The structure of the native α-synuclein ensemble determined using a combination of structural proteomics and discrete molecular dynamics simulations

by

Nicholas Ian Brodie
Bachelor of Science, University of Victoria, 2013

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of

Doctor of Philosophy

in the Department of Biochemistry and Microbiology

Supervisory Committee

The structure of the native α-synuclein ensemble determined using a combination of structural proteomics and discrete molecular dynamics simulations

by

Nicholas Ian Brodie
Bachelor of Science, University of Victoria, 2013

Supervisory Committee

Dr. Christoph H. Borchers, Department of Biochemistry and Microbiology
Co-Supervisor

Dr. John E. Burke, Department of Biochemistry and Microbiology
Co-Supervisor

Dr. Christopher J. Nelson, Department of Biochemistry and Microbiology
Departmental Member

Dr. Patrick von Aderkas, Department of Biology
Outside Member

Abstract

In Parkinson's disease and other Lewy Body disorders, aggregation of the protein α-synuclein results in the degeneration of nervous tissue. Under normal conditions, the α-synuclein protein is abundant in neurons, where it assists in the formation of vesicles and the reuptake of neurotransmitters. However, under some conditions the protein will undergo a prion-like misfolding conversion and ultimately be converted into a fibrillar form, which makes up the bulk of the protein content of Lewy bodies. Currently, our understanding of the initial structural changes involved in the conversion of this protein into a toxic oligomeric form is hindered by the limited availability of structural data on the native, intrinsically disordered protein. Helping to define a structural ensemble for this protein would be a first step towards the development of a model for the misfolding and oligomerization process of this protein.

The research hypothesis for this dissertation is that the α-synuclein protein adopts a conformational ensemble of structures which can be elucidated using structural proteomics, and that some of these conformations have features which may lead to an increased propensity to form oligomers. In order to test this hypothesis, I utilized a variety of structural proteomics tools. These included chemical crosslinking for the discovery of distance constraints which can be used for molecular modelling, surface modification experiments which determine the propensity for particular residues to reside on the protein surface, hydrogen-deuterium exchange measurements for determining the presence or absence of secondary structure, and molecular modelling, which will be performed by collaborators at the University of North Carolina.

In order to help answer these difficult structural questions, I developed a variety of new structural proteomics techniques including photo-reactive, non-specific crosslinking reagents, ultraviolet photo-dissociation for protein fragmentation during hydrogen deuterium exchange experiments, and, most importantly, in collaboration with the University of North Carolina, I developed a computational pipeline for determining protein structures by directly incorporating distance constraints into discrete molecular dynamics simulations. These new techniques were first tested on several model proteins in order to verify their effectiveness, and were then used in combination with already-established structural proteomics techniques to model new ensembles for the native synuclein protein. This ensemble structure indicates that *in vitro* the synuclein protein adopts an ensemble of 4 distinct structures, each with some transient secondary structure. In particular, the most populated structures in the ensembles possessed secondary structure motifs in regions known to be important for oligomerization, and stabilization of these transient structures is likely to be a key component of the conversion to the oligomeric form of the protein.

Table of Contents

List of Tables

List of Figures

List of Equations

Acknowledgments

Firstly, I would like to give thanks to Dr. Christoph Borchers, for taking me on as a student, and to Dr. Evgeniy Petrotchenko for giving my first opportunity and my first introduction to work in the proteomics field. When I first began my work in their lab, I had never even heard of proteomics; since working here at the Proteomics Centre I have grown quite fond of it, and for that early introduction to the field I am quite grateful. They presented me with quite a few opportunities, to publish, to engage with the proteomics field by attending conferences, and of course a wealth of instrument time on the mass spectrometers. These were extremely useful to me as a student to gain experience in the field.

Naturally I would also like to thank my committee members Dr. Chris Nelson, Dr. John Burke and Dr. Patrick von Aderkas. Their advice and support over the years has been much appreciated. I would give an especially big thanks to Dr. Nelson, who has allowed me to "borrow" quite a lot of equipment over the years, especially his centrifuge and incubators, without which a lot of my work would not have been possible. Additional thanks to him and Dr. Geoff Gudavicius, who provided me with protein material for a number of the experiments used in this thesis. Another especially large thank you is for Dr. John Burke, who stepped up to be my co-supervisor, for which I am extremely grateful.

I would also like to take the opportunity to thank my collaborators, in particular Dr. Nikolay Dokholyan and Dr. Konstantin Popov for their work on discrete molecular dynamics modelling of my target proteins. Without them this thesis would just be a list of crosslinked residues. Additionally I'd like to thank Dr. Carol Parker for her exceptional work as editor on my papers; without her I'd be linguistically lost, but don't let her read that, she never liked alliterative flourish.

I’d like to give huge thanks to all of the staff at the University of Victoria Genome BC Proteomics Centre, but especially to Darryl Hardy, Dr. Jason Serpa and Karl Makepeace. Darryl in particular has been instrumental, pun very much intended, in teaching me virtually everything I know about the practical ins and outs of operating a mass spectrometer. There is no doubt in my mind that without his guidance I’d have never discovered the sheer fiddly joy of operating a nanospray ESI mass spec.

Lastly of course I’d like to thank my family for being endlessly supportive of this endeavor. Without them it might have taken even longer, if that can be believed. I’ll thank Michael Brodie first, for instilling in me an early love of science. I still remember that first proper experiment we ran with those tomato plants. And thanks in particular to Carol Brodie, for giving me a lift that one time the alternator in the Jeep died in the woods behind Mt. Doug. Walking the rest of the way back to the lab in the dark would have been a right mess.

Abbreviations

ACN	Acetonitrile
ABAS	Azido-benzoic-acid-succinimide
CASP	Critical Assessment of protein Structure Predictions
CBS	Carboxy-benzophenone-succinimide
CBDPS	Cyanurbiotin-dimercaptopropionyl-succinimide
CD	Circular dichroism
CID	Collision-induced dissociation
CL	Crosslink
CL-DMD	Crosslinking-discrete molecular dynamics
CLMS	Crosslinking mass spectrometry
DCC	N,N'-dicyclohexylcarbodiimide
DDA	Data dependant acquisition
DMSO	Dimethyl sulfoxide
DLB	Dementia with Lewy bodies
DMD	Discrete molecular dynamics
DSA	Disuccinimidyl adipate
DSG	Disuccinimidyl gluterate
DSSO	Disuccinimidyl sulfoxide
ECD	Electron capture dissociation
EDC	1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride
EPR	Electron paramagnetic resonance
ESI	Electrospray ionization
ETD	Electron transfer dissociation
FA	Formic acid
FDR	False discovery rate
FKBP	FK506-binding protein
FRET	Fluorescence resonance energy transfer
FTICR-MS	Fourier transform ion cyclotron resonance mass spectrometer
FTMS	Fourier transform mass spectrometer
HDX	Hydrogen-deuterium exchange
HDX-MS	Hydrogen-deuterium exchange mass spectrometry
HSA	Human serum albumin
ICPL	Isotope-coded protein label
K	Lysine
LBD	Lewy body disorders
LC	Liquid chromatography
LD-CL	Long-distance crosslinking
MALDI	Matrix-assisted laser desorption ionization
Mb	Myoglobin
MD	Molecular dynamics
MS	Mass spectrometry
MSA	Multiple System Atrophy
NAC	Non-amyloid β component

NHS	*N*-hydroxy-succinimide
Nic-NHS	*N*-nicotinoyloxy-succinimide
NMR	Nuclear magnetic resonance
PCAS	Pyridine carboxylic acid succinimide
PD	Parkinson's disease
PRE-NMR	Paramagnetic relaxation enhancement nuclear magnetic resonance
PrP	Prion protein
PTM	Post-translational modifications
REX	Replica exchange
RMSD	Root-mean-square deviation
SAXS	Small-angle X-ray scattering
SCX	Strong cation exchange
SDA	Succinimidyldiazerine
SEC	Size exclusion chromatography
SNCA	α-Synuclein gene
Sulfo-SDA	Sulfo-succinimidyl-diazerine
T	Threonine
TATA	Triazidotriamine
ThT	Thioflavin T
UV	Ultraviolet
UVPD	Ultraviolet photo-dissociation

Chapter 1: Introduction

1.1. Structural Proteomics and Mass Spectrometry

Structural proteomics is a set of techniques which combine the use of mass spectrometry (MS) and the chemical modification of proteins in order to discover new information about protein structure. There are a variety of techniques which fall within this category including: chemical crosslinking, surface modification, hydrogen-deuterium exchange (HDX), affinity labelling, and limited proteolysis [1] (Table 1). In each of these types of experiments, proteins are chemically modified in some way, and the results are analyzed using MS. New structural information can be obtained by analyzing and examining the location and extent of modification to the protein caused by each of these methods. A major advantage of structural proteomics over other, more traditional, structural biology techniques, such as X-ray crystallography and nuclear magnetic resonance (NMR), is that it relies specifically on protein chemistry to obtain new information, and can therefore be applied to nearly any protein system, with comparatively few limitations. While crystallography can often be hindered by the small amount of available protein, or a protein's inability to form ordered crystals from solution, structural proteomic methods such as crosslinking or surface modification can deliver structural information on proteins under a variety of conditions, and requires only a very small quantity of protein. Even a small amount of heterogeneity can interrupt the collection of crystal diffraction data; in contrast, structural proteomics excels at the examination of heterogeneous or disordered proteins. While NMR can often be limited by spectrum complexity, and thus is limited in its application to larger proteins, structural

proteomics can be applied to every protein system, from something as small as a peptide, all the way to large mega-Dalton complexes [2, 3] and even whole proteomes [4]. Structural proteomics techniques do not have to be used in isolation; they can also be combined with other types of data in order to clarify the results and provide additional information on the orientation and the relative arrangement of proteins and domains.

Table 1: Structural proteomics techniques and their uses for elucidating protein structures

Table detailing the type of data available from different structural proteomics experiments and how these data may be used to asses a protein's conformation or structure.

Technique	Type of Structural Data	Use
Chemical Crosslinking	Inter-residue distance constraints	Assist the determination of protein structure
Surface Modification	Relative surface accessibility of residues	Observe changes in protein structure, including burying or exposure of residues
Hydrogen-Deuterium Exchange	Secondary structure and hydrogen bonding	Observe structurization events and confirm the presence or absence of secondary structure in particular regions
Affinity Labelling	Residues modified by a photoactivatable binding partner	Binding site residues

Prior to the advent of soft ionization methods in the late 1980s, analysis of proteins by mass spectrometry was relatively limited. The ionization procedure in a typical hard ionization experiment generally leads to the fragmentation of peptides – not just along the peptide bond – and also leads to the loss of side-chain atoms, making the direct reading of peptide sequences impossible. Soft ionization techniques such as electrospray ionization (ESI) [5, 6] and matrix-assisted laser desorption ionization (MALDI) [7, 8] allow the analysis of peptides without significant loss of amino acid sequence information. Peptide or protein ions can thus be fragmented *specifically* along the peptide backbone, preserving information on the identity of the amino acid from side-chains. This

can be done by using a variety of fragmentation techniques including collision-induced dissociation (CID) [9, 10], electron capture dissociation (ECD) [11, 12], electron transfer dissociation (ETD) [13] or ultra-violet photo-dissociation (UVPD) [14]. The resulting mass spectrum of the secondary fragment ions is typically referred to as an MS/MS spectrum; these MS/MS spectra can then be interpreted on their own or the peaks can be compared to expected sequence fragment ions from a protein sequence database in order to establish the identity of the parent ion. Since these techniques preserve the information regarding the status of the side-chain of each residue, they can also be used to determine any chemical modifications to residues have occurred, including natural post-translational modifications (PTMs) or experimentally induced modifications to the residues. It is these induced modifications and their detection which enable these techniques to be used for structural proteomics.

1.2. Crosslinking for analysis of protein structures

At the most basic level, chemical crosslinking of proteins followed by mass spectrometric detection and identification of the crosslinked species (CLMS) is used to identify the particular residues on proteins which were spatially proximate, based on the formation of a chemical reaction between the two residues which are now covalently linked by a crosslinking reagent [15, 16]. A crosslinking reagent typically has two reactive moieties separated by a backbone structure, which provides a rigid limit on the maximum possible spatial distance between two residues so crosslinked. This distance thus corresponds to a type of distance constraint, conceptually similar to the types of constraints which can be generated during NMR or fluorescence resonance energy transfer (FRET) experiments. The crosslinker's spacer arm can vary in length, so these

constraints can range from zero-length (those in which residues react directly with one another with no linker space) [17] to 14 Å linker arms or more on some of the longer range crosslinkers [18]. Once generated, these constraints can be used in a variety of ways, including helping to align members of a complex in the correct orientation [19, 20], validation of protein models generated *in silico*[21, 22], or incorporating the constraints directly into molecular modelling simulations.

In a typical crosslinking experiment, proteins are incubated with crosslinking reagents under native conditions, in whichever state (drug-bound, in-complex, etc.) is being analyzed (Figure 1). Typically, crosslinker chemistry uses n-hydroxy-succinimide (NHS) [23, 24] ester chemistry for at least one of the linkages (Figure 2). This moiety reacts to form a covalent bond between the crosslinker and a nucleophile. In a protein sample, this is typically the primary amine present on the side-chain of lysine residues. Since lysine residues are often somewhat limited in availability across a protein sequence [25], additional crosslinking chemistries can also be employed. These include photoreactive non-specific groups [25-28], thiol-reactive groups [29], and 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride (EDC)-based crosslinking strategies which modify acidic residues with an acylisourea group which can be directly attacked by nucleophiles [17] (Figure 2). Crosslinking is performed at concentrations designed to minimize the number of non-specific crosslinking reactions – usually by limiting the protein:crosslinker molar ratio to 1:20, with most experiments occurring in the 1:5 to 1:10 range [15]. Crosslinking reaction times vary considerably based on the target material,

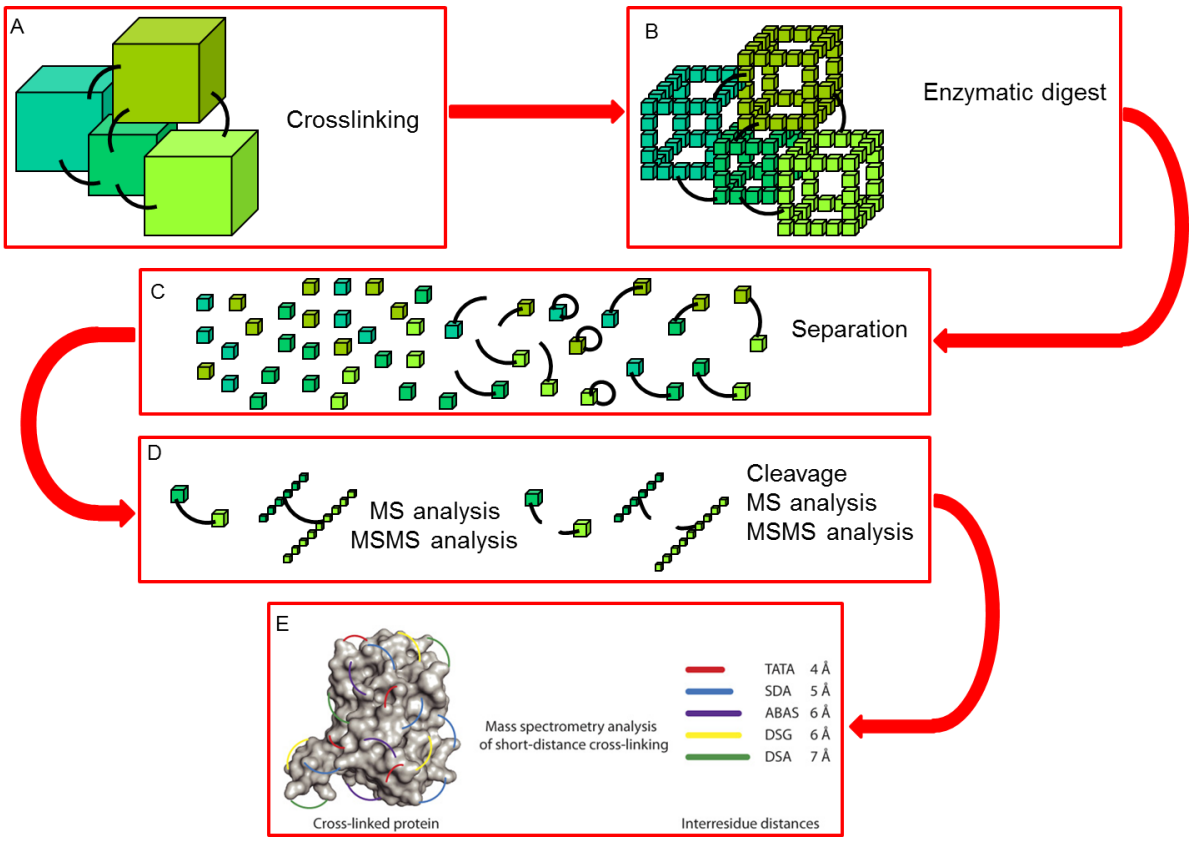


Figure 1: Workflow of a typical crosslinking experiment

A: Proteins are crosslinked using a single crosslinking reagent. B: Crosslinked proteins are enzymatically digested. C: Crosslinked and non-crosslinked peptides are separated, typically by reverse-phase liquid chromatography in line with a mass spectrometer. D: MS and MS/MS data are collected for peptides, and crosslinking sites are identified. E: Results of crosslinking experiments from multiple reagents with different lengths and reactivities are aggregated together to form one data set for modelling protein structures.

but for single proteins or simple complexes, 10-15 minutes is appropriate. After crosslinking is complete, reactions can be quenched using a buffer which includes an excess of substrate, typically ammonia bicarbonate in the case of NHS-ester reactions [15].

The preparation of the proteins for MS analysis can then be approached in several ways; the simplest approach is an in-solution digestion of the proteins using a proteolytic

enzyme, typically one with high specificity such as trypsin. This high degree of specificity will significantly simplify data analysis during the later stages of this pipeline. The proteins can also be separated prior to enzymatic digestion, typically by size using either SDS-PAGE or size-exclusion chromatography. The excised protein gel bands or specific chromatographic fractions can then be subjected to proteolytic digestion as described above [30].

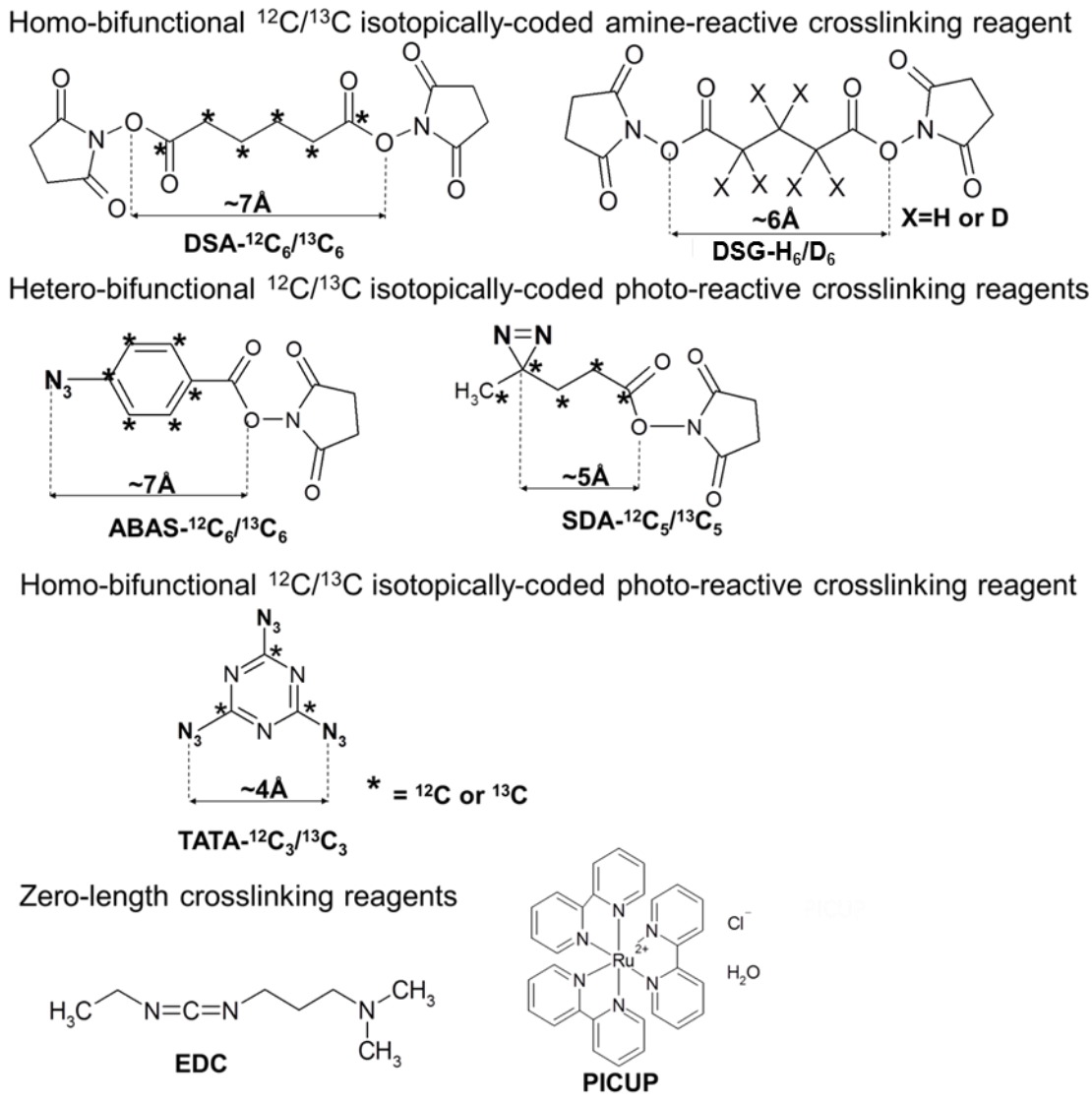


Figure 2: Crosslinkers commonly used in protein crosslinking experiments

Once a sample of digested peptides has been obtained, MS analysis is typically carried out using a reversed-phase liquid chromatography (LC) separation system attached to an ESI or MALDI mass spectrometer. Mass spectrometers typically used for analysing crosslinked peptide samples will share a number of important characteristics: 1) a soft ionization method such as MALDI or ESI; 2) a fragmentation method (CID and ETD are most common) for sequencing the peptides [16, 31, 32]; 3) a mass analyzer that provides high mass accuracy as well as relatively rapid scan rates, such as an Orbitrap or a time-of-flight mass analyzer; 4) high mass accuracy, of at least 10 ppm, are required to assist both the accuracy and speed of data analysis. The high sensitivity and scan rates will also allow the acquisition of very low-intensity signals. Since crosslinking is a relatively rare event, signals representing crosslinks are often of very low intensity and high sensitivity allows more crosslink (CL) ions to be detected; higher scan rates allow more ions to be acquired in MS/MS events, which is required for identification of the crosslink.

In addition to this, the crosslinkers themselves can be engineered with features that assist in the detection and identification of crosslinked peptides. Isotopic labelling of the crosslinker can be used to create a crosslinker-specific signature in the MS spectrum which can then be used to target the crosslinked peptides for MS/MS acquisition during an experiment [15, 33]. Heavy-labelled atoms, such as deuterium or carbon-13, may be incorporated into the crosslinker structure, and these heavy-labelled reagents are used in a 1:1 ratio with the regular crosslinker. As a result, they produce a distinctive doublet signature in the spectrum that corresponds to the mass difference. The software used on many commercial mass spectrometers incorporates the ability to search for such mass

differences during data-dependant acquisition (DDA) modes. This can increase the number of crosslinker-specific acquisition events. Crosslinkers can also include features which aid in identification after MS/MS fragmentation has occurred, specifically in the form of crosslinker-specific cleavage sites engineered directly into the reagent [34, 35]. These cleavage sites differ from typical peptide cleavages, and will produce product ions with crosslinker-specific mass additions, thereby increasing the specificity of identification [36]. Finally, the inclusion of specific enrichment tags may be used to increase the number of crosslinked peptides in the final sample before it is analyzed. The most commonly used tag is biotin, which is easy to incorporate into the relatively small crosslinker molecules [37]. Enrichment can also be performed by taking advantage of some of the typical features of crosslinked peptides. For example, crosslinked peptides typically carry an increased charge as a result of having two N-termini (one from each peptide) and usually a second tryptic cleavage site. Thus most crosslinks will have a charge of +3 to +6 at a pH of 1-2, as compared to peptides and crosslinker-modified single peptides, which usually have charges between +1 and +3. As a result of this difference in charge, strong cation exchange (SCX) [38-40] chromatography can be used to at least partially separate the two populations. Since crosslinked peptides are twice the size of their non-crosslinked counterparts, size exclusion chromatography (SEC) [41] can also be used to separate them and to generate fractions which are enriched for crosslinked peptides.

The desired end product of a crosslinking experiment is a set of constraints which can assist in the process of modelling a structure of a protein or complex. However, additional challenges occur if the complex is a homo-complex, consisting of the multiple units of one protein. Because the sequences of the different components of the complex are identical, it is impossible to distinguish those crosslinks which occur *within* a single protein sequence from those which occur *between* two proteins within the complex. In order to solve this problem, a heavy-labelled version of a protein can be used in a $^{14}N/^{15}N$ crosslinking experiment [20, 42-44]. Light- and heavy-labelled proteins are mixed in a 1:1 ratio, and are allowed to equilibrate and form mixed complexes (Figure 3). The proteins are then crosslinked, digested, and analyzed by LC-MS as usual. In the case of a crosslink which occurs between two subunits of a $^{14}N/^{15}N$ oligomer, there is a distinctive signal consisting of 4 peaks with mass differences corresponding to the number of nitrogen atoms within the crosslinked peptides. The two outer peaks of the signature quadruplet are the all-^{14}N and all-^{15}N peaks, which appear regardless of whether the crosslink is inter- or intra-protein. The inner two peaks are those produced by the mixed $^{14}N/^{15}N$ composition of the crosslink, and are only present in crosslinks which are inter-protein. In most cases these two inner and the two outer peaks will be in a 1:1 ratio, indicating that the crosslink formed predominantly between 2 protein molecules. These 4 peaks can also provide additional confirmatory evidence. Identification of multiple peaks in a cluster is indicative of single crosslinked peptide pair with the corresponding combination of nitrogen atoms on each peptide. This provides excellent evidence for the identity of the crosslink.

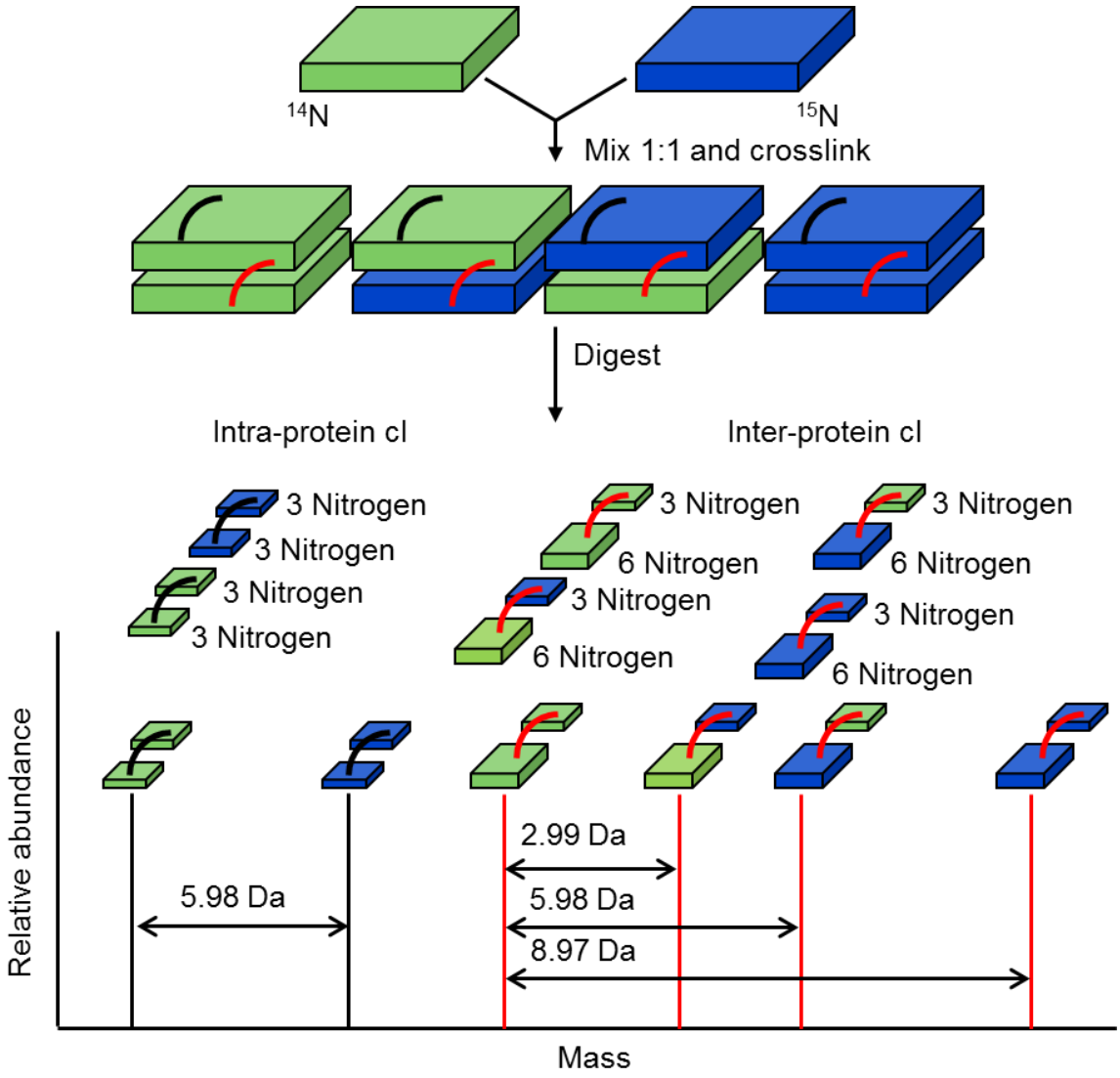


Figure 3: $^{14}N/^{15}N$ crosslinking scheme

Light and heavy forms of a protein are mixed at a 1:1 ratio and crosslinked. There are two distinct results of this mixing. Represented in black are intra-protein crosslinks; these result when the crosslinked residues are on the same protein molecule. They will have a distinct doublet signature in the MS spectrum, either all light, or all heavy. Represented in red are inter-protein crosslinks; these crosslinks form between two protein molecules. They will have a quadruplet signature in the MS spectrum, representing a mix of ^{14}N and ^{15}N peptide pairs.

1.3. Software for analyzing crosslinking data

After the data has been collected, there are now a variety of options for its analysis. These include the DXMSMS Match [45], Kojak [46], Stavrox [47], XlinX [48], XiSearch [49], xQuest/xProphet [41, 50], and many others. All of these programs incorporate the

basic feature of matching peptide fragment ions from MS/MS spectrum to crosslinked peptide pairs which match the observed MS parent-ion mass. The resulting match is then scored, using a variety of different methods depending on the software, with the idea that the highest resulting score is the most likely pair of crosslinked peptides that generated that spectrum. Some of these software packages also take into account additional crosslinker features which help with the identification. XlinX for example was designed with the cleavable crosslinker disuccinimidyl sulfoxide (DSSO) in mind. This crosslinker contains a cleavable sulfoxide bond which fragments under CID and ETD conditions to generate a set of distinctive product ions which can be used to enhance confidence in the crosslink identification. In order to increase the confidence of the assignment of the MS/MS spectrum, DXMSMS Match uses a similar approach, and can combine this with isotopic labelling as well.

Output from the majority of these software packages is subsequently statistically analyzed and/or manually validated. This is called post-processing. The usual method for post-processing involves the calculation of a false discovery rate (FDR). A database of decoy proteins is generated by reversing the peptide sequence of the target protein database. Hits on these decoy proteins constitute false-positive results, and can be used to calculate the FDR. In addition, manual validation of high scoring spectra may be performed in order to verify that the software is producing appropriate high-quality spectra in its high-scoring output. Additional post-processing can also come in the form of machine learning algorithms such as Percolator [51]. Percolator uses a semi-supervised learning approach to compare decoy spectra to target spectra to generate weights for user-defined features which can include virtually any piece of information known about the

potential identification, including, for example, the presence or absence of cleavage ions or the presence of an MS doublet from isotopic labelling of the crosslinker.

1.4. Hydrogen-Deuterium exchange

Hydrogen-deuterium exchange mass spectrometry (HDX-MS) is a structural proteomics method for examining the secondary structure and the hydrogen bonding of proteins in solution [52, 53]. When a protein is immersed in deuterium at a neutral pD, the majority of the exchangeable hydrogen atoms on the protein will be replaced with deuterium over time, with the rate of the exchange a function of pD, temperature, and the chemical environment of the hydrogen atom. If a hydrogen atom is involved in a hydrogen bond for example, the exchange rate will be significantly slower. Each exchanged deuterium atom also provides a 1 Da increase in mass. Thus the number of exchanged hydrogen atoms can be deduced by comparing mass spectra of non-deuterated samples with those that have been subjected to HDX.

There are two different approaches to HDX-MS measurement that can be taken: top-down [54] and bottom-up [52]. In a bottom-up HDX-MS experiment, the protein sample is subjected to HDX and then rapidly digested under acidic conditions using an acid stable protease such as pepsin (Figure 4A), followed immediately by MS analysis. The deuterium incorporation levels in identified peptides can then be determined by comparing the distribution to that of an analogous non-exchanged peptide acquired in a separate analysis. These levels can then be compiled and the deuteration levels across the entire protein can be calculated [55]. Thus, the deuteration level is ultimately determined on the peptide level. In a top-down HDX experiment, intact protein ions are injected into the mass spectrometer with no prior digestion, and are then fragmented using fast

fragmentation techniques such as ECD [54, 56], ETD [57], or UVPD [58] (Figure 4A). The deuteration levels can then be read from each fragment ion by comparing them to a fragmentation spectrum from the non-HDX protein. The primary advantage of top-down over bottom-up HDX experiments is that for small sized proteins (< 45 kDa) fragmentation typically occurs between most of the individual peptide bonds, and thus residue-level coverage over a large proportion of the protein can be obtained. For larger proteins, however, fragmentation will not be nearly as complete; bottom-up approaches will yield better coverage in this case.

Top-down HDX is usually measured with a continuous flow system consisting of three connecting syringes. The first two syringes contain either sample or D_2O, respectively (Figure 4B). The outflows from these two syringes combine at a T-junction, and the flows are typically combined in a 1:4 ratio, yielding 80 % D_2O. The mixture then flows through another and capillary, whose volume determines the exchange time experienced by the protein.

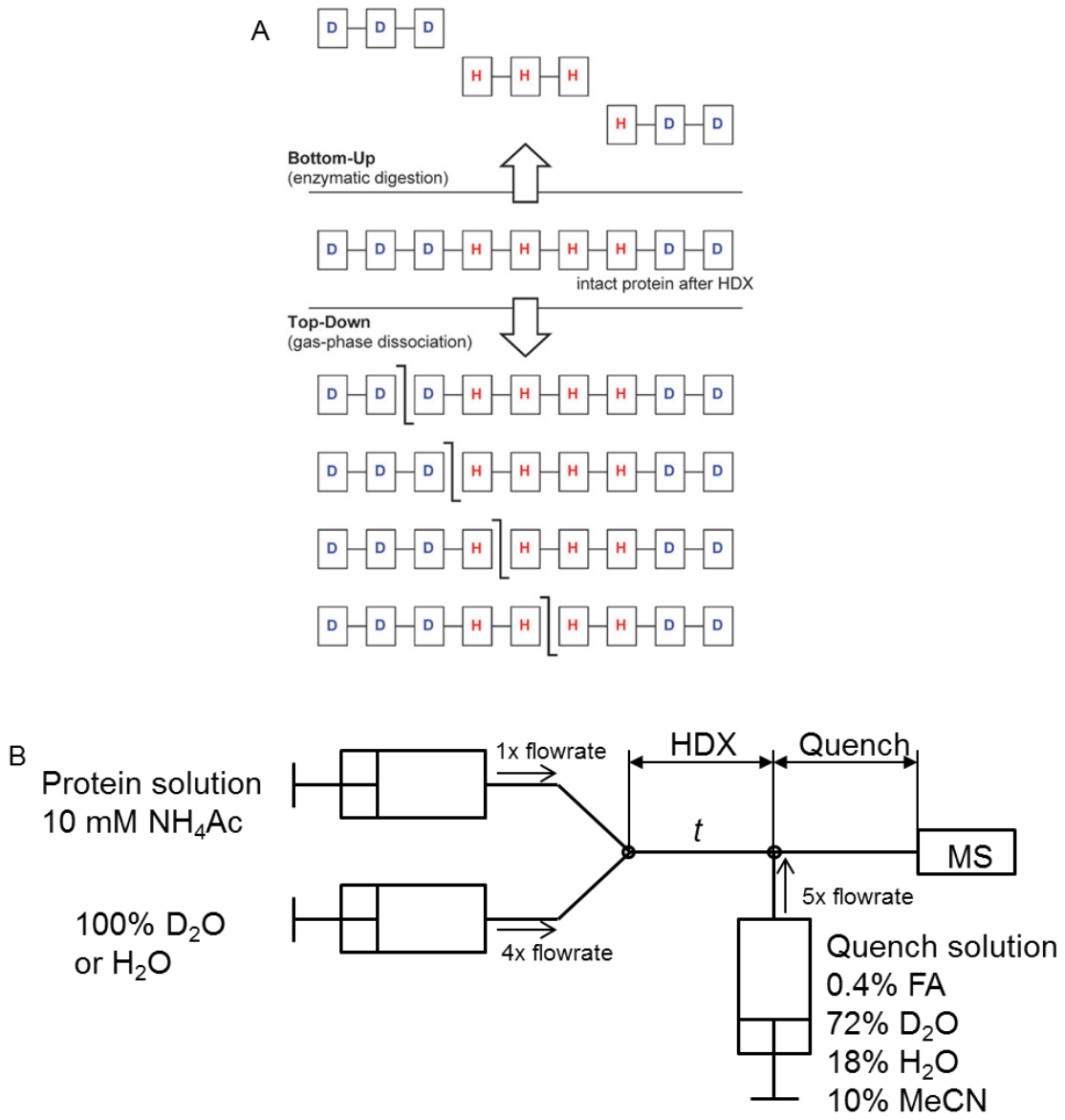


Figure 4: Hydrogen Deuterium Exchange scheme

A: Comparison between bottom-up and top-down HDX schemes. In a bottom-up experiment, peptide level resolution of HDX is achieved by enzymatically digesting the exchanged protein. In a top-down experiment, amino acid level resolution of exchange is provided by fragmentation of whole proteins in the gas phase. B: Schematic diagram of a typical top-down HDX setup.

A second T-junction introduces the quenching solution with the same H_2O:D_2O ratio.

This arrests any further exchange by reducing the pD such that exchange is minimized as

[53]. Additional organic components such as acetonitrile (ACN) can be added to the quench to assist with ionization and spray stability. The entire process takes place on-line with the mass spectrometer, with the outflow from the syringes leading to the ESI source for ionization and delivery to the mass spectrometer.

The resulting MS spectrum may then be used to determine the total deuteration of the protein ion. Such information is useful on its own, as changes to this value represent changes in the total number of protected hydrogen atoms of the whole protein.

This spectrum will typically include multiple charge states of the protein ion, at least one of which can then be selected and isolated for an MS/MS experiment. Selection of fragmentation type, however, must be more specific than for a typical experiment. The usual fragmentation method used for peptide ions is CID. However, this collisional process is too slow; during this time the molecule can undergoes scrambling, i.e., a rearrangement of the backbone amide hydrogen atoms. Unfortunately these are precisely the atoms that we wish to measure by HDX [59, 60]. Thus, other fragmentation methods must be used, such as ECD, ETD, or UVPD. These better methods result in more rapid formation of fragment ions, thereby minimizing scrambling.

HDX is used for accurate examination of changes in secondary structure as a result of protein conformational change [20]. Top-down HDX combined with fast fragmentation can then be used to isolate the particular regions of the protein in which this change occurred, often down to the residue level. These regions tend to be where important functions of proteins occur, such as binding sites for other proteins or small molecules, or as part of an enzyme's active site. Protein disorder in general, and changes in protein

structure related to protein activation or deactivation, are also a fruitful avenue for research, and can be examined using HDX techniques.

1.5. Surface modification for determining surface accessibility of residues

Surface modification of proteins is used to determine the relative exposure of different residues to the solvent, and, ultimately, to analyze the topology of a protein or protein complex using a small probe. These measurements provide a basis for comparison between different states of a protein, in order to determine whether a given residue is more or less exposed between these states [61]. Any event that changes the topology of a protein's surface can be measured [62-64]. This information can be used to determine which regions of the protein undergo structural changes as a result of whatever treatment has been performed.

In a typical surface-modification experiment, the light isotopic form of an isotopically-labelled probe will be added to a protein or system of proteins (Figure 5). Its corresponding heavy-labelled form will be added to that same protein when it is in a perturbed or modified state. This reaction labels the protein with the probe via covalent attachment of the probe to the protein. After the reaction has been quenched, the two samples are mixed at a 1:1 ratio, digested with a protease, and prepared for mass spectrometric analysis. Analysis usually uses a standard LC-MS/MS approach that emphasises the identification of as many peptides from the target proteins as possible. Identified peptides that contain either the covalently-bound heavy or light versions

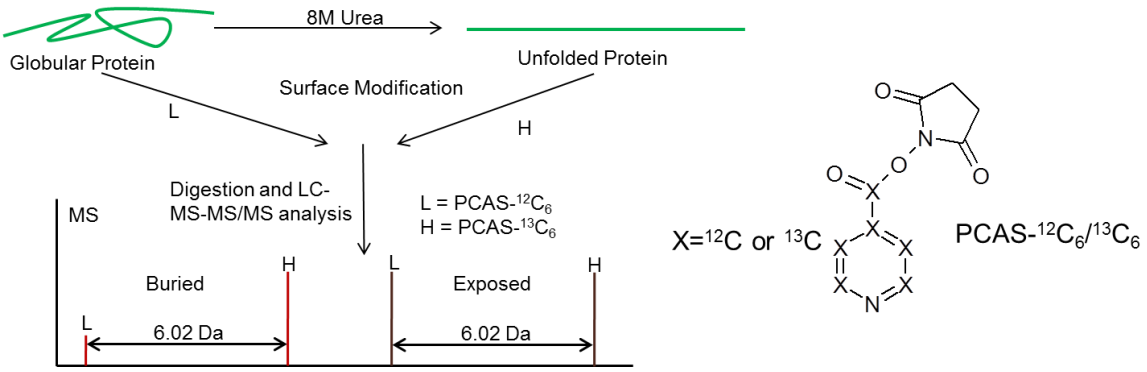


Figure 5: Schematic representation of a surface modification experiment using 8 M Urea

In this surface modification experiment, a protein unfolded with 8 M urea is compared with the native protein. If the native protein has a lesser degree of modification than the unfolded protein, we can determine that the modified residue is buried in the native structure.

of the probe can then be compared with their oppositely-labelled counterpart. This comparison usually takes place at the MS level, typically by comparing the relative intensity or the relatively peak area of the two labelled peptides.

In most cases, the easiest and most accessible protein chemistry to use for this probe is NHS-ester, which labels lysine residues. There are several variations of these reagents, including pyridine carboxylic acid succinimide (PCAS) [62] and the isotope-coded protein label (ICPL) reagent *N*-nicotinoyloxy-succinimide (Nic-NHS) [65], which include a variety of coding options as well as the potential for multiplexing using a variety of differently labelled variants of the same reagent. Lysines are typically surface-exposed because of their positive charge under most biological conditions. However, because lysines may not occur in regions of interest, the may not necessarily be useful for tracking changes in structure and surface accessibility. Probes targeting cysteine residues are also common, but are similarly restrictive given that cysteine residues are very often

involved in disulphide bonds, which limit their availability for reaction. Recently, some new probes have been developed which can target other residues, including oxidative labelling probes which target methionine and tryptophan [64] and photoreactive probes which may have the potential to target many more residues. In particular, these reagents can target the hydrophobic residues which make up the interior of a folded protein, and which may provide more contrast between different forms than probes which target lysine or cysteine.

Another useful tool for investigating disordered proteins in particular is to combine surface modification with deliberate unfolding of the protein in order to determine regions possessing residual structure. Many disordered regions of proteins retain some level of transient structure, and this transient structure may appear as weak protection from solvent in a surface modification experiment using the right strategy. In such cases, the comparison is between the protein under native-like conditions, and a protein that has been completely disordered through the application of some chaotropic agent such as urea or guanidinium hydrochloride. If the light form of the probe is used for the native protein, and the samples are mixed in a 1:1 ratio, then the ratio between the light and heavy forms of a modified peptide from the native vs. unfolded protein may be below a 1:1 L:H ratio (Figure 5). If the ratio is below 1, this indicates that the folded form of the protein has a lysine is at least partially more protected than when the protein is completely disordered.

1.6. Using structural proteomics data for assisting protein structure determination

The goal of applying any of the above techniques is to obtain a complete picture of a protein's structure – even under conditions which traditional structural techniques would not be able to provide sufficient detail. The data obtained from proteomics experiments must therefore be transformed, one way or another, into structural data. Structural proteomics data has generally been used to put other experimental data into a new context. This allowed the development of more useful models of proteins and their complexes. Structural proteomics was primarily used to provide additional data which could be used for validation or clarification of existing models. The additional data provided by crosslinking provides valuable evidence for modelling and docking of complexes by providing information on orientation or relative positioning within a complex. HDX may be used to indicate regions of the protein which were undergoing changes in structure during a binding process. Surface modification was used to label individual residues which may become protected from solvent as a result of complex formation. However, it has previously been difficult to generate models using structural proteomics data alone.

The complementarity of structural proteomics to other techniques is one of its great strengths that can, eventually be used to define an entire structure. Structural proteomics could provide a high-throughput method for structure determination, as the low requirements for protein amount per experiment and ease of automation of experiments and data analysis would result in shorter times for the modelling of diverse and difficult-to-handle proteins. Crosslinking is the most prominent technique for this purpose, as the distance constraints created by these experiments are analogous to the constraints generated by various NMR experiments, and can be used in a similar way. HADDOK,

for example, is an algorithm for protein docking originally based on NMR data, but which was easily adapted for use in the docking of structures based on crosslinking data [66]. This has been of great utility for modelling protein complexes, and has been used numerous times in the past to assist in that process.

More and more, crosslinking data is being used to refine or interpret data obtained from other more traditional experiments in order to generate new models of protein complexes. One example of a recent success of structural proteomics is the modelling of the RNA pol-II-Mediator core initiator complex [67]. The majority of the model was composed of well-resolved crystal structure and cryo-electron microscopy data. However, the middle module of the complex remained unresolved. Crosslinking of this complex allowed the determination of the topology of that module as well as its relative orientation within the overall complex. Crosslinking has also been used to examine a whole host of complexes including the anaphase promoting complex [2], the type-7 secretion system [68], and the yeast 19S proteasomal regulatory complex [69]. In each of these cases, crosslinking provided valuable constraints to be used for docking and for the construction of models of protein complexes.

There are two potential paths for trying to determine an individual protein structure using only structural proteomics data. The first is to use structural proteomics as a way to validate models generated by *de-novo* protein structure determination using algorithms such as Rosetta [22]. In this case, Rosetta is first used to generate tens of thousands of models of a protein. Then, in order to reduce the number of valid models, the crosslinking data is used to filter for the generated models, and the total number of potential structures is reduced. This can be repeated successively until a single valid structure is obtained.

This technique can also be used to model smaller portions of a protein at a time in order to reduce complexity. This particular strategy was used to model the structure of human serum albumin (HSA) using the sulfo-succinimidyl-diazerine (sulfo-SDA) crosslinker [21]. Here, Belsom *et al.* divided the HSA protein into three domains in order to divide the 576 amino acid residues of the protein into smaller sections that could be more easily modeled. They then used a total of 479 distance constraints to model the protein. The resulting model of HSA has a root-mean-square deviation (RMSD) of 2.9, 5.8, and 2.6 Å for each of the three domains.

This process does have limitations. Generating de-novo structures is extremely time-consuming. Modelling strategies that work for NMR constraints are not capable of producing models of proteins using the lower density of constraints produced by crosslinking experiments [21]. This is especially true as the number of amino acid residues to model becomes large. Even a 150 amino-acid protein modelled using this method requires an extraordinary amount of computation time. While there are a variety of methods for generating low-resolution structures more quickly, including discrete molecular dynamics (DMD) simulations amongst others, the ultimate goal of simulating any large protein or complex can be daunting.

This strategy has, been somewhat successfully employed previously in several Critical Assessment of protein Structure Prediction (CASP) contests. During the CASP 11 contest the target proteins were crosslinked using the photoreactive crosslinker sulfo-SDA [70]. This approach was able to generate approximately 0.6-1.2 crosslinks per residues depending on the protein structure. A total of 19 groups of modellers out of 146 total groups used the crosslinking data generated to assist in model refinement and validation,

including one group who used the crosslinking data to determine the correct orientation of the domains of the protein. However, at least some of the groups who did not utilize crosslinking data were still able to generate more accurate models of the target proteins. This indicated that refinements to the techniques used for incorporating crosslinking constraints into molecular dynamics (MD) simulations were clearly necessary, particularly when trying to model a single protein.

One approach to trying to refine and accelerate this modelling process for single proteins is to use the structural proteomics data explicitly within the modelling algorithm. If crosslinking constraints can be incorporated directly into the modelling process, there is great potential for accelerating the rate at which the protein approaches an energy minimum. Currently the *de novo* modelling of proteins with greater than 100 amino acids is severely limited by time and available computational power [71, 72]. It is difficult to sample such a large space in a practical period of time – thus any data which could restrict the search space would be useful for modellers. However, computational strategies can also be used to decrease the required computation time. For example, discrete molecular dynamics simulations model proteins using discretized energy functions rather than continuous functions could be used. This would dramatically reduce the number of computations required, and allow the calculation of substantially larger systems [73, 74]. These simulations can also readily be modified in such a way as to include crosslinking constraints directly into the energy function used in the simulation, thereby making crosslinking data explicit in the modelling process.

The success of modelling a protein using *any* set of constraints is dependent upon the number and length of these constraints. It is estimated that the number of pair-wise constraints needed to model a given protein fold should be approximately 1 constraint per 10 residues [23]. At this density, it should be possible to determine the correct protein fold using crosslinking constraints alone. Modelling a protein at very high resolution requires a much higher density of crosslinks [75]. Most crosslinking reagents utilize NHS esters to reliably generate intra-protein crosslinks. This is inhibited by the low number of lysine residues on a single protein. It may therefore be difficult to generate crosslinks at the density required to fully map the protein with no additional information from modelling efforts. Therefore, several additional strategies may be necessary in order to obtain a sufficient number of distance constraints. The easiest method might be to perform the crosslinking and quenching as normal, and then to maintain the protein at an acidic pH during the digestion process. Normally NHS esters are capable of reacting with lysine as well as serine, threonine, and tyrosine. These last three reactions are slowly reversible under the normal conditions used for protein digestion (i.e., pH 8.0, overnight time course) [76]. If the digestion is instead maintained at an acidic pH, this reaction reverses more slowly, and crosslinks can be found even after digestion [76].

The best solution would be to develop non-specific crosslinking reagents that can react with any residue. The most useful chemistries for such reagents are those that utilize a variety of radicals capable of crosslinking aliphatic residues via proton abstraction. Three such chemistries have previously been found to be effective: aryl azide reactions that utilize the photolysis of azyl nitrogens to generate nitrene radicals [77, 78], diazirine reactions resulting in photolysis that yields a reactive carbene species [25], and finally

benzophenone reagents which utilize ultraviolet (UV)-absorbing benzophenone groups capable of repeated excitation and relaxation [79]. Some of these reagents do have preferences for particular residues. In particular, the benzophenones are known to have a preference for the terminal methyl group of methionine [80], and aryl azides can undergo a ring expansion that leaves them open to nucleophilic attack, principally by lysine [81]. Diazirine reactions, however, appear not to have particularly strong preferences for any individual amino acid [21, 25]. Despite these preferences, all of these photoreactive groups are capable of reacting non-specifically and are thus capable of increasing the density of crosslinking constraints in an individual protein molecule.

1.7. Parkinson's Disease

Parkinson's disease (PD) is characterized by the death of dopimanergic neurons in the substantia nigra, a region of the midbrain important for controlling movement. It is one of several neurodegenerative diseases referred generally to as Lewy body disorders (LBD). Other diseases in this category include dementia with Lewy bodies (DLB), and multiple system atrophy (MSA). In each of these cases, the formation of proteinaceous plaques accompanies the death of the associated neurons [82, 83]. The primary differentiating factor between each of these diseases is generally the location of the primary region where cell death is occurring.

In Parkinson's disease, cell death occurs predominantly in the substantia nigra [82]. This region of the brain is deeply involved in dopamine-induced inhibition of motor signals, and it is the death of these neurons which lead to the specific symptoms of Parkinson's disease. More specifically, projections from the substantia nigra pars compacta travel to the dorsal putamen of the striatum [84], and it is the loss of these

projections which result in the disease symptoms, particularly the bradykinesia and the rigidity experienced by most individuals with Parkinson's disease [84].

Parkinson's disease was first identified in the early nineteenth century by James Parkinson who recorded the first detailed analysis of the disease, then referred to as paralysis agitans [85]. He conducted several case studies, and for the first time linked these diseases to damage that he observed in the medulla. Observing the connection between damage to the lower regions of the brain and the pathology [85]. It would take another 100 years for the first observation of proteinaceous plaques as a potential cause of the disease. In 1912, Fritz Heinrich Lewy isolated aggregates from the brains of patients with Parkinson's disease which he correctly identified as being amyloid protein by noting its similarity to the corpora amylacae, another amyloid protein deposit found throughout the body [86]. The association with synuclein was made in 1997, by Spillantini and collaborators. They used an anti-synuclein antibody to demonstrate the presence of this protein in Lewy bodies, which provided evidence that Lewy bodies were primarily composed of synuclein aggregates [87]. There are a variety of other genes related to Parkinson's disease, often those related to protein degradation, such as *LRRK2* and *Parkin* (Figure 6) [84].

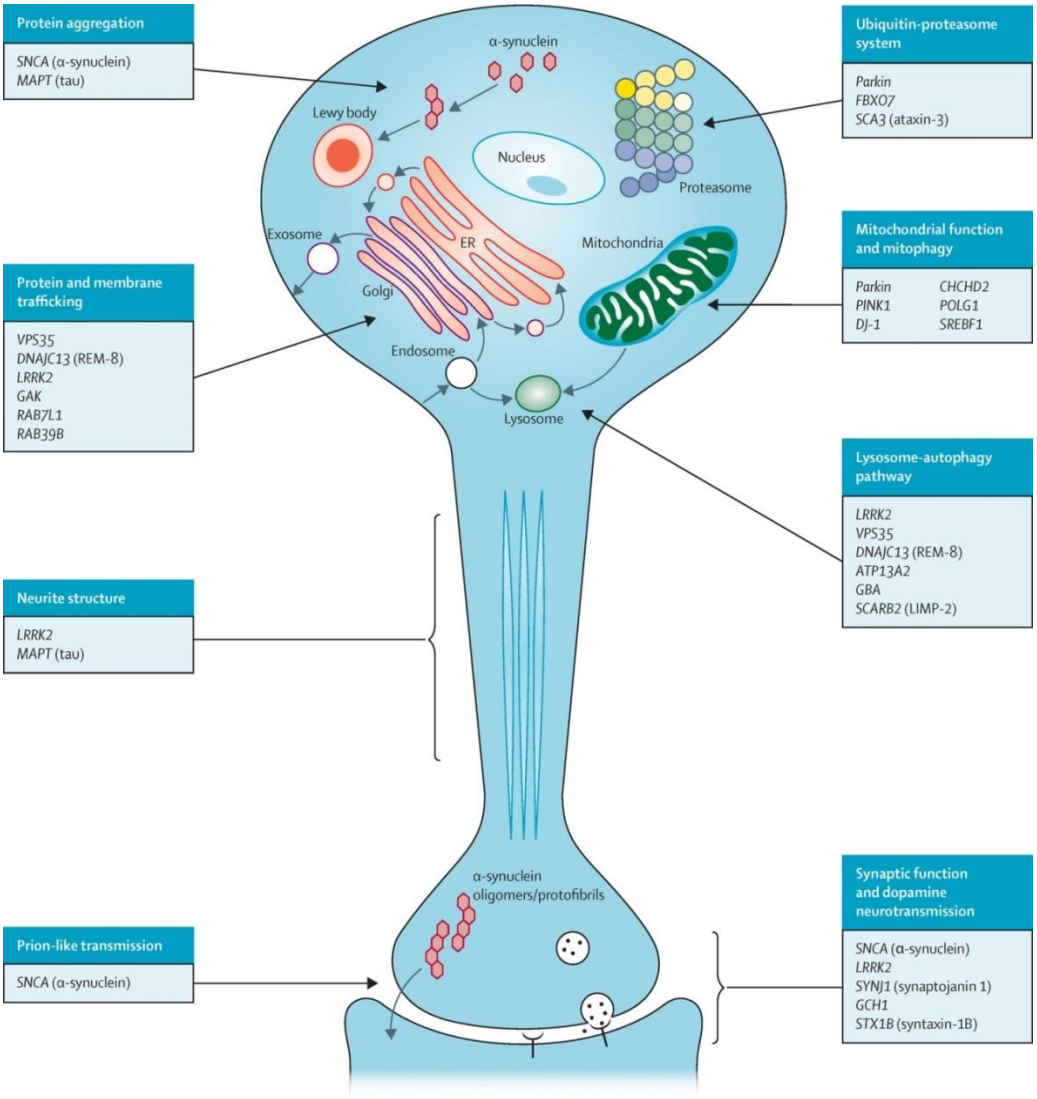


Figure 6: Genes and cellular pathways implicated in Parkinson's disease

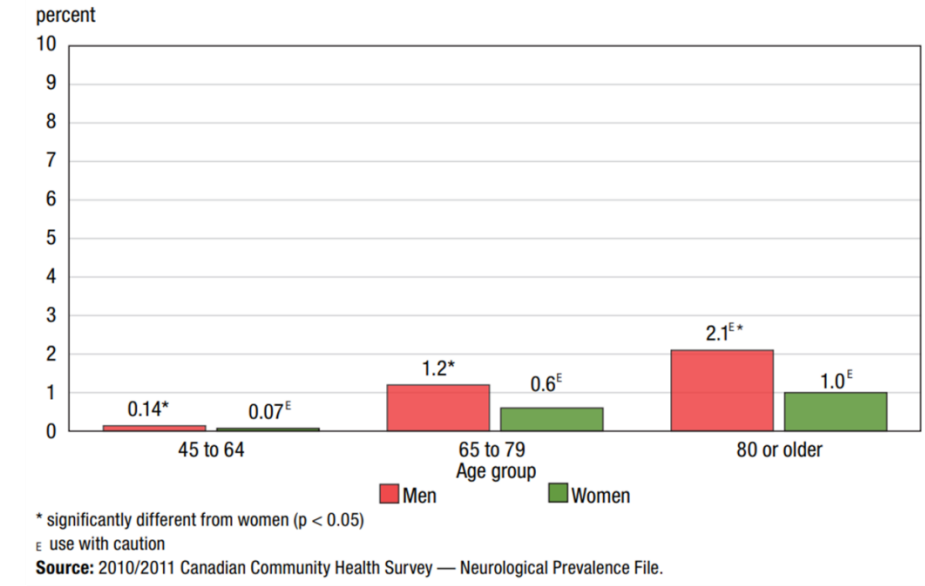
While α-synuclein remains a key gene responsible for Parkinson's disease, several other genes have been found to be risk factors, including LRRK2 and Parkin. Adapted from Kalia and Lang, 2015 [84].

Parkinson's disease affects nearly 1 in 500 Canadians, and approximately 1 % of all individuals over the age of 65. This number is expected to grow as the average age of the population increases [88]. The first diagnosis of Parkinson's is typically at around the age of 65, and thus, as the population ages, we will see an increase in Parkinson's disease. Patients with Parkinson's disease typically live for many years subsequent to diagnosis.

While the disease itself is certainly progressive, the onset of symptoms and the full course of degeneration is a long process. As it typically occurs late in life, most patients will succumb to other diseases before the effects of Parkinson's result in specific disease-related mortality. Parkinson's disease has a disastrous impact on quality of life. Many individuals with the disease incapable of maintaining themselves independently, and require significant levels of additional care. Approximately 10 % of individuals in assisted living facilities in Canada have been diagnosed with Parkinson's disease [88] (Figure 7).

Currently, there is no treatment for Parkinson's disease that targets the root causes of the disease. All current treatments alleviate only the symptoms of the disease. A typical patient with Parkinson's disease is treated with levodopa, a precursor of dopamine that is capable of crossing the blood brain barrier, where it is then converted into active dopamine. Other dopamine agonists such as Ropinirol are also prescribed. Both of these strategies alleviate symptoms by increasing the supply of dopamine in the brain, counteracting the loss of dopaminergic neurons [88]. Although this treatment is able to overcome some of the symptoms of the disease, due to the progressive nature of Parkinson's disease an individual's condition will inevitably worsen until no intervention provides relief.

A



B

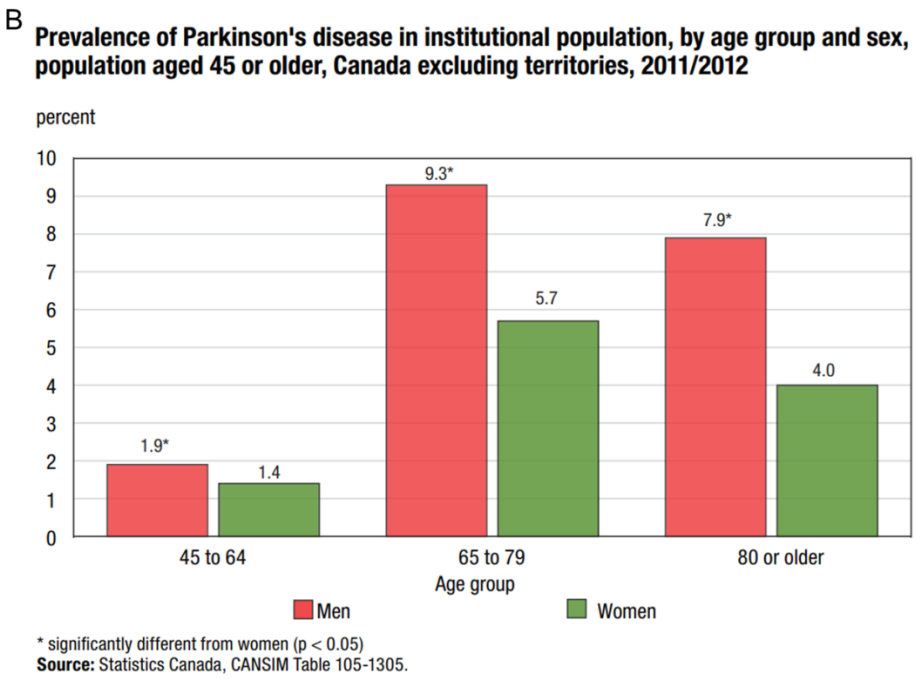


Figure 7: Prevalence of Parkinson's disease among the at home and institutional populations of Canada

A: Prevalence of Canadians aged 45 and older with Parkinson's disease, segregated by sex. B: Prevalence of Parkinson's disease among Canada's institutional population, aged 45 and older, segregated by sex. Adapted from Wong *et al.*, 2014 [88].

A variety of newer treatments are being proposed which may extend the period of time that symptoms can be effectively managed. Surgically-implanted deep-brain stimulation electrodes are another option for those patients with mid-stage Parkinson's or those patients for whom the side effects from chemical methods of disease stabilization are severe [89].

1.8. α-Synuclein

The first association between the synuclein protein and neurodegenerative disease was the discovery of a 35-amino-acid peptide from α-synuclein, consisting of residues 61-95 of the full length protein. It was sequenced from plaques of Aβ found in patients suffering from Alzheimer's. Uéda *et al.* [90] discovered the cDNA associated with this protein component, and they named the peptide the non-amyloid β component (NAC). The full length cDNA would later be determined to be the product of the α-synuclein gene (*SNCA*), the protein α-synuclein [91]. This protein was soon found to aggregated in a number of other diseases, most notably the Lewy body disorders [87]. The connection between synuclein and Alzheimer's disease was not lost, although subsequent studies have tended not to implicate synuclein in the development of this disease.

The full length α-synuclein protein can be broken down into essentially 3 regions. The N-terminal region of residues 1-60 is positively charged and highly repetitive. It consists of numerous lysine-threonine-lysine (KTK) repeats [90] (Figure 8). The next region, the NAC region, consists of residues 61-95, and is highly hydrophobic. Finally, residues 96-140 contain a region of negative charge consisting of primarily acidic residues, although there is a short region of positive charge at the beginning of this region, consisting of

lysines 96, 97, and 102. These unusual qualities combine to make synuclein an intrinsically disordered protein.

α-Synuclein belongs to the synuclein gene family, and is one of three such genes found in humans [92]. The others are β-synuclein and γ-synuclein. The sequences and expression patterns of these two genes are very similar to α-synuclein but a key difference is in the NAC region. This region is missing 11 amino acids in β-synuclein (Figure 8). As a result, β-synuclein does not aggregate spontaneously [93], unlike α-synuclein. These 3 proteins are most highly expressed in nervous tissue, and α-synuclein alone makes up nearly 1% of the total amount of protein in these cells [94]. These proteins are also very soluble owing to their highly charged nature, which makes the inherent aggregation of α-synuclein at least somewhat mysterious.

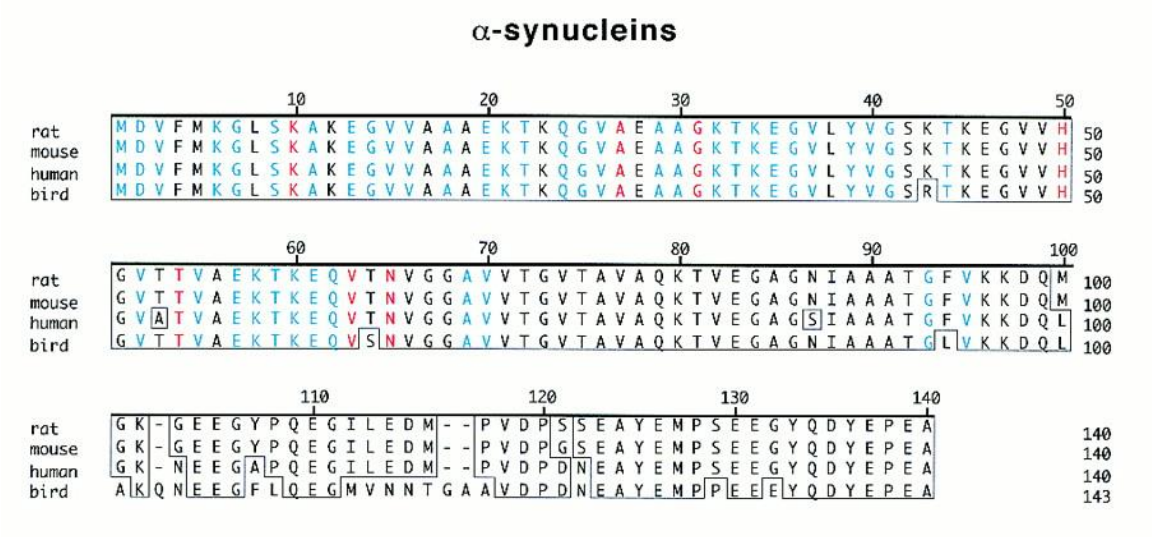


Figure 8: Alignment of α-synuclein sequence between rat, mouse, human and bird. Residues in blue are conserved among all proteins in the synuclein family including β- and γ-synuclein. Residues in red are unique to α-synuclein. Adapted from Lavedan, 1998 [92].

1.9. α-Synuclein structure and function

Synuclein is generally considered to be an intrinsically disordered protein. Proteins with intrinsic disorder do not adopt a singular globular fold, unlike most other proteins. Many proteins will contain regions of intrinsic disorder, but synuclein is one of a much smaller group of proteins which under native conditions do not have even a single well-ordered domain [95]. Intrinsic disorder can serve a number of unique purposes in proteins, often related to promiscuity in binding or recognition. In the case of synuclein, however, it is not particularly clear what purpose its intrinsic disorder serves. The lack of clarity here may be at least partly related to the lack of clarity regarding the purpose of synuclein in the cell. The clearest case for synuclein's function can be made at the surface of membranes. When bound to the surface of membranes or micelles, synuclein adopts a helical conformation [96] (Figure 9); this can be either a single long helix in the case of membrane surfaces with lower curvature [97], or a pair of helixes folded back on each other [96]. In both cases, the C-terminal portion of the protein (residues 90-140) remains disordered. The purpose of this ordering remains unclear. Synuclein does, however, gather at the pre-synaptic terminal of neurons, interacting strongly with vesicles in this region [98, 99], which does suggest a role in vesicle trafficking. Synuclein has also been shown to interact with a number of proteins within these regions, including acting as a chaperone for the SNARE protein synaptobrevin-2 [100]. This interaction, as well as synuclein's presynaptic localization and strong interactions with vesicular membranes support this hypothesis. However, these potential functions are at odds with several other observed interactions, including α-synuclein's interaction with tyrosine hydroxylases [101, 102], in which it acts to inhibit their activity and expression levels.

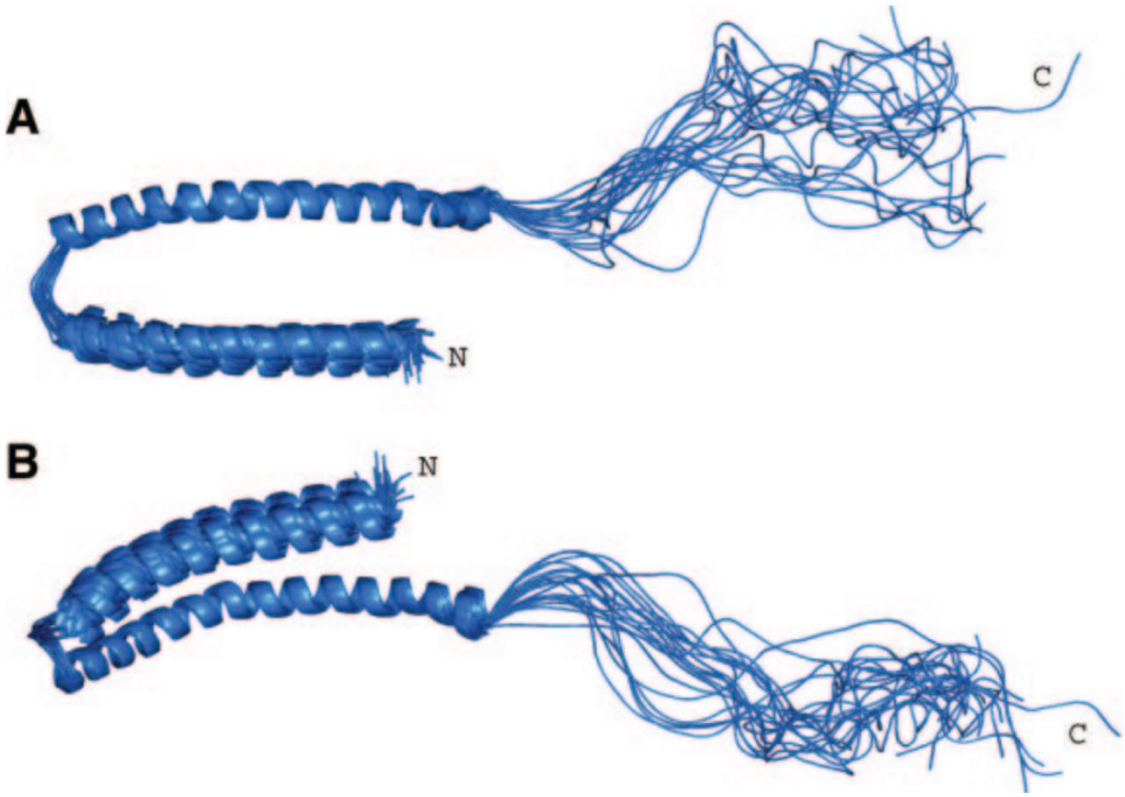


Figure 9: Ensemble structure of micelle bound α-synuclein based on NMR data.

A: Micelle-bound α-synuclein adopts an α-helical structure. The first helix spans residues 3-37, followed by a short linker, then the second helix spans residues 45-92. The C-terminal remains disordered. B: 120° rotation of A around the x-axis. Adapted from Ulmer *et al.*, 2004 [96].

All of these attempts to understand synuclein's functions are also undermined by the fact that synuclein does not appear to be strictly necessary for any of these functions. Knockouts of α-synuclein in mice show no apparent dysfunction [103], as whatever function α-synuclein performs is compensated for by the remaining β- and γ-synuclein [104]. Triple knockouts however do display a distinctive phenotype including reduced SNARE-complex formation, impaired vesicle trafficking, shortened longevity, and lower dopamine production [100]. These factors ultimately suggest a role in the presynaptic terminal, even if it is unclear precisely what that role is.

1.10. α-Synuclein Misfolding and Disease

Despite the dearth of information on synuclein's actual function within the cell, there is one consistent activity for which the protein is responsible: misfolding and the generation of disease-causing amyloid aggregates. Shortly after the discovery that synuclein was an important cause in the generation of Lewy bodies various attempts were made to characterize the misfolded form of the protein. It had long been known that the protein component which made up Lewy bodies was in an amyloid fibril configuration consisting of layers of β-sheets. These could be readily identified by thioflavin-T staining (ThT). α-Synuclein was identified as the amyloidogenic component of the aggregates soon after its discovery [105, 106].

There was conflicting data on the exact cause of both the misfolding event, as well as conflict over what exactly the most toxic form of the protein was. Fibrils themselves are highly stable and relatively inert, and it is unlikely that they represent the most toxic form of the protein. Instead, it was hypothesized that pre-fibrillar oligomeric forms of the protein might be responsible for the death of the affected neurons [107-109].

The genesis and subsequent spread of the disease was also unclear, although recent evidence indicates that misfolded synuclein represents the first true "prion" protein since the discovery of the first prion protein (PrP) [110]. It has since been demonstrated by Pruisner et al., that misfolded synuclein material contains all of the information necessary for the seeding of new misfolded protein and the spread of the disease [111]. Structures of the mature synuclein fibril have also been determined recently using solid state NMR and cryo-electron microscopy [112, 113]. The synuclein protein adopts a unique "reverse Greek key" fold consisting of several beta sheets from residues 42-95 in a flat stacking pattern (Figure 10).

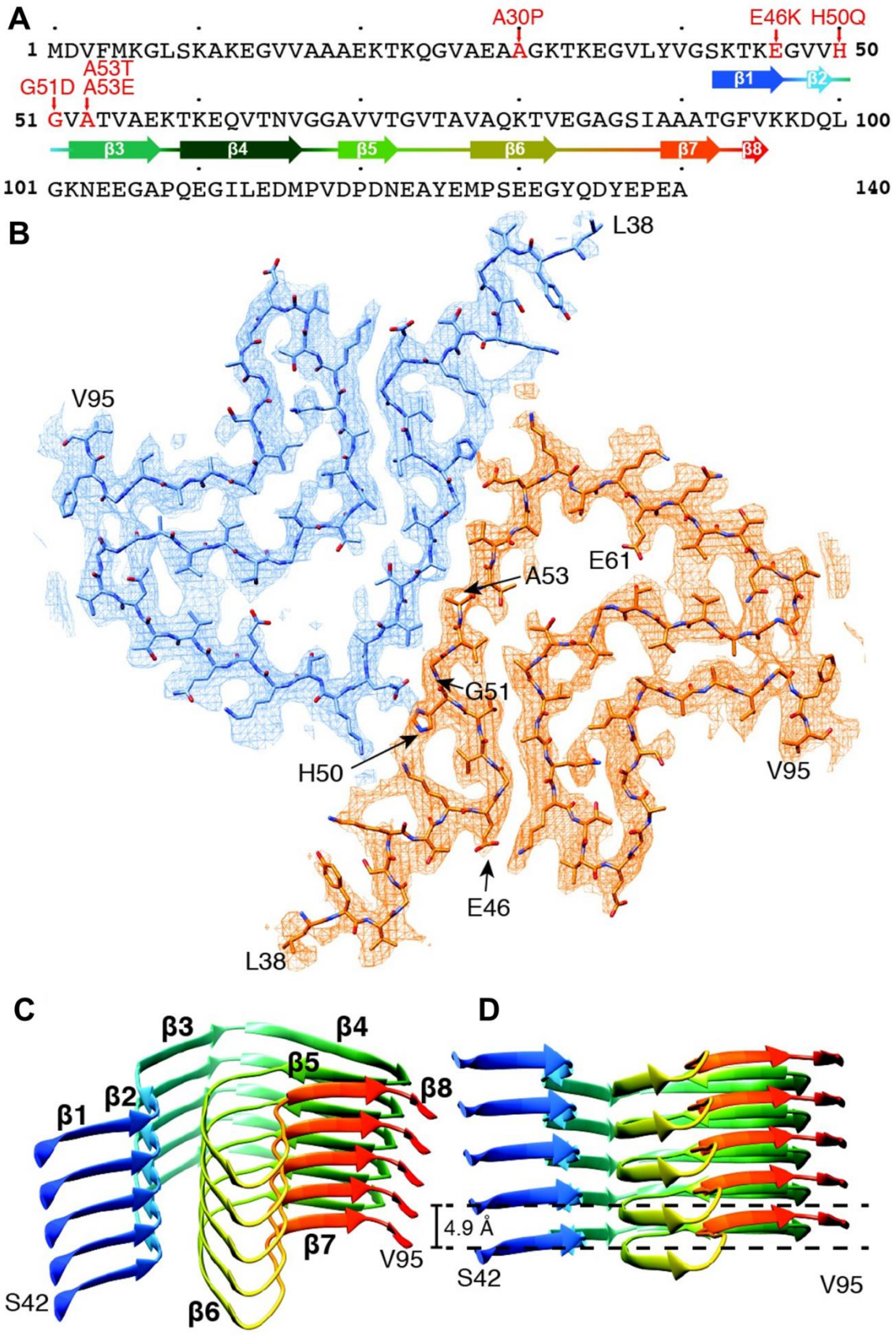

A
A30P
E46K H50Q
1 MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVH 50
A53T
G51D A53E
β1
β2
51 GVATVAEKTKEQVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDQL 100
β3
β4
β5
β6
β7
β8
101 GKNEEGAPQEGILEDMPVDPDNEAYEMPSEEGYQDYEPEA 140
B
L38
V95
E61
A53
G51
H50
V95
E46
L38
C
β3
β4
β5
β8
β2
β1
β7
V95
S42
β6
D
4.9 Å
S42
V95

Figure 10: Structure of the α-synuclein fibril core determined by cryo-electron microscopy.

A: Sequence of α-synuclein. Highlighted in red are residues with common synuclein mutations. Color coded in blue to red are the locations of the β-sheets. B: Cross-sectional structure of the synuclein fibril. The two protofilaments, in blue and orange, interact primarily through residues outside of the NAC region (H50-K58). C: A single filament of α-synuclein. The β-sheets from each monomer stack in-parrallel to form the fibril. D: Measurement of the height of the fibril repeats, showing some variation in inter-monomer distances along the vertical axis. Adapted from Guerrero-Ferreira *et al.,* 2018 [113].

These flattened molecules are then able to stack and form the overall structure of the fibrils. The pattern of beta sheets within these structures include not just the NAC region, but also the approximately 20 amino acids of the protein preceding it. In particular, it's been observed in cryo-em strucutres of mature fibrils (Figure 10) that residues outside of the NAC region, particularly residues H50-K58, are responsible for stabilizing the interaction between the two protofilaments in a fibril with a 2_1 screw-symmetry. It should be noted, however, that this is just one potential strain of synuclein fibril. In the solid-state NMR structure of synuclein fibrils published by Tuttle *et al.* the synuclein fibrils were of approximately half this diameter, and were determined to consist of a single stranded fiber.

There are several competing theories with regard to the neurotoxicity of the synuclein oligomers or fibrils. The first of these is the potential for synuclein oligomers to form pore-like structures on membrane surfaces [114, 115]. These pores then lead to an influx of Ca^{2+} ions, which kills the cells. There are also some groups who believe that a similar mechanism is responsible for mitochondrial dysfunction present in some cases. Another leading theory is that α-synuclein misfolding results in a misfolded protein endoplasmic-

reticulum stress response [116]. While this initially results in the production of chaperones and other misfolding stress proteins, ultimately apoptosis will be triggered.

Whatever the mechanism is that leads to cell death, it is clear that the misfolding event which transforms normal, native protein into an oligomeric form is a key process in the development and propagation of the disease [111]. A likely course of events is that native, molten globule synuclein protein is induced by other misfolded oligomers to adopt the misfolded conformation. This process likely involves the exposure of the NAC region of the protein, which will form the core of the synuclein fibril. In order to assist in modelling this process, a detailed ensemble of the native synuclein protein should be generated to serve as a starting point. Several attempts have been made previously, all utilizing NMR or FRET [117-119]. None of these studies provided detailed atomic-level descriptions of the synuclein ensemble, which is a necessary starting point for simulations of the potential misfolding mechanism.

1.11. Hypothesis and approach

The primary hypothesis of my dissertation is that the native α-synuclein protein adopts an ensemble of states in solution, the majority of which will provide some degree of protection for the NAC region from solvent. This ensemble will be determined using a combination of protein crosslinking and discrete molecular dynamics simulations. Intra-protein crosslinks detected by mass spectrometry can be interpreted as distance constraints to be used in these simulations. In order to obtain a sufficient number of crosslinking constraints, it was critical to develop new non-specific, isotopically-coded photoreactive crosslinking reagents. The development of these reagents is detailed in Chapter 2: Development of isotopically-labelled photoreactive crosslinkers for use in

structural proteomics. It was also critical to establish a protocol for the inclusion of this data into an algorithm for discrete molecular dynamics simulations, and to demonstrate that these simulations are capable of modelling proteins. This will be discussed in Chapter 3: Crosslinking combined with discrete molecular dynamics simulations for determining protein structures. Finally, a combination of crosslinking, surface modification, and HDX were performed on the α-synuclein protein, and the crosslinking constraints were used to generate ensemble models of the synuclein protein by their incorporation into discrete molecular dynamics simulations of the protein structure. These experiments are discussed in Chapter 4: Determination of an ensemble structure of the native synuclein protein using crosslinking and discrete molecular dynamics simulations. The synuclein ensemble was found to occupy four clusters of approximately equal energy. Each had a level of secondary structure which correlated with the total HDX observed, and in each case, the NAC region of the protein was protected from solvent in accordance with surface modification data, or in some cases, additionally stabilized by transient secondary structure.

Chapter 2: Development of new photoreactive crosslinkers for use in studying protein structures

Work in this chapter was performed exclusively in the lab of Dr. Christoph Borchers. The synthesis of the crosslinking reagents was performed by Dr. Evgeniy Petrotchenko. The crosslinking of the test peptides and the proteins RNase A and α-synuclein was performed by Nicholas Brodie. Crosslinking of the Spy-Im7 protein was performed by Karl Makepeace. All of the LC-MS/MS experiments and the data analysis thereof was performed by Nicholas Brodie. Experimental design was performed by Nicholas Brodie, Evgeniy Petrotchenko, and Christoph Borchers. Christoph Borchers oversaw this project.

This chapter was adapted from the publication:

Brodie, Nicholas; Makepeace, Karl; Petrotchenko, Evgeniy; Borchers, Christoph. **Isotopically-coded short-range hetero-bifunctional photo-reactive crosslinkers for studying protein structure**. Journal of Proteomics. Volume 118, pp. 12-20. April 6, 2015. **Doi:** 10.1016/j.jprot.2014.08.012

2.1. Introduction to Photocrosslinking

Chemical crosslinking is a useful technique for obtaining structural information on proteins and protein complexes which are difficult to characterize by other methods. The crosslinking reaction creates covalent linkages between two amino acid residues, and the subsequent enzymatic digestion and mass spectrometric analysis allows identification of the crosslinked sites. These crosslinks provide distance constraints which can be used in the molecular modelling of protein structures [23, 120]. In order to increase the resolution of crosslinking data for this purpose, shorter-distance crosslinks are required on the order of 5-8 Å. Crosslinkers traditionally use NHS-ester chemistry, which is specific for primary amino groups. However, the availability of two lysines that can be crosslinked diminishes as the distance is reduced, and thus new strategies are required for studies involving short-distance crosslinking.

To mitigate this problem, non-specific crosslinking chemistry can be used. Photo-reactive groups – such as phenyl azide, benzophenone, and diazirines – generate radicals upon stimulation by UV light, and these radicals react in a non-specific manner with amino acid residues (Figure 11), typically by insertion of the radical into a methyl or methylene group [28]. By combining these groups with NHS-ester chemistry, heterobifunctional crosslinkers can be produced, which offer short-range crosslinking capabilities not available with homo-bifunctional NHS-ester reagents. Three such crosslinkers are azido-benzoic-acid-succinimide (ABAS), carboxy-benzophenone-succinimide (CBS), and succinimidyl-diazirine (SDA) which utilize phenyl azide, benzophenone, and diazirine photo-reactive groups, respectively. Use of these reagents has been a common strategy in protein crosslinking for many years [27, 28, 78, 121, 122], primarily for the purpose of conjugating a protein or peptide to another molecule in a

photo-activatable manner. These three crosslinking reagents result in short-distance crosslinks of ~5 Å or 7 Å in length, as measured from the nitrogen atom of the newly formed amide from the NHS-ester reaction to the atom of the target residue modified by the photo-reactive group (Figure 11).

One of the major limiting factors in a crosslinking experiment is the small number of crosslinked peptides compared to their non-crosslinked counterparts [31]. This problem is

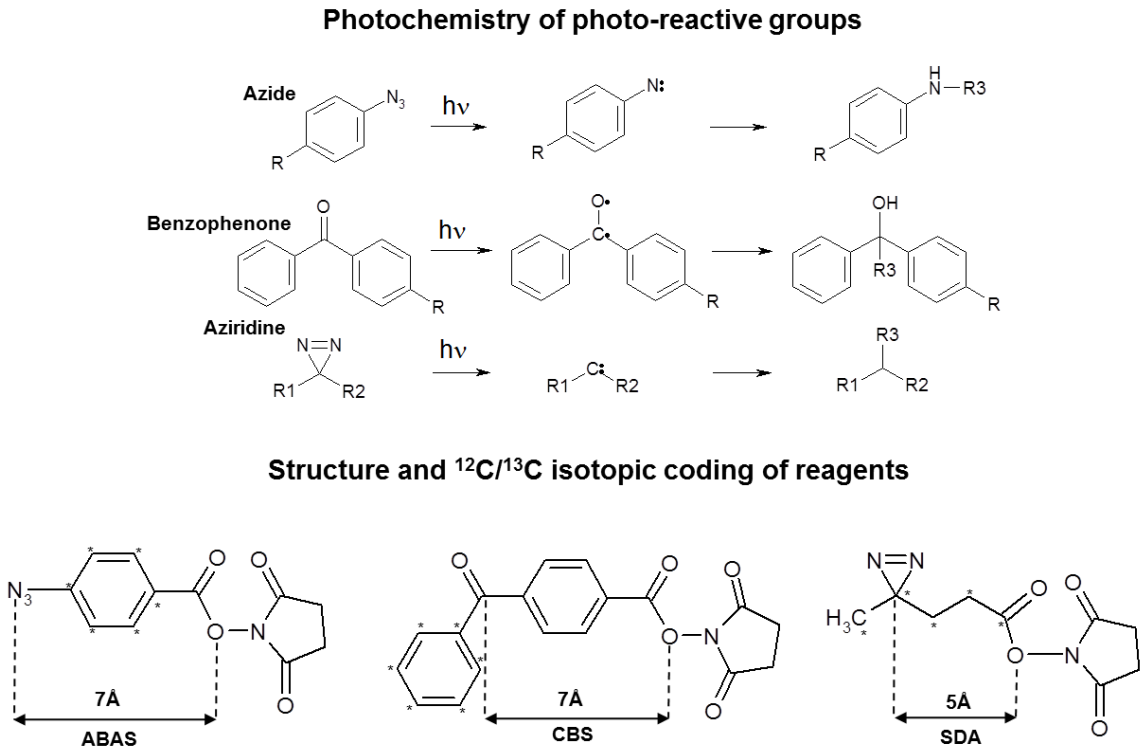


Figure 11: Photochemistry of reactive groups chosen for new isotopically-labelled crosslinkers and their labelling strategy

A. Photochemistry of three photo-reactive functional groups [28]. B, Structure, isotopic coding, and spacer length of the three photo-reactive, hetero-bifunctional, isotopically-coded crosslinkers ABAS-$^{12}C_6/^{13}C_6$, CBS-$^{12}C_6/^{13}C_6$ and SDA-$^{12}C_5/^{13}C_5$.

worsened in the case of non-specific crosslinkers, which may form multiple products, each of which further dilutes the signal of any one crosslinked species. The low signal intensity of these crosslinks makes them difficult to detect using traditional intensity based data-dependent MS/MS acquisition methods. Isotopic coding of the crosslinking reagents, however, generates a distinctive doublet signature in the mass spectra that can be utilized to improve the probability of acquiring an MS/MS spectrum for those peptides which have been modified by the crosslinker [123]. By ensuring that both the heavy and light precursor ions are fragmented simultaneously, the MS/MS fragment ions also reveal doublet signatures which can also be utilized for increased confidence in the crosslink assignment [15].

In order to initially assess the capabilities of these crosslinkers, they were applied to three model systems: 1) a test peptide with sequence Ac-TRTESTDIKRASSREADYLINKER, 2) the protein RNase S, and 3) the protein-peptide complex Spy-Im7. These crosslinkers can be used for the structural determination of misfolded proteins, including α-synuclein which is involved in the development of Parkinson's disease and in other Lewy body diseases [82, 124]. The transition from native α-synuclein to the misfolded and aggregated form is still poorly understood at the molecular level. [125, 126]. Non-specific photocrosslinking of α-synuclein can be used to obtain information about the conformations the native structure can adopt, and can elucidate important interactions within the protein.

2.2. Materials and Methods

All chemicals were obtained from Sigma-Aldrich, unless noted otherwise.

2.2.1 Crosslinker synthesis

ABAS-$^{12}C_6$/$^{13}C_6$. 4-bromobenzoic acid was converted to 4-aminobenzoic acid as in reference [127], 4-azidobenzoic acid was obtained as in reference [128], and activated with N-hydroxysuccinimide (NHS). Briefly, 0.1 mmol of an equimolar mixture of $^{12}C_6$- and $^{13}C_6$ -4-bromobenzoic acid with 2 mg of copper in 0.5 mL 28 % ammonia was stirred at 100°C overnight. The reaction mixture was acidified with 6 M HCl and product was recovered by partitioning with ethyl acetate. The 4-aminobenzoic acid obtained from the organic layer was resuspended in water and kept on ice under stirring while adding concentrated HCl and equimolar amounts of 3.3 M solution of $NaNO_2$ in water for 30 minutes, and 1.32 M solution of NaN_3 in water for 90 minutes. Volumes of water and ethyl acetate equal to the volume of the reaction mixture were added, and the organic layer was washed with 1 M NaOH. The aqueous layers were acidified, partitioned against ethyl acetate, the organic layer was collected and dried *in vacuo*. The 4-azidobenzoic acid obtained was activated with equimolar amounts of NHS and N,N'-dicyclohexylcarbodiimide (DCC) in dimethylsulfoxide (DMSO) overnight, the reaction mixture was then filtered, the product was precipitated with water, and dried *in vacuo*.

CBS-$^{12}C_6$/$^{13}C_6$. 4-methylbenzophenone was obtained from p-toluoylchloride and benzene as previously described [129], oxidized to 4-carboxybenzophenone as previously described [130] and activated with N-hydroxysuccinimide. Briefly, 1 mmol p-toluoylchloride, 2 mmol of benzene, 2 mmol of $^{13}C_6$-benzene (Cambridge Isotopes Laboratories), and 2 mmol $AlCl_3$ were stirred for 3 hours at 65 °C, the product was precipitated with water, and dried *in vacuo*. One-hundred and five milligrams of 4-

methylbenzophenone in 500 μL acetic acid was oxidized upon addition of CrO_3 in acetic acid:water:H_2SO_4 (150 mg:500 μL:340 μL:105 μL) for 1.5 hours, after which the reaction mixture was supplemented with 4 mL of acetic acid and diluted with water. The precipitate was dissolved in 1M NaOH, re-precipitated by acidification with concentrated HCl, washed with water, and dried *in vacuo*. The 4-carboxybenzophenone obtained was activated with equimolar amounts of NHS and DCC in DMSO overnight, the reaction mixture was filtered, and the product was precipitated with water and dried *in vacuo*.

SDA-$^{12}C_5$/$^{13}C_5$. Levulinic acid was obtained from glucose as previously described [131], converted into 3-(3-methyl-3-diazirinyl)propanoic acid as described previously [121, 132], and activated with N-hydroxysuccinimide. Briefly, 1 mmol of an equimolar mixture of $^{12}C_6$- and $^{13}C_6$-glucose (Cambridge Isotopes Laboratories), 17 mg $AlCl_3$, and 10 mL of water were stirred in a sealed ampoule for 2 hours at 170-200 °C. The reaction mixture was cooled, extracted three times with ethyl ether, and the extracts were dried *in vacuo*. Levulinic acid was dissolved in methanol, and ammonia gas was bubbled through the solution for 2 hours while it was kept on ice. One and seven-tenths equivalents of hydroxylamine-O-sulfonic acid was added and the reaction mixture was stirred on ice for 2 hours, filtered, and dried *in vacuo*. The 3-(3-methyl-3-diaziridinyl)propanoic acid was oxidized with CrO3 in acetone:20% H2SO4 for one hour at 25°C, the aqueous phase was extracted with chloroform, and the organic phase was dried *in vacuo*. The diazirinyl-pentanoic acid obtained was activated with equimolar amounts of NHS and DCC in DMSO overnight, the reaction mixture was filtered, partitioned with chloroform:water and the organic phase was dried *in vacuo*.

2.2.2. Crosslinking of proteins and peptides.

Protein and peptide samples were typically incubated for 15 minutes with the crosslinker in the dark to allow for NHS-ester reaction, followed by 15 minutes of UV irradiation under a 25W UV lamp (Model UVGL-58 Mineralight lamp, UVG), with the wavelength set as appropriate for each crosslinker (254 nm for ABAS and CBS, 366 nm for SDA). The lamp was held at a distance of 2 cm from the sample. The reactions were quenched with NH_4HCO_3 at a final concentration of 10 mM.

The model peptide Ac-TRTESTDIKRASSREADYLINKER (from Creative Molecules, Inc.) was prepared at a concentration of 0.5 mM in PBS buffer (137 mM NaCl, 2.7 mM KCl, 8.0 mM Na_2HPO_4, 2.0 mM NaH_2PO_4) at pH 7.4. A 20-µL aliquot was then treated with 1 mM ABAS in the dark at room temperature. The sample was digested with 0.2 µg proteinase K for 30 minutes at 37°C, and digestion was halted with 4-(2-aminoethyl)benzenesulfonyl fluoride hydrochloride (AEBSF) at a final concentration of 20 mM. The sample was then acidified with 0.1% formic acid for mass spectrometric analysis.

Spy protein and Im7 peptide were mixed at 26 µM each in 20 mM Na_2HPO_4, pH 7.4. These were then crosslinked with either 100 nM ABAS, 1 mM SDA, or 1mM CBS. Samples were digested with either proteinase K for 60 minutes at 37°C, or with trypsin for 18 hours at 37°C. Digestion was halted by adding 500 mM AEBSF to give a final concentration of 10 mM, and the sample was reduced with 25 mM DTT and acidified.

RNase S was prepared at a concentration of 77 µM in HPLC-grade H_2O, pH 6.5, and crosslinked with 0.75 mM CBS, 0.5 mM ABAS, or 0.1 mM SDA. An initial 30-minute incubation for the NHS-ester reaction was followed by a 60 minute incubation with 254

nm UV light as above for ABAS and CBS, or a 10-minute incubation with 366-nm light for SDA. Samples were digested and prepared as for Spy-Im7.

α-Synuclein was obtained from a pET-21a vector graciously provided by Dr. Ladner of the Wishart group at the University of Alberta, and was expressed and purified as previously described [96]. Briefly, the protein was overexpressed in 1L LB cultures of BL21DE3 E. coli for 4 hours at 30 °C. Cells were lysed with a French press and the lysate was heated at 70 °C for 10 minutes and then centrifuged at 14000g for 30 minutes. The soluble fraction was precipitated for 1 hour in 2.1 M $(NH_4)_2SO_4$. The α-synuclein was then purified by fast protein liquid chromatography on a Mono Q 4.6/100 SAX column (GE Life Science). Purified α-synuclein was diluted to 50 μM in 20 mM Na_2HPO_4 buffer and was then crosslinked with 1 mM ABAS as above. The crosslinked protein was then digested with either proteinase K for 30 minutes at 37 °C, or with trypsin for 18 hours at 37 °C, and then reduced and acidified as above.

2.2.3. Mass spectrometry analysis of crosslinked peptides

Samples were prepared for HPLC separation by acidifying with 0.1 % TFA. Mass spectrometric analysis was then performed using a nano-HPLC system (Easy-nLC II, ThermoFisher Scientific), coupled to the ESI-source of an LTQ Orbitrap Velos (ThermoFisher Scientific), using conditions previously described [15].

Briefly, samples were injected onto a 100 μm ID x 360 μm OD trap column packed with Magic C18AQ (Bruker-Michrom, Auburn, CA) 100 Å particles, 5 μm pore size (prepared in-house), and desalted by washing with Solvent A (2 % acetonitrile:98 % water, both 0.1 % formic acid (FA)). Peptides were separated with a 60-min gradient (0-60 min: 4–40 % solvent B (90 % acetonitrile, 10 % water, 0.1 % formic acid (FA)), 60–

62 min: 40-80 % B, 62-70 min: 80 % B), on a 75 μm ID x 360 μm OD analytical column with an IntegraFrit (New Objective Inc., Woburn, MA). The column was packed with Magic C18AQ 100 Å particles, 5 μm pore size (prepared in-house) and equilibrated with solvent A. MS data were acquired using the data-dependent Mass Tags method of Xcalibur software, incorporating a Δ mass specific to the isotopic coding of the crosslinker (ABAS and CBS Δ mass: 6.02013, 3.01006, 2.00671, 1.50503, 1.20403; SDA Δ mass: 5.01677, 2.50839, 1.67226, 1.25419). The Mass Tags acquisition instructs the instrument to search MS data on the fly looking for pairs of monoisotopic peaks matching these mass deltas, then schedules both peaks for MS/MS acquisition. The data-dependent acquisition (DDA) also utilized dynamic exclusion, with an exclusion window of 10 ppm and exclusion duration of 60 seconds. MS and MS/MS events used 60000 and 30000 resolution Fourier transform mass spectrometry (FTMS) scans, respectively, with a scan range of m/z 400-2000 in the MS. For MS/MS, CID collision energy was set to 35%. Data were analyzed using the DXMSMS Match program from the ICC-CLASS software package [133]. For scoring and assignment of the MS/MS spectra, b and y ions were primarily used, with additional confirmation from CID-cleavage of the peptide bond formed by the NHS-ester reaction with lysine.

2.3. Results and Discussion

In order to assess the efficacy and amino acid reactivities of the photo-reactive crosslinkers ABAS-$^{12}C_6/^{13}C_6$, CBS-$^{12}C_6/^{13}C_6$ and SDA-$^{12}C_5/^{13}C_5$, several model systems were used: 1) a test peptide with sequence Ac-TRTESTDIKRASSREADYLINKER, 2) the protein/peptide complex Spy-Im7, and 3) the protein RNase S. The crosslinking reactions were allowed to proceed in the dark first, in order to initially react the NHS-

ester, followed by irradiation with UV light (254nm in the case of CBS and ABAS, 366nm in the case of SDA). Samples were then digested and analyzed on an Orbitrap Velos. The DXMSMS Match program from ICC-CLASS was used to analyze the resulting MS and MS/MS data [133].

2.3.1. Evaluation of SDA

The SDA crosslinker was evaluated on the test peptide, RNase S, and on the Spy-Im7 complex. While doublets usually numbered around one hundred in the Spy-Im7 and RNase S samples, and forty in the test peptide samples, all identified crosslinks were found to be either a unique type of dead end crosslink, or a more typical dead-end crosslink which included the addition of water. These unique SDA reaction products had masses indicating the loss of N_2 and formation of the carbene radical with subsequent insertion into -CH_2-, but analysis of the MS/MS data revealed that the modification resided entirely upon one lysine residue in the peptide examined. This result indicated that the carbene radical had inserted into the same lysine that had already reacted with the NHS-ester, producing a cyclization of the crosslinker on the lysine residue. Dead-end crosslinks resulting in the addition of water have two pathways for formation, indistinguishable by mass – they may result from NHS-ester hydrolysis followed by carbene-insertion into -CH_2-, or from the addition of water to the carbene after an NHS-ester reaction with lysine. Since these dead ends occurred on lysine, it seems probable that they represent the latter case.

2.3.2. Evaluation of ABAS

Three model systems were used to validate the ABAS crosslinker: 1) a test peptide, 2) RNase S, and 3) Spy-Im7. Six inter-peptide crosslinks were found for the test peptide.

Four of these were crosslinks between K9 and K22, but a single crosslink was discovered from E23 to K22, as well as a crosslink from S13 to K9. These crosslinks had masses indicative of either of two proposed mechanisms: 1) formation of a nitrene radical followed by proton abstraction from the target amino acid and formation of a bond with the crosslinker, or 2) ring expansion followed by nucleophilic attack on the ring. Both of these reactions would result in products with the same mass [77]. These initial experiments showed no clear preference for the targets of this reaction, with S13, E23, K9, and K22 all being targets of the crosslinker (Table 2). Dead-end crosslinks from this data set indicated that photolysis of the azide results in the formation of either a simple primary amine or in the incorporation of water, both of which are supported by the literature [77]. This incorporation of water is indistinguishable by mass from the result of an initial hydrolysis of the NHS-ester followed by a nitrene insertion. The majority of the dead ends were found on lysine residues, however, so which reactive group was actually incorporated could not be determined.

In the RNase S dataset, a total of five inter-peptide crosslinks were found. Two of these crosslinks were from the N-terminus to K41, and another was found between K41 and K91. The remaining two crosslinks were non-specific, and crosslinked N24 to K31, and Q60 to K104. All of these crosslinks matched the crystal structure of RNase S within the 17Å maximum range of the ABAS crosslinker.

In the Spy-Im7 datasets, a trend of crosslinker reactivity began to emerge. The Spy protein is a chaperone homo-dimer containing a binding pocket in which the peptide Im7 is bound [134]. There were a total of 11 crosslinks found in the Spy-Im7 system, six of which were found between the N-terminus of the Spy protein and other residues in Spy.

Five of the 11 crosslinks were either between two lysines or between the N-terminus and a lysine, while the remaining six crosslinks were between alanine, valine, phenylalanine, or glycine. The occurrence of crosslinked products for both the nucleophilic lysine ε-amino group and a collection of hydrophobic residues indicates that both nitrene insertion into a methylene or methyl group, and ring expansion followed by addition of a lysine ε-amino group are viable pathways for crosslink formation, and that there is a tendency to favor ring expansion. Structurally, most of the crosslinks were found between the flexible N-terminus of the protein and several residues within the ordered core. Those crosslinks which were found within the core of the protein and within the Im7 peptide were in good agreement with the reported crystal structure [135].

2.3.3. Evaluation of CBS

The CBS crosslinker was tested using a test peptide, RNase S, and Spy-Im7. These were tested using the same method of crosslinking, data collection, and data analysis as was used for the ABAS crosslinking of Spy-Im7. In the test peptide, no definitive crosslinks were identified. However, three masses corresponding to crosslinks were detected on RNase S (Table 2). Two of these were lysine to methionine crosslinks (K1-M13), and one was a lysine to lysine crosslink (K91-K98). All 3 were consistent with the X-ray crystal structure of RNase S [136]. All three of these crosslinks were found with a mass

Table 2: Table of inter-peptide crosslinks detected using new isotopically labelled photoreactive crosslinkers

Inter-peptide crosslinks detected in several proteins and peptides using the photo-reactive crosslinkers ABAS-$^{12}C_6/^{13}C_6$ and CBS-$^{12}C_6/^{13}C_6$. In addition to the systems depicted here, crosslinking with CBS and SDA in TP1, crosslinking of Spy-Im7 with SDA and CBS, and crosslinking of RNase S with SDA were also attempted, but these yielded no crosslinks.

System	Crosslinker	Mass	ppm error	Sequence	Sequence #	Modified aa	Sequence	Sequence #	Modified aa
TP1	ABAS	923.4576	1.2	(K)ER(-)	23-24	E23	(L)INKE(R)	20-23	K22
TP1	ABAS	941.4679	1.4	(S)SR(E)	13-14	S13	(E)STDIK(R)	5-9	K9
TP1	ABAS	1111.5743	0.4	(N)KER(-)	22-24	K22	(E)STDIK(R)	5-9	K9
TP1	ABAS	1338.7108	1.5	(N)KER(-)	22-24	K22	(E)STDIKRA(S)	5-11	K9
TP1	ABAS	1425.7439	0.8	(N)KER(-)	22-24	K22	(E)STDIKRAS(S)	5-12	K9
TP1	ABAS	2260.1167	1.9	(A)SSREADYLINKER(-)	12-24	K22	(E)STDIK(R)	5-9	K9
Spy	ABAS	1557.6977	0.3	(H)KG(K)	19-20	K19	(K)FGPHQDMMFK(D)	22-31	F22
Im7	ABAS	2761.3912	0.9	(K)EIEKEN(V)	16-21	K19	(N)VAATDDVLDVLLEHFVK(I)	22-38	V22
Spy	ABAS	1378.6309	-0.1	(T)AAPADAKPM(M)	7-15	K13	(P)ATAE(-)	136-139	A136
Spy	ABAS	1581.7756	0.3	(-)SADTTTAAPADAK(P)	1-13	N-term	(K)VK(A)	77-78	K78
Spy	ABAS	1582.7616	0.4	(A)PADAKPM(M)	9-15	K13	(M)MHHKGK(F)	16-21	K19
Spy	ABAS	1695.7926	0.7	(-)SADTTTAAPADAK(P)	1-13	N-Term	(K)GQR(D)	49-51	G49
Spy	ABAS	1895.915	1.2	(-)SADTTTAAPADAK(P)	1-13	N-term	(A)KGKMP(A)	131-135	K133
Spy	ABAS	1906.9497	0.7	(-)SADTTTAAPADAK(P)	1-13	N-term	(R)PAAKGK(M)	128-133	K131
Spy	ABAS	1923.8866	0.7	(-)SADTTTAAPADAKPM(M)	1-15	N-term	(K)GQR(D)	49-51	G49
Spy	ABAS	2151.0573	-1.3	(R)AMHDIIASDTFDK(V)	64-76	A64	(R)PAAKGK(M)	128-133	K131
Spy-Im7	ABAS	1853.8761	0.4	(-)SADTTTAAPADAK(P)	1-13	N-term	(K)EIEK(E)	16-19	K19
α-Synuclein	ABAS	1489.754	0.5	(-)MDVFMK(G)	1-6	N-Term	(K)GLSKAK(E)	7-12	K10
α-Synuclein	ABAS	1539.7491	-0.9	(-)MDV(F)	1-3	N-Term	(K)TKQGVAEAAGK(T)	22-32	K23
α-Synuclein	ABAS	1699.8365	0	(-)MDVF(M)	1-4	N-Term	(K)AKEGVVAAAEK(T)	11-21	K10
α-Synuclein	ABAS	1729.8464	0.4	(-)MDVF(M)	1-4	N-Term	(K)EGVVAAAEKTK(Q)	13-23	K21
α-Synuclein	ABAS	2100.1072	0.8	(K)KDQLGK(N)	97-102	K97	(K)EGVVHGVATVAEK(T)	46-58	H50
α-Synuclein	ABAS	2278.2044	-0.1	(K)EGVVAAAEKTK(Q)	13-23	K21	(K)QGVAEAAGKTK(E)	24-34	K32
α-Synuclein	ABAS	945.4140	1.2	(-)MVDF(M)	1-4	N-Term	(L)GKN(E)	101-103	K102
α-Synuclein	ABAS	861.3806	1.2	(-)MDVF(M)	1-4	N-Term	(L)SK(A)	9-10	K10
α-Synuclein	ABAS	905.3895	0.7	(-)MDVF(M)	1-4	N-Term	(F)MK(G)	5-6	K6
α-Synuclein	ABAS	1077.507	0.9	(-)MDVF(M)	1-4	N-Term	(T)GFVK(K)	93-96	K96
RNase S	CBS	1231.4927	0.8	(-)KETA(A)	1-4	K1	(Q)HMDSS(T)	12-16	M13
RNase S	CBS	1261.5017	2	(-)KET(A)	1-3	K1	(Q)HMDSST(S)	12-17	M13
RNase S	CBS	1942.8862	-0.5	(R)ETGSSK(Y)	86-91	K91	(N)CAYKTTQANK(H)	95-104	K98

indicating the formation of the ketyl radical by hydrogen abstraction from the target methyl or methylene group, followed by formation of the alcohol. In the case of dead ends, the NHS-ester was typically able to react with a lysine, and the benzophenone remained stable and unreacted throughout the UV irradiation period and the sample preparation procedure. Some of the detected dead ends did indicate the addition of water as well. These could result from hydrolysis of the NHS-ester, and benzophenone insertion. The fact that many of the latter type of dead-end crosslinks were found on methionine further indicates the preference of the crosslinker for this specific amino acid.

In addition to the abundant dead ends, there were a large number of unidentified doublets. The CBS crosslinked peptides tended to fragment poorly, despite the high intensity of many of the doublets. This hindered the MS/MS analysis of many potential crosslinks to the point that the true identity of the crosslink could not be determined. The poor quality of the MS/MS spectra acquired from the Spy-Im7 and test peptide samples also played a role in the negative results in these systems. In the Spy-Im7 complex, no crosslinks were detected which could be unambiguously verified based on the MS/MS spectrum. It is also probable that the lack of methionine in the test peptide contributed to the negative results observed in that system.

2.3.4. Comparison of SDA, ABAS, and CBS

The diazirine group in SDA is highly reactive, but its tendency to produce dead ends is a problem which might be able to be addressed by limiting the mobility of the photo-reactive group (possibly by including a rigid spacer such as a benzene ring in the structure, similar to ABAS). CBS-crosslinked peptides exhibited poor fragmentation under CID. This led to difficulties in verifying the identity of the crosslinked peptides, as

well as in confidently assigning the location of the crosslink. Additionally, all of these crosslinkers might benefit from the addition of an affinity group for enrichment of crosslinks, such as biotin. Unfortunately, the use of an avidin-enrichable version of the ABAS crosslinker, which we called cyano-biotin-azido-propionyl-succinimide (CBAPS), resulted in complications with the crosslinking reaction, including a photo-reactive reaction with the affinity tag. All these considerations need to be taken into account in future designs of "next generation" photo-reactive crosslinkers for structural proteomics applications.

In summary, of the three photo-reactive crosslinkers, ABAS proved to be the most effective. The use of this crosslinker allowed confident assignment of crosslinks with a variety of targets for the photo-reactive group, including alanine, valine, phenylalanine, glycine, lysine, glutamic acid serine, glutamine, and asparagine (Table 2). Therefore, we chose to apply this crosslinker for studying the native structure of α-synuclein as the first "real world" sample.

2.3.5. Crosslinking of α-synuclein using ABAS

α-Synuclein is an apparently disordered protein, which makes studying the possible interactions in its native structure difficult to perform by crystallography or NMR. The best model available is derived from a micelle-bound NMR structure, in which the only ordered region is an arrangement of two anti-parallel alpha helices [96]. In order to study the structure of this protein, recombinantly expressed α-synuclein was crosslinked using the ABAS photocrosslinker, and the resulting crosslinked products were analyzed by mass spectrometry. Crosslinked samples showed the formation of a small amount of dimer product by SDS-PAGE. Crosslinked protein was digested in-solution by either

trypsin or proteinase K. The resulting peptide digests were separated by RP-HPLC and analyzed on the Orbitrap Velos using the Mass Tags method. This analysis yielded a total of 10 inter-peptide crosslinks (Table 2) with confidently assigned MS/MS spectra indicating the crosslinked residues (Figure 12).

When mapped onto the NMR structure, nearly every one of the ABAS crosslinks was longer than 20 Å, a distance incompatible with the 15 Å maximum length of the ABAS crosslinker (Figure 13). One crosslink in particular, H50-K97, had a distance on the structure of 74 Å. H50 is located within the beginning of the second alpha helix of the protein, and its crosslinking to K97 presents a particular problem in that it necessitates the unfolding of the second alpha helix in order for these two residues to be brought within the 17 Å distance of the crosslinker. Crosslinks between the N-terminus and K21 and K23 are also at a long range based on the NMR structure, 31 and 32 Å, respectively.

In order to satisfy the observed distance constraints, these two crosslinks require that the first alpha helix be deformed as well. A third pair of significant crosslinks between K98 or K102 and the N-terminus had a distance of 36 Å and 46 Å, respectively, in the structure. In order to satisfy these distance constraints, the two alpha helices would have to be moved substantially closer to each other; alternatively, the loop between the two helices would have to be more compact in order to accommodate this crosslink.

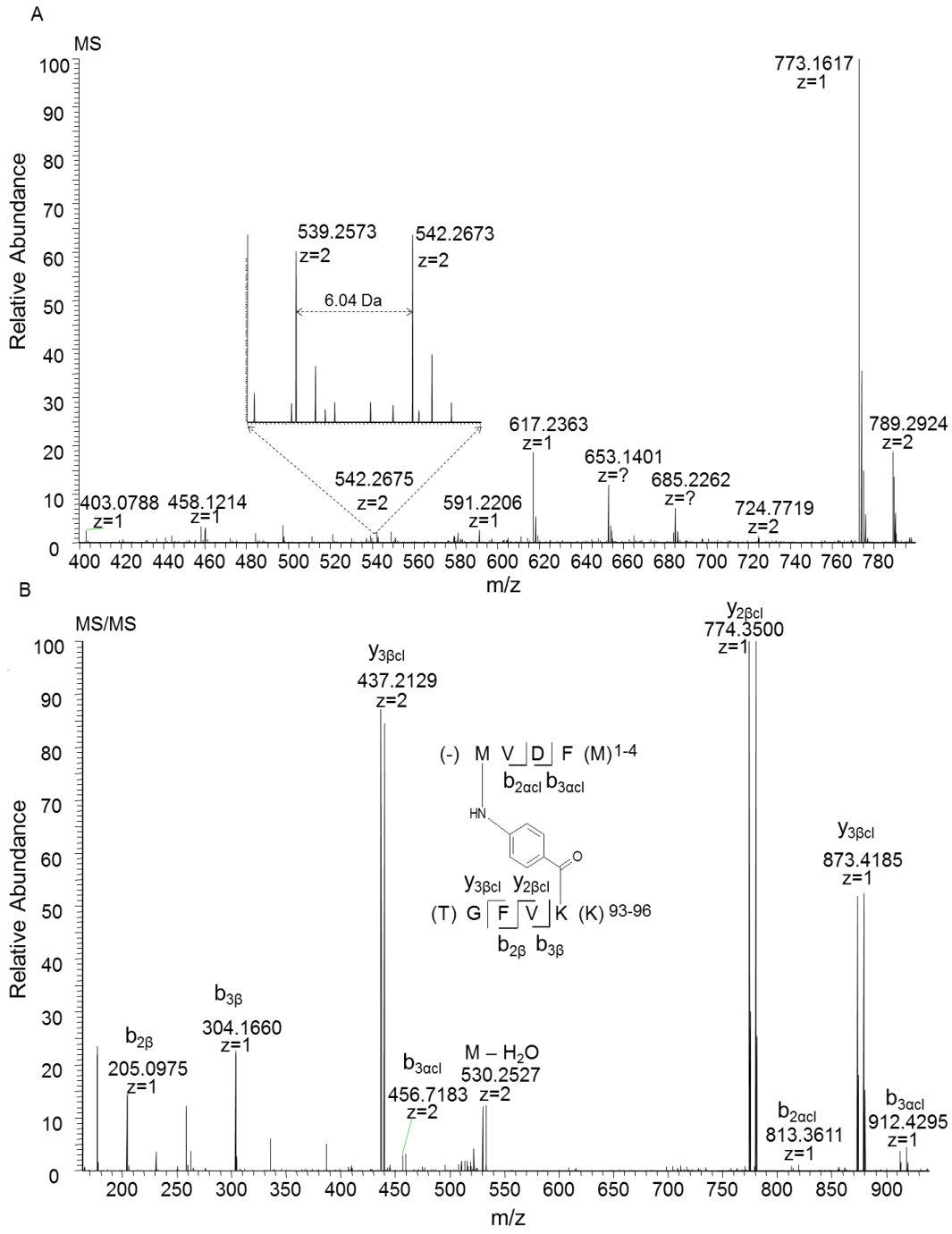


Figure 12: MS and MS/MS spectrum of ABAS crosslink

A. MS spectrum of ABAS crosslinked and proteinase K digested α-synuclein, with an observed crosslink mass at m=539.2573, z=2. ABAS-$^{12}C_6$/$^{13}C_6$ crosslink appears as a doublet of signals 6.02 Da apart (Inset). B. Merged MS/MS spectra of the m/z 539.2573 and 542.2673 precursors. The b and y fragment ions containing the crosslinker moiety appear in the spectrum as a 6.02 Da doublet, confirming the identity of the crosslink and facilitating the identification of the crosslinked amino acid residues.

In order to provide a comparison with the data obtained with the photoreactive crosslinker, α-synuclein was also crosslinked using the isotopically-labelled homo-bifunctional NHS-ester crosslinker cyanurbiotin-dimercaptopropionyl-succinimide (CBDPS) [18]. Thirty-one crosslinks were found using this crosslinker. Many of these crosslinks were found between the N-terminal region of the protein and K96/97, as well as between some of the closely situated lysine residues including K21-K32/34, K34-K45, and K96-K102. Compared to CBDPS, the non-specific reactivity of ABAS proved to be advantageous because it allowed crosslinking of regions of the protein that were unable to be crosslinked by traditional NHS-ester reagents, and also provided shorter (i.e., tighter) distance constraints.

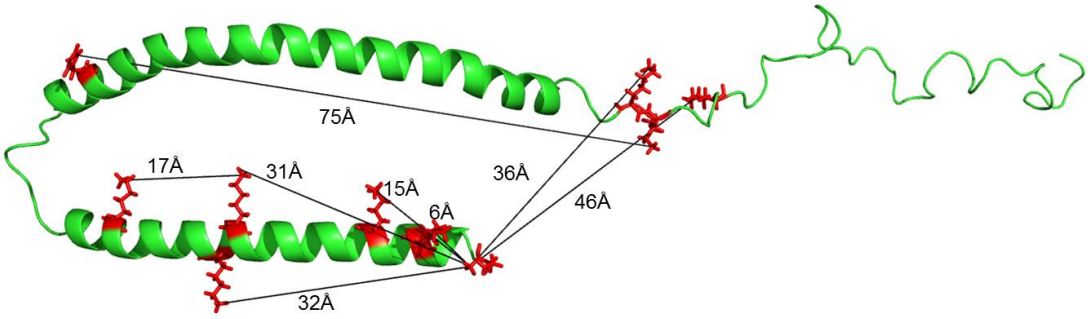


Figure 13: α-synuclein ABAS crosslinks

The ABAS crosslinker has a length of only 5Å, and crosslinks found are incompatible with the current NMR structure [96]. Short-distance crosslinks indicate multiple interactions between several regions of α-synuclein.

Overall, this crosslinking data indicates that the in-solution structure of α-synuclein may be significantly different from that of the micelle-bound structure of monomeric α-synuclein. There are three possibilities for how this difference can be explained. If all of

the crosslinks are assumed to exist within one single conformer, then the structure must be highly compact. These crosslinks may also, however, represent a heterogeneous mixture of possible conformations, each contributing one or more crosslinks. Finally, the observed crosslinks may also be of either inter- or intra-protein origin. In order to dissect these possibilities, we are currently applying a $^{14}N/^{15}N$ crosslinking strategy [42]. Whichever the case, these crosslinks point toward regions of the α-synuclein molecule which may interact with each other. These interactions may play a role in the conversion of the protein from its native to the mis-folded, aggregated state, and may also assist in characterizing the structural transitions necessary for this transformation.

2.4. Conclusions

Three new isotopically-coded short-range hetero-bifunctional photo-reactive crosslinkers have been presented here: ABAS-$^{12}C_6/^{13}C_6$, CBS-$^{12}C_6/^{13}C_6$, and SDA-$^{12}C_5/^{13}C_5$. Each crosslinker was designed to take advantage of non-specific reactivity and isotopic labelling to allow for the detection of short-distance crosslinks with a more diverse range of reactivities than NHS-ester alone. Isotopic labelling allowed for confident detection of the low-intensity non-specific crosslinks generated by these reagents, which would most likely have been missed by traditional intensity-based data-dependent MS/MS acquisition methods. When both the light and heavy isotopic forms of the crosslinks are fragmented, characteristic isotopic doublets of the fragments in overlaid MS/MS spectra of light and heavy forms increase the confidence of the identification of the crosslinking sites. The combination of short range, broad reactivity, and isotopic labelling allowed the detection of crosslinks using both ABAS and CBS in well-established systems such as RNase S, providing evidence for the usefulness of these

crosslinkers as structural probes. Finally, the crosslinker ABAS was used for crosslinking studies on the neurodegenerative disease related protein, α-synuclein. Crosslinking in α-synuclein indicated a potential for a multiple interactions within the structure that may be important for the understanding of the folding of the native structure as well as for the mis-folding and aggregation of this protein.

Chapter 3: Solving protein structures using short-distance crosslinking constraints as a guide for discrete molecular dynamics simulations

Work in this chapter was performed in the laboratories of Dr. Christoph Borchers and Dr. Nikolay Dokholyan. The crosslinking, surface modification, and hydrogen deuterium exchange of the proteins myoglobin and FKBP25 was performed by Nicholas Brodie. All LC-MS/MS experiments and data analysis thereof was performed by Nicholas Brodie. All discrete molecular dynamics simulations were performed by Konstantin Popov. Experimental design was performed by Nicholas Brodie, Evgeniy Petrotchenko, Konstantin Popov, Nikolay Dokholyan and Christoph Borchers. Christoph Borchers oversaw this project.

This chapter was adapted from the publication:

Brodie, Nicholas; Popov, Konstantin; Petrotchenko, Evgeniy; Dokholyan, Nikolay; Borchers, Christoph. **Solving protein structures using short-distance cross-linking constraints as a guide for discrete molecular dynamics simulations**. Science Advances. Volume 3, no. 7, e1700479. July 7, 2017. **Doi:** 10.1126/sciadv.1700479

3.1. Introduction to crosslinking and discrete molecular dynamics

Since the publication in 2000 of the landmark paper by Young *et al.* on fold recognition using crosslinking data [137], the idea of solving protein structures using crosslinking distance constraints has attracted the attention of researchers worldwide. Indeed, it seems intuitively obvious that the 3-dimensional structure of a protein should

be able to be unequivocally defined by a collection of pairwise short inter-residue distances or inter-residue contacts. Unfortunately, there are only a few, rare, opportunities for the formation of zero-length crosslinks, which can be directly translated into inter-residue contacts in proteins. Zero-length crosslinks can be amide bond formation between adjacent amino- and carboxy-groups [17], crosslinks between adjacent tyrosine residues [138], and/or disulfide bond formation between cysteine residues. Traditional amine-reactive crosslinking reagents can provide only long-distance constraints (>15 Å) between amine groups. Unfortunately amine groups have a relatively sparse distribution and are usually only found on the protein surface. Recently, non-specific short-distance hetero-bifunctional [139] and homo-bifunctional [140] photo-reactive reagents have been designed for this purpose (See Chapter 2). These reagents have the potential to form crosslinks between pairs of nearby amino acid residues and therefore should be able to provide the required number of short-distance constraints for determining the true protein structure.

For the past decade, computational approaches have provided an alternative for protein structure determination [141-143] and have become powerful and widely used tools for computational structural biology [144-148]. Great progress is currently being made in knowledge-based prediction methods which take advantage of both solved protein structures and homology-detection algorithms [148]. *De novo* structure prediction methods rely solely on energy-based calculations and are attractive because they are knowledge-independent approaches for protein structure prediction [149, 150]. For smaller proteins (<100 residues), because of the smaller conformational space that needs to be sampled, computational methods have been able to accurately predict native-like

structures [74, 151]. Larger proteins (>100 residues), however, undergo folding on a microsecond time scale which makes prediction of the structure for such protein computationally unrealistic even if highly efficient computational algorithms and specialized hardware are used [71, 72]. Energy functions themselves can cause biasing toward or against specific protein structural features during protein folding simulations [152-154]. The inclusion of experimental data as constraints on the modelling process has the potential to overcome these issues and increase the accuracy of the predictions. Experimentally derived data on a protein's structure simultaneously decreases the allowed protein conformational space and prevent computational bias toward incorrect protein folds or configurations.

Of the available types of experimentally derived structural data, residue-level or atom-level structural data are preferred. Crosslinking analysis in combination with modern MS techniques provides inter-residue distances that can be incorporated into the modelling process. However, crosslinking results can produce inconsistent data because of fluctuations in the solution structure of the protein during the experiment [155, 156]. Thus the incorporation of crosslinking constraints will represent a structural ensemble, rather than a single protein structure. This must be taken into consideration when selecting the "best fit" models from computationally generated ensembles of conformations [21, 157] and when directly incorporating distance constraints into an energy-based simulation process [155, 158].

Here I will present a method for predicting protein structures by adding experimental crosslinking distance constraints into discrete molecular dynamics (DMD) simulations [73, 159], which I will call CL-DMD from here on. The incorporation of these

experimental data considerably reduces the allowed conformational space during simulations, helping to guide the folding of the protein toward conformational ensembles with minimum energies at shorter time scales. I consider this workflow to be the first step in an ongoing effort that will allow multiple types of residue-level experimental constraints – derived from structural proteomics – to be incorporated into the modelling process.

I have validated the workflow on proteins with well-known and well-defined structures, and have shown that this approach successfully predicts model structures that agree with known x-ray structures. I have also independently validated the predicted structures with additional experimental structural proteomics techniques, such as HDX, chemical surface modification, and long-distance crosslinking (LD-CL).

3.2. Materials and Methods

Materials and Reagents. All materials were from Sigma-Aldrich unless noted otherwise. The FKBP protein was a gift from C.J. Nelson (University of Victoria, Canada), and was expressed and purified as in reference [160].

3.2.1. Short-distance Crosslinking

40 µL-aliquots of horse myoglobin at a concentration of 1 mg/mL in PBS were crosslinked using either 0.40 mM azidobenzoic acid succinimide (ABAS), 0.6 mM disuccinimidyl adipate (DSA), or 0.6 mM disuccinimidyl glutatrate (DSG) (all from Creative Molecules, Inc.). The structures of these crosslinking reagents are shown in Figure 2. A 0.28 mg/mL solution of myoglobin or a 0.328 mM solution of succinimidyldiazirine (SDA) was used for the SDA crosslinking reactions. FKBP25 was prepared at a concentration of 0.14 mg/mL, and 105 µL was crosslinked using either

DSA or triazidotriazine (TATA), at a concentration of 0.46 mM. These reaction mixtures were incubated for 10 minutes in the dark to allow the NHS-ester reaction to take place, followed by 10 minutes of UV irradiation under a 25 W UV lamp (Model UVGL-58 Mineralight lamp, UVG) at 254 nm for ABAS and TATA, or 366 nm for SDA. Samples were then acidified with formic acid prior to LC-MS/MS analysis.

3.2.2. Computational Methods

An all-atom protein model was used with a united atom representation in which all heavy atoms and polar hydrogens are explicitly represented. The discrete Medusa force field used in DMD approximates atomic interactions, such as van der Waals, electrostatics and hydrogen bonding by multistep square-well potentials [161, 162]. The Lazaridis-Karplus implicit solvation model [163] was adopted to account for the solvation energy. In addition, we use the Anderson thermostat [164] to control the temperature during simulations.

In order to incorporate inter-residue proximity constraints into the DMD simulations, we introduced additional square-well potentials between the crosslinked atoms into the Medusa force field, where the first term is the Hamiltonian corresponding to the original Medusa force field [161, 162].

$$H = H^{Medusa} + \sum_{i<j}^{Ncl} E(r_{ij})$$

Equation 1: Medusa Force Field Equation

Equation of the Medusa force field used to calculate protein models in DMD simulations.

The second term represents the sum of pairwise interactions for the crosslinked atoms. Ncl is the number of crosslinks. For each pair of crosslinked atoms $E(r_{ij})$ has a well-like shape:

$$E(r_{ij}) = \begin{cases} e, & r_{ij} \le r_{\min}^{ij} \\ 0, & r_{\min}^{ij} < r_{ij} < r_{\max}^{ij} \\ e, & r_{ij} \ge r_{\max}^{ij} \end{cases}$$

Here, r_{ij} is the distance between two crosslinked atoms during the simulations; $r_{\min}^{ij}$ and $r_{\max}^{ij}$ are the minimum and maximum inter-atom distances allowed by each particular crosslinker; e is the energetic value assigned for the depth of the well (in this work we used 20 kcal/mol).

This potential allows the atoms to freely move within the wells and will energetically penalize any motion outside the potential wells. Thus, the addition of a set of these crosslink-based potentials will make the corresponding portion of the conformational space energetically prohibitive for trajectories during the protein folding simulations.

To reduce the degree of complexity of the folding protein, we did not explicitly model the heme group during the myoglobin simulations. Instead, we introduced a few additional structural constraints between those myoglobin residues which directly interact with the heme group (based on the X-ray structure of the protein (PDB: 2V1H)).

Using these constraints, we used a replica exchange (REX) approach [165, 166] for the DMD simulations. Starting with the unfolded conformation, we ran multiple parallel simulations for different replicas of the same system at different temperatures. The replicas periodically exchange their temperatures, allowing the system to overcome local energy barriers and explore a larger conformational space. For each run, we analyzed 24

parallel replicas with temperatures within a range from 0.375 to 0.605 kcal/(mol kB), which corresponds to ~187°K to 302°K. We ran simulations for 2 x 10^6 time steps, and saved snapshots of the structures every 1000 steps per replica.

3.2.2.1. Clustering

The trajectories obtained were then analyzed, and the 10% of the structures that had the lowest energy were selected. We performed a clustering analysis on these structures using the algorithm implemented in Wordom and GROMACS [167, 168].

We calculated the distribution for the pairwise RMSD's between Cα atoms of the selected structures, and defined the highest peak of the obtained distribution as a threshold value for distances between the structures within a single cluster. Finally, we selected as the model a centroid of the most populated cluster that had the lowest average energy.

3.2.2.2. Model Dynamics

The strength of this new approach lies in the incorporation of experimentally derived constraints as part of the force field that is used to computationally predict the protein structure, instead of using these constraints as filters during the last stage of structure determination. This approach allows the user to (i) identify a native-like protein conformation and (ii) to capture its intrinsic dynamic and structural fluctuations.

The final models for FKBP and myoglobin and the dynamics of the simulation are presented in Figure 19. Figure 19B and E shows the regions and amplitudes of the fluctuations in the models. Figure 19C and F illustrates how often different residue-residue contacts appeared during the simulation of the structure.

A static-contact map is a binary two-dimensional matrix in which a value of 1 is assigned for every two residues (i and j) of the protein, if the distance between their Cα's

is less than a specified cut-off distance (8 Å in our case). Contacts of a residue with itself, i=j, are omitted. The map is symmetrical with respect to i and j – thus, only half of the map is needed in order to show the contacts for an entire protein. The static-contact maps for our predicted models are plotted below the diagonals in Figure 19C and F.

Above the diagonals in Figure 19C and F. we have shown the contact-frequency maps for the residues in those structures within the most populated cluster, for which our predicted model is the centroid (see clustering section above). This contact-frequency map is similar to the static-contact map, but instead of binary values (1 or 0) for contacts between residues i and j, we have a number between 0 and 1 which corresponds to the frequency of this contact in the structures within this cluster. In order to calculate this value, we counted the number of the structures within this cluster for which residues i and j are in the contact, normalized by the total number of structures within the cluster.

3.2.3. LC-MS/MS analysis

Mass spectrometric analysis was performed using a nano-HPLC system (Easy-nLC II, ThermoFisher Scientific), coupled to the ESI-source of an LTQ Orbitrap Velos or Fusion (ThermoFisher Scientific), using conditions described in [1]. Briefly, samples were injected onto a 100 μm ID 3 x 60 μm OD trapping column packed with Magic C18AQ, 100 Å, 5 μm pore size particles (Bruker-Michrom, Auburn, CA), prepared in-house, and desalted by washing with 5 μL solvent A (2 % acetonitrile:98 % water, both containing 0.1 % formic acid (FA). Peptides were separated with a 60-min gradient (0–60 min: 4–40 % solvent B (90 % acetonitrile, 10 % water, 0.1 % FA), 60–62 min: 40–80 % B, 62–70 min: 80 % B), on a 75 μm ID, 360 μm OD analytical column packed with Magic C18AQ 100 Å 5 μm pore size particles (prepared in-house) with an IntegraFrit (New Objective

Inc., Woburn, MA), and equilibrated with solvent A. MS data were acquired using a data dependent method. The DDA utilized dynamic exclusion, with an exclusion window of 10 ppm and exclusion duration of 60 seconds. MS and MS/MS events used 60000 and 30000 resolution FTMS scans, respectively, with a scan range of 400-2000 m/z in the MS mode. For MS/MS, the collision energy was set to 35 %. Data were analyzed using the DXMSMS Match program [45] from our ICC-CLASS software suite, or with Kojak [43]. For scoring and assignment of the MS/MS spectra, b and y ions were primarily used, with additional confirmation from CID-cleavage of the crosslinker whenever this was available.

3.2.4. Circular dichroism

Circular dichroism (CD) spectra were recorded on Jasco J-720 spectrometer in a stream of nitrogen. The α- and β-structure contents were calculated using the BeStSel web server [169].

3.2.5. Hydrogen/deuterium exchange

Top-down ECD-FTMS hydrogen/deuterium exchange was performed as described previously [56] (see Figure 4 for the workflow). Briefly, protein solution and D2O in separate syringes were continuously mixed in a 1:4 ratio (with a final concentration of 80% D2O) via a 3-way tee which was connected to a 100 µm x 5 cm capillary, providing a labelling time of 2 s. The outflow from this capillary was mixed with a quenching solution containing 0.4% formic acid: 20% acetonitrile :64% D2O and 16% H2O from a third syringe via a second 3-way tee, and injected into a Bruker 12 T Apex-Qe hybrid Fourier Transform ion cyclotron resonance mass spectrometer (FTICR-MS), equipped with an Apollo II electrospray source. In-cell ECD fragmentation experiments were

performed using a cathode filament current of 1.3 amps and a grid potential of 13 V. Approximately 800 scans were accumulated over the m/z range 200-2000, corresponding to an acquisition time of approximately 20 minutes for each ECD spectrum. Deuteration levels of the amino acid residues were determined using the HDX Match program [170].

3.2.6. Surface modification

Chemical surface modification with pyridine carboxylic acid N-hydroxysuccinimide ester (PCAS) was performed as previously described (as shown in Figure 5) [62]. Briefly, myoglobin was prepared as a 50 µM solution in PBS containing 8 M urea, pH 7.4 to generate the unfolded state, or in PBS only to generate the folded state. Either the light or the heavy form of the isotopically-coded reagent (PCAS-$^{12}C_6$ or PCAS-$^{13}C_6$) (Creative Molecules, Inc.) was then added to give a final concentration of 10 mM. Reaction mixtures were incubated for 5 minutes, and then quenched with 50 mM ammonium bicarbonate. Next, samples were mixed at a 1:1 ratio, combining folded (PCAS-^{12}C) with unfolded (PCAS-^{13}C) samples. Samples were acidified with 150 mM acetic acid and digested with pepsin at a 20:1 protein:enzyme ratio overnight at 37 ^{o}C. After digestion, samples were prepared for MS analysis using C18 zip-tips (Millipore). Zip-tips were equilibrated with 30 µL 0.1 % TFA, the sample was introduced, then washed with 30 µL 0.1 % TFA and eluted with 2 µL of 0.1 % aqueous formic acid/50 % acetonitrile. Samples were analyzed by LC-MS/MS as described above.

3.2.7. Long distance crosslinking using CBDPS.

For the CBDPS reactions, FKBP and myoglobin were prepared at 1 mg/mL and crosslinked with 0.1 mM CBDPS. Reactions were quenched with 10 mM ammonium bicarbonate. Aliquots were then split and digested with either trypsin or proteinase K at

an enzyme:protein ratio of 1:20. Samples were then acidified with formic acid prior to LC-MS/MS analysis.

3.3. Results and Discussion

The workflow for the CL-DMD method is shown in Figure 14. The overall workflow consisted of three main steps: 1) the acquisition of the short-distance crosslinking data, 2) the performance of CL-DMD simulations guided by these crosslinking constraints, and 3) the validation of the obtained structures with additional structural-proteomics methods. If the model did not meet the validation criteria, the workflow could be repeated after adding additional sets of crosslinking data.

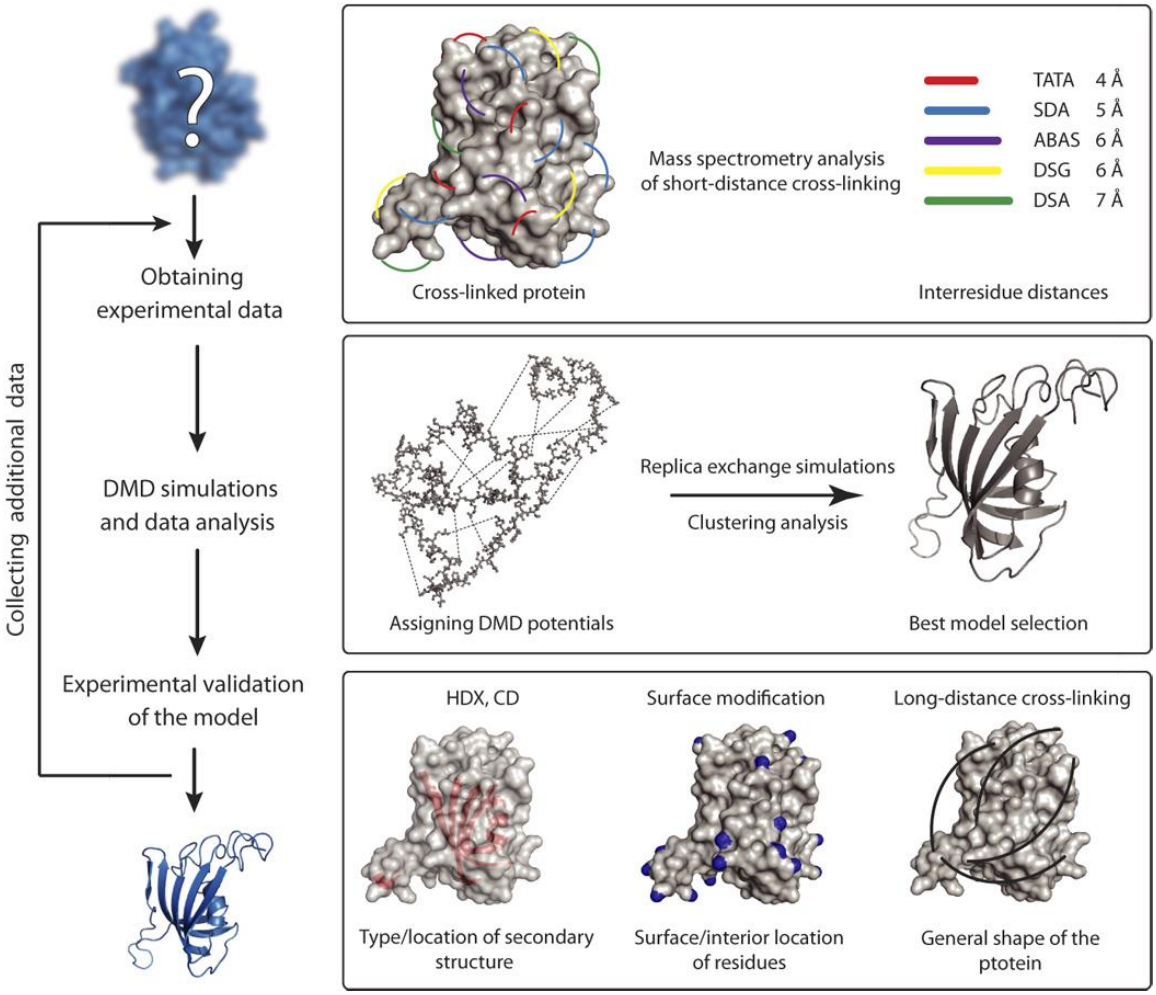


Figure 14: CL-DMD workflow schematic

3.3.1. Short-distance crosslinking

The key to this approach is to obtain multiple inter-residue short-distance crosslinking constraints covering most of the protein. To obtain these distance constraints, I used a panel of crosslinking reagents that can produce zero-length (i.e., no crosslinker spacer) or short (~5 Å) crosslinks. To obtain numerous crosslinks for every region of the protein, non-selective photo-reactive hetero- and homo-bifunctional crosslinkers were used [139, 140]. For the proof-of-concept experiments shown here, I used myoglobin (Mb) and the FK506 binding domain of the FKBP25 protein (hereafter, FKBP), models for alpha-helix and beta-sheet rich proteins, respectively. To generate the crosslinks, I used a panel of crosslinking reagents consisting of disuccinimidyl adipate (DSA), disuccinimidyl glutarate (DSG), succinimidyl 4,4'-azipentanoate (SDA) [139], azido-benzoic acid succinimide (ABAS) [139], and triazidotriazirine (TATA) [140]. DSA and DSG are amine-reactive reagents, SDA and ABAS are hetero-bifunctional amino group- and photo-reactive reagents, and TATA is a homo-bifunctional photoreactive reagent. Crosslinked proteins were digested with proteolytic enzymes (trypsin or proteinase K [171]), and the resulting peptides were analyzed by LC-MS/MS (Figure 15, Tables 3 and 4). Crosslinks were found to be rather evenly distributed throughout the protein structures, connecting the secondary-structure motifs and loops (Fig. 15) that were known to be adjacent. These short-distance crosslinks were used as constraints for the DMD simulations.

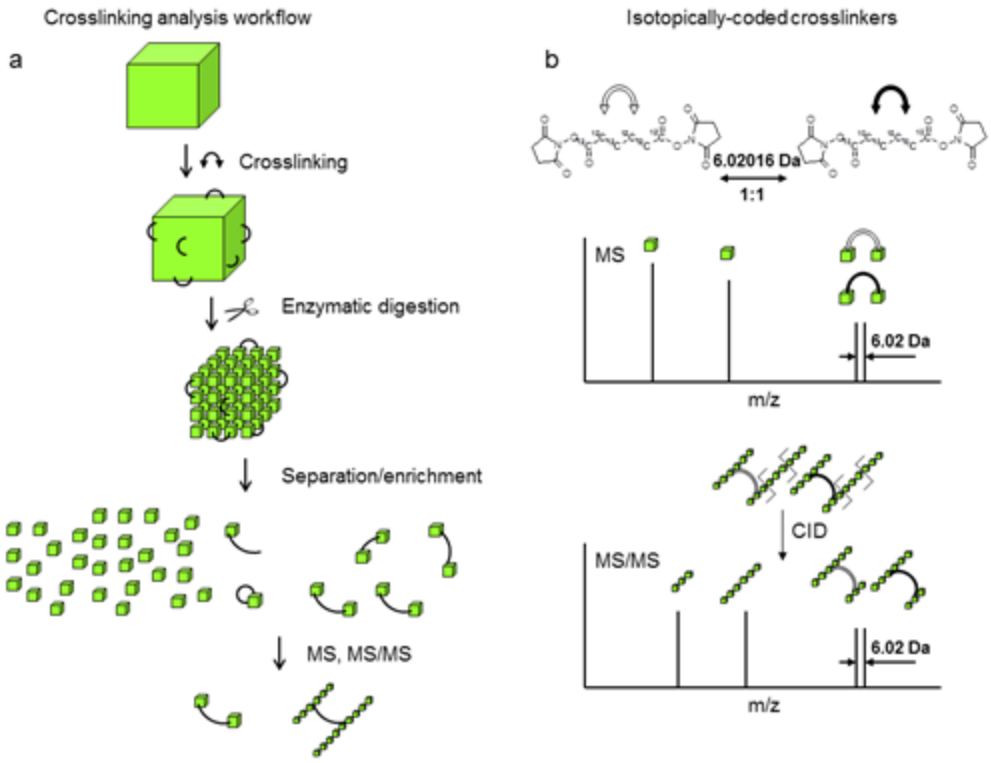


Figure 15: Crosslinking Analysis Workflow

Panel a) Proteins are crosslinked, enzymatically digested, and analyzed by LC-MSMS. Panel b) Isotopic labelling of crosslinking reagents allows for the specific detection and acquisition of crosslinked peptides in MS1. The doublets present in the combined MS/MS spectra of the light and heavy forms of the crosslink also provide additional confirmation of the correct assignment of the crosslink.

Table 3: Myoglobin inter-peptide crosslinks

Myoglobin inter-peptide crosslinks were used as constraints in DMD simulations. ABAS and TATA crosslinked samples were digested with proteinase K, while SDA, DSA, and DSG samples were digested with trypsin.

Crosslinker	M+H	m/z	z	ppm error	Residues 1	Sequence	Residues 2	Sequence	Crosslink
ABAS	1094.5079	547.7579	2	0.8	55-56	E.MK.A	118-125	S.KHPGDF.G	K56-K118
ABAS	1225.5476	613.2777	2	0.7	133-134	T.KA.L	1-8	M.GLSDGEWQ.Q	N-term-K133
TATA	787.3334	394.1706	2	1.1	1-2	-.GL.S	80-83	K.GHHE.A	N-term-G80
TATA	821.4743	411.2408	2	-0.3	88-90	K.PLA.Q	142-145	D.IAAK.Y	A90-K145
TATA	922.5115	461.7597	2	-1.1	37-39	H.PET.L	98-101	H.KIPI.K	T39-I99
TATA	1031.4278	516.2178	2	1.1	1-7	-.GLSDGEW.Q	129-130	Q.GA.M	N-term-A130
TATA	1035.5952	518.3015	2	0.7	45-48	D.KFKH.L	98-100	H.KIP.P	K45-K98
TATA	1090.4008	1090.4008	1	0	3-8	L.SDGEWQ.Q	129-131	Q.GAM.T	S3-A130
DSA	1785.9883	893.4978	2	0.5	97-102	K.HKIPIK.Y	146-153	K.YKELGFQG.-	K98-K147
DSA	2853.4494	951.8213	3	0.5	1-16	-.GLSDGEWQQVLNVWGK.V	140-147	R.NDIAAKYK.E	N-term-K145
DSA	2904.4621	968.8255	3	0.4	80-96	K.GHHEAELKPLAQSHATK.H	146-153	K.YKELGFQG.-	K87-K147
DSA	3013.5848	1005.1998	3	0	79-96	K.KGHHEAELKPLAQSHATK.H	140-147	R.NDIAAKYK.E	K87-K145
DSG	2492.3603	623.8455	4	-2.5	97-102	K.HKIPIK.Y	32-45	R.LFTGHPETLEKFDK.F	K98-K42
SDA	3194.8229	799.4612	4	0.8	64-78	K.HGTVVLTALGGILKK.K	17-31	K.VEADIAGHGQEVLIR.L	K77-E18
SDA	3194.8225	799.4611	4	0.9	64-78	K.HGTVVLTALGGILKK.K	17-31	K.VEADIAGHGQEVLIR.L	K77-V17
SDA	3167.6229	792.6612	4	0.6	51-63	K.TEAEMKASEDLKK.H	17-31	K.VEADIAGHGQEVLIR.L	K56-E27
SDA	3167.6252	792.6617	4	-0.2	51-63	K.TEAEMKASEDLKK.H	17-31	K.VEADIAGHGQEVLIR.L	K56-Q26
SDA	3167.6245	792.6616	4	0	51-63	K.TEAEMKASEDLKK.H	17-31	K.VEADIAGHGQEVLIR.L	K56-H24
SDA	3349.7411	838.1907	4	0.3	32-45	R.LFTGHPETLEKFDK.F	17-31	K.VEADIAGHGQEVLIR.L	K42-L29
SDA	3750.9015	938.4808	4	-0.9	80-96	K.GHHEAELKPLAQSHATK.H	1-16	-.GLSDGEWQQVLNVWGK.V	N-term-H81

Table 4: FKBP inter-protein crosslinks

Myoglobin inter-peptide crosslinks were used as constraints in DMD simulations. ABAS and TATA crosslinked samples were digested with proteinase K, while SDA, DSA, and DSG samples were digested with trypsin.

Crosslinker	M+H	m/z	z	ppm	Residues 1	Sequence	Residues 2	Sequence	Crosslink
DSA	1144.5990	572.8034	2	0.6	30-38	K.KGDKTNFPK.K	-	-.i.-	K30-K33
DSA	1193.5875	597.2977	2	-0.3	23-25	K.YTK.S	16-22	R.GSMGPPK.Y	K22-K25
DSA	1473.8036	491.9398	3	1.6	113-125	K.KGQPDAKIPPNAK.L	-	-.i.-	K113-K119
DSA	1679.9786	420.7505	4	1	113-119	K.KGQPDAK.I	78-85	K.VGVGKVIR.G	K82-K113
DSA	1701.9623	851.4851	2	1.4	97-101	K.GEKAR.L	69-77	K.KNAKPLSFK.V	K72-K99
DSA	1709.0443	855.0261	2	1.7	71-85	N.AKPLSFKVGVGKVIR.G	-	-.i.-	K72-K77
DSA	1748.0042	437.7569	4	1.3	67-69	K.KKK.N	114-125	K.GQPDAKIPPNAK.L	K119-K68
DSA	1901.9437	951.4758	2	-0.6	86-101	R.GWDEALLTMSKGEKAR.L	-	-.i.-	K99-K96
DSA	1917.9720	480.2489	4	1	113-119	K.KGQPDAK.I	16-25	R.GSMGPPKYTK.S	K113-K22
DSA	1982.0948	496.2796	4	0.5	23-29	K.YTKSVLK.K	30-38	K.KGDKTNFPK.K	K25-K30
DSA	2077.1019	693.0392	3	-1	67-69	K.KKK.N	86-99	R.GWDEALLTMSKGEK.A	K96-K68
DSA	2271.2424	757.7527	3	-1.8	78-85	K.VGVGKVIR.G	102-112	R.LEIEPEWAYGK.K	K82-Y110
DSA	2351.2040	588.5569	4	-1	34-39	K.TNFPKK.G	3-15	R.GSHHHHHHGLVPR.G	S3-K38
DSA	2399.3342	600.5894	4	-0.4	78-85	K.VGVGKVIR.G	102-113	R.LEIEPEWAYGKK.G	K112-K82
DSA	2472.2611	824.7589	3	-0.8	16-22	R.GSMGPPK.Y	100-113	K.ARLEIEPEWAYGKK.G	S17-K112
DSA	2523.2475	421.3811	6	1	31-38	K.GDKTNFPK.K	3-15	R.GSHHHHHHGLVPR.G	S3-K33
DSA	2637.3286	1319.1682	2	-0.6	16-25	R.GSMGPPKYTK.S	102-113	R.LEIEPEWAYGKK.G	K22-K112
DSA	2637.3296	879.7817	3	-1	16-25	R.GSMGPPKYTK.S	102-113	R.LEIEPEWAYGKK.G	K22-K113
DSA	2807.4643	702.6220	4	-1	102-113	R.LEIEPEWAYGKK.G	114-125	K.GQPDAKIPPNAK.L	K113-K119
DSA	2864.4653	716.8722	4	0	16-25	R.GSMGPPKYTK.S	100-113	K.ARLEIEPEWAYGKK.G	K22-K112
TATA	1591.7847	531.2668	3	-1.5	102-107	R.LEIEPE.W	113-119	K.KGQPDAK.I	E107-G114
TATA	1743.8944	581.9700	3	0.8	37-39	F.PKK.G	86-96	R.GWDEALLTMSK.G	K39-E89
TATA	1866.9105	622.9754	3	1.5	23-25	K.YTK.S	102-112	R.LEIEPEWAYGK.K	K25-W108
TATA	2164.0559	722.0239	3	0.5	102-109	R.LEIEPEWA.Y	83-91	K.VIRGWDEAL.L	A109-I84

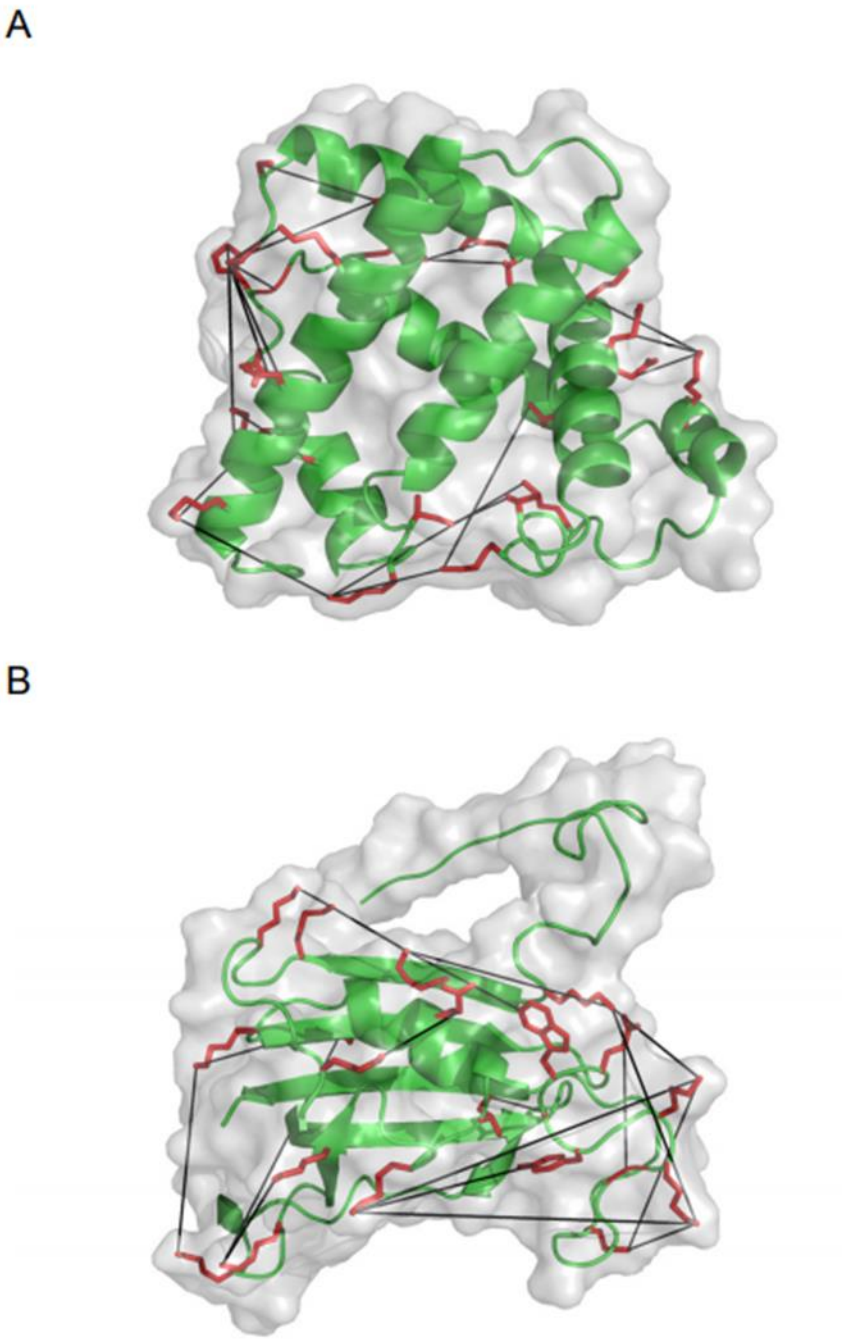


Figure 16: Crosslinking results for Myoglobin and FKBP

Crosslinking results overlaid onto the crystal structures of A) Myoglobin and B) FKBP. Crosslinked residues are shown in red. Distances are within the maximum distance expected for each reagent.

3.3.2. Discrete molecular dynamics simulations

DMD is a physics-based efficient computational algorithm for the structural simulation of proteins and complexes.[73, 74, 159, 162]. DMD uses physical principles of ballistic motion to describe the time evolution of the atom positions. In the event of a collision, the atoms involved instantaneously change their position and velocities according to energy and momentum conservation laws [74]. This algorithm has been shown to provide more efficient sampling of the protein conformational space than traditional MD simulations, allowing more rapid folding of large proteins. Also, the discrete energy representation allows for the incorporation of experimental pairwise atom-proximity constraints [161, 172] – thus for each experimental constraint we have introduced an additional potential to the force field developed [161, 173]. The combination of these potentials constrains the positions of the crosslinked atoms during the simulations. The width of the potential well is defined by the spacer length of the crosslinker (see Section 3.2.2. Computational Methods for details).

For each protein, all-atom replica exchange (REX) simulations were performed, starting from the unfolded conformation. During data analysis, the 10% of the structures that had the lowest energies in the DMD simulations were selected, and distance-based clustering was performed among them. These clusters represent different predicted conformational ensembles for a given protein. Because of the discrete nature of potentials in DMD, there are no continuous forces driving the system to satisfy all of the constraints at the same time. Thus each of these ensembles might satisfy only some of the constraints. For

further study, centroids of the most populated clusters with the lowest energies were selected and scored by our energy function as our best models.

In order to visualize how the predicted models aligned with known structures (Figures 17 and 18), the root-mean-square deviation (RMSD) values were determined for all of the structures generated during the DMD simulations of Mb and FKPB. The RMSD of the Cα atom positions provides a quantitative measurement of the similarity of the models to the X-ray structures of the proteins (PDB: 2V1H, 2MPH). The RMSD values were plotted versus the corresponding energy scores as provided by the Medusa force field energy function [73, 173], (Figures 17A and 18A). Each point on the plot corresponds to a snapshot of the structure taken during the simulation. In general, it can be seen that during the simulation, the structures cluster in areas with small RMSDs and low energies. This indicates that our approach can accurately explore the conformational space of these proteins. The data in Figures 17B and 18B represent the states with the 10% lowest energy. It can be seen that these structures populate several major clusters (see Section 3.2.2. Computational Methods for analysis). The models corresponding to each of these clusters, aligned with the corresponding X-ray structures, are presented in Figures 17C, 17D, 17E and 18C, 18D. The RMSD of the lowest energy models compared to the X-ray structures were 5.4 Å for myoglobin and 2.7 Å for FKBP.

Another strength of the approach presented here is that, based on the structures found in each ensemble, we can show the possible dynamics and fluctuation of the protein structure in the vicinity of the predicted model (Figure 19).

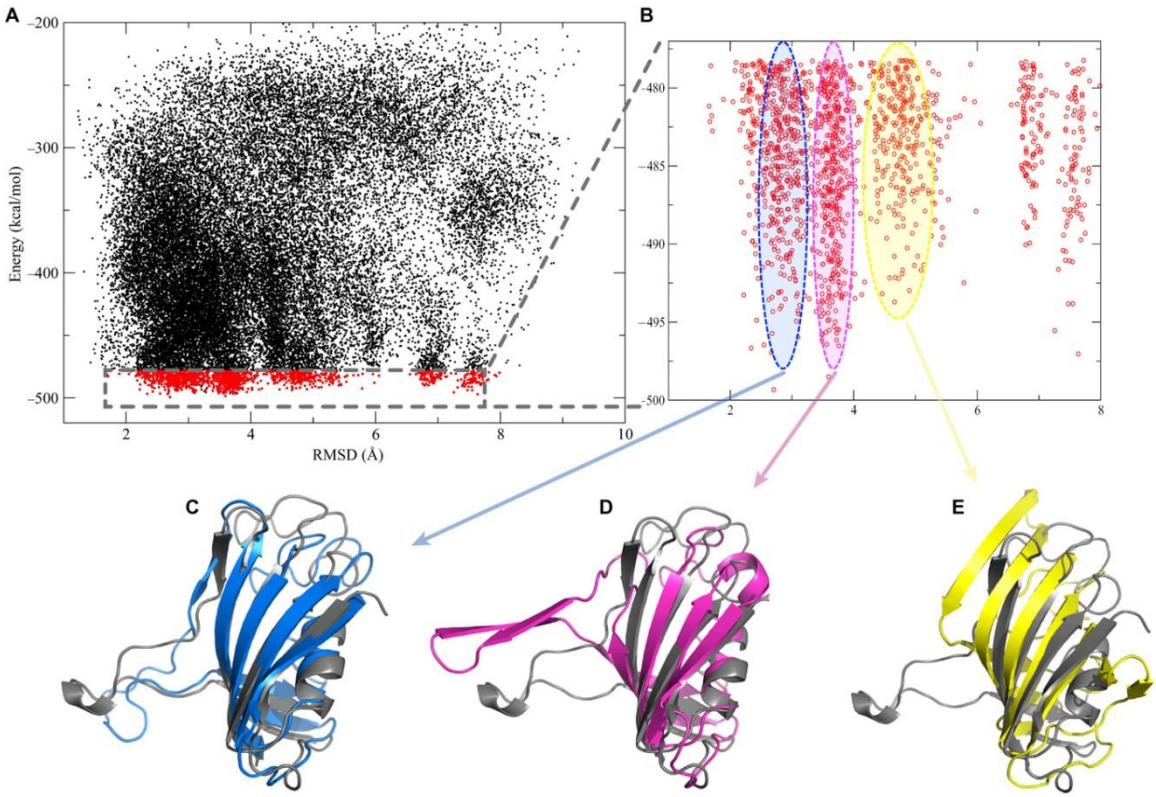


Figure 17: CL-DMD modelling of FKBP.

A) Scatter plot of the Medusa force field energy versus the RMSD (in angstroms) from the x-ray structure obtained from a CL-DMD simulation of FKBP with external experimental short-distance crosslinking constraints. B) Clusters found among the 10% of the structures that had the lowest energies. C to E) Models, corresponding to each cluster from (B), aligned to the x-ray structure of FKBP (PDB ID: 2MPH).

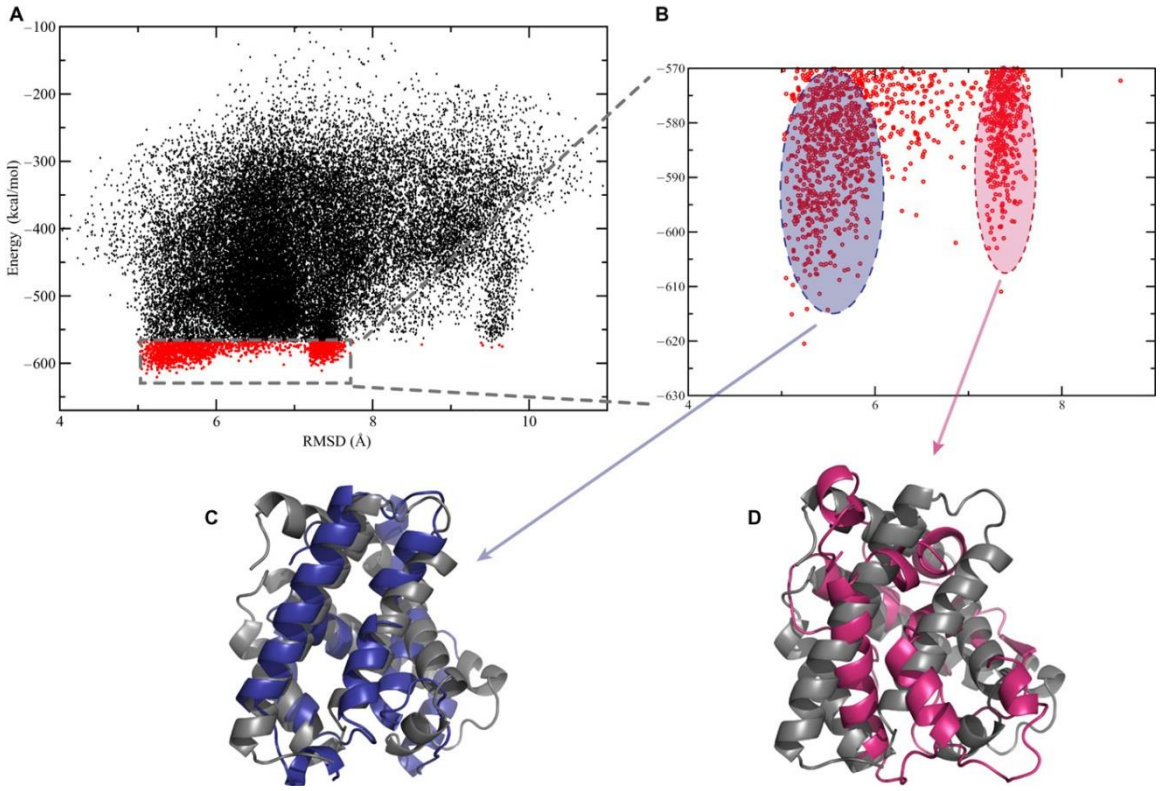


Figure 18: CL-DMD modelling of Myoglobin.

A) Scatter plot of the Medusa force field energy versus RMSD from x-ray structure obtained from simulation of Mb with external experimental short-distance crosslinking constraints. B) Clusters found among the 10% of the structures that had the lowest energies. C and D) Models, corresponding to each cluster from (B), aligned to the x-ray structure of Mb (PDB: 2V1H).

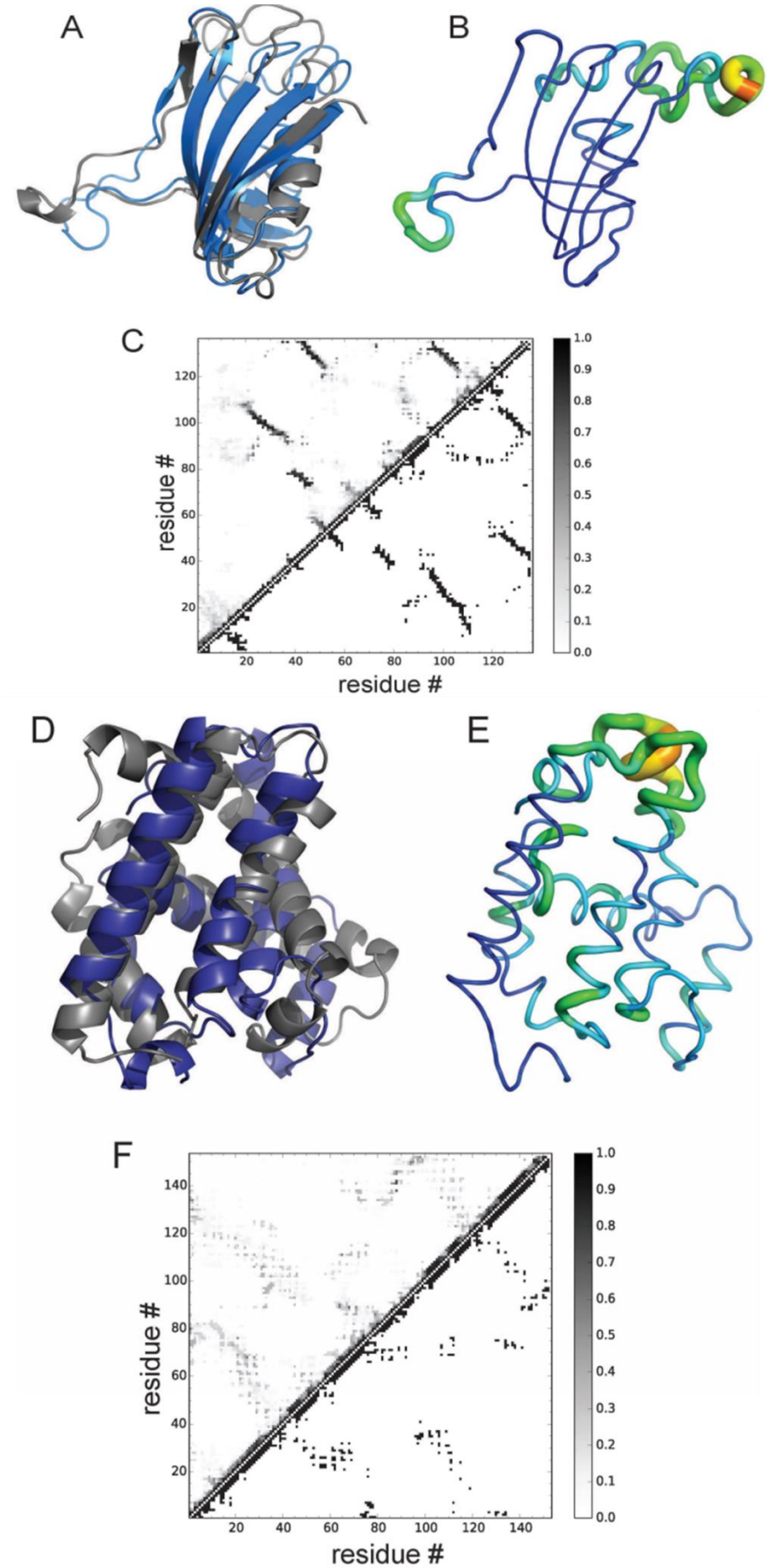

A
B
C
residue #
residue #
D
E
F
residue #
residue #

Figure 19: Conformational dynamics of predicted structures.

Shown here are the predicted models (blue) for (A) FKBP and (D) myoglobin aligned with their respective X-ray structures (grey), along with a tube representation of the fluctuations of the (B) FKBP and (E) myoglobin models. The thickness and color of the tubes indicate the dynamics of the corresponding regions during the simulations. The tubes are colored from blue (low flexibility) to red (high flexibility). The data points below the diagonals in plots (C) and (F) represent binary static-contact maps between the residues of the predicted structures shown in blue in (A) and (D), respectively. Two residues form a contact if their Cα atoms are within 8Å of each other; every black dot indicates a contact between corresponding residues. Data points above the diagonal in panels (C) and (F) show how often each particular contact between two residues can be found within the clusters for which our models (blue in (A) and (D) are centroids. The grayscale color code indicates the frequencies of the corresponding contact, where black = always in contact and white = never in contact.

3.3.3. Experimental validation of the models

For experimental validation of the final models, we used CD, hydrogen-deuterium exchange (HDX), chemical surface modification (SM), and long-distance crosslinking (LD-CL) techniques (for details of these methods, see Chapter 3.2. Materials and Methods). In the current form of the workflow (Figure 14), for well-structured proteins, such as used in this study, all of the validation criteria has to be met to pass final model validations.

Similar to CD, the HDX method provides data on the secondary-structure content, and the location of the secondary-structure motifs within the protein sequence. Here, I used our recently developed top-down HDX method, which combines FTMS with fragmentation via ECD [56]. A key advantage of this approach is the ability to determine the degree of HDX on the individual-residue level. The secondary structure content, as determined by both CD (Figure 20 and 21) and HDX (Figure 22), was in good agreement with that obtained from the final models CL-DMD (Figure 19). Delineation of the secondary structure motifs by ECD-FTMS HDX analysis also was in good agreement with the location of the α-helices and β-strands in the protein sequences (Figure 23) in

the final DMD models. According to the FKBP CL-DMD model, 78 backbone amides are involved in hydrogen bonding, compared to 72 in the crystal structure. Of these, 38 residues are involved in the formation of β sheets in the model, compared to 37 in the crystal structure. These β sheet residues represent 27 and 26% of the entire protein, respectively. On the basis of the CD data, 35% of the residues are involved in β sheets. On the basis of both the model and the crystal structure, seven residues (representing 5% of the protein) are involved in the formation of a single α helix. CD data indicate that 4% of the protein is involved in α helices. According to the model, the remaining 40 protected backbone amides form hydrogen bonds with other parts of the protein and are not involved in secondary structure, compared to the 35 in the crystal structure. The model is in agreement with the HDX data, which indicate that 79 residues of FKBP are protected from exchange. For Mb, the agreement between the model and CD and HDX data was similarly good. In the CL-DMD model of Mb, there are 88 hydrogen bonds within the α-helices, whereas in the crystal structure, there are 84. These correspond to 58 and 55% of the total number of residues, respectively. From our CD data, we observed that 55% of the protein was α-helical. On the basis of the HDX experiments, 90 of 153 residues are protected from exchange.

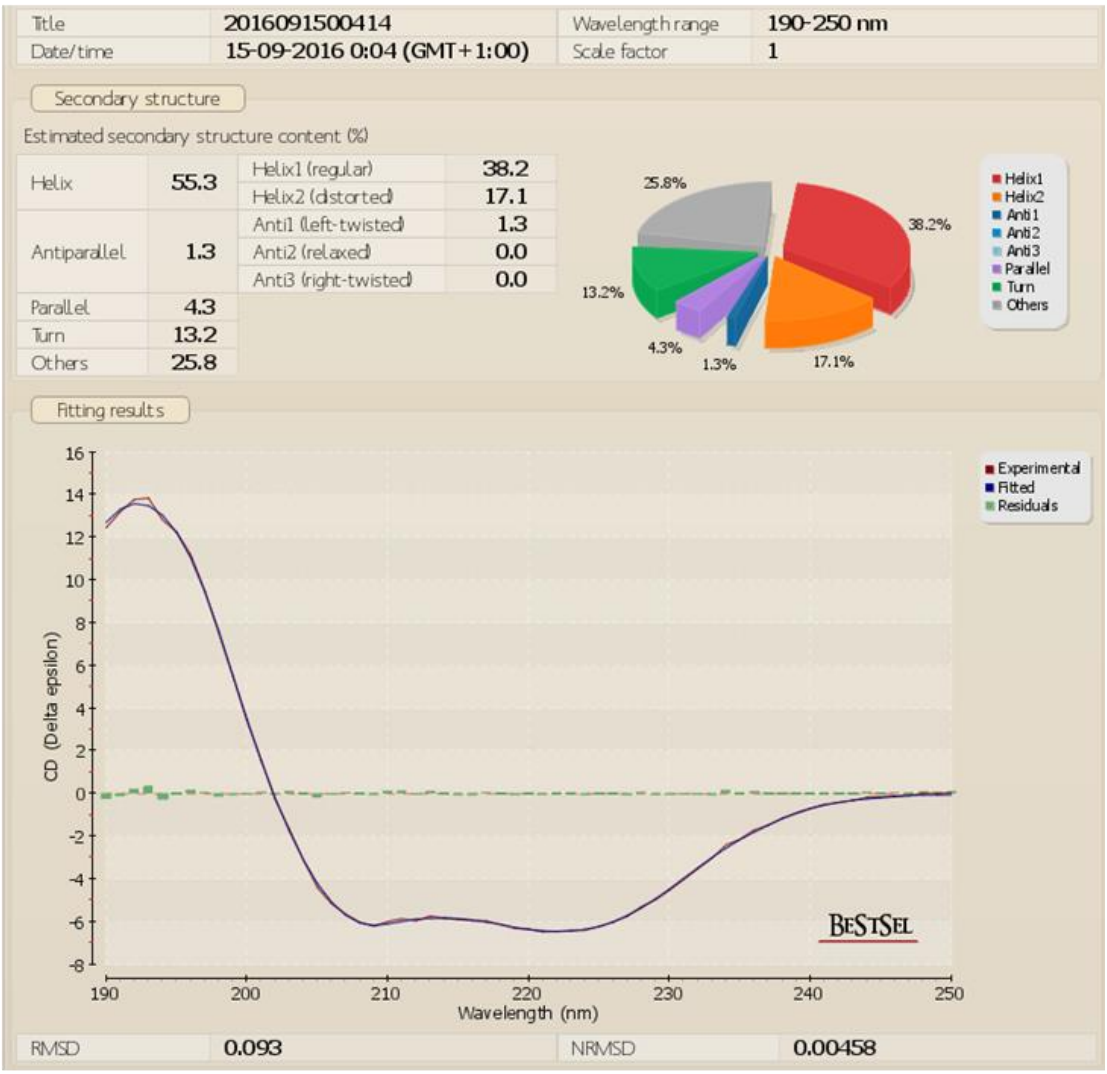


Figure 20: Circular dichroism results for myoglobin

Circular dichroism results for Mb, analyzed using the BeStSel server.

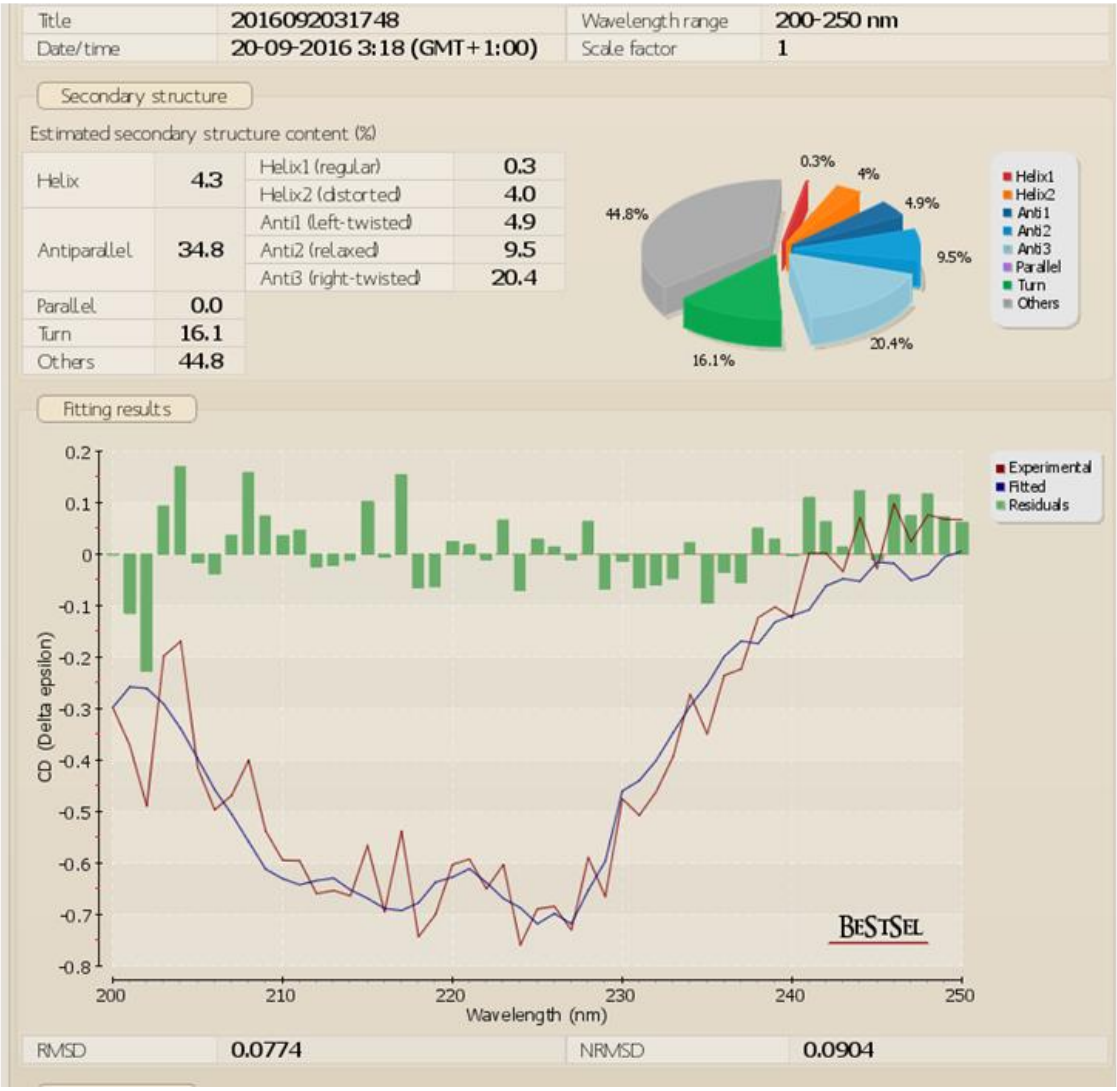


Figure 21: Circular dichroism results for FKBP

Circular dichroism results for FKBP, analyzed using the BeStSel server.

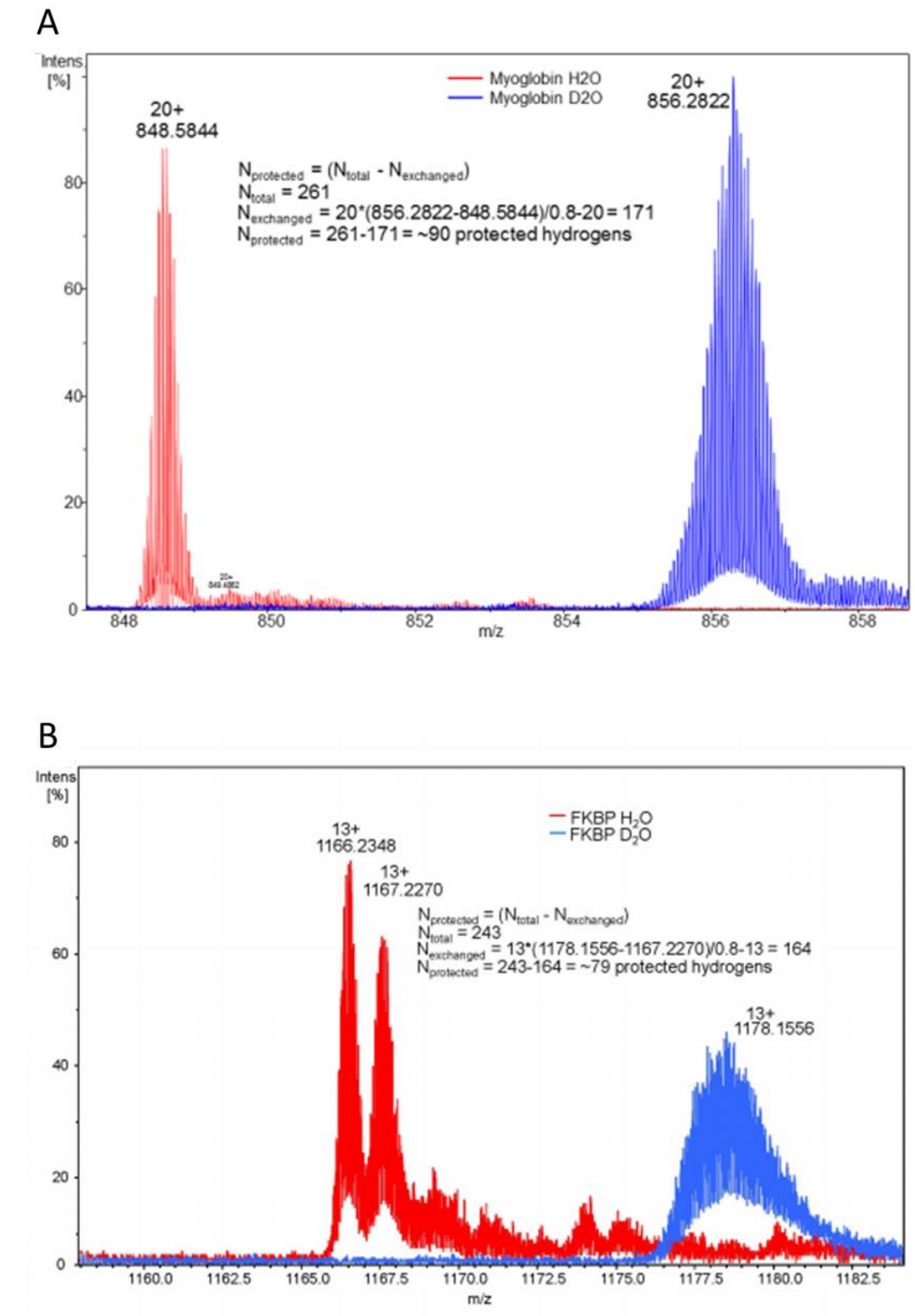


Figure 22: HDX of intact proteins

(A) Total HDX of intact myoglobin. Spectra of Myoglobin under H_2O and D_2O conditions are overlayed, and the calculation of the total number of backbone hydrogens protected from exchange is shown. (B) Total HDX of intact FKBP. Spectra of FKBP under H_2O and D_2O conditions are overlayed, and the calculation of the total number of backbone hydrogens protected from exchange is shown.

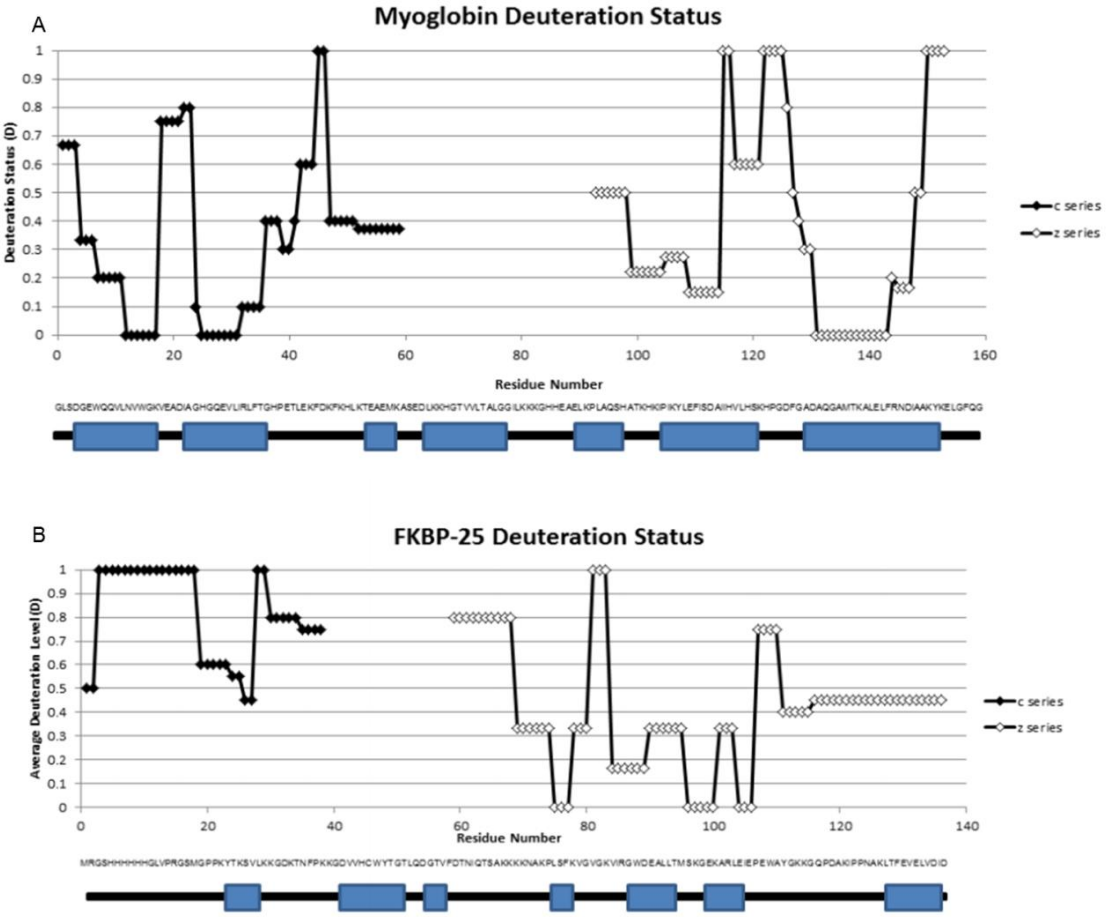


Figure 23: Deuteration status of backbone amides

The deuteration status of backbone amides for myoglobin (A) and FKBP (B). Filled dots represent c-series ions and hollow dots represent z-series ions. A deuteration status of one represents complete exchange, a status of zero represents complete protection.

To further evaluate the models, differential surface modification (SM) experiments were performed with isotopically-coded reagents, comparing the folded state with the unfolded state which was generated by denaturing the protein with 8 M urea. This differential labelling allows us to quantitatively determine the degree of surface exposure

of amino acid residues of the protein. The protein samples in the folded and unfolded states were modified with light and heavy isotopic forms of the reagent, respectively. Reactions were quenched, mixed in a 1:1 ratio, digested with pepsin, and the resulting peptides were analyzed by LC-MS/MS. In this experimental design (Figure 5), surface-exposed residues are equally modified in both folding states appear as doublets of ion signals with equal intensities in the mass spectra. In contrast, buried residues show a higher degree of modification in the unfolded state, resulting in doublet of peaks with unequal intensities in the mass spectra. For this study, we used the isotopically-coded reagent PCAS-$^{12}C_6$ or PCAS-$^{13}C_6$ [62] which modifies Lys, Tyr, Ser, and Thr residues. The SM method showed good agreement with X-ray structures, and allowed the detection of specific buried or exposed residues in the proteins (Table 5). The locations of all of these residues were in good agreement with the models (Figure 24).

Long-distance Lys-Lys crosslinking using amino-reactive reagents with spacer lengths of >10Å generally cannot be directly used for the DMD simulations, as these constraints are too loose to be reasonably employed. Nevertheless, these long-range constraints can be used to validate the in-solution protein structures predicted by our method. In this study, I used the CBDPS amine-reactive crosslinker (~14 Å spacer length) [18]. The long-distance intra-protein CBDPS crosslinks were in good concurrence with the final models of the proteins (Figure 25).

In summary, application of this new CL-DMD procedure for de novo protein structure prediction of myoglobin and FKBP gave results that were in good agreement with their known crystal structures. In-solution experiments with HDX, SM, and LD-CL, performed to compare the experimental results with the DMD-predicted structures, consistently

confirmed the modelling results, indicating that CL-DMD can be successfully used for predicting unknown protein structures.

Table 5: Residues modified by PCAS-$^{12}C_6/^{13}C_6$ in urea-PCAS experiments

The Heavy/Light (H/L) ratio indicates the ratio of the exposure a specific residue to solvent when treated with urea compared to its exposure in the folded state. An H/L ratio of > 1.5 indicates that the residue was protected from solvent in the folded, native protein.

Myoglobin				
Peptide	**Mass**	**ppm error**	**H/L ratio**	**Residue**
L.NVWGK(+105.02)VE.A	935.4501	0.3	1.30	17
K.FK(+105.02)HLKT.E	877.4810	0.0	2.63	48
L.K(+105.02)TEAEM.K	812.3375	-0.8	1.43	51
A.SEDLK(+105.02)KHGTVVL.T	1429.7570	-1.2	2.24	63
M.KASEDLKK(+105.02)HGTVV.L	1515.8050	-0.1	1.43	64
K.K(+105.02)GHHEAEL.K	1024.4730	-1.6	1.46	80
L.KPLAQSHATK(+105.02).H	1184.6300	-0.2	2.20	97
I.IHVLHSK(+105.02)HPGDF.G	1490.7420	-0.3	2.54	119
D.AQGAMTK(+105.02)A.L	881.4066	0.0	5.14	134
A.K(+105.02)YKELG.F	841.4334	-0.1	2.88	146
K.YK(+105.02)ELGF.Q	860.4069	0.1	1.31	148

FKBP				
Peptide	**Mass**	**ppm error**	**H/L ratio**	**Residue**
L.KK(+105.02)GDKTNF.P	1041.5240	-1.0	3.05	30
F.K(+109.05)VGVGKVIRG.W	1120.7060	-4.1	1.89	77
G.K(+105.02)VIRGWD.E	977.5083	0.3	1.89	82
M.SK(+105.02)GEKARL.E	992.5403	-0.5	1.72	96
L.TMSKGEK(+109.05)ARL.E	1228.6580	-4.5	2.87	99
L.EIEPEWAY(+105.02)G.K	1197.4980	-0.6	1.33	110
A.YGK(+105.02)KGQPDAKIPPNAKLTF.E	2177.1630	0.3	1.48	112

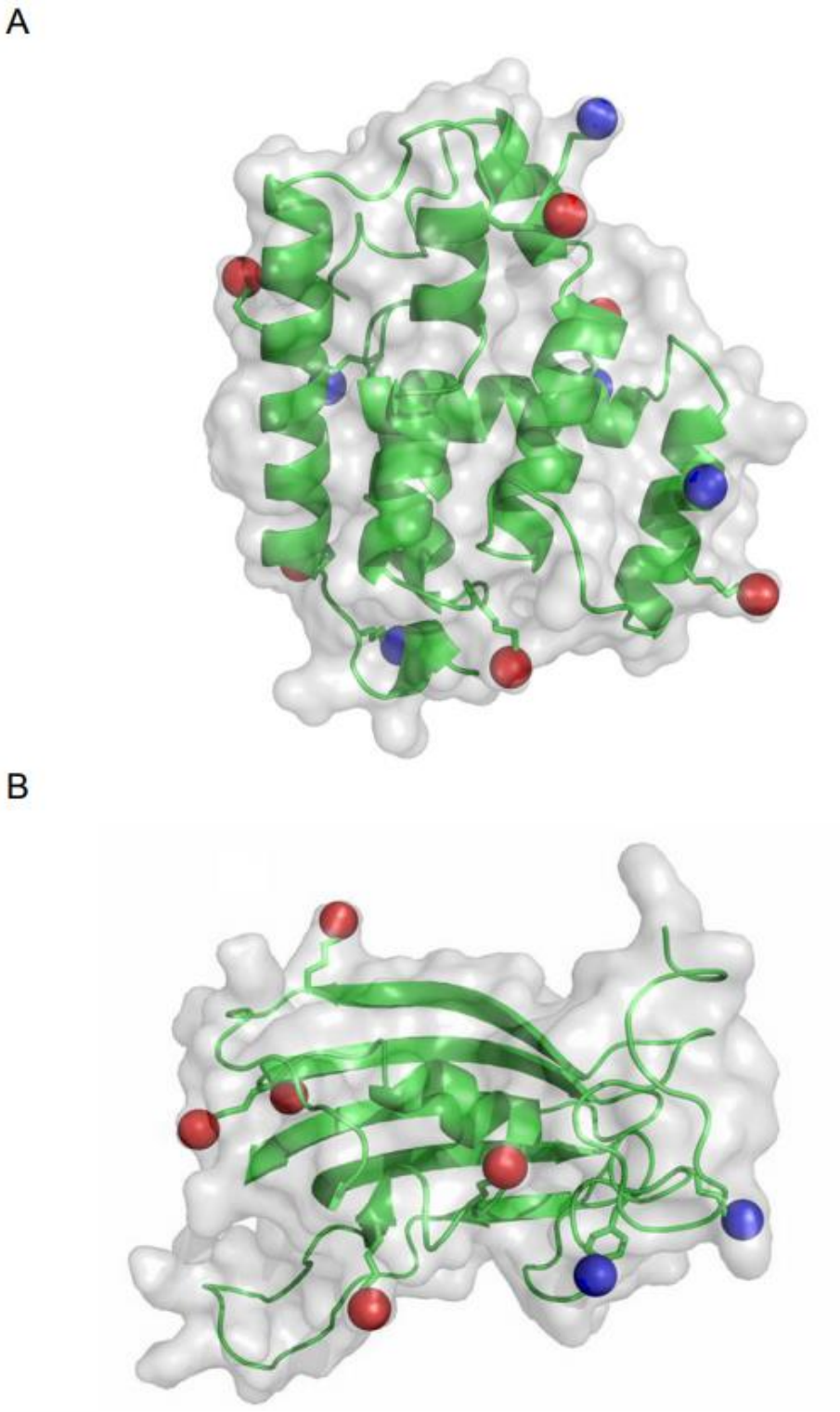


Figure 24: Surface modification results for Myoglobin and FKBP

Residues with heavy/light peak intensity ratios of ≥1.5 are shown as red spheres, while those with heavy/light peak intensity ratios below 1.5, shown here as blue spheres.

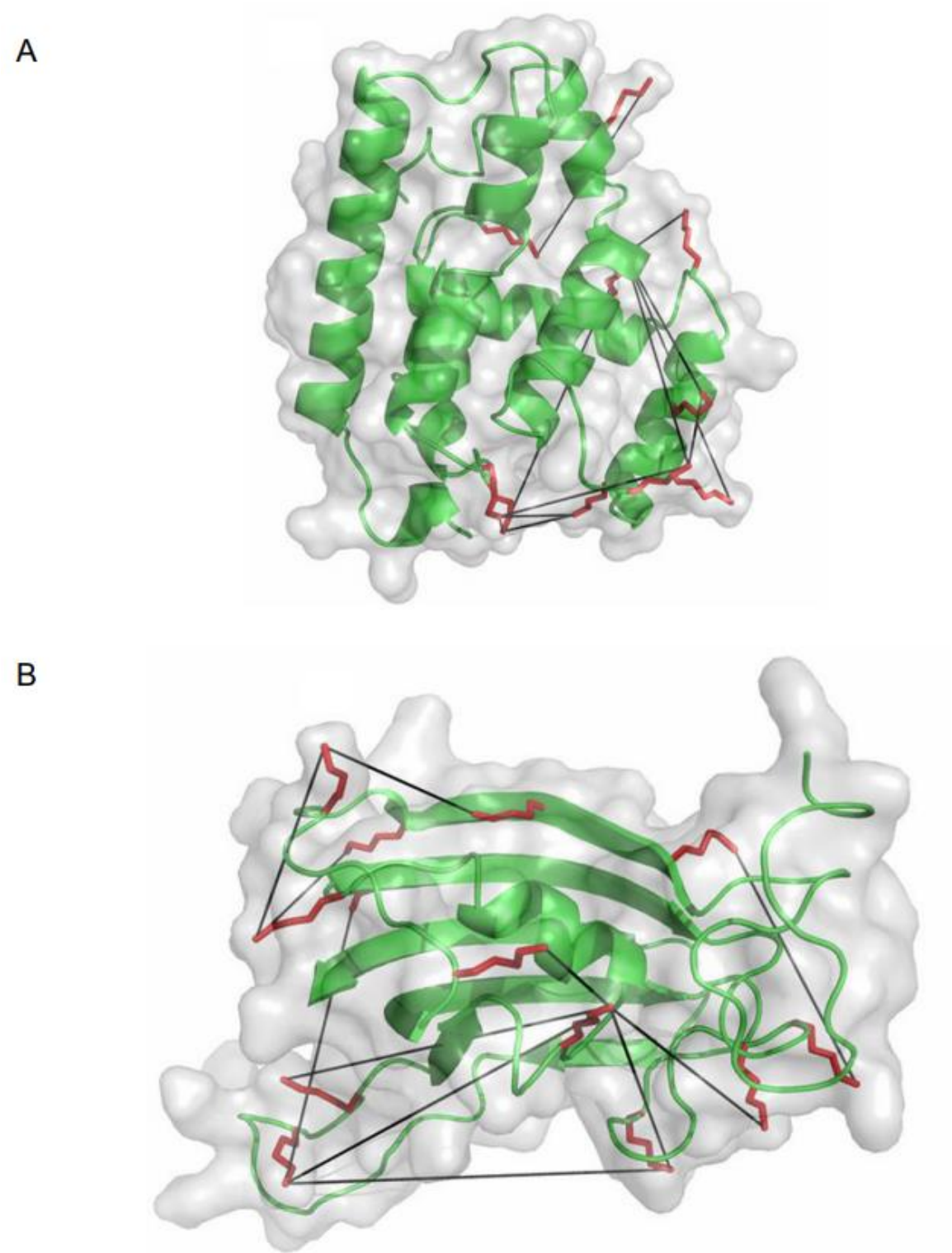


Figure 25: Long-distance crosslinking of Myoglobin and FKBP with CBDPS

Myoglobin (A) and FKBP (B) were crosslinked with CBDPS. All crosslinks were within the 25 Å maximum range of the crosslinker on the structure.

3.4. Conclusions

In this chapter, I have described a method for the determination of protein structures based on short-distance crosslinking constraints-guided DMD simulations followed by validation of the obtained structures by CD, HDX, SM, and LD-CL data. I have tested the proposed approach on the mainly α- and mainly β-structure proteins – Mb and FKBP respectively – and have obtained agreement of the calculated results with known structures of these proteins. Experimentally-determined inter-residue distance data provide valuable structural information and have the potential to be helpful in any computational approach, including in conventional MD or NMR structure prediction algorithms [157, 174].

DMD provides computational efficiency and discrete representation of potential energies. DMD simulations naturally allow the generation of conformational ensembles satisfying only a portion of the constraints. This makes DMD a perfect computational platform for the methodology proposed in this study. Short-distance constraints were directly incorporated into the DMD force field energy function, thus influencing the entire folding process. This allows the software to restrict the conformational space and achieve the folding of native structures on a practical time scale.

I believe that both short-distance crosslinks and the DMD algorithm are essential for the success achieved by the predictions. In fact, when we attempted simulations of the Mb and FKBP proteins without any constraints, we were not able to find any close-to-native structures for a simulation time of 3×10^6 steps. We also found that the incorporation of long-distance constraints (>25 Å) did not have any noticeable effect on the simulations (when compared to non-constrained simulations) because the length of

the crosslink was comparable to the size of the protein. Long-distance constraints can, however, be used for additional validation of the predicted models.

The coexistence of ensembles of structures, where only a portion of the constraints are satisfied, is an intrinsic property of the discrete energy representation of DMD. The algorithm energetically penalizes structures where the distances between the atoms do not satisfy the experimentally obtained constraints (see Section 3.2.2. Computational Methods). However, there is no continuous force during the simulations which would drive the system to a single state that satisfies all of the constraints. Instead, our method allows the generation of possible conformational ensembles, to which different energy scores are assigned by the Medusa force field function.

For well-structured proteins, such as those presented in this chapter, we observe clear separation of the low-energy clusters and a narrow distribution of structures within the clusters. However, in the case of intrinsically disordered proteins, multiple CL-DMD clusters with similar energies are observed (See Chapter 4). These clusters probably represent coexisting ensembles of conformations. Analysis of the structures within each cluster reveals some aspects of structural dynamics. In Figure 19, we show a tube diagram that indicates particular regions of the proteins that have higher flexibility. Regions of increased flexibility can be located by thicker tubes or by "blurring" of areas of the contact map. The contact frequency map in Figure 19 indicates how often each particular inter-residue contact appears in the different structures within the cluster.

In the current form of the approach, only crosslinking data were used as constraints. It may be possible to add other types of the experimental data to the DMD simulations as additional constraints. Previously, it was shown that limited proteolysis data can be

formalized for incorporating into DMD [175]. Secondary-structure information from HDX, in cases where it is possible to distinguish between α- and β-structures, can also potentially be incorporated into the algorithm, especially if the data are from high-resolution experiments where it is possible to delineate the boundaries of the secondary structure motifs at single-residue resolution [56]. If residue exposure information from SM experiments can be formalized and incorporated into the algorithm, this would be another valuable addition to the procedure. All of this experimental information would enhance the computational power of the approach, which would be advantageous for the application of CL-DMD to larger proteins or systems.

In summary, I have introduced CL-DMD as a method for the determination of unknown protein structures. I hope that this method will find its place in the protein structure-determination field, especially for cases where standard structural biology methods are not applicable.

Chapter 4: Conformational ensemble of native α-synuclein in solution as determined by short-distance crosslinking constraint-guided discrete molecular dynamics simulations

Work in this chapter was performed in the laboratories of Dr. Christoph Borchers and Dr. Nikolay Dokholyan. The crosslinking, surface modification and hydrogen deuterium exchange of α-synuclein was performed by Nicholas Brodie. All LC -MS/MS experiments and data analysis thereof was performed by Nicholas Brodie. All discrete molecular dynamics simulations were performed by Konstantin Popov. Experimental design was performed by Nicholas Brodie, Evgeniy Petrotchenko, Konstantin Popov, Nikolay Dokholyan and Christoph Borchers. Christoph Borchers oversaw this project. This chapter was adapted from the publication:

Brodie, Nicholas; Popov, Konstantin; Petrotchenko, Evgeniy; Dokholyan, Nikolay; Borchers, Christoph. **Conformational ensemble of native α-synuclein in solution as determined by short-distance crosslinking constraint-guided discrete molecular dynamics simulations**. PLOS Computational Biology. e1006859. March 27, 2019. **Doi:** 10.1371/journal.pcbi.1006859

4.1. Introduction to the crosslinking of α-synuclein

α-Synuclein is involved in the pathogenesis of misfolding-related neurodegenerative diseases, particularly Parkinson's disease [82, 83]. A misfolding event leads to the formation of oligomers which are believed to result in cell toxicity and which eventually leads to the death of neuronal cells [176]. α-Synuclein is thought to interact with lipid vesicles *in vivo* [177] and the toxicity is thought to be caused by membrane disruption initiated by misfolded oligomers [178]. Moreover, a prion-like spread of the pathology via the conversion of native α-synuclein molecules by toxic oligomers has been suggested [179]. Native α-synuclein is considered to be an intrinsically disordered protein, although there is evidence that some globular structure exists in solution, this globular structure might serve as a basis for understanding the misfolding and oligomerization pathways. A number of biophysical methods, such as NMR, electron paramagnetic resonance (EPR), FRET, and small-angle X-ray scattering (SAXS) – in combination with computational methods – have been applied to the study of intrinsically disordered proteins, including the structure of α-synuclein in solution [118, 119, 180, 181]. In all of these cases, even a limited amount of experimental structural data was necessary for the characterization of the conformational ensemble of α-synuclein in solution.

In the previous chapter, I described the development of a method for determination of protein structures, termed short-distance crosslinking constraint-guided discrete molecular dynamics simulations (CL-DMD), where the folding process is influenced by short-distance experimental constraints which are incorporated into the DMD force field [182]. Adding constraints to DMD simulations results in a reduction of the possible conformational space and allows the software to achieve protein folding on a practical

time scale. This approach was tested on well-structured proteins including Mb and FKBP, and a clear separation between low-energy clusters and a narrow distribution of structures within the clusters was observed. The conformational flexibility of intrinsically disordered proteins, such as α-synuclein, brings additional challenges to the computational process [156]. In cases like this, proteins exist as a collection of inter-converting conformational states, and the crosslinking data represents multiple conformations of a protein rather than a single structure. In addition, recent research indicates that traditional force fields, with their parameterization, are not ideal for providing an accurate description of disordered proteins, and tend to produce more compact structures [183]. Recent research has focused on improving traditional state-of-the-art force fields and their ability to predict the structures of disordered proteins without losing their accuracy for structured proteins [184]. In this work, my collaborators used a Medusa force field [73, 74, 185] in the DMD simulations and this force field was discretized to mimic continuous potentials. The Medusa force field utilizes 6 different forces including attractive and repulsive van der Waals interactions, a solvation energy, and 3 different forces for hydrogen bonding (those between backbone atoms, those between side chains, and those between backbone and sidechain atoms). DMD uses a united atom representation for the protein, where all heavy atoms and polar hydrogens are explicitly accounted for. The solvation energy is described in terms of the discretized Lazaridis-Karplus implicit solvation model [163] and inter-atomic interactions, such van der Waals and electrostatics, are approximated by a series of multistep square-well potentials.

Additional potentials, such as pair-wise distance constraints [172, 186] and solvent accessibility information [187, 188] can also be readily integrated. During CL-DMD simulations there are no continuous forces that would drive the atoms to satisfy all constraints. Instead, conformational ensembles are generated which satisfy an optimal number of the constraints. This, to some degree, naturally resolves conflicting experimental constraints. Thus, CL-DMD simulations are a viable computational platform for the structural analysis of intrinsically disordered proteins in general [189], and of α-synuclein in particular.

Here, my collaborators and I used the CL-DMD approach [182] to determine conformational ensembles of the α-synuclein protein in solution. During this process, α-synuclein was crosslinked with a panel of short-range crosslinkers, crosslinked proteins were enzymatically digested, crosslinked residues were determined by LC-MS/MS analysis, and the resulting data on inter-residues distances were introduced into the DMD force field as external constraints. To experimentally validate the predicted structures, we analyzed α-synuclein using surface modification, circular dichroism, hydrogen-deuterium exchange, and long-distance crosslinking.

4.2. Materials and Methods

All materials were from Sigma-Aldrich, unless noted otherwise.

4.2.1. Structural Proteomics

α-Synuclein was crosslinked with a panel of short-range reagents azido-benzoic acid succinimide (ABAS-$^{12}C_6$/$^{13}C_6$), succinimidyl 4,4'-azipentanoate (SDA), [139] triazidotriazirine (TATA-$^{12}C_3$/$^{13}C_3$), [140] and 1-ethyl-3-(3-

dimethylaminopropyl)carbodiimide (EDC) [17]. ABAS and SDA are hetero-bifunctional amino group-reactive and photo-reactive reagents, TATA is a homo-bifunctional photo-reactive reagent, and EDC is a zero-length carboxyl-to-amino group crosslinker. Crosslinked proteins were digested with the proteolytic enzymes proteinase K or trypsin, and the digest was analyzed by LC-MS/MS to identify crosslinked peptides (Table S1). I used an equimolar mixture of ^{14}N- and ^{15}N-metabolically labelled α-synuclein to exclude potential inter-protein crosslinks from the analysis and to facilitate the assignment of crosslinked residues based on the number of nitrogen atoms in the crosslinked peptides and in the MS/MS fragments [190]. The distances between the crosslinked residues were based on the length of each crosslinker reagent, and were introduced as constraints into the DMD potentials (see the section below and [182] for additional details). A total of 30 crosslinking constraints were used in these DMD simulations (Appendix A). In addition, α-synuclein was characterized by top-down ECD- and UVPD-FTMS HDX and CD to determine the secondary-structure content (Figure 24 and Appendix B). Quantitative differential surface modification experiments were performed with and without 8 M urea to determine whether the residues were exposed or buried (Appendix C). LD-CL was used to estimate the overall protein topology (Appendix D).

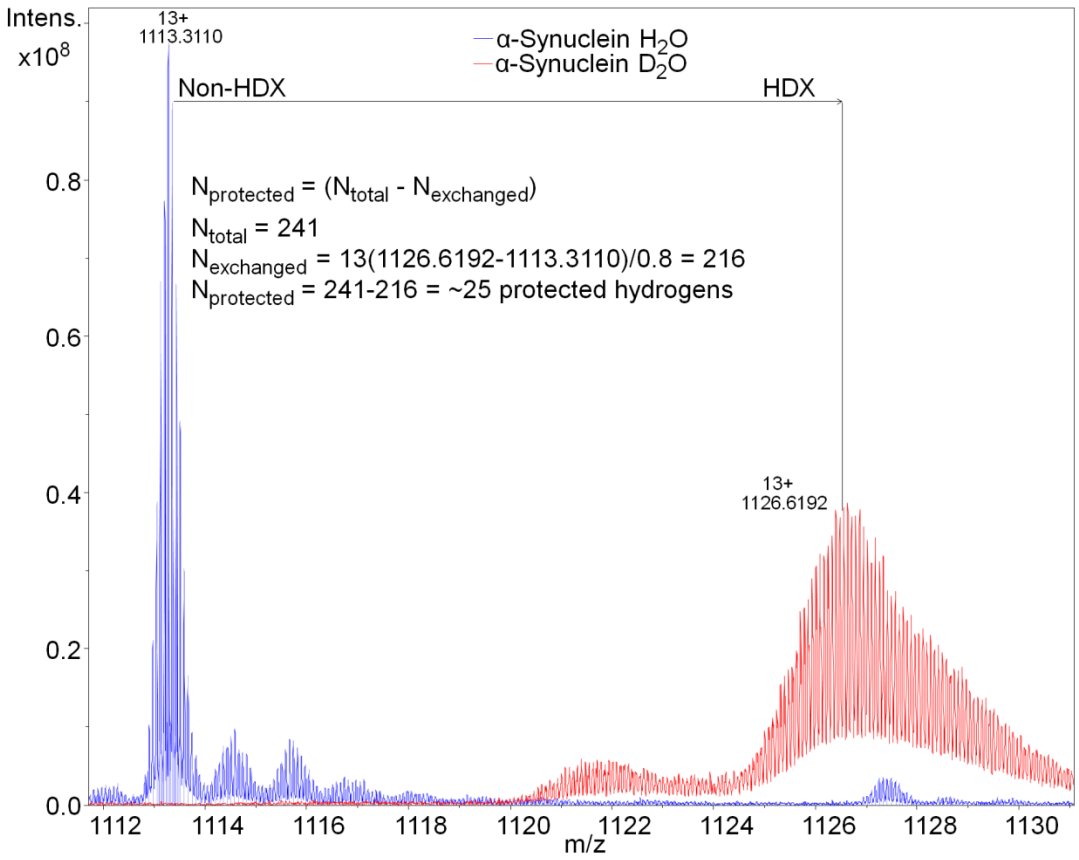


Figure 26: Hydrogen-deuterium exchange of α-synuclein

The total number of protected hydrogens was measured to be approximately 25 at 2s of exchange, indicating the existence of transient secondary structure.

4.2.2. Expression and purification of α-synuclein

α-Synuclein was expressed using a pET21a vector provided by Dr. Carol Ladner of the University of Alberta. The protein was expressed in *E. coli* BL21(DE3) bacteria and was purified as in [140]. Briefly, the protein was overexpressed with 1 mM IPTG in 1L LB cultures of BL21DE3 E. coli for 4 hours at 30 °C. Cells were lysed with a French press and the lysate was heated at 70 °C for 10 minutes and then centrifuged at 14000 g for 30 minutes. The soluble fraction was precipitated for 1 hour in 2.1 M $(NH_4)_2SO_4$. α-Synuclein was then purified by fast protein liquid chromatography on a Mono Q 4.6/100

SAX column (GE Life Science), using a gradient from 50-500 mM NaCl, in 50 mM Tris at pH 8.0. Elution fractions containing α-synuclein were further purified by size exclusion on a Superdex 200 30/100 GL column (GE Life Science). For the expression of metabolically labelled ^{15}N α-synuclein, 1 L of M9 Minimal media was prepared with 1 g/L $^{15}NH_4Cl$ (Cambridge Isotopes) as the sole source of nitrogen. BL21(DE3) cells were grown overnight in 50 mL of this media, then seeded into 1 L, grown to an A600 of approximately 0.8, and induced using 1 mM IPTG. After expression overnight at 30 °C, ^{15}N α-synuclein was purified as described above.

4.2.3. Crosslinking

Unlabelled and ^{15}N metabolically-labelled α-synuclein were mixed in a 1:1 ratio at a concentration of 20 µM in 50 mM Na2HPO4 and incubated overnight at room temperature prior to crosslinking. α-Synuclein aliquots of 38 µL each were then crosslinked using either 1 mM of the ABAS-$^{12}C_6/^{13}C_6$ crosslinker (Creative Molecules) or 30 mM of the EDC crosslinker. ABAS crosslinking reaction mixtures were incubated for 10 minutes in the dark to allow the NHS-ester reaction to take place, followed by 10 minutes of UV irradiation under a 25 W UV lamp (Model UVGL-58 Mineralight lamp, UVG) with a 254 nm wavelength filter. EDC reaction mixtures were incubated for 20 minutes. Reaction mixtures were quenched with 10 mM ammonia bicarbonate. A portion of each crosslinking reaction mixture was checked by SDS-PAGE to determine the extent of potential intermolecular crosslinked products. Aliquots were subsequently split and digested with either trypsin or proteinase K at an enzyme:protein ratio of 1:10. Digestion was stopped by adding AEBSF (ApexBio) to a final concentration of 10 mM, and the samples were then acidified with formic acid for analysis by mass spectrometry. For

TATA, 100 µM synuclein in 50 mM sodium phosphate buffer was reacted with 0.5 mM TATA-$^{12}C_3/^{13}C_3$ (Creative Molecules). Samples were incubated for 5 minutes with 254 nm UV light from the same lamp as was used for the ABAS reactions. Samples were then split and digested with either proteinase K or trypsin at an enzyme:protein ratio of 1:20. For the SDA reactions, 20 µL of 1 mg/mL α-synuclein was crosslinked using 1 mM SDA (Creative Molecules, Inc.). Aliquots were incubated for 15 minutes in the dark prior to incubation under the same UV lamp as used previously for ABAS reactions, but changing the wavelength to 366 nm. Samples were quenched with 10 mM ammonium bicarbonate then run by SDS-PAGE, and bands representing the α-synuclein monomer were excised and subjected to in-gel trypsin digestion. After in-gel digestion, samples were acidified using formic acid prior to mass spectrometric analysis. The CBDPS crosslinking reaction mixture consisted of 238 µL of 50 µM α-synuclein, with 0.12 mM CBDPS. Samples were split and digested with either proteinase K or trypsin at an enzyme:protein ratio of 1:10. Digests were quenched with 10 mM AEBSF and samples were enriched using monomeric avidin beads (Thermo Scientific). Enriched samples were acidified for mass spectrometric analysis using formic acid.

4.2.4. LC-MS/MS analysis

Mass spectrometric analysis was then performed using a nano-HPLC system (Easy-nLC II, ThermoFisher Scientific), coupled to the ESI-source of an LTQ Orbitrap Velos or Fusion (ThermoFisher Scientific), using conditions described previously [182]. Briefly, samples were injected onto a 100 µm ID, 360 µm OD trap column packed with Magic C18AQ (Bruker-Michrom, Auburn, CA), 100 Å, 5 µm pore size (prepared in-house) and desalted by washing with Solvent A (2 % acetonitrile:98 % water, both 0.1 % formic acid

(FA)). Peptides were separated with a 60-min gradient (0–60 min: 4–40 % solvent B (90 % acetonitrile, 10 % water, 0.1 % FA), 60–62 min: 40–80 % B, 62–70 min: 80 % B), on a 75 μm ID, 360 μm OD analytical column packed with Magic C18AQ 100 Å, 5 μm pore size (prepared in-house), with IntegraFrit (New Objective Inc., Woburn, MA) and equilibrated with solvent A. MS data were acquired using a data-dependent method. The data dependent acquisition also utilized dynamic exclusion, with an exclusion window of 10 ppm and exclusion duration of 60 seconds. MS and MS/MS events used 60000- and 30000-resolution FTMS scans, respectively, with a scan range of m/z 400-2000 in the MS scan. For MS/MS, the CID collision energy was set to 35 %. Data were analyzed using the 14N15N DXMSMS Match program from the ICC-CLASS software package [190]. SDA crosslinking data was analyzed using Kojak [46] and DXMSMS Match [45]. For scoring and assignment of the MS/MS spectra, b- and y-ions were primarily used, with additional confirmation from CID-cleavage of the crosslinker where this information was available.

4.2.5. Differential surface modification

Chemical surface modification with pyridine carboxylic acid N-hydroxysuccinimide ester (PCAS) (Creative Molecules, Inc.) was performed as described previously [62]. Briefly, α-synuclein was prepared at 50 μM in 8 M urea in PBS, pH 7.4 (unfolded state), or in only PBS (folded state). Either the light or the heavy form of the ^{13}C-isotopically-coded reagent (PCAS-$^{12}C_6$ or PCAS-$^{13}C_6$) was then added to give a final concentration of 10 mM. Reaction mixtures were incubated for 30 minutes and quenched with 50 mM ammonium bicarbonate. Samples were then mixed at a 1:1 ratio, combining folded (PCAS-$^{12}C_6$) with unfolded (PCAS-$^{13}C_6$) samples, as well as in reverse as a control.

Samples were acidified with 150 mM acetic acid and digested with pepsin at a 20:1 protein:enzyme ratio overnight at 37 °C. After digestion, samples were prepared for mass spectrometry analysis using C18 zip-tips (Millipore). Zip-tips were equilibrated with 30 µL 0.1 % TFA, sample was introduced, then washed with 30 µL 0.1 % TFA and eluted with 2 µL of 0.1 % formic acid/50 % acetonitrile. Samples were analyzed by LC-MS/MS as described above.

4.2.6. Hydrogen/deuterium exchange

Top-down ECD-FTMS hydrogen/deuterium exchange was performed as described previously [56]. Briefly, protein solution and D2O from separate syringes were continuously mixed in a 1:4 ratio (80% D2O final) via a three-way tee which was connected to a 100 µm x 5 cm capillary, providing a labelling time of 2 seconds. The outflow from this capillary was mixed with a quenching solution containing 0.4 % formic acid in 80% D_2O from the third syringe via a second three-way tee, and injected into a Bruker 12T Apex-Qe hybrid Fourier Transform mass spectrometer, equipped with an Apollo II electrospray source. In-cell ECD fragmentation experiments were performed using a cathode filament current of 1.2 A and a grid potential of 12 V. Approximately 800 scans were accumulated over the m/z range 200-2000, corresponding to an acquisition time of approximately 20 minutes for each ECD spectrum. Deuteration levels of the amino acid residues were determined using the HDX Match program [170] (Appendix B).

Synuclein UVPD spectra were collected on a Thermo Scientific Orbitrap Fusion Lumos Tribrid mass spectrometer equipped with a 2.5-kHz repetition rate (0.4 ms/pulse) 213 nm Nd:YAG (neodymium-doped yttrium aluminum garnet) laser (CryLas GmbH)

with pulse energy of 1.5 ± 0.2 μJ/pulse and output power of 3.75 ± 0.5 mW for UVPD. The solution was exchanged with deuterium using the same three-way tee setup, although in this case a 50 μm x 7cm capillary provided a labelling time of ~1 s. Spectra were acquired for 8 or 12 ms, and resultant spectra were averaged and used for the data analysis with the HDX Match program as above.

4.2.7. Circular dichroism

CD spectra were recorded on Jasco J-715 spectrometer under a stream of nitrogen. The content of α-helical and β-sheet structures was calculated using a BeStSel web server [169].

4.2.8. Discrete molecular dynamics modelling

CL-DMD simulations were performed according to the protocol described in the previous chapter (See Chapter 3). Briefly, discrete molecular dynamics (DMD) is a physically based and computationally efficient approach for molecular dynamics simulations of biological systems [73, 74]. In DMD, continuous inter-atom interaction potentials are replaced with their discretized analogs, allowing the representation of interactions in the system as a series of collision events where atoms instantaneously change their momenta according to conservation laws. This approach significantly optimizes computations by replacing integration of the motion equations at fixed time steps with the solution of conservation-law equations at event-based time points [159]. In order to incorporate experimental data for inter-residue distances between corresponding atoms into DMD simulations, my collaborators introduced a series of well-shape potentials that energetically penalize atoms whose interatomic distance do not satisfy experimentally determined inter-atom proximity constraints. The widths of these

potentials are determined by the crosslinker spacer length and the side chain flexibility of crosslinked amino acids [182]. Starting from the completely unfolded structure of α-synuclein molecule, we performed an all-atom Replica Exchange (REX) [165] simulations of the protein where 24 replicas with temperatures equally distributed in the range from 0.375 to 0.605 kcal/(mol kB) are run for 6 x 10^6 DMD time steps. The simulation temperature of each of the replicates periodically exchanged according to the Metropolis algorithm allowing the protein to overcome local energetic barriers and increase conformational sampling. During the simulations, the specific heat curve of the system energy distribution was calculated by a Weighted Histogram Analysis Method (WHAM) [167] which was used as the indicator of system equilibration. The first 2 x 10^6 time steps of system equilibration during the analysis were discarded. Next, all of the structures among all of the trajectories were ranked, and the ones with the lowest 10% of the energies, as determined by the DMD Medusa force field [173], were selected. These structures were then clustered using the GROMACS [168] distance- based algorithm described by Daura et al. [191]. It uses root-mean-square deviation (RMSD) between backbone Cα atoms as a measure of the structural similarities between the structures within a cluster. An RMSD cut-off was chosen to correspond to the peak of the distribution of pair-wise RMSDs for all of the low-energy structures. Because the energies of the resulting centroids – which are representative of the clusters – are very close to each other (Appendix E), and picking one of them would potentially introduce a bias related to our scoring energy function, we present all of them as our predicted models of the α-synuclein globular structure. The root-mean-square deviation of atomic positions within each cluster was calculated and this was used as a measure of

fluctuations of the structures of corresponding centroids (Figures 27 and 28). In order to obtain information on the global folding of α-synuclein, we performed clustering analysis on the lowest-energy structures obtained during CL-DMD simulations.

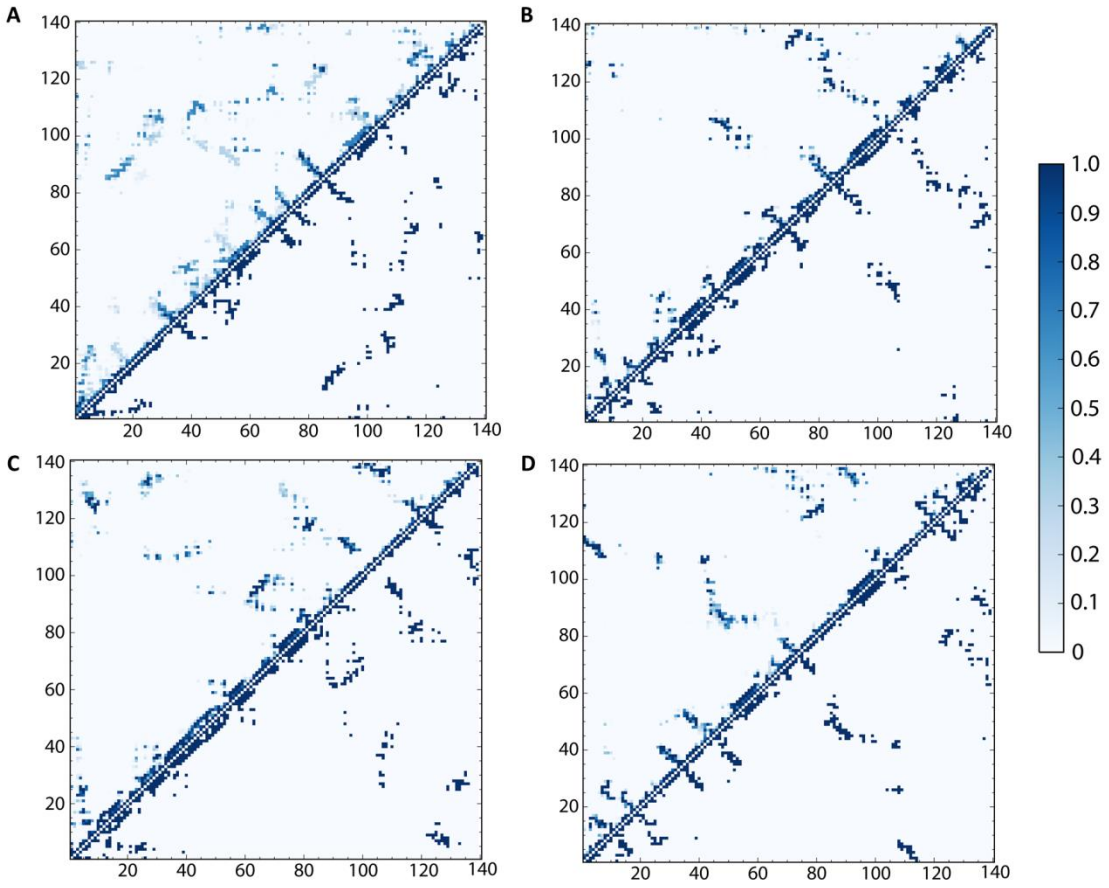


Figure 27: Contact frequency maps for representative clusters of α-synuclein models

Plots (A-D) represent static and frequency contact maps for the major representative clusters of protein structures during the simulations (populations: cluster 1 ~37%, cluster 2 ~28%, cluster 3 ~20%, cluster 4 ~8%). Points below the diagonals correspond to contacts between residues. Note that 0 means that the atoms are not in contact; 1 means that the two atoms are in contact. Two residues form a contact if their Cα atoms are within 8 Å of each other. Data points above the diagonal in plots (A-D) indicate how often each particular contact between two residues can be found within the clusters for the corresponding centroid structures.. The colour map quantifies how often these contact form, where white/zero means that contacts never form, and dark blue/one means that contacts always form).

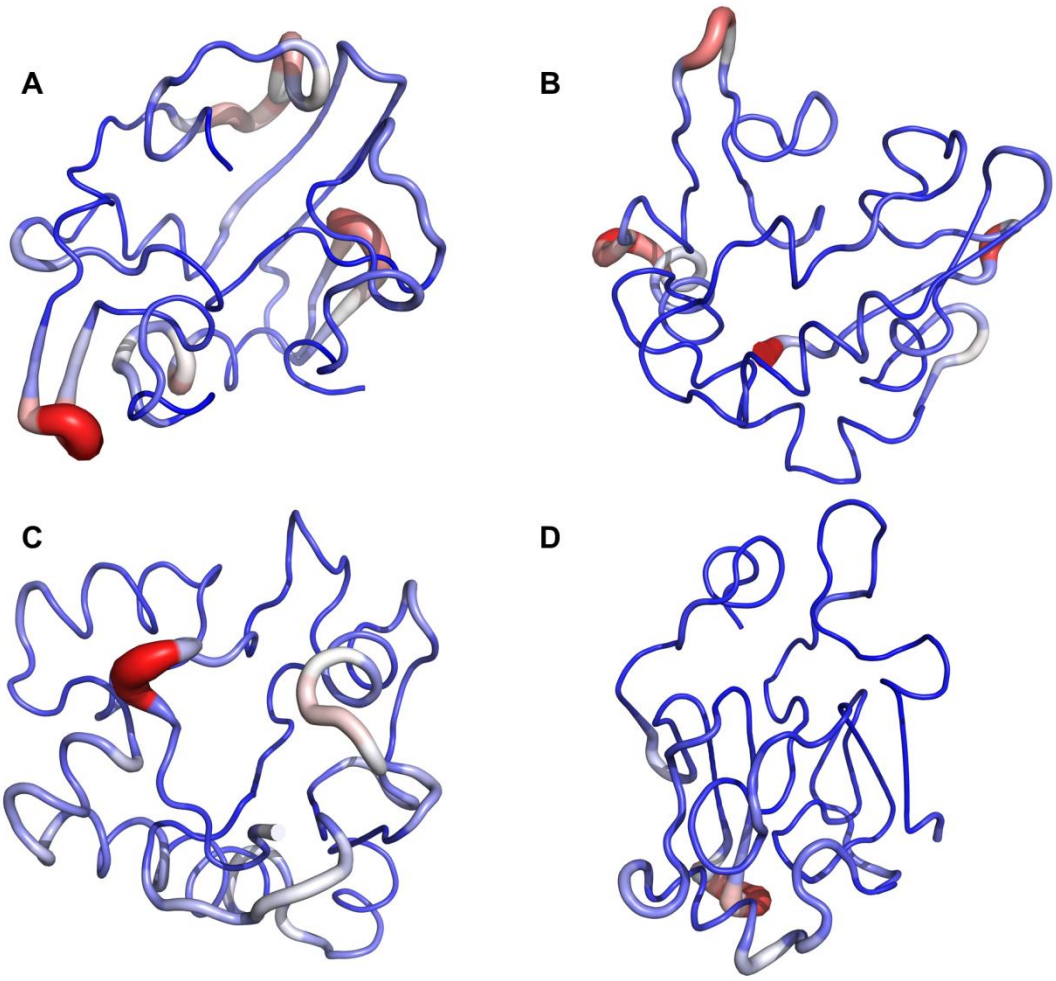


Figure 28: Tube representation of the fluctuations of the clusters

Tube models A-D are representative structures of the clusters which represent 37% (A), 28% (B), 20% (C), and 8% of the total lowest energy structures , respectively. The thickness and colour of the tubes indicate the dynamics of the corresponding regions within the each cluster during the simulations. The tubes are coloured from blue (low flexibility) to red (high flexibility).

4.3. Results and Discussion

In contrast to ordered proteins, where the lowest energy structures are usually represented by a few centroids with distinct conformations close to the corresponding native structure [182], disordered proteins are usually represented by a broader variety of structures, which are distinct from each other, reflecting the conformational freedom of the intrinsically disordered protein [192]. In this case, the lowest energy structures

forming the conformational ensemble for α-synuclein are shown as overlays in Figure 29A.

4.3.1. The α-synuclein ensemble

The α-synuclein ensemble we have modelled can be described as 4 compact globular structures with a common general topology, containing a few distinct secondary structure fragments. Overall, a distinct topology can be observed for all these conformers (Figure 29B). The protein forms a three-prong closed claw-like shape with the N-terminal (blue), intermediate (yellow), and C-terminal (red) subdomains having converged at the top, and with the connecting subdomain (green) being located at the bottom of the structure (Figure 29). The C-terminal portion of the molecule was found in close proximity to the N-terminal portion, possibly reflecting long-range electrostatic interactions between these subdomains (Figure 29) [118]. The C-terminal portion was also in contact with the intermediate subdomain – in some conformers it protruded deeper into the structure and was positioned between the N-terminal and the intermediate subdomains. It should be noted that I used non-N-acetylated α-synuclein protein in this study, although in the cell, α-synuclein is predominantly N-acetylated [117]. This post-translational modification may have some effect on the protein structure by increasing the propensity of the N-terminus to form an α-helix [193]. This may affect the aggregation of synuclein under some conditions [194, 195]. Acetylation, however, does not appear to have a significant effect on the propensity of synuclein to aggregate [196] under lipid-free conditions, which seems to indicate that the structural features important needed for predicting aggregation and designing small molecule inhibitors of misfolding are also present in the absence of acetylation.

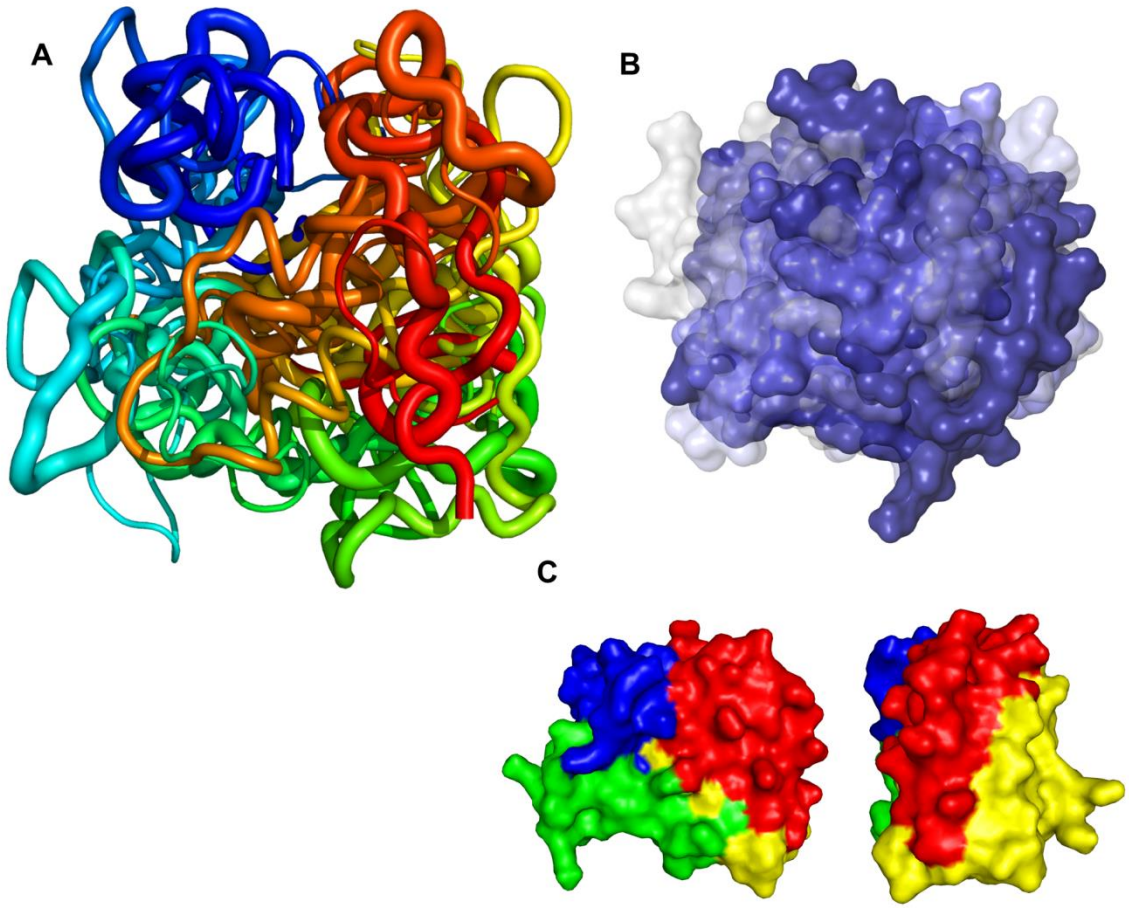


Figure 29: Structure of native α-synuclein in solution as determined by CL-DMD

A. Conformational ensemble of α-synuclein. Representatives of the four major clusters are aligned and coloured N-to-C from blue-to-red. The thickness of the cartoon representations correspond to the content of approximately 37%, 28%, 20% and 8% of the overall population of the structures in each cluster. B. Surface representation of the ensemble. Representatives of the four major clusters as in panel A are aligned and coloured from blue to white according to the percentage of all structures found in each cluster. C. Subdomain topology of the α-synuclein conformational ensemble. Residues are coloured as follows: 1-27 blue, 28-52 green, 53-108 yellow, 109-140 red.

Some of the experimental constraints may conflict with each other due to the fact that proteins exist in multiple conformations. During DMD simulations, the total potential is minimized, satisfying an optimal number of constraints and generating an ensemble of structures satisfying various subsets of the experimental constraints. This ensemble is further clustered into substates that represent subsets of non-conflicting constraints.

Depending on the conformation and the number of constraints that are satisfied, these structures will have different energy scores assigned by the force field calculation [73, 173]. This allows the software to account for both the nature of conformational changes of the protein in solution (i.e., while the structure of the protein is "breathing" and transitioning between conformations), and possible errors that occurred during the determination of the experimental constraints [156]. Nevertheless, it was found that the lowest-energy conformers satisfy most of the experimental distance constraints (Appendix F), while the models with higher energy satisfy a lower number of constraints. No quantitative crosslinking data or weighting of observed crosslinks was used in the simulations, but more highly-populated conformations have a higher chance of being represented in the resulting crosslinking dataset, given that crosslinks are not randomly distributed and are most likely to form between residues which spend a significant amount of time within close contact. A comparison of predicted structures within the overall ensemble revealed that the conformational changes to the protein structure were mainly due to movement of the large loops and hairpins constituting the N-terminal, C-terminal, and intermediate subdomains.

4.3.2. Ensemble validation

The validity of this ensemble was tested by removing all of the crosslinking constraints and allowing the structure to relax over $2x10^6$ steps, to see if the structure was stabilized only by the potentials enforcing satisfaction of the experimental crosslinks. The predicted representatives of the clusters were allowed to fluctuate in this way, and the result was a relatively minor expansion of the protein (results for the centroid of the most representative lowest energy cluster, Figure 28A, are shown for clarity), from a radius of

gyration of 14.1 Å initially, to an average of 15 Å (Appendix G). This indicates that the structure has not been overfit to the experimental data during the simulation, which is often a problem when modelling intrinsically disordered proteins with experimental distance constraints [184].

The crosslinking analysis applied here is conceptually similar to paramagnetic relaxation enhancement nuclear magnetic resonance (PRE-NMR) [117, 118, 197, 198] or FRET [199, 200] techniques used for the prediction of the α-synuclein conformational ensembles, with pairwise distances between crosslinked residues being analogous to the pairwise distances between the spin-label and the atomic nuclei of the protein in case of PRE-NMR (Appendix H) or between fluorophores, in the case of FRET. The centroids of the low-energy clusters in our ensemble tend to be more compact than those based on PRE-NMR (Appendix H), FRET ensembles, or SAXS data. It has been suggested that the higher values of the radius of gyration determined by these other techniques may be caused by an existing equilibrium between the monomeric and multimeric states of the synuclein protein under the experimental conditions used [201, 202]. SAXS data can be uniquely sensitive to the more-extended conformations of the ensemble [201]. Unconstrained all-atom molecular dynamics simulations produced multimodal distributions which included compact states [201], although it was noted that modelling disordered proteins may require the development of specialized force fields [195, 201]. It was also shown that the addition of experimental distance constraints to the all-atom Monte Carlo conformational search leads to more compact structures [201]. Thus, representation of the slightly more compact states may still be caused by bias due to the

short-range crosslinking reagents used in this study, as well as by the traditional parameterization of the Medusa force field [73, 74, 185].

In addition, the modellers selected the lowest-energy structures for the clustering and model selection, which eliminated some of the unfolded less-compact states, while in the PRE-NMR based approach, the authors used all of the generated structures for the analysis. The other difference between these two techniques is the range of distance constraints produced. The PRE-NMR technique uses distances up to ~2 nm and local residue-specific conformational preferences are not well reproduced in the PRE-RMD ensembles [198]. When applying CL-DMD to well-structured model proteins, i.e., proteins with known crystal structure, my collaborators had also found that long-distance constraints (>15 Å) were not particularly helpful in finding the true protein structure solutions [182]. Short-distance constraints were critical in finding the correct local conformations for well-structured proteins.

4.3.3. α-Synuclein secondary structure

Some transient secondary-structure was observed in the conformers of the ensemble (Figure 30). The extent and the location of the predicted secondary-structure motifs in the lowest-energy conformer were in good agreement with the experimental data obtained by HDX and CD. Thus, ~25 of the total number of protected protons in the whole protein were detected by HDX (Figure 26), and ~2.4% of α- and 29.1% of β-structure was detected by CD. Interestingly, I was able to detect amide HDX protection only when the exchange time was reduced from ten seconds to two seconds, which indicates the transient nature of the secondary structure observed. The location of the predicted secondary structure within the N-terminal portion of the protein was confirmed by

determining the protection status of individual residues using the c-ion fragment series (i.e., the fragments starting from N-terminus) as obtained by ECD MS/MS (Figure 26 and Appendix B). Here, I observed some propensity for the protein to form α-helical secondary structure near the N-terminus, particularly between residues 25-55. However, it is not nearly as extensive as the secondary structure formed in the presence of detergent micelles (Figure 9) [96]. In particular, two centroids (Figure 30B and 30C) showed a greater tendency towards α-helicity, and together represent some 48% of the ensemble.

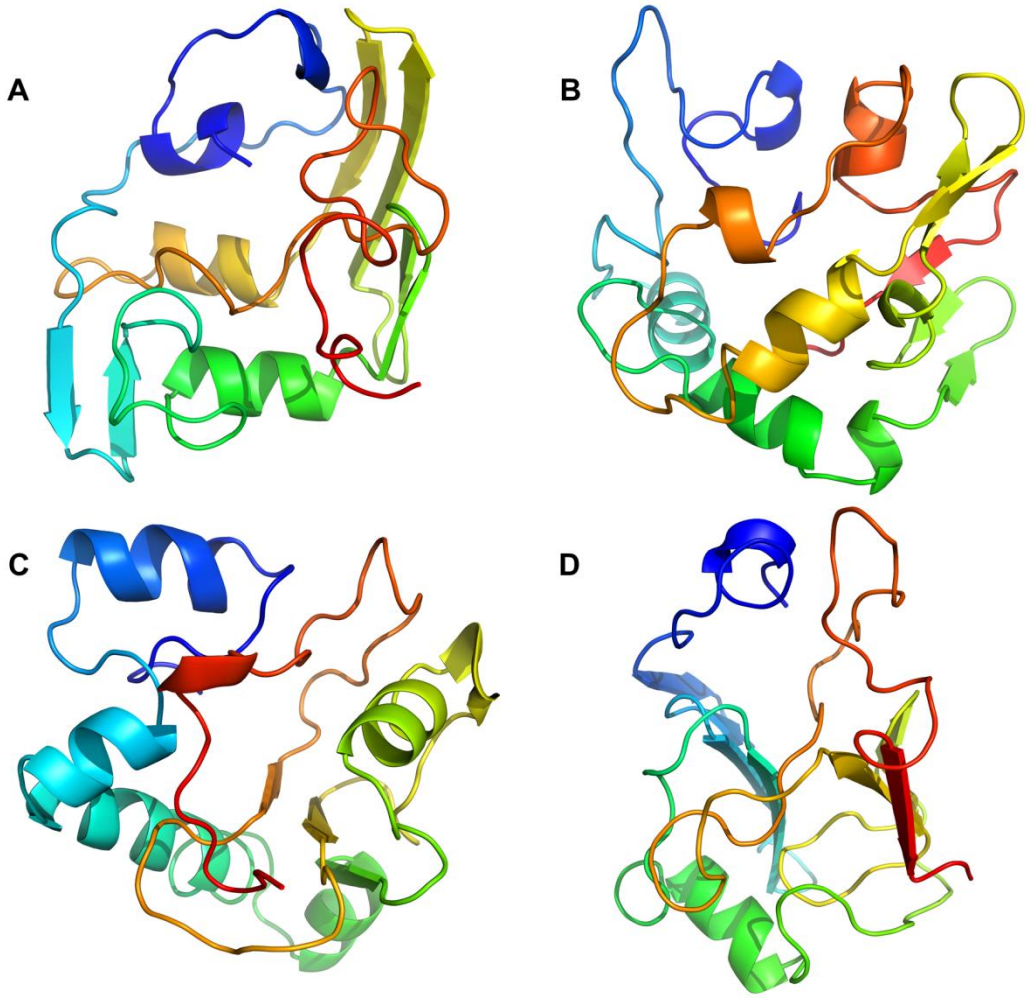


Figure 30: Comparison of the transient secondary structure in the α-synuclein conformational ensemble

Shown here are representative conformers of the α-synuclein ensemble shown in Figure 28. Structures A, B, C, and D represent clusters containing approximately 37%, 28%, 20%, and 8% of the overall population of the structures, respectively. Structures are coloured from blue to red, from the N- to C-terminus; the NAC region 61-95 corresponds to the light green-to-yellow section.

Unfortunately, for α-synuclein, I was unable to obtain a reliable z-ion fragment series (i.e., fragments starting from the C-terminus) using ECD or ETD fragmentation which would have spanned the predicted β-structural elements, probably due to the high net negative charge on the C-terminal portion of the molecule. However, the amount of protection predicted in the structure of the C-terminal part of the model, added to observed protection of the N-terminal part from ECD-HDX data, is in agreement with the total protection value of the protein (Figures 25, 27, and Appendix B).

I then used the recently developed top-down UVPD-HDX method [58] in an attempt to determine the possible locations of secondary structure in the C-terminal portion of the α-synuclein molecule. I was able to observe a number of UVPD-specific fragment ions spanning the C-terminal portion of the protein sequence – a region which had previously not been covered. Deuteration analysis of these fragments allowed me to confirm the presence of secondary structure in this region and to further narrow the location of the secondary-structure elements to residues 116-119 and 128-138, and ~10 protected residues in the residue 71-116 segment of the sequence (Figures 26, 28, and Appendix B and F). Surface modification results were in agreement with the final α-synuclein structures. As expected, lysine residues which ended up being exposed to the solvent in the model were equally modified with PCAS in the unfolded and folded states (i.e., with and without 8 M urea), and residues which were partially buried in the model exhibited a larger extent of modification in the fully unfolded state with 8 M urea (Figure 31). The

long-distance intra-protein CBDPS crosslinks were also in good agreement with the final models of the protein.

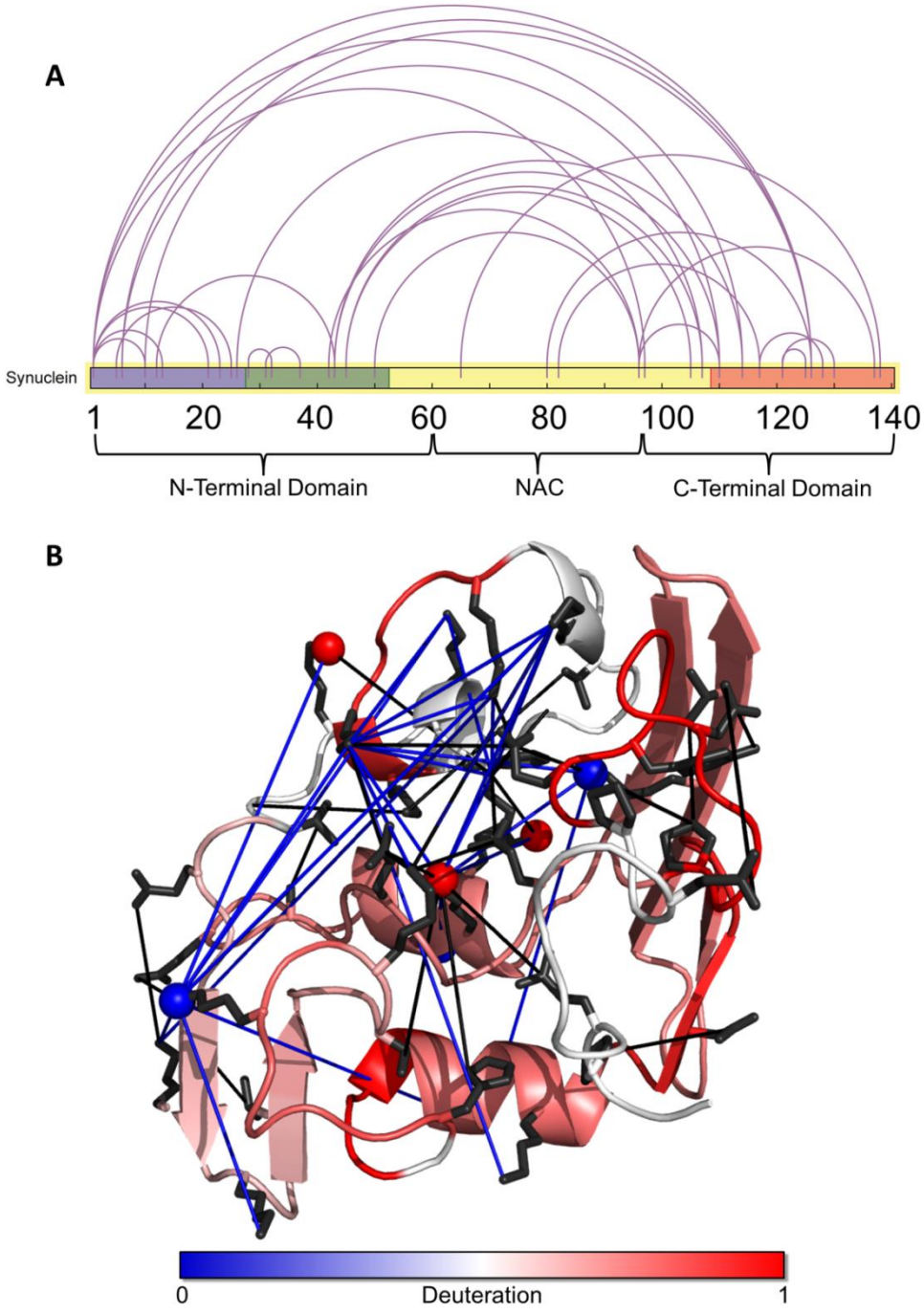


Figure 31: Experimental validation of the α-synuclein structure with SM, HDX, and LD-CL

A. A schematic diagram of the synuclein protein which each crosslink used for the CL-DMD simulation shown as a line. The diagram is coloured in accordance with the subdomain topology described in Figure 29. B. The lowest energy conformer from the ensemble as in Figure 28 is shown, with crosslinks and other validating data indicated on the structure. Short-distance crosslinks used in the CL-DMD simulations are shown as black lines. The main chain is coloured from blue to red according to the experimentally observed backbone amide protection values from the HDX experiments. Atoms that are equally modified with PCAS in the native and denatured states are shown in blue; atoms that are preferentially modified after denaturation with 8M urea atoms are shown in red. LD-CL CBDPS crosslinks are shown as blue lines.

4.3.4. Location and conformation of the NAC region in the structure

Interestingly, the aggregation-prone non-amyloid-β component (NAC) region, which is reportedly involved in the nucleation of the aggregation [203], was consistently predicted to contain hairpins in extended conformation, which were in contact with the C-terminal loop, which, in turn, interacted with the N-terminal sub-domain. It would be intriguing to see if a distortion of any of these interactions could lead to an "opening up" of this region which might facilitate stacking interactions with similar portions of other monomers during the aggregation initiation. I am currently performing CL-DMD-based differential structural characterization of α-synuclein oligomers, to compare to the native structures reported here, in order to localize critical mis-folding events leading to the pathological aggregation.

Two recent studies, by Lautenschlager et al. [177] and Fusco et al [178] have indicated that the cellular toxicity of synuclein oligomers may be dependent on both the N- and C-terminal domains and their interactions with lipid bilayers. In the toxic oligomeric species, the synuclein N-terminus has less interaction with the membrane surface, and this interaction is replaced by an interaction with the NAC region, leading to toxicity. In the case of the C-terminus, increased concentrations of calcium may create a charge-

neutralizing effect, which leads to increased toxicity and oligomer formation, possibly by disrupting its interaction with the NAC region.

The β-structure in the NAC-containing intermediate domain and in the C-terminal subdomain of the predicted ensemble is in general agreement with the C-terminal half of the protein's propensity towards formation of β-structure, as determined by NMR [180]. The fact that, in the ensemble, the NAC region was repeatedly found to contain β-structure-like hairpins may explain the propensity of this region to form inter-molecular β-structures which appear in the aggregation process [112]. These hairpins also resemble those observed in the aggregating form of Aβ1-40 found in fibrils [204], and in the mature fibrillar form of the synuclein protein [112]. In agreement with previous studies [118, 197], I also found that the NAC region was in contact with the C- terminal portion of the molecule, but in my ensemble it is not quite shielded by it, as had been suggested. Rather, it is stabilized by β-sheet-like contacts. Additionally, residues V74-V82 within the NAC region, which has been shown to be critical for incorporation into newly forming fibrils [113], remain somewhat exposed in the most favorable conformations of the ensemble, and with this region already incorporating important β-sheet contacts it seems primed for incorporation into newly-forming fibrils, which may help to explain the tendency for native synuclein to spontaneously form fibrils after mere agitation.

The structures of the major α-synuclein conformational clusters determined here and the presence of the secondary structure detected in the NAC region and C-terminal portions of the molecule allow one to hypothesize a possible mechanism for the misfolding conformational change and the early aggregation events. The existence of the transient hairpin/β-structure in the NAC and C-terminal subdomains predisposes the

protein to form the β-nucleation sites of the early oligomers, which may further mature to the cross-β-structure of the fibrils. The mis-folding conformational change that is necessary for the formation of inter-molecular contacts would then be the detachment of the NAC-containing hairpin (possibly in concert with stabilizing C-terminal beta-strands) from the core of the molecule. This would provide a stacking template for the similarly detached NAC hairpin of the other α-synuclein molecule. The residues 54-61 are also prone to forming an α-helix, which may provent their interaction with other protofibrils during fibril formation. This hypothesis can be tested by determining the structure of the early oligomers, which is currently underway in the Borchers laboratory.

The locations of the mutations within the proposed structure, which are known to influence the aggregation of α-synuclein is interesting (Figure 32) [205]. The familial mutations A30P, E46K, H50Q, G51D, and A53 T/E [206] were found to be located on the same surface of the connecting subdomain. Although the effects of these mutations could possibly be produced via different mechanisms (such as modified membrane binding or fibril stability), the effect of the oligomer-promoting mutations A30P and A76P, at least, can be explained in terms of the structure described here, where they would increase the segmental flexibility and possibly relax the interaction of the N-terminal subdomain with the core of the molecule. The A76P oligomer-promoting mutation similarly would facilitate segmental flexibility and the detachment of the NAC-containing β-hairpin postulated above. The S87E mutation, which blocks α-synuclein oligomerization and fibrilogenesis [207], would possibly disrupt NAC β-hairpin via electrostatic repulsion with the E83 residue.

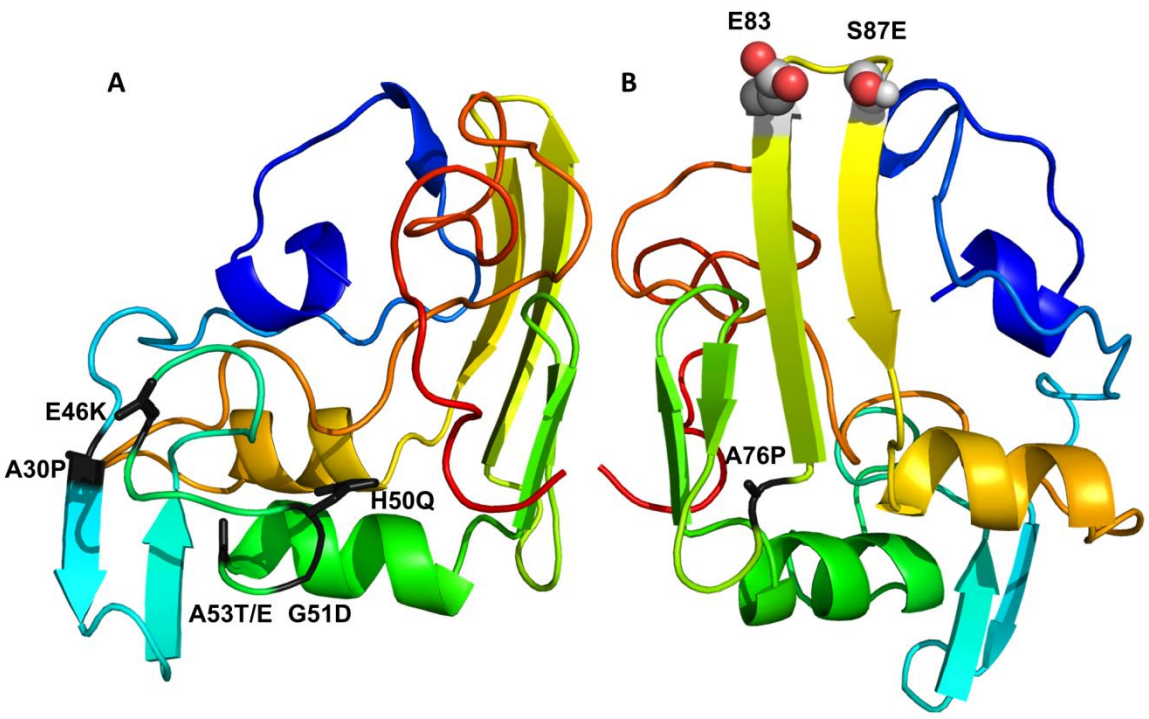


Figure 32: Location of the mutations effecting the aggregation of α-synuclein

A. Familial mutations located in the connecting subdomain are highlighted in black in the centroid structure of most populated cluster. B. Mutations potentially influencing the NAC β-hairpin.

4.4. Conclusions

In summary, my collaborators and I have determined *de novo* the conformational ensemble of native α-synuclein in solution by short-distance crosslinking constraint-guided DMD simulations, and validated this structure with experimental data from CD, HDX, SM, and LD-CL experiments. The predicted conformational ensemble is represented by rather compact globular conformations with transient secondary-structure

elements. The obtained structure can serve as a starting point for understanding the misfolding and oligomerization of α-synuclein.

Chapter 5: Conclusions and Future Directions

5.1. Summary of research objectives

The primary hypothesis of this dissertation was that the native α-synuclein protein adopts an ensemble of states in solution, the majority of which will provide some degree of protection of the NAC region from solvent. In order to address this hypothesis, I utilized structural proteomics techniques, some of which I developed or helped to develop in order to specifically address this question.

In Chapter 2: Development of new photoreactive crosslinkers for use in studying protein structures, I demonstrated the use of three new isotopically-labelled, hetero-bifunctional short-range crosslinking reagents for generating new crosslinking constraints. These reagents, in particular the phenyl-azide reagent ABAS, were able to crosslink a variety of protein targets, and several new amino acid targets were identified for the reagent. The primary target for ABAS, however, was lysine, which is less useful from some perspectives, but other targets such as histidine were also identified, indicating ABAS has some applicability as a non-specific reagent. Unfortunately, I was not able to generate very many crosslinks with the other two reagents I tested – the benzephenone crosslinker CBS and the diazirine crosslinker SDA. Of the crosslinks that I did observe, the preference for methionine by CBS was in accordance with the literature, as was the preference of ABAS for lysine. Some of these new crosslinkers would prove useful for further studies which are included in Chapters 3 and 4, as they provide a different set of constraints from other, more-standard, crosslinkers which exclusively use homo-

bifunctional NHS-ester linkages between lysine residues. I also demonstrated, to some extent, the applicability of the ABAS crosslinker to the synuclein protein.

In Chapter 3, I demonstrated a method of determining protein structures that was developed in collaboration with Drs. Nikolay Dokholyan and Konstantin Popov. This method utilizes crosslinking constraints in combination with discrete molecular dynamics simulations to model protein structures. I gathered a variety of constraints using short-distance crosslinking reagents including lysine-specific homobifuncitonal reagents such as DSA, and non-specific reagents such as SDA and TATA. These constraints were used by my collaborators in their DMD simulations of the protein structures. This method was able to predict the structures of myoglobin and the FK506 binding domain of another protein, FKBP-25, to within 5.4 Å and 2.7 Å, respectively.

I used a variety of structural proteomics methods, including HDX, LD-CL and surface modification in order to confirm some of the predicted features of the models. I used a top-down HDX approach to measure the secondary structure of the target proteins. This was combined with ECD fragmentation of the proteins in order to determine, on a residue-by-residue level, the exchange of individual hydrogens in the proteins. In my surface modification experiments, I unfolded the proteins using urea and then modified lysine, tyrosine, serine, and threonine residues with the PCAS reagent, and compared these to the untreated, native protein in order to determine the surface exposure of these residues.

We found this technique to be useful enough that it could be used to model other proteins which currently do not yet have well-defined structures in the literature.

In Chapter 4, I used the CL-DMD method to study the α-synuclein protein. I generated short distance crosslinking constraints, including several zero-length constraints using the EDC reagent. These were again given to my collaborators Drs. Nikolay Dokholyan and Konstantin Popov for modelling with discrete molecular dynamics. They produced an ensemble structure of the synuclein protein with four distinct clusters. These clusters represented roughly 37%, 28%, 20%, and 8% of the total lowest energy structures. These clusters had a number of salient features. In most of the structures, the NAC region is part of an extended hairpin, a stabilizing pair of beta sheets. This hairpin is somewhat exposed to solvent, although lysines 96 and 97 at the end of the NAC region were somewhat buried in later surface modification experiments, indicating that the region may have some protection.

There is also some secondary structure in most of the centroids. I examined this using top-down HDX, with both ECD and UVPD fragmentation to try to localize regions of protection. No region was found to be completely protected; instead, low levels of protection from exchange were observed across the protein, indicating a great degree of flexibility, confirmed by the relative differences in extent and positioning of secondary structure within the different clusters.

Surface modification using PCAS and comparing the native protein to protein that had been treated with 8 M urea was also performed, and several lysine residues were found to be somewhat protected from solvent exposure. These lysines also tended to be those which were crosslinked most heavily during earlier crosslinking experiments, indicating they were participating in intra-molecular interactions.

Dr. Popov also compared the synuclein structure to those generated using PRE-NMR, and found the synuclein structures generated using CL-DMD were somewhat more compact than those generated by PRE-NMR. Therefore, I think it can be concluded that the α-synuclein protein adopts an ensemble of structures in solution. These share a feature; the NAC region of the protein is at least somewhat stabilized by either secondary structure or other intra-molecular interactions within the protein itself, and that the frequent exposure of this region to solvent contributes to the ease with which synuclein can be converted into its pathogenic, oligomeric form.

5.2. Future Directions

The field of structural proteomics has a great deal of room to grow, as new techniques are constantly being added to the toolbox. A major area for structural proteomics to expand into is its combination with other techniques, including Cryo-EM and X-ray crystallography. Structural proteomics techniques, crosslinking in particular, represent a relatively simple way to obtain additional information which can be used for generating protein structures. In the Borchers laboratory, this has been previously demonstrated with respect to crystallography [19]. Additionally, new crosslinker chemistries are being developed and refined, which will lead to increasingly larger sets of constraints for modelling, resulting in improvement of techniques such as CL-DMD.

In addition to advancing the field of structural proteomics, there is also interesting work to be done on the α-synuclein protein as well. The next step will obviously be to study the oligomeric form of the protein, and compare how this protein differs from the native synuclein. This would need to be done under controlled conditions, using a defined oligomeric species. Since there are a great many ways to convert the synuclein protein to

an aggregated form, it would be useful to use one which converts the protein into a stable, toxic oligomer. For this purpose, one could use a reagent such as FN075 [208], which does just that. This oligomeric protein could then be crosslinked using the $^{14}N/^{15}N$ methodology, generating a collection of inter and intra-protein crosslinking constraints for use in CL-DMD simulations of the oligomer. Comparisons between these oligomer structures and the native protein could yield structural insights into which interactions are critically important for toxic oligomer formation. They may reveal potential drug targets in the oligomeric form. Surface modification could also be performed, using PCAS labelling to compare the two forms. HDX also has a role here, as changes in secondary structure upon the adoption of the more stable oligomeric form likely involves an increase in secondary structure content.

A potential clinical application may also be available for crosslinking technology. Synuclein and other disordered proteins such as tau are known to adopt different "strains" during their conversion into aggregates. Crosslinking may be able to differentiate among these variable strains, which may have implications for disease pathology and treatment. The goal of all of these studies is to assist in the development of new treatments for diseases that share these protein-misfolding characteristics, such as Parkinson's, Alzheimer's, Huntington's, and even potentially for prion disease, that have thus far proven recalcitrant to standard structural techniques. There is potential here to develop novel therapies, which may one day target all of these diseases, if the prion-like nature of each proves to be similar enough for them all to be amenable to a similar therapeutic approach.

Bibliography

1. Petrotchenko, E.V. and C.H. Borchers, *Modern mass spectrometry-based structural proteomics.* Adv Protein Chem Struct Biol, 2014. **95**: p. 193-213.
2. Shi, Y., et al., *A strategy for dissecting the architectures of native macromolecular assemblies.* Nat Methods, 2015. **12**(12): p. 1135-8.
3. Shi, Y., et al., *Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex.* Mol Cell Proteomics, 2014. **13**(11): p. 2927-43.
4. Liu, F., et al., *The interactome of intact mitochondria by cross-linking mass spectrometry provides evidence for coexisting respiratory supercomplexes.* Mol Cell Proteomics, 2018. **17**(2): p. 216-232.
5. Fenn, J.B., et al., *Electrospray ionization for mass spectrometry of large biomolecules.* Science, 1989. **246**(4926): p. 64.
6. Whitehouse, C.M., et al., *Electrospray interface for liquid chromatographs and mass spectrometers.* Anal Chem, 1985. **57**(3): p. 675-9.
7. Karas, M., et al., *Matrix-assisted ultraviolet laser desorption of non-volatile compounds.* Int J of Mass Spec and Ion Proc, 1987. **78**: p. 53-68.
8. Bahr, U., M. Karas, and F. Hillenkamp, *Analysis of biopolymers by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry.* Fresenius' J Anal Chem, 1994. **348**(12): p. 783-791.
9. Hunt, D.F., et al., *Protein sequencing by tandem mass spectrometry.* Proc Natl Acad Sci U S A, 1986. **83**(17): p. 6233-7.
10. Wells, J.M. and S.A. McLuckey, *Collision-induced dissociation (CID) of peptides and proteins.* Methods Enzymol, 2005. **402**: p. 148-85.
11. Zubarev, R.A., N.L. Kelleher, and F.W. McLafferty, *Electron capture dissociation of multiply charged protein cations. A nonergodic process.* J Am Chem Soc, 1998. **120**(13): p. 3265-3266.
12. Stensballe, A., et al., *Electron capture dissociation of singly and multiply phosphorylated peptides.* Rapid Commun Mass Spectrom, 2000. **14**(19): p. 1793-800.
13. Syka, J.E., et al., *Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry.* Proc Natl Acad Sci U S A, 2004. **101**(26): p. 9528-33.
14. Bowers, W.D., et al., *Fragmentation of oligopeptide ions using ultraviolet-laser radiation and fourier-transform mass-spectrometry.* J Am Chem Soc, 1984. **106**(23): p. 7288-7289.
15. Petrotchenko, E.V., et al., *Analysis of protein structure by cross-linking combined with mass spectrometry.* Methods Mol Biol, 2014. **1156**: p. 447-63.
16. O'Reilly, F.J. and J. Rappsilber, *Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology.* Nat struct mol biol, 2018. **25**(11): p. 1000-1008.

17. Boulikas, T., J.M. Wiseman, and W.T. Garrard, *Points of Contact between Histone H-1 and the Histone Octamer.* Proc Natl Acad of Sci U S A, 1980. **77**(1): p. 127-131.
18. Petrotchenko, E.V., J.J. Serpa, and C.H. Borchers, *An isotopically coded CID-cleavable biotinylated cross-linker for structural proteomics.* Mol Cell Proteomics, 2011. **10**(2): p. M110 001420.
19. Tonkin, M.L., et al., *Structural and biochemical characterization of Plasmodium falciparum 12 (Pf12) reveals a unique interdomain organization and the potential for an antiparallel arrangement with Pf41.* J Biol Chem, 2013. **288**(18): p. 12805-17.
20. Quan, S., et al., *Super Spy variants implicate flexibility in chaperone action.* Elife, 2014. **3**: p. e01584.
21. Belsom, A., et al., *Serum albumin domain structures in human blood serum by mass spectrometry and computational biology.* Mol Cell Proteomics, 2016. **15**(3): p. 1105-16.
22. Kahraman, A., et al., *Cross-link guided molecular modeling with ROSETTA.* PLoS One, 2013. **8**(9): p. e73411.
23. Leitner, A., et al., *Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics.* Mol Cell Proteomics, 2010. **9**(8): p. 1634-1649.
24. Morpurgo, M., E.A. Bayer, and M. Wilchek, *N-hydroxysuccinimide carbonates and carbamates are useful reactive reagents for coupling ligands to lysines on proteins.* J Biochem Biophys Methods, 1999. **38**(1): p. 17-28.
25. Gomes, A.F. and F.C. Gozzo, *Chemical cross-linking with a diazirine photoactivatable cross-linker investigated by MALDI- and ESI-MS/MS.* J Mass Spectrom, 2010. **45**(8): p. 892-9.
26. Chong, P.C. and R.S. Hodges, *A new heterobifunctional cross-linking reagent for the study of biological interactions between proteins. II. Application to the troponin C-troponin I interaction.* J Biol Chem, 1981. **256**(10): p. 5071-6.
27. Brunner, J., *New photolabeling and crosslinking methods.* Annu Rev Biochem, 1993. **62**: p. 483-514.
28. Tanaka, Y., M.R. Bond, and J.J. Kohler, *Photocrosslinkers illuminate interactions in living cells.* Mol Biosyst, 2008. **4**(6): p. 473-80.
29. Partis, M.D., et al., *Cross-linking of protein by ω-maleimido alkanoylN-hydroxysuccinimido esters.* J Prot Chem, 1983. **2**(3): p. 263-277.
30. Rosenfeld, J., et al., *In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis.* Anal Biochem, 1992. **203**(1): p. 173-179.
31. Sinz, A., *Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions.* Mass Spectrom Rev, 2006. **25**(4): p. 663-82.
32. Petrotchenko, E.V. and C.H. Borchers, *Crosslinking combined with mass spectrometry for structural proteomics.* Mass Spectrom Rev, 2010. **29**(6): p. 862-76.
33. Muller, D.R., et al., *Isotope-tagged cross-linking reagents. A new tool in mass spectrometric protein interaction analysis.* Anal Chem, 2001. **73**(9): p. 1927-34.

34. Liu, F., et al., *Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry.* Nat Methods, 2015. **12**(12): p. 1179-84.
35. Petrotchenko, E.V., V.K. Olkhovik, and C.H. Borchers, *Isotopically coded cleavable cross-linker for studying protein-protein interaction and protein complexes.* Mol Cell Proteomics, 2005. **4**(8): p. 1167-79.
36. Petrotchenko, E.V. and C.H. Borchers, *ICC-CLASS: isotopically-coded cleavable crosslinking analysis software suite.* BMC Bioinformatics, 2010. **11**: p. 64.
37. Fujii, N., et al., *A novel protein crosslinking reagent for the determination of moderate resolution protein structures by mass spectrometry (MS3-D).* Bioorg Med Chem Lett, 2004. **14**(2): p. 427-9.
38. Chen, Z.A., et al., *Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry.* EMBO J, 2010. **29**(4): p. 717-26.
39. Fritzsche, R., et al., *Optimizing the enrichment of cross-linked products for mass spectrometric protein analysis.* Rapid Commun Mass Spectrom, 2012. **26**(6): p. 653-8.
40. Tinnefeld, V., et al., *Enrichment of cross-linked peptides using charge-based fractional diagonal chromatography (ChaFRADIC).* J Proteome Res, 2017. **16**(2): p. 459-469.
41. Leitner, A., T. Walzthoeni, and R. Aebersold, *Lysine-specific chemical cross-linking of protein complexes and identification of cross-linking sites using LC-MS/MS and the xQuest/xProphet software pipeline.* Nat Protoc, 2014. **9**(1): p. 120-37.
42. Taverner, T., et al., *Characterization of an antagonist interleukin-6 dimer by stable isotope labeling, cross-linking, and mass spectrometry.* J Biol Chem, 2002. **277**(48): p. 46487-92.
43. Petrotchenko, E.V., et al., *(14)N(15)N DXMSMS Match program for the automated analysis of LC/ESI-MS/MS crosslinking data from experiments using (15)N metabolically labeled proteins.* J Proteomics, 2014. **109**: p. 104-10.
44. Groitl, B., et al., *Protein unfolding as a switch from self-recognition to high-affinity client binding.* Nat Commun, 2016. **7**: p. 10357.
45. Petrotchenko, E.V., K.A. Makepeace, and C.H. Borchers, *DXMSMS Match program for automated analysis of LC-MS/MS data obtained using isotopically coded CID-cleavable cross-linking reagents.* Curr Protoc Bioinformatics, 2014. **48**: p. 8 18 1-19.
46. Hoopmann, M.R., et al., *Kojak: Efficient analysis of chemically cross-linked protein complexes.* J Proteome Res, 2015. **14**(5): p. 2190-2198.
47. Gotze, M., et al., *StavroX--a software for analyzing crosslinked products in protein interaction studies.* J Am Soc Mass Spectrom, 2012. **23**(1): p. 76-87.
48. Liu, F., et al., *Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification.* Nat Commun, 2017. **8**: p. 15473.
49. Chen, Z.A. and J. Rappsilber, *Quantitative cross-linking/mass spectrometry to elucidate structural changes in proteins and their complexes.* Nat Protoc, 2019. **14**(1): p. 171-201.
50. Rinner, O., et al., *Identification of cross-linked peptides from large sequence databases.* Nat Methods, 2008. **5**(4): p. 315-8.

51. Kall, L., et al., *Semi-supervised learning for peptide identification from shotgun proteomics datasets.* Nat Methods, 2007. **4**(11): p. 923-5.
52. Smith, D.L., Y. Deng, and Z. Zhang, *Probing the non-covalent structure of proteins by amide hydrogen exchange and mass spectrometry.* J Mass Spectrom, 1997. **32**(2): p. 135-146.
53. Konermann, L., J.X. Pan, and Y.H. Liu, *Hydrogen exchange mass spectrometry for studying protein structure and dynamics.* Chem Soc Rev, 2011. **40**(3): p. 1224-1234.
54. Pan, J., et al., *Electron capture dissociation of electrosprayed protein ions for spatially resolved hydrogen exchange measurements.* J Am Chem Soc, 2008. **130**(35): p. 11574-5.
55. Chalmers, M.J., et al., *Probing protein ligand interactions by automated hydrogen/deuterium exchange mass spectrometry.* Anal Chem, 2006. **78**(4): p. 1005-14.
56. Pan, J., et al., *Hydrogen/deuterium exchange mass spectrometry with top-down electron capture dissociation for characterizing structural transitions of a 17 kDa protein.* J Am Chem Soc, 2009. **131**(35): p. 12801-8.
57. Pan, J.X., et al., *Subzero temperature chromatography and top-down mass spectrometry for protein higher-order structure characterization: method validation and application to therapeutic antibodies.* J Am Chem Soc, 2014. **136**(37): p. 13065-13071.
58. Brodie, N.I., et al., *Top-down hydrogen-deuterium exchange analysis of protein structures using ultraviolet photodissociation (UVPD).* Anal Chem, 2018. **90**(5): p. 3079-3082.
59. Jorgensen, T.J., et al., *Intramolecular migration of amide hydrogens in protonated peptides upon collisional activation.* J Am Chem Soc, 2005. **127**(8): p. 2785-93.
60. Boyd, R. and Á. Somogyi, *The mobile proton hypothesis in fragmentation of protonated peptides: A perspective.* J Am Soc Mass Spectrom, 2010. **21**(8): p. 1275-1278.
61. Serpa, J.J., et al., *Mass spectrometry-based structural proteomics.* Eur J Mass Spectrom (Chichester), 2012. **18**(2): p. 251-67.
62. Serpa, J.J., et al., *Using multiple structural proteomics approaches for the characterization of prion proteins.* J Proteomics, 2013. **81**: p. 31-42.
63. Hochleitner, E.O., et al., *Characterization of a discontinuous epitope of the human immunodeficiency virus (HIV) core protein p24 by epitope excision and differential chemical modification followed by mass spectrometric peptide mapping analysis.* Protein Sci, 2000. **9**(3): p. 487-496.
64. Serpa, J.J., et al., *Using isotopically-coded hydrogen peroxide as a surface modification reagent for the structural characterization of prion protein aggregates.* J Proteomics, 2014. **100**: p. 160-6.
65. Schmidt, A., J. Kellermann, and F. Lottspeich, *A novel strategy for quantitative proteomics using isotope-coded protein labels.* Proteomics, 2005. **5**(1): p. 4-15.
66. van Zundert, G.C.P., et al., *The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes.* J Mol Biol, 2016. **428**(4): p. 720-725.

67. Plaschka, C., et al., *Architecture of the RNA polymerase II-Mediator core initiation complex.* Nature, 2015. **518**(7539): p. 376-80.
68. Solomonson, M., et al., *Structure of EspB from the ESX-1 type VII secretion system and insights into its export mechanism.* Structure, 2015. **23**(3): p. 571-83.
69. Kao, A., et al., *Mapping the structural topology of the yeast 19S proteasomal regulatory particle using chemical cross-linking and probabilistic modeling.* Mol Cell Proteomics, 2012. **11**(12): p. 1566-77.
70. Belsom, A., et al., *Blind testing cross-linking/mass spectrometry under the auspices of the 11(th) critical assessment of methods of protein structure prediction (CASP11).* Wellcome Open Res, 2016. **1**: p. 24.
71. Shaw, D.E., et al., *Atomic-level characterization of the structural dynamics of proteins.* Science, 2010. **330**(6002): p. 341-6.
72. Pan, A.C., et al., *Demonstrating an order-of-magnitude sampling enhancement in molecular dynamics simulations of complex protein systems.* J Chem Theory Comput, 2016. **12**(3): p. 1360-7.
73. Ding, F., et al., *Ab initio folding of proteins with all-atom discrete molecular dynamics.* Structure, 2008. **16**(7): p. 1010-1018.
74. Shirvanyants, D., et al., *Discrete molecular dynamics: an efficient and versatile simulation method for fine protein characterization.* J Phys Chem B, 2012. **116**(29): p. 8375-82.
75. Havel, T.F., G.M. Crippen, and I.D. Kuntz, *Effects of distance constraints on macromolecular conformation. II. Simulation of experimental results and theoretical predictions.* Biopolymers, 1979. **18**(1): p. 73-81.
76. Madler, S., et al., *Chemical cross-linking with NHS esters: a systematic study on amino acid reactivities.* J Mass Spectrom, 2009. **44**(5): p. 694-706.
77. Bräse, S., et al., *Organic azides: an exploding diversity of a unique class of compounds.* Angew Chem Int Ed Engl, 2005. **44**(33): p. 5188-240.
78. Chong, P.C. and R.S. Hodges, *A new heterobifunctional cross-linking reagent for the study of biological interactions between proteins. I. Design, synthesis, and characterization.* J Biol Chem, 1981. **256**(10): p. 5064-70.
79. Kauer Jc Fau - Erickson-Viitanen, S., et al., *p-Benzoyl-L-phenylalanine, a new photoreactive amino acid. Photolabeling of calmodulin with a synthetic calmodulin-binding peptide.* (0021-9258 (Print)).
80. Wittelsberger, A., et al., *Methionine acts as a "magnet" in photoaffinity crosslinking experiments.* FEBS Lett, 2006. **580**(7): p. 1872-6.
81. Brase, S., et al., *Organic azides: an exploding diversity of a unique class of compounds.* Angew Chem Int Ed Engl, 2005. **44**(33): p. 5188-240.
82. Lashuel, H.A., et al., *The many faces of α-synuclein: from structure and toxicity to therapeutic target.* Nat Rev Neurosci, 2013. **14**(1): p. 38-48.
83. Bendor, J.T., T.P. Logan, and R.H. Edwards, *The Function of alpha-Synuclein.* Neuron, 2013. **79**(6): p. 1044-1066.
84. Kalia, L.V. and A.E. Lang, *Parkinson's disease.* Lancet. **386**(9996): p. 896-912.
85. Parkinson, J., *An Essay on the Shaking Palsy.* J Neuropsychiatry Clin Neurosci, 2002. **14**(2): p. 223-236.
86. Holdorff, B., *Friedrich Heinrich Lewy (1885–1950) and His Work.* J Hist Neurosci, 2002. **11**(1): p. 19-28.

87. Spillantini, M.G., et al., *[alpha]-Synuclein in Lewy bodies.* Nature, 1997. **388**(6645): p. 839-840.
88. Wong, S.L., H.L. Gilmour, and P.L. Ramage-Morin, *Parkinson's disease: Prevalence, diagnosis and impact*2014: Statistics Canada.
89. Charles, D., et al., *Subthalamic nucleus deep brain stimulation in early stage Parkinson's disease.* Parkinsonism & related disorders, 2014. **20**(7): p. 731-737.
90. Uéda, K., et al., *Molecular cloning of cDNA encoding an unrecognized component of amyloid in Alzheimer disease.* Proc Natl Acad of Sci U S A, 1993. **90**(23): p. 11282-11286.
91. Chen, X., et al., *The human NACP/α-synuclein gene: chromosome assignment to 4q21.3–q22 and TaqI RFLP analysis.* Genomics, 1995. **26**(2): p. 425-427.
92. Lavedan, C., *The synuclein family.* Genome Res, 1998. **8**(9): p. 871-80.
93. Uversky, V.N., et al., *Biophysical properties of the synucleins and their propensities to fibrillate: inhibition of alpha-synuclein assembly by beta- and gamma-synucleins.* J Biol Chem, 2002. **277**(14): p. 11970-8.
94. Iwai, A., et al., *The precursor protein of non-A beta component of Alzheimer's disease amyloid is a presynaptic protein of the central nervous system.* Neuron, 1995. **14**(2): p. 467-75.
95. Uversky, V.N., C.J. Oldfield, and A.K. Dunker, *Intrinsically disordered proteins in human diseases: Introducing the D2 concept.* Annu Rev of Biophys, 2008. **37**(1): p. 215-246.
96. Ulmer, T.S., et al., *Structure and dynamics of micelle-bound human α-synuclein.* J Biol Chem, 2005. **280**(10): p. 9595-9603.
97. Jao, C.C., et al., *Structure of membrane-bound alpha-synuclein studied by site-directed spin labeling.* Proc Natl Acad Sci U S A, 2004. **101**(22): p. 8331-6.
98. Kahle, P.J., et al., *Subcellular localization of wild-type and Parkinson's disease-associated mutant alpha -synuclein in human and transgenic mouse brain.* J Neurosci, 2000. **20**(17): p. 6365-73.
99. Burre, J., *The synaptic function of alpha-synuclein.* J Parkinsons Dis, 2015. **5**(4): p. 699-713.
100. Burre, J., et al., *Alpha-synuclein promotes SNARE-complex assembly in vivo and in vitro.* Science, 2010. **329**(5999): p. 1663-7.
101. Perez, R.G., et al., *A role for alpha-synuclein in the regulation of dopamine biosynthesis.* J Neurosci, 2002. **22**(8): p. 3090-9.
102. Yu, S., et al., *Inhibition of tyrosine hydroxylase expression in alpha-synuclein-transfected dopaminergic neuronal cells.* Neurosci Lett, 2004. **367**(1): p. 34-9.
103. Abeliovich, A., et al., *Mice lacking alpha-synuclein display functional deficits in the nigrostriatal dopamine system.* Neuron, 2000. **25**(1): p. 239-52.
104. Chandra, S., et al., *Double-knockout mice for alpha- and beta-synucleins: effect on synaptic functions.* Proc Natl Acad Sci U S A, 2004. **101**(41): p. 14966-71.
105. Iwai, A., et al., *Non-A beta component of Alzheimer's disease amyloid (NAC) is amyloidogenic.* Biochemistry, 1995. **34**(32): p. 10139-45.
106. Hashimoto, M., et al., *Human recombinant NACP/alpha-synuclein is aggregated and fibrillated in vitro: relevance for Lewy body disease.* Brain Res, 1998. **799**(2): p. 301-6.

107. Conway, K.A., et al., *Acceleration of oligomerization, not fibrillization, is a shared property of both α-synuclein mutations linked to early-onset Parkinson's disease: Implications for pathogenesis and therapy.* Proc Natl Acad Sci U S A, 2000. **97**(2): p. 571-576.
108. Conway, K.A., et al., *Accelerated oligomerization by parkinson's disease linked α-synuclein mutants.* Ann N Y Acad Sci, 2000. **920**(1): p. 42-45.
109. Winner, B., et al., *In vivo demonstration that α-synuclein oligomers are toxic.* Proc Natl Acad Sci U S A, 2011. **108**(10): p. 4194-4199.
110. Luk, K.C., et al., *Pathological alpha-synuclein transmission initiates Parkinson-like neurodegeneration in nontransgenic mice.* Science, 2012. **338**(6109): p. 949-53.
111. Prusiner, S.B., et al., *Evidence for α-synuclein prions causing multiple system atrophy in humans with parkinsonism.* Proc Natl Acad Sci U S A, 2015. **112**(38): p. E5308-E5317.
112. Tuttle, M.D., et al., *Solid-state NMR structure of a pathogenic fibril of full-length human α-synuclein.* Nat Struct Mol Biol, 2016. **23**(5): p. 409-415.
113. Guerrero-Ferreira, R., et al., *Cryo-EM structure of alpha-synuclein fibrils.* Elife, 2018. **7**.
114. Lashuel, H.A., et al., *Neurodegenerative disease: Amyloid pores from pathogenic mutations.* Nature, 2002. **418**(6895): p. 291-291.
115. Danzer, K.M., et al., *Different species of α-synuclein oligomers induce calcium influx and seeding.* J Neurosci, 2007. **27**(34): p. 9220.
116. Hoozemans, J.J., et al., *Activation of the unfolded protein response is an early event in Alzheimer's and Parkinson's disease.* Neurodegener Dis, 2012. **10**(1-4): p. 212-5.
117. Theillet, F.X., et al., *Structural disorder of monomeric alpha-synuclein persists in mammalian cells.* Nature, 2016. **530**(7588): p. 45-50.
118. Dedmon, M.M., et al., *Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations.* J Am Chem Soc, 2005. **127**(2): p. 476-477.
119. Nath, A., et al., *The Conformational Ensembles of alpha-Synuclein and Tau: Combining Single-Molecule FRET and Simulations.* Biophysical J, 2012. **103**(9): p. 1940-1949.
120. Mouradov, D., et al., *Protein structure determination using a combination of cross-linking, mass spectrometry, and molecular modeling.* Methods Mol Biol, 2008. **426**: p. 459-74.
121. Church, R.F.R. and M.J. Weiss, *Diazirines. II. Synthesis and properties of small functionalized diazirine molecules. Observations on the reaction of a diaziridine with the iodine-iodide ion system.* J Org Chem, 1970. **35**(8): p. 2465–2471.
122. Parker, J.M.R. and R.S. Hodges, *I. Photoaffinity probes provide a general method to prepare synthetic peptide-conjugates.* J Prot Chem, 1984. **3**(5-6): p. 465-478.
123. Müller, D.R., et al., *Isotope-tagged cross-linking reagents. A new tool in mass spectrometric protein interaction analysis.* Anal Chem, 2001. **73**: p. 1927-1934.
124. Weinreb, P.H., et al., *NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded.* Biochemistry, 1996. **35**(43): p. 13709-15.

125. Uversky, V.N., J. Li, and A.L. Fink, *Evidence for a partially folded intermediate in alpha-synuclein fibril formation.* J Biol Chem, 2001. **276**(14): p. 10737-44.
126. Biere, A.L., et al., *Parkinson's disease-associated alpha-synuclein is more fibrillogenic than beta- and gamma-synuclein and cannot cross-seed its homologs.* J Biol Chem, 2000. **275**(44): p. 34574-9.
127. Jiao, J., et al., *A facile and practical copper powder-catalyzed, organic solvent- and ligand-free Ullmann amination of aryl halides.* J Org Chem, 2011. **76**(4): p. 1180-3.
128. Barqawi, H. and W.H. Binder, *Azide/alkyne-"click"-reactions on amino resin materials: An LC-ESI-TOF analysis.* J Polym Sci A Polym Chem, 2010. **48**(21): p. 4855-4866.
129. Sechi, M., et al., *Design, synthesis, molecular modeling, and anti-HIV-1 integrase activity of a series of photoactivatable diketo acid-containing inhibitors as affinity probes.* Antivir Res, 2009. **81**(3): p. 267-76.
130. Wertheim, E., *Benzylbenzaldoxime.* J Am Chem Soc, 1933. **55**(6): p. 2540–2543.
131. Peng, L., et al., *Catalytic conversion of cellulose to levulinic acid by metal chlorides.* Molecules, 2010. **15**(8): p. 5258-72.
132. Wagner, G., et al., *Structure-reactivity relationships: Reactions of a 5-substituted aziadamantane in a resorcin[4]arene-based cavitand.* Org Lett, 2010. **12**(2): p. 332-335.
133. Petrotchenko, E.V. and C.H. Borchers, *ICC-CLASS: Isotopically-Coded Cleavable CrossLinking Analysis Suite.* BMC Bioinformatics, 2010. **11**(1): p. 64.
134. Quan, S., et al., *Super Spy variants implicate flexibility in chaperone action.*. Elife, 2014. **3**: p. e01584.
135. Quan, S., et al., *Genetic selection designed to stabilize proteins uncovers a chaperone called Spy.* Nat Struct Mol Biol, 2011. **18**(3): p. 262-9.
136. Kim, E.E., et al., *Refinement of the crystal structure of ribonuclease S. Comparison with and between the various ribonuclease A structures.* Biochemistry, 1992. **31**(49): p. 12304-14.
137. Young, M.M., et al., *High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry.* Proc Natl Acad Sci U S A, 2000. **97**(11): p. 5802-6.
138. Bitan, G., A. Lomakin, and D.B. Teplow, *Amyloid beta-protein oligomerization - Prenucleation interactions revealed by photo-induced cross-linking of unmodified proteins.* J Biol Chem, 2001. **276**(37): p. 35176-35184.
139. Brodie, N.I., et al., *Isotopically-coded short-range hetero-bifunctional photo-reactive crosslinkers for studying protein structure.* J Proteomics, 2015. **118**: p. 12-20.
140. Brodie, N.I., E.V. Petrotchenko, and C.H. Borchers, *The novel isotopically coded short-range photo-reactive crosslinker 2,4,6-triazido-1,3,5-triazine (TATA) for studying protein structures.* J Proteomics, 2016.
141. Baker, D. and A. Sali, *Protein structure prediction and structural genomics.* Science, 2001. **294**(5540): p. 93-6.
142. Zhang, Y., *Protein structure prediction: when is it useful?* Curr Opin Struct Biol, 2009. **19**(2): p. 145-55.

143. Dill, K.A. and J.L. MacCallum, *The Protein-Folding Problem, 50 Years On.* Science, 2012. **338**(6110): p. 1042-1046.
144. Webb, B. and A. Sali, *Comparative Protein Structure Modeling Using MODELLER.* Curr Protoc Bioinformatics, 2016. **54**(5.6.1-5.6.37).
145. Kelley, L.A. and M.J. Sternberg, *Protein structure prediction on the Web: a case study using the Phyre server.* Nat Protoc, 2009. **4**(3): p. 363-71.
146. Cole, C., J.D. Barber, and G.J. Barton, *The Jpred 3 secondary structure prediction server.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W197-201.
147. Zhang, Y. and J. Skolnick, *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic Acids Res, 2005. **33**(7): p. 2302-9.
148. Rohl, C.A., et al., *Protein structure prediction using Rosetta.* Methods Enzymol, 2004. **383**: p. 66-93.
149. MacCallum, J.L., A. Perez, and K.A. Dill, *Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference.* Proc Natl Acad Sci U S A, 2015. **112**(22): p. 6985-90.
150. Perez, A., et al., *Blind protein structure prediction using accelerated free-energy simulations.* Sci Adv, 2016. **2**(11): p. e1601274.
151. Lindorff-Larsen, K., et al., *Picosecond to Millisecond Structural Dynamics in Human Ubiquitin.* J Phys Chem B, 2016. **120**(33): p. 8313-8320.
152. Palazzesi, F., et al., *Accuracy of current all-atom force-fields in modeling protein disordered states.* J Chem Theory Comput, 2015. **11**(1): p. 2-7.
153. Rauscher, S., et al., *Structural ensembles of intrinsically disordered proteins depend strongly on force field: A comparison to experiment.* J Chem Theory Comput, 2015. **11**(11): p. 5513-5524.
154. Henriques, J., C. Cragnell, and M. Skepo, *Molecular dynamics simulations of intrinsically disordered proteins: Force field evaluation and comparison with experiment.* J Chem Theory Comput, 2015. **11**(7): p. 3420-3431.
155. Ravera, E., et al., *A critical assessment of methods to recover information from averaged data.* Phys Chem Chem Phys, 2016. **18**(8): p. 5686-5701.
156. Bonomi, M., et al., *Principles of protein structural ensemble determination.* Curr Opin Struct Biol, 2017. **42**: p. 106-116.
157. Chen, Y., S.L. Campbell, and N.V. Dokholyan, *Deciphering protein dynamics from NMR data using explicit structure sampling and selection.* Biophysical Journal, 2007. **93**(7): p. 2300-2306.
158. Boomsma, W., J. Ferkinghoff-Borg, and K. Lindorff-Larsen, *Combining experiments and simulations using the maximum entropy principle.* Plos Comput Biol, 2014. **10**(2).
159. Dokholyan, N.V., et al., *Discrete molecular dynamics studies of the folding of a protein-like model.* Fold Des, 1998. **3**(6): p. 577-587.
160. Gudavicius, G., et al., *The prolyl isomerase, FKBP25, interacts with RNA-engaged nucleolin and the pre-60S ribosomal subunit.* RNA, 2014. **20**(7): p. 1014-22.
161. Ding, F. and N.V. Dokholyan, *Discrete Molecular Dynamics Simulation of Biomolecules*, in *Computational Modeling of Biological Systems: From Molecules to Pathways*, N.V.E. Dokholyan, Editor 2012, Springer-Verlag: NewYork. p. 55–73.

162. Proctor, E.A. and N.V. Dokholyan, *Applications of discrete molecular dynamics in biology and medicine.* Curr Opin Struct Biol, 2016. **37**: p. 9-13.
163. Lazaridis, T. and M. Karplus, *Effective energy functions for protein structure prediction.* Curr Opin Struct Biol, 2000. **10**(2): p. 139-45.
164. Andersen, H.C., *Molecular-dynamics simulations at constant pressure and-or temperature.* J Chem Phys, 1980. **72**(4): p. 2384-2393.
165. Okamoto, Y., *Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations.* J Mol Graph Model, 2004. **22**(5): p. 425-439.
166. Zhou, R.H., B.J. Berne, and R. Germain, *The free energy landscape for beta hairpin folding in explicit water.* Proc Natl Acad of Sci U S A, 2001. **98**(26): p. 14931-14936.
167. Chodera, J.D., et al., *Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations.* J Chem Theory Comput, 2007. **3**(1): p. 26-41.
168. Pronk, S., et al., *GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit.* Bioinformatics, 2013. **29**(7): p. 845-854.
169. Micsonai, A., et al., *Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy.* Proc Natl Acad of Sci U S A, 2015. **112**(24): p. E3095-103.
170. Petrotchenko, E.V. and C.H. Borchers, *HDX match software for the data analysis of top-down ECD-FTMS hydrogen/deuterium exchange experiments.* J Am Soc Mass Spectrom, 2015. **26**(11): p. 1895-8.
171. Petrotchenko, E.V., et al., *Use of proteinase K nonspecific digestion for selective and comprehensive identification of interpeptide cross-links: application to prion proteins.* Mol Cell Proteomics, 2012. **11**(7): p. M111 013524.
172. Chen, Y., F. Ding, and N.V. Dokholyan, *Fidelity of the protein structure reconstruction from inter-residue proximity constraints.* J Phys Chem B, 2007. **111**(25): p. 7432-8.
173. Yin, S., et al., *MedusaScore: An accurate force field-based scoring function for virtual drug screening.* J Chem Inf Model, 2008. **48**(8): p. 1656-1662.
174. Best, R.B. and M. Vendruscolo, *Determination of protein structures consistent with NMR order parameters.* J Am Chem Soc, 2004. **126**(26): p. 8090-8091.
175. Proctor, E.A., et al., *Nonnative SOD1 trimer is toxic to motor neurons in a model of amyotrophic lateral sclerosis.* Proc Natl Acad Sci U S A, 2016. **113**(3): p. 614-619.
176. Cremades, N., S.W. Chen, and C.M. Dobson, *Structural Characteristics of alpha-Synuclein Oligomers.* Int Rev Cell Mol Biol, 2017. **329**: p. 79-143.
177. Lautenschläger, J., et al., *C-terminal calcium binding of α-synuclein modulates synaptic vesicle interaction.* Nat Commun, 2018. **9**(1): p. 712.
178. Fusco, G., et al., *Structural basis of membrane disruption and cellular toxicity by α-synuclein oligomers.* Science, 2017. **358**(6369): p. 1440-1443.
179. Brundin, P., J.Y. Ma, and J.H. Kordower, *How strong is the evidence that Parkinson's disease is a prion disorder?* Curr Opin Neurol, 2016. **29**(4): p. 459-466.

180. Eliezer, D., *Biophysical characterization of intrinsically disordered proteins.* Curr Opin Struct Biol, 2009. **19**(1): p. 23-30.
181. Li, J., V.N. Uversky, and A.L. Fink, *Conformational behavior of human alpha-synuclein is modulated by familial Parkinson's disease point mutations A30P and A53T.* Neurotoxicology, 2002. **23**(4-5): p. 553-567.
182. Brodie, N.I., et al., *Solving protein structures using short-distance cross-linking constraints as a guide for discrete molecular dynamics simulations.* Sci Adv, 2017. **3**(7): p. e1700479.
183. Piana, S., et al., *Water dispersion interactions strongly influence simulated structural properties of disordered protein states.* J Phys Chem B, 2015. **119**(16): p. 5113-23.
184. Robustelli, P., S. Piana, and D.E. Shaw, *Developing a molecular dynamics force field for both folded and disordered protein states.* Proc Natl Acad of Sci U S A, 2018. **115**(21): p. E4758-E4766.
185. Ding, F. and N.V. Dokholyan, *Emergence of protein fold families through rational design.* Plos Comput Biol, 2006. **2**: p. e85
186. Proctor, E.A., F. Ding, and N.V. Dokholyan, *Discrete molecular dynamics.* Wiley Interdiscip Rev Comput Mol Sci, 2011. **1**(1): p. 80-92.
187. Dixon, R.D., et al., *New insights into FAK signaling and localization based on detection of a FAT domain folding intermediate.* Structure, 2004. **12**(12): p. 2161-71.
188. Ding, F., et al., *Three-dimensional RNA structure refinement by hydroxyl radical probing.* Nat Methods, 2012. **9**(6): p. 603-8.
189. Szöllősi, D., et al., *Discrete molecular dynamics can predict helical prestructured motifs in disordered proteins.* PLoS One, 2014. **9**(4): p. e95795.
190. Petrotchenko, E.V., et al., *$^{14}N^{15}N$ DXMSMS Match program for the automated analysis of LC/ESI-MS/MS crosslinking data from experiments using ^{15}N metabolically labeled proteins.* J Proteomics, 2014. **109**: p. 104-110.
191. Daura, X., et al., *Peptide folding: When simulation meets experiment.* Angew Chem, 1999 **38**(1-2): p. 236-240.
192. Pauwels, K., P. Lebrun, and P. Tompa, *To be disordered or not to be disordered: is that still a question for proteins in the cell?* Cell Mol Life Sci, 2017.
193. Rossetti, G., et al., *Conformational ensemble of human α-synuclein physiological form predicted by molecular simulations.* Phys Chem Chem Phys, 2016. **18**(8): p. 5702-6.
194. Bartels, T., et al., *N-alpha-acetylation of α-synuclein increases its helical folding propensity, GM1 binding specificity and resistance to aggregation.* PLoS One, 2014. **9**(7): p. e103727.
195. Ruzafa, D., et al., *The influence of N-terminal acetylation on micelle-induced conformational changes and aggregation of α-Synuclein.* PLoS One, 2017. **12**(5): p. e0178576.
196. Maltsev, A.S., J. Ying, and A. Bax, *Impact of N-terminal acetylation of α-synuclein on its random coil and lipid binding properties.* Biochemistry, 2012. **51**(25): p. 5004-13.

197. Bertoncini, C.W., et al., *Release of long-range tertiary interactions potentiates aggregation of natively unstructured alpha-synuclein.* Proc Natl Acad of Sci U S A, 2005. **102**(5): p. 1430-1435.
198. Allison, J.R., et al., *A relationship between the transient structure in the monomeric state and the aggregation propensities of α-synuclein and β-synuclein.* Biochemistry, 2014. **53**(46): p. 7170-83.
199. Grupi, A. and E. Haas, *Segmental conformational disorder and dynamics in the intrinsically disordered protein α-synuclein and its chain length dependence.* J Mol Biol, 2011. **405**(5): p. 1267-83.
200. Trexler, A.J. and E. Rhoades, *Single molecule characterization of α-synuclein in aggregation-prone states.* Biophys J, 2010. **99**(9): p. 3048-55.
201. Schwalbe, M., et al., *Predictive atomic resolution descriptions of intrinsically disordered hTau40 and α-synuclein in solution from NMR and small angle scattering.* Structure, 2014. **22**(2): p. 238-49.
202. Gurry, T., et al., *The dynamic structure of α-synuclein multimers.* J Am Chem Soc, 2013. **135**(10): p. 3865-72.
203. Giasson, B.I., et al., *A hydrophobic stretch of 12 amino acid residues in the middle of alpha-synuclein is essential for filament assembly.* J Biol Chem, 2001. **276**(4): p. 2380-2386.
204. Nasica-Labouze, J., et al., *Amyloid β protein and Alzheimer's disease: When computer simulations complement experimental studies.* Chem Rev, 2015. **115**(9): p. 3518-63.
205. Breydo, L., J.W. Wu, and V.N. Uversky, *A-synuclein misfolding and Parkinson's disease.* Biochim Biophys Acta, 2012. **1822**(2): p. 261-85.
206. Sahay, S., et al., *Alteration of structure and aggregation of α-synuclein by familial Parkinson's disease associated mutations.* Curr Protein Pept Sci, 2017. **18**(7): p. 656-676.
207. Oueslati, A., et al., *Mimicking phosphorylation at serine 87 inhibits the aggregation of human α-synuclein and protects against its toxicity in a rat model of Parkinson's disease.* J Neurosci, 2012. **32**(5): p. 1536-44.
208. Horvath, I., et al., *Mechanisms of protein oligomerization: Inhibitor of functional amyloids templates α-synuclein fibrillation.* J Am Chem Soc, 2012. **134**(7): p. 3439-3444.
209. Kumar, S., et al., *THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method.* J Comput Chem, 1992. **13**: p. 1011–1021.
210. Allison, J.R., et al., *Determination of the free energy landscape of alpha-synuclein using spin label nuclear magnetic resonance measurements.* J Am Chem Soc, 2009. **131**(51): p. 18314-18326.

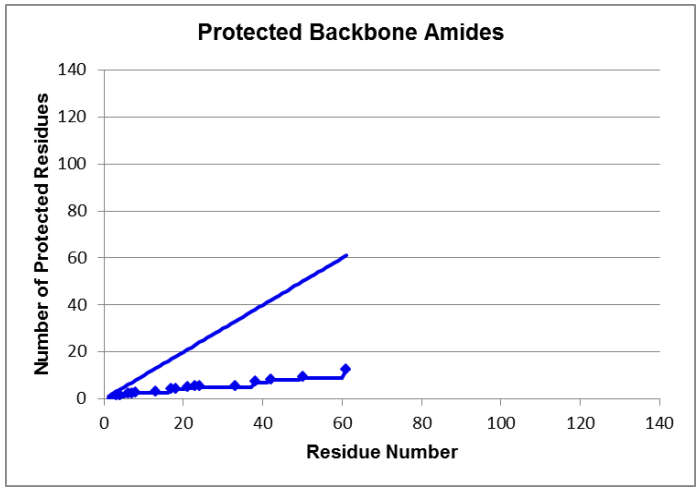
Appendix A: Crosslinks used for CL-DMD of α-synuclein

Crosslinker	Constraint (Å)	Mass (M+H)	ppm error	Residues 1	Sequence 1	Residues 2	Sequence 2	Crosslink
EDC	0	819.4291	-0.6	5-6	(F)MK(G)	110-114	(Q)EGILE(D)	K6-E110
EDC	0	862.5024	1.6	110-113	(Q)EGIL(E)	93-96	(T)GFVK(K)	K96-E110
EDC	0	901.5341	1.9	110-113	(Q)EGIL(E)	95-98	(F)VKKD(Q)	K96-E110
EDC	0	963.5496	2	110-113	(Q)EGIL(E)	92-96	(A)TGFVK(K)	K96-E110
EDC	0	980.4061	0.8	5-6	(F)MK(G)	126-131	(Y)EMPSEE(G)	K6-E126
EDC	0	990.5969	1.9	110-113	(Q)EGIL(E)	93-97	(T)GFVKK(D)	K96-E110
EDC	0	991.5458	0.6	93-96	(T)GFVK(K)	110-114	(Q)EGILE(D)	K96-E110
EDC	0	1004.5613	1.5	42-45	(G)SKTK(E)	110-114	(Q)EGILE(D)	K43-E114
EDC	0	1030.5772	1.2	95-98	(F)VKKD(Q)	110-114	(Q)EGILE(D)	K96-E110
EDC	0	1044.5591	-1.3	97-100	(K)KDQL(G)	110-114	(Q)EGILE(D)	K96-E110
EDC	0	1078.5423	-0.2	95-98	(F)VKKD(Q)	136-140	(D)YEPEA(-)	K96-E137
EDC	0	1106.5031	0.8	7-10	(K)GLSK(A)	126-131	(Y)EMPSEE(G)	K10-E126
EDC	0	1111.545	0.7	1-4	(-)MDVF(M)	9-14	(L)SKAKEG(V)	N-term-E13
EDC	0	1165.5933	-0.5	1-4	(-)MDVF(M)	11-17	(K)AKEGVVA(A)	N-term-E13
EDC	0	1178.5242	0.8	80-83	(Q)KTVE(G)	126-131	(Y)EMPSEE(G)	K80-E126
EDC	0	1190.5695	-0.1	93-96	(T)GFVK(K)	104-110	(N)EEGAPQE(G)	K96-E110
EDC	0	1213.4738	1.6	1-4	(-)MDVF(M)	126-131	(Y)EMPSEE(G)	N-term-E126
EDC	0	1234.6668	1.2	110-114	(Q)EGILE(D)	93-98	(T)GFVKKD(Q)	K96-E110
EDC	0	1291.6174	-0.3	92-96	(A)TGFVK(K)	104-110	(N)EEGAPQE(G)	K96-E110
EDC	0	1306.582	1.3	80-85	(Q)KTVEGA(G)	126-131	(Y)EMPSEE(G)	K80-E126
EDC	0	1309.7474	0.4	110-113	(Q)EGIL(E)	43-50	(S)KTKEGVVH(G)	K43-E110
EDC	0	1361.6249	0.7	42-45	(G)SKTK(E)	114-121	(L)EDMPVDPD(N)	K43-E114
EDC	0	1409.5626	-1.5	1-4	(-)MDVF(M)	114-121	(L)EDMPVDPD(N)	N-term-E114

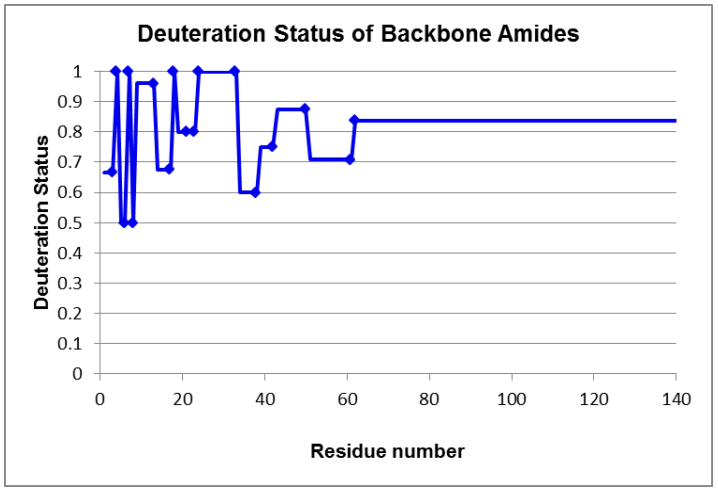
TATA	5	697.32574	1	82-85	(T)VEGA(G)	117-118	(M)PV(D)	V82-P117
TATA	5	724.33746	-0.2	26-28	(G)VAE(A)	107-109	(G)APQ(E)	V26-A107
TATA	5	834.40641	1.1	5-8	(F)MKGL(S)	16-17	(V)VA(A)	M5-V16
TATA	5	1080.51757	0.5	65-72	(T)NVGGAVVT(G)	138-139	(E)PE(A)	N65-P138
TATA	5	1102.54943	0.6	45-50	(T)KEGVVH(G)	107-109	(G)APQ(E)	K45-A107
TATA	5	1199.63008	1.1	5-11	(F)MKGLSKA(K)	24-27	(K)QGVA(E)	M5-G25
TATA	5	1260.66819	1.1	31-36	(A)GKTKEG(V)	37-41	(G)VLYVG(S)	G31-V37
TATA	5	1472.60277	0.6	117-125	(M)PVDPDNEAY(E)	128-130	(M)PSE(E)	P117-P128
TATA	5	1547.58079	-2	114-121	(L)EDMPVDPD(N)	125-128	(A)YEMP(S)	D121-Y125
TATA	5	1595.57311	0.7	119-125	(V)DPDNEAY(E)	130-134	(S)EEGYQ(D)	D121-E130
ABAS	7	1001.51165	1.3	95-97	(F)VKK(D)	1-4	(-)MDVF(M)	K97-N-term
ABAS	7	1058.49889	-0.8	1-4	(-)MDVF(M)	23-26	(T)KQGV(A)	N-term-K23
ABAS	7	861.380602	0.6	9-10	(L)SK(A)	1-4	(-)MDVF(M)	N-term-K10
ABAS	7	932.417833	0.4	9-11	(L)SKA(K)	1-4	(-)MDVF(M)	N-term-K10
ABAS	7	1060.513597	-0.4	9-12	(L)SKAK(E)	1-4	(-)MDVF(M)	N-term-K10
ABAS	7	1075.476732	-0.2	19-22	(A)AEKT(K)	1-4	(-)MDVF(M)	N-term-K21
ABAS	7	1016.5508	1.5	47-50	(E)GVVH(G)	95-98	(F)VKKD(Q)	H50-K96
SDA	5	1450.776371	1.7	25-28	(Q)GVAE(A)	31-39	(A)GKTKEGVLY(V)	E28-K32
SDA	5	1722.901401	0.4	105-109	(E)EGAPQ(E)	40-50	(Y)VGSKTKEGVVH(G)	K45-E105
SDA	5	2621.417515	1.7	35-43	(K)EGVLYVGSK(T)	5-20	(F)MKGLSKAKEGVVAAAE(K)	K12-K43
SDA	5	2639.429416	-0.1	85-96	(G)AGSIAAATGFVK(K)	35-48	(K)EGVLYVGSKTKEGV(V)	S42-K96

Appendix B: ECD and UVPD fragmentation results for exchanged synuclein

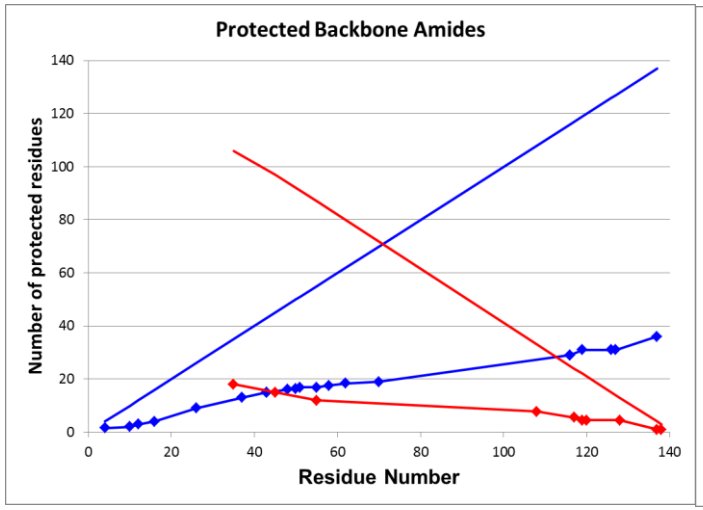
A



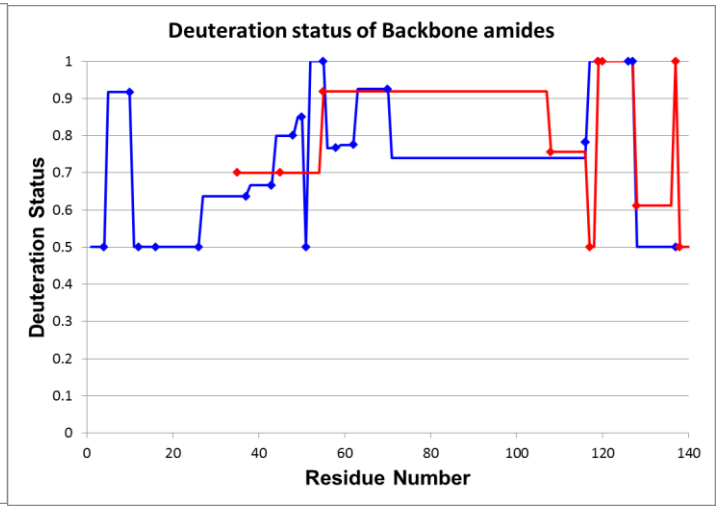
B



C



D



A. The number of protected hydrogen atoms in c-ions detected by ECD-FTICR mass spectrometry. The upper blue line represents the theoretical situation of perfect protection of all backbone hydrogen atoms. The lower line represents the number of actual protected hydrogen atoms detected at that fragment. B. The deuteration status of backbone amides from ECD data. A value of 1 indicates that the residue is unprotected; a value of zero represents complete protection from exchange with deuterium. C. The number of protected hydrogen atoms in fragment ions detected by UVPD-FT mass spectrometry. D. The deuteration status of backbone amides from UVPD data. N-terminal and C-terminal fragment data are shown in blue and red, respectively.

Appendix C: Surface modification results for α-synuclein

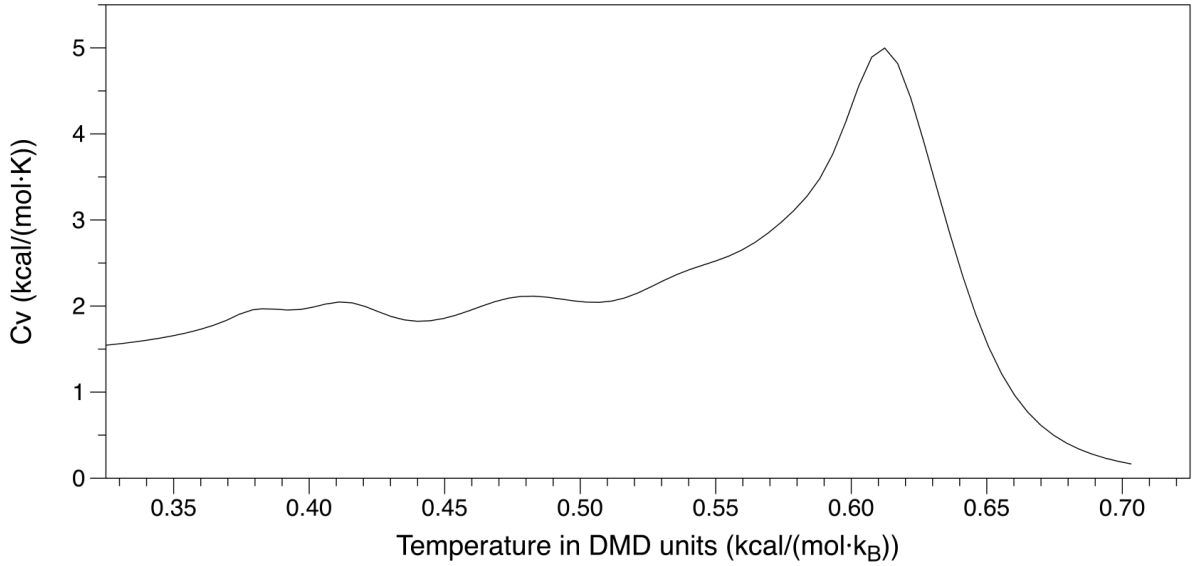
Peptide	Mass	ppm error	H/L ratio	Lysine
A.AAEKTK(+105.02)QGVAE.A	1235.615	0.6	2.599	23
T.K(+105.02)EGVVHGVAT.V	1100.562	2.1	1.025	45
T.AVAQK(+105.02)TVE.G	949.4869	0.2	0.812	80
F.VK(+105.02)KDQL.G	834.46	0.3	1.852	96
F.VKK(+105.02)DQL.G	834.46	0.5	1.932	97
L.GK(+105.02)NEEGAPQEGIL.E	1445.679	-0.2	0.903	102

Appendix D: Long-distance crosslinking results for α-synuclein

Mass (M+H)	ppm error	Residues 1	Sequence 1	Residues 2	Sequence 2	Constraint
1175.51140	0.4	9-10	(L)SK(A)	31-34	(A)GKTK(E)	K10-K32
1220.50327	-1.9	93-96	(T)GFVK(K)	97-98	(K)KD(Q)	K96-K97
1237.46263	-0.4	10-11	(S)KA(K)	1-4	(-)MDVF(M)	N-K10
1265.49303	0.3	95-96	(F)VK(K)	1-4	(-)MDVF(M)	N-K96
1297.46311	1.8	5-6	(F)MK(G)	1-4	(-)MDVF(M)	N-K6
1337.56407	-1.4	5-6	(F)MK(G)	92-96	(A)TGFVK(K)	K6-K96
1390.57371	-0.2	5-6	(F)MK(G)	80-85	(Q)KTVEGA(G)	K6-K96
1437.57871	-0.3	8-11	(G)LSKA(K)	1-4	(-)MDVF(M)	K6-K80
1452.58965	-0.4	9-12	(L)SKAK(E)	1-4	(-)MDVF(M)	N-K10
1454.61598	0	5-6	(F)MK(G)	45-50	(T)KEGVVH(G)	K6-K45
1485.74860	0.1	95-97	(F)VKK(D)	7-12	(K)GLSKAK(E)	K12-K96
1495.66149	-0.8	11-12	(K)AK(E)	44-50	(K)TKEGVVH(G)	K12-K45
1511.65315	1.4	42-44	(G)SKT(K)	45-50	(T)KEGVVH(G)	K43-K12
1516.65551	-1.8	5-7	(F)MKG(L)	11-17	(K)AKEGVVA(A)	K6-K12
1582.69335	-0.6	31-33	(A)GKT(K)	44-50	(K)TKEGVVH(G)	K32-K45
1603.70299	-0.3	54-59	(A)TVAEKT(K)	78-81	(V)AQKT(V)	K58-K80
1622.69469	0	7-12	(K)GLSKAK(E)	1-4	(-)MDVF(M)	N-K10
1622.69548	-0.5	1-4	(-)MDVF(M)	7-12	(K)GLSKAK(E)	N-K10
1638.65325	-0.1	1-4	(-)MDVF(M)	9-14	(L)SKAKEG(V)	N-K10
1674.75820	0.1	5-6	(F)MK(G)	9-17	(L)SKAKEGVVA(A)	K6-K10
1692.70000	0.1	1-4	(-)MDVF(M)	11-17	(K)AKEGVVA(A)	N-K12
1707.71177	-0.4	97-102	(K)KDQLGK(N)	1-4	(-)MDVF(M)	N-K97
1770.78916	0.8	5-6	(F)MK(G)	42-50	(G)SKTKEGVVH(G)	K6-K43 or 45

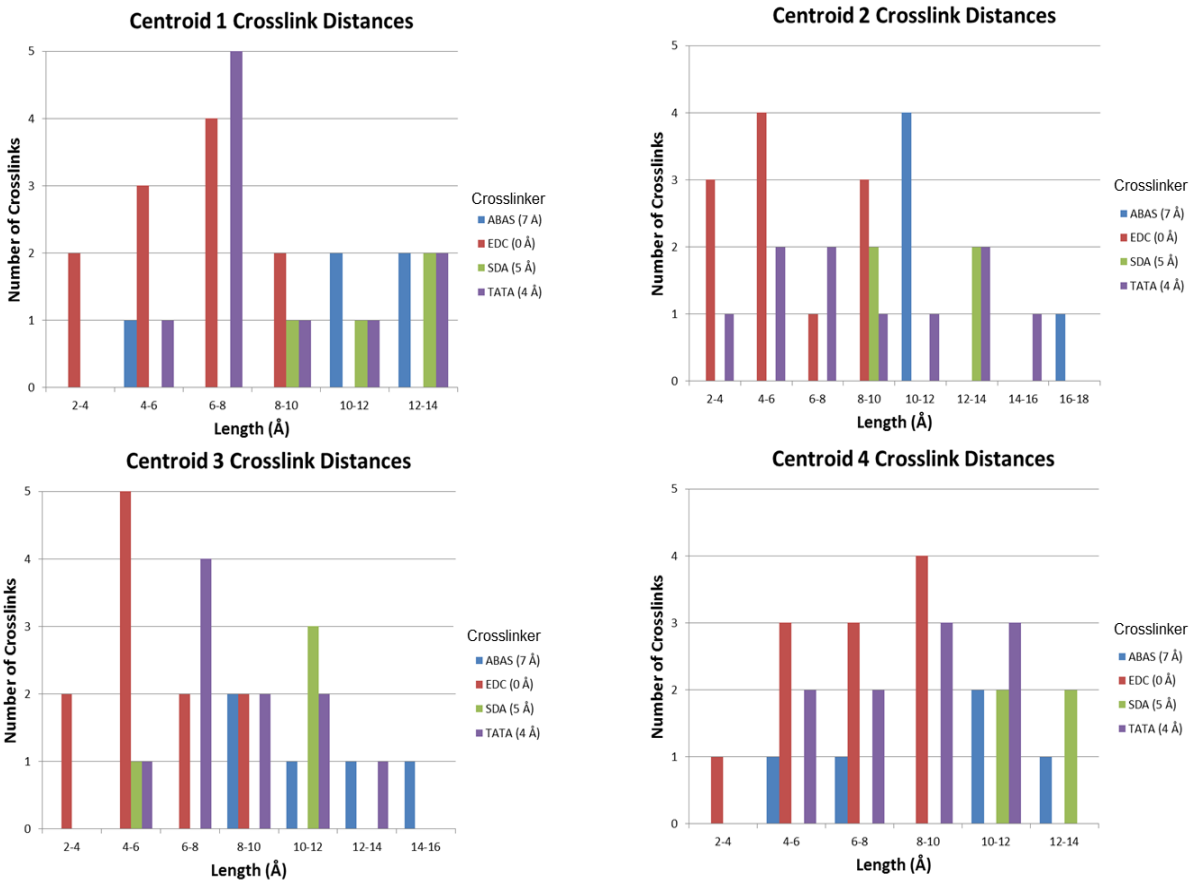
1788.76683	-1.1	5-6	(F)MK(G)	60-69	(T)KEQVTNVGGA(V)	K6-K60
1796.78855	-0.5	18-22	(A)AAEKT(K)	44-50	(K)TKEGVVH(G)	K21-K45
1805.81431	-0.6	80-85	(Q)KTVEGA(G)	93-98	(T)GFVKKD(Q)	K80-K96 or 97
1807.80556	-1	23-28	(T)KQGVAE(A)	45-50	(T)KEGVVH(G)	K23-K45
1822.84191	-1.2	34-38	(T)KEGVL(Y)	44-50	(K)TKEGVVH(G)	K34-K45
1824.82063	-0.9	44-50	(K)TKEGVVH(G)	55-59	(T)VAEKT(K)	K45-K58
1984.93979	0.2	6-9	(M)KGLS(K)	11-21	(K)AKEGVVAAAEK(T)	K6-K12
2000.94476	0.8	9-12	(L)SKAK(E)	22-32	(K)TKQGVAEAAGK(T)	K10-K32
2016.91384	-0.8	5-8	(F)MKGL(S)	19-28	(A)AEKTKQGVAE(A)	K6-K21
2028.94701	0.8	5-8	(F)MKGL(S)	11-21	(K)AKEGVVAAAEK(T)	K6-K10
2091.91261	-0.4	1-4	(-)MDVF(M)	11-21	(K)AKEGVVAAAEK(T)	N-K21
2113.03324	0.9	6-10	(M)KGLSK(A)	11-21	(K)AKEGVVAAAEK(T)	K6-K12
2121.92064	0.8	1-4	(-)MDVF(M)	13-23	(K)EGVVAAAEKTK(Q)	N-K12
2121.92357	-0.5	13-23	(K)EGVVAAAEKTK(Q)	1-4	(-)MDVF(M)	N-K21
2274.08695	-0.4	5-10	(F)MKGLSK(A)	13-23	(K)EGVVAAAEKTK(Q)	K6-K21
2351.05039	-1.3	1-6	(-)MDVFMK(G)	11-21	(K)AKEGVVAAAEK(T)	N-K12
2580.20341	-1.2	95-96	(F)VK(K)	97-113	(K)KDQLGKNEEGAPQEGIL(E)	K96-K97 or 102
2803.37143	-0.9	80-96	(Q)KTVEGAGSIAAATGFVK(K)	97-102	(K)KDQLGK(N)	K80-K96

Appendix E: Heat capacity curve of native α-synuclein



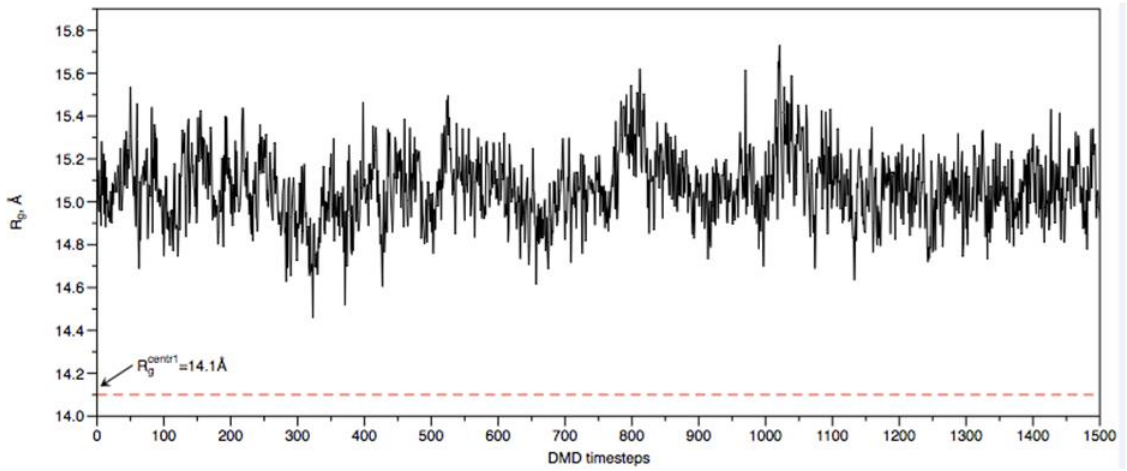
The heat capacity (Cv) curve was computed using WHAM [209] on the REX/DMD trajectories for α-synuclein in the range of 0.375 to 0.605 kcal/(mol k_B) DMD temperature units. The large peak corresponds to the unfolding temperature of the protein. States corresponding to all representative structures are located on the wide shoulder of the curve on the left. The absence of major peaks in this area indicates that the protein can coexist in multiple compact states.

Appendix F: Comparison of crosslinking constraints satisfied by each cluster



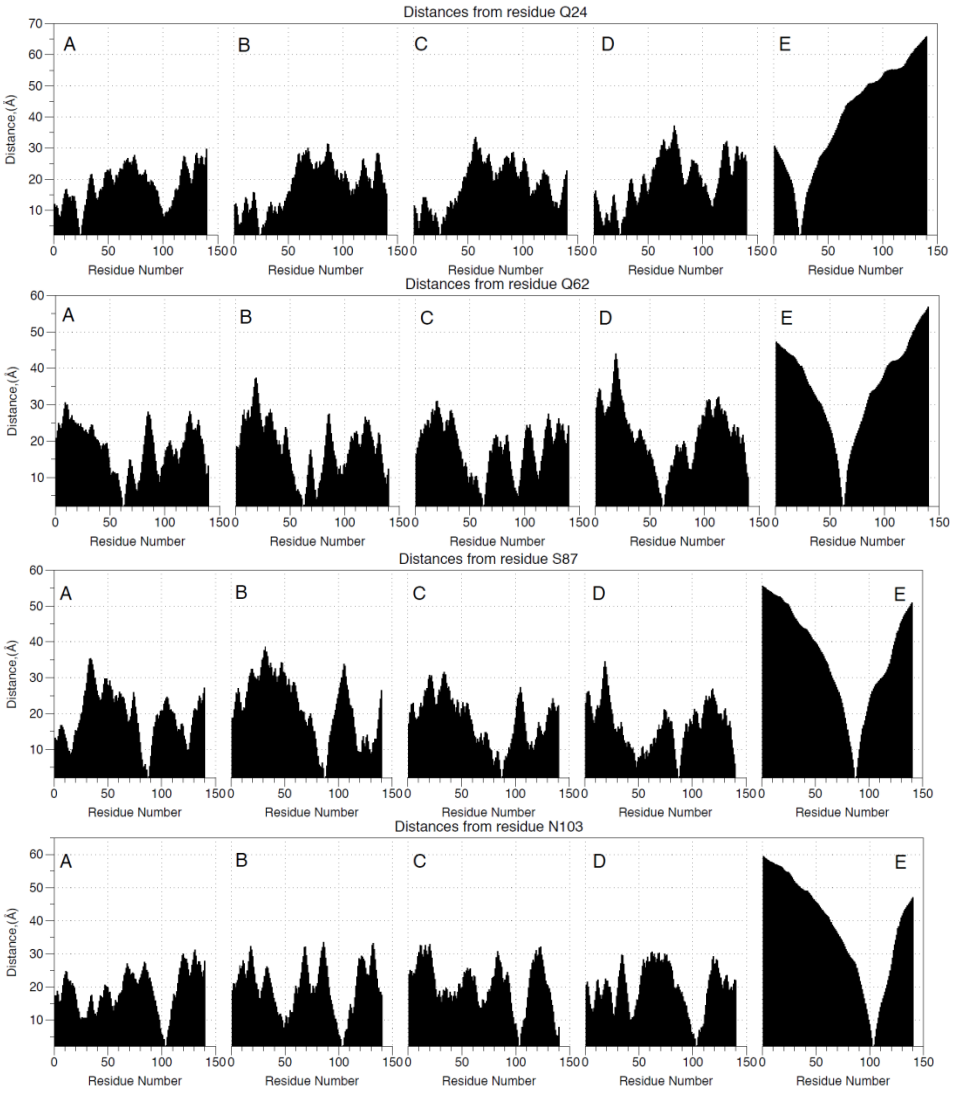
The distances at which a crosslink is considered satisfied for ABAS, EDC, SDA, and TATA crosslinks are < 17 Å, < 10 Å, < 15 Å , and < 14 Å, respectively. Unsatisfied crosslinks were counted in the simulation as a penalty.

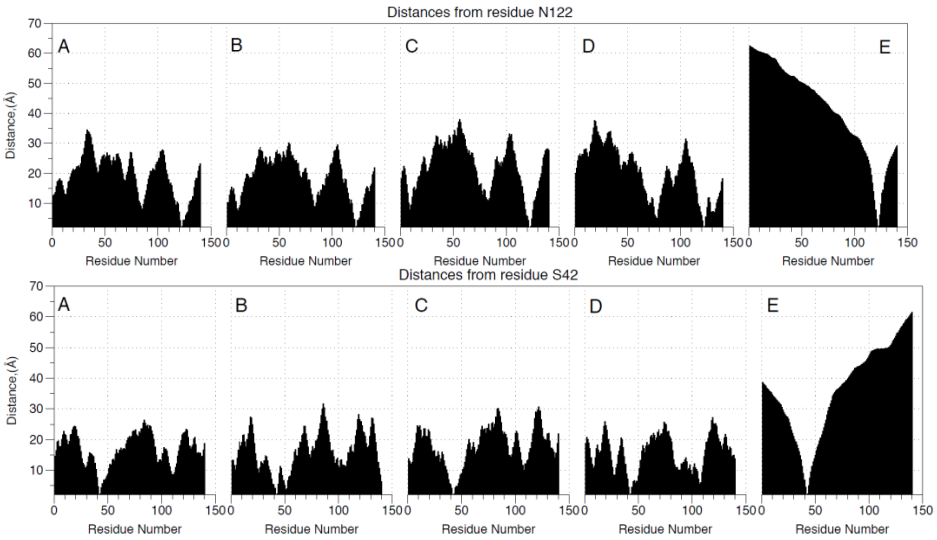
Appendix G: α-synuclein structure fluctuations in the absence of crosslinker restraints



The centroid structure from the lowest energy cluster in the ensemble (Figure 28A) was allowed to relax during simulation without any crosslinking constraints at the temperature of 0.45 (kcal/mol/kb DMD units) for 2 x 10^6 DMD time steps. The first 500k steps were discarded as the system equilibration.

Appendix H: Comparison of CL-DMD and PRE-NMR ensembles





PRE-NMR α-synuclein ensemble corresponding to the publication by Allison et al., JACS 2009 [210] was obtained from Protein Ensemble Database http://pedb.vib.be. Ensemble average inter-residue distances from the residues, which were spin labelled in the PRE-NMR study. Each graph shows the average inter-residue differences between all residues in the ensemble and the spin-labelled residues in the PRE-NMR experiment. Panels A-D correspond to the four major CL-DMD clusters presented in this work, the E panels correspond to the PRE-NMR ensemble.