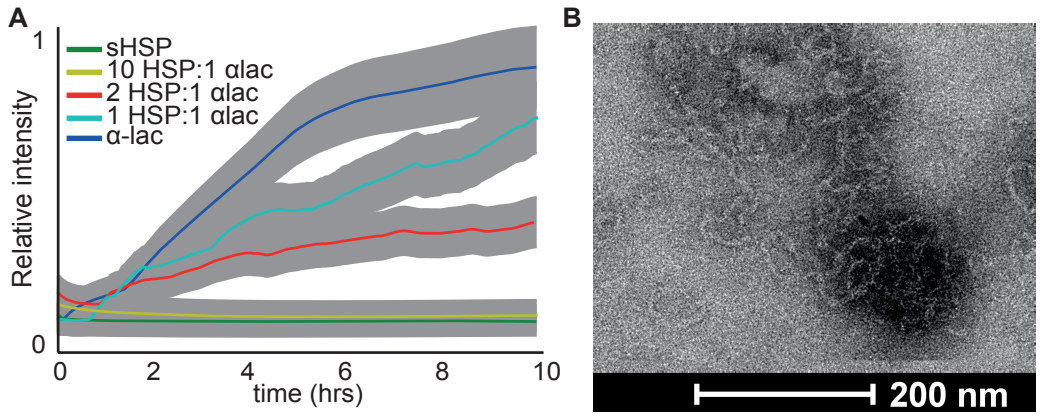


Figure 4.19: This figure is taken from Michael Barber's thesis [78], who performed the experiments, to demonstrate the chaperone ability of the core construct of HSP16.5. The light scattering time course of 200 μM α-lac with various concentrations of HSP16.5 is shown (A). EM micrographs showed no clear differences in the samples at the end time point (B). More information on the sample details are available in the original document [78]

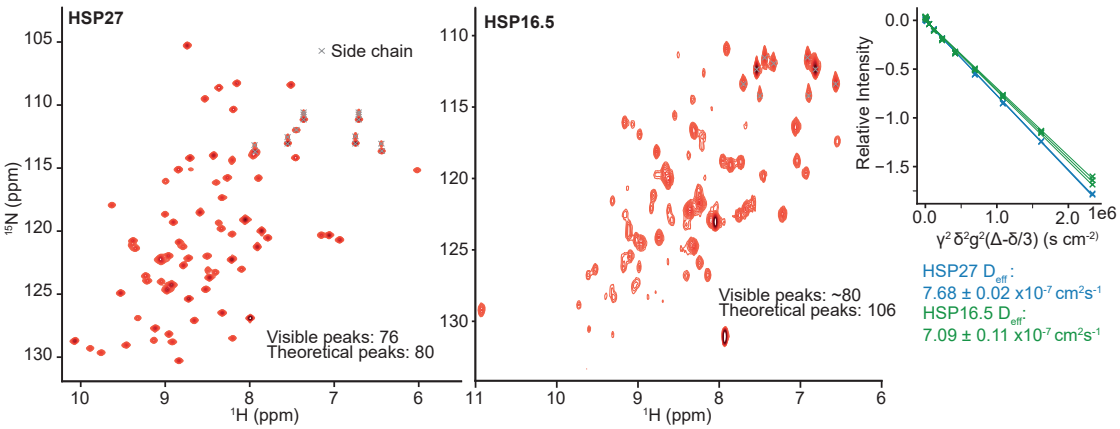


Figure 4.20: NHSQC spectra are shown for both hsp27 [25] and hsp16.5 at 25 degrees, showing substantial differences in spectral quality. HSP27 has much sharper line shapes, while HSP16.5 is much lower signal-to-noise and resolution, indicative of higher relaxation rate or chemical exchange. Diffusion measurements show very similar diffusion coefficient for the two spectra, consistent with dimers. This suggests the major state is a dimer, while the low quality NHSQC is indicative of chemical exchange with another species. The C-terminal peaks are very sharp and visible for HSP16.5.

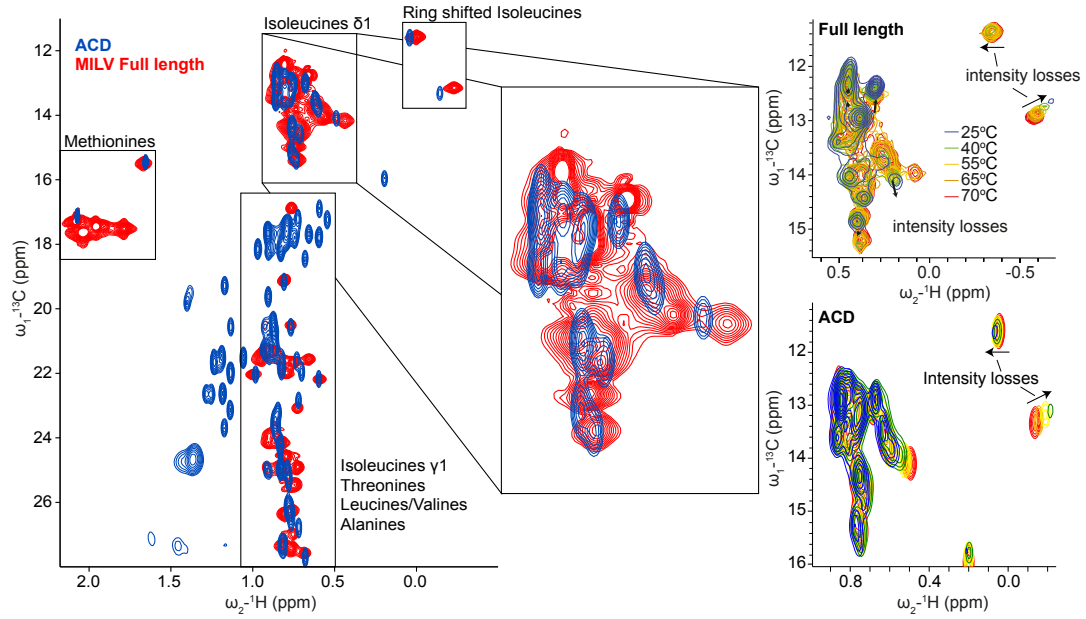


Figure 4.21: HMQC spectra are overlaid for the ACD and full length constructs, showing great agreement. Temperature ramps are also shown with similar behaviour shown in the isolated ring-shifted peaks.

best defined peaks are the methionines, so these were chosen for the -0.5J offset. It is also important to note the sequence differences between the two constructs: the removed N-terminus contains 1 Isoleucine, 4 methionines and 1 Leucine. The removed C-terminus contains two Isoleucines. Therefore, we expect some missing peaks. Overlaying the full length spectrum on the ACD shows striking similarity with both ring-shifted methyls visible, 2 of the 6 methionines present as expected, a similar pattern in the central region of Isoleucines and generally good overlap in the valine/leucine region. Varying the temperature showed similar intensity losses and chemical shift perturbations for the ring shifted methyls, indicative of similar structural response with temperature. The CCCTOCSY and HCCCNH experiments were used to assign the ACD methyls. They pass magnetisation along the carbon side-chains using a carbon-carbon TOCSY sequence, recording the side chain carbons and hydrogens respectively along with the N-H i-1 chemical shift information. If an N-H assignment is present, it is then possible to assign the methyls.

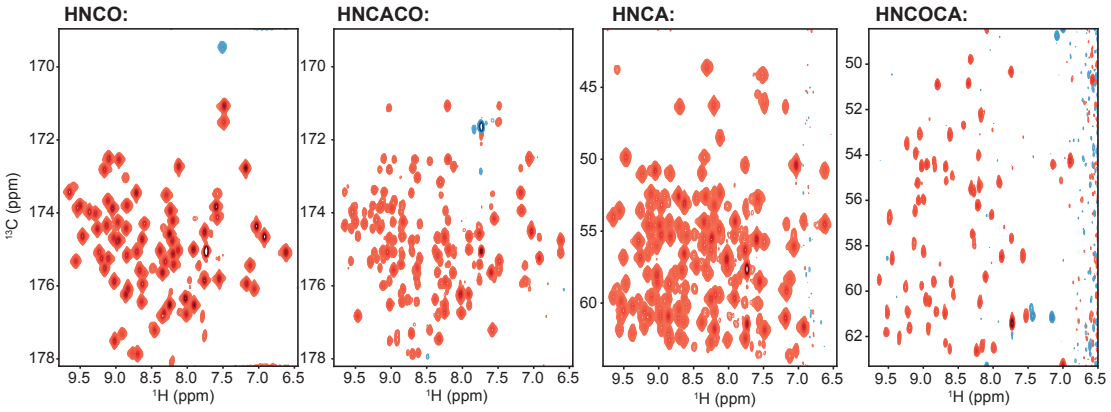


Figure 4.22: HNCO, HNCACO, HNCA, HNCOCA spectra are shown at 50 degrees C performed on a Varian 600 MHz spectrometer with room temperature probe.

4.6.2 Assignment of ACD

A full suite of assignment spectra were recorded at 50 degrees, including HNCO, HNCA, HNCOCA, HNCACO, HNCOCACB, HNCACB and CCCTOCSY experiments. Some peaks were missing at this temperature, but the vast majority were substantially sharper than at lower temperatures. Using the algorithm and environment detailed in Chapter 5, we correlated the assignment spectra (Figure 4.22). This produced an incomplete assignment due to peak disappearances. Using the CCCTOCSY and HCCCNH experiments, we were able to map and successfully confirm many of our methyl assignments.

4.6.3 Temperature Ramp

A more thorough temperature ramp was recorded from 7 to 50 degrees (Figure 4.23B), with the assignment enabling peak movements to be quantified. At 7 degrees the spectra completely degraded excluding the C-terminus, while the peak intensities at 50 degrees were relatively uniform. Peak intensities generally improved with temperature, excluding the terminus (141K), and flexible loop residues 87M and 90E. The loop was involved in dynamics in the full-length construct, and it might be exhibiting similar behaviour at higher temperatures here. However, these residues are also more exposed to solvent exchange, which would increase with temperature and might explain the trend.

Missing assignments correspond to dynamical regions

Assignments for all residues on the top and bottom turns of the beta sandwich are missing. Peak disappearances of this type are generally explained by chemical exchange. Two reasonable sources of this behaviour are protein structural rearrangements or solvent exchange. Referring to the oligomer crystal structure, the residues on the top turn are capped by the C-Terminal domain, while the residues on the bottom turn are capped by the monomer-dimer loop. The peak disappearances from the former could indicate that the CTD increases the rigidity of the top turn of the beta sandwich. The combination of the dimeric diffusion data and the partial disappearance of the latter, monomer-dimer interface, could indicate the same dynamic behaviour as the full construct. The generally poor signal at 25 degrees combined with the diffusion data can only be explained by exchange.

Probing exchange specific diffusion

Similarly to chemical shift for a spin undergoing exchange, the measured diffusion value can be dependent on exchange regime. In the slow exchange regime, the peaks will be split and each will have its own diffusion coefficient. In the fast exchange regime, the singular peak will have a population weighted diffusion coefficient. In the intermediate exchange regime (which is the only explanation for the peak intensity losses we observe), the faster the exchange regime the more population weighted the diffusion coefficient. Assuming that we have a 2 state chemical exchange mechanism (in the absence of any data), we assume all residues will have the same $k_e x$, and therefore the diffusion coefficients will become a product of the $\Delta\omega$. Any more complex chemical exchange mechanism would make each peak's diffusion coefficient dependent on multiple $\Delta\omega$ values. Long diffusion measurements were made at 20 degrees and scaled according to the sharpest peak (141K terminus). Any deviations from this value might represent a different state. It can be observed here that the spectra overlay almost perfectly, suggesting we aren't able to capture $\Delta\omega$ encoded diffusion differences. There might be a slightly faster diffusion contribution around the major peak, which would represent the unfolded region, but this is probably

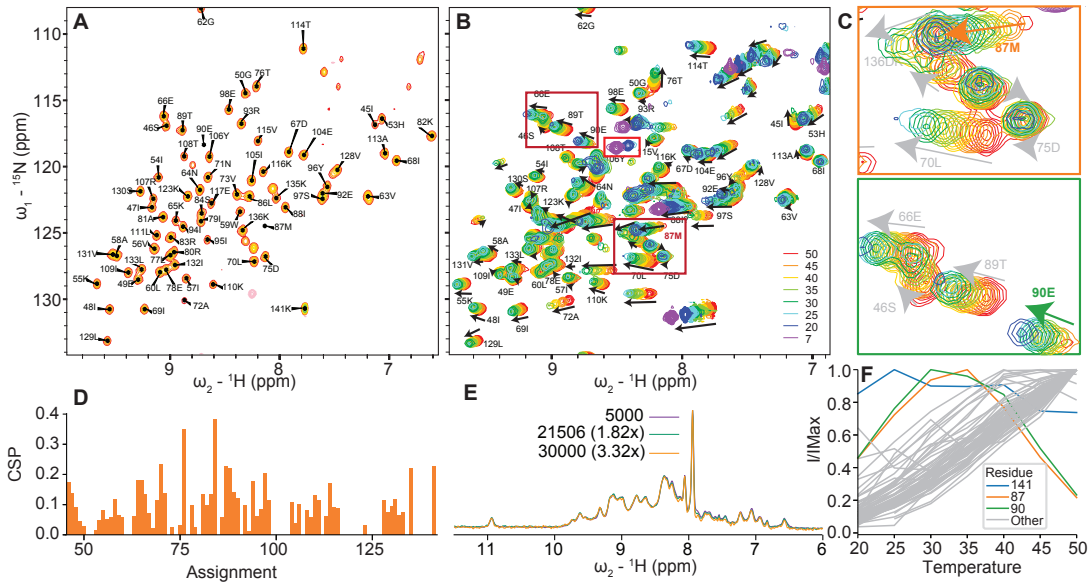


Figure 4.23: The assigned NHSQC is shown next to a temperature ramp from 7 degrees to 50. The majorifty of peaks move away from the solvent peak (4.3-4.8ppm dependent on temperature) with increasing temperature, although some display a different trend. The majority of peaks increase in intensity with increasing temperature, although three residues that don't are highlighted in the bottom right-hand figure. 87M and 90E are both on the flexible loop regions that are involved in the CPMG fits in the larger construct, and there is a possibility they are involved in something similar here. However, the flexible loop is likely to be very solvent exposed, enabling amide solvent exchange that could also explain the intensity losses. The assigned chemical shift perturbation with temperature is indicated (D). Diffusion spectra are probing residue specific changes to the diffusion constant are shown.

within noise. The chemical shift perturbation between the 20 and 50 degree NHSQC spectra is shown (Figure 4.23D). There were no clear structural trends.

Long 1D CPMG data

We recorded long 1D CPMG spectra in an attempt to quantify the chemical exchange. As the signal degrades with temperature, it takes longer to record spectra at lower temperatures. However, we discovered here that the CPMG spectra do show some very limited evidence of relaxation dispersion at 25 degrees, indicative of an excited state. Unfortunately at that temperature we did not have enough signal on our spectrometer to acquire a psuedo-3D CPMG experiment in a feasible time. We recorded a CEST experiment at both 25 and 50 degrees but there was no evidence of an excited state here. The combination of these results might suggest

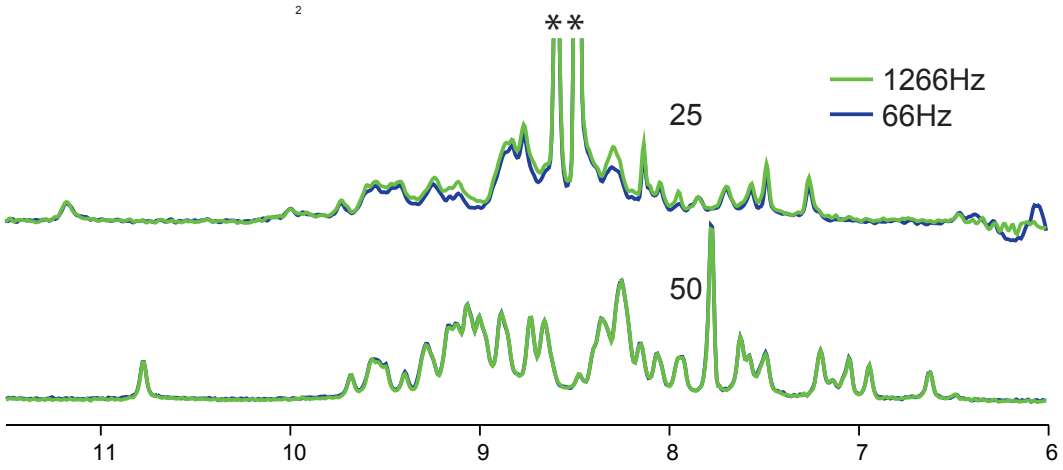


Figure 4.24: 1D CPMG spectra are shown at two temperatures. There is evidence of relaxation dispersion at higher CPMG frequencies at 25 degrees, while at 50 degree the spectra are indiscernable.

an exchange regime that is too fast to create a second saturatable frequency in the CEST experiment or be even partially quenched by CPMG pulse trains.

4.7 Conclusions

This multiple timescale dynamical study of HSP16.5 has enabled us to begin to understand the dynamics of this very large, exchanging protein in the solution state. Starting with the assignment, we then progressed to temperature ramps and dynamical studies. The CPMG measurements demonstrate that the residues on the monomer-dimer interface on the interior of the sphere are exchanging with some state with different chemical shifts. A model consistent with this would be motion in the loop that uncaps the β1 and β2 strands of the Ig ACD domain (Figure 4.2 B). Further dynamics were found on the interior of the sphere through the N-terminal domain. Alphafold models suggest an alpha-helical secondary structure for residues 6-20, which also agrees with some limited NOESY data. This correlates with previous work investigating the N-terminus [53]. Calculations demonstrate a densely packed interior of the sphere, while clashes within the extrapolation from Alphafold dimer to full 24-mer which together may suggest the cavity might not have many viable packing conformations. Combining single quantum and spin-state

selective CPMG data indicates a disorder-order NTD transition. These data are consistent with a ground state alpha-helical NTD that points into the interior of the sphere in a locally mobile state without tertiary structure, while the excited state tucks into the side of the sphere, perhaps stabilising the dynamically uncapped B·S and becoming less mobile through tertiary interactions. There is no data to confirm the NTD excited state stabilising the uncapped beta sandwich, it may be possible to investigate this in the future by measuring excited state residual dipolar couplings, which would give us relative orientation data [79]. This interior region has been indicated in chaperone activity [39] [77] but is poorly accessible to client proteins on the interior of the sphere (Figure 4.2 B). Our dynamical model could be rationalised by a combination of stabilising interactions between the ACD and N-terminal domain, and steric constraints that prevent all monomers doing this simultaneously. Methodologically this work would have been signficantly more challenging without UnidecNMR. Many of the required elements to repeat this work are now included within the UnidecNMR package, including the MAGMA assignment pipeline, FuDA fitting and CATIA Bloch-McConnell simulation of the dynamics experiments. We are hopeful that further smoothing of this pipeline and open-source accessibility should enable more researchers to benefit from the excellent basos of spin-physics and molecular biology that was established over a decade ago [9][31][64][80] and underpins this work.

4.8 Future work

Further improvements to the UnidecNMR pipeline that enabled this work will enable higher throughput studies of similarly complicated and relevant proteins, including membrane-bound G-protein coupled receptors (GPCRs) [81], kinases [59] and plant small heat shock proteins.

One aim that has not yet been completed is to link the exchange of the dimeric construct to the 24-mer construct. We intend to probe this with a new experiment. The methyl groups have substantially more signal than the N-H protons and do not suffer from solvent exchange effects. In order to probe their dynamics, we could

record the CPMG experiment on the methyl groups. The difficulty is the requirement for pulse selectivity [80]. Without selective $^{13}CH_3$ labelling, homonuclear $^{13}C-^{13}C$ coupling can evolve between the methyl group and the adjacent Carbon. Fortunately in Chapter 5 we outline new pulse design software that enables us to selectively excite certain frequencies of spins while avoiding others.

Spectra in the presence of α-lactalbumin promote some intriguing questions and diffusion indicates smaller species interacting. More generally this presents, to the best of our knowledge, the first time the dynamics of a small heat shock protein oligomer have been studied with such a high density of atomic resolution probes.

4.9 Materials and Methods

As previously stated, these samples were not made by me. I include this section for reproducability and more biological context. The following protocol was used to express the proteins used in this chapter (taken from Mike Barber's thesis [78]):

"Isotopically enriched proteins were produced with E. coli BL21 strain cells in minimal media (m9) supplemented with 15NH4Cl and 13C6H12O6. Core construct HSP16.5 protein was expressed with N-terminal tobacco etch virus (TEV) protease-cleavable His-tags and purified by nickel-affinity chromatography. The N-terminal His-tag was removed with TEV protease, and the protein was further purified using a Superdex 75 gel filtration column (GE Healthcare). Protein purity was assayed by SDS-PAGE and found to be 99% pure after purification. α-lac was purchased from Sigma."

The samples were made up in buffer with 50mM NaPi, 100mM NaCl and 2mM Sodium Azide at pH 7.

5

Optimised pulsing during NMR experiments: Seedless

Contents

5.1 Acknowledgements 78
5.2 Introduction 78
5.2.1 Pulses and the Bloch equation 79
5.3 Seedless pulses 82
5.3.1 Pulse optimisation 82
5.3.2 GRAPE pulses 82
5.3.3 Point-to-point vs Unitary 85
5.3.4 B1 inhomogeneity 86
5.4 Using Flourine NMR as a Seedless proof of concept . 86
5.5 A critical component of heteronuclear NMR: Insensitive Nuclei Enhanced by Polarisation Transfer 88
5.6 Triple resonance experiments 89
5.6.1 The requirement for carbon selectivity 91
5.6.2 Alpha-Beta carbon decoupling 92
5.6.3 Pulse sequence analysis 93
5.6.4 The benefits of Seedless pulses 94
5.7 Triple resonance results 96
5.7.1 HNCO 97
5.7.2 B1 inhomogeneity 100
5.7.3 Introducing Hydrogen and Nitrogen pulses 102
5.7.4 Full triple resonance results 103
5.8 NHSQC results 105
5.9 Conclusions 105
5.10 Materials and methods 107

5.1 Acknowledgements

This work has been previously written into paper format but is yet to be submitted. All figures were made by me, with minor editing by Prof. Baldwin. Some of the figures and captions have come directly from the paper draft, which were all written by me with minor editing by Prof. Baldwin. My great thanks to Gogulan Karunanithy and Virginia Casablancas-Antras for producing an isotopically labelled protein at the beginning of this project. Thank you also to Adeline Poh for preparing the Fluorine sample.

5.2 Introduction

NMR employs radiofrequency pulses to manipulate nuclear spin states in a magnetic field. The introduction of short, higher power pulses to excite a broader range of frequencies enabled Fourier transform (FT) NMR [82], which greatly increased sensitivity and was partially responsible for Richard Ernst's Nobel Prize in 1991. This principle has since underpinned an astonishing amount of chemical and biochemical research. Within the field of Biomolecular NMR, FT NMR has expanded into a vast array of multidimensional experiments [83][27] [28] [30]. These concatenate sequences of alternating pulses and delays to select particular spin states and extract biologically relevant information from proteins. The time evolution of spin states during these experiments can be described using the Liouville-von Neumann equation and thought of as control theory for a statistical ensemble of spin states. Kurt Wüthrich, Ad Bax and Lewis Kay then developed many of the isotopically labelled experiments that have become standard in the biological NMR community [84] [27] [29] [85]. The resulting biological information has contributed greatly to the fields of structural biology and biochemistry. The improvements to sensitivity through pulse sequence design led us to consider that optimising the underlying pulses could result in further sensitivity gains. Rectangular

pulses are short and simple, but they have undesirable off-resonance effects and little compensation for the inhomogeneity of the B1 field. To address this, a number of pulse shapes have been designed, including Q pulses[86], BURP and REBURP[87]. More recently, methods from optimal control theory have been used to calculate pulses [88]. Here, we build on this to develop highly selective, B1 compensated and phase only pulses, that we call 'seedless GRAPE pulses'. These can be calculated on the fly in a few seconds. We demonstrate these pulses can increase the NHSQC sensitivity by 45%.

5.2.1 Pulses and the Bloch equation

A pulse is produced by a radiofrequency amplifier at an amplitude and phase for a given duration. The frequency of the pulse is always set to the rotating frame frequency (i.e. the middle of the Fourier transformed spectrum). The amplifier produces the pulse by pushing the given amplitude and phase through the probe coils for the given duration. The action of a pulse on a single spin state can be calculated using the Bloch equations [89]. To demonstrate this, we start with the simplest, short, high power pulses defined by Richard Ernst, which have a rectangular shape in the time domain, representing constant amplitude and phase.

The Bloch equations can be expressed as a 4x4 matrix that describes the time dependence of the single spin magnetisation vector, including T1 and T2 relaxation. Therefore, this matrix contains rates to describe auto-relaxation and interconversion rates. The diagonal elements represent the auto-relaxation rates of each vector axis (R2 when transverse, R1 when longitudinal), while off-diagonal rates describe interconversion, or movement of the vector between axes. X/Y interconversion is described by Ω, which is $\omega_0 - \omega_{rf}$ in the rotating frame, where ω_0 is the Larmor precession frequency and ω_{rf} is the carrier frequency at which the pulse is applied. Therefore, Ω becomes the difference in frequency between the pulse and the spin. Where the difference is large, this component of the matrix, representing x/y interconversion, akin to chemical shift evolution, becomes large. ω_x represents the power of the RF pulse on the x-axis. It can be related to the pulse as follows:

$\omega_x = \omega_1 \cos\phi$ where theta is a description of the phase of the pulse. ω_1 is then related to the field of the pulse $B_1 = -\frac{\omega_1}{\gamma}$. ω_x is therefore the number of times per second the spin oscillates around the x. This is intuitively sensible, as it represents the rotation rate from the z axis onto the y-axis. This can be calculated by considering the rate of a 360-degree rotation. For example, if our pulse performs a 90-degree rotation in 12us at a given power level, the field will be $1/(4 * 12 \times 10^{-6}) = 20833Hz$. $\omega_y = \omega_1 \sin\phi$ provides the other axis. The final column vector with one element, $M_0 R_1$ returns magnetisation to thermal equilibrium Z at long times.

$$\frac{d}{dt}\begin{pmatrix} M_x \\ M_y \\ M_z \\ M_E \end{pmatrix} = \begin{pmatrix} -R_2 & -\Omega & \omega_Y & 0 \\ \Omega & -R_2 & -\omega_X & 0 \\ -\omega_Y & \omega_X & -R_1 & M_0 R_1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} M_x \\ M_y \\ M_z \\ M_E \end{pmatrix} \tag{5.1}$$

Magnetisation evolution of a single spin under the influence of pulses of defined powers can be followed using this matrix. In order to understand the state after the application of a pulse duration t, it is necessary to find a solution to the differential equation. In the rotating frame, none of the elements in the Bloch equation matrix rely on time. Therefore, we can rearrange the equation to:

$$\frac{d}{dt}\vec{M} = R\vec{M} \tag{5.2}$$

Which has a simple solution:

$$\vec{M}(t) = e^{Rt}\vec{M}(0) \tag{5.3}$$

If we can describe the matrix of parameters and initial state, therefore, we can determine the state at time t. Assuming magnetisation starts thermally equilibrated on Z, the initial state is easily defined. Following the trajectory of a single spin during a pulse then becomes simulatable (Figure 5.1). The pulse in Figure 5.1 has a Y phase, and therefore it causes rotation in the XZ plane, with zero offset showing perfect oscillation while 5000Hz offset shows imperfections, known as off-resonance effects.

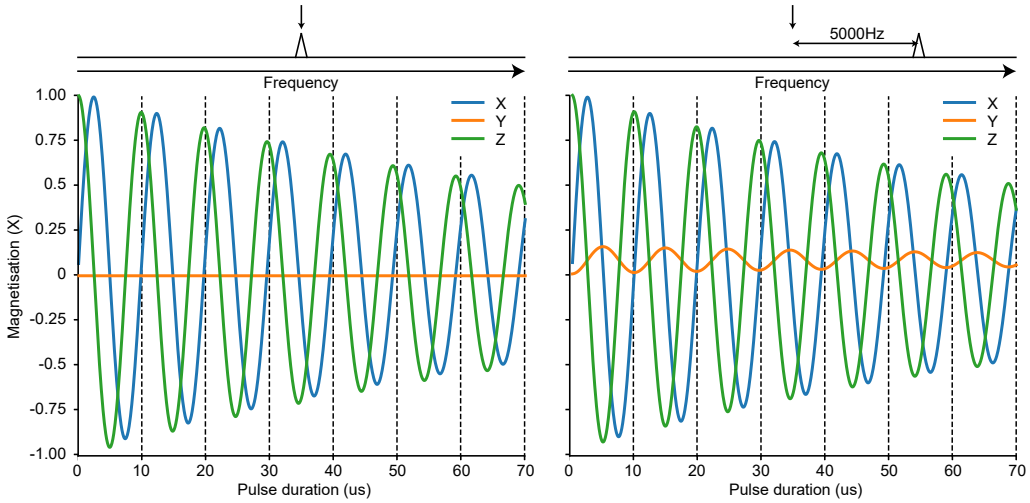


Figure 5.1: Tracking the evolution of an on- and off-resonance spin during a rectangular Y pulse. Off-resonance there are significant off-resonance effects

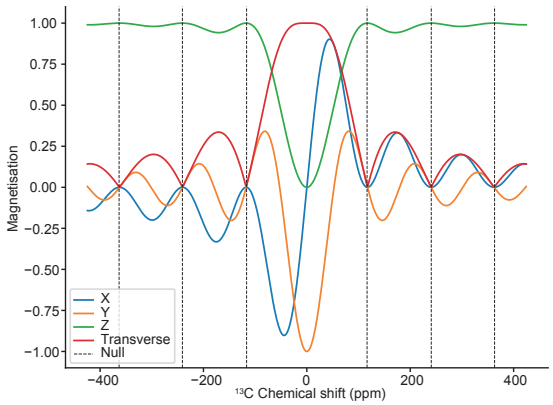


Figure 5.2: The excitation profile of a 55us 90 degree pulse with x phase showing the sinc function and corresponding nulls

The excitation profile of a time-domain pulse shape can be approximated as the Fourier transform. The Fourier transform of a rectangular pulse is a sinc excitation function (Figure 5.2).

To account for these off-resonance effects, pulse elements with variable phase and amplitude can be concatenated to produce a more complex overall pulse shape and a more desirable excitation profile. In this chapter we present Seedless, software that employs GRadient Ascent Pulse Engineering (GRAPE) to calculate pulses in Hilbert space by acting on 2x2 matrices (as opposed to the 4x4 above), as described in the original optimal control theory paper[88]. This is demonstrated to lead to

increased sensitivity and perfect phasing for a variety of pulse sequences.

5.3 Seedless pulses

5.3.1 Pulse optimisation

Historically, optimised pulses have been pre-computed for a more desirable excitation window, rather than designed for the particular manipulation required in a pulse sequence. Examples include SEDUCE[90], REBURP and Q pulses, which often need careful calibration and additional time-reversed pulses to prevent significant off-resonance effects. They are also generally longer and substantially lower power than can be achieved with modern, cryogenically cooled probes and digital amplifiers[91] [92]. However, improvements in computational power present the opportunity to calculate optimal pulse shapes in real time without these problems.

5.3.2 GRAPE pulses

GRadient Ascent Pulse Engineering (GRAPE) has been used to create pulse profiles that are much more desirable, with significantly better control over the spins. This is an application of optimal control theory, which attempts use control variables to optimise a state variable against some cost function. In our case, the control variable are the pulse design parameters, and the state variables are the initial and final states of the spin system. The GRAPE algorithm is designed to optimise the elements comprising a composite pulse (the smaller pulse elements sequentially concatenated to produce a longer, 'shaped' pulse). It optimises these against some cost function related to the input and output states. This cost function can be made from the quantum mechanical 'fidelity', which can be thought of as describing the 'overlap' of any two states. Where the fidelity is equal to 1, the states are identical. Therefore, if we desire an initial Hz magnetisation vector to be transformed to a final -Hy magnetisation vector (similar to a rectangular 90 degree pulse on the x axis), the fidelity function can be used to compare the simulated final state upon application of a pulse (for a given initial state (i.e. Hz)). The higher the fidelity (i.e. the closer the desired-final and simulated-final states), the more optimal the pulse.

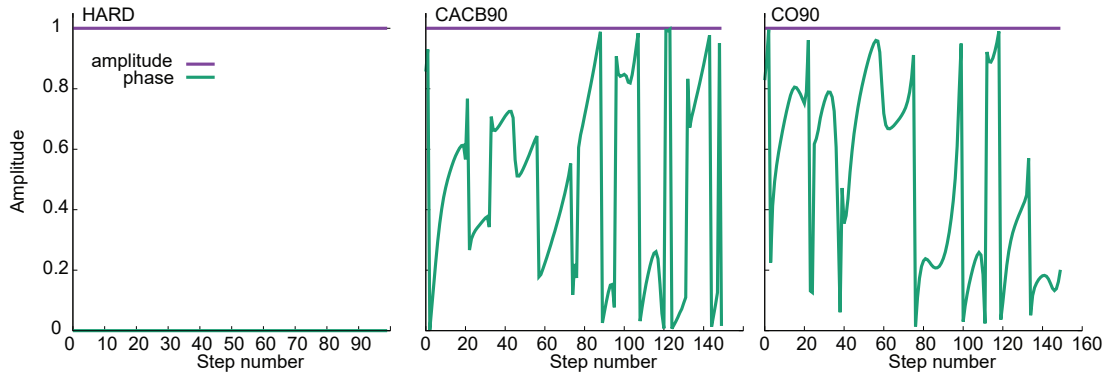


Figure 5.3: The Amplitude and Phase of Seedless pulses

Where the pulse elements can be further optimised, the optimiser will 'ascend' the gradient surface to attempt to improve the final simulated state.

Jonathan Jones' group focuses on Optimal Control theory using NMR, and has developed a GRAPE algorithm to control isolated spin states accurately [93]. We have heavily optimised and packaged this algorithm into user-friendly software, called Seedless, which can flexibly generate pulses according to user inputted requirements. We took code written in Matlab, and translated it into C++. Following significant optimisation, we achieved a performance boost of several thousand times, enabling these pulses to be produced in seconds.

As the optimisation described above suggests, the resulting Seedless pulses are a class of composite pulses, meaning they are concatenations of very short hard pulses of a given frequency, phase and amplitude (Figure 5.3). The resolution of the pulse composition was set at 2μs[1], meaning for a 200μs pulse, the optimiser has 100 steps of phase and amplitude to adjust to produce the desired transformation. These steps can constructively/destructively reinforce actions. The final state of each 2us step is calculated by the Bloch equations, and the output fed into the next 2us step. Therefore, the trajectory of a spin during the pulse can be monitored (Figure 5.4). The simulation in this figure was taken from an on-resonance spin,

[1]This can be easily controlled within the software, but for our purposes we found that this duration was short enough. Going too short on the pulses may start to challenge the speed at which the amplifier can faithfully produce the required phase and amplitude. We rationalised that 2μs was not too far off the shortest 90 degree pulse times (circa 6μs), and the experiment ended up matching simulation well.

but can be repeated for many offsets. This is similar to how the Seedless pulses are created, although rather than optimising in Louivillian space (3x3 matrix), the time evolution matrix neglects relaxation and is expressed in Hilbert space (2x2 matrix). As the optimisation is designed to produce a final state, any off-resonance B0 evolution does not occur. The longer the pulse, the more steps the optimiser can use to constructively/destructively reinforce total action of the pulse.

There are two optimisable axes in pulse design, phase and amplitude. We find that constant amplitude, phase modulated pulses, where the amplitude is set to the highest possible value, generates the most effective pulses. During testing, amplitude modulation gave us very little benefit, and in fact because of amplitude transients, could reduce the efficacy of the pulse. Therefore, we removed amplitude modulation from the optimisation.

Differences to rectangular pulses

Implementation within rectangular sequences required an understanding of the differences between Seedless and rectangular pulses. During a rectangular pulse a transverse component will be building throughout. For off-resonant spins, this component has a torque applied to it by the B0 field for the duration of the pulse, alongside the evolution of any scalar coupling. This means that by the end of the pulse, the final state is dependent on its duration. This requires the subtraction of some duration from the surrounding delays[2].

It is also generally necessary to move rectangular, or pre-computed shaped, pulses off-resonance to achieve a desired manipulation. This requires the application of a time-dependent increase or decrease in phase over the duration of the pulse, called a phase ramp, which enables a single shape to be used to achieve multiple spin-state transformations. This is not required for Seedless pulses.

We have not tested the performance of Seedless pulses extensively against other shaped pulses. There has been some excellent work in this field, and we think it highly likely that in some circumstances Seedless may be matched on performance.

[2]It has been shown that the best subtraction is $\frac{\pi}{2} \times t$ where t is the pulse duration

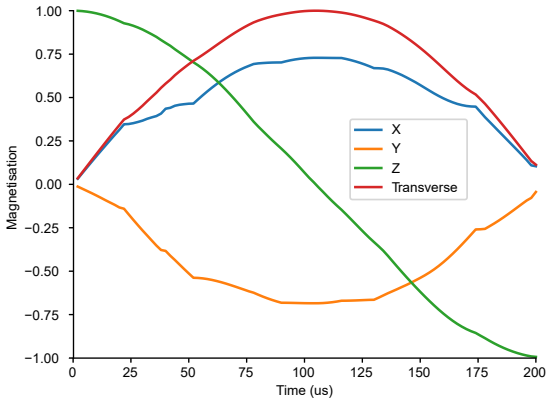


Figure 5.4: The trajectory of a Seedless unitary on-resonance pulse on the three axes and the 'Transverse' component

However, there has been a lack of widespread adoption of shaped pulses, and they are generally a single shape for a single operation (e.g. the REBURP pulse which performs a general 180-degree rotation). To achieve band-selectivity, these pulses must have a defined duration (longer duration means a narrower excitation window) and the amplitude must be adjusted to provide the correct rotation. This means they tend to be low amplitude and require long durations, which is unfavourable for biomolecules with relatively fast relaxation properties. Seedless enables a user to define whatever pulse requirements they have, enabling the specific transform to utilise full power and require a much shorter duration.

5.3.3 Point-to-point vs Unitary

Two types of Seedless pulses were defined: Unitary and point-to-point (P2P) pulses. Unitary pulses perform a full rotation of the Bloch sphere, meaning the starting position of the desired spin can be unknown. P2P pulses transform an initial state to a final state, while any spins with a different initial state are not optimised and behave unpredictably. The benefit of P2P pulses is their comparably shorter duration, as fewer states are controlled. This is beneficial for relaxation. They can be employed when the initial state is known, for example, when rotating a spin 90 degrees from longitudinal to transverse. The initial state of a 180 pulse during a spin-echo, however, is dependent on the preceding spin-specific transverse evolution. Therefore, this pulse must be Unitary. P2Ps were used where possible

to ensure the sequences were applicable to faster relaxing proteins. Both pulses can be thought of as 'zero-time' as the optimiser is designed to produce a final state. Therefore, there is no need to subtract pulse durations from surrounding delays, simplifying the pulse sequences.

5.3.4 B1 inhomogeneity

Our pulse, or B1 field, is intended to apply uniformly throughout the sample coil volume. In practice, probe coil design and sample geometries mean that the homogeneity of the B1 field is definitely not perfect; some regions of the sample receive less amplitude than other parts. While we can carefully experimentally calibrate the response of a hard pulse through nutation experiments to achieve a near-perfect 90-degree rotation, we are effectively calibrating this for the mean B1 value within the sample. Any deviation from this B1 value causes deviation in the resulting rotation of a rectangular pulse. Compensation for B1 inhomogeneity means that for any B1 value between, for example, 90% and 110%, the pulse will perform a 90-degree rotation. There is a rich history of designing pulses that may include compensation for B1 inhomogeneity [94][95][96][97][98]. As our Seedless algorithm allows us to code any conditions into the pulse, we chose to include B1 inhomogeneity.

5.4 Using Flourine NMR as a Seedless proof of concept

We were looking for a simple test case for Seedless pulses. There have been substantial developments in the application of ^{19}F as an orthogonal biological and chemical NMR probe[61][99]. A defining ^{19}F feature is very large chemical shift dispersion, which enables high resolution study. The NMR experiments generally employ rectangular pulses that produce a suboptimal excitation profile, meaning quantification is complicated and even spectra of small molecules must be recorded in chunks. We therefore asked Seedless to design broadband excitation pulses with perfect phasing that covered over 300ppm. This was implemented into a 90 degree

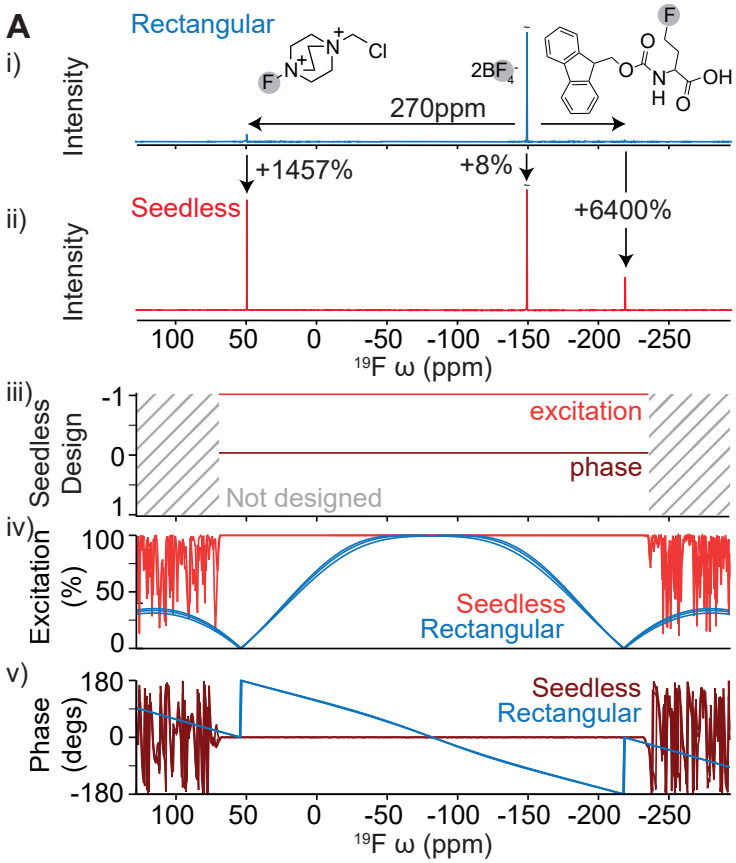


Figure 5.5: Employing Seedless pulses (red, Aii) facilitates study of Fluorine compounds (Ai) when compared with hard pulses (blue, Ai). The remarkable improvements in off-resonance signal-to-noise ratios are shown at up to 6000%. The pulses were designed with even excitation and phase characteristics (Aiii), which are demonstrated in Bloch simulations at 3 B1 fields: 95%, 100% and 103% (Aiv, Av). The sample was prepared in 90% H2O + 10% D2O, and included SelectFluor (with 2 BF4- counterions) and Fmoc protected fluorinated Alanine (the structures are shown in Ai). These compounds were chosen entirely for their large chemical shift range. The experiment recorded was a simple 90 degree pulse-acquire, with either rectangular or Seedless pulses.

sequence, which is the simplest possible. The results confirmed perfect phasing and improved the sensitivity of ^{19}F NMR (Figure 5.5) by up to 6400%.

Having demonstrated an excellent performance on our test case, our attention turned to more complicated, heteronuclear experiments. The most widely used approach to study proteins with NMR is the ^{15}N-^{1}H single quantum coherence experiment (NHSQC). This experiment relies heavily on the Insensitive Nuclei Enhanced by Polarisation Transfer (INEPT) pulse block. It was important to understand the action of the INEPT before applying Seedless pulses to the het-

eronuclear experiments.

5.5 A critical component of heteronuclear NMR: Insensitive Nuclei Enhanced by Polarisation Transfer

As the sensitivity of an NMR experiment is related to the gyromagnetic ratio of the spin, direct detection of heteronuclei is often prohibitively time-consuming. ^{1}H and ^{19}F have the highest nuclear gyromagnetic ratio, while the proteinaceous heteroatoms ^{13}C and ^{15}N have a gyromagnetic ratio of roughly 1/4 and a 1/10 of ^{1}H respectively. Fortunately, Ray Freeman introduced the INEPT block in 1979, which greatly increases the sensitivity of heteronuclear NMR (Figure 5.6). As the name suggests, the INEPT is designed to transfer magnetisation (or polarisation) from sensitive nuclei to insensitive nuclei through scalar coupling. Heteronuclear NMR heavily relies on this phenomenon to increase sensitivity. The INEPT combines a spin-echo on the sensitive nucleus and a population inversion on the insensitive nucleus. Chemical shift evolution of the sensitive nucleus is refocused due to the spin echo, while the population inversion of the insensitive nucleus prevents refocusing of the scalar coupling evolution. Overall, only the scalar coupling operator evolves. When the spin echo delay is set to a total time of 1/2J, magnetisation transfer to the insensitive nucleus is maximised. The chemical shift of the insensitive nucleus can then be recorded during an indirect acquisition block (Figure 5.6). The most sensitive experimental protocol is as follows: Initially exciting the sensitive nucleus, performing an INEPT, recording the insensitive nucleus indirectly, then transferring back to the sensitive nucleus via an INEPT for direct acquisition. By concatenating INEPT and indirect acquisition blocks, magnetisation can be transferred from an initially sensitive ^{1}H nucleus to a ^{15}N nucleus, then a ^{13}C nucleus and back again through scalar coupling with minimal losses (assuming perfect pulses, no relaxation and perfect J values).

The efficiency of the INEPT transfer is dependent on two components: the strength of the ^{13}C ^{15}N scalar coupling constants and efficient refocusing of the

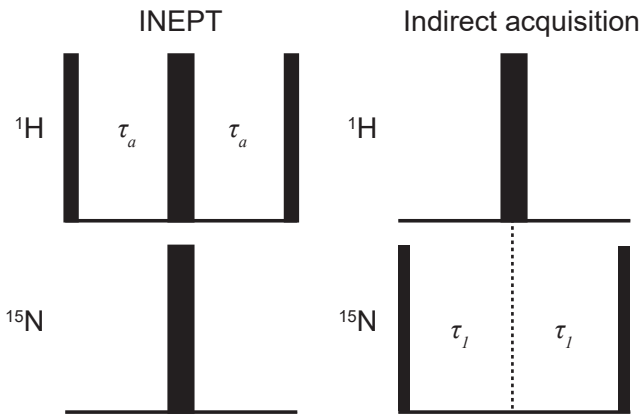


Figure 5.6: The INEPT block combines a spin-echo on the sensitive (in this case ^{1}H) nucleus and a population inversion on the insensitive (^{15}N) nucleus. The initial state is Hz, and final state is HzNz. During the indirect acuqisition block, the insensitive nucleus evolves under chemical shift while refocusing scalar coupling evolution. The time τ_1 is incremented between slices, encoding the amplitude of the direct dimension FID.

spin of interest (otherwise scalar coupling does not evolve). Equally, it is important to avoid refocusing other spins that are scalar coupled to prevent leakage to off-pathway spins. This introduces the need for frequency selective pulses, which can be designed using Seedless.

By concatenating INEPT and acquisition blocks, protein spin states can be manipulated and their precession around the B0 field recorded. Each spin state has its own properties which are encoded by motion and structure. An understanding of these spin states and their properties enables us to tease out biologically relevant information.

5.6 Triple resonance experiments

Armed with an understanding of the INEPT, we turned our attention to the widely used NHSQC, which records scalar coupled Hydrogen and Nitrogen chemical shifts in the form of a 2D interferogram. This can be Fourier transformed in both dimensions to give a 2D contour map with one peak per N-H moiety within the protein. There is roughly 1 N-H group per residue (excluding Prolines and the side chains of Asn, Gln, Arg, Lys residues). This residue specific map of the protein can be used to probe its structure and function under different conditions. However, initially

the structure map has no relation to the sequence, so peaks must be 'assigned' to specific residues in order to access the atomic resolution information. The most common assignment strategy is to label the protein with further NMR active ^{13}C isotopes, making it possible to transfer magnetisation to the $C\alpha$, Carbonyl and side chain carbons via INEPT blocks. This extends the 2D NHSQC experiment to record the chemical shift of a third, carbon dimension. This suite of 'triple resonance' experiments, developed by Ad Bax, has been extremely widely used and is some of the most cited literature. The triple resonance experiments, including the HNCO, HNCA, HNCOCA, HNCACO, HNCACB and CCCTOCSY (etc.), enable the acquisition of backbone carbon chemical shifts. A central component of this assignment approach is the N-Cα coupling constant.

Cα coupling enables assignment

The N_i-Cα_i (-10.7Hz) and N_i-Cα_{i-1} (-7.7Hz) coupling constants are very similar. Therefore, the magnetisation splits when transferred from N->Cα. When the chemical shift is then indirectly recorded, there are 2 carbon frequencies for every N-H peak. Considering the carbon dimension as a set of 1D 'bores' extending from NHSQC peaks, there will be two Cα peaks per bore. Equally, we can see the same carbon chemical shift in 2 N-H bores. One of these corresponds to the current residue N-H, the other to the previous residue. Using this logic, chains of peaks can be connected through the $C\alpha$ chemical shift and other experiments used to determine the amino acid type. The splitting of magnetisation between the i and i-1 $C\alpha$ atoms causes lower intensity peaks, and highlights the need for increased sensitivity.

More sensitivity needed

In general, the triple resonance experiments are carefully designed to record spin states as cleanly as possible to enable highest sensitivity. However, acquisition can still take weeks of expensive spectrometer time and requires careful setup to maximise sensitivity. Low sensitivity can then make the data analysis problem substantially more difficult when peaks have low signal:noise ratios. Any improvements in either the experimental sensitivity or setup, therefore, should improve the accessibility

of NMR protein study to both new and experienced researchers. One example of potential sensitivity improvements we thought possible using Seedless pulses was frequency selectivity during the Nitrogen-Carbon INEPT blocks.

5.6.1 The requirement for carbon selectivity

Due to ^{13}C labelling and INEPT blocks, magnetisation can be transferred along the entire spin active backbone. During the HNCO experiment, the N->C INEPT must transfer magnetisation to carbonyl spins, whilst during the HNCA experiment, it must transfer to Cα spins. Selection of one path or the other requires a selective +Z->-Z transformation of one spin, while the other spin stays on +Z. The BMRB chemical shift statistics show large frequency differences between carbonyl and Cα spins (Fig 5.7).

Rectangular pulse selectivity

The rectangular pulses employed in the original sequences achieved this carbon selectivity through the use of 'soft' rectangular pulses, where the duration is chosen such that one of the oscillating sinc function nulls (Figure 5.2) coincides with the undesirable spins. For example, during an N->CO inept, the null is chosen to coincide with the centre of the CA frequencies (CO is generally between 170 and 180ppm and CA between 45 and 70ppm). This produces a relatively even excitation of CO frequencies but also avoids the excitation of CA frequencies. Spins that are further away from the centres of these two frequencies, for example glycines (45ppm in CA and 170ppm in CO), will not have perfect excitation, and can lose significant intensity. Furthermore, the homogeneity of the B1 field over the sample is never perfect, meaning the amplitude of the pulses has a spatial dependence against which the rectangular pulses have no resilience. It is easy to generate excellent selectivity using Seedless pusles between the carbonyl and alpha/beta carbon chemical shifts. The Cα spin is strongly coupled to the neighboring Cβ spin, and during the HNCA experiment the Cα must be transverse for indirect

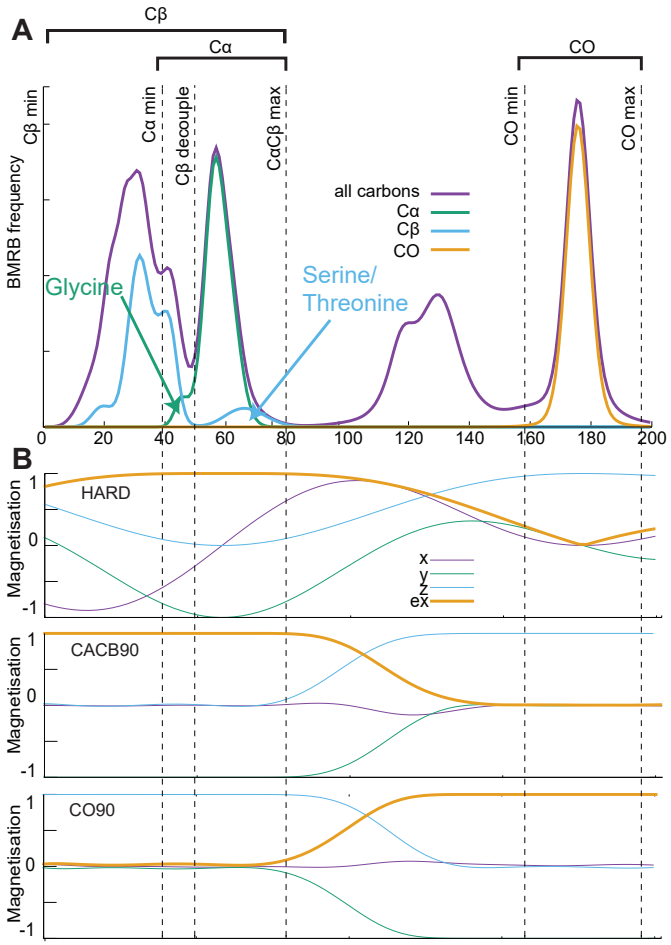


Figure 5.7: The distribution of carbon chemical shifts, including carbonyl, alpha carbon and beta carbon (A). Glycine CA and Threonines/Serines CB are indicated, as their chemical shifts prevents perfect selectivity between alpha and beta carbons. The pulse excitation profiles of hard and two of 90 degree pulses with different selectivities is shown (B), showing a much more perfect excitation profile, even with comparable durations. The Rectangular pulse time was 49us while the Seedless times were 80us.

acquisition. This enables scalar coupling to evolve between Cα and Cβ, which unless decoupled, lowers the sensitivity.

5.6.2 Alpha-Beta carbon decoupling

The BMRB statistics show alpha and beta carbons to have similar chemical shifts, making them difficult to decouple using rectangular pulses (Figure 5.7). Some attempts have been made at this using other shaped pulses, but none have been perfect. This imperfection manifests in some of the peaks showing doublets (not refocused) while others are singlets (successfully refocused) (Figure 5.9). Often a

'solution' is to record the spectrum at insufficient resolution to resolve the coupling. This limits the CA dimension to circa 35Hz (0.16ppm on a 600MHz). A previous attempt to decouple alpha and beta carbons using GRAPE pulses has been published, which confirmed that some peaks provided singlets while others doublets. This complicates data analysis further, as the expected peak shape becomes unpredictable. The algorithm developed by Jonathan Jones' group to generate GRAPE pulses is significantly more efficient than previous attempts, enabling pulse generation in under 2 minutes. We could recapitulate the previous results, but because of the positions of the glycines, serines and threonines (Figure 5.7A), we recognised perfect decoupling was not possible by optimising pulses in the triple resonance experiments.

5.6.3 Pulse sequence analysis

Detailed pulse sequence analysis was necessary to understand how triple resonance sequences might be improved by Seedless pulses. Performing product operator expansion and cancellation would be extremely time-consuming and detailed work. Fortunately, analysis can be greatly expedited by using the symmetry of pulse blocks (Fig 5.8). The first step is to break the sequence down into blocks separated by longitudinal breaks. This usually translates into breaks at 90 degree pulses. Taking the first INEPT block of the HNCO, this analysis reveals two functions: firstly, the H_zN_z operator is allowed to evolve for $1/2J_{HN}$ which brings nitrogen in-phase with respect to Hydrogen for aquisition; secondly, the N_zCO_z operator is also allowed to evolve for $1/2J_{NCO}$, which transfers magnetisation onto the carbonyl spin. This is a good example of the logic of pulse sequence development, as two functions are achieved from one block, saving relaxation time. The total action of this block is to move H_zN_z magnetisation to N_zCO_z magnetisation. Antiphase acquisition of single quantum coherent carbonyl chemical shift is then possible, which occurs in the next block.

There is a single carbon pulse in this block that could be replaced by a Seedless pulse. We hypothesised that the more even excitation of this pulse would lead to more efficient INEPT transfers, particularly at the edges of the CO and CA

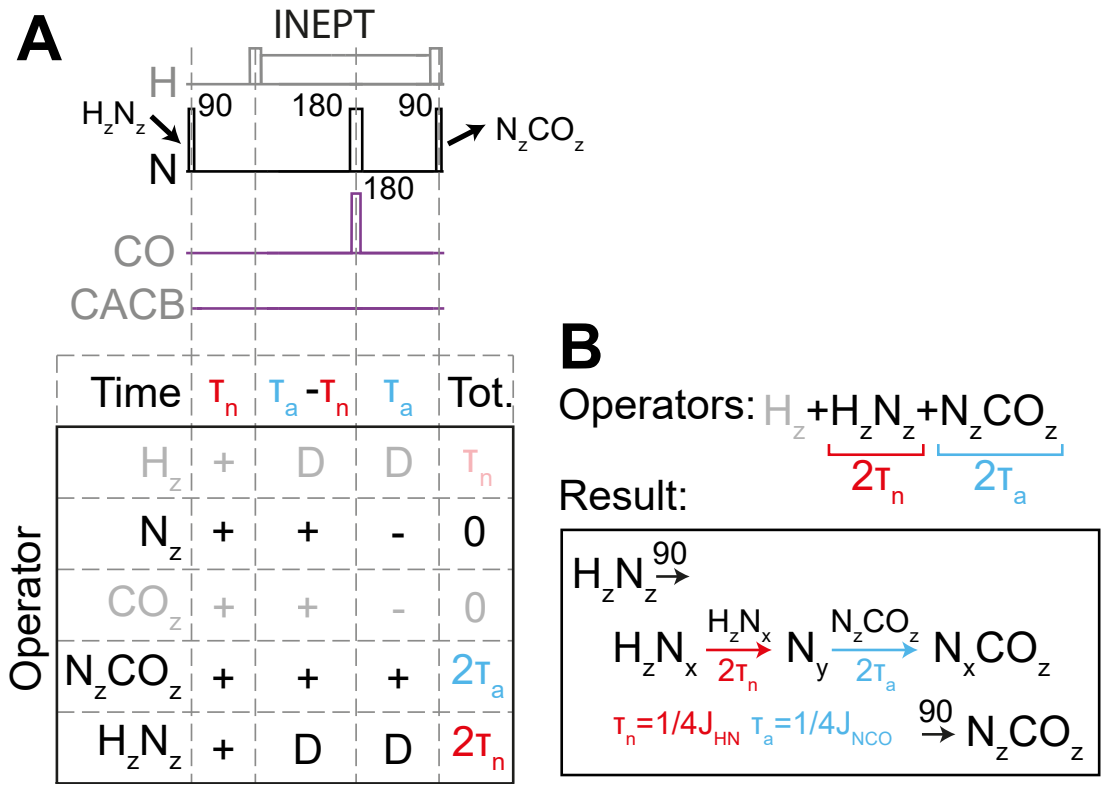


Figure 5.8: The N->CO INEPT block is shown (A). The block begins and ends with 90 degree pulses on Nitrogen. The block is then broken down by 180 degree pulses, which refocus particular spins, effectively cancelling the effect of certain operators. Our initial state for this block (pre-90 degree pulse) is H_zN_z, which is followed by the 90 degree pulse on Nitrogen. As both Hydrogen and Carbonyl spins are never transverse, their chemical shift operators (light grey) do not have any effect. Interpreting 180 degree pulses as flipping the '+' sign in the operator box, we can consider the CO 180 will flip any operator that has a CO_z component (i.e. N_zCO_z and CO_z (although this doesn't apply). We can then apply the resulting signs in these boxes to the duration of that delay to give the total time evolution. The N_z chemical shift operator refocuses to 0 evolution as we have $+\tau_n + \tau_a - \tau_n - \tau_a$. N_zCO_z and H_zN_z do not refocus, and have time evolution. This results in scalar coupling evolution and coherence transfer (B) resulting in H_zN_z being transformed to N_zCO_z.

chemical shift ranges. Applying this logic to the full HNCO pulse sequence revealed two further optimisable blocks involving carbon: carbonyl indirect acquisition and the constant time block (Figure 5.10).

5.6.4 The benefits of Seedless pulses

Having completed the pulse sequence analysis, there were a couple of general considerations.

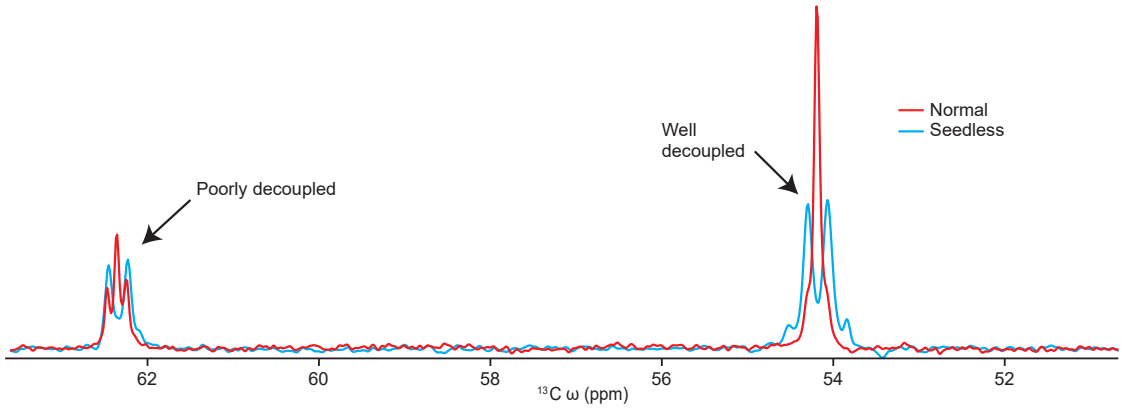


Figure 5.9: Two Cα peaks are shown for residues Glu 33 (54ppm) and Ile 32 (62.5ppm) respectively in abp1p using Seedless and rectangular pulses. The seedless pulse attempts to decouple the Cβ coupling to increase the intensity of the Cα peak. This is performed relatively well for the Glutamine peak (although there does appear to be some smaller longer range coupling that is not refocused), which has a greatly increased intensity. However Ile 32 shows poor decoupling due to imperfect refocusing of the Cβ. The Cβ chemical shifts (determined from a seperate HNCACB experiment) are shown. 39ppm is very close to the 40-45ppm range of the Cα chemical shift of our glycine residues. With a reasonable pulse duration (when considering the relaxation of the protein), we cannot both refocus a peak at 39ppm region and perfectly record the glycine residues. This experimentally demonstrates the overlapped BMRB statistics (Figure 5.7), where CA and CB regions can be overlapped. We contend that perfectly decoupling the Cβ residues while also recording all Cα residues is unfeasible for protein molecules with relatively fast transverse relaxation rates.

1. Rectangular time Subtractions: Much detailed work has been done previously to apply the hard pulse subtractions from the rectangular sequences. As previously discssed, 'zero-time' seedless pulses required these subtractions to be removed. Similar considerations were needed when replacing hetero- and homo-nuclear decoupling sequences on carbon spins with Seedless refocusing pulses. For example, application of homonuclear decoupling of CO during CA acquistion would effectively hold the CA positions stationary during the pulse, as their initial and final states are set to be the same, while CO initial and final states differ by 180 degrees.

2. Carrier definition: For all Seedless sequences the carbon carrier is defined to be the middle of the acquisition carbons, for example 176ppm for the ABP1P HNCO. All pulses are then calculated around that carrier, so the CAs

(40-70ppm), are defined to be -136ppm to -106ppm with respect to the carrier (or -20400Hz to -15900Hz for a 600Mhz spectrometer). Therefore there is no need to ever move the carrier frequency mid-sequence.

Next, we considered the potential benefits of Seedless pulses, as follows:

1. Seedless pulses greatly lower the amount of calibration required for each sample (only the 90 degree full power pulse time is required).

2. The excitation profile is much more even over the required ranges (Figure 5.7 B).

3. The pulses are designed to compensate for the inhomogeneity of the B1 field.

The first benefit requires little analysis; the advantages are evident and will bring us closer to fully automating NMR experiments. To test the second, we explored whether Seedless pulse improvements were a function of chemical shift. To unpick the third, pulses were designed with the same excitation requirements but with and without B1 inhomogeneity compensation.

Other members of the Baldwin Group made uniformly $^{13}C -^{15} N$ labelled actin binding protein ABP1P SH3 domain. This was used to test the performance of the Seedless pulses on triple resonance sequences.

5.7 Triple resonance results

The results throughout this chapter are from correct final pulse sequences, the development of which required the acquisition of hundreds of spectra. To record and analyse these in a reasonable time we employed the two following strategies:

1. To expedite acquisition, we hypothesised that acquiring the 2D ^{1}H-^{13}C planes, rather than full 3D spectra, would expedite the improvements while still demonstrating the 3D improvements. This was confirmed for the HNCO and the results are shown (Figure 5.13).

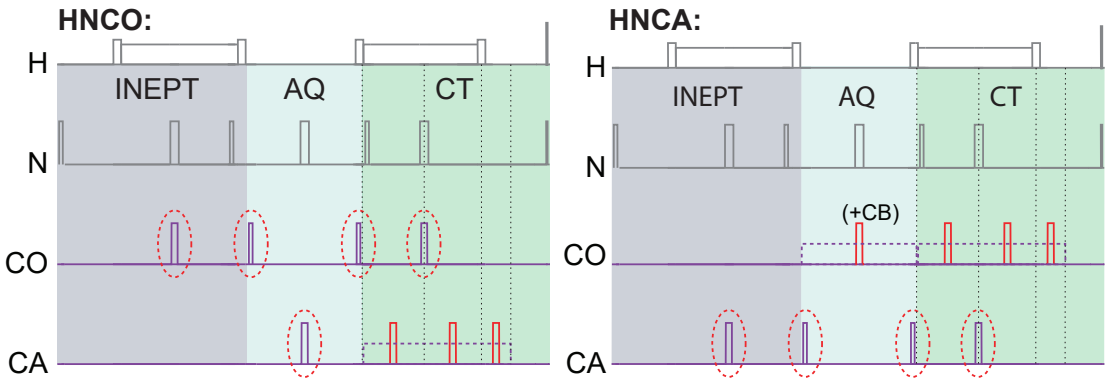


Figure 5.10: There are 3 regions in the carbon region of the HNCO experiment: the N->CO INEPT, the CO acquisiton and the Constant time block. There are 5 carbon pulses identified that can be replaced with equivalent Seedless pulses and one carbon decoupling block which can be replaced with carefully placed Seedless refocusing pulses.

2. To expedite spectral analysis, we developed a pipeline with UnidecNMR 2D mode to peak pick the spectra and output intensities. This gave reliable intensity results, and proved critical in expediting the project.

5.7.1 HNCO

Initially, only carbon Seedless pulses were engineered to account for off-resonance effects. Assuming any intensity gain would be related to better control of excitation frequencies, modest gains were anticipated for the HNCO, as CO-Cα selectivity is not challenging. Application of carbon Seedless pulses to the HNCO gave a 11% improvment to the intensities using Seedless pulses. If these improvements were due to superior excitation profiles, they should be proportional to chemical shift. Using the assignment, a dependence is demonstrated on the Cα adjacent to the Carbonyl, particularly for the off-resonance glycines (Figure 5.11).

To understand the origin of Seedless improvements, the sequence was broken down into blocks, and pulses replaced sequentially. Due to the symmetry of the out-and-back experiments, we grouped the carbon pulses as follows:

1. INEPT refocusing pulses

2. Carbonyl acquisition 90 degree pulses

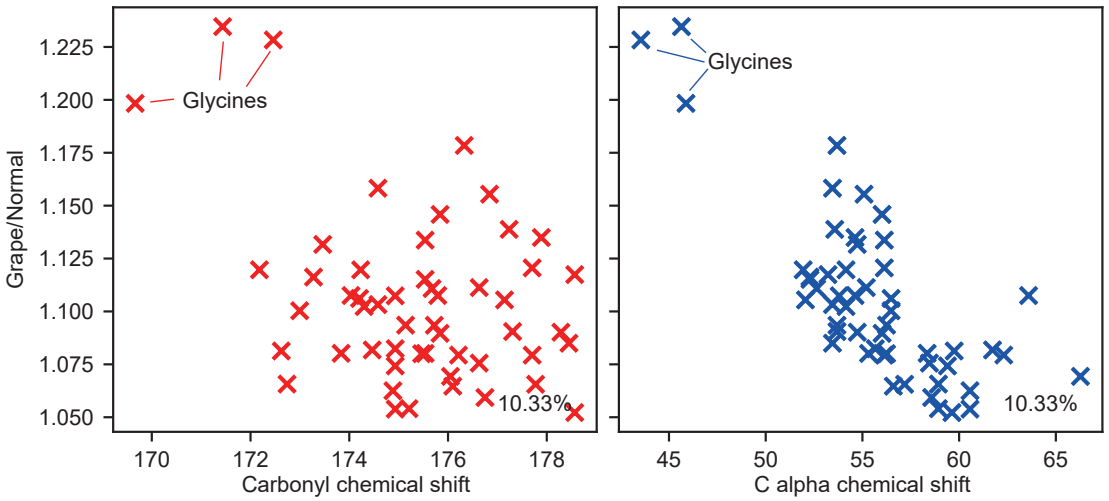


Figure 5.11: The trend of Seedless improvement over hard pulse sequences for given chemical shifts is not clear for carbonyl carbon, but is much clearer for C alpha.

3. $C\alpha$ acquisition refocusing pulse

4. $C\alpha$ constant time refocusing pulses

Considering these blocks we note the following potential losses:

1. INEPT refocusing pulses: N-CA magnetistation transfer during the N-CO INEPT will be lost and should be minimised

2. Carbonyl acquisition: CO-Cα coupling should be refocused during CO aqusition to prevent splitting/peak broadening

3. Constant time block: N-Cα coupling must be refocused and N-CA transfer prevented during the INEPT

These 180 degree pulses can be grouped into two classes: refocusing and INEPT pulses. Refocusing pulses require accurate C alpha +Z->-Z while avoiding CO perturbation. INEPT pulses require no perturbation on C alpha and accurate CO inversion. The rectangular INEPT pulses are applied on resonance at CO, with the power adjusted to achieve a null at the CA frequencies. Conversely the Cα refocusing pulses require a phase ramp to effectively move a pulse off-resonance. The

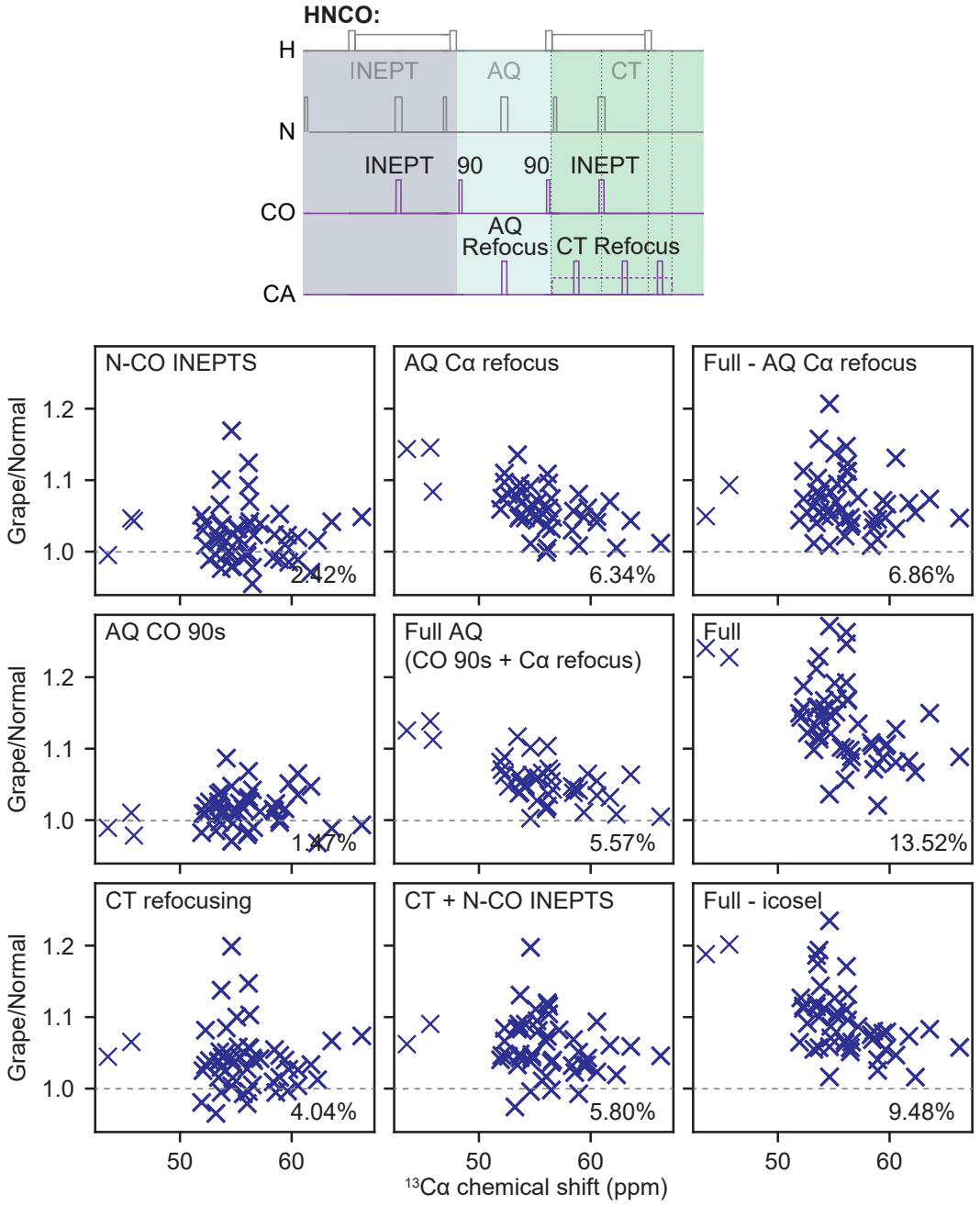


Figure 5.12: Replacement of individual carbon pulses within the HNCO probes the improvements. INEPT replacement includes the CO 180s during the N-CO INEPT. The acquisition (AQ) Cα refocusing only is replaced. The acquisition 90 degree pulses are replaced, and the constant time (CT) block refocusing pulses are replaced. Combinations of these blocks are then taken. 'Full' refers to fully replaced spectra, and 'icosel' refers to the replacement of a gradient pulse during the constant time block to enable better decoupling of Carbonyl spins from the transverse N. 'Full-icosel' refers to the fully replaced spectra without the icosel correction.

overall Cα chemical shift trend must be a combination of the sinc nulls (on-resonance CO INEPT 180), and the top of the sinc function (off-resonance Cα refocusing 180).

The individual hard pulses were then replaced with their Seedless equivalents. The off-resonance glycines were used to probe the chemical shift dependency of each pulse. We computed the average improvement over all peaks. Using Seedless CO 180 degree pulses during the INEPT blocks does not demonstrate a significant improvement (Figure 5.12). Replacing the Carbonyl 90 degree acquisition pulses also does not make a significant difference. Replacing the C alpha 180 degree refocusing pulse, however, makes a substantial difference. During the constant time block, continuous off resonance C α seduce decoupling was replaced with carefully timed refocusing 180s. This proved to give modest improvements. Seemingly replacing the off-resonance 180 degree pulses with Seedless equivalents produced the most substantial effects. The chemical shift dependence also became clearer in these experiments, indicative of a more even excitation profile. This was particularly clear for the Glycine residues, which are at the edge of the C alpha range (45 ppm). This indicates that off-resonance hard pulse refocusing during acquisition and the constant time block are lossy components of the rectangular pulse sequences. There is a roughly additive effect of the individual changes to produce both the overall and Glycine improvements in the fully Seedless spectra.

Finally, an improvement was made to the constant time block. Previously, the continuous decoupling could not extend throughout the block because of a gradient pulse required for the Rance-Kay indirect acquisition mode. The seedless sequences, however, employ 180 degree pulses, enabling refocusing to account for this gradient delay. The benefits of this are shown in (Figure 5.12) where 'Full-icosel' (meaning 'Full' without gradient delay correction) performs worse than 'Full' (meaning 'Full' with gradient delay correction).

5.7.2 B1 inhomogeneity

On-resonance, the frequency excitation of Seedless pulses cannot be superior to well-calibrated hard pulses. However, the results show a significant improvement

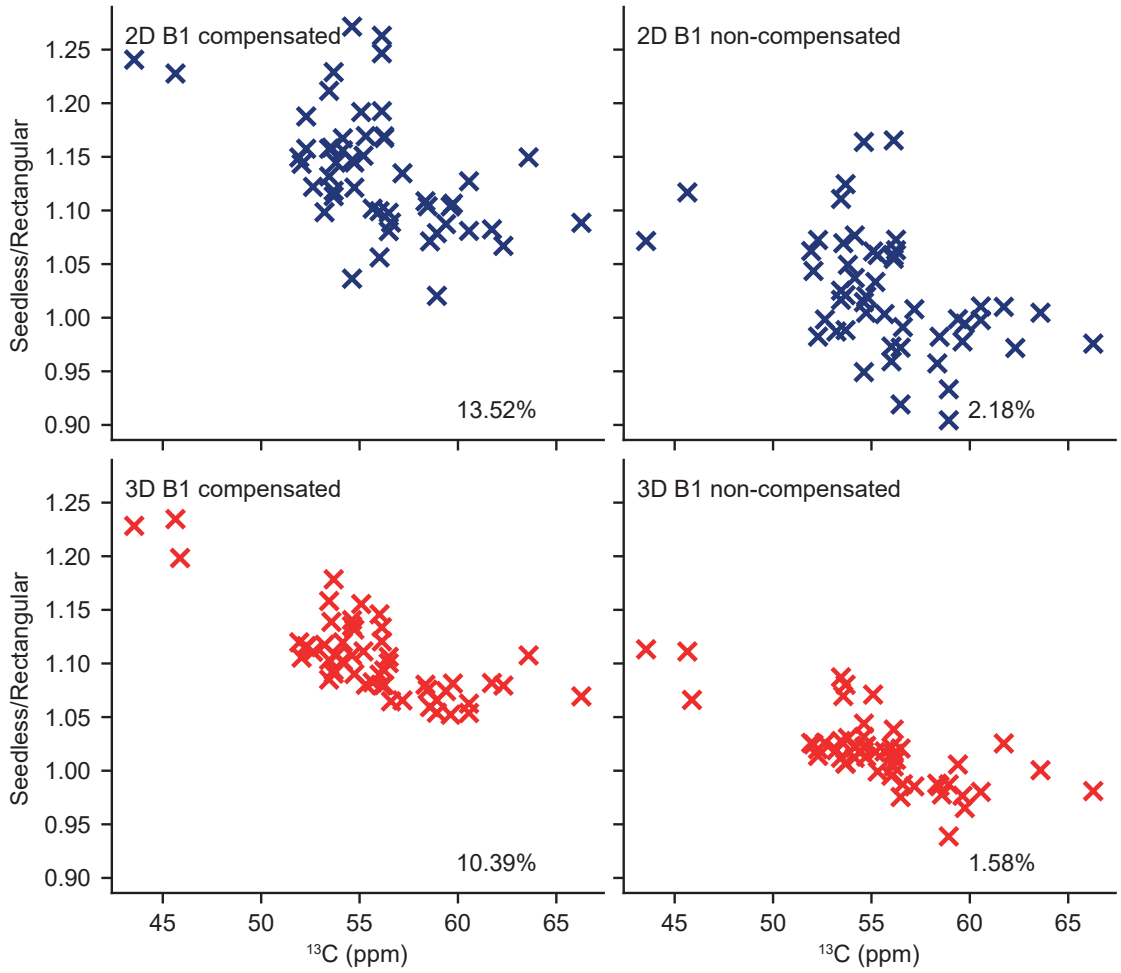


Figure 5.13: 2D and 3D HNCO experiments were recorded to demonstrate the difference between Seedless and rectangular pulse sequences. The peaks were quantified using UnidecNMR and the intensity of the Seedless peak was divided by the corresponding intensity of the rectangular peak. These fractions are then plotted against the Cα chemical shift of the same residue as the CO chemical shift (i.e. the i-1 residue). The results are compared here, showing significantly less scatter in the 3D experiments, but the acquisition time for the 3Ds was 4.5 hours, while the 2Ds was 9 minutes. The percentages correspond to the average improvement from all the peaks. The approach was then taken that broad features would be visible in the 2D experiments, but 3D experiments should be recorded to show final improvements where trends were identified.

on resonance. Another significant benefit of the Seedless pulses was considered; compensation for B1 inhomogeneity. To test this, Seedless pulses were generated without B1 compensation. These pulses are required to transform fewer states for the same pulse duration, and therefore it might be expected, if the B1 compensation were not a significant factor, that the sensitivity of the pulse sequence would increase. Replacing all pulses with like-for-like non-B1-compensated equivalents, however, revealed that B1 compensation is a significant contribution to Seedless pulse improvements. The improvements over the hard pulse sequence were effectively eradicated by removing B1 compensation. B1 compensation should have no dependence on chemical shift, which is confirmed when plotting by a relatively clean y-axis offset between the B1 compensated and non B1 compensated sequences (Figure 5.13). Further, these results were compared with the bruker standard Q pulse HNCO. The Q pulses have an even excitation profile, but poor tolerance to B1 inhomogeneity. The comparison produced very similar results to the Seedless experiment without B1 compensation (Figure 5.14 A i+ii). Simulating the on resonance response of 180 degree CO pulses at different B1 fields sheds light on this effect. The Q and hard pulses are shown to have a very narrow field where they produce the desired 180 degree pulse, while the Seedless is much more tolerant (Figure 5.14 B).

5.7.3 Introducing Hydrogen and Nitrogen pulses

Noting that effective B1 compensation should not be limited to carbon pulses, Hydrogen and Nitrogen Seedless pulses were produced in an attempt to increase the possible improvements to spectra. The same pulse sequence optimisation was repeated as required for the introduction of Carbon pulses. The results demonstrated further significant improvements from using Seedless pulses on all nuclei for a total 28% improvement. We therefore successfully unpicked the two features of Seedless pulses that provide the most sensitivity improvements to the pulse sequence; B1 compensation and an enhanced pulse excitation profile. There has been far more emphasis in the literature on creating pulses with a desirable excitation profile than

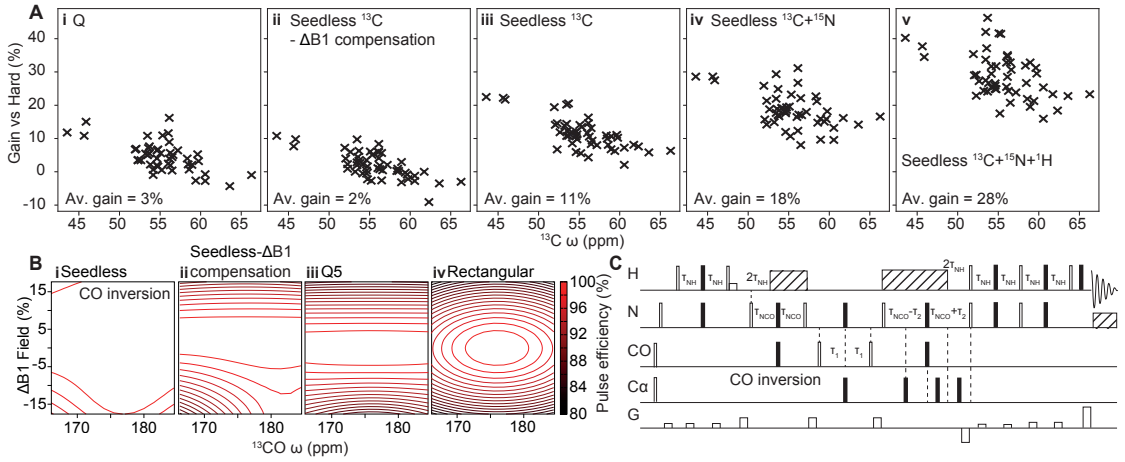


Figure 5.14: Variations of the HNCO with Seedless pulses were compared against a rectangular pulse experiment for the ABP1P sample at 25 degrees Celsius. The results are plotted against Cα chemical shift, where there is a trend. This can be explained by the selectivity against Cα during the INEPT blocks and the Cα refocusing pulse during acquisition. The average improvement is indicated. Q pulses (Ai) showed a minor improvement to the off-resonance Glycine peaks (45ppm). A similar trend was seen for a non-B1 compensated Carbon-only GRAPE experiment (Aii). Introducing B1 compensation for these carbon pulses made a substantial improvement (Aiii), with further improvements (but increased scatter on the Cα trend) when B1 compensated pulses were also applied to Hydrogen (Aiv) then Nitrogen (Av). Simulating the effect of a CO inversion pulse on the CO pulses (B) at various B1 field strengths (ΔB1, y-axis) and chemical shift positions (ppm, x-axis) shows differences in effectiveness (contour and colors) for different pulse types. The pulse efficiency is defined as the 180 degree (+Z to -Z only) inversion, so 96% corresponds to 172.8 degree rotation. The Seedless with (Bi) and without (Bii) B1 compensation shows differences on the ΔB1 axis. The response of the Q5 pulse is similar to the non-compensated Seedless, and the rectangular falls off both in band selectivity and B1 axes. The pulse sequence for the HNCO experiment is shown (C) with the relevant position for the simulated CO inversion (B) indicated (part of the first INEPT block)

on B1 compensation, and the application of this to the triple resonance sequences appears novel. Having developed the understanding and control over these pulses, they were applied to the rest of the triple resonance experiments.

5.7.4 Full triple resonance results

The same strategy was applied to the HNCA, HNCACO and HNCOCA pulse sequences. Sections of the pulse sequence where Cα is transverse were considered carefully due to the Cα-Cβ coupling and frequency overlap (See supplementary information). The resulting triple resonance sequences are shown to improve

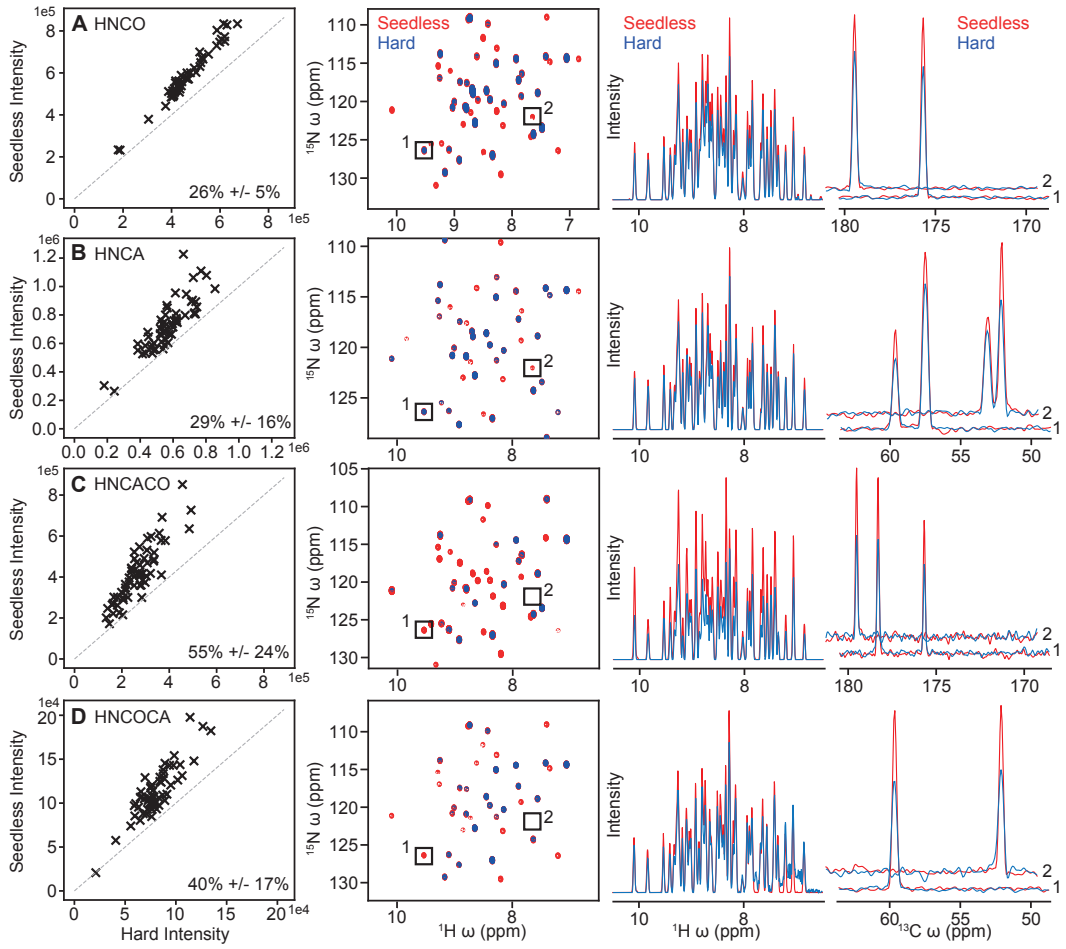


Figure 5.15: All rectangular pulses were replaced with Seedless equivalents and sequences optimised for the HNCO, HNCA, HNCACO and HNCOCA. Some representative improvements are indicated peakwise in the first column (the Seedless intensity is plotted against the rectangular intensity and the y=x line is indicated). The 2D projections are shown in the second column, indicating some improvements to the peak intensities (these maps are contoured to demonstrate the improvements). Skyline projections (noise thresholded before summation to prevent the summation of noise) are shown in the third column and also illustrate the improvements. Selected bores (indicated in the second column) show the improvements for each spectrum in the final column.

remarkably. The number of pulses increases with the sequence complexity, with the HNCOCA and HNCACO more complicated than the HNCO or HNCA. The more complex experiments correspondingly show increased benefit from Seedless.

5.8 NHSQC results

As all the above results were demonstrated on a 600Mhz room temperature probe, we wanted to test the generality of Seedless. Therefore, we applied Seedless pulses to the bruker standard NHSQC pulse sequence on a 950Mhz spectrometer equipped with cryoprobe for comparison. The remarkable results demonstrate a substantial improvement on application of the Seedless pulses (Figure 5.16). The requirement for broader excitation frequencies at higher field is likely important here. Equally, the bruker cryoprobes use a larger sample coil volume than the varian room temperature probes, which is likely decrease to B1 homogeneity.

5.9 Conclusions

Seedless pulses have been shown to produce remarkably improved spectra for a wide range of experiments, including a simple ^{19}F excitation, the full suite of 3D triple resonance experiments, and the NHSQC. These improvements are shown to be a result of two features of Seedless pulses: more even excitation profiles (e.g. ^{19}F) and B1 inhomogeneity compensation (the complex triple resonance and NHSQC spectra). These sequence-specific responses demonstrate the need for closer integration between pulse optimisation and pulse sequence design. Seedless is presented as a tool to enable this, through rapid calculation of pulses that should improve the sensitivity of almost all NMR experiments. The pulse sequences have been simplified through zero-time pulsing and the loss of offset changes. Further, the experimental setup requires only one 90 degree, max power calibration per heteroatom to produce the most sensitive experiment. The resulting spectra also require less processing due to near-perfect phase profiles. The experiments could easily be included in an automated setup workflow to record the highest quality spectra with minimal user input. The flexibility of Seedless enables pulses to be calculated 'on the fly' (pulses are generated in seconds) according to sample-specific parameters, for example chemical shift ranges. Integration with UnidecNMR and the assignment algorithm (Chapters 2 and 3 respectively) could enable a fully automated

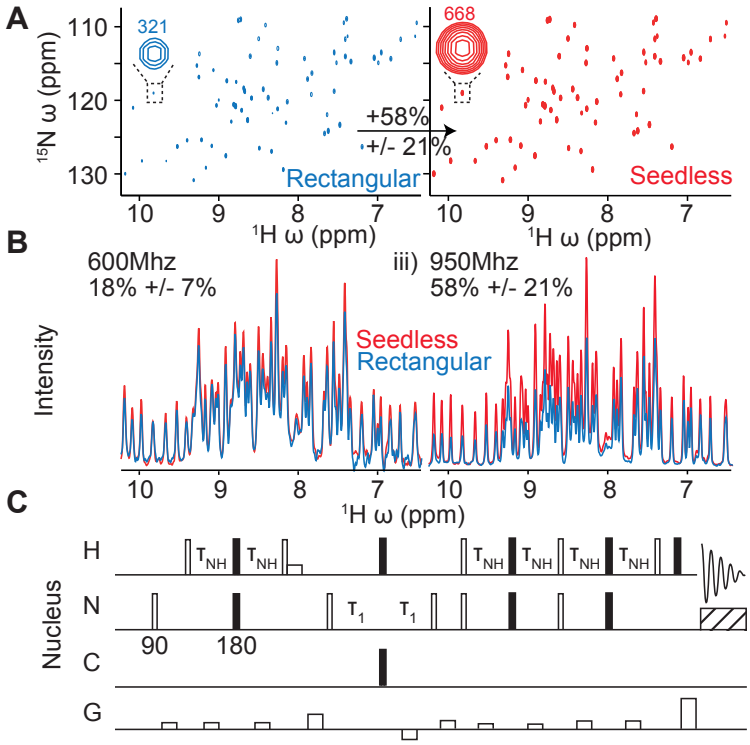


Figure 5.16: Application of Seedless pulses to the NHSQC experiment is shown, with the resulting 950Mhz Rectangular and Seedless spectra compared (A). The improvements are shown to increase between 600Mhz and 950Mhz spectrometers, with the error being a standard deviation of the scatter around the mean improvement. The design and simulation of 180 degree N and H pulses are shown (Biv). The Seedless NHSQC pulse sequence is also shown including an optional carbon refocusing pulse in case of 1^{13}C labelling (C). All 90 degree pulses were state-to-state excitation and de-excitation pulses, while 180s were +Z->-Z pulses where the state is well described (for example the HzNz magnetisation that is evolving during the first INEPT is known to be on +Z and therefore we do not care what happens to any Nitrogen in the X and Y plane). The remaining 180s were Unitary (for example the Hydrogen 180 during the first INEPT, where the position of the spin at the beginning of the pulse is dependent on spin specific chemical shift evolution).

assignment workflow, with real-time feedback from the data analysis platforms to enable input-free assignment. The introduction of Seedless pulse calculation could therefore present a more modern, optimised class of pulse sequences that greatly increase both sensitivity and usability.

5.10 Materials and methods

The ABP1P sample was dissolved in PBS buffer and is at pH 7.4. The NHSQC experiments overlaid well with previous experiments from the group (not shown), indicating the protein was produced as desired. All experiments reported here were performed at 298K on a 14.1 T Varian Inova spectrometer equipped with a 5 mm z-axis gradient triple resonance room temperature probe except the 950 MHz NHSQC spectrum. This was recorded at 298K on a 950-MHz spectrometer with Bruker Avance III HD console and 5-mm TCI CryoProbe, running TopSpin 3.6.1. The experimental parameters were varied between experiments but kept constant for Seedless/Rectangular comparisons. Non-Uniform sampling was employed to expedite 3D data collection with 25% of data points collected, and the rest interpolated using the SMILE algorithm [100]. Data were processed with the nmrPipe software package within the UnidecNMR GUI [10]

6

Universal Saturation Transfer Analysis

Contents

6.1 Acknowledgements 109
6.2 Introduction 110
6.3 Development of uSTA 112
6.4 BSA-Tryptophan application 114
6.5 Haemaglutinnin influenza application 115
6.6 Cryptic SARS-CoV-2 spike sialoside binding 117
6.7 Cryo-EM pinpoints the sialoside attachment site . . . 121
6.8 Discussion and conclusions 126
6.9 Future Work 128
6.10 Materials and Methods 128

6.1 Acknowledgements

This section is rephrased in my own words from our published Science manuscript [101]. All figures have been taken directly from the manuscript. The figure captions have been mildly edited but kept largly as they were in the manuscript. All of these figures were made by me and edited by Andrew Baldwin and Ben Davis, with the exception of the Cryo-EM figure that was made by Prof. James Naismith's group with editing by Prof. Ben Davis. This is included to help understand and visualise the Cryo-EM data. This was an extremely large inter-disciplinary collaboration

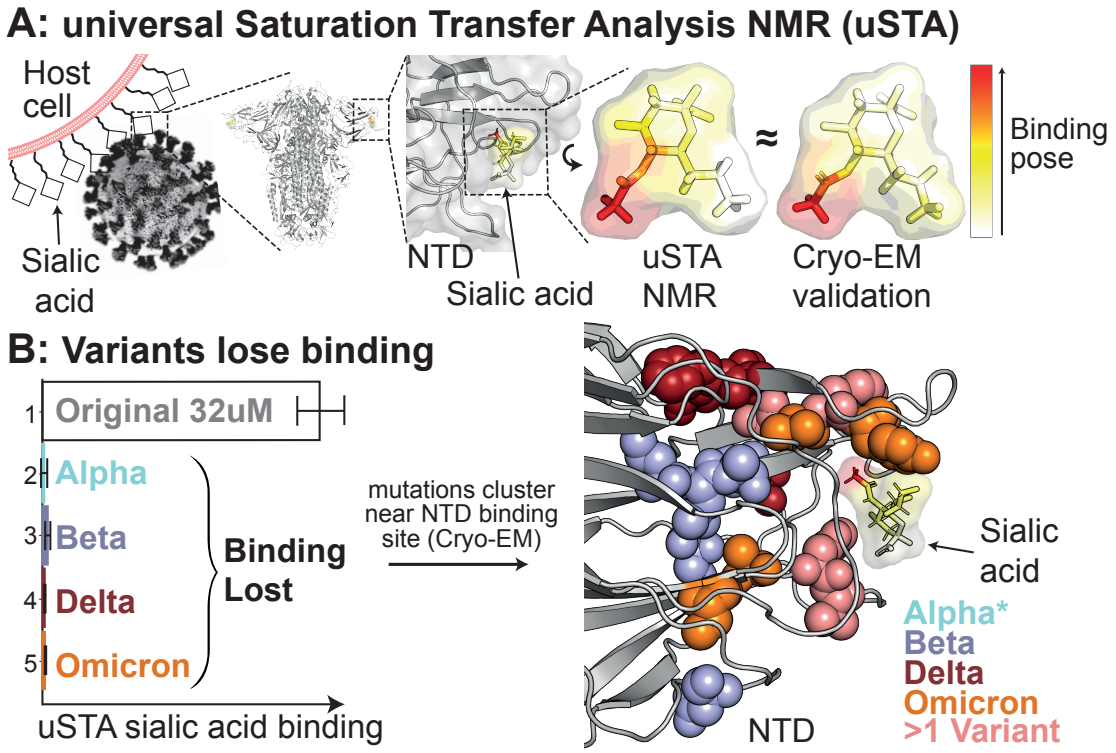


Figure 6.1: (A) This site and pose, which were confirmed by cryo-electron microscopy, are heavily mutated in variants of concern. (B) Analyses reveal a loss of sialoside-spike binding, rationalized by clustering of mutations around the binding site. *All alpha-variant mutations are in >1 variant.

between multiple groups in mutliple countries. It would not have been possible to do this work without the devoted team we had around us. My contribution was the design and implementation of UnidecNMR to perform the data analysis, which constitutes much of the method, and the 'heat maps' of molecules with atomic resolution informs on how small molecules interact with proteins.

6.2 Introduction

Proteins found on the surfaces of cells and pathogens are critical components in the life-cycle of viruses, including entry, exit, progression and transmission. Mammalian cell types decorate their proteins with post-translational modifications, an example of which is glycosylation. These glycans can be linked to Asparagine (N-linked) or Serine/Threonine (O-linked) residues and are essential for both stability and function. N-linked glycans can terminate with a Sialic acid monosaccharide,

which many viruses have been shown to use as host cell-surface attachment sites. For example, Influenza uses the Haemaglutinnin protein to attach to sialosides on cell entry and the Neuraminidase protein to cleave the sialic acids on exit, freeing the replicated viral particle [102]. Middle Eastern Respiratory Syndrome Coronavirus (MERS-CoV)[103], which is related to the 2019 severe acute respiratory syndrome (SARS-CoV-2) virus, has also been shown to bind sialosides through the N-terminal domain (NTD) as part of a cell-entry strategy [103] [104][105][106][107]. Conversely, limited information was available on the interaction between SARS coronaviruses (SARS-CoV-1 and SARS-CoV-2) and sialosides, but both are known to bind to the cell surface protein Angiotensin Converting Enzyme-2 (ACE-2) through their receptor binding domain (RBD) [108] [109]. Other work that had investigated sugar interactions had identified the RBD as the potential binding site[110][111], despite a sugar binding fold in the N-terminal domain and the aforementioned NTD binding domain of MERS[106]. There were no confirmed functions of the SARS-CoV-2 NTD, although many NTD-specific neutralising antibodies had been discovered[112]. The heterogeneity/disorder of the NTD, alongside heavy glycosylation and low expressibility/solubility all contribute to make the SARS-CoV-2 spike ligand interaction hard to investigate. More generally, atomic resolution information about protein-ligand binding is very valuable, but difficult to establish through traditional techniques. There is also no technique that would currently enable the thermodynamics and kinetics of the interaction to be explored simultaneously to atomic resolution binding information.

Principally, NMR, as a solution state atomic resolution technqiue, could solve this problem. Saturation Transfer Difference NMR has been widely used to qualitatively determine ligand binding[113][114], although we believed that armed with more incisive analysis and modelling technqiues we could greatly increase the output information. We therefore devloped a new workflow called universal Saturation Transfer Analysis, or uSTA. This workflow includes application of UnidecNMR (Chapter 2) to extract atomic resolution information and the modified Bloch-McConnell equations [67] which unveil kinetic and thermodynamic parameters.

Atomic resolution information then leads to structural restraints that could be used to inform modelling using HADDOCK [115] [116]. We first demonstrate the potential of uSTA using a simple system with a known protein-ligand binding pair, Bovine Serum Albumin (BSA) and tryptophan [117]. We then move onto Influenza Haemaglutinnin (HA) [102] and 2,3 sialic acid before looking at the binding of a panel of ligands with the SARS-CoV-2 spike protein, including natural and unnatural ligands. This is despite a strong confounding background signal arising from the mobile glycans. We confirm the binding of sialosides with the spike protein and identify a separate binding site for the modified ligands. Fitting the Bloch-McConnell equations, we determine kinetic and thermodynamic parameters in agreement with surface plasmon resonance (SPR) measurement. Our discoveries were further confirmed by collaboration with the Naismith and Berger groups, who generated a very high resolution cryo-EM structure of the SARS-CoV-2 spike protein with a substantial increase in NTD resolution. Upon the discovery of variants of concern, several NTD mutations were noted but their role was unknown. We demonstrated through uSTA application that the sialic acid binding was ablated. Our results correlated with collaborator genetic data that identified two genes as important for patient outcome in pre-variant early pandemic. These genes are known to code proteins that install sialylated glycans on host-proteins.

Taken as a whole, our results present the development of a very relevant analysis technique that help biochemists understand the interactions between large proteins and small molecules.

6.3 Development of uSTA

Starting with the traditional STD experiment to examine the spike-sialic interaction during the early pandemic, we noted some significant difficulties which led to the development of uSTA. Firstly, there were significant differences between previously NMR derived kinetic and thermodynamic parameters (K_d of $190 \pm 50\mu M$)[118] and other biophysical technqiues (e.g Isothermal calorimetry showing a K_d of $30 \pm 9\mu$M) [119].

The Bloch-McConnell[67] equations have been discussed previously in this work and here they are used to model the observables in the experiment. Some limited work has employed the BM equations in STD NMR, leading to the development of CORCEMA [120]. This software requires a crystal structure and attempts to calculate the cross-relaxation rate of all spins in the system using distance restraints. In many real-world cases, including the initial work on the spike-sialic system, we do not have a crystal structure of the ligand bound to the protein. There is also not yet a model that can reliably reproduce NOE, or cross-relaxation, intensities from a given crystal structure. Taken together, this indicates that CORCEMA will not produce accurate results. We developed our own approach to extract the k_{off}, k_{on} and K_d parameters. We found that our approach produced results that were in line with other biophysical techniques. Another approach based on intial magnetisation rates was simulated, and our BM approach was shown to be more accurate (Figure A.4)

As discussed above, relevant mammalian proteins are often covered in mobile glycans that can produce resonances that directly obscure the ligand resonances. Relaxation filters cannot be used without compromising the sensitivity [121]. We employed a routine within uSTA to subtract the protein baseline, enabling quantification of the resulting ligand signal.

We demonstrated the importance of accounting for the direct effects of the saturation pulse on the ligand resonances. Application of saturation at a frequency too close to a ligand resonance could cause a false positive STD response affecting the result. Therefore, a control experiment was recorded to quantify the necessary gap between the saturation pulse and nearest ligand resonance. This demonstrated a >0.3ppm distance between ligand resonance and saturation pulse was required to produce a near invariant heat-map (shown later in Figure A.7).

Sugar molecules display complex multiplet resonances resulting in very complicated spectra. In order to reliably quantify them, we employed UnidecNMR in 1D to determine the peak locations in the off-resonance 1D spectrum. The resulting locations were then used as a restraint to deconvolve the STD spectra, which produced excellent quantification of the often noisy data. It also resulted

in clean data structures, as the STD intensities directly corresponded to some value in the raw 1D spectrum.

uSTA then automatically calculates the transfer efficiency of each resonance. It can map this to an inputted assignment to produce per-atom transfer efficiencies and align these to a pymol structure of the molecule. It will then output a 'heatmap' of binding based on each residue. As we are only measuring protons in this experiment, we interpolate the predicted heteroatom transfer efficiency. This is done by iterating through each heteroatom and calculating the set of distances to all protons as per the following formula:

$$S_i = \sum_{j=1}^{j=n} \frac{S_j}{r_{ij}^6} \tag{6.1}$$

Where i is the ith Heteroatom, j is the jth proton, S is the transfer efficiency of atom and r_{ij} is the distance between heteroatom i and proton j.

6.4 BSA-Tryptophan application

Before studying the relevant system (SARS-CoV-2), it was important to validate the method by applying it to a simpler system to test our improvements. Bovine Serum Albumin is known to bind to tryptophan[119] [117] and the hydrophobicity is thought to affect the plasticity of the interaction. There is no crystal structure demonstrating tryptophan bound to BSA, but it has been perceived to have been delineated well by previous STD studies[122]. Three experiments were recorded, on a protein only sample (P), a ligand only sample (L) and a protein-ligand mixture sample (P+L). The experiment was recorded with an on- and off-resonance saturation pulse at 5.44ppm and 38ppm respectively. We chose the 5.44ppm to saturate the protein (which can be seen by the presence of the broad protein baseline in the STD spectrum (Figure 6.2 E i, ii, iii)) while not saturating the unbound ligand. A broader search of BSA-tryp excitation frequencies is shown in Appendix A Figure A.8). This demonstrates that upon subtraction of any residual ligand response, the extracted heatmap is effectively independent of excitation frequency

(assuming the saturation pulse is applied > 0.3ppm from a ligand resonance). The duration of the saturation pulse was incrementally increased using the following times: 0.1s, 0.3s, 0.5s, 0.7s, 0.9s, 1.1s, 1.3s, 1.5s, 1.7s, 1.9s, 2s, 2.5s, 3s, 4s and 5s. For each time there was a on- and off-resonance spectrum produced. The STD spectrum is a subtraction of the on-resonance spectrum from the off-resonance (referred to as 1D in figures). Positive signal in the STD spectrum comes from cross-relaxation between the protein and ligand, meaning an interaction must have occurred. UnidecNMR quantified the 1D spectrum (Figure 6.2D dark blue) and the accuracy was confirmed by an excellent match between the simulated spectra (red) and raw data. By applying these locations to quantify the STD spectra at a high ligand concentration (Figure 6.2Ei) we were able to faithfully recapitulate the spectrum and, having mapped this to an assignment, uSTA provided an atomic resolution heat map of the interaction (Figure 6.2Fi). We lowered the concentration of the ligand while keeping the protein concentration constant which significantly lowered the signal-to-noise ratios, but would also provide the ability to fit the Bloch-McConnell equations later. We were satisfied to see a striking similarity in the heat maps even at the very low signal-to-noise in the 40uM sample (Figure 6.2Eiii + Fiii). This gave us confidence in the ability of uSTA to extract correct numbers. These heat maps all indicated a strong interaction with the heteroaromatic indole side chain of the Tryptophan. This proved consistent with X-ray structures of BSA and other hydrophobic ligands ([117] here they looked at Naproxen, which has both a carboylate tail and two aromatic rings, chemically similar to tryptophan). We then extracted the build up curves for the protein, and used that to fit the BM equations (AppendixB). We extracted $K_d = 38 \pm 15\mu M$, $k_{off} = 1.6(\pm 0.6) \times 10^5 M^{-1}s^{-1}$ and $k_{off} = 6.0 \pm 2.0s^{-1}$, which was in excellent agreement with previously determined ITC measurement $(30 \pm 9\mu M)$ [119].

6.5 Haemaglutinnin influenza application

We next applied uSTA to an interaction between the NC99 H1N1 Influenza Haemaglutinnin (HA) trimer, which is well established to bind sialic acids [123]. 2

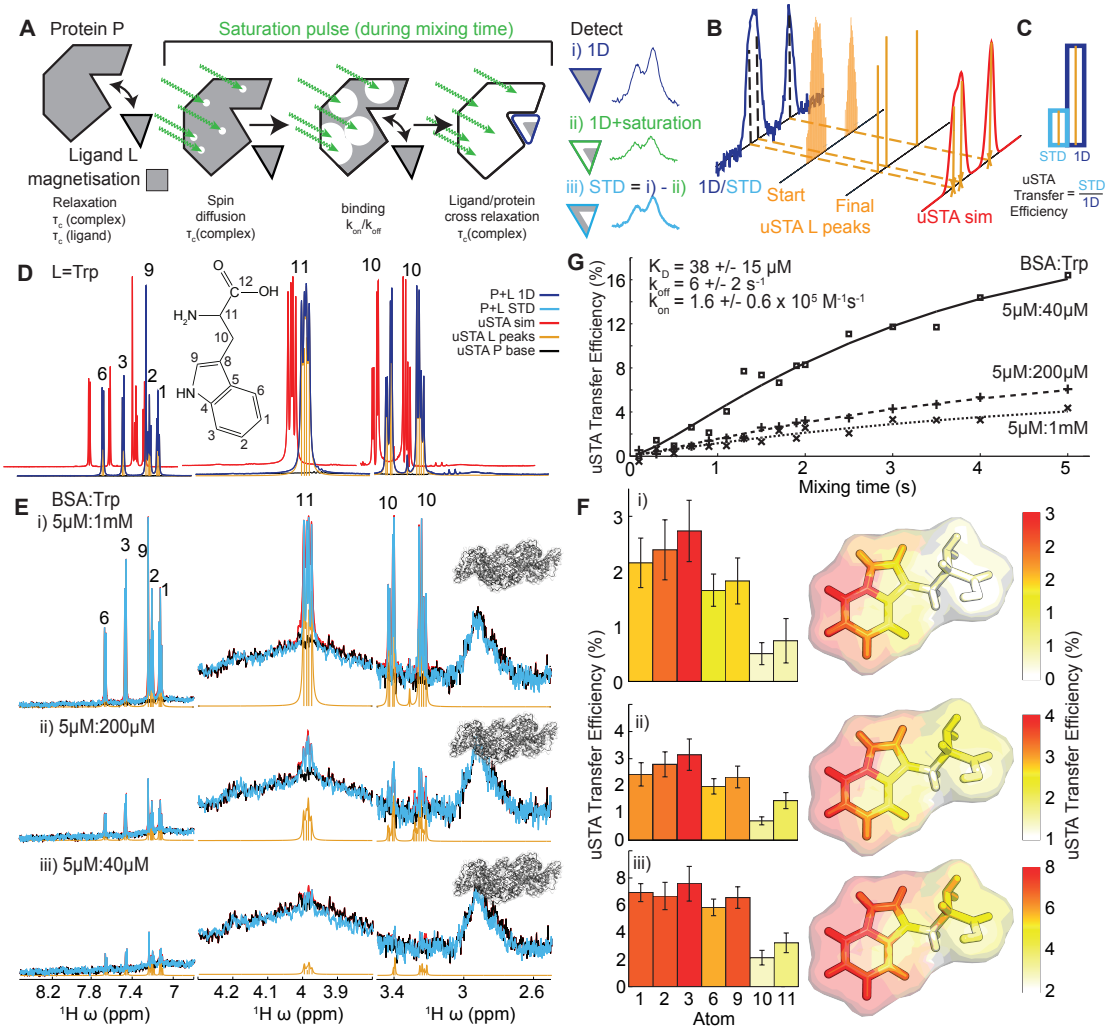


Figure 6.2: (A to C) Schematic of uSTA including comprehensive numerical analysis of relaxation and ligand binding kinetics (A) using full and automatically quantified signal intensities in NMR spectra (B) and calculation of per-resonance transfer efficiencies (C). In (B), signal analysis determined the number of peaks that can give rise to the signal, and returned simulated spectrum by convolving these with peak shape function. The precise peak positions returned are in excellent agreement with the known positions of resonances identified using conventional means. Magnetization transfer NMR experiments compare two 1D NMR spectra, where the second involves a specific saturation pulse that aims to "hit" the protein but "miss" the ligand in its excitation. This is accomplished by acquiring the 1D spectrum with the saturation pulse held off resonance at -35 ppm such that it will not excite protons in either ligand or protein [labelled "1D" in (B) and (C)]. The uSTA method requires these two spectra to be analysed as described in (B), in pairs, one that contains the raw signal, and the second that is the difference between the two. We define the "transfer efficiency" as the fractional signal that has passed from the ligand to the protein. (D to F) Application of uSTA to study the interaction between bovine serum albumin (BSA) and L-tryptophan (Trp). In (D), the 1D 1H-NMR spectrum of the mixture at 200 mM Trp and 5 mM BSA (= P + L, blue) is dominated by ligand, yet the ligand (L) and protein (P) can still be deconvolved by universal deconvolution, using a reference obtained from a sample containing protein only.

Figure 6.2 ***(previous page)*****:** This reveals contributions from individual multiplets originating from the ligand (yellow) and the protein-only baseline (black), allowing precise recapitulation of the sum (red). In (E), application of universal deconvolution to STD spectra with varying concentrations of tryptophan allows uSTA using ligand resonances identified in (D). This in turn allows signal intensity in the STD spectrum (P + L STD, light blue) to be determined with high precision. Although signal-to-noise in the STD increases considerably with increasing ligand concentration, the measured atom-specific transfer efficiencies as determined by uSTA are consistent [(F), left, bar charts; right, transfer efficiency binding "maps"], showing that the primary contact between protein and ligand occurs on the distal edge (C-1, 2, 3, 4; N-7 and C-9 using the numbering shown) of the indole aromatic ring. Application of the same uSTA workflow also allowed precise determination of even weakly binding sugar ligand trehalose (Glc-a1, 1a-Glc) to E. coli trehalose repressor TreR. Again, uSTA allows determination of transfer efficiencies with atom-specific precision (see fig. S3F). (G) Quantitative analysis of the STD build-up curves using a modified set of Bloch-McConnell equations that account for binding and cross-relaxation allows us to determine thermodynamic and kinetic parameters that describe the BSA-Trp interaction, KD, k_{on}, and k_{off}. The values obtained are indicated and are in excellent accord with those obtained by other methods (25, 32). Errors come from a bootstrapping procedure (see Methods). (Caption and figure taken and mildly edited from [101]).

forms were expressed, wild type HA and HA-ΔRBS, which introduces a glycosylation site in the sialic acid binding pocket, aiming to disrupt binding [124] (Figure 6.3 A). uSTA demonstrated that the wild type HA binds 2,3-sialyllactose (henceforth called 2,3) in an end-on manner, specifically binding the terminal sialic acid (Figure 6.3 B, C). This specificity rationalises the virion-surface Neuraminidase, which only cleaves the sialic acid. 2,3 binding to ΔRBS mutation was still picked up by uSTA, but was substantially reduced, but more experiments would be required to determine the reason (some explanations might include incomplete occupation of the glycan site, decreased binding affinity for occupied pockets or a secondary binding site).

6.6 Cryptic SARS-CoV-2 spike sialoside binding

Having established the efficacy of the method, we probed the binding of sialosides to SARS-CoV-2. An initial, long 1D spectrum of the spike trimer revealed extensive glycosylation that were mobile enough to produce significant signal in the ^{1}H region between 3.4 to 4.0ppm (data shown later in Appendix A Figure A.3 C i). They were not sharp, they were consistent with complex N-glycosylation profiles present

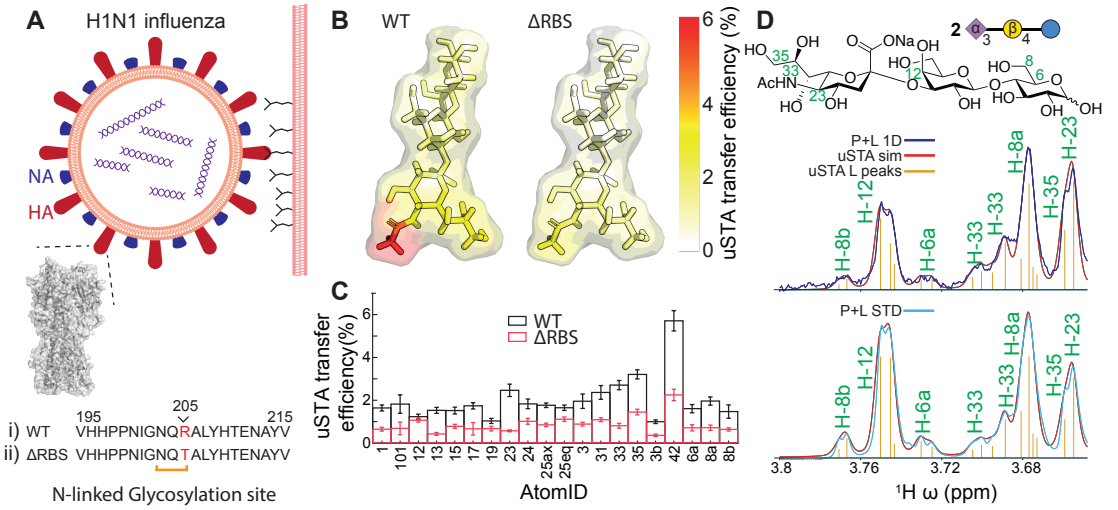


Figure 6.3: (A) HA presents on the surface of the viral membrane and has been shown to bind with sialic acid surface glycans to mediate host cell entry. A designed (38) ΔRBS variant is generated by the creation of an N-linked glycosylation site via the creation of needed sequon NQT from wild-type NQR by the R205T point mutation in HA adjacent to the sialic acid-binding site. In this way, disruption of sialic acid binding through designed "blocking" in HA-ΔRBS is intended to ablate the binding of HA wild-type to sialosides in this synthetic variant. (B and C) Notably, wild-type and DRBS variants of H1N1 HA in fact show a remarkably similar overall binding pattern (B) for the 2,3-sialo-trisaccharide 2 (focused through engagement with the sialoside) but a significant intensity moderation (C) for the DRBS variant, indicative of a partial (but not complete) loss of binding consistent with design [102]. (D) Raw spectral data demonstrate that atom-specific differences in intensity in the 1D versus the difference spectra can be discerned using uSTA. Note that atom numbering shown and used here is generated automatically by uSTA. (Caption and figure taken and mildly edited from [101]).

on SARS-CoV-2 proteins after expression in human cells [125]. We used uSTA to evaluate a panel of both natural and unnatural sialosides (Figure 6.4). The simplest monosaccharide, Neu5Ac equilibrates as a mixture of alpha and beta mutarotating anomers. uSTA revealed that only the minor, alpha form bound in an end-on manner, primarily by the acetamide group. The alpha state is only present at around 5% under these conditions, so this presented another example of the sensitivity of the uSTA algorithm (Figure 6.4 E, F). It also correlates with the near total natural occurrence of alpha states on cell-surfaces. We then applied the algorithm to more extended alpha-sialoside trisaccharides, including 2,3- and 2,6-linked sialyllactose. Both of these molecules displayed a striking similarity to the monosaccharide ligand, displaying the end-on sialic binding, similar to the HA

binding, but with slightly more extensive binding. The interaction of the glycerol side chain to the sialic acid were probed with unnatural ligand preparation, initially by replacing the OH-9 hydroxyl group with an azide N_3 group (Figure 6.4E ii, F iii+iv). This was well tolerated, with very similar binding maps. Again the beta anomer showed no binding. The introduction of a larger biphenyl (BPC) moiety at the same OH-9 position led to an abrupt shift in the binding mode that was dominated by the hydrophobic modification (Figure 6.4 G i+ii). Turning our attention to the kinetic and thermodynamic parameters of the 2,3-spike interaction, variable ratio protein:ligand samples were made, including 6μM spike with 2,3 concentrations of 60μM, 200μM, 1mM and 2mM ((Figure 6.5)). We extracted $K_D = 32 \pm 12\mu M$, $k_{on} = 6300 \pm 2300 M^{-1}s^{-1}$ and $k_{off} = 0.2 \pm 0.08 s^{-1}$. To confirm this we probed the binding using surface plasmon resonance (SPR) [126]. This generated a corresponding $K_D = 23.7 \pm 3.6\mu M$ with a $k_{on} = 1004 \pm 290 M^{-1}s^{-1}$.

Having established the 'end-on' binding mode of the natural sialosides, we were able to use the transfer efficiencies to score the output of the 'High ambiguity-driven docking' (HADDOCK) method and probe likely regions within the spike protein where the end-on mode was likely [115][116]. The modelling was performed by two collaborators, Alexandre Bonvin and Panagiotis I. Koukos. They homology modelled from the MERS structure [106] and then performed docking with the sialosides. Using the uSTA atomic restraints to score the output models, a glycan-binding pocket was identifed.

Previous work on SARS-CoV-2 spike glycan attachment identified binding on the RBD rather than NTD specifically for heparin sulfate[127]. Using tetrasaccharide heparin constructs, we were able to identify a substantially different binding mode to the sialosides, with an apparently non-localised extended 'side-on' binding mode (Figure 6.6 A, B). Next we turned our attention to four variants of concern which emerged later during the pandemic, alpha, beta, delta and omicron[128]. uSTA demonstrated that the sialoside binding had been completely ablated. This was in excellent agreement with corresponding mutations that are near to the identified binding pocket. Lastly, we probed the effects of an RBD-binding nanobody on the

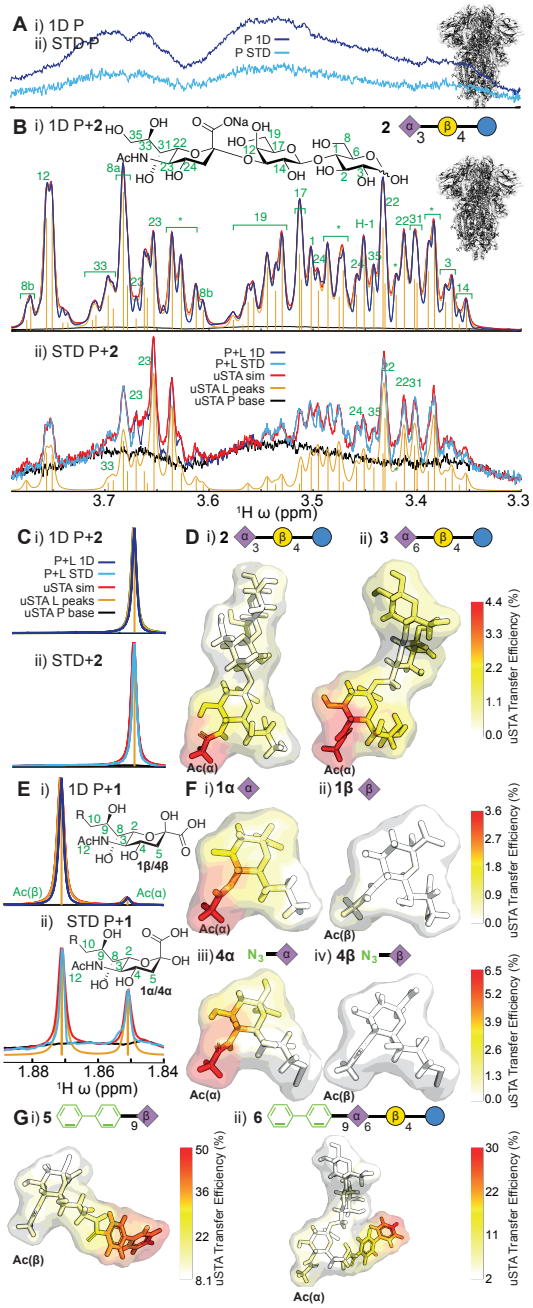


Figure 6.4: A panel of natural, unnatural, and hybrid variant sialoside sugars 1-6 (shown throughout the figure in glycan representation) was used to probe interaction between sialosides and spike. (A) The 1D 1H-NMR of SARS-CoV-2 spike protein shows considerable signal in the glycan-associated region despite protein size, indicative of mobile internal glycans in spike protein. This effectively masks traditional analyses, as without careful subtraction of the protein's contributions to the spectrum (shown later in FigureA.8), the ligand cannot be effectively studied. (B) Application of the uSTA workflow (shown later in Figure A.6) to SARS-CoV-2 spike protein (shown in detail for 2). The uSTA process of ligand peak assignment and deconvolution → P + L peak assignment and deconvolution → application to P + L STD yields precise atom-specific transfer efficiencies (Figure shown later in A.6).

Figure 6.4 ***(previous page)*****:** Note how in (ii) individual multiplet components, have been assigned (yellow); the back-calculated deconvolved spectrum (red) is an extremely close match for the raw data (purple). The spectrum is a complex superposition of the ligand spectrum and protein only yet uSTA again accurately deconvolves the spectrum, revealing the contribution of protein-only (black) and the ligand peaks (yellow). Using these data, uSTA analysis of the STD spectrum pinpoints ligand peaks and signal intensities. Spectral atom numbering shown and used here is generated automatically by uSTA; all other numbering in sugars follows carbo- hydrate nomenclature convention. (C and D) Application of the uSTA workflow (more detail of the workflow is shown in Figure A.6) reveals atom-specific binding modes to spike protein for both natural (e.g., sialoside) trisaccharides Siaα2,3Galb1,4Glc (2) and Siaα2,6Galb1,4Glc (3). Comparison of the uSTA method focused on the NHAc methyl resonance shows excellent agreement (C). The uSTA method allowed determination of binding surfaces for both trisaccharides 2 [D(i)] and 3 [D(ii)]. (E and F) STD spectrum (E) and mapped atom-specific transfer efficiencies (F) for sialic acid (1) and 9-N_3 azido variant 4. Both interconverting α and β anomeric forms could be readily identified. Despite the dominance of the b form [94%, E(i)], application of the uSTA method following assignment of resonances from the two forms allowed determination of binding surfaces simultaneously [E(ii), F(i, ii)]. Spike shows strong binding preference for the α anomers [F(i, iii) versus F(ii, iv)] despite this strong population difference. Binding surfaces were also highly similar to those of extended trisaccharides 2 and 3. (G) Using these intensities, atom-specific transfer efficiencies can be determined with high precision, shown here for hybrid sialoside 5. The details of both the unnatural BPC moiety and the natural sialic acid moiety can be mapped; although the unnatural aromatic BPC dominates interaction, uSTA nonetheless delineates the subtleties of the associated contributions from the natural sugar moiety in this ligand (see also figs. S5 and S6). (Caption and figure taken and mildly edited from [101]).

spike-sialic binding [129]. Samples were made up at saturating conditions where all the spike proteins will be bound by the nanobody. If the binding site occupied the same site on the RBD, we would expect the binding to be ablated. However, we observed identical binding maps and only mildly reduced transfer efficiencies.

The combination of these results suggested that end-on binding between spike and natural sialosides could be localised to the NTD, and does not require RBD ACE-2 binding.

6.7 Cryo-EM pinpoints the sialoside attachment site

Up to this point, our analyses had been hampered by the low-resolution structures available of the spike NTD region [130]. Undoubtedly more flexible and hetero-

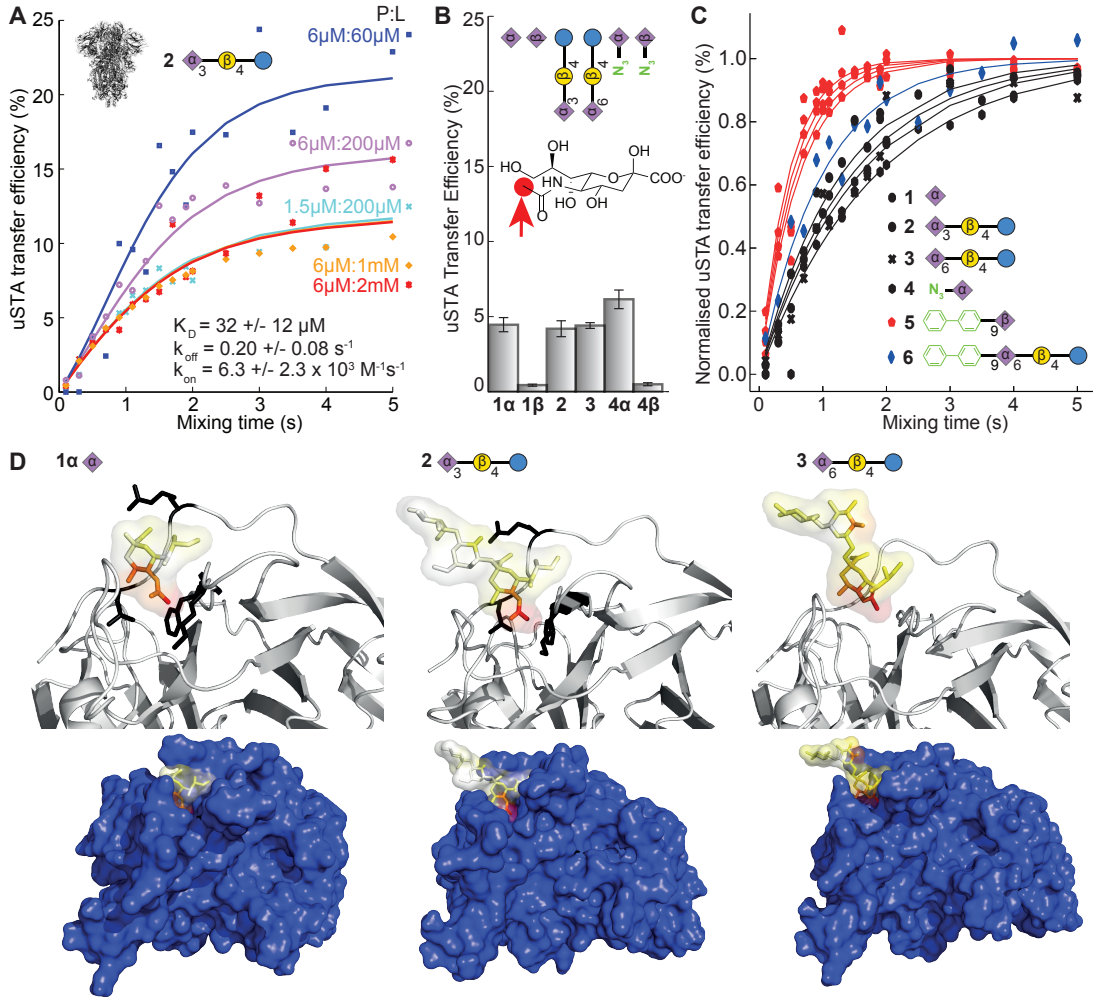


Figure 6.5: (A) Data (various points) show different Protein:Ligand concentrations (shown in the same color). The simulated fits (lines) of these data are produced by iteratively manipulating input parameters for a modified set of Bloch-McConnell equations that account for binding and cross-relaxation. As some of these input parameters are thermodynamic and kinetic (See appendix B for a more detailed explanation), we can find K_d, k_{off} and k_{on} parameters that explain the multiple concentration data. The values that we obtain for the SARS-CoV-2 • 2,3 interaction are indicated. Errors come from a bootstrapping procedure. The lower ligand concentrations yield data that are of lower sensitivity than higher concentrations. The transfer efficiencies will be higher in this case, as more molecules are effectively involved in the binding. Thus, data at lower concentrations will in general have more scatter and higher transfer efficiencies. These data points are desirable for the analysis, as it is here where we expected the greatest variation of transfer efficiency with ligand concentration. The analysis is applied globally and so the uncertainties in the final fitted parameters from the bootstrapping analysis (see Methods) provide a direct and confident measure of the goodness of fit. (B) Normalized uSTA transfer efficiencies of the NAc-5 methyl protons can be determined for each ligand studied here. This allowed relative contributions to "end on" binding to be assessed via uSTA in a "mode-specific" manner. This confirmed strong α- over β-sialoside selectivity.

Figure 6.5 ***(previous page)*****:** Errors were determined through a bootstrapping procedure where mixing times were sampled with replacement, allowing for the construction of histograms of values in the various parameters that robustly reflect their fitting errors. (Caption and figure taken and mildly edited from [101]). (C) Normalized build- up curves for the most intense resonances allowed two distinct modes of binding to be identified in natural (1-4) and hybrid (5, 6) sugars. Data are shown at constant protein and ligand concentrations. With the BPC moiety present, the build-up of magnetisation occurs significantly faster than when not; various such hybrid ligands give highly similar curves. By contrast, natural ligands have a much slower build-up of magnetisation. This, together with the absolute transfer efficiencies being very different, and the overall pattern on the interaction map combine to reveal that the ligands are most likely binding via two different modes and possibly locations on the protein. (D) Coupling uSTA with an integrative modelling approach such as HADDOCK [116] [115] allowed generation and, by quantitative scoring against the experimental uSTA data, selection of models that provide atomistic insights into the binding of sugars to the SARS-CoV-2 spike protein, as shown here by superposition of uSTA binding "map" onto modelled poses. uSTA mapping the interaction between SARS-CoV-2 spike [based on RCSB 7c2l [112]] with ligands 1a, 2, and 3 identifies the NHAc-5 methyl group of the tip sialic acid residue making the strongest interaction with the protein. By filtering HADDOCK models against this information, we obtain structural models that describe the interaction between ligand and protein (fig. S14). Most strikingly, we see the same pattern of interactions between protein and sialic acid moiety in each case, where the NAc methyl pocket is described by a pocket in the spike NTD. Although sequence and structural homology are low (shown later in Figure A.1), MERS spike protein possesses a corresponding NHAc-binding pocket characterized by an aromatic (Phe39)-hydrogen-bonding (Asp36)-hydrophobic (Ile132) triad [106]. (Caption and figure taken and mildly edited from [101]).

geneous, the lack of resolution made crystallographic work with the protein very difficult. Our collaborators in the Naismith Group worked for 18 months to get a structure of spike that had sufficient resolution in the NTD. Eventually, using a specific spike construct that is bound by linoleic acid[131], the NTD resolution improved to 2.3 Å(The density associated with the linoleic acid is shown in Figure A.15). Two samples were made, one soaked in a sialoside containing ligand, the other not. The NTD region had a region of density that matched very well to a sialic acid group, including the glycerol and NAc moieties (Figure A.14). Indeed, the sialic acid group pose agreed strikingly with the uSTA heatmaps (Figure 6.7 E). Overlaying the structure with the MERS pocket that we had used to inform homology modelling demonstrated that the sugar binding pockets are 12 Åaway and do not appear to be similar (Figure 6.7 D). The precise ligand used for the

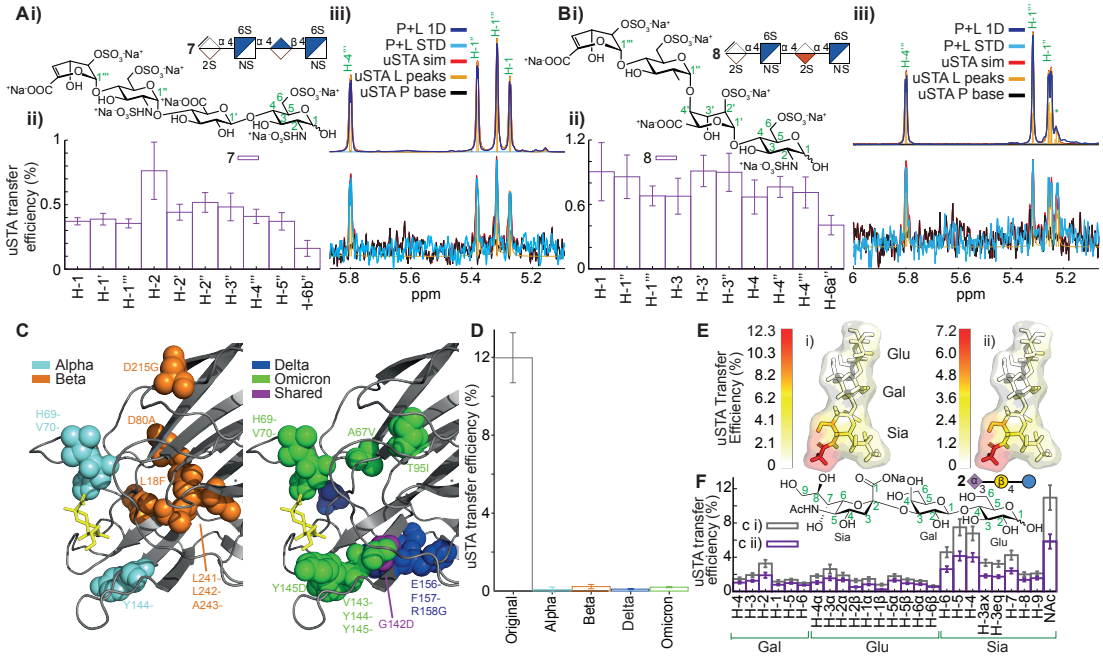


Figure 6.6: (A and B) Two heparin tetrasaccharides (7, 8) are shown by uSTA [A(iii), B(iii)] to bind B-origin-lineage SARS-CoV-2 spike protein ("original" spike) in a "side-on" mode [A(ii), B(ii)]. Atom specific binding is shown in A(ii) and B(ii). Assignments shown in green use conventional glycan numbering. (C) Substantial numbers of mutations arise in the NTD region identified by uSTA in the B.1.1.7/alpha (cyan), B.1.351/beta (orange), B.1.617.2/delta (dark blue) and B.1.1.529/omicron (green) lineage variants of SARS-CoV-2 spike. (D) Ablated binding of the a-2,3-sialo-trisaccharide 2, as measured by the transfer efficiency of the NHAc protons, is identified by uSTA in the lineage variants of SARS-CoV-2 spike [see (C) for colors]. This proves consistent with mutations that appear in the sialoside-binding site of NTD identified in this study [see (C) and Fig. 6]. (E and F) uSTA of sialoside 2 with B-origin-lineage SARS-CoV-2 spike in the presence of the potent RBD-neutralizing nanobody C5 [spike E(i), nanobody-plus-spike E(ii)] shows essentially similar binding patterns with uniformly modulated atomic transfer efficiencies (F). (Caption and figure taken and mildly edited from [101]).

EM work had a hydrophobic aromatic moiety attached distal to the sialic acid (Figure A.16). The aromatic ring was modified with three iodo groups to produce an iodobenzamide moiety. This was intended to increase the electron density and make it easier to see in the EM density maps. Interestingly, there was additional density in a distal site that could be well explained by the tri-iodo aromatic moiety (Figure A.16). This second, hydrophobic binding site might explain the change in uSTA binding pose from introducing the biphenyl moiety in ligand 2.

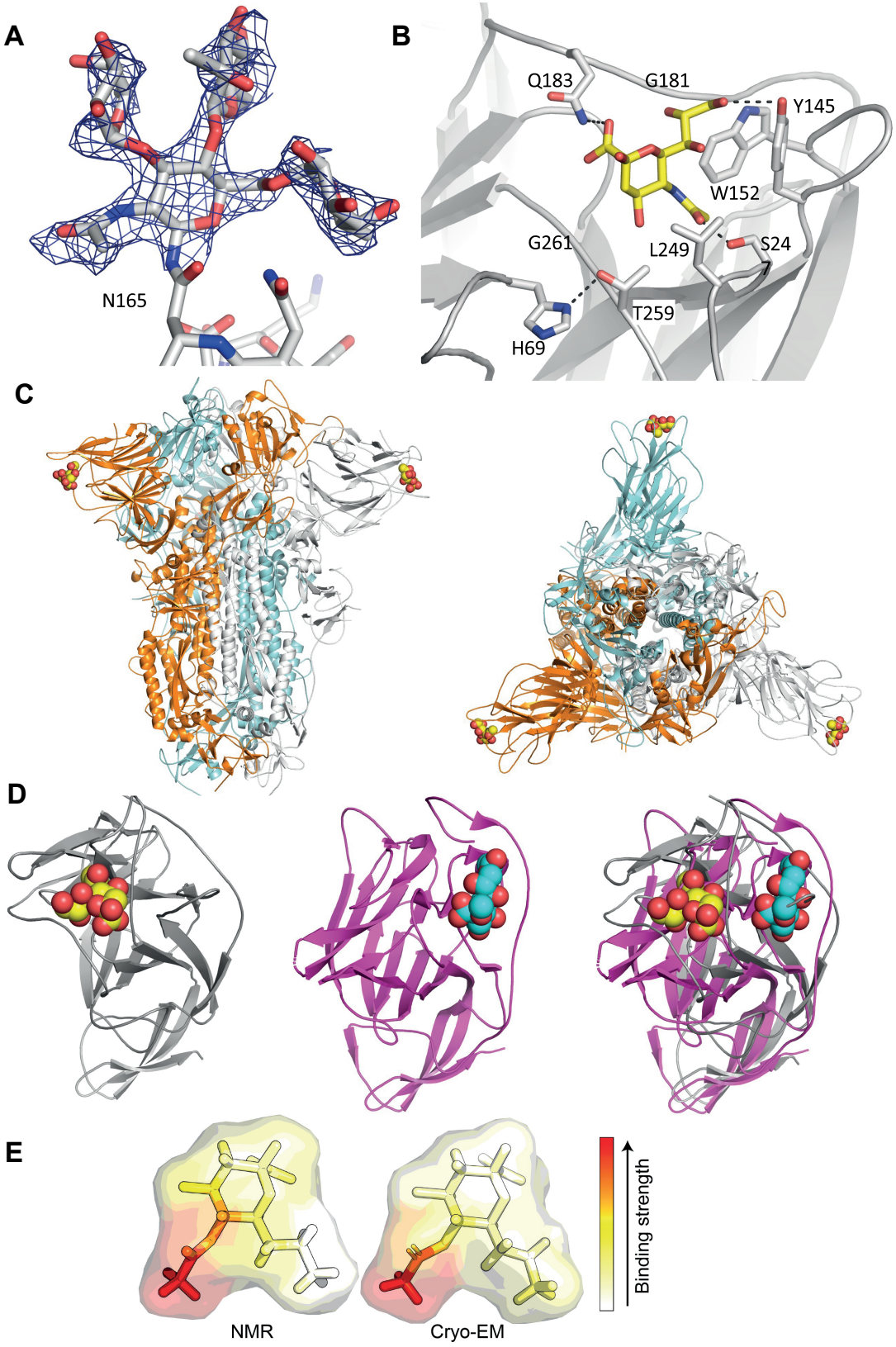


Figure 6.7: (A) In the presence of a modified sialoside (shown in more detail in Figure A.16), a well-ordered structure with resolution (2.3 Å) was obtained, sufficient to identify even usually mobile N-linked glycans within spike.

Figure 6.7 ***(previous page)*****:** (B) The sialoside-binding site (with Coulombic maps shown in Figure A.14) in the NTD is bounded by His69, Tyr145, Trp152, Ser247, and Gln183. Gln183 appears to engage COOH-1, which perhaps rationalises the stereoselective recognition of α-sialosides. Tyr145 engages the C7-C9 glycerol side chain of sialoside. Ser247 engages the NHAc-5. Notably, Tyr145 and His69 are deleted in the α variant of SARS-CoV-2 which loses its ability to bind sialoside (see Figure 6.6). (C) Spike's sialoside binding site is found in a distinct region of the NTD (left, from side, right from above) that is at the "edge" of the spike (see text). (D) Superposition (right) of SARS-CoV-2-NTD (left, grey) with MERS-NTD (middle, magenta) shows distinct sites that are 12 Å apart. (E) A comparison of the ligand-binding mode measured by uSTA NMR and a map calculated from the cryo-EM structure. For each hydrogen environment in the resolved sialic acid, an array of distances to all protein hydrogens, r, was calculated. The interaction of each ligand environment with the spike protein was defined as the summation of all 1/r6 environment-protein distances. Values were interpolated in a 1/r6 manner onto heteroatoms following the same procedure according to the described NMR methods, and the colour bar was scaled from zero to the maximum value. (Caption and figure taken and mildly edited from [101], with particular thanks to the Naismith group who made this figure.).

6.8 Discussion and conclusions

There are few methods to dissect glycan-protein interactions besides the useful surface display 'glycoarrays'. The uSTA workflow has been built from more simple systems up to the very relevant spike protein, demonstrating its capability to provide poses and binding parameters even in complex and delicate systems. The application to H1N1 HA demonstrated the workflow's ability to dissect protein-ligand interaction even when there was a designed 'blocking glycosylation'. The spike trimer of the B-origin lineage SARS-CoV-2 presented the most complicated system due to the heavy glycosylation of similar type to the ligands we were trying to dissect. The pose of the sialosides was unambigiously unpicked as 'end-on', similar to HA, and this was confirmed by a very high resolution Cryo-EM structure obtained by our collaborators. Our original structural work was slightly misled by earlier low-resolution spike structures that did not resolve the heterogenous NTD. This EM also confirmed a second, distal hydrophobic binding mode indentified by uSTA, although the physiological relevance was not the focus of this study. It is worth noting that we cannot discount a second sialoside binding site, but there is no structural evidence for it. The identified binding site is coincident for

numerous mutations and deletions in the variants-of-concern. Alpha, delta and omicron (His^{79}, Val^{70} and Tyr^{145}) and beta (Leu^{241}-Leu^{242}-Ala^{243}-Leu^{244}) changes remove key interacting resides or perturb structural motifs that form the pocket. The binding mode that we have identified ('end-on') is structurally distinct from the heparin binding mode that was identified by uSTA and has been reported from computation in the literature. The ablation of binding in the variants of concern may suggest a dynamic role for sugars within viral evolution.

The second question raised by this work is why the binding is ablated in later variants. Human endothelial cells are covered by the glycocalyx, which contains glycoproteins that are littered with sialosides. Functionally, therefore, this binding will act to make the SARS-CoV-2 viral particle 'stick' to the airways. The role of this stickiness and the ablation need more study. It is worth noting that several neutralising antibodies that bind the NTD region are no longer able to bind to the variants of concern, which perhaps indicates an evolutionary driving force to structurally adapt this region. The cost-benefit function of sialoside binding is therefore not straightforward, and we hope that uSTA presents a new technique to probe the interaction at high resolution *in vitro*, perhaps enabling a broader study of the viral protein • sialoside interaction landscape. Once the uSTA protocol has been run for a particular ligand (for example 2,3-sialyllactose), it is much easier to repeat the process with a new protein. This could enable a broader study of the interaction between 2,3-sialyllactose and other coronavirus spike proteins (MERS-CoV, SARS-CoV and Human Coronaviruses would all be good targets). This may help answer the intriguing question of whether there is any functional importance to the interaction.

While the uSTA method has here been focused on a relatively specific set of ligands and proteins, this approach can be taken with most ^{1}H containing ligands (notably excluding paramagnetic ligands), including many drugs. We are hopeful that these developments complement drug-design techniques such as fragment-based design. Indeed, recently there has been some attention on the difficulty of computationally predicting viable drugs based on structural information

alone. The inherently dynamic information present in the solution-state NMR measurement, which enables simultaneous ensemble binding information and kinetic and thermodynamic parameter extraction, may offer a complementary approach to build on the excellent structural biology work.

More generally, STD NMR requires no isotopic labels and the uSTA analysis pipeline now utilises UnidecNMR to enable higher fidelity and resolution binding information alongside a more accurate Bloch-McConnell treatment of kinetic and thermodynamic parameters. We are hopeful that this technique will bring more biochemists and glycobiologists to NMR.

6.9 Future Work

We would like to model the importance of atom-specific T1 and T2 relaxation in the free ligand. Once the ligand has unbound, T1 relaxation would cause the saturation to be lost, meaning that if the T1 is shorter on one atom of the molecule, that would decrease the uSTA transfer efficiency without corresponding to lower binding. We hope this will lead to more accurate heat maps. It should be noted that the samples were prepared in D2O solvent to facilitate water suppression during the NMR experiments. This will affect the relaxation times by lowering the proton density.

6.10 Materials and Methods

The protein expression and purification was not my work, but is detailed in the Materials and methods section of the manuscript [24]. NMR experiments were recorded on 3 magnets: 600 MHz Bruker AVANCE NEO equipped with QCIF cryoprobe and sample changer, 600 MHz Bruker AVANCE III equipped with TCI Cryoprobe and 950Mhz Bruker Avance III equipped with TCI Cryoprobe and sample changer. All STD experiments were recorded at 15 degrees and 16 saturation times. All samples were made up in Phosphate Buffered Saline (PBS) at pH 7.4.

7

BAR-1, a sugar binding antibody

Contents

7.1 Acknowledgements 129
7.2 Introduction 130
7.3 Glycan binding antibody design 131
7.4 uSTA application 132
7.4.1 Comparison with X-Ray data 132
7.4.2 Kd calculations 134
7.4.3 Mutational study 134
7.5 Conclusions 135

7.1 Acknowledgements

This Chapter is an ongoing collaboration alongside the Davis, Sattentau and Naismith groups. The majority of the immunology work was done by Lachlan Deimel from the former two groups. The majority of the synthetic sugar work was done by Aziz Khan. My work was to characterise the interaction of the monoclonal antibody BAR-1 and various sugars.

7.2 Introduction

Antibodies are one of the mammalian lines of defence against pathogenic threats. They are a solubilised form of a membrane protein, called a B cell receptor (BCR), which are found on the surface of B-cells. A particular host has a base repertoire of B-cell germlines (which each produce a distinct BCR) that can undergo maturation through a process called Somatic Hypermutation (SHM)[132]. Particular regions of the BCR genetic code, stored within the B-cells, are replicated during SH with a mutation rate many orders of magnitude higher than normal host cells [133]. Through this process, the BCRs have access to a vast number of potential structures [133]. When B-cells are presented with an antigen, this iterative process begins, and B-cells with tighter antigen binding BCRs are upregulated. These BCRs can also be expressed as the solubilised, Immunoglobulin (Ig) form (antibodies)[134]. This iterative process leads to tight antibody-antigen binding that ideally neutralises the pathogen.

As an example, numerous antibodies that bind and neutralise the SARS-COV-2 spike protein from Chapter 6 have been well characterised. Many have evolved to bind to the spike receptor binding domain (RBD), preventing SARS-CoV-2 from then binding to the host receptor (ACE-2) [135], while others neutralise by binding to the N-terminal Domain (NTD)[136]. The SARS-CoV-2 spike, however, exhibits a heavy glycosylation pattern[137], exemplifying an immune-evasion strategy employed by viral surface proteins called a 'glycan shield'[138]. As these glycans decorate both host-cell and pathogen machinery, they are inherently not immunogenic.

Another classic pathogenic protein that utilises this shield is the human immunodeficiency virus (HIV) envelope protein GP120 [138]. A small subset of people produce broadly neutralising antibodies which interact with either the glycans or both glycans and proteins. Sequencing the antibodies demonstrates that they underwent substantial somatic hypermutation [139] [140]. This means the B-cells producing these antibodies have deviated significantly from the base germline repertoire. As somatic hypermutation is a pseudo-random process, neutralising antibodies produced by B-cells that are close to the base repertoire are statistically

more accessible to the host immune system. This includes SHM-dependent HIV neutralising antibodies. Over half of the GP120 glycoprotein mass is glycans [138], shielding the underlying proteinaceous structure. It has been demonstrated neutralising antibodies must incorporate these glycan chains spatially, which requires extensive SHM [139] [140]. The base germline produced antibodies do not bind to the glycoprotein [141]. The requirement for extensive SHM makes antibody production less replicable and more difficult to ellicit in potential vaccine design. An important step is to create a construct with a lower reliance on this pseudo-random process and bring the immune response closer to the germline repertoire, requiring a novel approach.

7.3 Glycan binding antibody design

There have been major advancements in understanding protein-antibody binding[142]. Antibodies that bind to the glycan chains directly, however, are known, but difficult to structurally characterise[143]. The Sattentau and Davis groups decided to investigate this by creating unnatural linkers to encourage BCR binding to the glycans themselves. Natively, glycosylations are incorporated through either N- (Asp) or O- (Ser/Thr) linkages. They hypothesised that creating an unnatural linker would enable B-cells to bind to the glycan chains. They coupled a relevant viral glycan, 2,3-sialyllactose, to hen egg lysozyme through an unnatural amidine linker of the lysine ϵ amines. The aim was to distinguish the amidine-linked glycan-chain from the inherently immunorecessive host-like N-linked glycans.

This modified glycoprotein (HEL-amidine-SiaLac) was then injected into mice with an adjuvant to raise an immune response. While a native SiaLac presentation without the amidine linker failed to elicit a substantial response, a significant IgG response to HEL-amidine-SiaLac was detected. There was no IgG response upon injection of HEL alone, indicative of glycan-specific binding. In order to unpick this response, the corresponding B cells were isolated and sequenced, enabling monoclonal antibodies to be expressed. A representative example, BAR-1, was chosen with the aim of structurally characterising the antibody-carboydrate interaction.

7.4 uSTA application

Having recently developed a tool for unpicking carboydrate-protein interactions[101], samples were created with BAR-1 and the glycan chain without HEL, i.e. lysine-amidine linked 2,3-SiaLac. An initial experiment demonstrated a clear STD response, indicating binding. The system (including protein, ligand and mixture samples) was run through the uSTA pipeline to produce a heatmap (Figure 7.1 Ai). This indicated that BAR-1 was binding the glycan in a different way to our previous, SARS-CoV-2 2,3-SiaLac work (Chapter 6). Rather than 'end-on' binding, there was appreciable binding all along the trisaccharide, while the lysine moiety was less engaged. This suggested the protein was successfully binding the sugar moiety and was not specifically binding the protein backbone. To confirm the lysine residue was not necessary for binding, we performed the same uSTA experiment on the 2,3 SiaLac without amidine linker or lysine and found a very similar heatmap - again the 2,3 was engaged along the entirety of the trisaaccharide (Figure 7.1 Aii).

7.4.1 Comparison with X-Ray data

The Naismith group crystallised the fragment antigen-binding region (Fab) of BAR-1 in complex with the SiaLac-amidine-lysine, producing a structure which could be compared with the uSTA results. This confirmed our prediction of side-on binding and agreed with the uSTA work. The lysine residue makes contacts with a groove formed by Y97 and T98 of the BAR-1 light chain that were important in the binding (Figure 7.2). As a solution state technique, the uSTA method probes the binding in an ensemble through cross-relaxation. In order for cross-relaxation to occur between two protons, they must be stationary with respect to each other for a duration. Our results are consistent with the trisaccharide protons being stationary with respect to BAR-1 for a duration, indicating binding. The indication from the heatmaps is that the lysine moiety is less engaged to the protein. There is a further control to do here, as the relatively increased proton density in the lysine moiety (geminal hydrogens at multiple side chain carbons) will increase the R1 relaxation rates of the molecule, which would artificially lower the uSTA response.

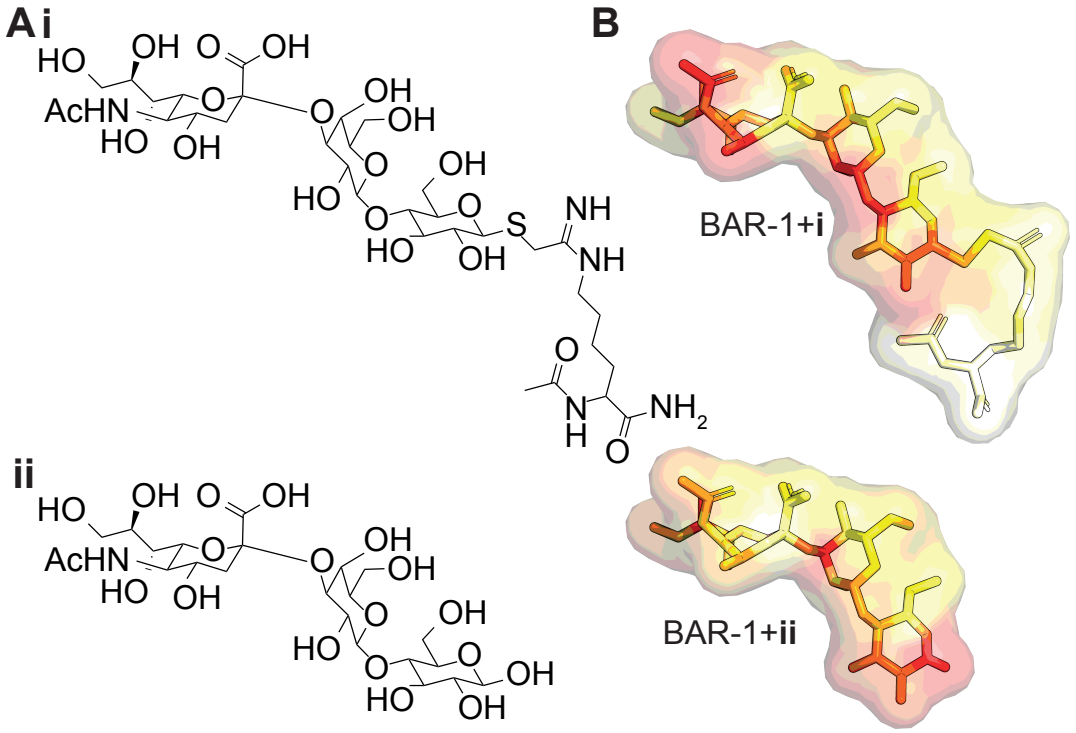


Figure 7.1: The chemical structures (A) and corresponding heatmaps (B) of amidine-linked-2,3-sialylactose (i) and 2,3-sialylactose are shown. The binding map for the lysine modified variant demonstrates the relative diengagement of the lysine moeity. On the shared trisaccharide, there are similarities between the two binding maps, showing engagement along the trisaccharide for both. More precisely, there are more subtle differences in where they light up, perhaps indicating a structural rearrangement of either pocket or ligand pose when the lysine and linker moeities do not have to be accommodated.

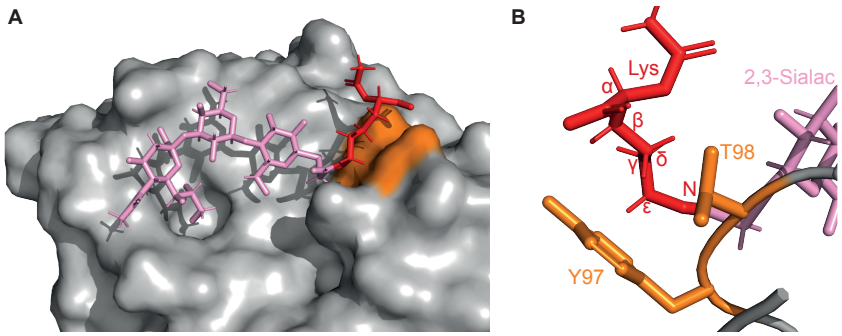


Figure 7.2: The fragment antigen-binding region (FAB) of BAR-1 was crystallised by the Naismith group in the presence of amidine-lysine linked 2,3. This shows the

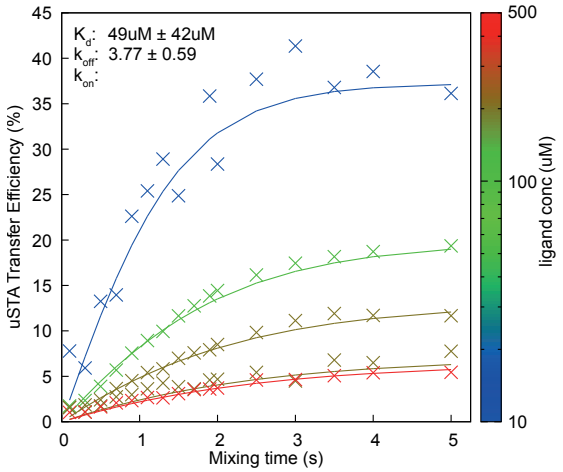


Figure 7.3: Multiple protein:ligand concentrations (8uM:5uM, 100uM:5uM, 200uM:5uM, 500uM:5uM and 200uM:2.5uM) of BAR-1 and amidine-lysine linked sialyllactose were studied at 298K, pH 7.4 in Phosphate Buffered Saline. The build up curves were experimentally determined (crosses) and the build up curves simulated using the modified Bloch-McConnell formalism discussed in Appendix B. The extracted kinetic and thermodynamic parameters are shown.

7.4.2 Kd calculations

We wanted to determine the kinetic and thermodynamic parameters of the BAR-1 glycan interaction. A number of glycan:antibody ratios were therefore produced, including 8uM:5uM, 100uM:5uM, 200uM:5uM, 500uM:5uM and 200uM:2.5uM. The build-up curves and transfer efficiencies of these were fitted using the Bloch-McConnell equation fitter (see appendix B), with excellent results. Error calculations were also performed (not shown), and our results were in a similar regime to SPR measurements ($17 \pm 2\mu M$). The fits are demonstrated in Figure 7.3 and the kinetic/thermodynamic values are $K_D = 49 \pm 42\mu M$, $k_{off} = 3.77 \pm 0.59$.

7.4.3 Mutational study

In order to probe the binding further, Lachlan produced mutants of BAR-1. Using the crystal structure, residues close to the binding site were chosen. While this included some on the light chain, the majority were on the heavy chain, which appeared to be interacting most with the trisaccharide. W57A, V59A, A58G, F37A, Y99A and Y103A were all chosen to be studied by uSTA. An immediate discovery was that the W57A mutant completely ablated binding. The face of the aromatic

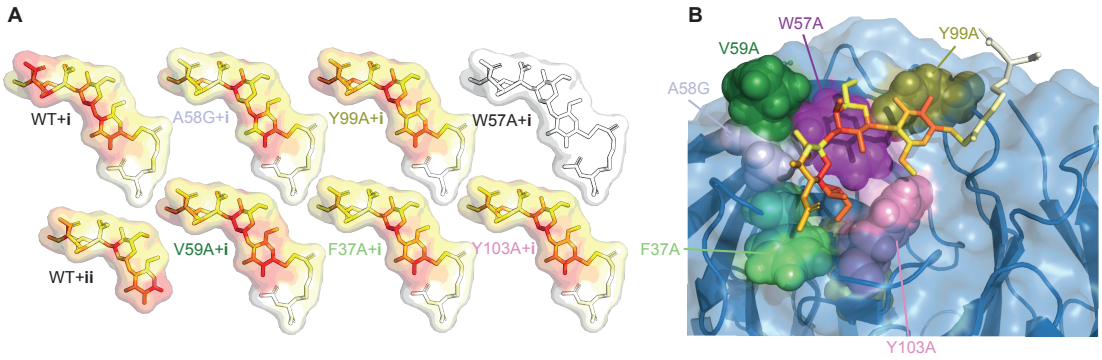


Figure 7.4: The heatmaps (A) and mutation sites from the WT crystal structure (B) are shown, demonstrating broadly similar binding features for all the lysine-amidine-SiaLac excluding W57A.

ring of W57 in the crystal structure makes contact with the second saccharide, galactose. Most intriguingly, by this point in the research, Lachlan had sequenced thousands of HEL-SiaLac binding B-cells, and all the corresponding antibodies contained this tryptophan residue. The correlation between uSTA mutational work and the preservation of W57A confirmed its importance. Further, the rest of the mutants showed remarkably similar poses to the wild type BAR-1 all along the trisaccharide. Overall, these results suggest the interaction between the galactose and W57 pins the middle of the trisaccharide. This prevents much motion in the adjacent, rotationally restricted saccharides. Mutations in the regions surrounding the sialic and glucose groups could further increase affinity, but structurally the pose appears to be governed by the pinning tryptophan.

7.5 Conclusions

Unpicking the structural basis of glycan-binding antibodies helps to illuminate the intriguing role host-like glycans can play within immunology. The identification of the critical role that W57 plays in centrally pinning the glycan unpicks the structural basis of this interaction, and its conservation might suggest a broader role in IgG-carbohydrate binding repetoire. The initial result from X-Ray cystallography suggested the importance of the unnatural amidine-lysine linker in the antibody-glycan interaction. The ability of uSTA to probe the ensemble at atomic resolution

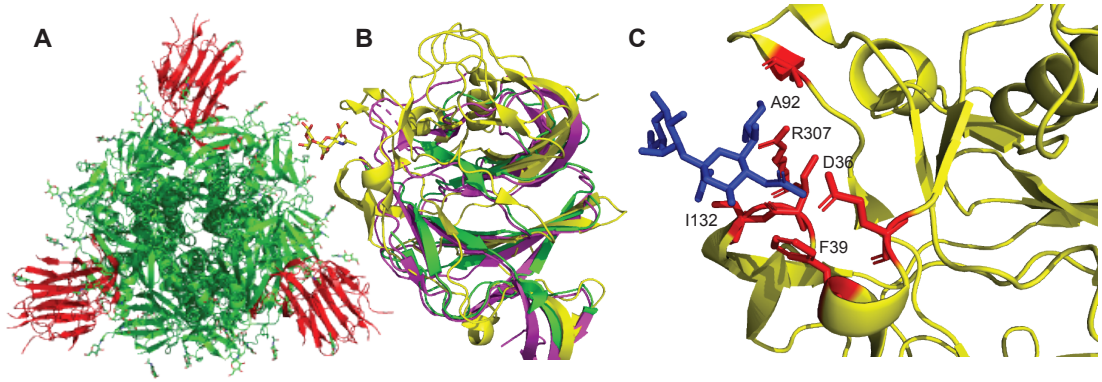
suggests a glycan-centric, as opposed to linker-centric binding. The similarity of the 2,3-SiaLac interaction confirmed that BAR-1 binding does not rely on an unnatural linker. To confirm this, extracting Kd information for 2,3 sialyllactose would be a good next step. If the reliance on 2,3 sialyllactose is minimal, this would greatly increase the generality of using an unnatural linker to raise a glycan-binding antibody response. Combined with the substantially lower requirement for somatic hypermutation, our results indicate this approach could be used to design glycan specific vaccines and potentially neutralise viral glycan-shield evasion. Experimental determination of T1 and T2 relaxation and Bloch-McConnell simulation is an important control that will help underpin this work.

Appendices

A

uSTA Supplementary figures

This appendix contains the supplementary figures and corresponding captions from the uSTA manuscript [24]. All of these figures were made by me and edited by Prof. Andrew Badwin and Prof. Ben Davis with the exception of the homology modelling Figure A.1 and the cryo-EM Figures A.14, A.15, A.16, which were respectively made by our collaboraters led by Alexandre Bonvin, and the Naismith group. Both are included to support statements in the main text.



D

```
K0BRG7 K0BRG7_MERS    1  MIHSVFLLMFLLTPTESYVDVGPDSVKSACIEVDIQQTFFDKTWPRP-IDVSKADGIIYP    59
P59594 SPIKE_CVHSA    1  ----MFIFLLFLTLT-S----------GSDLD--RCTTFDDVQAPNYTQHTSSMRGVYYP    43
P0DTC2 SPIKE_SARS2    1  ----MFVFLVLLPLV-S----------SQCVN--LTT--RTQLPPAY--TNSFTRGVYYP    39
                             :*::::.:*  . *          .  ::           *      *   *: **

K0BRG7 K0BRG7_MERS   60  QGRTYSNITITYQGLF-PYQGDHGDMYVYSAGHATGTTPQKLFVANYSQDVKQFANGFVV   118
P59594 SPIKE_CVHSA   44  DEIFRSDTLYLTQDLFLPFYSNVT---GFHTIN-------HT-----FGNPVIPFKDGIYF    89
P0DTC2 SPIKE_SARS2   40  DKVFRSSVLHSTQDLFLPFFSNVT---WFHAIHVSGTNGTKR----FDNPVLPFNDGVYF    92
                          :    *.     *.** *: .:     : : :      :      :.: *  * :*. .

K0BRG7 K0BRG7_MERS  119  RIGAAANSTGTVIISPSTSATIRKIYPAFMLGSSVGNFSDGKMGRFFNHTLVLLPDGCGT   178
P59594 SPIKE_CVHSA   90  AATE-----------------KSNVVRGWVFGSTMNNKSQ---------SVIIINNSTNV   123
P0DTC2 SPIKE_SARS2   93  ASTE-----------------KSNIIRGWIFGTTLDSKTQ---------SLLIVNNATNV   126
                                               ::  .:::*:::.. ::         ::::: :. ..

K0BRG7 K0BRG7_MERS  179  LLRAF--YCILEPRSGNHCPAGNSYTSFATYHTPATDCSDGNYNRNASLNSFKEYFNLRN   236
P59594 SPIKE_CVHSA  124  VIRACNFELCDNPFFAVSKPMGT-------------------------QTHTMIFDNAFN   158
P0DTC2 SPIKE_SARS2  127  VIKVCEFQFCNDPFLGVYYHKNN---------------------KSWMESEFRVYSSANN   165
                         ::::.       :*  .     ..                          .    : .  *

K0BRG7 K0BRG7_MERS  237  CTFMYTYNITEDEILEWFGITQTAQG-VHLFSSRYVDLYGGN---------------MFQ   280
P59594 SPIKE_CVHSA  159  CTFEYISDAFSLDVSEKSGNFKHLREFVFKNKDGFLYVYKGYQPIDVVRDLPSGFNTLKP   218
P0DTC2 SPIKE_SARS2  166  CTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEP   225
                         *** *  :    ::    *  :  :  *.   . :. :*                 :

K0BRG7 K0BRG7_MERS  281  FATLPVYDTIKYYSIIPHSIR---SIQSDRKAW----AAFYVYKLQPLTFLLDFSVDGYI   333
P59594 SPIKE_CVHSA  219  IFKLPLGINITNFRAILTAFS------PAQDIWGTSAAAYFVGYLKPTTFMLKYDENGTI   272
P0DTC2 SPIKE_SARS2  226  LVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTI   285
                         :  **:  .*. :  :             . *    **::*  *:* **:*.:. :* *
```

Figure A.1: A) The N-terminal domain of the SARS-CoV-2 spike protein (green, 6vxx). The N-terminal domain comprises the first 280 residues (red). B) The structural homology of MERS (yellow, 6q06), sars-cov1 (magenta, 5x58) and sars-cov2 (green, 6vxx) spike proteins in this region is partial. The C-terminal domains show substantial similarities, but the N-terminus show substantial differences. C) The interaction between Gal-Sia and MERS NTD. The interaction can be largely attributed to 5 amino acid residues with the NAc binding F39, I132 and D36, and the glycerol arm binding to A92 and R307. D) Sequence homology of the N- terminal domains of MERS, SARS-CoV-1 and SARS-CoV-2 using BLASTP. Substantial differences between the three proteins are observed. The five key residues for sialic acid binding in MERS are indicated. These all fall within regions in SARS-CoV-1 and SARS-CoV-2 that differ substantially from MERS, thus rendering the likelihood of generating a reliable homology model exceedingly unlikely.

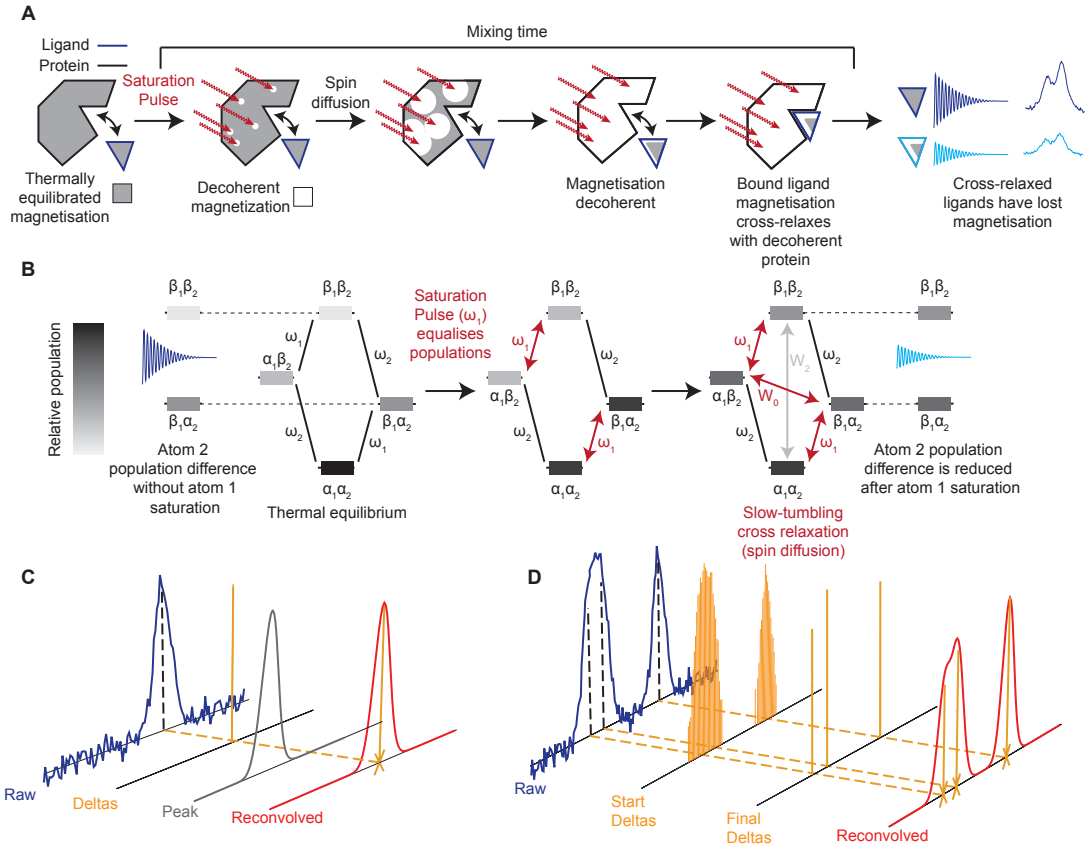


Figure A.2: A): A 'saturation pulse' is applied to a protein and ligand mixture that covers a specific but narrow range of ppm frequencies selected to avoid excitation of the ligand. The nuclei excited in the protein will have specific spatial locations within the protein, for example methyl groups or aromatics. Spins in the protein that have been excited will cross-relax with their neighbours, aiming to restoring the initially excited spins to equilibrium. This process continues during the period in which the excitation pulse is applied and the net magnetisation on each spin within the protein is gradually reduced to zero, a process termed 'spin diffusion'. The rate of the many interior cross relaxations increases as tumbling time of the protein decreases. When a ligand binds rigidly to the protein, intermolecular cross-relaxation can occur if adjacent spins in the protein have been knocked out of equilibrium by either the initial excitation, or spin diffusion. The action will be to transfer unperturbed longitudinal magnetisation from the ligand to partially restore the now unequilibrated longitudinal magnetization on the various spins of the protein. After the mixing time, the quantity of magnetization on the ligand is detected, following a 90^o pulse. The action of the saturation pulse therefore is to reduce the magnetisation on the ligand in a manner that depends on the on and off rates, the various tumbling times of ligand and protein that govern intrinsic (R1/R2) and cross (intra/inter) relaxation and the details of the saturation pulse. A 'build-up curve' follows the amount of lost signal from the ligand versus the duration of the mixing time. These processes can be effectively modelled using modified Bloch-McConnell equations and by recording 'build-up curves' at a range of varying ligand and protein concentrations, and fitting them to an appropriate model based on the modified Bloch-McConnell equations (see methods) concentration independent

Figure A.2 ***(previous page)*:** effects such as relaxation rates, and concentration dependent effects, such as binding, can be separated, as is demonstrated in this article. B): The details of the experiment can be understood with a series of single spin energy level illustration. Initially, protein and ligand spins are thermally equilibrated. The application of a saturation pulse at $\omega 1$ (always chosen to be a protein spin) induces coherence between the a and b spin states, and with sustained application will eventually equalises populations in the two $\alpha 1$ to $\beta 1$ transitions (the transition becomes 'saturated', net magnetisation for the spin in the ensemble has gone to zero). In our slowly-tumbling protein, the W0 cross-relaxation pathway is most efficient, providing a path to move magnetization from the state closes to equilibrium (ligand) to the state furthest from equilibrium (protein), here pushing population from the $\beta 1\alpha 2$ state to the $\alpha 1\beta 2$. The combination of continuous application of saturation pulse and cross-relaxation (both intra to ligand and inter within the protein) lowers the population differences between the interacting spin states. This is the nuclear Overhauser effect (NOE). The rate for cross relaxation, both inter and intra depends on the tumbling time of the complex, and the distance between the spins. If atom 2 is a nearby protein spin, we are interrogating part of intra-protein cross relaxation or spin diffusion, if atom 2 is a bound ligand spin, we are interrogating intermolecular cross-relaxation. These complex processes can be well accounted for in a simple two spin model as described in the methods. Our analysis is validated in this work by demonstrating that fitted KDs using our picture returns values that are equal to those obtained by independent means, together with the on and off rates. C),D): Schematic of process for automatically quantifying the signal intensities in NMR spectra using deconvolution. The analysis determined the number of peaks that can give rise to the signal and return a simulated spectrum by convolving these with a peak shape function. Precise peak positions and their intensities are returned, which can be attributed to single protons observed in a spectrum.

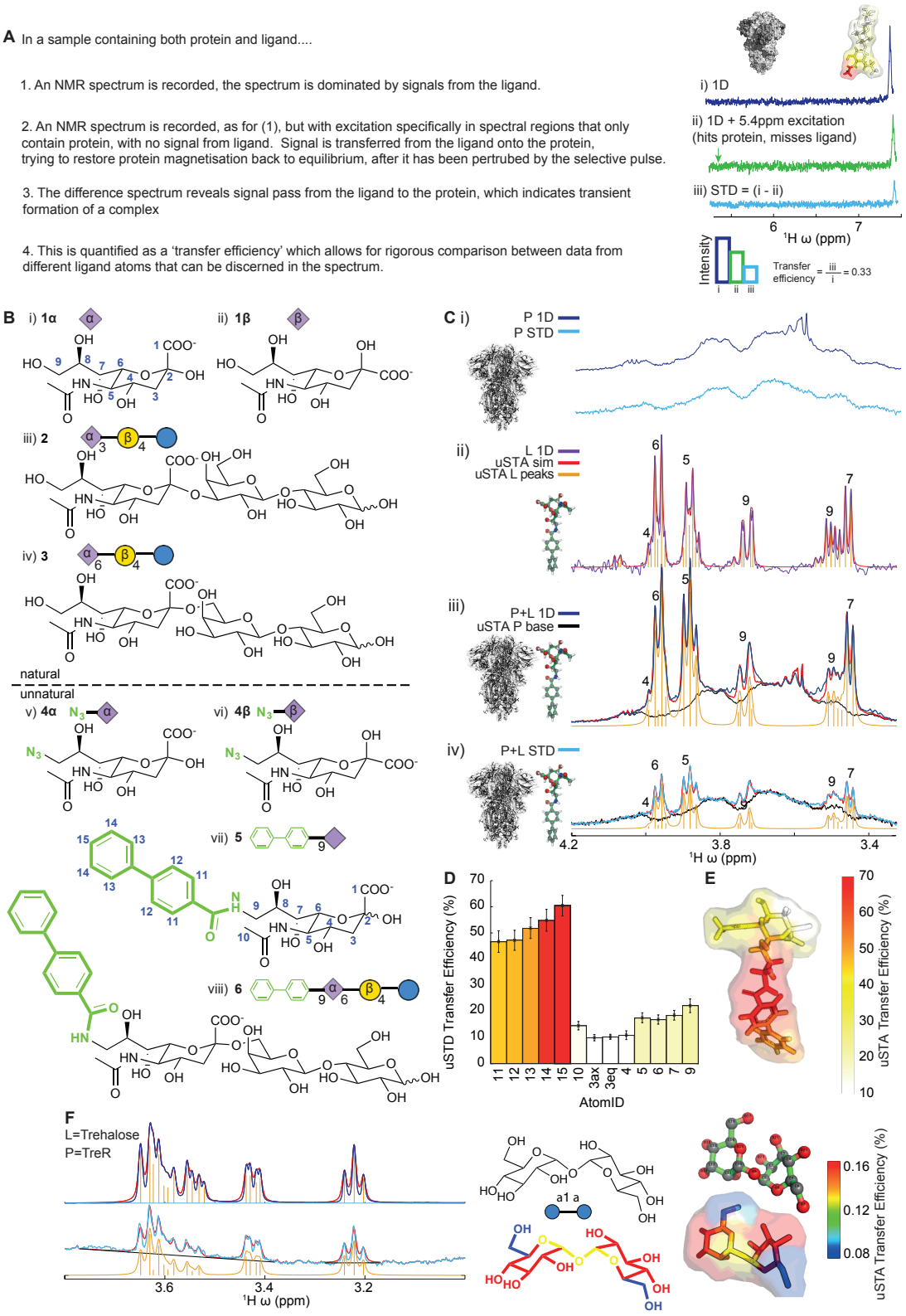


Figure A.3: A): When the protein is selectively excited, its magnetisation is pushed far from its equilibrium value. When a ligand binds the protein when in this state, cross relaxation occurs where signal from the ligand moves onto the protein, effectively trying to restore the protein to equilibrium.

Figure A.3 ***(previous page)*****:** By comparing the intensity of ligand resonances with, and without the selective excitation, the quantity of magnetisation that has been transferred can be determined, which indicates binding. In order to compare accurately between different atoms in the same spectrum, the degree of transferred magnetisation needs to be expressed relative to the initial magnetisation. This ratio is referred to in this work as the 'transfer efficiency', which can be physically interpreted as a fraction of magnetisation from an atom on the ligand that has moved onto the protein. Increasing the power and the duration of the selective excitation will increase this signal. The specific transfer efficiency will depend on the proximity of the ligand atom and the protein when the complex forms, the overall correlation time of the complex, and the chemical kinetics of the complex formation, the on and the off rate. With the uSTA method, these factors are all treated rigorously using the Bloch-McConnell equations in order to ascertain a reliable KD. B): A panel of natural, unnatural and hybrid variant sialoside sugars 1-6 was used to probe interaction between sialic acid moieties and spike. Unnatural variations (4-6, green) allowed mapping of C7-C9 side-chain interactions in the sialoside whilst use of extended sugars probed differing cell-surface glycan structures C): Application of the uSTA workflow (A.6) to SARS-CoV-2-spike protein (shown in detail for 5, see also Figures 1,2) i) The 1D 1H-NMR of SARS-CoV-2-spike protein shows considerable signal in the glycan-associated region despite protein size, indicative of mobile internal glycans in spike protein. This effectively masks traditional analyses, as without careful subtraction of the protein's contributions to the spectrum (A.7), the ligand cannot be effectively studied. The uSTA process of: ii) ligand peak assignment and deconvolution → iii) p + L peak assignment and deconvolution → iv) application to p + L STD yields precise atom-specific transfer efficiencies (A.6). Note how in ii) individual multiplet components, have been assigned (yellow); the back-calculated deconvolved spectrum (red) is an extremely close match for the raw data (purple). In iii) the spectrum is a complex superposition of the ligand spectrum (and protein only yet uSTA again accurately deconvolves the spectrum revealing the contribution of protein only (black) and the ligand peaks (yellow). Using these data, uSTA analysis of the STD spectrum in iv) pinpoints ligand peaks and signal intensities. D),E): Using these intensities, atom-specific transfer efficiencies can be determined with high precision and reveal in hybrid 5 the details of both the unnatural BPC moiety and the natural sialicacid moiety. Although the aromatic BPC dominates the interaction for the unnatural ligand 5, the subtleties of the associated sugar contribution in this ligand can nonetheless be determined (Figures A.5, A.6, A.7). F): Application of the same uSTA workflow allows precise determination of even weakly binding sugar ligand trehalose (Glc-α1,1α-Glc) to E. coli trehalose repressor TreR. Again, the uSTA allows determination of transfer efficiencies with atom-specific precision.

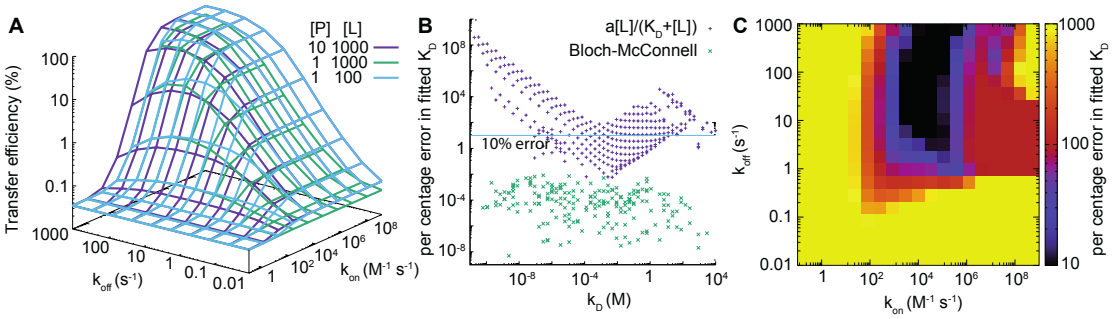


Figure A.4: A) The region of validity of the saturation transfer experiment is wider than has been realised, and that by taking both variable protein and ligand concentrations, the two rates, kon and koff, can be numerically separated from the concentration independent factors that describe the relaxation processes. Transfer efficiency from the saturation transfer experiment was simulated using the Bloch- McConnell equations (see Methods) using parameters optimised for the alpha3/spike interaction that describe the concentration independent relaxation effects (tG 2.5 ns, tE 85 ns, rISP 1.6 A, rISl=3.1 A, rIS(mix) 2.4 A, fac 18) simulated using a Gaussian pulse train of peak B1 field of 200 Hz with each element running for 50 ms for a total duration of 5 s, as utilised in the experiments, as a function of kon and koff for various protein and ligand concentrations (specified on figure, units of µM). The specific value of the parameters that affect relaxation alter the expected transfer efficiency, but fitting data with varying protein and ligand concentration is able to uniquely determine on and off rates. Experimentally, the ability to measure a transfer efficiency depends on the specific signal to noise, which can always be increased with more scans. In this work, we reliably determine transfer efficiencies on the order of 0.5%. At this sensitivity level, kon needs to exceed 100 M-1 s-1, and koff needs to exceed 0.1 s-1. Above these limits, the transfer efficiency is increases with increasing on and off rates. Above these limits, suitable combinations of ligand and protein concentration can give sufficiently different values of transfer efficiency at different mixing times to enable the concentration independent intrinsic relaxation rates to be reliably separated from the concentration dependent parameters that describe the protein/ligand interaction, kon and koff during numerical fitting. Notably, we obtain two physically sensible limits from these simulations. With high koff, the transfer efficiency becomes independent of ligand concentration (green and blue converge), whereas at high kon, the transfer efficiency becomes independent of the overall protein to ligand ratio, L/P (purple and blue converge). In a general case, determining errors in the fitting parameters is sufficient to tell a user whether or not they are well defined. If insufficient ligand/protein concentration dependence has been sampled to accurately determine kon and koff, the numerical uncertainties in these parameters obtained from a bootstrapping procedure will be large.

Figure A.4 ***(previous page)*:** B) It has been previously proposed that an approximate treatment can be applied to reliably obtain KD values.(24, 97) Applications of this method have resulted in KD values being significantly underestimated, by orders of magnitude or more, suggesting limitations in this approach.(98) By using this approach to fit simulated data, we establish that this method has a very narrow region where it can be considered effective. As has been argued previously,(97) a Bloch-McConnell treatment that covers the chemical kinetics and the relaxation processes, exactly as we perform in this work, is the method by which saturation transfer data should be analysed, and with appropriate choices of protein and ligand concentrations, is expected to yield accurate exchange parameters for the range of on and off rates described in (A). In brief, the method described by Angelo et al. determines the initial gradient of the build-up curve. These values are then followed versus ligand concentration to create an isotherm, which is fitted to A[L]/(KD+[L]) to obtain KD values. Here, we perform this analysis by simulating isotherms with 10 ligand concentrations spaced between 10 µM and 1 mM for a wide range of pairs of KD and koff. The simulated data is fitted according to the protocol described above, and the per centage error in the fitted KD versus the simulated KD is obtained. This fitting method is shown to be effective only in a narrow window of parameter space (black, errors below 10%). These points are roughly described by a locus centred on 50 µM KD and an off rate of 10 s-1. In the cases encountered in this work, the off rate is determined to be in the vicinity of 0.1 s-1, a region of parameter space where the expectation for reliable parameters using this method is expected to give, at a minimum errors of 1,000%. Fundamentally, the assumptions that underpin this method requires that the saturation transfer is proportional to the percentage of bound complex, and the proportionality constant is itself independent of protein/ligand combination. This is reliable only for a narrow region of parameter space. The differences shown here, between 'actual' and 'fitted' KD demonstrate that 'non-specific binding' is not required to explain the discrepancy.(98) The fitting methods described in this work using the Bloch-McConnell equations do not need to make assumptions about the binding regime and provide a reliable and rigorous method to obtain parameters such as kon, koff and KD.

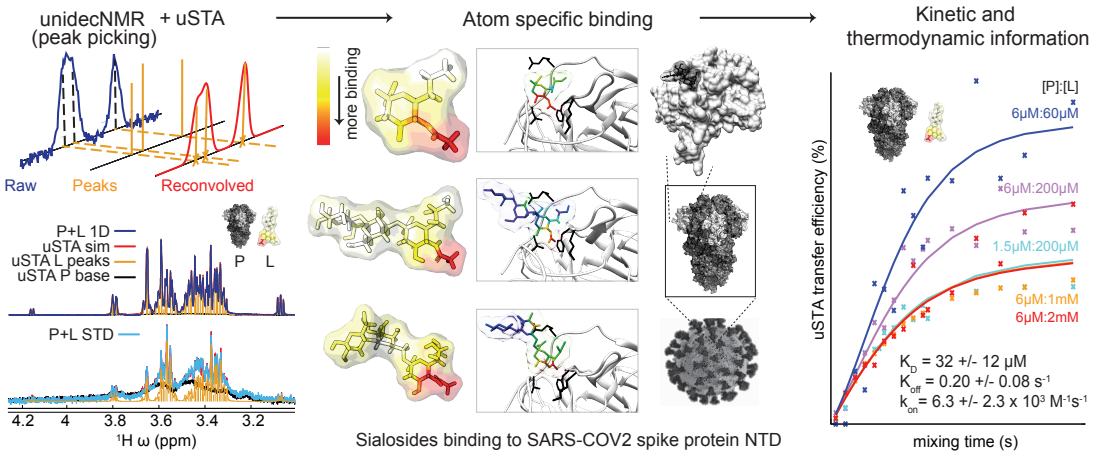


Figure A.5

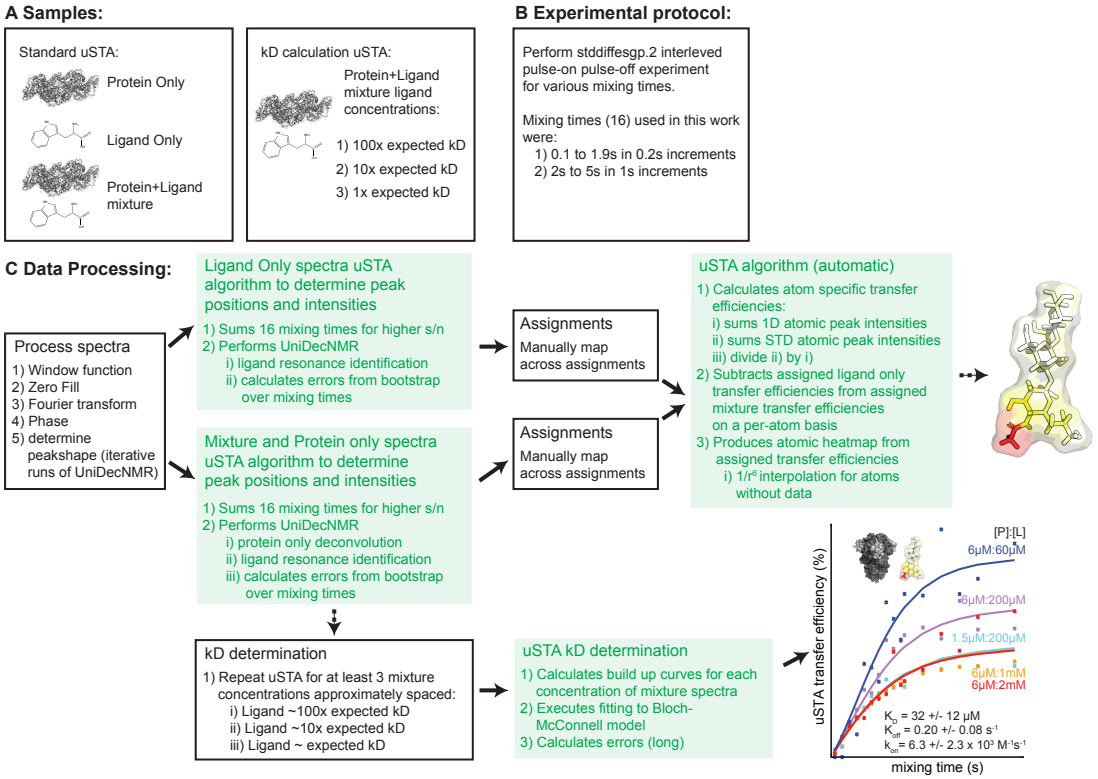


Figure A.6: Flowchart describing the specific samples required for a uSTA analysis, and the data processing steps involved. The boxes with a white inlay are the parts that require manual intervention, and the boxes in green show the automated analysis by the uSTA pipeline including figure generation.

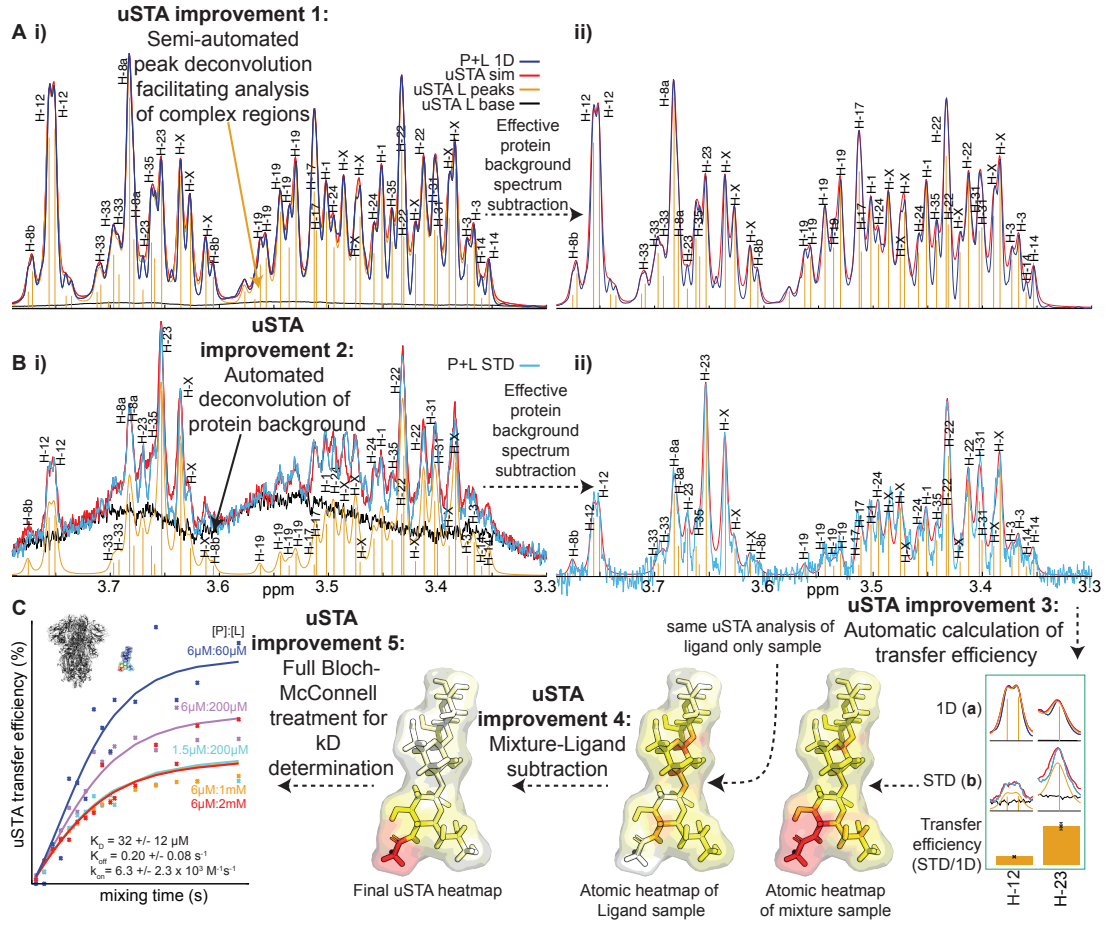


Figure A.7: The UnidecNMR algorithm allows for highly efficient and automated peak detection, including the ability to subtract the protein background. The analysis when combined with a manual assignment of the individual resonances directly computes the transfer efficiencies which when combined with ligand-only data, provide a heat map of the molecule of interest. Finally, where data are available with variable protein/ligand concentrations, the method computes a KD via a model using the Bloch-McConnell equations.

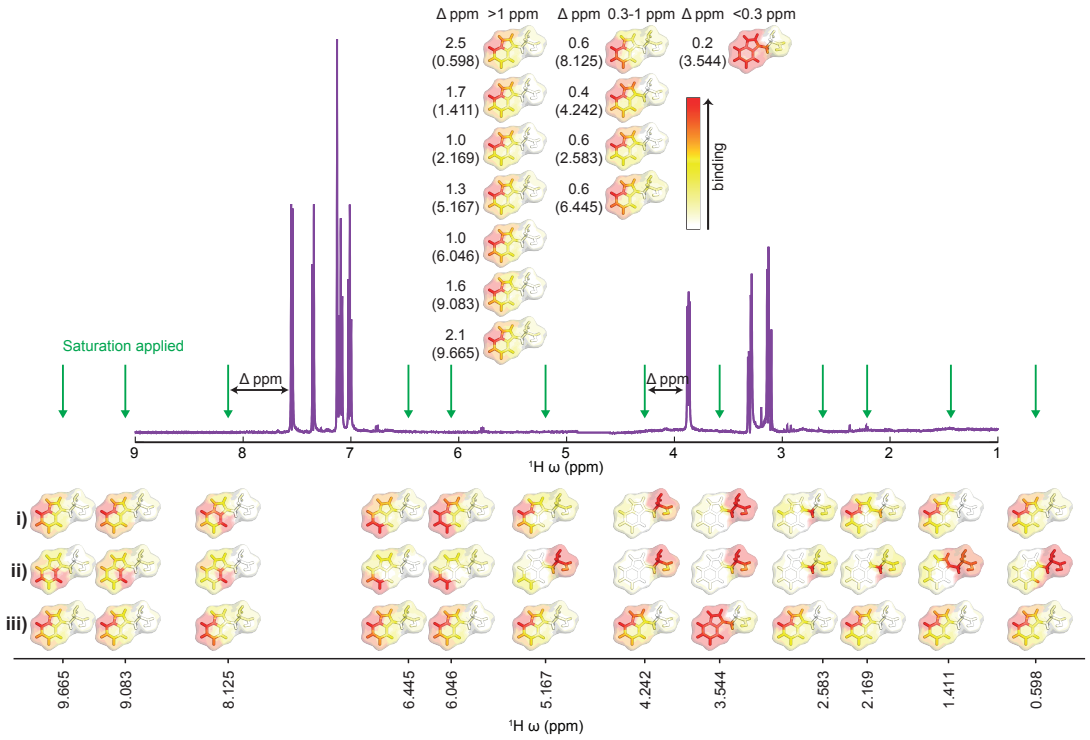


Figure A.8: Here we present the variation of the uSTA surface depending on excitation frequency. uSTA surfaces for before (i) and after (iii) subtracting the ligand-only contribution (ii), exciting at multiple frequencies indicated by the green arrows. After subtracting the ligand contributions, the final surfaces are similar except at saturation frequencies that are within 1ppm of the ligand. Between 0.3 and 1ppm, the surfaces start to change slightly, while less than 0.3ppm away the resulting heatmap becomes substantially affected. Provided one excites >1ppm from a ligand residue with a saturation pulse with this bandwidth and applies the ligand subtraction, the uSTA data appears be invariant of excitation location. This will be dependent on the correlation time of the protein, but BSA, at around 66kDa, is not that large for a STD-type protein.

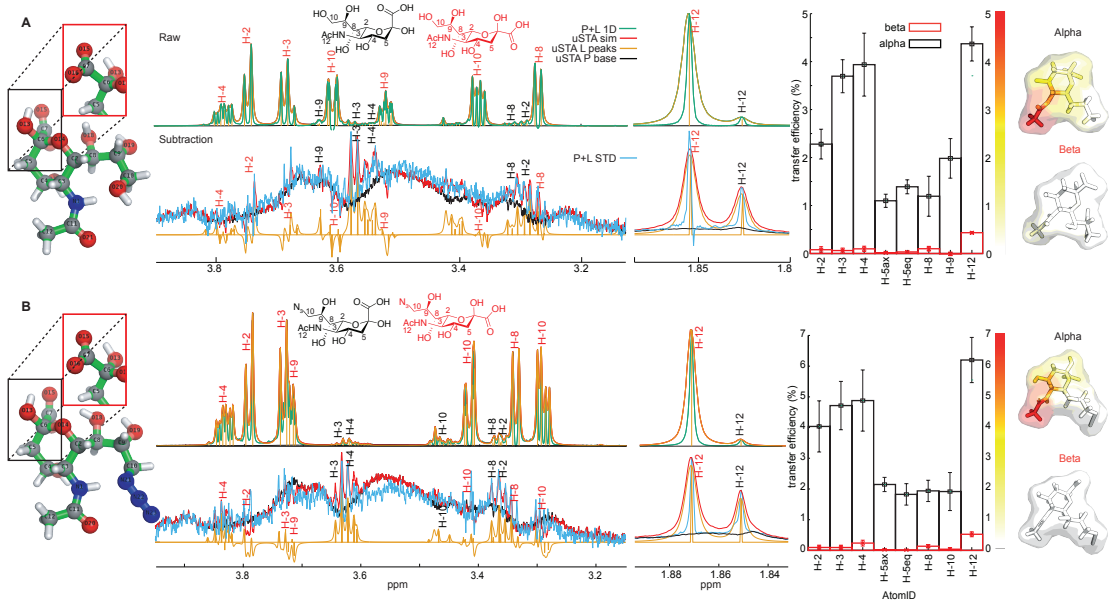


Figure A.9: In the spectra of sialic acid (1) and azido-sialic acid (4), both α and β anomeric forms could be readily identified, with the overall population being dominated by the β form (94 and 95%, respectively). Even despite this strong population difference, application of the uSTA method following assignment of resonances from the two forms allowed determination of binding surfaces simultaneously. Spike shows strong binding preference for the α anomers, as revealed by surfaces (A, C) even though its population is minor overall; both surfaces were highly similar for these two simpler monosaccharides but closely resemble those of extended trisaccharides 2 and 3 (Figure 3). While the β form is dominant in terms of population and overall contribution to the 1D NMR spectra, its ability to bind spike (B, D), and hence its proportional contribution to the difference spectrum, was found to be significantly lower than the α form. (E) For α and β anomers of 1 and 4. While the 1D spectrum is dominated by the β form of both ligands, the STD spectrum is dominated by the α form. The uSTA surface shows a similar pattern for these, and ligands 2 and 3, where the NAc from the methyl group and adjacent protons interact most strongly with the protein. Given the low signal intensities for the resonances from the β form, we cannot conclusive demonstrate any binding of the β form from these data.

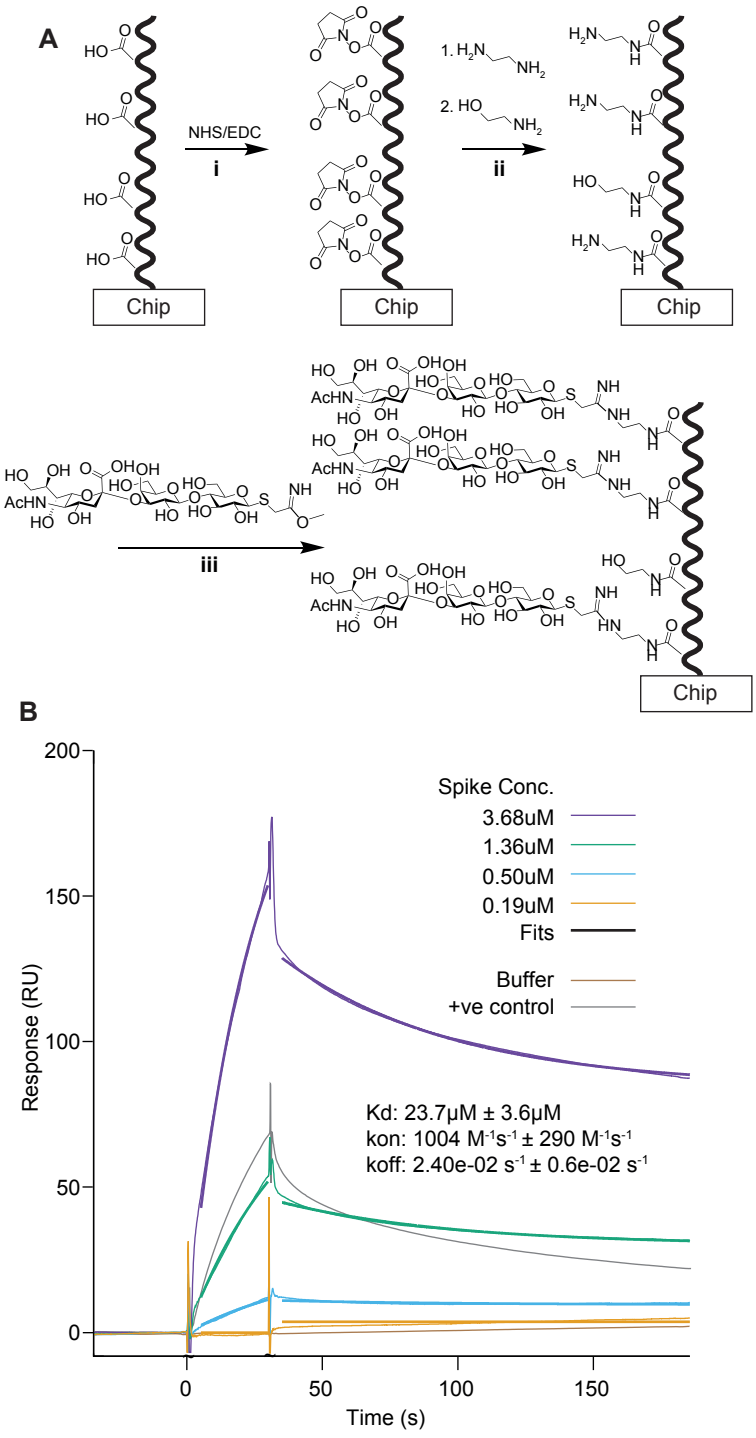


Figure A.10: A) Immobilisation of 2 onto carboxymethylated SPR sensor chip. A carboxymethylated CM5 sensor chip was activated with N-hydroxysuuccinimide (NHS) and N-ethyl-N'-(3-dimethylaminopropyl)carbodiimide hydrochloride (EDC • HCl) and subsequently coupled to ethylenediamine followed by blocking of unreacted esters with ethanolamine. In order to introduce the sugar, a solution of SiaLac-IME was then injected into the flow cell, where it reacts with the surface bound free amines generating an amidine linkage to the glycan (SiaLac-Chip). Specifically, in step i, a CM5 chip activated with NHS (50 mM) and EDC (200 mM) 10 min at 10 µL/min.

Figure A.10 ***(previous page)*:** In step ii, ethylenediamine (1 M) was injected to functionalise chip with free amine groups for 7 min at 10 µL/min. Then ethanolamine (1 M) was injected over 10 min to block any unreacted NHS activated esters at 10 µL/min. In step iii, SiaLac- IME reagent 10 (5.6 mM) was injected over 10 min at 10 µL/mL followed by washing with HBP-EP buffer (wash steps with buffer between each step and during sensor equilibration). B) Binding curves and fitted binding curves of spike at (3.68, 1.36, 0.50 and 0.19 µM) with 30 s association and 150 s dissociation at 16 ^{o}C, with a positive control (anti-SiaLac mouse serum) and corrected by subtracting BSA negative control. When accounting for background, the binding curves could be fitted to calculate apparent kon and koff rates at each spike concentration giving $K_D = 23.7\mu M \pm 3.6\mu M$; $k_{on} = 1004M^{-1}s^{-1} \pm 290M^{-1}s^{-1}$; $k_{off} = 0.024s^{-1} \pm 0.06s^{-1}$ from a global analysis of all 4 binding curves.

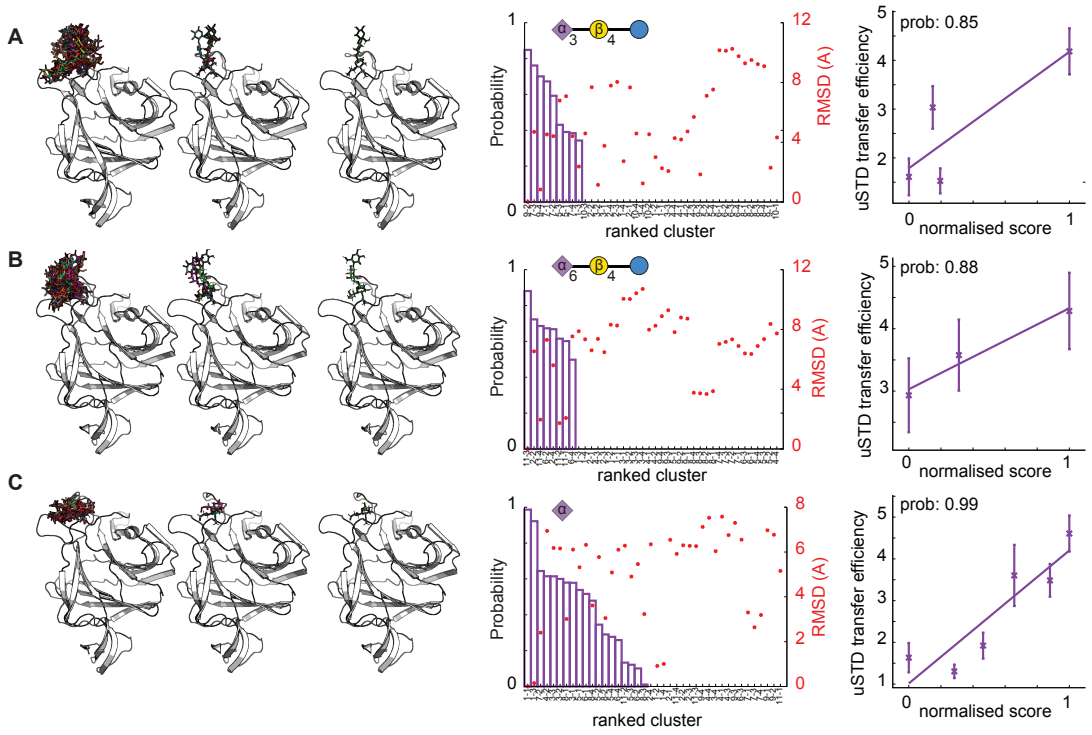


Figure A.11: A) A series of models describing the interaction between relevant sugars and spike were calculated using HADDOCK2.4 following standard methods. The docking resulted in 13 clusters ranked by the score of their top four models (i). After comparison to the NMR data, the top 4 members of all clusters were scored based on uSTA data, which allowed selection of 3 very similar models showing excellent agreement with the data (ii) of which the top scorer is shown (iii, main text Figure 5). The interactions between Sia2,3Lac and the protein are highly similar for the sialic acid moiety for the top scoring models, with variation occurring predominantly in the flexible Glc residue that has relatively little interaction with the protein. B) For each proton in the ligand, the following score was calculated through summing all adjacent protons from the protein: $(< 1/r^6 >)^{(1/6)}$. Thus any proton in close proximity to a large number of protons from the protein will receive a high score. This score is expected to correlate with the cross-relaxation rate between the protein and the ligand,[144] and so a high score should correlate with a high transfer efficiency, as measured using STD NMR spectroscopy. The Pearson 'R' correlation coefficient was calculated for each model for the correlation between the score for each atom in the ligand and converted to a probability that the correlation is not generated through random noise,[145] and the transfer efficiency, for the atoms with the highest transfer efficiencies (mapped to 5 heteroatoms, atoms 22, 23, 24 (Sia) 14 (Gal) and 42 (NAc)). The top-scoring model was an outlier, with a probability level of 98% making it statistically highly significant. The top three models form a cluster of models with a probability >90%, with all falling within 2Åall atom RMSD relative to the model with the highest R value (calculated with the proteins aligned). C) The correlation plots between a normalised score and the % transfer efficiency from the uSTA data for the best model (green, i, P(r)=98%), and two excluded models with P(r) values of 75% (blue, ii) and 39% (orange, iii) respectively. Qualitatively, the correlation line passes through the error bars of the selected models (P(r)>90%) and does not for the models that were excluded.

Figure A.11 ***(previous page)*:** Overall, all the top scoring models indicate highly similar interactions between the sialic acid and the protein, and they all predict that the NAc methyl group should be located in the vicinity of the highest proton density from the protein, in agreement with the measurements.

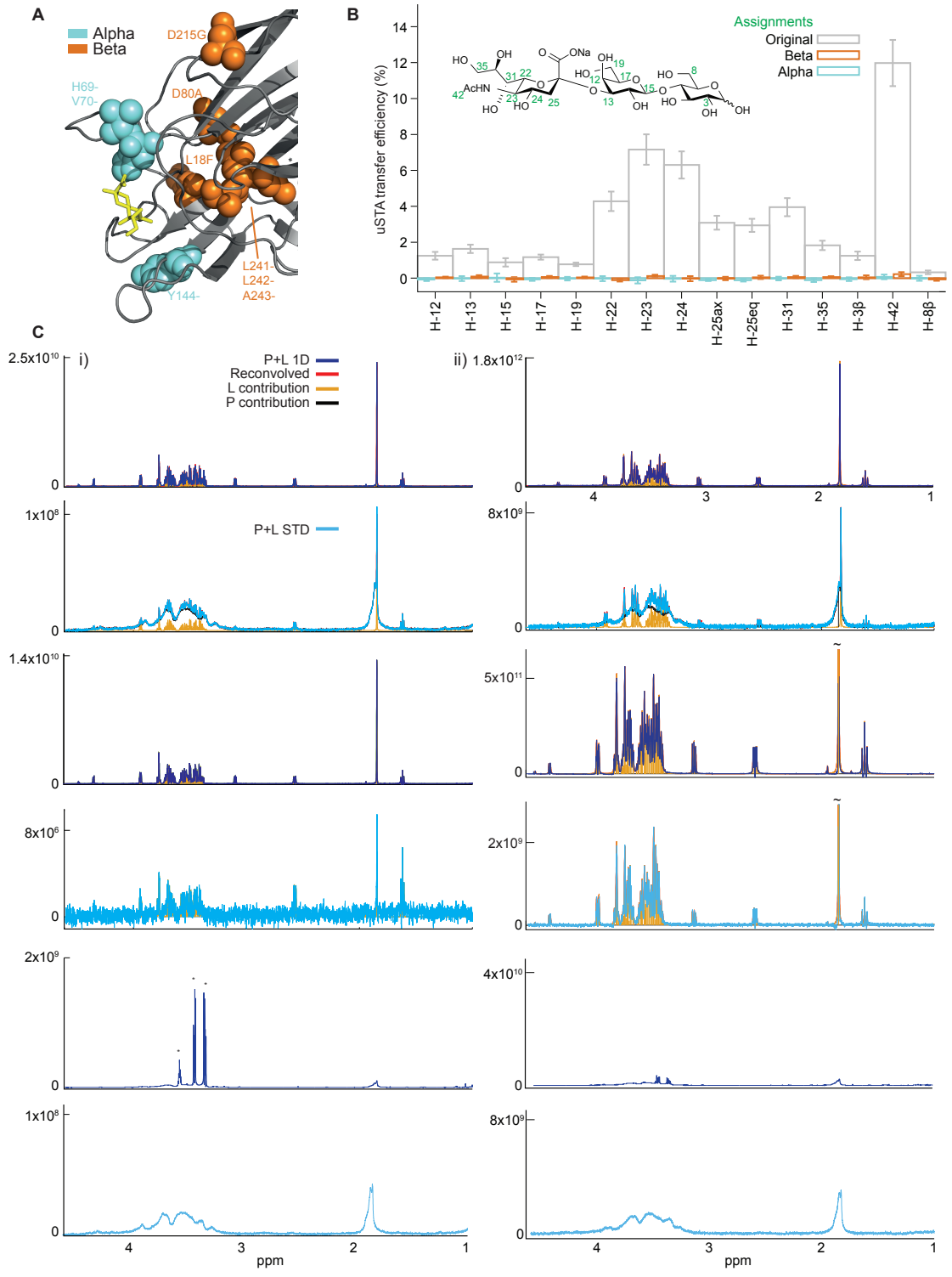


Figure A.12: Sialic acid binding pocket indicating site of B.1.1.7 and B.1.351 mutations. B) atom specific binding comparison of the three variants showing greatly diminished binding in the B- lineage variants. C) raw data for B.1.351 (i) and B.1.1.7 (ii), showing mixture 1D, STD, ligand 1D, STD and protein 1D and STD.

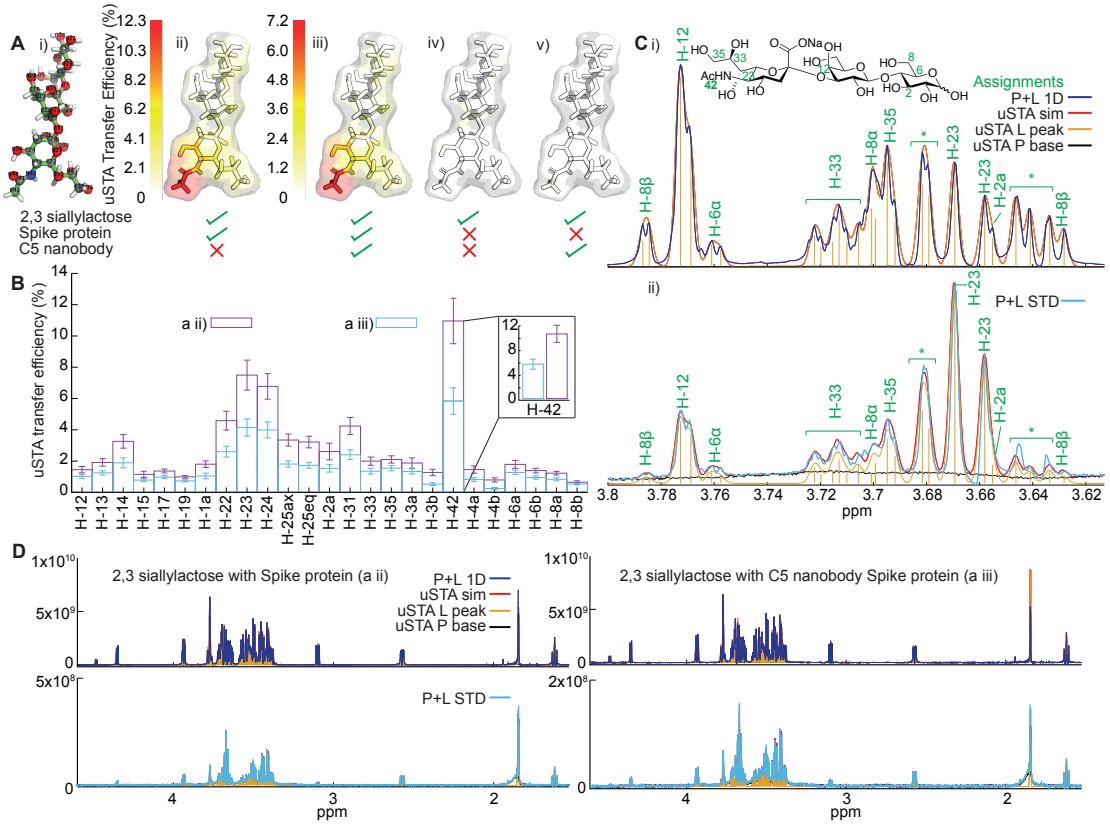


Figure A.13: A i) a wireframe with atom specific numbering ii) heatmap showing the uSTA binding of 2,3 sialyllactose to the SARS-CoV-2 Spike protein iii) heatmap showing an almost identical binding pose, but mediated transfer efficiency of the C5 nanobody bound spike protein iv) heatmap showing essentially no uSTA response from 2,3 sialyllactose without spike or C5 nanobody present, v) heatmap showing essentially no uSTA response from 2,3 sialyllactose with C5 nanobody present but no spike protein. B) the uSTA transfer efficiencies of 2,3- sialyllactose in spike and spike and nanobody systems. C i) a section of a 1-H spectrum of 2,3- sialyllactose with SARS-CoV-2 Spike protein. C ii) difference spectrum of 2,3-sialyllactose with SARS-CoV-2 Spike protein. The assignments are marked in green. * indicates a region with unassignable overlap. D) the reference spectra and difference spectra for mixtures without and with C5 nanobody

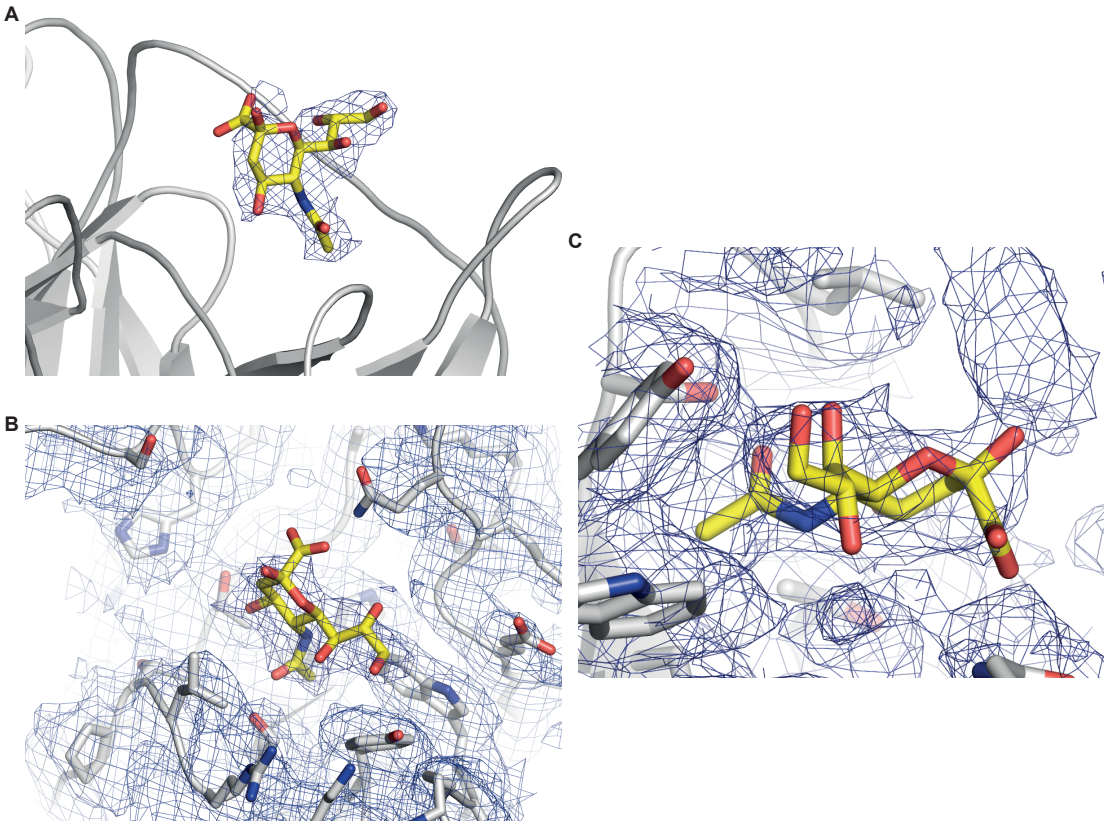


Figure A.14: The map of the sialoside portion is shown at a contour level of 2.1 σ, indicating the presence of a ligand that fits well with a super-imposed sialic acid structure (A). Increasing the contour level to 2.6 ω shows that the sialic acid moiety is similarly structured to the surrounding protein (B). Lowering the contour level to 1.6 σ suggests evidence for a linked galactoside moiety (C).

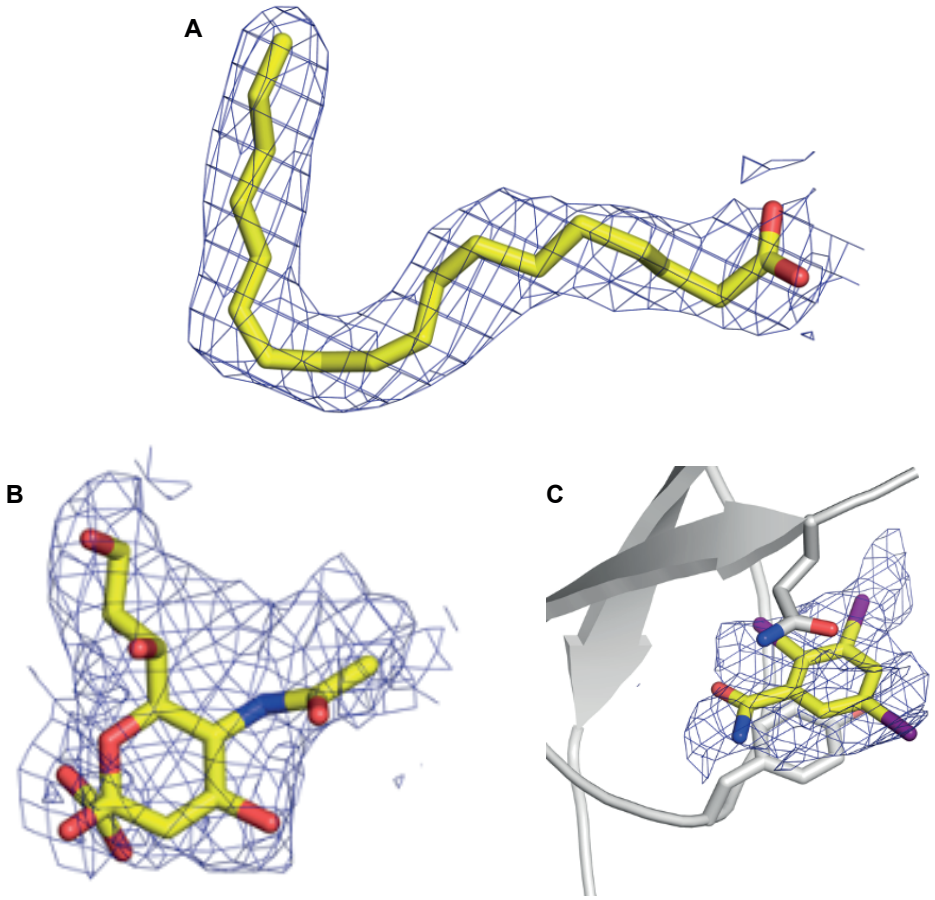


Figure A.15: The Coulombic map is shown for a region which is well described by the super-imposed Linoleic acid structure (A). Using a package called EMDA [146], the native Coulombic map was subtracted from the sialic-acid soaked map which, when contoured at 4 σ, demonstrates positive residual density that is well explained by a sialic acid moiety (B). Further residual density is shown when contoured at 12 σ, which is well explained by the tri-iodo aromatic species.

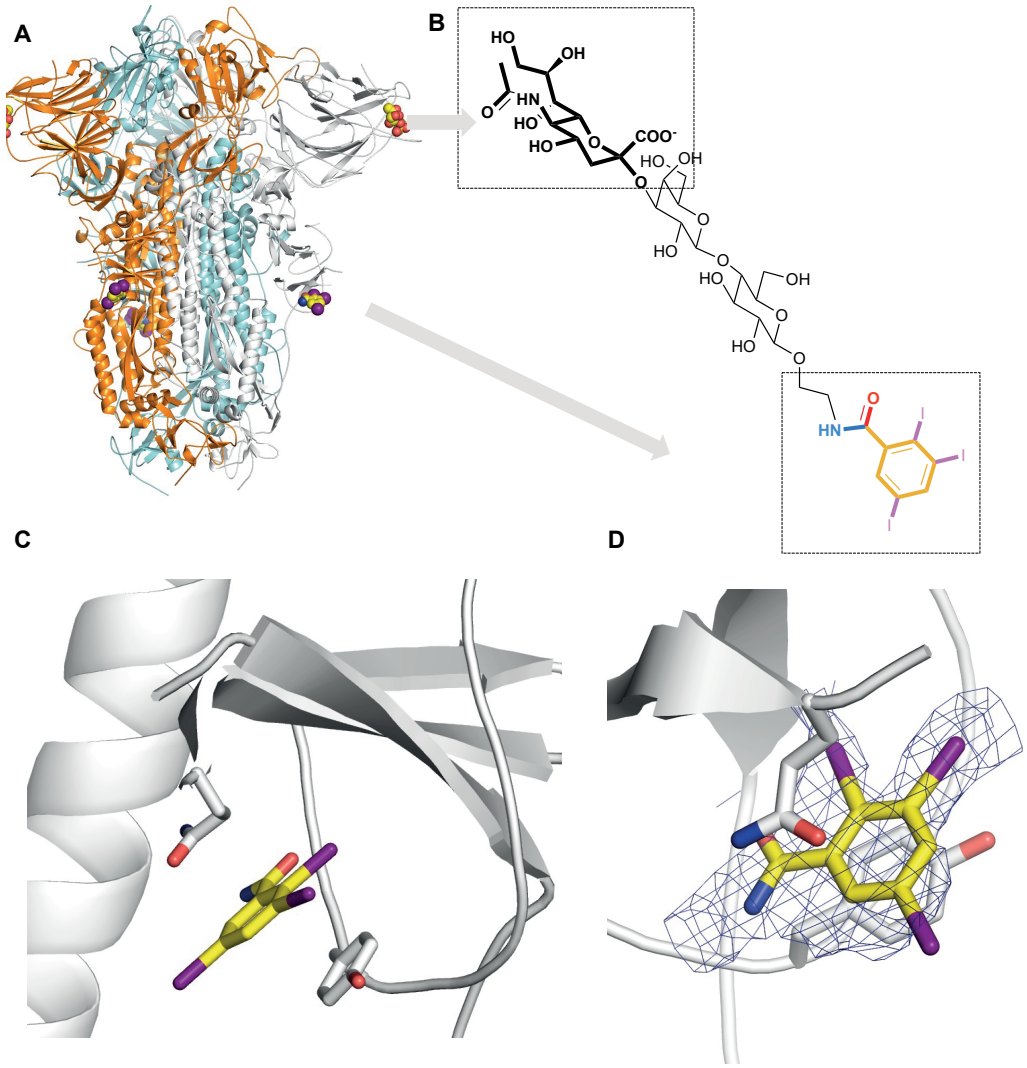


Figure A.16: The two binding locations are shown on the full spike structure, with the sialic acid binding site localised to the NTD and the iodobenzamide aromatic moiety localised to halfway up the structure (A). The sugar that was used to soak the structure is shown with the aromatic group distal to the sialic acid (B). A close up of the iodobenzamide interaction with the protein is shown (C) alongside the coulombic map contoured at 4 σ (D)

B
uSTA Theory

Contents

B.1 uSTA theory 161
B.1.1 Kinetics 164
B.1.2 Combining kinetics and spin evolution 165
B.1.3 Relaxation treatment 167
B.1.4 Using the matrix 168
B.2 Simulating 170

This appendix contains the theory that underpins our analysis of the Bloch-McConnell equations in the uSTA work.

B.1 uSTA theory

Starting with quantum mechanics, one can develop mathematical systems to understand the NMR observables. Expressing our system in terms of a wavefunction is not possible, so we employ density matrices. A density matrix describes the ensemble of spins at a point in time. The Louiville-von-Neumann equation describes how this matrix evolves with time.

$$\frac{\delta\rho}{\delta t} = -i[H, \rho] \tag{B.1}$$

$$\frac{\delta\rho}{\delta t} = \hat{\hat{H}}_0\rho + \hat{\hat{R}}(\rho - \rho_{eq}) \tag{B.2}$$

Where ρ is the density matrix, $\hat{\hat{H}}_0$ describes the evolution of the system and $\hat{\hat{R}}$, called the Relaxation super-operator, describes relaxation processes including cross-relaxation between dipolar coupled spins. The density matrix must be expressed as a sum of orthonormal, time-independent operators with correpsonding time-dependent coefficients. Taking an example of a one spin system, the operators can be the normalised cartesian basis, I_x^N, I_y^N, I_z^N, and E^N, meaning that our density matrix is expressed as a sum of these operators.

$$\rho = \frac{1}{\sqrt{2}}c_x \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + \frac{1}{\sqrt{2}}c_y \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} + \frac{1}{\sqrt{2}}c_z \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + \frac{1}{\sqrt{2}}c_E \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{B.3}$$

We have normalised the pauli spin matrces using the following formula: $\sqrt{Tr[\phi\phi]}$. We can express the coefficients in a row vector form $C = (c_x, c_y, c_z, c_E)$ that when multiplied by the time independent row matrix $\phi = (I_x^N, I_y^N, I_z^N, E^N)$ returns the above equation.

$$\rho = C\phi \tag{B.4}$$

$$\frac{d}{dt}\rho = \frac{d}{dt}C\phi \tag{B.5}$$

We have conveniently split the equation into time-dependent and independent components, meaning we need only solve $\frac{d}{dt}C$. The approach we will take is to find the square matrix, V, called the Evolution matrix, that describes the relationship between $C(0)$ and $C(t)$. V itself is by definition time independent.

$$\frac{d}{dt}C = VC \tag{B.6}$$

Solving this differential equation gives us a matrix exponential:

$$C(t) = e^{Vt}C(0) \tag{B.7}$$

To understand the composition of the evolution matrix, we must look at the derviative of the individual coefficients with time. As our density matrix is composed

of orthonormal matrix operators, we can extract the coefficient by acting on the density matrix with the corresponding operator:

$$c_i =< \rho\phi_i > \tag{B.8}$$

In this sense we are projecting; in our example along the cartesian axis of interest. Initially neglecting relaxation and differentiating gives us:

$$\frac{d}{dt}c_i = -i < \rho\hat{\hat{H}}_0\phi_i > \tag{B.9}$$

The ρ can be expanded out here to its constituents to give:

$$\frac{d}{dt}c_i = -i\sum_j < \phi_j\hat{\hat{H}}_0\phi_i > \tag{B.10}$$

This allows us to populate our matrix in the i and j dimensions. Including relaxtion and allowing for thermal equilibration and long times we get the following for each element:

$$E_{i,j} =< \phi_j(\hat{\hat{H}}_0 + \hat{\hat{R}}_0)\phi_i > -c_i^{eq} < \phi_i\hat{\hat{R}}\phi_i >< \phi_j\hat{E} > \tag{B.11}$$

We can then return these coefficients to the density matrix and, in our simple one spin case, find the state of the system in the cartesian basis at time T. For an isolated spin:

$$EC = \begin{pmatrix} -R_2 & \omega_z & -\omega_y & 0 \\ -\omega_z & -R_2 & \omega_x & 0 \\ \omega_y & -\omega_x & -R_1 & R_1M_0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} c_x \\ c_y \\ c_z \\ c_E \end{pmatrix} \tag{B.12}$$

In these equations R_2 and R_1 represent transverse and longitudinal relaxation, respectively. ω_z represents larmor oscillation, or chemical shift, which causes an interconversion of states between x and y with time. In the rotating frame, ω_z is the difference between the true oscillation frequency Ω and carrier ω_{rf} frequencies. ω_x and ω_y are defined as the frequency of oscillation on action of a pulse with x or y phase. This is dependent on the B1 field strength.

The spin states that must be considered for this system are: protein bound and free (P_B and P_F) and ligand bound and free (L_B and L_F). Each has a 4x4 matrix in cartesian basis (I_x, I_y, I_z, E).

B.1.1 Kinetics

In order to create a model for the exchanging system, we must consider the kinetics of the reaction:

$$\mathrm{P_F} + \mathrm{L_F} \underset{k_{off}}{\overset{k_{on}}{\leftrightarrows}} \mathrm{PL}$$

Where P_F and L_F refer to free protein and ligand respectively. The forward reaction is bimolecular while the backwards in monomolecular, meaning the rate constants k_{on} and k_{off} have units $mol^{-1}s^{-1}$ and s^{-1} respectively. The thermodynamic dissociation coefficient, K_d is defined as $\frac{k_{off}}{k_{on}}$. For NMR, chemical exchange timescales are governed by the k_{ex} parameters. In this case, $k_{ex}^L = k_{off} + k_{on}[P]$ while $k_{ex}^P = k_{off} + k_{on}[L]$. At equilibrium:

$$k_{on}[L][P] = k_{off}[LP] \tag{B.13}$$

$$[P] + [LP] = [P]_{tot} \tag{B.14}$$

$$[L] + [LP] = [L]_{tot} \tag{B.15}$$

We can use these equations to solve for [P], [L] and [LP] given K_d and $[L]_{tot}$ and $[P]_{tot}$, which are whare are used to make the sample. We end up with:

$$[P]^2 + [P]([L]_{tot} - [P]_{tot} + K_d) - K_d P_{tot} = 0 \tag{B.16}$$

Therefore:

$$[P] = \frac{-[L]_{tot} + [P]_{tot} - K_d + \sqrt{(-[L]_{tot} + [P]_{tot} - K_d)^2 + 4(K_d P_{tot})}}{2} \tag{B.17}$$

And equivalently

$$[L] = \frac{+[L]_{tot} - [P]_{tot} - K_d + \sqrt{([L]_{tot} - [P]_{tot} - K_d)^2 + 4(K_d P_{tot})}}{2} \tag{B.18}$$

We can formulate the kinetics into matrices:

$$\frac{d}{dt}\begin{pmatrix}[P_F]\\[P_B]\end{pmatrix} = K_P\begin{pmatrix}[P_F]\\[P_B]\end{pmatrix} \tag{B.19}$$

$$\frac{d}{dt}\begin{pmatrix}[L_F]\\[L_B]\end{pmatrix} = K_L\begin{pmatrix}[L_F]\\[L_B]\end{pmatrix} \tag{B.20}$$

Where F refers to the free (unbound) and B refers to the bound states respetively. The rate matrices K_P and K_W are then defined as follows:

$$K_P = \begin{pmatrix}-k_{on}[L_F] & k_{off}\\ k_{on}[L_F] & -k_{off}\end{pmatrix} K_L = \begin{pmatrix}-k_{on}[P_F] & k_{off}\\ k_{on}[P_F] & -k_{off}\end{pmatrix} \tag{B.21}$$

These rate matrices are then time-independent and describe the rate of change of the ground and excited state of each species given 4 paramters, k_{on}, k_{off}, $[P_F]$ and $[L_F]$.

B.1.2 Combining kinetics and spin evolution

Initially considering just the ligand spin rate matrix, we can take the Kronecker product of our ground and excited ligand spins. This will produce a 16x16 matrix, but we can collapse this because we are neglecting the effects of scalar coupling in our model. This reduces the axes from 16 to 8. The K_L rate matrix is not in the cartesian basis, and indeed is a 2x2, but we need to incorporate it into our spin evolution matrix: we can perform an outer product with an identity 4x4. This produces an 8x8 matrix that has the same species ordering as the spin evolution matrix, to which it can therefore be added to produce a full matrix for evolution under chemical exchange (in this case just for the ligand). This is included landscape on the next page because it is too big to include in portrait:

$$\frac{d}{dt}\begin{bmatrix} c_x, P_F \\ c_y, P_F \\ c_z, P_F \\ c_E, P_F \\ c_x, P_B \\ c_y, P_B \\ c_z, P_B \\ c_E, P_B \end{bmatrix} = \begin{bmatrix} -R_2^F - k_{on}[P_F] & \omega_z^F & -\omega_y^F & 0 & +k_{off} & 0 & 0 & 0 \\ -\omega_z^F & -R_2^F - k_{on}[P_F] & \omega_x^F & 0 & 0 & +k_{off} & 0 & 0 \\ \omega_y^F & -\omega_x^F & -R_1^F - k_{on}[P_F] & R_1^F M_0 & 0 & 0 & +k_{off} & 0 \\ 0 & 0 & 0 & -k_{on}[P_F] & 0 & 0 & 0 & +k_{off} \\ k_{on}[P_F] & 0 & 0 & 0 & -R_2^B - k_{off} & \omega_z^B & -\omega_y^B & 0 \\ 0 & k_{on}[P_F] & 0 & 0 & \omega_z^B & -R_2^B - k_{off} & \omega_x^B & 0 \\ 0 & 0 & k_{on}[P_F] & 0 & \omega_y^B & \omega_x^B & -R_1^B - k_{off} & R_1^B M_0 \\ 0 & 0 & 0 & k_{on}[P_F] & 0 & 0 & 0 & -k_{off} \end{bmatrix} \begin{bmatrix} c_x, P_F \\ c_y, P_F \\ c_z, P_F \\ c_E, P_F \\ c_x, P_B \\ c_y, P_B \\ c_z, P_B \\ c_E, P_B \end{bmatrix} \tag{B.22}$$

We note here the two identity rows, one of which can be removed to produce a 7x7 matrix. We can then produce a second, equivalent 7x7 matrix for the protein contribution (with its own spin parameters, but sharing the k_{on} and k_{off} rates), before performing the Kronecker product again to produce a large 49x49 matrix. The vast majority of terms in this matrix can be collapsed, as they represent both inter- and intra- state scalar coupling which are unphysical and neglected, respectively. We are then left with a 14x14, from which we can again remove one identity row/column to give a final 13x13 matrix. We need to add two moroe terms to the matrix to account for cross-correlated relaxation between our bound Protein and ligand [1], called the NOE effect. This term is the very reason the STD experiment works. This is a Z-Z transition and can be expressed as follows:

$$\frac{d}{dt}\begin{pmatrix} c_z, P_B \\ c_z, L_B \end{pmatrix} = \begin{pmatrix} 0 & -\sigma \\ -\sigma & 0 \end{pmatrix}\begin{pmatrix} c_z, P_B \\ c_z, L_B \end{pmatrix} \tag{B.23}$$

Including these terms in their respective places in the 13x13 matrix gives us our Bloch-McConnell equations.

B.1.3 Relaxation treatment

Remembering that we are attempting to fit these equations to data, the relaxation parameters contained within the matrix above, R_1, R_2, σ could be treated as distinct parameters. However, unrestricting them like this would allow the optimiser to push them into unphysical regions that would deprioritise the kinetic and thermodynamic parameters that we are trying to extract. To ameliorate this, we turn to Redfield theory, which attempts to relate the relaxation rates to motion within the molecule. In this work we use the simplest model for relaxation, caused by dipole-dipole interactions between two spins and only allowing global tumbling.

$$R_2(\tau_c) = \frac{d^2}{20}(4J(\tau_c, 0) + J(\tau_c, 0) + 9J(\omega_H) + 6J(\tau_c, 2\omega_H)) \tag{B.24}$$

$$R_1(\tau_c) = \frac{d^2}{10}(J(\tau_c, 0) + 3J(\tau_c, \omega_H) + 6J(\tau_c, 2\omega_H)) \tag{B.25}$$

[1] This effect is proportional to $\frac{1}{r^6}$ and can be neglected for unbound states

$$\sigma(\tau_c) = \frac{d^2}{10}(6J(\tau_c, 2\omega_H) - J(\tau_c, 0)) \tag{B.26}$$

Where $d = \frac{\hbar\gamma_I\gamma_S}{4\pi\mu_0 r^3}$ where r is a proton-proton distance. Therefore, the relaxation rates can be calculated from a single variable, τ_c, which is the correlation time of the molecule. This is the average amount of time taken for the molecule to travel one radian. While not a rigorous treatment of relaxation, we do not need to extract accurate distance and relaxation rate parameters and we are employing a simple model to prevent the optimiser finding values that are not physical. We can assume that the protein-ligand complex will have a very similar tumbling time to the free protein, meaning that we have two tumbling times to optimise, τ_c^P and τ_c^L referring to protein and free ligand respectively. There are also three distances to optimise: average distance within the protein for R_1 and R_2 calculations of the P and PL complex, average distance within the ligand for the R_1 and R_2 calculations of the free ligand, and average distance between the ligand and protein spins within the complex for the σ calculation.

B.1.4 Using the matrix

Having defined the matrix for our system, we can apply it to simulate the effects of our NMR experiment. If we assume our starting position is thermal equilibrium, we can express the equilibrated coefficient of c_z for each species as the equilibrated concentration of that species using the concetration equations above B.17 B.18. The assumption that the protein is a singular spin breaks down here: our initial magnetisation must be multiplied by a factor for the density of spins within the protein. This is the final parameter for our model, joining 2 correlations times, 4 chemical shifts, k_{on} and k_{off} and 3 distances to give 12 parameters.

Simulating the saturation pulse

The most substantial element of the pulse sequence is the saturation pulse. It has a gaussian pulse shape of duration 50ms, which is calibrated to a 90^o pulse. We can use the 13x13 matrix to calculate how the system evolves during this pulse.

Similarly to the GRAPE pulses, the gaussian pulse shape can be thought of as composed of many smaller pulse elements. We calibrated the maxmium amplitude pulse element at the spectometer, which gave a 200Hz field, meaning the field of element x in the gaussian pulse is $200Hz \times \frac{A}{A_M}$ where A is the amplitude of the element and A_M is the maximum amplitude of the gaussian pulse. The actual pulse shape used at the spectrometer had 1000 elements, which would slow down the calculation substantially for little benefit. Therefore, the pulse shape is calculated at the beginning of the code with a user-defined number of steps. This ensures that we capture the gaussian pulse shape while keeping a reasonable calculation time. To calculate the effect of the pulse, we must simulate each of these elements in turn. The pulse is applied on the x-axis with an offset change statement (i.e. no phase ramp). Therefore $\sigma_x^{P_B}$ and $\sigma_x^{P_F}$ are given the following values: $2\pi 200Hz\frac{A}{A_M}$.

Matrix exponentials

Having defined our evolution matrix for each step of the saturation pulse, we now have the tools to determine the evolution of the spin over that duration. We return to equation B.7 and taking our initial equilibrated magnetisation vector of coefficients, calculate the coefficients at time T:

$$C_{STD} = \prod_n e^{V_n \tau} \tag{B.27}$$

Where n is the element number of the gaussian pulse and τ is the duration of the pulse elements ($= \frac{50ms}{n}$). The final element is a 90 degree excitation pulse; we have excluded the water suppression block here to keep the experiment simple: it should affect the on- and off-resonance pulses equally.

$$C_{STD} = e^{E_H \tau_{90}} \prod_n e^{V_n \tau} \tag{B.28}$$

Where E_H is a similar matrix corresponding to a hard pulse element of duration τ_{90}.

B.2 Simulating

Using this framework, the STD experiment can be simulated under any exchange scenario. The result is an FID that needs processing before quantification of the resulting spectrum through lineshape fitting. The simulated data is then fitted to experimental data using a chi-squared cost function.

References

[1] A. C. Finlay et al. “Terramycin, a New Antibiotic”. In: *Science* 111.2874 (1950), pp. 85–85. URL: https://www.science.org/doi/abs/10.1126/science.111.2874.85.a.

[2] A. Baicus. “History of polio vaccination”. In: *World J Virol* 1.4 (2012), pp. 108–14.

[3] S. Riedel. “Edward Jenner and the history of smallpox and vaccination”. In: *Proc (Bayl Univ Med Cent)* 18.1 (2005). Riedel, Stefan Journal Article 2005/10/04 Proc (Bayl Univ Med Cent). 2005 Jan;18(1):21-5. doi: 10.1080/08998280.2005.11928028., pp. 21–5.

[4] Sara Y. Tartof et al. “Effectiveness of mRNA BNT162b2 COVID-19 vaccine up to 6 months in a large integrated health system in the USA: a retrospective cohort study”. In: *The Lancet* 398.10309 (2021). doi: 10.1016/S0140-6736(21)02183-8, pp. 1407–1416. URL: https://doi.org/10.1016/S0140-6736(21)02183-8.

[5] Seyed K. Imam. “Advancements in cancer therapy with alpha-emitters: a review”. In: *International Journal of Radiation Oncology*Biology*Physics* 51.1 (2001), pp. 271–278. URL: https://www.sciencedirect.com/science/article/pii/S0360301601015851.

[6] Christoph D. Spinner et al. “HIV pre-exposure prophylaxis (PrEP): a review of current knowledge of oral systemic HIV PrEP in humans”. In: *Infection* 44.2 (2016), pp. 151–158. URL: https://doi.org/10.1007/s15010-015-0850-2.

[7] Andrew J. Baldwin et al. “Quaternary Dynamics of αB-Crystallin as a Direct Consequence of Localised Tertiary Fluctuations in the C-Terminus”. In: *Journal of Molecular Biology* 413.2 (2011), pp. 310–320. URL: https://www.sciencedirect.com/science/article/pii/S0022283611007716.

[8] Algirdas Velyvis, Howard K. Schachman, and Lewis E. Kay. “Assignment of Ile, Leu, and Val Methyl Correlations in Supra-Molecular Systems: An Application to Aspartate Transcarbamoylase”. In: *Journal of the American Chemical Society* 131.45 (2009). doi: 10.1021/ja906978r, pp. 16534–16543. URL: https://doi.org/10.1021/ja906978r.

[9] D. Flemming Hansen, Pramodh Vallurupalli, and Lewis E. Kay. “Measurement of Methyl Group Motional Parameters of Invisible, Excited Protein States by NMR Spectroscopy”. In: *Journal of the American Chemical Society* 131.35 (2009). doi: 10.1021/ja903897e, pp. 12745–12754. URL: https://doi.org/10.1021/ja903897e.

[10] Frank Delaglio et al. “NMRPipe: A multidimensional spectral processing system based on UNIX pipes”. In: *Journal of Biomolecular NMR* 6.3 (1995), pp. 277–293. URL: https://doi.org/10.1007/BF00197809.

[11] T. Reid Alderson and Lewis E. Kay. "Unveiling invisible protein states with NMR spectroscopy". In: *Current Opinion in Structural Biology* 60 (2020), pp. 39–49. URL: https://www.sciencedirect.com/science/article/pii/S0959440X19301186.

[12] Pramodh Vallurupalli, Guillaume Bouvignies, and Lewis E. Kay. "Studying "Invisible" Excited Protein States in Slow Exchange with a Major State Conformation". In: *Journal of the American Chemical Society* 134.19 (2012). doi: 10.1021/ja3001419, pp. 8148–8161. URL: https://doi.org/10.1021/ja3001419.

[13] A. Deiana et al. "Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell". In: *PLoS One* 14.8 (2019), e0217889.

[14] Remco Sprangers and Lewis E. Kay. "Quantitative dynamics and binding studies of the 20S proteasome by NMR". In: *Nature* 445.7128 (2007), pp. 618–622. URL: https://doi.org/10.1038/nature05512.

[15] Babak Alipanahi et al. "PICKY: a novel SVD-based NMR spectra peak picking method". In: *Bioinformatics* 25 (2009), pp. 268–275.

[16] Z. Liu et al. "WaVPeak: picking NMR peaks through wavelet-based smoothing and volume-based filtering". In: *Bioinformatics* 28 (2012), pp. 914–920.

[17] Piotr Klukowski et al. "NMRNet: a deep learning approach to automated peak picking of protein NMR spectra". In: *Bioinformatics* 34.15 (2018). Klukowski, Piotr Augoff, Michal Zieba, Maciej Drwal, Maciej Gonczarek, Adam Walczak, Michal J eng Research Support, Non-U.S. Gov't England Bioinformatics. 2018 Aug 1;34(15):2590-2597. doi: 10.1093/bioinformatics/bty134., pp. 2590–2597. URL: https://www.ncbi.nlm.nih.gov/pubmed/29547986.

[18] T.D. Goddard and D.G. Kneller. *SPARKY 3*. University of California, San Francisco.

[19] William Hadley Richardson. "Bayesian-Based Iterative Method of Image Restoration*". In: *Journal of the Optical Society of America* 62.1 (1972), pp. 55–59. URL: https://opg.optica.org/abstract.cfm?URI=josa-62-1-55.

[20] L. B. Lucy. "An iterative technique for the rectification of observed distributions". In: *Astronomical Journal* 79 (June 1974), p. 745.

[21] Michael T. Marty et al. "Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles". In: *Analytical Chemistry* 87.8 (2015). doi: 10.1021/acs.analchem.5b00140, pp. 4370–4376. URL: https://doi.org/10.1021/acs.analchem.5b00140.

[22] Michael T Marty et al. "Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles". In: *Anal Chem* 87.8 (2015), pp. 4370–6. URL: https://www.ncbi.nlm.nih.gov/pubmed/25799115.

[23] P A Bromiley. "Products and Convolutions of Gaussian Probability Density Functions". In: (2014).

[24] Charles J. Buchanan et al. "Pathogen-sugar interactions revealed by universal saturation transfer analysis". In: *Science* 377.6604 (2022), eabm3125. URL: https://www.science.org/doi/abs/10.1126/science.abm3125.

[25] T. Reid Alderson et al. “Local unfolding of the HSP27 monomer regulates chaperone activity”. In: *Nature Communications* 10.1 (2019). Alderson, T Reid Roche, Julien Gastall, Heidi Y Dias, David M Pritisanac, Iva Ying, Jinfa Bax, Ad Benesch, Justin L P Baldwin, Andrew J eng Research Support, N.I.H., Intramural England Nat Commun. 2019 Mar 6;10(1):1068. doi: 10.1038/s41467-019-08557-8., p. 1068. URL: https://www.ncbi.nlm.nih.gov/pubmed/30842409.

[26] Georg K. A. Hochberg et al. “The structured core domain of αB-crystallin can prevent amyloid fibrillation and associated toxicity”. In: *Proceedings of the National Academy of Sciences* 111.16 (2014), E1562–E1570. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1322673111.

[27] Lewis E. Kay et al. “Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins”. In: *Journal of Magnetic Resonance (1969)* 89.3 (1990), pp. 496–514. URL: https://www.sciencedirect.com/science/article/pii/0022236490903335.

[28] Michael Salzmann et al. “TROSY-type Triple-Resonance Experiments for Sequential NMR Assignments of Large Proteins”. In: *Journal of the American Chemical Society* 121.4 (1999). doi: 10.1021/ja9834226, pp. 844–848. URL: https://doi.org/10.1021/ja9834226.

[29] Stephan Grzesiek and Ad Bax. “Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein”. In: *Journal of Magnetic Resonance (1969)* 96.2 (1992), pp. 432–440. URL: https://www.sciencedirect.com/science/article/pii/002223649290099S.

[30] Vitali Tugarinov et al. “Cross-Correlated Relaxation Enhanced 1H-13C NMR Spectroscopy of Methyl Groups in Very High Molecular Weight Proteins and Protein Complexes”. In: *Journal of the American Chemical Society* 125.34 (2003). doi: 10.1021/ja030153x, pp. 10420–10428. URL: https://doi.org/10.1021/ja030153x.

[31] Patrik Lundström et al. “A single-quantum methyl 13C-relaxation dispersion experiment with improved sensitivity”. In: *Journal of Biomolecular NMR* 38.1 (2007), pp. 79–88. URL: https://doi.org/10.1007/s10858-007-9149-7.

[32] Iva Pritišanac et al. “Automatic Assignment of Methyl-NMR Spectra of Supramolecular Machines Using Graph Theory”. In: *Journal of the American Chemical Society* 139.28 (2017). doi: 10.1021/jacs.6b11358, pp. 9523–9533. URL: https://doi.org/10.1021/jacs.6b11358.

[33] Iva Pritišanac et al. “Automatic Assignment of Methyl-NMR Spectra of Supramolecular Machines Using Graph Theory”. In: *Journal of the American Chemical Society* 139.28 (2017), pp. 9523–9533. URL: https://www.ncbi.nlm.nih.gov/pubmed/28691806.

[34] Vincenzo Venditti, Nicolas L. Fawzi, and G. Marius Clore. “Automated sequence- and stereo-specific assignment of methyl-labeled proteins by paramagnetic relaxation and methyl-methyl nuclear overhauser enhancement spectroscopy”. In: *Journal of Biomolecular NMR* 51.3 (2011), p. 319. URL: https://doi.org/10.1007/s10858-011-9559-4.

[35] Konstantin Pervushin et al. “Attenuated T₂ relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution”. In: *Proceedings of the National Academy of Sciences* 94.23 (1997), pp. 12366–12371. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.94.23.12366. URL: https://www.pnas.org/doi/abs/10.1073/pnas.94.23.12366.

[36] Jonathan J. Helmus and Christopher P. Jaroniec. “Nmrglue: An open source Python package for the analysis of multidimensional NMR data”. In: *Journal of Biomolecular NMR* 55.4 (2013). Helmus, Jonathan J Jaroniec, Christopher P eng R01 GM094357/GM/NIGMS NIH HHS/ R01GM094357/GM/NIGMS NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov’t Research Support, U.S. Gov’t, Non-P.H.S. Netherlands J Biomol NMR. 2013 Apr;55(4):355-67. doi: 10.1007/s10858-013-9718-x. Epub 2013 Mar 2., pp. 355–67. URL: https://www.ncbi.nlm.nih.gov/pubmed/23456039.

[37] John D. Hunter. “Matplotlib: A 2D graphics environment”. In: *Computing in science and engineering* 9 (2007), pp. 90–95.

[38] O Tkachenko. “Polydisperse chaperone proteins and the mechanisms by which they inhibit aggregation”. PhD thesis. University of Oxford, 2018.

[39] T. Reid Alderson et al. “Local unfolding of the HSP27 monomer regulates chaperone activity”. In: *Nature Communications* 10.1 (2019), p. 1068. URL: https://doi.org/10.1038/s41467-019-08557-8.

[40] M Barber. “Mechanisms and prevention of protein aggregation”. PhD thesis. University of Oxford, 2016.

[41] Ad Bax et al. “Removal of F1 baseline distortion and optimization of folding in multidimensional NMR spectra”. In: *Journal of Magnetic Resonance (1969)* 91.1 (1991), pp. 174–178. URL: https://www.sciencedirect.com/science/article/pii/002223649190422P.

[42] B. Han et al. “SHIFTX2: significantly improved protein chemical shift prediction”. In: *J Biomol NMR* 50.1 (2011), pp. 43–57.

[43] Andrew J. Baldwin et al. “Metastability of Native Proteins and the Phenomenon of Amyloid Formation”. In: *Journal of the American Chemical Society* 133.36 (2011). doi: 10.1021/ja2017703, pp. 14160–14163. URL: https://doi.org/10.1021/ja2017703.

[44] Gillian R. Hilton et al. “Small Heat-Shock Proteins: Paramedics of the Cell”. In: *Molecular Chaperones*. Ed. by Sophie Jackson. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 69–98. URL: https://doi.org/10.1007/128_2012_324.

[45] S. E. Conroy et al. “Antibodies to heat-shock protein 27 are associated with improved survival in patients with breast cancer”. In: *British Journal of Cancer* 77.11 (1998), pp. 1875–1879. URL: https://doi.org/10.1038/bjc.1998.312.

[46] J. Xiong et al. “Small Heat Shock Proteins in Cancers: Functions and Therapeutic Potential for Cancer Therapy”. In: *Int J Mol Sci* 21.18 (2020).

[47] T Reid Alderson et al. "A weakened interface in the P182L variant of HSP27 associated with severe Charcot-Marie-Tooth neuropathy causes aberrant binding to interacting proteins". In: *The EMBO Journal* 40.8 (2021), e103811. URL: https://www.embopress.org/doi/abs/10.15252/embj.2019103811.

[48] Tom Scheidt et al. "The binding of the small heat-shock protein αB-crystallin to fibrils of α-synuclein is driven by entropic forces". In: *Proceedings of the National Academy of Sciences* 118.38 (2021), e2108790118. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2108790118.

[49] Georg K. A. Hochberg et al. "Structural principles that enable oligomeric small heat-shock protein paralogs to evolve distinct functions". In: *Science* 359.6378 (2018), pp. 930–935. URL: https://www.science.org/doi/abs/10.1126/science.aam7229.

[50] R. Van Montfort, C. Slingsby, and E. Vierling. "Structure and function of the small heat shock protein/α-crystallin family of molecular chaperones". In: *Advances in Protein Chemistry* 59 (2001). Export Date: 04 October 2022; Cited By: 375, pp. 105–156. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0035718677&doi=10.1016%2fS0065-3233%2801%2959004-X&partnerID=40&md5=cb6ff779e8bebc1948fa4c635325425e.

[51] Martin Haslbeck et al. "Some like it hot: the structure and function of small heat-shock proteins". In: *Nature Structural and Molecular Biology* 12.10 (2005), pp. 842–846. URL: https://doi.org/10.1038/nsmb993.

[52] Helen M. Berman et al. "The Protein Data Bank". In: *Nucleic Acids Research* 28.1 (2000), pp. 235–242. URL: https://doi.org/10.1093/nar/28.1.235.

[53] Jian Shi et al. "Cryoelectron Microscopy Analysis of Small Heat Shock Protein 16.5 (Hsp16.5) Complexes with T4 Lysozyme Reveals the Structural Basis of Multimode Binding". In: *Journal of Biological Chemistry* 288.7 (2013). doi: 10.1074/jbc.M112.388132, pp. 4819–4830. URL: https://doi.org/10.1074/jbc.M112.388132.

[54] Kyeong Kyu Kim, Rosalind Kim, and Sung-Hou Kim. "Crystal structure of a small heat-shock protein". In: *Nature* 394.6693 (1998), pp. 595–599. URL: https://doi.org/10.1038/29106.

[55] S. Kundu and R. L. Jernigan. "Molecular mechanism of domain swapping in proteins: an analysis of slower motions". In: *Biophys J* 86.6 (2004). 1542-0086 Kundu, Sibsankar Jernigan, Robert L Journal Article 2004/06/11 Biophys J. 2004 Jun;86(6):3846-54. doi: 10.1529/biophysj.103.034736., pp. 3846–54.

[56] Konstantin Pervushin et al. "Attenuated T_2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution". In: *Proceedings of the National Academy of Sciences* 94.23 (1997), pp. 12366–12371. URL: https://www.pnas.org/doi/abs/10.1073/pnas.94.23.12366.

[57] Vitali Tugarinov and Lewis E. Kay. "An Isotope Labeling Strategy for Methyl TROSY Spectroscopy". In: *Journal of Biomolecular NMR* 28.2 (2004), pp. 165–172. URL: https://doi.org/10.1023/B:JNMR.0000013824.93994.1f.

[58] Joël Janin, Susan Miller, and Cyrus Chothia. "Surface, subunit interfaces and interior of oligomeric proteins". In: *Journal of Molecular Biology* 204.1 (1988), pp. 155–164. URL: https://www.sciencedirect.com/science/article/pii/0022283688906067.

[59] Tao Xie et al. "Conformational states dynamically populated by a kinase determine its function". In: *Science* 370.6513 (2020), eabc2754. URL: https://www.science.org/doi/abs/10.1126/science.abc2754.

[60] Yajun Jiang, Paolo Rossi, and Charalampos G. Kalodimos. "Structural basis for client recognition and activity of Hsp40 chaperones". In: *Science* 365.6459 (2019), pp. 1313–1319. URL: https://www.science.org/doi/abs/10.1126/science.aax1280.

[61] Jan H. Overbeck et al. "Observation of conformational changes that underlie the catalytic cycle of Xrn2". In: *Nature Chemical Biology* 18.10 (2022), pp. 1152–1160. URL: https://doi.org/10.1038/s41589-022-01111-6.

[62] Geerten W Vuister et al. "Increased resolution and improved spectral quality in four-dimensional 13 C/13 C-separated HMQC-NOESY-HMQC spectra using pulsed field gradients". In: (1993).

[63] Vitali Tugarinov and Lewis E. Kay. "Relaxation Rates of Degenerate 1H Transitions in Methyl Groups of Proteins as Reporters of Side-Chain Dynamics". In: *Journal of the American Chemical Society* 128.22 (2006). doi: 10.1021/ja060817d, pp. 7299–7308. URL: https://doi.org/10.1021/ja060817d.

[64] Vitali Tugarinov, Remco Sprangers, and Lewis E. Kay. "Probing Side-Chain Dynamics in the Proteasome by Relaxation Violated Coherence Transfer NMR Spectroscopy". In: *Journal of the American Chemical Society* 129.6 (2007). doi: 10.1021/ja067827z, pp. 1743–1750. URL: https://doi.org/10.1021/ja067827z.

[65] D. Flemming Hansen. *FuDA*. Computer Program.

[66] Andrew J. Baldwin and Lewis E. Kay. "NMR spectroscopy brings invisible protein states into focus". In: *Nature Chemical Biology* 5.11 (2009), pp. 808–814. URL: https://doi.org/10.1038/nchembio.238.

[67] Harden M. McConnell. "Reaction Rates by Nuclear Magnetic Resonance". In: *The Journal of Chemical Physics* 28.3 (1958), pp. 430–431. URL: https://aip.scitation.org/doi/abs/10.1063/1.1744152.

[68] D. Flemming Hansen. *CATIA*. Computer Program.

[69] Eman Basha, Kenneth L. Friedrich, and Elizabeth Vierling. "The N-terminal Arm of Small Heat Shock Proteins Is Important for Both Chaperone Activity and Substrate Specificity*". In: *Journal of Biological Chemistry* 281.52 (2006), pp. 39943–39952. URL: https://www.sciencedirect.com/science/article/pii/S0021925820768286.

[70] Emily E. Selig et al. "N- and C-terminal regions of αB-crystallin and Hsp27 mediate inhibition of amyloid nucleation, fibril binding, and fibril disaggregation". In: *Journal of Biological Chemistry* 295.29 (2020). doi: 10.1074/jbc.RA120.012748, pp. 9838–9854. URL: https://doi.org/10.1074/jbc.RA120.012748.

[71] Jacob P. Brady et al. "Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation". In: *Proceedings of the National Academy of Sciences* 114.39 (2017), E8194–E8203. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1706197114.

[72] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (2021), pp. 583–589. URL: https://doi.org/10.1038/s41586-021-03819-2.

[73] Mihaly Varadi et al. "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models". In: *Nucleic Acids Research* 50.D1 (2022), pp. D439–D444. URL: https://doi.org/10.1093/nar/gkab1061.

[74] Milot Mirdita et al. "ColabFold: making protein folding accessible to all". In: *Nature Methods* 19.6 (2022), pp. 679–682. URL: https://doi.org/10.1038/s41592-022-01488-1.

[75] *Vast.ai*. Web Page. 2022. URL: https://vast.ai/.

[76] Kunihiro Kuwajima et al. "Kinetics of disulfide bond reduction in .alpha.-lactalbumin by dithiothreitol and molecular basis of superreactivity of the Cys6-Cys120 disulfide bond". In: *Biochemistry* 29.36 (1990). doi: 10.1021/bi00488a007, pp. 8240–8249. URL: https://doi.org/10.1021/bi00488a007.

[77] Emily E. Selig et al. "The Monomeric α-Crystallin Domain of the Small Heat-shock Proteins α B-crystallin and Hsp27 Binds Amyloid Fibril Ends". In: *Journal of Molecular Biology* 434.16 (2022), p. 167711. URL: https://www.sciencedirect.com/science/article/pii/S0022283622003035.

[78] M Barber. "Mechanisms and prevention of protein aggregation". Thesis. 2016.

[79] Andrew J. Baldwin et al. "Measurement of Methyl Axis Orientations in Invisible, Excited States of Proteins by Relaxation Dispersion NMR Spectroscopy". In: *Journal of the American Chemical Society* 131.33 (2009). PMID: 19627152, pp. 11939–11948. eprint: https://doi.org/10.1021/ja903896p. URL: https://doi.org/10.1021/ja903896p.

[80] Nikolai R. Skrynnikov et al. "Probing Slow Time Scale Dynamics at Methyl-Containing Side Chains in Proteins by Relaxation Dispersion NMR Measurements: Application to Methionine Residues in a Cavity Mutant of T4 Lysozyme". In: *Journal of the American Chemical Society* 123.19 (2001). doi: 10.1021/ja004179p, pp. 4556–4566. URL: https://doi.org/10.1021/ja004179p.

[81] Daniel M. Rosenbaum, Søren G. F. Rasmussen, and Brian K. Kobilka. "The structure and function of G-protein-coupled receptors". In: *Nature* 459.7245 (2009), pp. 356–363. URL: https://doi.org/10.1038/nature08144.

[82] R. R. Ernst and W. A. Anderson. "Application of Fourier Transform Spectroscopy to Magnetic Resonance". In: *Review of Scientific Instruments* 37.1 (1966), pp. 93–102. URL: https://dx.doi.org/10.1063/1.1719961.

[83] Richard R. Ernst, Geoffrey Bodenhausen, and Alexander Wokaun. *Principles of nuclear magnetic resonance in one and two dimensions*. The International series of monographs on chemistry. 85026014 £60.00 (Richard R. Ernst, Geoffrey Bodenhausen, and Alexander Wokaun. ill. ; 25 cm. Bibliography: p. [565]-596. Includes index. Oxford Oxfordshire New York: Clarendon Press ; Oxford University Press, 1987, xxiv, 610 p. URL: Publisher%20description%20http://www.loc.gov/catdir/enhancements/fy0639/85026014-d.html%20Table%20of%20contents%20only%20http://www.loc.gov/catdir/enhancements/fy0639/85026014-t.html.

[84] Kurt Wüthrich. "NMR with proteins and nucleic acids". In: *Europhysics News* 17.1 (1986), pp. 11–13.

[85] DR Muhandiram and Lewis E Kay. "Gradient-enhanced triple-resonance three-dimensional NMR experiments with improved sensitivity". In: *Journal of Magnetic Resonance, Series B* 103.3 (1994), pp. 203–216.

[86] Ray Freeman. "Shaped radiofrequency pulses in high resolution NMR". In: *Progress in Nuclear Magnetic Resonance Spectroscopy* 32.1 (1998), pp. 59–106. URL: https://www.sciencedirect.com/science/article/pii/S0079656597000241.

[87] Helen Geen and Ray Freeman. "Band-selective radiofrequency pulses". In: *Journal of Magnetic Resonance (1969)* 93.1 (1991), pp. 93–141. URL: https://www.sciencedirect.com/science/article/pii/002223649190034Q.

[88] S. J. Glaser et al. "Unitary Control in Quantum Ensembles: Maximizing Signal Intensity in Coherent Spectroscopy". In: *Science* 280.5362 (1998), pp. 421–424. URL: https://www.science.org/doi/abs/10.1126/science.280.5362.421.

[89] F. Bloch. "Nuclear Induction". In: *Phys. Rev.* 70 (7-8 Oct. 1946), pp. 460–474. URL: https://link.aps.org/doi/10.1103/PhysRev.70.460.

[90] M. A. McCoy and L. Mueller. "Selective shaped pulse decoupling in NMR: homonuclear [carbon-13]carbonyl decoupling". In: *Journal of the American Chemical Society* 114.6 (1992). doi: 10.1021/ja00032a026, pp. 2108–2112. URL: https://doi.org/10.1021/ja00032a026.

[91] Jithender G. Reddy et al. "Simultaneous determination of fast and slow dynamics in molecules using extreme CPMG relaxation dispersion experiments". In: *Journal of Biomolecular NMR* 70.1 (2018), pp. 1–9. URL: https://doi.org/10.1007/s10858-017-0155-0.

[92] Colin A. Smith et al. "Population Shuffling of Protein Conformations". In: *Angewandte Chemie International Edition* 54.1 (2015), pp. 207–210. URL: https://dx.doi.org/10.1002/anie.201408890.

[93] Gaurav Bhole and Jonathan A. Jones. "Practical pulse engineering: Gradient ascent without matrix exponentiation". In: *Frontiers of Physics* 13.3 (2018), p. 130312. URL: https://doi.org/10.1007/s11467-018-0791-1.

[94] Helen Geen and Ray Freeman. "Band-selective radiofrequency pulses". In: *Journal of Magnetic Resonance (1969)* 93.1 (1991), pp. 93–141. URL: https://www.sciencedirect.com/science/article/pii/002223649190034Q.

[95] Lyndon Emsley and Geoffrey Bodenhausen. "Optimization of shaped selective pulses for NMR using a quaternion description of their overall propagators". In: *Journal of Magnetic Resonance (1969)* 97.1 (1992), pp. 135–148. URL: https://www.sciencedirect.com/science/article/pii/002223649290242Y.

[96] Mari A. Smith, Haitao Hu, and A. J. Shaka. "Improved Broadband Inversion Performance for NMR in Liquids". In: *Journal of Magnetic Resonance* 151.2 (2001), pp. 269–283. URL: https://www.sciencedirect.com/science/article/pii/S109078070192364X.

[97] E. Kupce and R. Freeman. "Wideband Excitation with Polychromatic Pulses". In: *Journal of Magnetic Resonance, Series A* 108.2 (1994), pp. 268–273. URL: https://www.sciencedirect.com/science/article/pii/S1064185884711235.

[98] Jean-Marl Bohlen, Martial Rey, and Geoffrey Bodenhausen. "Refocusing with chirped pulses for broadband excitation without phase dispersion". In: *Journal of Magnetic Resonance (1969)* 84.1 (1989), pp. 191–197. URL: https://www.sciencedirect.com/science/article/pii/0022236489900188.

[99] Keith E. Arntson and William C. K. Pomerantz. "Protein-Observed Fluorine NMR: A Bioorthogonal Approach for Small Molecule Discovery". In: *Journal of Medicinal Chemistry* 59.11 (2016). doi: 10.1021/acs.jmedchem.5b01447, pp. 5158–5171. URL: https://doi.org/10.1021/acs.jmedchem.5b01447.

[100] Jinfa Ying et al. "Sparse multidimensional iterative lineshape-enhanced (SMILE) reconstruction of both non-uniformly sampled and conventional NMR data". In: *Journal of Biomolecular NMR* 68.2 (2017), pp. 101–118. URL: https://doi.org/10.1007/s10858-016-0072-7.

[101] Charles J. Buchanan et al. "Pathogen-sugar interactions revealed by universal saturation transfer analysis". In: *Science* 377.6604 (2022), eabm3125. URL: https://www.science.org/doi/abs/10.1126/science.abm3125.

[102] Shinya Yamada et al. "Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors". In: *Nature* 444.7117 (2006), pp. 378–382. URL: https://doi.org/10.1038/nature05264.

[103] Wentao Li et al. "Identification of sialic acid-binding function for the Middle East respiratory syndrome coronavirus spike glycoprotein". In: *Proceedings of the National Academy of Sciences* 114.40 (2017), E8508–E8517. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1712592114.

[104] Christel Schwegmann-Weßels and Georg Herrler. "Sialic acids as receptor determinants for coronaviruses". In: *Glycoconjugate Journal* 23.1 (2006), pp. 51–58. URL: https://doi.org/10.1007/s10719-006-5437-9.

[105] Ruben J. G. Hulswit et al. "Human coronaviruses OC43 and HKU1 bind to 9-O-acetylated sialic acids via a conserved receptor-binding site in spike protein domain A". In: *Proceedings of the National Academy of Sciences* 116.7 (2019), pp. 2681–2690. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1809667116.

[106] Young-Jun Park et al. "Structures of MERS-CoV spike glycoprotein in complex with sialoside attachment receptors". In: *Nature Structural and Molecular Biology* 26.12 (2019), pp. 1151–1157. URL: https://doi.org/10.1038/s41594-019-0334-7.

[107] Enya Qing et al. “Distinct Roles for Sialoside and Protein Receptors in Coronavirus Infection”. In: *mBio* 11.1 (2020), e02764–19. URL: https://journals.asm.org/doi/abs/10.1128/mBio.02764-19.

[108] Wenhui Li et al. “Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus”. In: *Nature* 426.6965 (2003), pp. 450–454. URL: https://doi.org/10.1038/nature02145.

[109] Jun Lan et al. “Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor”. In: *Nature* 581.7807 (2020), pp. 215–220. URL: https://doi.org/10.1038/s41586-020-2180-5.

[110] Alexander N. Baker et al. “The SARS-COV-2 Spike Protein Binds Sialic Acids and Enables Rapid Detection in a Lateral Flow Point of Care Diagnostic Device”. In: *ACS Central Science* 6.11 (2020). doi: 10.1021/acscentsci.0c00855, pp. 2046–2052. URL: https://doi.org/10.1021/acscentsci.0c00855.

[111] Linh Nguyen et al. “Sialic acid-containing glycolipids mediate binding and viral entry of SARS-CoV-2”. In: *Nature Chemical Biology* 18.1 (2022), pp. 81–90. URL: https://doi.org/10.1038/s41589-021-00924-1.

[112] Xiangyang Chi et al. “A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2”. In: *Science* 369.6504 (2020), pp. 650–655. URL: https://www.science.org/doi/abs/10.1126/science.abc6952.

[113] Moriz Mayer and Bernd Meyer. “Characterization of Ligand Binding by Saturation Transfer Difference NMR Spectroscopy”. In: *Angewandte Chemie International Edition* 38.12 (1999). https://doi.org/10.1002/(SICI)1521-3773(19990614)38:12<1784::AID-ANIE1784>3.0.CO;2-Q, pp. 1784–1788. URL: https://doi.org/10.1002/(SICI)1521-3773(19990614)38:12%3C1784::AID-ANIE1784%3E3.0.CO;2-Q.

[114] Ignacio Pérez-Victoria et al. “Saturation transfer difference NMR reveals functionally essential kinetic differences for a sugar-binding repressor protein”. In: *Chemical Communications* 39 (2009), pp. 5862–5864. URL: http://dx.doi.org/10.1039/B913489A.

[115] Cyril Dominguez, Rolf Boelens, and Alexandre M. J. J. Bonvin. “HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information”. In: *Journal of the American Chemical Society* 125.7 (2003). doi: 10.1021/ja026939x, pp. 1731–1737. URL: https://doi.org/10.1021/ja026939x.

[116] G. C. P. van Zundert et al. “The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes”. In: *Journal of Molecular Biology* 428.4 (2016), pp. 720–725. URL: https://www.sciencedirect.com/science/article/pii/S0022283615005379.

[117] Anna Bujacz, Kamil Zielinski, and Bartosz Sekula. “Structural studies of bovine, equine, and leporine serum albumin complexes with naproxen”. In: *Proteins: Structure, Function, and Bioinformatics* 82.9 (2014), pp. 2199–2208. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.24583.

[118] Jesús Angulo, Pedro M. Enríquez-Navas, and Pedro M. Nieto. "Ligand-Receptor Binding Affinities from Saturation Transfer Difference (STD) NMR Spectroscopy: The Binding Isotherm of STD Initial Growth Rates". In: *Chemistry - A European Journal* 16.26 (2010), pp. 7803–7812. URL: https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/chem.200903528.

[119] Rapier H. McMenamy and J. L. Oncley. "The Specific Binding of l-Tryptophan to Serum Albumin". In: *Journal of Biological Chemistry* 233.6 (1958), pp. 1436–1447. URL: https://www.sciencedirect.com/science/article/pii/S0021925818493532.

[120] V. Jayalakshmi and N. Rama Krishna. "Complete Relaxation and Conformational Exchange Matrix (CORCEMA) Analysis of Intermolecular Saturation Transfer Effects in Reversibly Forming Ligand-Receptor Complexes". In: *Journal of Magnetic Resonance* 155.1 (2002), pp. 106–118. URL: https://www.sciencedirect.com/science/article/pii/S1090780701924991.

[121] Moriz Mayer and Bernd Meyer. "Group Epitope Mapping by Saturation Transfer Difference NMR To Identify Segments of a Ligand in Direct Contact with a Protein Receptor". In: *Journal of the American Chemical Society* 123.25 (2001). doi: 10.1021/ja0100120, pp. 6108–6117. URL: https://doi.org/10.1021/ja0100120.

[122] Lee Fielding, Samantha Rutherford, and Dan Fletcher. "Determination of protein-ligand binding affinity by NMR: observations from serum albumin model systems". In: *Magnetic Resonance in Chemistry* 43.6 (2005), pp. 463–470. URL: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/mrc.1574.

[123] Hsin-Yu Liao et al. "Differential Receptor Binding Affinities of Influenza Hemagglutinins on Glycan Arrays". In: *Journal of the American Chemical Society* 132.42 (2010). doi: 10.1021/ja104657b, pp. 14849–14856. URL: https://doi.org/10.1021/ja104657b.

[124] Daniel Lingwood et al. "Structural and genetic basis for development of broadly neutralizing influenza antibodies". In: *Nature* 489.7417 (2012), pp. 566–570. URL: https://doi.org/10.1038/nature11371.

[125] Yasunori Watanabe et al. "Site-specific glycan analysis of the SARS-CoV-2 spike". In: *Science* 369.6501 (2020), pp. 330–333. URL: https://www.science.org/doi/abs/10.1126/science.abb9983.

[126] Kosuke Nakamura et al. "Immobilized glycosylated Fmoc-amino acid for SPR: comparative studies of lectin-binding to linear or biantennary diLacNAc structures". In: *Carbohydrate Research* 382 (2013), pp. 77–85. URL: https://www.sciencedirect.com/science/article/pii/S0008621513003674.

[127] Thomas Mandel Clausen et al. "SARS-CoV-2 Infection Depends on Cellular Heparan Sulfate and ACE2". In: *Cell* 183.4 (2020), 1043–1057.e15. URL: https://www.sciencedirect.com/science/article/pii/S0092867420312307.

[128] Andrew Rambaut et al. "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology". In: *Nature Microbiology* 5.11 (2020), pp. 1403–1407. URL: https://doi.org/10.1038/s41564-020-0770-5.

[129] Jiandong Huo et al. “A potent SARS-CoV-2 neutralising nanobody shows therapeutic efficacy in the Syrian golden hamster model of COVID-19”. In: *Nature Communications* 12.1 (2021), p. 5469. URL: https://doi.org/10.1038/s41467-021-25480-z.

[130] Daniel Wrapp et al. “Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation”. In: *Science* 367.6483 (2020), pp. 1260–1263. URL: https://www.science.org/doi/abs/10.1126/science.abb2507.

[131] Christine Toelzer et al. “Free fatty acid binding pocket in the locked structure of SARS-CoV-2 spike protein”. In: *Science* 370.6517 (2020), pp. 725–730. URL: https://www.science.org/doi/abs/10.1126/science.abd3255.

[132] F. Nina Papavasiliou and David G. Schatz. “Somatic Hypermutation of Immunoglobulin Genes: Merging Mechanisms for Genetic Diversity”. In: *Cell* 109.2 (2002). doi: 10.1016/S0092-8674(02)00706-7, S35–S44. URL: https://doi.org/10.1016/S0092-8674(02)00706-7.

[133] Susumu Tonegawa. “Somatic generation of antibody diversity”. In: *Nature* 302.5909 (1983), pp. 575–581. URL: https://doi.org/10.1038/302575a0.

[134] B. Alberts et al. *Molecular Biology of the Cell*. 4th. Garland, 2002.

[135] Yan Wu et al. “A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2”. In: *Science* 368.6496 (2020), pp. 1274–1278. URL: https://www.science.org/doi/abs/10.1126/science.abc2241.

[136] Lihong Liu et al. “Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike”. In: *Nature* 584.7821 (2020), pp. 450–456. URL: https://doi.org/10.1038/s41586-020-2571-7.

[137] Yasunori Watanabe et al. “Site-specific glycan analysis of the SARS-CoV-2 spike”. In: *Science* 369.6501 (2020), pp. 330–333. URL: https://www.science.org/doi/abs/10.1126/science.abb9983.

[138] Weston B. Struwe et al. “Site-Specific Glycosylation of Virion-Derived HIV-1 Env Is Mimicked by a Soluble Trimeric Immunogen”. In: *Cell Reports* 24.8 (2018). doi: 10.1016/j.celrep.2018.07.080, 1958–1966.e5. URL: https://doi.org/10.1016/j.celrep.2018.07.080.

[139] T. Zhou et al. “A Neutralizing Antibody Recognizing Primarily N-Linked Glycan Targets the Silent Face of the HIV Envelope”. In: *Immunity* 48.3 (2018), 500–513.e6.

[140] Christopher O. Barnes et al. “Structural characterization of a highly-potent V3-glycan broadly neutralizing antibody bound to natively-glycosylated HIV-1 envelope”. In: *Nature Communications* 9.1 (2018), p. 1251. URL: https://doi.org/10.1038/s41467-018-03632-y.

[141] Tongqing Zhou et al. “Structural Basis for Broad and Potent Neutralization of HIV-1 by Antibody VRC01”. In: *Science* 329.5993 (2010), pp. 811–817. URL: https://www.science.org/doi/abs/10.1126/science.1192819.

[142] Mohammed M. Al Qaraghuli et al. "Antibody-protein binding and conformational changes: identifying allosteric signalling pathways to engineer a better effector response". In: *Scientific Reports* 10.1 (2020), p. 13696. URL: https://doi.org/10.1038/s41598-020-70680-0.

[143] J. Sebastian Temme, Dorothy L. Butler, and Jeffrey C. Gildersleeve. "Anti-glycan antibodies: roles in human disease". In: *Biochemical Journal* 478.8 (2021), pp. 1485–1509. URL: https://doi.org/10.1042/BCJ20200610.

[144] Abragam. "Principles of nuclear magnetism". In: (1961), p. 599.

[145] John Taylor. *Introduction to error analysis, the study of uncertainties in physical measurements*. 1997.

[146] Rangana Warshamanage, Keitaro Yamashita, and Garib N. Murshudov. "EMDA: A Python package for Electron Microscopy Data Analysis". In: *Journal of Structural Biology* 214.1 (2022), p. 107826. URL: https://www.sciencedirect.com/science/article/pii/S1047847721001313.