THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Biologie Santé

École doctorale Sciences Chimiques et Biologiques pour la Santé (CBS2)

Centre de Biologie Structurale (CBS) – CNRS UMR 5048 – UM – INSERM U1054

Structure and Dynamics of Huntingtin. A Segmental Labelling Approach

Présentée par Xamuel Loft LUND
Le 15 décembre 2025

Sous la direction de Pau BERNADÓ, Frank GABEL, et Anne MARTEL

Devant le jury composé de

Mme. Véronique RECEVEUR-BRÉCHOT, Dr., BIP - Marseille (France)	Rapportrice
Mme. Annette Eva LANGKILDE, Dr., University of Copenhagen - Copenhagen (Denmark)	Rapportrice
Mme. Virginie REDEKER, Dr., Institut de Biologie Francois Jacob - Fontenay-aux-Roses (France)	Examinatrice
Mme. Sophie COMBET, Dr., Laboratoire Léon-Brillouin – Gif-sur-Yvette (France)	Examinatrice
Mme. Zoë FISHER, Dr., European Spallation Source - Lund (Sweden)	Invitée
M. Pau BERNADÓ, Dr., CBS – Montpellier (France)	Directeur de thèse
Mme. Anne MARTEL, Dr., Institut Laue-Langevin – Grenoble (France)	Co-directrice de thèse
M. Frank GABEL, Dr., Institut Laue-Langevin – Grenoble (France)	Invité

UNIVERSITÉ DE MONTPELLIER

PREFACE

The work included in this PhD dissertation was performed at the Centre de Biologie Structurale (CBS), University of Montpellier, France, and at the Institut Laue-Langevin (ILL), Grenoble, France. The experimental work was performed between July 2020 and December 2023, with the initial two years mainly being performed at the CBS and the latter one and a half year at the ILL.

The two years at the CBS focused on acquiring knowledge about cell-free protein expression and atomistic ensemble generation under the main supervision of Pau BERNADÓ. The one and a half year at the ILL was spent measuring SANS data and optimizing data analysis of SEC-SANS data under the supervision of Anne MARTEL & Frank GABEL.

The Project was funded as a collaboration between the CBS and ILL, with one third of the funding coming from a Consolidator Grant from the European Research Council and the two thirds being funded by the ILL PhD program.

Small Angle X-ray Scattering experiments were performed at the SWING beamline, Soleil Synchrotron, France, and Small Angle Neutron Scattering experiments were performed at the D22 Beamline, ILL, France.

The dissertation combines the efforts into expression, purification, and validation of protein samples of the Huntingtin Exon1 protein fused to sfGFP, realistic generation of atomistic structures for ensemble refinement, experimental measures of both Small Angle X-ray Scattering and Small Angle Neutron Scattering, and ensemble refinement analysis combining several scattering datasets to improve informational content.

No manuscript has been written as of submission of this dissertation, but it is expected that the method developed during this thesis project and the results obtained from the combined analyses will be collected into a publication.

RESUMES

ENGLISH RESUME

Huntington's Disease (HD) is a genetic neurodegenerative disorder caused by a mutation in the gene encoding the protein Huntingtin (Htt). The mutation causes an increase in the number of CAG trinucleotides in the first exon that, after the translation, increases the number of glutamines in the poly-glutamine (Poly-Q) tract of the intrinsically disordered N-terminal region of the protein. HD symptoms only manifest in individuals with a Poly-Q tract of more than 35 consecutive glutamines, which is named the pathological threshold. Importantly, the length of the Poly-Q tract beyond the threshold is correlated with the age of onset and the severity of the pathology. The exon-1 of Htt is a low complexity region that contains a 17-residue long N-terminal region (N17), the Poly-Q tract and a proline rich region (PRR). Significantly, Htt Exon-1 can recapitulate the disease and is normally used for mechanistic and biophysical studies.

The PhD project aims at elucidating the structural differences between non-pathogenic and pathogenic Htt exon-1 constructs using Small-Angle Neutron Scattering (SANS) measurements in amino-acid specific deuterated samples. Capitalizing on the distinct scattering properties of deuterium and hydrogen, we aim at extracting valuable structural information of the disordered region. Constructs with specific deuteration patterns of glutamine (Q) and proline (P) residues have been produced using the Cell-Free (CF) protein expression system. The accuracy of the labelling process and homogeneity of the samples were confirmed by mass spectrometry.

Atomistic ensemble models of Htt with 16 and 36 glutamines (H16/H36) were computed with the specific deuterium patterns and deuterium exchange mimicking different H_2O/D_2O buffer solutions. The SANS profiles for all the conformations and deuteration states were calculated for subsequent analyses. The theoretical ensembles showed an enrichment in structural information depending on the labelling pattern and the D_2O% of the buffer solution. The analysis of this synthetic data also guided the selection of the optimal samples to be produced and measured by SANS.

Size exclusion chromatography (SEC)-SANS data collected at the D22 Beamline at ILL and Small-Angle X-ray Scattering (SEC-SAXS) data measured at the Swing beamline at Soleil Synchrotron were integrated with the previously mentioned atomistic models using the

ensemble optimization method (EOM). 28 SANS experimental profiles for the pathogenic (H36) and the non-pathogenic (H16) constructs, as well as SAXS data for both H16 and H36 in 0% and 100% D_2O have been collected. The quality of the SANS/SAXS data was determined through extensive cross-validation, and the EOM fitting, combining up to five SANS datasets with SAXS, was performed. The resulting ensembles, compatible with all the experimental datasets, indicate that Htt Exon-1 is an elongated protein with partial structuration in the Poly-Q tract. The pathogenic version of the protein displays an enhanced expansion suggesting that stabilization of the secondary structure in the Poly-Q could be at the origin of the aggregation and the disease.

In summary, I have developed an original method that, exploiting the control of the isotopologues provided by the CF, the precise positioning of deuterium in atomistic models and the capacity to simultaneously fit multiple datasets, enhances the structural information content of SANS data. This strategy paves the way to study the low-complexity regions, a family of proteins that has remained out of the reach to structural biology due to the lack of appropriate methodologies.

FRENCH RESUME

La maladie de Huntington (HD) est une maladie génétique neurodégénérative causée par une mutation du gène codant pour la protéine Huntingtine (Htt). Cette mutation entraîne une augmentation du nombre de trinucléotides CAG dans le premier exon qui, après traduction, augmente le nombre de glutamines dans la poly-glutamine (Poly-Q) de la région N-terminale intrinsèquement désordonnée de la protéine. Les symptômes de la HD ne se manifestent que chez les personnes dont la Poly-Q comporte plus de 35 glutamines consécutives, ce qui est appelé le seuil pathologique. Au-delà de ce seuil, la longueur du Poly-Q est corrélée à l'âge d'apparition et à la sévérité de la pathologie. L'exon-1 de l'Htt est une région de complexité qui contient la région N-terminale longue de 17 résidus (N17), le tractus Poly-Q et une région riche en prolines. Ce fragment de la protéine est habituellement utilisé pour des études mécanistiques et biophysiques.

Mon projet vise à élucider les différences structurelles entre les versions pathogénique et non pathogénique de l'exon-1 de Htt en utilisant des mesures de diffusion des neutrons aux petits angles (SANS) sur des échantillons dont seuls certains acides aminés sont deutérés. Profitant des propriétés de diffusion distinctes du deutérium et de l'hydrogène, nous visons à extraire des informations structurelles sur cette région désordonnée. Des fusions de cet exons avec la GFP (green fluorescent protéine) ont été construites et exprimées dans un système « cell-free » avec des schémas de deutération spécifique des glutamines et prolines. Le succès du marquage et l'homogénéité des échantillons ont été confirmés par spectrométrie de masse.

Les modèles atomiques d'ensembles conformationnels de ces constructions, avec 16 et 36 glutamines (H16/H36), ont été générés et deutérés *in silico* pour reproduire la deutération et l'échange d'hydrogène des échantillons dans leurs tampons. Leurs profils SANS théoriques ont été calcules pour les analyses ultérieures. Les ensembles théoriques ont montré un enrichissement des informations structurelles en fonction du schéma de marquage et du pourcentage de D_2O de la solution tampon guidant le choix des mesures expérimentales de SANS à effectuer.

Les données de diffusion aux petits angles couplée à la chromatographie d'exclusion de taille ont été recueillies sur SWING (SOLEIL) pour les rayons X, et D22 (ILL) pour les neutrons. 28 profils expérimentaux SANS pour les constructions pathogénique (H36) et non pathogénique (H16), ainsi que des données SAXS pour H16 et H36 dans 0% et 100% D_2O ont été collectés et analysés par le Ensemble Optimization Method (EOM) à partir des modèles

décrits ci-dessus. La qualité des données SANS/SAXS a été déterminée par validation croisée et l'ajustement EOM, combinant jusqu'à cinq ensembles de données SANS avec SAXS, a été fait. Les ensembles obtenus, compatibles avec tous l'ensemble de données expérimentales, indiquent que l'exon-1 de Htt est une protéine allongée avec une structuration partielle dans le Poly-Q. La version pathogénique de la protéine augmentation de la partie Poly-Q, ce qui suggère que une stabilisation de la structure secondaire dans le Poly-Q pourrait être à l'origine de l'agrégation et de la maladie.

En résumé, j'ai développé une méthode originale qui, en exploitant la deutération spécifique permise par la synthèse *in vitro*, le positionnement précis des atomes de deutérium dans les modèles atomiques et la capacité d'ajuster simultanément plusieurs types de données, améliore le contenu de l'information structurale des données SANS. Cette stratégie ouvre la voie à l'étude des régions de faible complexité, une famille de protéines qui est restée hors de portée de la biologie structurale en raison du manque de méthodologies appropriées.

ACKNOWLEDGEMENTS

Firstly, I would like to acknowledge and sincerely thank Dr. Pau Bernadó, Dr. Anne Martel, and Dr. Frank Gabel for the opportunity and privilege to do this PhD work under their guidance and tutelage. Naturally, I could not have done this without them, but their support and enthusiasm during what ended up being a five-year project is what kept me motivated. I am lucky to have learned from them, how to be a great scientist.

Dr. Pau Bernadó is a valuable teacher when it comes to cell free expression and theoretical ensemble production, whose knowledge I am fortunate to have benefitted from. He has imparted a strong, scientific curiosity in me that I should be lucky to bring with me in my next professional adventures. Additionally, he has taught me about patience, when correcting the written mis-adventures in this dissertation as well as when letting me be human at work. He always has an ear to lend and has an open door, for which I have been and continue to be grateful.

I am thankful for Dr. Anne Martel's strong guidance when it comes to the practical application of SANS experiments in their setup and conduction, steering me true and clear. In addition, she has also been a voice of reason in not setting the bar too high for myself, when I have been overwhelmed; a force in tempering my perfectionism.

I have benefitted greatly from Dr. Frank Gabel's theoretic knowledge of SAXS and SANS. Likewise, his structured approach and ability to keep the grand overview in focus is something, which I appreciate, and without which I could not have completed this project.

On the same note I would also like to extend my thanks to Dr. Véronique Receveur-Bréchot, Dr. Annette Eva Langkilde, Dr. Sophie Combet, Dr. Virginie Redeker, and Dr. Zoë Fisher for agreeing to be jury members for my PhD defense – their time invested and interest in my work is much appreciated.

I thank the individual monitoring committee, consisting of Dr. Annette Eva Langkilde and Dr. Andrey Kajava, for their vested time and interest in my progress as a PhD student and for keeping me on point and on track.

Heartfelt thanks also go out to Dr. Annika Urbanek, Dr. Carlos Elena-Real, Dr. Amin Sagar, and Dr. Anna Morató for sharing their knowledge (about their areas of research and where to get the best ice cream in Montpellier), for inspiring me, and supporting my growth as a scientist. Community cannot be overrated in accomplishing my dissertation.

On a similar note, an equally heartfelt thanks to the PhD group (the coffee group) at Institut Laue-Langevin (ILL) for their friendship and ongoing support. The work environment, we created together, set me up for success.

Speaking of work environment, I would like to specifically thank Institut Laue-Langevin (ILL) for believing in the project despite COVID-19 lockdowns and reactor remodelling, and extending my contract to allow me ample time to do the necessary experiments. And a similar thanks to Université de Montpellier for listening to me and allowing my humanity in this project by letting me extend my writing period into 2025.

In the same vein, I extend sincere thanks to Dr. Michael Gajhede from University of Copenhagen for also seeing my potential and doing everything in his power to secure me a desk at the university premises in Copenhagen, so I would have a proper working space while in Denmark. And for also giving me the opportunity for further experience by including me in his work and articles on small molecule docking.

For this project, my supervisors and I had the privilege of receiving support from a Consolidator Grant from the European Research Council (ERC-CoG) in combination with the Institut Laue-Langevin PhD program funding.

On a more personal note, to Christian Kragh Pedersen I say thanks for catching me, when I needed it. Your support in managing my stress was and is more important than I can express.

I would like to thank Dr. Rasmus Agerholm for being an understanding listener, when I needed to be mirrored in the work as PhD student; your thoughtful inputs and friendship is greatly valued by me.

And to my mum, Ingelise Loft: though you were not here to see me accomplish this and will never read this dissertation, I am so thankful for the woman and mother you were. For teaching me that one can do hard things. And for inspiring to me do the master thesis I did, which set me on the path for this PhD project.

Finally, to my wife, Nanna Fjellerad: you were my rock in this project. Your patience, stability, mental guidance, understanding, care, and love is what kept me afloat, when the project was challenging. Words cannot describe how grateful I am and how much I treasure you. With you by my side, I can accomplish anything.

TABLE OF CONTENT

ABBREVIATIONS

AF	Alphafold
ATP	Adenosine triphosphate
cAMP	3',5'-cyclic adenosine monophosphate
CF	Cell-free
CG	Coarse-Grained
CP / CK	Creatine Phosphate and Creatine Kinase
Cryo-EM	Cyrogenic electron microscopy
D_{max}	Maximum intramolecular distance
DNA	Deoxyribonucleric Acid
EOM	Ensemble Optimisation Method
FDA	Food and Drug Administration
FPLC	Fast Protein Liquid Chromatography
H16	Huntingtin Exon1 Construct with 16 glutamines
H36	Huntingtin Exon1 Construct with 36 glutamines
HAP40	40 kDa Huntingtin Associated Protein
HD	Huntington's Disease
HEAT	Huntingtin Elongation factor 3
Htt	Huntingtin
I(0)	Forward scattering intensity
IDP	Intrinsically Disordered Protein
IDR	Intrinsically Disordered Region
IMAC	Immobilized Metal-ion Affinity Chromatography
IPTG	Isopropyl β-D-1-thiogalactopyranoside
kDa	kilodalton
KGlu	Potassium Glutamate
KOAc	Potassium Acetate
LCR	Low-complexity Region
mHtt	Mutant Huntingtin
mRNA	Messenger Ribonucleric Acid
MS	Mass spectrometry

MSN	Medium Spiny Neurons
MW	Molecular weight
N17	N-terminal Initial 17 amino acids
Ni-NTA	Nickel-ion Nitriloacetic Acid
NMR	Nuclear Magnetic Resonance
OD	Optical Density
P(r)	Pair-wise Distance Distribution
pdb-file	Protein Database File Format
Poly-Q	Poly glutamine tract
PRR	Proline Rich Region
PURE	Protein Synthesis using Recombinant Elements
R_g	Radius of Gyration
RNA	Ribonucleic Acid
SANS	Small Angle Neutron Scattering
SAS	Small Angle Scattering
SAXS	Small Angle X-ray Scattering
SBMA	Spinal and Bulbar Muscular Atrophy
SCA	Spinocerebellar Ataxia
SDS-PAGE	Sodium Dodecyl Sulphate polyacrylamide gel electrophoresis
SEC	Size Exclusion Chromatography
SEC-SANS	Size Exclusion chromatography Small Angle Neutron Scattering
SEC-SAXS	Size Exclusion chromatography Small Angle X-ray Scattering
SEP	Synthetic Enzymatic Pathways
sfGFP	Superfolder Green Flourorescent Protein
SLD	Scattering Length Density
VMAT2	type-2 vesicular monoamine transporter
χ2-free	Comparison χ^2-value from cross-validation
χ2-work	Fitting χ^2-value from EOM

1 INTRODUCTION

1.1 HUNTINGTON'S DISEASE

Huntington's disease (HD) is an inheritable, neurodegenerative disease, which is also known as Huntington's Chorea. While symptoms consistent with HD had previously been described, the first accurate description of the disease was in 1872 by Dr. George Huntington (1,2). HD presents itself primarily in brain tissue as degeneration of medium spiny neurons, which is thought to be the cause of disease symptoms (2–4) (Figure 1.1). Symptoms consistent with HD include, but are not limited to, motor symptoms, such as muscle jerk and gait disturbance, with later stages presenting bradykinesia, rigidity, or loss of the ability to walk; psychiatric symptoms, such as depression, apathy, irritability, and change of sexual behavior; and cognitive symptoms, which may affect anything from memory and attention to mental flexibility, which can result in dementia and cognitive inhibition at later stages of the disease (2,3,5,6).

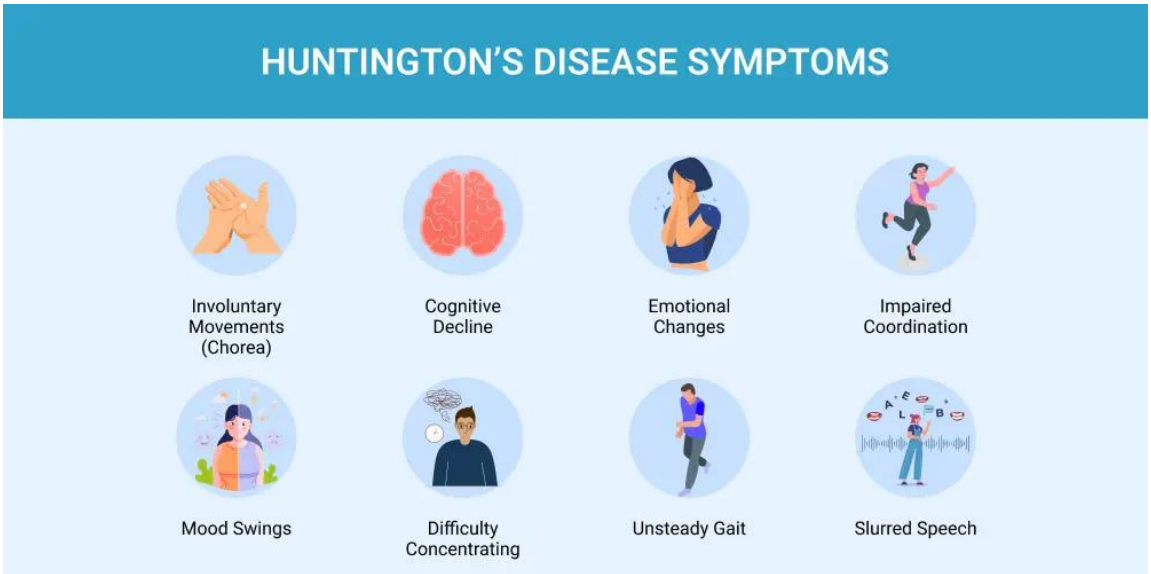


Figure 1.1: Common symptoms of Huntington's Disease. Extracted from https://lonestarneurology.net/blog/what-is-huntingtons-disease/.

The disease pathology is characterized by an abnormal extension of CAG-trinucleotide repeats in the exon 1 of the gene responsible for encoding the protein Huntingtin (Htt) (3). This is located on chromosome 4p16.3 (3). A pathological threshold of 36 CAG repeats has been identified (7,8), but full penetrance of the disease is not reached before 40 – 42 repeats (7–11). The mean age of onset of HD is 40 years, however, it has been demonstrated that the age of

onset is influenced by the length of the CAG repeats (5). Therefore, repeats exceeding >60 consecutive codons can lead to juvenile HD, which accounts for approximately 5% of individuals affected by HD (12,13). Patients affected by HD typically die 20 years following the onset of the disorder (3,14).

Furthermore, the hereditary nature of HD contributes to geographical variation in its prevalence (15). The approximate prevalence of HD is 4-12 cases per 100.000 in western populations, while Asian and African countries have reported numbers around 0.1- 0.7 per 100.000 (2,5,15). While the number of individuals affected by HD has increased, this trend could be linked to a rise in life expectancy and the decrease of stigmatization associated with HD patients (15–17).

1.1.1 Pathology of Huntington's Disease

The abnormal expansion of CAG codons in the *huntingtin* gene results in an elongation of a poly-glutamine (poly-Q) tract in the N-terminal region of the protein Htt (18). The expanded poly-Q tract is a trait that HD shares with eight other poly-Q disorders, such as Haw River syndrome, Spinal and bulbar muscular atrophy (SBMA) and Spinocerebellar Ataxia (SCA) (19,20). Poly-Q diseases have been linked to the presence of aggregates in regions of the brain and are known to cause neuronal death (21), which I will describe in further detail in the next paragraph. While Htt is a structured protein containing ~3,142 amino acids with a molecular weight of ~347.6 kDa (3), the pathogenicity has been attributed to smaller fragments (50-150 kDa) of Htt, the aggregation of which forms inclusion bodies in both the cytoplasm and nucleus of striatal neurons (21–24). Interestingly, it has been hypothesized that the presence of inclusion bodies could be the cell's mechanism to inhibit the degenerative effect of mutant Htt (mHtt) fragments in the nucleus (25). The occurrence of inclusion bodies have been suggested to possess protective qualities by reducing the levels of soluble mHtt aggregates (25–27). Soluble aggregates of Htt fragments are thought to drive the apoptosis of the nerve cells (26,28). Moreover, it has been observed that synthetic mHtt can be absorbed by cells (29,30). As in other prionic diseases, the soluble aggregates have also been shown to transit between neighboring cells, inducing cell death (31,32).

In healthy neuronal cells, the proteasome system and autophagy will clear damaged or misfolded proteins, and degrade protein complexes and damaged organelles, respectively. Conversely, in HD, both of these pathways are impacted, leading to a decrease of activity and accumulation of mHtt (33,34). It has been suggested that increasing the activity of these pathways or, alternatively, decreasing the levels of aggregated Htt would have a

neuroprotective effect in HD (35), potentially making the disease less severe. The most characteristic impact of the HD is the death of medium spiny neurons (MSN) in the striatum. Indeed, in later stages of HD, up to 90% of these neurons are dead (36,37). The molecular mechanism behind neuronal cell death in HD has been linked to glutamate excitotoxicity, which causes increased levels of glutamate in neuron synapses due to elevated glutamate release and inhibited clearance of glutamate from the synaptic cleft (38,39). The symptoms of HD can further be attributed to dopaminergic imbalance (40), mitochondrial dysfunction (41,42) and neuroinflammation (43). The prolonged decline of the number of MSN is linked to the early symptoms of HD, such as involuntary movement and chorea (37). While the death of MSNs is the major direct effect of HD, other regions of the brain, such as the cerebral cortex, are also impacted (44).

There has been an intense search for biological markers of HD, which would both be present in pre-symptomatic HD cases and preferably act as a measurement of disease progress throughout the disorder's development. Longitudinal studies, such as TRACK-HD and PREDICT-HD, have aimed at elucidating biological and clinical markers of HD development (45,46). These studies show that currently the most promising biomarkers are neurofilament light chain (NfL) (47,48) and soluble Htt in cerebrospinal fluid (49).

1.1.2 Treatment of Huntington's Disease

Over 240 clinical studies have been registered on clinicaltrials.gov (https://clinicaltrials.gov/search?cond=Huntington%27s%20Disease), investigating several approaches to remediate HD. A recent review showed that drug candidates had been tested in regards to excitotoxicity, the dopaminergic pathway, mitochondrial dysfunction, neuroinflammation, mHtt aggregation, transcriptional dysregulation and DNA/RNA targeting approaches (15). While many trials have failed to show significant efficacy, three drugs (tetrabenazine, deutetrabenazine and valbenazine) have been approved by the American Food and Drug Administration (FDA) for treatment of chorea in HD (50–52). The three drugs are type-2 vesicular monoamine transporter (VMAT2) inhibitors that decreases the bioavailability of synaptic dopamine and reduces dopaminergic signaling. The use of VMAT2 inhibitors in HD treatment has been shown to alleviate motor symptoms and reduce striatal neuron death (53).

Another interesting approach to HD treatment is the use of the CRISPR system, which is theorized to have the ability to offer a long-term solution for patients suffering from HD (54).

Positive results in animal studies have already demonstrated the concept of using RNA-targeting CRISPR-Cas13d to clear toxic CAG-rich RNA, reducing the mHtt levels in striatal neurons (55).

1.2 The Structure of Huntingtin

As aforementioned in the introductory paragraph, Htt is a structured protein with a molecular weight of ~347.6 kDa. Htt has three major domains, two domains of several HEAT (Huntingtin, elongation factor 3, protein phosphatase 2A and lipid kinase TOR1) repeats, separated by a smaller flexible domain acting as a bridge (56). The HEAT repeats consist of approximately 40 residues and their structure is primarily α-helical (56,57). The general structure of Htt was elucidated by cryo-electron microscopy in 2018 at an overall resolution of 4 Å (56).

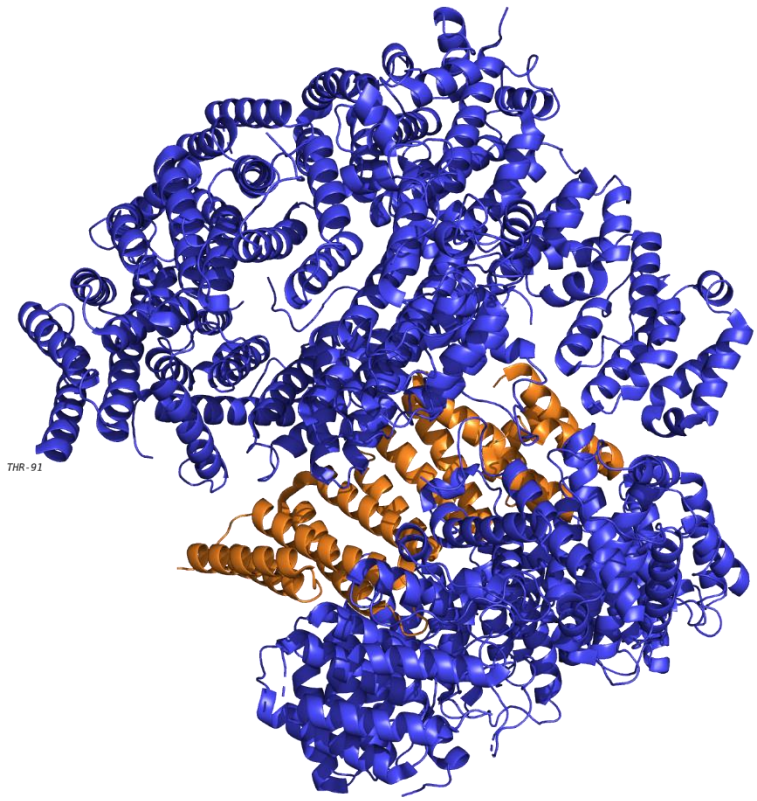


Figure 1.2: Molecular structure of Huntingtin in complex with HAP40 obtained from CryoEM at 3.6 Å (58). PDB-ID 7DXJ. Htt in blue and HAP40 in orange. The initial N-terminal residues were not modelled due to flexibility meaning the first residue of the structure is Thr-91.

This structure was obtained in the presence of the 40 kDa Htt-associated protein (HAP40) and the two proteins formed a complex (59). The smaller HAP40 protein is primarily α-helical and binds in the cleft formed between the N- and C-terminal HEAT domains. It is suggested that

the complex stabilizes the Htt, which is otherwise relatively dynamic (56,59). A structure of Htt has also been determined at a lower resolution of 18.2 Å by cryo-EM and this structure suggests that monomeric free Htt is significantly different from the complexed one (60). Cryo-EM experiments have further suggested that the movement of the Htt protein is impacted by the length of the poly-Q (60). The Htt exon-1 domain contains two low complexity regions, the poly-Q and proline-rich region (PRR). The poly-Q tract includes nine or more consecutive glutamines (3), while the PRR contains 28 prolines, including two stretches with 11 and 10 consecutive prolines. Importantly, the published structures of Htt, regardless of resolution, are omitting the N-terminal exon-1 containing the poly-Q and the proline-rich region (PRR). The disordered nature of the exon-1 domain renders it too flexible to be determined by cryo-EM (56,59–61).

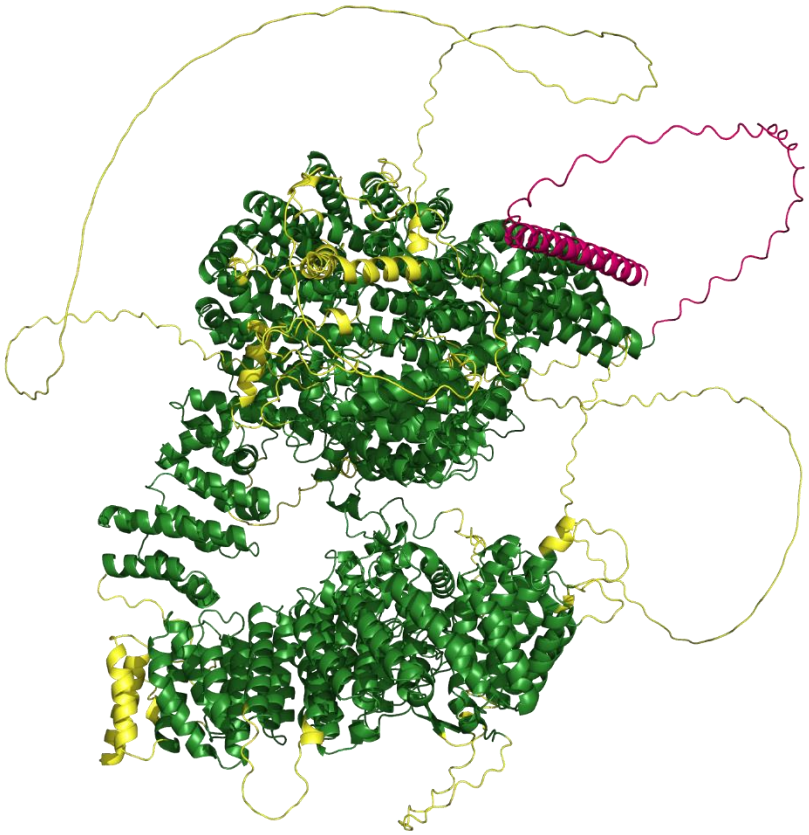


Figure 1.3: Alphafold (AF) Model of full length Huntingtin. Protein regions not accounted for in the cryo-EM structure (PDB ID: 7DXJ) are coloured in yellow, with the exon-1 coloured in pink. This prediction shows that several parts of the protein can not be predicted well as they become large unstructured loops in the AF model. The initial N17 and Poly-Q domain of exon-1 are both predicted to be fully alpha helical and in close proximity to the protein.

By predicting the structure with Alphafold (AF), we can visualise the flexibility of huntingtin (Figure 1.3). When compared to the experimental cryo-EM structure, several large loops (yellow loops in Figure 1.3) appear in the AF model. These loops are not explained by the cryo-EM due to flexibility and thus AF can not accurately predict their structure. The Exon-1 (pink domain in Figure 1.3) appears to depict the N17 and the poly-Q as fully alpha helical domains with the PRR region and the remaining residues between that and the folded domain to be fully disordered as a large loop. The large loops predicted from AF is a typical response to disordered or flexible regions (62). The fully alpha helical structure of the N17 and poly-Q domain and close proximity to the protein could render it stable enough to determine from the cryo-EM, but as this is not the case, this region is still to be considered flexible.

The structure of the exon-1 and poly-Q regions in general have been investigated in several studies, a definitive answer is not available. Previous studies of poly-Q regions have shown it to adopt both an alpha helical structure and a disordered structure depending on the flanking regions (63,64). This structural flexibility is also supported by NMR studies that show a decreasing propensity for alpha-helical content when approaching the PRR (65,66). These studies suggest that the poly-Q of Htt is in a constant equilibrium of conformations with varying degrees of alpha-helical structure. The PRR is suggested to be fully extended due to the rigidity of poly-proline tracts (67). This continuous shift of helical content in poly-Q does not support the model calculated from AF, which shows the entire glutamine tract as an alpha helix, but rather that it could be one of several conformations defining this region (63,65,66).

1.3 LOW-COMPLEXITY REGIONS AND HOMOREPEATS IN DISORDERED PROTEINS

Intrinsically disordered proteins (IDPs) are characterized by not being able to spontaneously fold into stable three-dimensional structures, although they can contain partially folded secondary structures (68). IDPs perform their biological function through interactions or by facilitating interactions with other biomolecules (68,69). The range of functions that has been reported for IDPs is very broad, facilitated by their conformational plasticity. Indeed, functions performed by disordered proteins complement those of their folded counterparts. Enrichment in disorder is typical in signalling pathways and regulation of cellular processes, such as translation, transcription, and cell-cycle (68,70–72). The amount of disorder varies in the different kingdoms of life. For instance, more than 30% of proteins found in eukaryotic proteomes present disordered regions of more than 30 consecutive residues, while this percentage decreases dramatically in bacteria and archaea (73). Proteins achieve their intrinsic

disorder because of amino acid compositions that differ from those found in globular proteins. IDPs/IDRs are enriched in small, polar, and charged amino acids, and they are depleted from large and hydrophobic ones (74).

Low-complexity regions (LCRs) are protein fragments that encompass a bias in their amino acid composition (75,76). The enrichment in a single or a few amino acids can significantly impact the structural and functional properties of proteins. Homorepeats, which are protein fragments composed of a single amino acid, are a very eye-catching family of LCRs. LCRs in general, and homorepeats in particular, are poorly structured and are most of the time found inserted in IDPs or intrinsically disordered regions (IDRs). The population of LCRs in eukaryotic proteomes is likewise large with around half of the proteins containing at least one LCR, which is accounting for up to 25% of the coding sequence (75,77).

Protein homorepeats with runs of at least 6 repeated amino acids are significantly more common in eukaryotic cells compared to prokaryotes. The most common homorepeats in eukaryotic cells are Poly-P, Poly-G, and Poly-S (followed by A, N, and Q), while prokaryotic cells are comprised of A, G, and P tracts, with the other residues only occurring seldomly (78). Homorepeats can exhibit transient structures or structural tendencies depending on the repeated amino acid. It has been shown that the structural tendency of pure amino acid stretches can be categorized into four groups (Figure 1.4) (75). The four groups include α-helical promoting, extended (β-strand and PP-II) structure promoting, other structural conformations, such as the different types of turns, and tendencies towards all three conformations (75). In addition to conformational preferences predicted for the 20 canonical amino acids, the length of the homorepeat is known to have an impact on the degree of secondary structure observed (66,79). Additionally, the residues flanking the homorepeat can also impact their structural properties, as demonstrated for Poly-Q (65,80). The modulation of the secondary structure by flanking regions is likely true for all types of homorepeats. Recent bioinformatics analyses of the sequence context observed for the 20 types of homorepeats indicate that all of them display compositional preferences for the flanking regions and that, in some cases, can be asymmetrical for the N- and C-flanking regions (65,75,81–83).

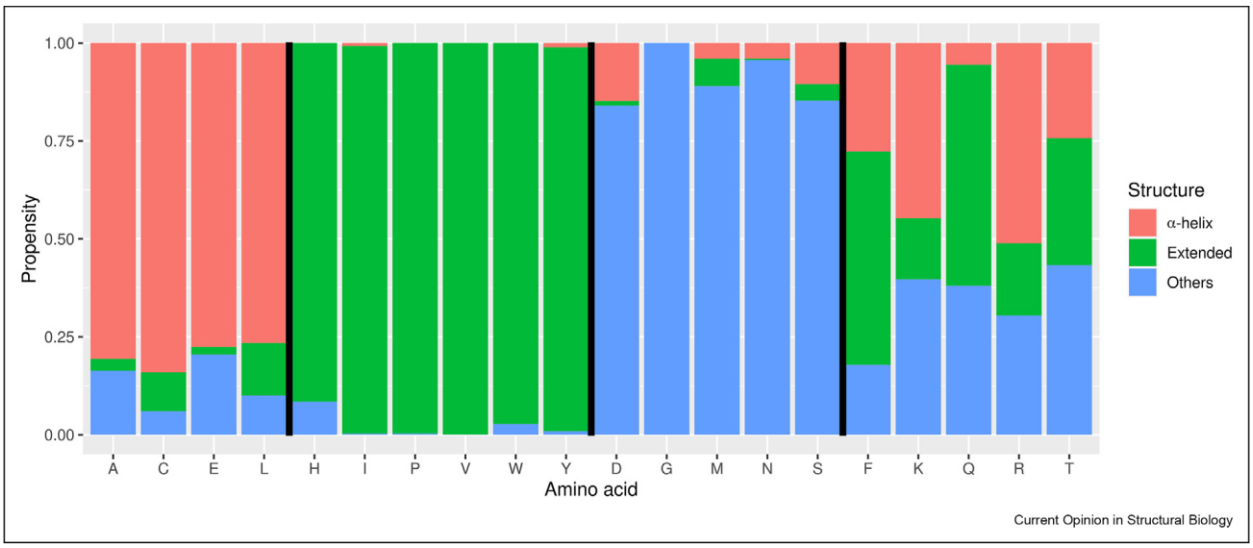


Figure 1.4: Extracted from Elena-Real et al. 2023 (75). Secondary structure preferences in pure stretches of amino acids. Fraction of predicted a-helical, extended (β-strand and PP-II) and Others, which joins all other secondary structure combinations that are neither purely a-helical nor extended. Predictions were performed for 30-residue long poly-X fragments with the LS2P program (84). Note that only the central stretch was analyzed. Amino acids are classified in four groups according to their preferred predicted conformation.

Homorepeats are often connected to human diseases, and especially abnormal expansions of Poly-Q and Poly-A tracts, which have been observed in several neurodegenerative and developmental diseases, respectively (75,85–87). The pathogenic expansion of poly-X tracts causes elongation of existing endogenous LCR sequences. For Poly-Q and poly-A, this change in length shifts the balance between the protein function and toxicity (75,88).

As mentioned above, Poly-Q repeats have been linked to nine different diseases including Huntington's disease, spinal and bulbar muscular atrophy, dentatorubral pallidoluysian atrophy, and several types of spinocerebellar ataxias (89,90). Poly-Q has an α-helical tendency in the case of Huntingtin, but in addition, upon its pathogenic expansion the α-helix is enlarged and stabilized, increasing aggregation propensity (75).

Unfortunately, the hypothesized, aforementioned structural preferences have not been experimentally determined for the vast majority of homorepeats. Due to the typical flexibility and the compositional bias of the homorepeats, the structure of these domains is difficult to ascertain. Additionally, high resolution structures obtained from protein crystallography or electron microscopy omit the flexible regions often including LCRs, leaving nuclear magnetic resonance (NMR) as the only high-resolution technique able to study this family of proteins. However, the compositional bias precludes the frequency assignment required for the structural and dynamic investigation of biomolecules by NMR. That being said, recently, a new strategy

enabling the isotopic labelling of individual positions within homorepeats has paved the way to the high resolution investigation of homorepeats by NMR independently of their length (91,92).

1.4 SMALL ANGLE SCATTERING

Small angle X-ray (SAXS) and neutron (SANS) scattering are methods used to analyse the nanometer-scale structure of a sample. These techniques can be used to investigate different types of matter, including polymers, biological macromolecules, metal alloys, and nanoparticles (93–100). Small angle scattering (SAS) experiments allow biological samples to be measured in solution and monitor a dynamic system over the course of a reaction. SAS represents a unique opportunity to obtain structural information of *in vitro* systems in solution, such as flexible proteins or large biomolecular complexes, which are difficult to obtain by techniques such as protein crystallography or cryogenic electron microscopy (cryoEM). In addition, SAS can be used in conjunction with higher resolution techniques, such as protein crystallography, cryoEM and nuclear magnetic resonance (NMR) to describe biomolecular systems.

SAXS was initially introduced in the original work of André Guinier from 1939 (101) and further described in a comprehensive textbook in 1955 (102). The SANS methods of solvent contrast variation and specific deuteration was developed in the 60's and 70's (103–105). SAXS experiments are primarily performed at large synchrotron facilities, such as the Synchrotron Soleil and the European Synchrotron Radiation Facility (ESRF) in France, or the Deutsches Elektronen-Synchrotron (DESY) in Germany, but smaller lab sources are also available to perform X-ray experiments in-house (106). Neutron experiments are conducted at either nuclear reactors or spallation sources (107). There are several high flux neutron sources in Europe including the Institut Laue-Langevin (ILL), Forschungsreaktor München II (FRMII), the Budapest neutron center (BNC), the Nuclear Physics Laboratory (NPL), the Reactor Institute Delft (RID), the TRIGA reactor at Johannes Gutenberg-Universität Mainz (TRIGA JGU), the TU Wien Atominstitut (ATI), the National Centre for Nuclear Research (MARIA), and three spallation sources: the Swiss Spallation Neutron Source (SINQ), the Institute of Science in Society Neutron and Muon Source (ISIS), and the European Spallation Source (ESS). All of these research neutron sources in Europe have formed the LENS consortium (League of Advanced European Neutron Sources) [https://lens-initiative.org/].

Nuclear reactors provide a constant stream of neutrons during operation through nuclear fission, which is characterized by hitting an atom with a neutron and the impact cleaves the element into lighter elements, free neutrons, and radiation. Such a chain-reaction is caused by the neutrons, released at a fission event, proceeding to cleave neighbouring atoms. These reactions within the reactors are controlled by absorption or diversion of neutrons (108). Fission reactors thereby generate energy by the chain reaction of neutron induced fission of heavy elements, typically uranium (isotope U235) (108). The research reactors then divert neutrons from the core via guides inserted into the core-vessel, which allow neutrons to travel to the experimental beamlines. The neutron beam is subsequently passed through a moderator material to adjust the wavelength to the experimental setup (109).

Spallation sources generally offer pulsed bursts of neutrons, which are released by hitting a neutron-rich element with a pulsed high-energy proton beam (110). In these instances, target materials are typically tungsten, lead, or mercury, and the number of neutrons released depends on both the energy of the proton beam and the chosen material (110). The released neutrons are then moderated by passing them through a thin layer of moderation material such as hydrogen, deuterium, H_2O etc., after which they are guided to the experimental stations (110).

While both methods offer low resolution structural information, SAXS has been used more due to the ease of access to facilities (111,112). Several software packages have also been published over the years to facilitate simulation, fitting and validation of SAS data (111,113–118). Further developments in sample environments, such as size exclusion chromatography SAXS (SEC-SAXS), have also increased the type of samples that can be measured (119–121). The on-line SEC-SANS has recently been developed at the D22 Beamline at the ILL (122,123). SEC-SAS is ideal for samples that are prone to aggregate, which were previously more difficult (or impossible) to measure.

In biological samples, the X-ray and neutron probe interacts with the electrons and nuclei of a given molecule, respectively. The scattered beam is subsequently measured by detectors as a function of the scattering angle 2θ and described by the function of Q described as $Q = \frac{4\pi}{\lambda}\sin\theta$, where λ is the wavelength (93,124).

1.4.1 Coherent Scattering

While SAXS and SANS are fundamentally different, the mathematical description of the scattering is the same and will therefore be described in parallel. Differences between both techniques will be underlined along the text.

X-ray photons used in structural studies typically have an energy, E, of approximately 10 keV. Because the wavelength of photons is proportional to the energy, the wavelength is given by $\lambda = \frac{12.56}{E} = \frac{12.56}{\sim 10KeV} = \sim 1.26$ Å. Similarly, the neutron wavelength can be calculated by de Broglie's relationship: $\lambda(nm) = 396.6 \,/\, v\,(ms^{-1}) \approx \lambda(\text{Å}) = 396.6 \,/\, v\,(ms^{-1}) * 10$. Here υ is defined as the velocity of neutrons. The experimental wavelength is typically selected by a diffraction monochromator when doing SAXS and SANS experiment, but SANS beamlines, such as the D22 at the ILL, use a velocity selector to select the experimental wavelength, as it reduces the loss of intensity of the beam (Figure 1.5). The velocity selector is a spinning drum with angled helical lamellae, which, while spinning, will allow only neutrons matching the rotational speed to pass. The velocity selector will not isolate a single wavelength, but rather a range of wavelengths, resulting in the wavelength at the D22 being around 6 Å $\pm 10 - 20\%$.

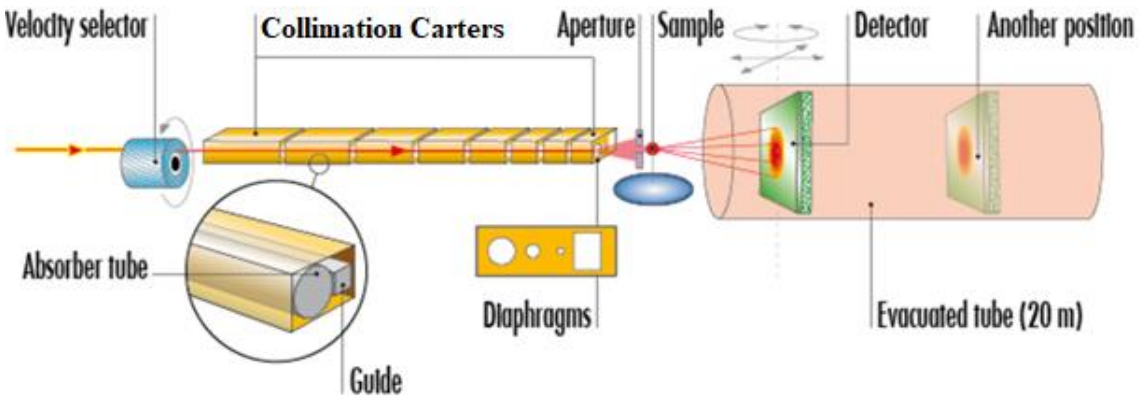


Figure 1.5: The instrumental setup at the D22 (https://www.ill.eu/users/instruments/instruments-list/d22/description/instrument-layout). The neutron guide funnels neutrons into the velocity selector, followed by the collimation carters, before reaching the sample environment and the subsequent scattering is recorded without the vacuum detector tube.

SANS instruments, such as the D16 at ILL, use a monochromator to select their wavelength with a better resolution, but have a lower maximum flux than using a velocity selector (D22 has a maximum flux of $1.2 * 10^8\ cm^{-2}s^{-1}$, while the D16 one is $2 * 10^7 cm^{-2}s^{-1}$) (125).

The basis of SAS is the scattering of the incident beam by an atom. The incident beam is described as a plane wave with the wavevector $k_0 = |\overrightarrow{k_0}| = 2\pi/\lambda$ and gives rise to a spherical scattered wave. At the scattering incident and assuming elastic scattering, the scattered wave is equal to the incident wave $k_1 = |\overrightarrow{k_1}| = |\overrightarrow{k_0}|$ (Figure 1.6).

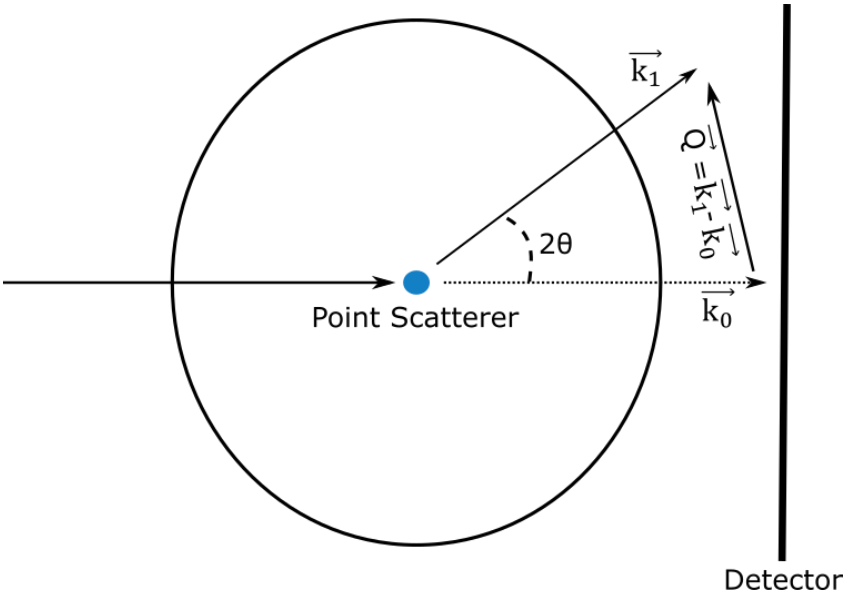


Figure 1.6: Illustration of the scattering event with the incident beam $\overrightarrow{k_0}$ and the scattered beam $\overrightarrow{k_1}$ at the angle 2θ. The vector $\vec{Q}$ is defined as $\overrightarrow{k_1} - \overrightarrow{k_0}$

Because the neutron wavelength is much larger than the size of the nuclei (variation of nuclear potential (f_p)), the atom nucleus can be considered as a point-like object in SANS. The amplitude of the scattered wave is dependent on the choice of technique and is described as the scattering length f. In SAXS, the scattering length of an atom from the sample depends on its number of electrons, $f_x = N_e r_0$, where N_e is the number of electrons and r_0 is the Thomson radius ($r_0 = 2.92 \cdot 10^{-13}\ cm$). When two scattering centres are present in the sample at a given distance (r), the wave scattered from each electron cloud can interfere with each other's waves constructively and destructively. The interferences result in the intensity wave, combined of both the coherent and incoherent signal. The incoherent part of the signal does not depend on the structure of the molecule and is considered a background scattering in SANS. In SANS, the scattering length depends on the isotopic composition. Neutrons interact with both the nuclear potential and nuclear spin, described as $f_N = f_p + f_s$ (93).

Table 1.1: Scattering length in X-ray and Neutron scattering. While the atomic number has a large impact on the X-ray scattering length of an atom, the neutron scattering length is dependent on isotopic content (93,126).

Atom	**H**	**D**	**C**	**N**	**O**	**P**	**S**	**Au**
Atomic mass	1	2	12	14	16	30	32	197
N electrons	1	1	6	7	8	15	16	79
f_x *(10^{-12} cm)*	0.282	0.282	1.69	1.97	2.16	3.23	4.51	22.3
$f_{N\text{-coherent}}$ *(10^{-12} cm)*	-0.374	0.667	0.665	0.940	0.580	0.510	0.280	0.760
$f_{N\text{-incoherent}}$ *(10^{-12} cm)*	2.527	0.404	0.00	0.202	0.00	0.020	0.00	-0.184

The nuclear potential (f_N), as shown in table 1.1, varies depending on the isotopic nature of the object. The very large difference between the coherent scattering length of hydrogen and deuterium is taken as an advantage for the selective labelling and contrast variation experiments, as it will be described in section 1.4.2.

When describing scattering experiments, we use the Fourier transformation to go from the real space with the coordinates $\vec{r}$, to the reciprocal space described by the scattering vector $\vec{Q} = \overrightarrow{k_0} - \overrightarrow{k_1}$. This reciprocal relationship implies that the larger interparticle distances are coded by the smallest measured values of Q. This relationship is given by $Q = 2\pi/r$. When the scattering event is considered as a single incident, the scattering length density (SLD) distribution, $\rho(r)$, is defined as the total scattering length of all atoms in the unit volume. In a solvent of a constant $\text{SLD}_{\text{solvent}}$ (ρ_s), the difference of density between a scatterer and its solvent in a specific volume is the contrast defined as $\Delta\rho = \rho(r) - \rho_s$. We can define the scattering amplitude by the following Fourier transformation:

$$A(Q) = \int_V \Delta\rho(\vec{r}) e^{-i\vec{Q}\cdot\vec{r}} d\vec{r}$$

Because the amplitude is a complex number it can not be directly measured in experimental settings, however, the intensity can. Indeed, the Intensity ($I(Q)$) is defined as the scattering amplitudes ($A(Q)$) multiplied by their complex conjugate ($A^*(Q)$): $I(Q) = A(Q)A^*(Q)$ and it is proportional to the number of scattered X-ray photons or neutrons of the incident beam in the direction of the scattering angle 2θ (127). For a particle of homogeneous contrast, the strongest intensity is typically observed at the zero-angle *I(0)*. The *I(0)* is commonly used to describe the total sum of scattering atoms and calculate molecular mass of a molecule (93,127).

In crystallographic scattering, the sample is considered an ideal crystal which implies that the sample orientation and distribution is defined throughout the sample. Imperfect crystals can

also be measured, but they will not be able to detect disordered or flexible regions of a protein. In solution SAS experiments, the sample is present in all orientations and the presence of flexibility in the molecule introduces different distances between scattering centers. The scattering intensity of an object can be described as the particle scattering or 'form factor' and the 'structure factor', which depends on interactions between particles within the sample. In ideal solutions, *i.e.* when dissolved molecules do not interact with each other, the structure factor will be equal to 1. In practice, the two factors of the intensity can be separated by conducting experiments at different concentrations (93). The experimentally measured intensity will be an averaged intensity of all scattering atoms in all orientations and distances as a function of the scattering angle. In addition to molecular interactions, solutions of flexible or disordered proteins will have a conformational polydispersity and the intensity will also be an average of all conformations present in the sample.

Additionally, the solution itself will add to the scattered intensity. Unless the solute contains characteristic macromolecular distances, its scattering is coherent and incoherent, and will yield a constant intensity at low Q-values, which can be subtracted from the measured intensity of the sample to extract the coherent part of the scattering signal. This will result in the coherent scattering profile of the target molecule in solution, which is subsequently analysed (Figure 1.7).

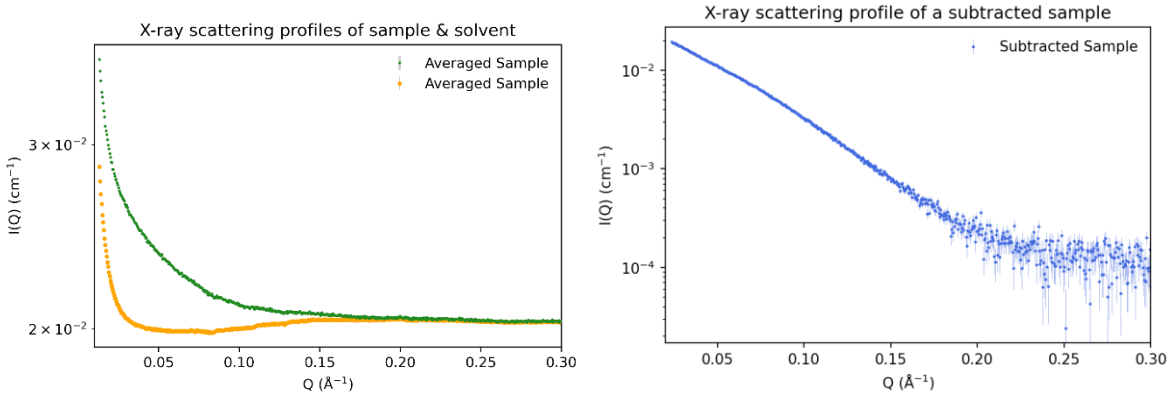


Figure 1.7: Example of X-ray scattering data before and after background subtraction. The profile of the solvent (orange) is subtracted from the sample (green) and yields the experimental profile (blue).

1.4.2 Contrast variation in SANS experiments.

As shown in table 1.1, the scattering length density of a molecule is highly dependent on the isotopic content of the sample. The large difference in SLD between hydrogen and deuterium has led to methods that exploit this difference. Complexes of protein-protein, protein-DNA and protein-RNA are common targets for this method and early examples of contrast variation date

back to the 1960s and 70s (104,128–130). As previously described, the contrast of a volume element can be written as $\Delta\rho = \rho(r) - \rho_s$ and when molecules are measured, $\rho(r)$ describes the average SLD of the molecule under investigation. The situation where the average $\rho(r) = \rho_s$, *i.e.* the contrast between the solute and the solvent is zero, is called the "match point".

In the case of multi-component solutions, the equation for averaged intensity can be alternatively written as:

$$I(Q) = (\Delta\rho_1)^2 I_1(Q) + 2\Delta\rho_1\Delta\rho_2 I_{12}(Q) + (\Delta\rho_2)^2 I_2(Q)$$

Here, 1 and 2 denote the first and second component of a system while I_{12} is the cross-term of the scattering (93). In the event that the SLD of one component becomes equal to the solution or "matched out", the recorded scattering intensity would only retain information about the visible (*i.e.* "non-matched") component (Figure 1.8 (chromatin isolation)).

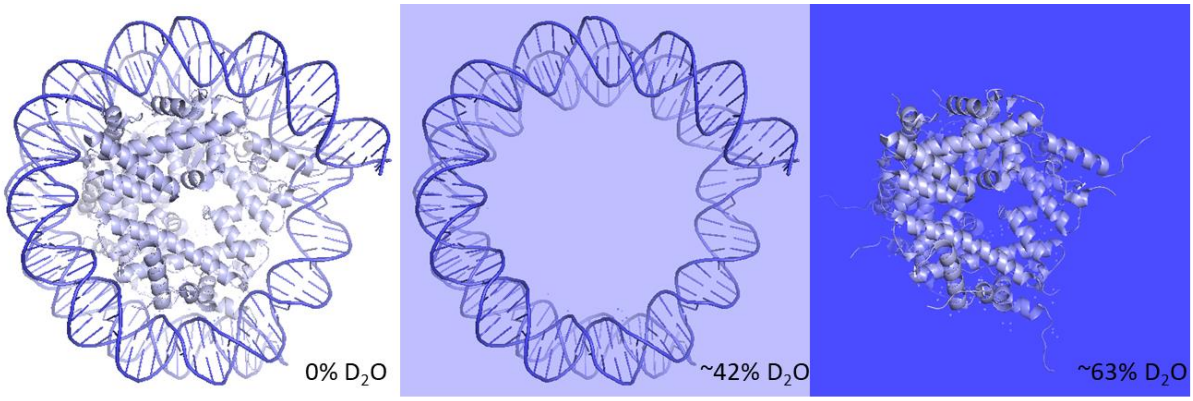


Figure 1.8: The theoretical use of match out experiments. Chromatin is a complex that contains a protein core of histones with DNA wrapped around it forming beads of condensed DNA. In a theoretical SANS experiment the scattering signal of either the protein core or the DNA can be isolated by matching the solution D_2O% to either of the components. The average values of SLD match-points have to be adjusted to the specific component as it is dependent on the specific residue/base composition of the complex.

An important factor in the solute and solvent SLD is the hydrogen content. Knowing the SLD of a molecule, one can use it to match the SLD of the molecule to that of the solvent, thereby masking the signal of the object (124,129,131). The SLD of an aqueous solvent can be adjusted between -0.56 x10^6 Å^{-2} and 6.34 x10^6 Å^{-2} by increasing its D_2O content from 0 to 100%, respectively. In general, the solution will match an average protein around ~42% D_2O and DNA around 63% D_2O as observed in Figure 1.9 (129). Note that the SLD of biomolecules also changes when increasing the D_2O content of the solution. This is due to the equilibration of the exchangeable hydrogens of biomolecules with the H/D content of the solution. This exchange must be considered when computing match points from sample composition. This is

applicable to proteins and nucleic acids, but not to lipids that do not have exchangeable hydrogens.

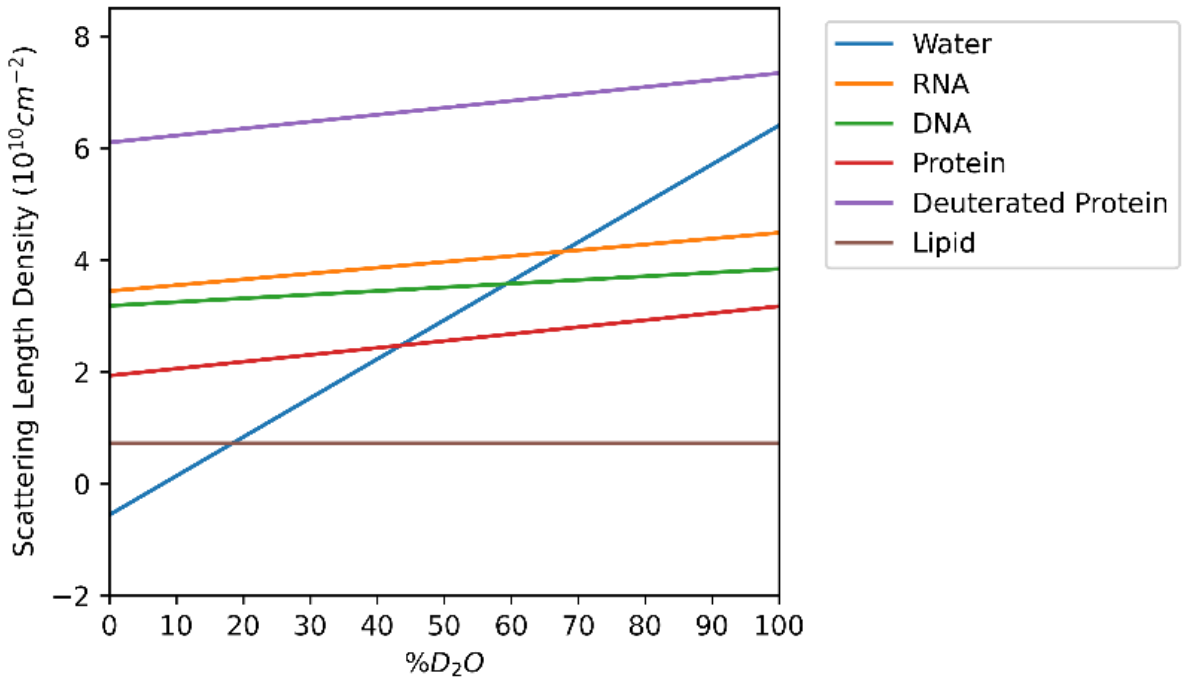


Figure 1.9: Average scattering length densities of examples of biomolecules as a function of solution content of D_2O. Water is exchanged for D_2O and at every intersection with a biomolecule line, the object is matched out which makes it indistinguishable from the solution. The slope of the object lines is caused by the exchange of labile hydrogens with deuterium from the solvent solution. Lipids generally do not contain exchangeable hydrogens and therefore present as a constant. The SLD of perdeuterated protein is higher than that of a 100% D_2O solution, which results in fully deuterated protein samples being unable to be matched out.

Another method to match a solution to a component of a complex is the use of deuteration. Experimental methods to achieve deuteration will be described in section 1.5. This approach is typically used to investigate protein-protein complexes. In this case, a protein can be matched to 100% D_2O solutions (132) by replacing around 72% of its non-labile hydrogens by deuterium atoms. In a complex of two components, one can isolate the scattering signal of a single component if the SLD of the two are sufficiently different, so that one can be measured, while the other is matched out. To increase the SLD difference between two partners, one partner could be deuterated, while keeping the other protonated. The complex can be measured in buffers with different D_2O contents to vary their respective contrasts. Importantly, the incoherent background scattering is reduced when measuring deuterated molecules due to the decreased hydrogen content of the solution. There are available programs that allow the structural modelling of SANS data measured at different contrast values (133). Note, however,

that this combined analysis assumes that the structure of the complex does not change due to deuteration, and that a (almost) perfect exchange of the labile atoms is achieved.

1.4.3 SEC-SAXS and SEC-SANS

In the case of aggregation-prone proteins, it can often be a significant challenge to have a monodisperse dilute sample. Aggregation of the sample will cause the structure factor to deviate from 1 and subsequently the capacity of interpretation of $I(Q)$ as a form factor would be incorrect. Size-exclusion chromatography has commonly been used to separate coexisting components of a sample before acquiring biophysical or functional data on pure species (119,134). By injecting the sample on a resin column, such as the Superdex® columns, the molecular species in the sample are separated by size. The separation range depends on the resin characteristics. In SEC-SAXS, the column is connected to a quartz capillary, which is exposed to the beam probe. The technique was implemented at synchrotron beamlines between the mid 2000's and 2010's, and automated systems were built to render this technique more widely available to users (119,135,136). BioSAXS beamlines are typically equipped with setups using automated sample injector and liquid chromatography (LC) pumps which are automatically coordinated with beam exposure. In order to follow the elution of protein from the SEC column, a UV-absorbance spectrophotometer measures sample absorbance at 280nm (or other wavelengths) before it reaches the capillary, which facilitates subsequent data reduction. Due to the short exposure time, SAXS data can be recorded during the elution of the column without modifying the flowrate.

SEC-SANS is less prominently used than the X-ray equivalent due to the reduced brilliance. Even so, the D22 beamline at ILL recently built a SEC-SANS sample environment (Figure 1.10) (123,137). The injection and the elution of a protein sample is the same as for X-ray studies, but due to the longer sampling time during SANS, the data collection is slightly different.

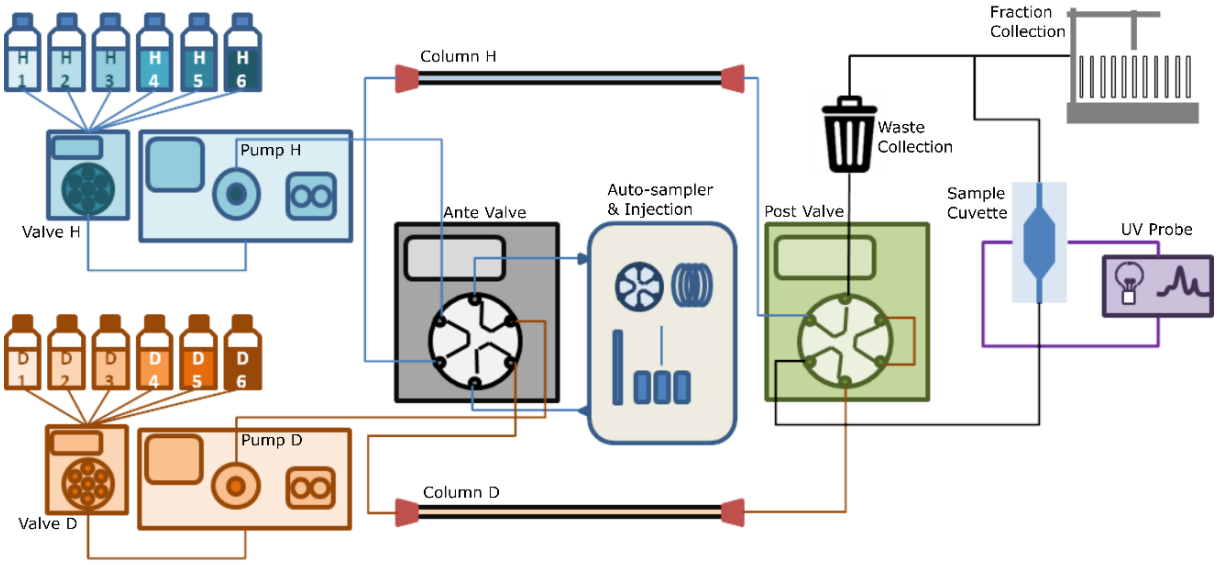


Figure 1.10: Schematic setup of the SEC-SANS sample environment at the D22 instrument. The SEC setup has two pumps and column slots, allowing for preparation of subsequent measurements while an experiment is running. The sample is led from the column into the quartz cuvette located in the beam and afterwards it can be collected by a fractionator (122).

The eluted protein is kept in a circular quartz cuvette, which is placed in the beam. UV-absorbance is measured simultaneously with SANS signal, directly in the quartz cuvette (Figure 1.11). Depending on the scattering intensity of the sample, the experiment can be run automatically or manually. The software can automatically slow down the flowrate (i.e. from 0.5 mL/min to 0.01 mL/min) of the LC pump once a specified increase in UV absorbance is observed and the sample is exposed in 30 seconds counts until the UV signal decreases below a given cut-off value. At this point, the pump comes back to the normal flow to finish the elution. In cases of low-scattering samples as occurs in my studies, the pump is manually shut-down once the protein peak is observed via the UV chromatogram. This allows longer exposures of the sample. Importantly, SANS does not cause sample degradation due to radiation damage, as it is often observed in synchrotron SAXS experiments (138). Therefore, in SEC-SANS experiments, samples can be recovered after the exposure and saved for other experiments, for instance using a buffer with different D_2O/H_2O composition.

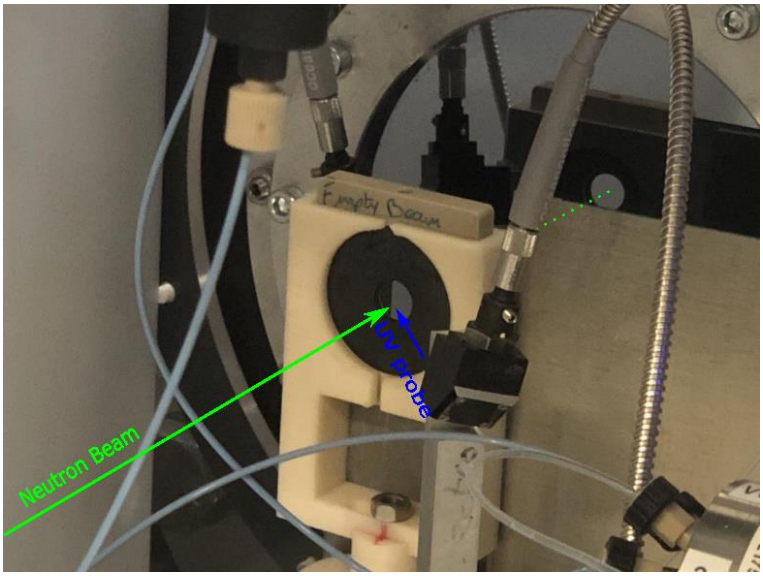


Figure 1.11: The D22 SEC-SANS sample cell. A circular quartz cell with a diameter of 13 mm and a volume of 150 µL is exposed to the neutron beam. The eluted sample is pumped into the bottom of the cuvette and flows out the top. A UV laser probe is positioned to the right of the cell with a UV detector to the left behind the cell, with a 45⁰ angle. This allows protein elution to be followed directly inside the cell during elution.

1.4.4 Model-Free Analysis of SAS data

Initial analysis of SAS data is usually done without using structural models (93,114,124,131,139). Such a model-free analysis refers to the information gathered from different profile representations and fits to simple models. The information extracted from this step includes the radius of gyration (R_g), the pairwise distance distribution ($p(r)$), the maximum intramolecular distance (D_{max}), the forward scattering intensity ($I(0)$) (containing information on the molecular weight), and a compactness estimation by Kratky plots (93,131,140). These parameters report on the size and shape of the sample in the absence of *a priori* structural knowledge, *e.g.* an atomic model.

The R_g is calculated from the Guinier approximation, which characterizes the mass distribution around the center of mass of a given structure (101,141).

Eq. 1.4.4.1 $$I_Q = I(0) \cdot \exp\left(-\frac{Q^2 \cdot R_g^2}{3}\right)$$

Under the assumption that the intensity could be represented as a straight line at low Q $\left(Q < \frac{1.3}{R_g}\right)$ in the Guinier plot (ln(I(Q)) vs. Q^2), the R_g-value can be extracted using the equation 1.4.4.1.

Besides the Guinier approximation, one can also utilize both Debye's approximation and the $p(r)$-distribution to extract the R_g from a given dataset (139,142,143). Regardless of the

approach used, two proteins of equal molecular weight could exhibit vastly different R_g-values depending on their structure (globular versus extended/rod-like protein structures).

The pairwise distance distribution (*p(r)*) describes the distribution of distances between any two given atoms of a protein, and is derived from a Fourier transformation of the scattering curve (Figure 1.12) (144). In addition to an alternative calculation of R_g, the *p(r)* enables the estimation of the maximum distance (D_{max}) within the protein, providing another insight into the extendedness of the particle. Importantly, the D_{max} can be difficult to determine for flexible systems, such as disordered proteins, meaning that structural interpretation has to be done carefully (144,145).

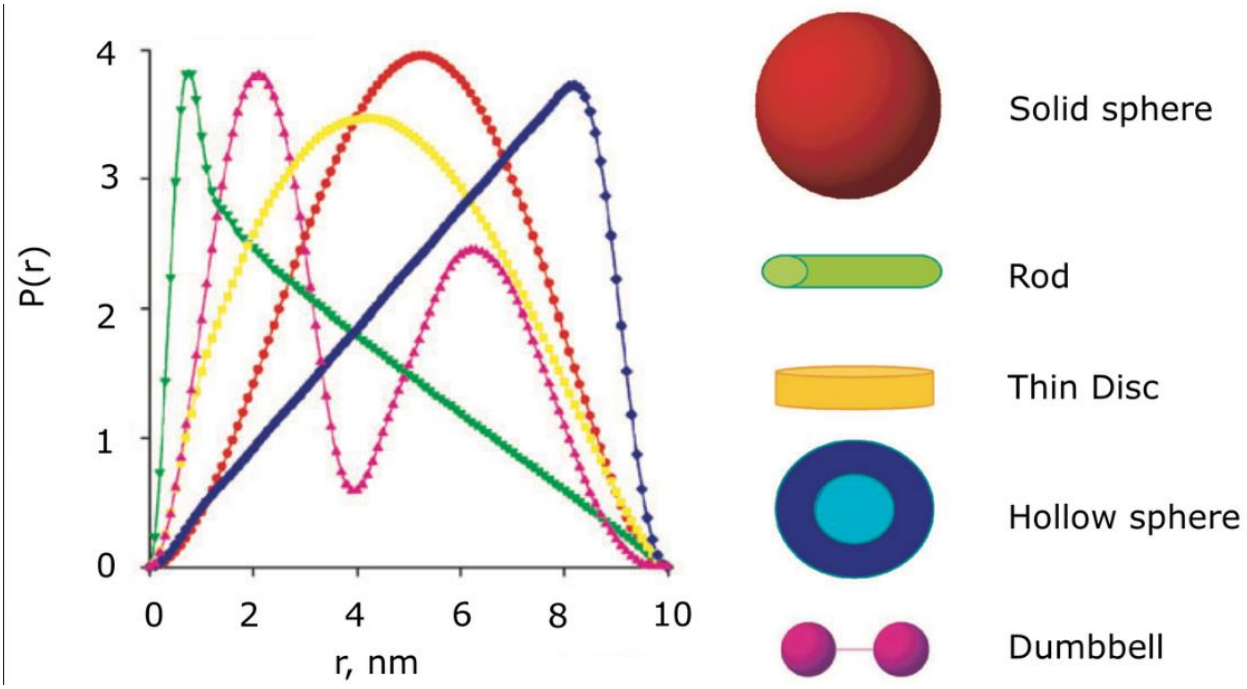


Figure 1.12: p(r)-distribution changes form depending on the overall shape of the sample examined. Figure extracted from https://bioxtas-raw.readthedocs.io/en/latest/_images/pr_shapes.png.

The forward scattering intensity, or intensity at $Q = 0$, *I(0)*, is an extrapolation of the scattering intensity expected at the 0-angle. However, in experimental setups, this value is unmeasurable: The 0-angle scattering intensity would be measured together with the much higher, non-scattered, direct beam intensity and may degrade the detector, which is why beam stops are implemented at SAS beamlines to block the beam intensity at the given angle (146). While the semi-transparent beamstop facilitates transmission detection, *I(0)* remains inaccessible, as the scattering contribution from the protein cannot be separated from the direct beam (147,148). Like R_g, the *I(0)* is also extrapolated using the Guinier fit (Eq. 1.4.4.1), the Debye law, or the *p(r)*-distribution with varying accuracy. Additionally, *I(0)* is sensitive to protein-protein

interactions and aggregation within the sample (149). The molecular weight of the sample can be calculated from the *I(0)* by:

Eq. 1.4.4.2 $$I(0) = \frac{cM_w}{N}[(\rho_P - \rho_S)v_P]^2$$

The equation incorporates the protein concentration (*c*), Avogadro's number (*N*), scattering length densities of the protein and solvent (ρ_P & ρ_S), and the partial specific volume of the protein (v_P) (149).

Because of its sensitivity to protein aggregation, *I(0)* is a useful value to calculate when initially evaluating experimental data, as this would be an indicator of the oligomeric state of the protein.

The Kratky plot is used to qualitatively differentiate between flexible and globular proteins. The Kratky plot is given by plotting $I(Q) \cdot Q^2$ against Q. A globular protein would yield a distinct bell-shaped profile, which returns toward 0, while a disordered protein would show a plateau followed by an increase at higher *Q*-values (139,149,150). Combinations of the two components yield varying shapes to the Kratky plot, allowing for a visual qualification of the compactness of a particle in solution from a given SAS dataset (139).

To better compare Kratky-plots of different samples, the dimensionless Kratky plot can be utilized (151). In this instance it would be calculated as

Eq. 1.4.4.3 $$x = QR_g$$

$$y = \frac{(QR_g)^2 \cdot I(Q)}{I(0)}$$

Figure 1.13 shows the dimensionless Kratky plot of the components and full construct employed in this project.

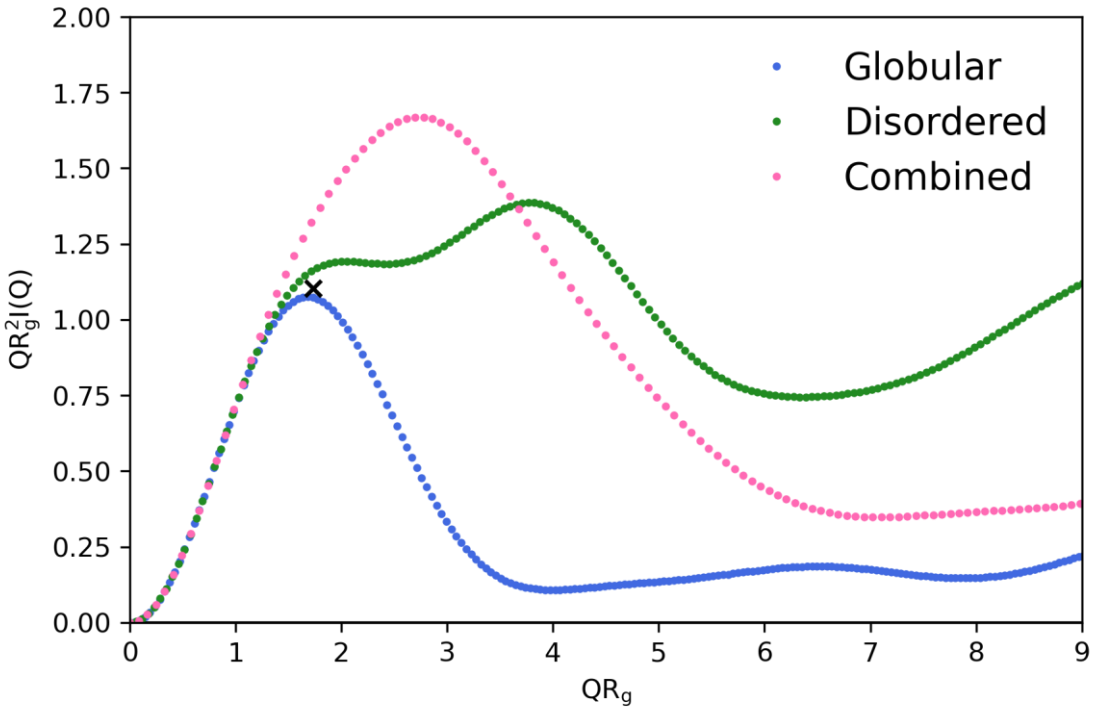


Figure.1.13:Dimensionless Kratky plot of fragments and full construct used in this project: The simulated sfGFP profile (blue) exhibits a clear globular signature with a bell-shaped peak close to that of the perfect globular protein (black cross at $\sqrt{3}$;1.104). The Huntingtin Exon1 region (green) reaches a plateau around QR_g = 2 and never returns to 0, a profile characteristic of disordered molecules (profile of a single conformation). The hH16 construct (pink) exhibits a combination of the two, with a broad bell-shaped peak that is not centered around $QR_g = \sqrt{3}$.

Fully globular proteins could be identified from the dimensionless Kratky plot by exhibiting a bell-shaped peak centered at $QR_g = \sqrt{3}$ with a maximum intensity of 1.104 (149,151). The dimensionless Kratky plot of disordered protein will often reach a plateau around $QR_g = 1.5 - 2$, and, depending on the degree of disorder, could continue to increase at larger QR_g (149).

1.4.5 Structural analysis of rigid particles in solution

The analysis of the SAS data in terms of structure can be done using different approaches, mainly depending on the amount of previous knowledge on the system of interest. These include, but are not limited to, *ab initio* and rigid-body modelling - where the protein is described by a single conformation in solution - as well as ensemble modelling approaches more adequate to describe flexible proteins.

Ab initio modelling is used to determine the low-resolution 3D structure of biomolecules in solution, when no previous structural information is available (152). The most popular and widely used *ab initio* modelling method rely on densely packed beads to represent a protein as a simple shape, meaning increased flexibility or extended structures will not be accurately

represented by this modelling approach. The final arrangement is optimized using either the scattering pattern or the *p(r)* (153–155). While several softwares have been developed, the most commonly used *ab initio* softwares are DAMMIN and DAMMIF of the ATSAS software package (155,156). Resolution of *ab initio* models can be assessed by the Fourier shell correlation function (157).

Rigid-body modelling incorporates the structure of known sub-units, for instance derived from X-ray crystallography, NMR, or Cryo-EM, to model larger structures or complexes (158–161). This approach has been used to optimize the relative position and orientation of domains within multi-domain proteins and to dock different biomolecules in macromolecular complexes. Even though some researchers have applied these strategies to describe flexible multi-domain structures, this should be avoided, as the validity of this type of modelling is restricted to rigid particles and would therefore not describe anything of significance when applied to flexible multi-domain structures (162).

1.4.6 Ensemble refinement of flexible proteins

In contrast to both *ab initio* and rigid-body modelling, ensemble approaches allow for analysis of flexible proteins, and has been the main structural method employed in the present project (139,149). In contrast to *ab initio* and rigid-body modelling, the ensemble approach does not result in a single structure solution, but rather a pool of structures that collectively describe the sample in solution. This is the best approach when investigating flexible proteins, such as intrinsically disordered proteins (IDPs) and regions (IDRs), which lack permanent secondary and tertiary structure and thus need a collection of structures to describe their behaviour in solution (163).

Ensembles of structures can be generated either from atomistic models or from coarse-grained models (CG), which represent groups of atoms (normally residues) with single beads (140,149,163–165). While atomistic structure ensembles might provide higher accuracy, the approximated CG models can speed up simulations, enabling broader conformational sampling. However, several these purely computational approaches cannot fully describe the conformational behaviour of flexible proteins in solutions. To address this limitation, several ensemble methodologies have been applied to refine and optimize these theoretical structural models using experimental data, such as NMR (166) and single molecule Förster Resonance Energy Transfer (smFRET) (167). SAS has also been a very common source of information used for this refinement (140,149,163).

SAS-driven structural refinement of ensembles requires exhaustive sampling of the conformational space probed by proteins, which can been done via several approaches, such as MD simulations (168), Flexible-Meccano (169), MoMa (84), TraDES (170) and IDPConformerGenerator (171). Once an ensemble of structures has been generated - either atomistic or CG models - the scattering patterns have to be calculated for each conformation in order to fit the experimental data (Figure 1.14). Multiple softwares have been produced since the 90's with the capacity to calculate a scattering pattern from an input structure file (pdb) with the most commonly referenced being Crysol (SAXS) and Cryson (SANS) from the ATSAS Software package (113,172–174). Alternatives to the ATSAS package include: FoXs (160), PepsiSAXS/PepsiSANS (118), and WAXSiS (175). The choice of software depends on the model resolution (Atomistic/CG) and the need or not to have an explicit description of the hydration layer of the protein.

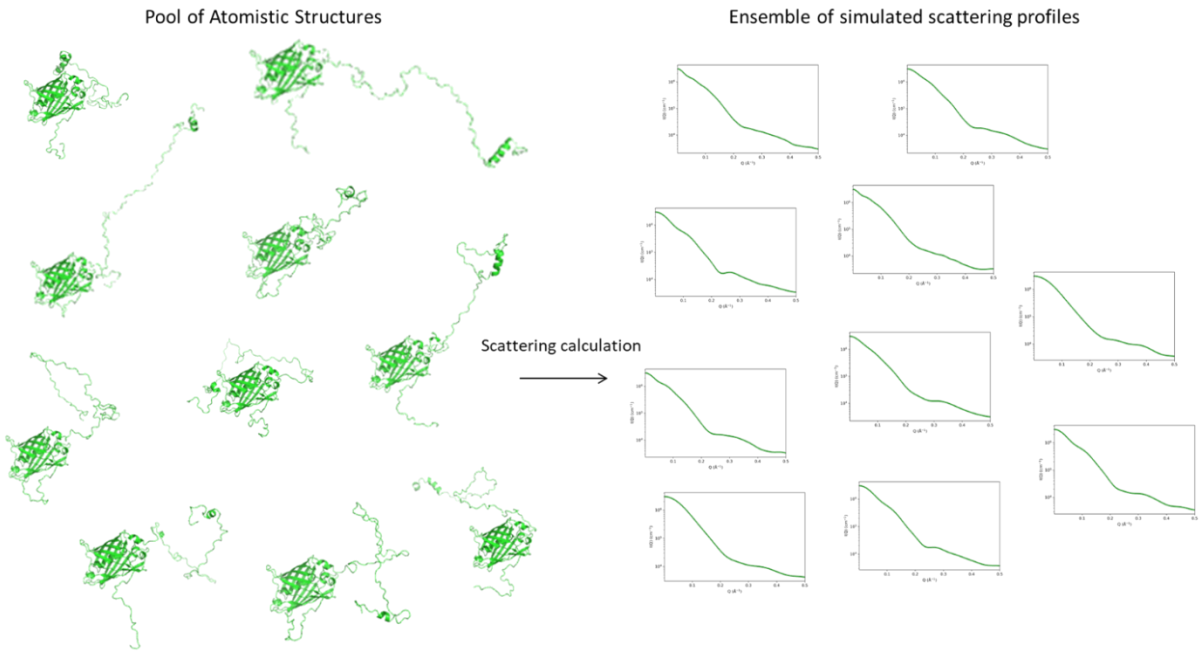


Figure 1.14: Atomistic structures of the target protein in a wide variation of conformations is used to calculate a theoretical ensemble of scattering profiles for ensemble refinement.

Although the impact of the hydration layer in the calculation of SAS properties of globular proteins has been investigated for different protein force field models (176), little is known about their effect for disordered chains. Normally, programs using a homogeneous hydration shell surrounding the protein are used to compute the SAS properties of disordered chains.

Ensemble refinement using SAS data has been used extensively to describe IDPs (177–181). SAS data allows the ensemble shape and size of an IDP/IDR to be characterized when combined with realistic computational models (165,182). In general, there are two major strategies to refine structural ensembles with SAS data. First: the maximum parsimony principle, which aims to describe experimental data using a very reduced number of conformations. Software based on this approach includes the ensemble optimization method (EOM), which refines a sub-ensemble of a given size (183,184), and the Minimum Ensemble Size (MES), which optimizes the sub-ensemble to the fewest possible structures, while still matching the scattering data (185). The second strategy is the maximum entropy principle. The maximum entropy principle is used to bias ensembles towards experimental data or re-weight the existing conformational ensembles and requires the initial structural ensemble to already fit the experimental data decently (186). Software based on this approach includes the Ensemble-Refinement of SAXS (EROS) (187) and the Bayesian Maximum Entropy (BME) (188).

The EOM program has been utilized as the main one for this project as it is very well suited to the simultaneous description of SAXS and SANS data (189). EOM is based on the maximum parsimony principle and employs a genetic algorithm to select a specified number of structures, called sub-ensembles, to collectively describe the experimental data (Figure 1.15).

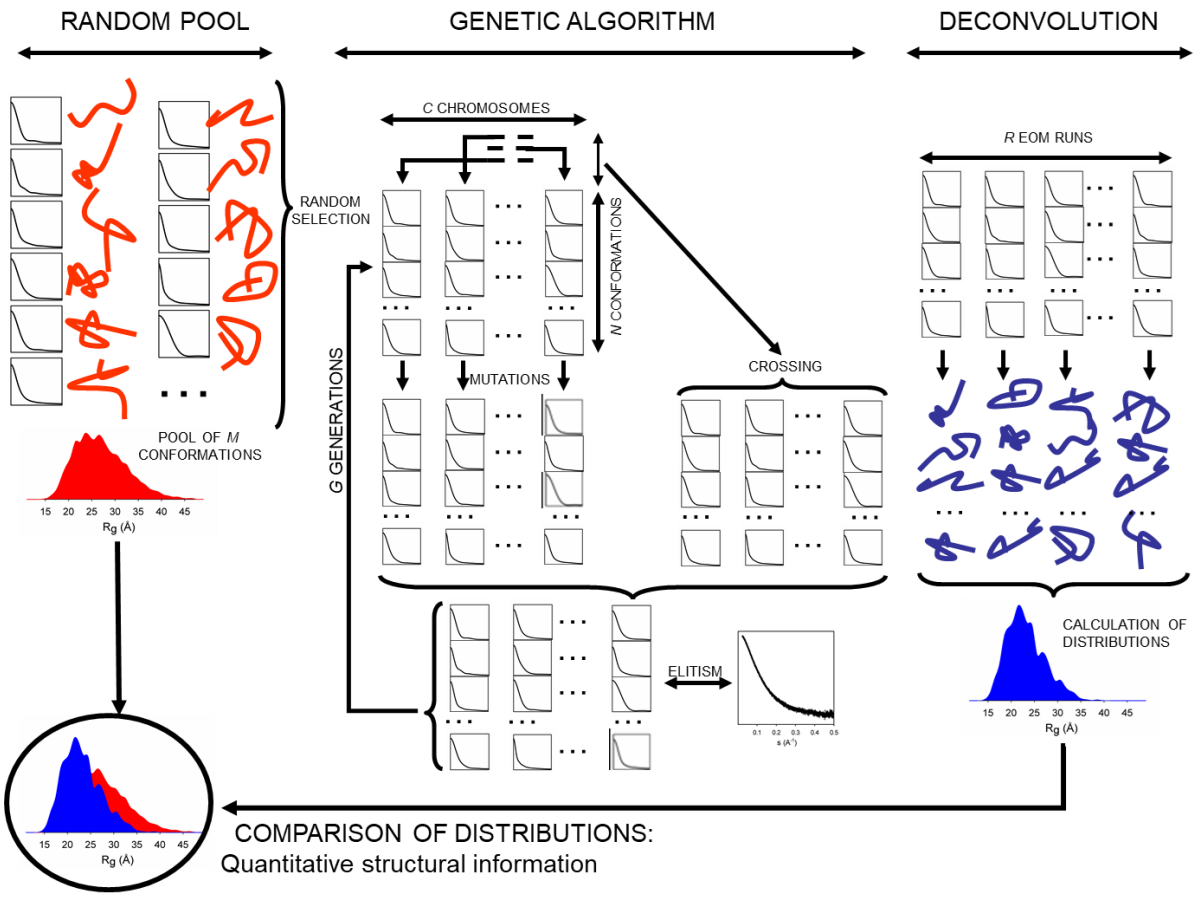


Figure 1.15: Overview of the EOM Genetic algorithm procedure extracted from Bernadó et al. 2011 (139). Schematic representation of the EOM strategy for the analysis of SAXS data in terms of Rg distributions. The M conformations/curves of the pool (random distribution), left part of the figure, are used to generate the initial C chromosomes and to feed the genetic operators (mutations, crossing and elitism) along the GA process that runs for G generations. The complete process is repeated R independent times, and each run provides N selected structures/curves that fit the experimental profile. The structural analysis of the resulting conformations is displayed on the right part of the scheme, the Rg distribution of the selected (N R) conformations is compared with that derived from the pool that is considered as a complete conformational freedom scenario. From this comparison it is possible to derive a quantitative structural estimation of the protein conformations coexisting in solution.

The genetic algorithm relies on picking sub-ensembles (chromosomes) of a pre-determined size (n-structures) from the theoretical ensemble and refining them through mutation (exchange of structures with the random pool) and crossing (exchange of structures between chromosomes). The average profile of each chromosome is calculated and fitted towards the experimental data (with a χ^2 metrics) with the best fitting sub-ensembles being saved and used

as starting points of the subsequent generation. This process is repeated for G generations and the chromosome (sub-ensemble) with the best fit is considered as the solution and it is stored. The same procedure is performed for R cycles (runs) and the best fitting chromosome of each cycle is collected. When putting together all the solutions, the resulting ensemble is considered as the structural description of the protein in solution and can be further analyzed. In the present work, sub-ensembles of 50 structures were averaged to account for the flexible nature of the Huntingtin exon1. The R_g values of the selected sub-ensemble (R_g-distribution) can be plotted against the overall distribution of the random pool of structures to visualize structural specificities of the flexible protein.

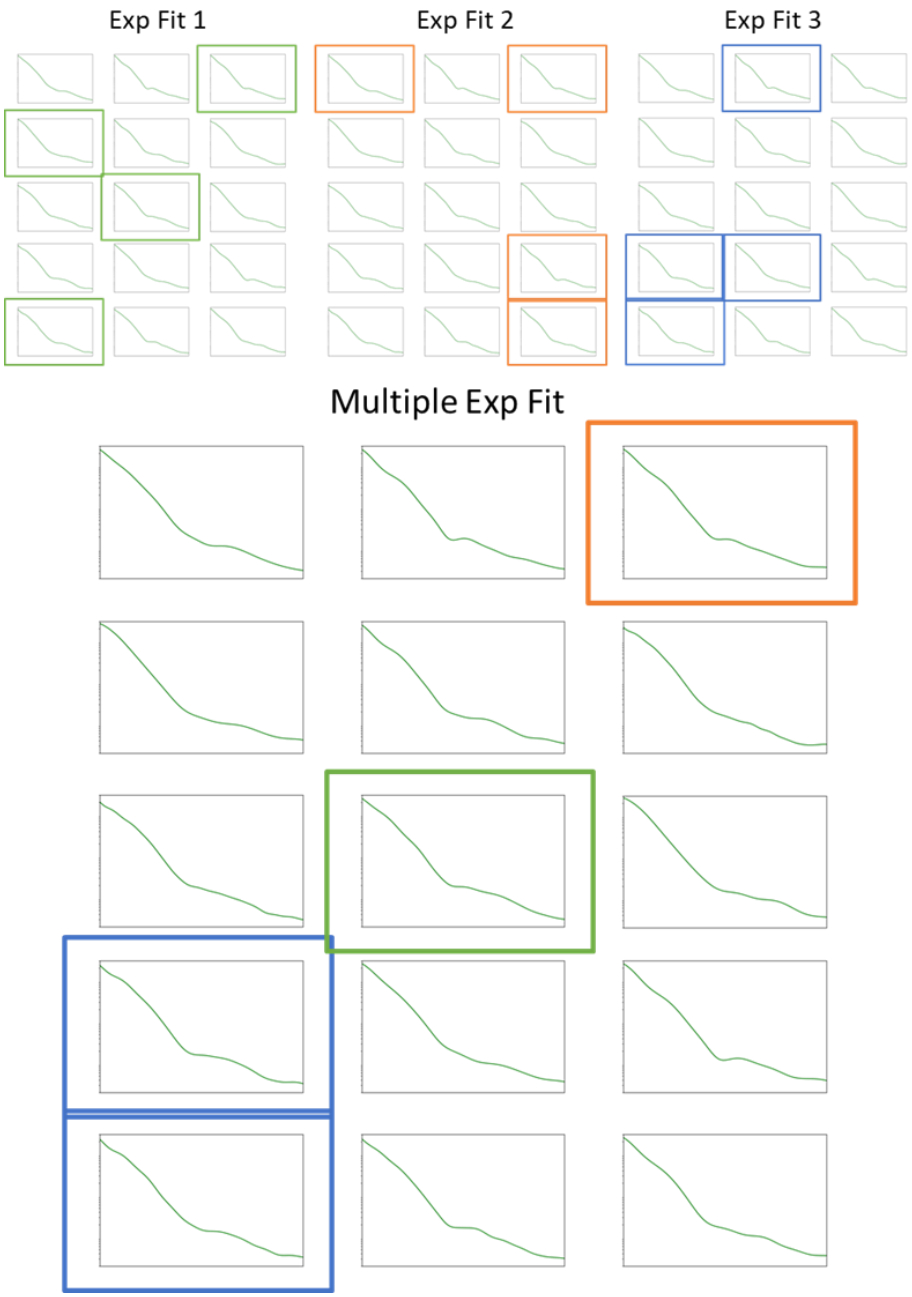


Figure 1.16: Visualization of ensemble fitting to experimental data. Fitting three different experimental datasets will result in three different sub-ensembles (Top, Exp Fit 1-3). If the three datasets are fitted simultaneously, the sub-ensemble will incorporate the information of each experimental dataset into one combined fit (Bottom, Multiple Exp Fit).

The advantage of implementing EOM in the present project was the software's ability to simultaneously analyse SAXS and SANS (189,190). Fitting multiple datasets simultaneously using EOM requires the average profile of each chromosome to be fitted to each of the experimental datasets included in the analysis (Figure 1.16). The combination of experimental data would increase the structural constraints of the fit and thereby decrease the number of degrees of freedom. Although the χ^2-values of a multi-dataset fit would increase, the integration of additional experimental data improves the overall information content and robustness of the

SAS analysis. The ability to simultaneously fit multiple dataset makes EOM ideal for data analysis of segmentally labelled samples, due to the ability to combine the data of several labelling patterns. Simultaneous analysis of multiple scattering datasets enables a more refined characterization of atomistic sub-ensembles than is achievable with individual experiments. As a consequence, the sub-ensemble chosen through the multiple dataset fitting provides a more robust description of the protein's conformational flexibility and disordered nature.

1.5 DEUTERATION

Deuteration is a powerful tool in neutron scattering to improve, highlight, or alter the structural and dynamic information gathered from an experiment. Biological samples contain a high amount of hydrogen. For instance, in H16 49.4% of atoms are hydrogens and, importantly, 20.8% of them are labile and can be exchanged by deuterium. Due to the above-mentioned difference in SLD, the exchange between hydrogen (H^1) and deuterium (H^2) in these labile positions has a major impact on the measured SLD of samples.

Table 1.2: Coherent and Incoherent scattering lengths of atoms commonly present in biological samples

Atom	Atomic Number	Coherent Neutron scattering length (10^{-12} cm)	Incoherent Neutron scattering length (10^{-12} cm)	X-ray scattering length (10^{-12} cm)
Hydrogen	1	-0.374	2.527	0.28
Deuterium	1	0.667	0.404	0.28
Carbon	6	0.665	0.000	1.69
Nitrogen	7	0.940	0.202	1.97
Oxygen	8	0.580	0.000	2.16

The difference in scattering length of hydrogen (H^1) and deuterium (H^2) (see Table 1.2) has allowed researchers to obtain detailed structural information for complicated systems. Consequently, deuteration has been widely used by the bio-structural scientific community. In neutron crystallography, deuterated protein crystals have been diffracted, allowing researchers to glean information that is not typically available from traditional X-ray crystallography experiments, such as the placement of hydrogen atoms, hydrogen bonds, and the identification of water molecules in binding sites (191,192). Deuterated samples have also allowed protein complexes to be further investigated using SANS, *i.e.* by using the mentioned contrast match effect to isolate the SLD of individual components of protein-protein complexes or larger multi-subunit nucleoprotein complexes, such as ribosomes (193–195).

When working with biological samples, deuteration can be done either by chemical synthesis or by biosynthesis (196). While chemical synthesis of peptides is an effective way to produce

pure deuterated peptide samples, the use of solution-based or solid-phase peptide synthesis is expensive and has generally been confined to short peptides, up to 30 residues (197–199). Protein biosynthesis, either *in vivo* or *in vitro*, is a more general approach to produce tailored deuterated samples.

In addition to in-house production, several neutron sources, including ILL (France), ANSTO (Australia), ISIS (England), ESS (Sweden) etc., have deuteration facilities connected to their scientific campuses, which, upon application acceptance, can assist in the production of deuterated biomolecular samples. The deuteration methods available depend on the laboratory. A collaboration between the different laboratories has been recently started with the aim of exchanging expertise in deuterated sample production. This project, called DeuNet (https://deuteration.org/), aims at promoting collaboration, development, and visibility of deuteration research.

The achievable deuteration level can differ depending on the procedure employed. Although the most classical approaches produce homogeneously deuterated samples, when the whole molecule presents the same degree of deuteration, more tailored strategies can produce segmentally and specifically deuterated molecules. The more sophisticated samples will allow specialised experiments probing specific structural information. In the following sections, the methodologies applied to produce these three deuteration patterns will be presented.

1.5.1 Homogeneous Deuteration

In vivo deuteration refers to the expression of deuterated proteins in living cells. It has been shown that several microorganisms are capable of growing in deuterated media, allowing deuteration to various levels (196,200,201). The choice of the expression strain is highly dependent on the characteristics of the target protein, but in addition to expression attributes, the planned deuteration method can also impact which strain should be chosen. *Escherichia coli* cells can be readily adapted to growth in both deuterated minimal and rich media. For these cases, deuterated media contain deuterated carbon sources, allowing for the expression of perdeuterated proteins. Strains of yeast, as well as insect and mammalian cells have also been adapted to deuterated media, allowing samples which require folding chaperones or post-translational modifications to be deuterated (196).

A perdeuterated protein is not always useful however and research has been conducted to improve the control of deuteration processes. Several strains have been developed to introduce specific deuteration levels and protocols have been developed for both *E. coli* and *Pichia*

pastoris (yeast). These protocols aim at producing partially deuterated protein samples at approximately 72% by growing the cultures in a minimal media dissolved 15%:85% H_2O:D_2O (202). With this level of deuteration, the SLD of the expressed protein matches the SLD of 100% D_2O solvents during SANS experiments, allowing contrast matching. When 72%-deuterated and protonated proteins are combined in complexes, the labelled protein can be masked by the 100% D_2O solvent, providing direct structural information of the non-deuterated component (see Figure 1.17). This allows neutron experiments to examine a single unit of a multi-protein complex, where one unit of the complex is masked by the solvent (203,204). Equivalently, following Figure 1.17, the match out of specific components of a complex can be achieved simply by optimizing the percentage of D_2O of the buffer. For instance, one can match out the average scattering signal of DNA at about 60% D_2O. Another advantage of mixing hydrogenated and 72% deuterated proteins in 100% D_2O solutions is the significant increase of the signal-to-noise ratio. This is due to the much lower number of hydrogen atoms in the solution, which in turn lowers the incoherent scattering. The incoherent scattering primarily contributes to the incoherent background noise in SANS experiments (127).

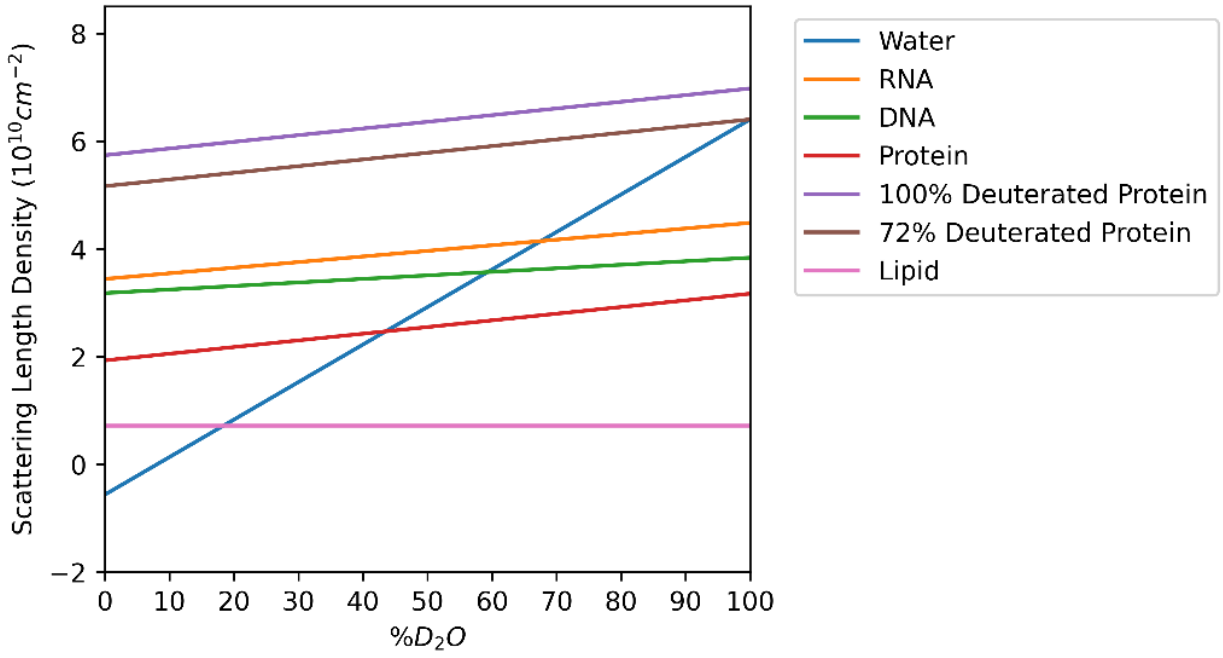


Figure 1.17: Average scattering length densities of examples of biomolecules as a function of solution content of D_2O. Water is exchanged for D_2O and at every intersection with a biomolecule line, the object is matched out, which makes it indistinguishable from the solution. Of special significance here, the change from 100% deuterated protein to 72% deuterated protein allows the sample to be matched out at 100% D_2O. The match-point of a protein is dependent on the specific amino acid sequence and so varies from protein to protein. The lipid used is phosphorylcholine and all values used were extracted from Jacrot et al. 1976 (129).

1.5.2 Segmental Deuteration

Segmental deuteration is the process of labelling specific fragments within a protein. The method is normally applicable in multi-domain proteins where the behaviour of the individual domains can be investigated in the context of the full-length protein. Segmentally labelled samples are generally produced by protein ligation of the different domains after expression. Different strategies have been developed to produce suitable precursors and for the protein ligation. Historically, the most popular approach to ligate protein fragments is the use of inteins, which have been reviewed in David et al. 2004 (205). Inteins can be described as parasitic genetic fragments, which are excised from a protein by the intermolecular process of protein splicing, followed by the ligation of the flaking regions of the protein (205).

This self-splicing behaviour is similar to that of exons and introns during RNA transcription. The ability of inteins to link flanking protein regions has been used to covalently bind unlabelled and labelled protein fragments in order to create protein samples with specific segmental labelling patterns (205,206). While there are several specific methods of obtaining segmental labelling by native ligation or intein-induced ligation (figure 1.18), the resulting sample will be the product of fusing the N- and the C-terminal protein fragments (205,206)

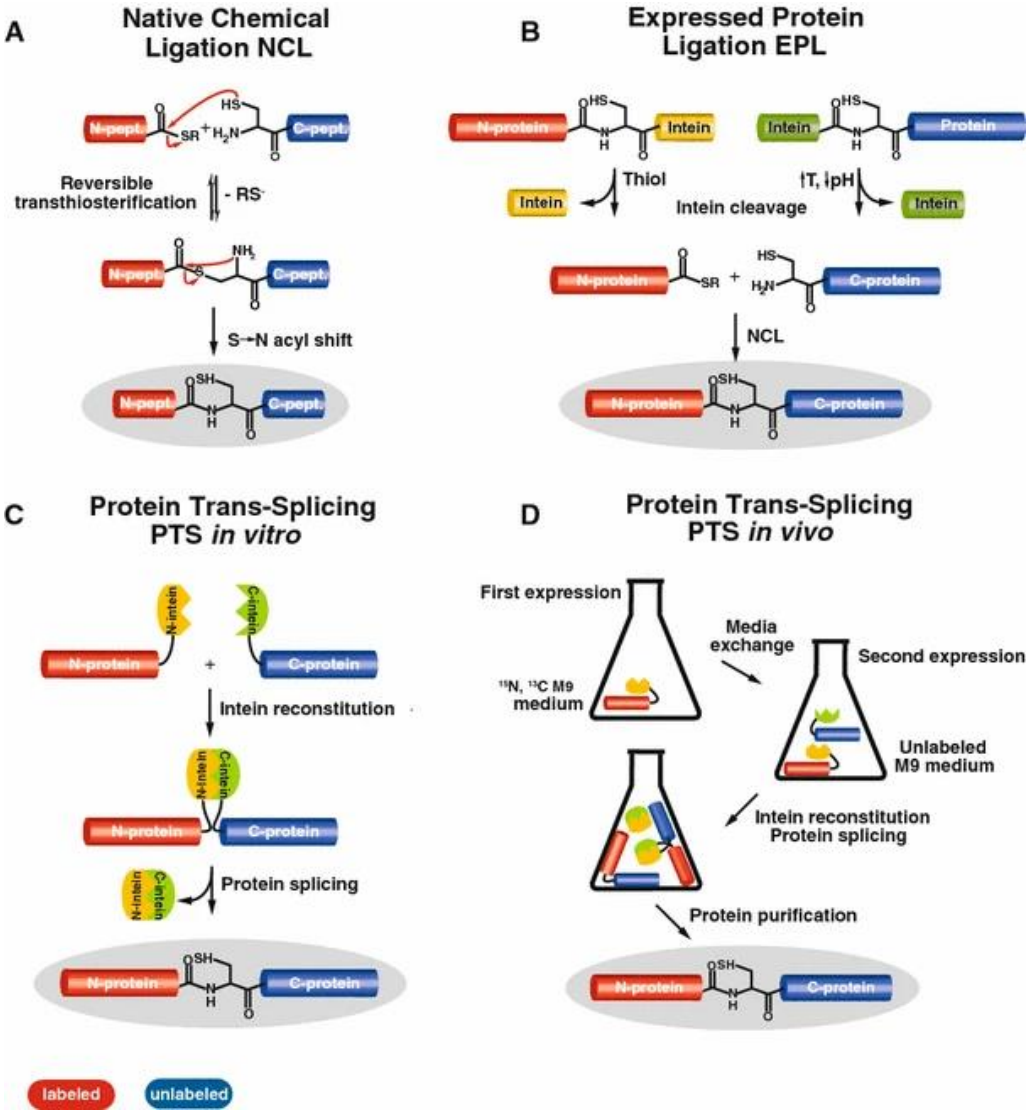


Figure 1.18: Overview of the methods for segmental isotope labelling of proteins. ***a*** *Mechanism of the Native Chemical Ligation.* ***b*** *Principle of Expressed Protein Ligation.* ***c*** *Principle of Protein Trans-Splicing by split inteins* ***d*** *Schematic representation of the protocol for in vivo Protein Trans-Splicing. Figure extracted from: Skrisovska L, Schubert M, Allain FHT. Recent advances in segmental isotope labelling of proteins: NMR applications to large proteins and glycoproteins. J Biomol NMR. 2010 Jan 1;46(1):51–65.*

A more recent method of segmental labelling is Sortase A mediated ligation. This method uses the enzyme Sortase A, which will ligate protein fragments if the Sortase A recognition motif is present (206–208). This method has been shown to produce high-quality samples for SANS measurements in regards to both purity and yield (209). For instance, in Sonntag et al., a multi-domain protein was segmentally perdeuterated allowing SANS contrast matching experiments to yield insight into the behaviour of the individual domains of the RNA binding protein TIA-1 (RBP TIA-1) (209).

1.5.3 Residue Specific Deuteration

Further structural selectivity can be obtained by specific amino acid deuteration. This can be achieved by either adding deuterated amino acids to a hydrogenated minimal medium or, by reverse labelling, incorporating hydrogenated amino acids in a deuterated minimal medium. Importantly, the addition of specific deuterated amino acids to *in vivo* cultures can result in endogenous amino acid biosynthesis and residue scrambling (210). Auxotrophic cell-strains can be used to specifically deuterate single amino acids which would otherwise not be possible using *in vivo* expression. For instance, *E. coli* strain B834(DE3) can be used to specifically incorporate deuterated methionine residues in proteins (210–212). Auxotrophic labelling requires the use of large amounts of either deuterated precursors or deuterated amino acids, which makes this production expensive (210). In addition to the *in vivo* expression methods, the open system of cell-free (CF) expression allows *in vitro* expression of protein samples with residue specific deuteration. Similar to auxotrophic cell-lines, CF residue deuteration can be controlled by the addition of the selected deuterated amino acids to the expression mixture. The specific advantage to deuteration here being that the labelling is not limited by the choice of strain, resulting in the possibility of simultaneously labelling multiple amino acids in a single expression. Moreover, as mentioned in section 1.6, in CF expression amino acid scrambling processes are notably reduced.

Residue specific labelling enables neutron experiments the ability to increase the signal of residues in key locations throughout proteins (213). However, the scattering increase produced by the specific incorporation of deuterated amino acids is very limited (210). That being said, this concept can be theoretically applied to proteins with strong compositional bias, such as the LCRs, where multiple deuterated residues can be grouped together to increase the scattering of a specific location. For this family of proteins, one can match out the non-labelled part of the protein or increase/decrease the SLD derived from the deuterated region in order to yield additional structural information. The work presented along this manuscript has used CF expression to produce LCR protein samples with residue specific labelling. The procedures and advantages of this protein production method are described in the following section.

1.6 CELL-FREE PROTEIN SYNTHESIS

Cell-free (CF) protein expression is an *in vitro* technique, which uses the transcription-translation machineries purified from cell extracts to produce recombinant proteins (214,215). Cell extracts used along this work have been derived from *E. coli*, but techniques for obtaining

extracts from other sources, such as insect cells, rabbit reticulocytes or wheat germ, have also been developed (216,217). The first utilization of cell extracts dates back to 1897, when yeast extract was employed to convert sugar into ethanol (218). The use of CF has since played an important role in the production of both DNA and proteins (219,220). Conventional *in vivo* expression remains the dominant method for protein production, but with recent advances and the availability of commercial kits, CF has become a very useful technique in multiple fields, including structural and synthetic biology (214,221,222).

Contrary to *in vivo* expression, CF is an open system that allows direct synthetic manipulation of the expression of proteins and these modifications include the introduction of chaperones (223), nano-disks (224), unnatural amino acids (92,225,226), or crowding agents (227). Another advantage of CF is that proteins which have toxic properties or are prone to aggregation, can be produced with fewer problems (214). Similarly, proteins that would be expressed in inclusion bodies, can be produced without the need of refolding (228). In the context of this thesis, the biggest advantage of CF is the ability to control the amino acid composition in order to obtain specifically labelled protein samples. This feature has been employed in NMR projects with the purpose of gleaning structural information from proteins, which would otherwise be difficult to analyse (91,92,225,229–232). Introducing unnatural amino acids at specific positions enables the structural analysis to be guided towards the domain of interest within the protein (233–235). This possibility has been recently extended to soluble expression of membrane proteins (236). A practical advantage of CF is the short reaction time and scalability of the expression mixture, enabling everything from screening of plasmids to sample production and industrial expression (214,237,238).

1.6.1 *E. coli* Cell Free Systems

While various lysate sources are available, this description will exclusively concentrate on the utilization of *E. coli*-based lysates, as they were the ones employed in this thesis.

CF protein expression derived from *E. coli* can be divided into crude extract CF systems (239,240), purified systems known as the synthetic enzymatic pathways (SEP) (241,242) and the protein synthesis using recombinant elements (PURE) systems (243). The SEP system is a fully synthetic approach where the enzymatic machinery of the prokaryotic cell is assembled from individually purified components (222,241). In contrast to the crude extract, the PURE technology allows complete control of the reaction, but also significantly increases the cost of the production steps due to the need for component purification prior to CF expression (244).

As a consequence of this issue with cost of production steps, the use of crude extract systems is more common and has also become commercially available, with the capacity to be scaled up to industrial expression volumes (237,245).

Crude extract *E. coli* CF requires several important components during expression. The most important ones are:

- **Lysate:** Cell extract obtained from the *E. coli* strain of choice. Lysate can be either a crude extract or a purified system, but importantly contains the translational machinery allowing protein synthesis. The lysate used in this project was S30 extracts of BL21 Star (DE3): RF1-CBD3 cells. This strain allows the extraction of the released factor 1 (RF1), which was necessary for the introduction of isotopically labelled amino acids in a site-specific manner in proteins using the tRNA suppression technology (65,92,225). Note that for the purposes of this thesis, the extraction of the RF1 was not necessary, but it was systematically done as this lysate was used by multiple members of the team.
- **Plasmid**: Plasmids used in protein production are genetic vectors that encode the target protein. Plasmids can exhibit significant variability based on factors such as organism origin, antibiotic resistance, and expression promoter sequence (74,246). The origin and promoter are tied to the cell-line chosen, while the resistance is used as a marker during protein production to verify plasmid integration (74). Genes encoding Htt with a poly-Q of 16 and 36 repeats were cloned into a pIVEX 2.3d vector during this project. pIVEX vectors are very often used in CF expression as they are high-copy vectors and they can be produced in large quantities. Note that CF production requires relatively high amounts of purified plasmid to enhance yields.
- **Polymerase:** To initialize the protein transcription, the presence of RNA polymerase in the CF is required. The polymerase should specifically bind the promoter sequence within the plasmid and ideally exhibit high selectivity towards this sequence. T7 RNA polymerase has been commonly used in *E. coli* expression for decades due to its selectivity and translation efficiency (247). The polymerase can be added to the expression mixture or, as it is the case for our laboratory, expressed by inducing the cell culture using IPTG, before harvesting and lysate production (215).
- **Precursors**: Precursors are small molecules added to the CF reaction that are needed for the synthesis of mRNA and the expression of the recombinant protein. These precursors include ribonucleotide tri-phosphates (rNTOPs) and amino acids.

- **Energy regeneration system**: Energy, in the form of ATP, is essential for all the steps of the protein synthesis (248). Therefore, a system that regenerates the ATP during the expression is needed to extend the reaction time and increase the protein yield (249). Energy sources can be divided into three major classes: *(i)* Traditional high-energy phosphate bond substrates, *(ii)* multi-step enzymatic pathways and *(iii)* full energy regeneration system (249). The yield of the CF expression is correlated to the stability of the energy system and during this project the creatine phosphate (CP) / creatine kinase (CK) system (class *i*) has been used. This energy regeneration of CK has been shown to improve the protein yield compared to other traditional phosphate bond substrates (230).

In addition to these major components, other chemicals are added to the CF reaction mixture, including co-factors (folinic acid and magnesium acetate), reducing agents (dithiothreitol-DTT and protease inhibitors), and buffer compounds. The procedure followed during this project used the following buffer composition: Hepes, ammonium acetate (NH_4OAc), and potassium glutamate (KGlu) / potassium acetate (KOAc). Some components are not essential to the reaction, but they have been reported to enhance the yield, such as the addition of tRNA (250). The optimal concentrations of some components are subject to calibration and can vary for each lysate production. In our group the concentration of magnesium acetate and KGlu/KOAc are optimized for each lysate and protein produced (215). CF expression has undergone many optimizations including the use of creatine phosphate / creatine kinase as an energy regeneration system and reducing the lysate volume using PEG-solutions by dialysis, among others (230). The specific method and component concentrations are described in material and methods (chapter 9.4).

CF expression systems are divided into two major families. Continuous flow CF and batch CF. While not used in this project, continuous flow will be briefly explained, followed by a more detailed description of batch CF, which has been the one used in this project.

1.6.2 Continuous Flow Cell Free

The equipment used for continuous flow CF employs two reaction chambers; an inner and an outer chamber, separated by a membrane. The inner reaction chamber contains the cell-extract and the DNA template, while the outer chamber, which contains around ten times more volume, supplies the inner chamber with a sustained flow of reactants, such as salts, nucleotides, ATP, and amino acids (251,252). The expressed protein is generally kept in the inner chamber of the

reaction setup, but in certain cases it can be removed from the reaction chamber by diffusion over the membrane and can be purified from the substrate solution (253). This method allows prolonged reaction times (10-40 hours) compared to batch CF, significantly increasing the final production yield. The drawback is the larger amount of certain chemicals, such as amino acids and nucleotides, that must be employed.

1.6.3 Batch Cell Free

The conventional batch CF method typically uses smaller volumes and, as a consequence, it is more suited for isotopic labelling that uses expensive labelling reagents, such as isotopically labelled amino acids (254). While the protein yield is lower than conventional *in vivo* expression and the continuous flow CF method, the low reaction volume and complete reaction control of the CF batch expression is well suited for samples with expensive chemicals. The cost of lysate production has decreased over time and combined with the optimization of the method, this has led to commercially available CF kits, using *E. coli* lysates (255).

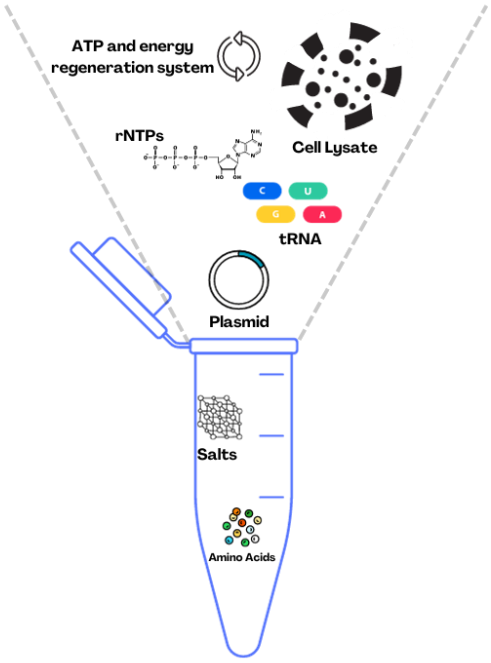


Figure 1.19: Cell-free expression is performed directly in the solution, by mixing the reagents and incubating the mixture. Illustration inspired by Hong, Kwon, Jewett 2014.

The batch CF can be run for screening, optimization, and production. As such, the reaction can be prepared in different reservoirs from well plates to test tubes depending on the reaction volume (Figure 1.19). The temperature is typically kept stable during the 2-4 hour expression (215). Incubation time and temperature are important factors of CF expression, in which the optimal condition is dependent on the specific expression mixture and target protein.

One drawback associated with CF expression is a considerable variance in protein yield, which might lead to low reproducibility between experimental replicates. Research indicates that a meticulous optimization of the CF protocol can mitigate some of this variance, making the method easier to use even for novice operators. Notable improvements include pre-solubilizing reagents and thorough mixing of the CF mixture prior to incubation (256). In this project, the batch method has been used to express most protein samples measured, including those labelled with deuterated glutamine, glutamic acid, and proline residues.

1.6.4 Amino acid scrambling

The cellular enzymatic machinery can produce several of the amino acids through well characterized pathways, such as the citric acid cycle, and some amino acids can act as precursors of others (248). For example, glutamate can act as a pre-cursor for glutamine, proline, and arginine or serine, which can be transformed into cysteine and glycine (248). During lysate production most of the enzymes are removed or their activity is compromised, considerably reducing amino acid scrambling and affording CF reactions with much higher control of the amino acid incorporation. Nonetheless, some of scrambling-related enzymes, such as the glutamine and asparagine transaminases, are retained in *E. coli* lysates (229,257). These enzymes lead to the conversion between glutamine and glutamic acid, as well as asparagine and aspartic acid. In this context, it is important to note that potassium glutamate (KGlu) is used as an osmolyte / buffer compound in the expression mixture. Therefore, along with the CF reaction, glutamic acid and glutamine (through glutamine synthetase) originating from the buffer are incorporated into the protein. Although this scrambling is not a problem if the aim is the production of an unlabelled protein, it has a negative effect when the aim is the production of an isotopically labelled protein.

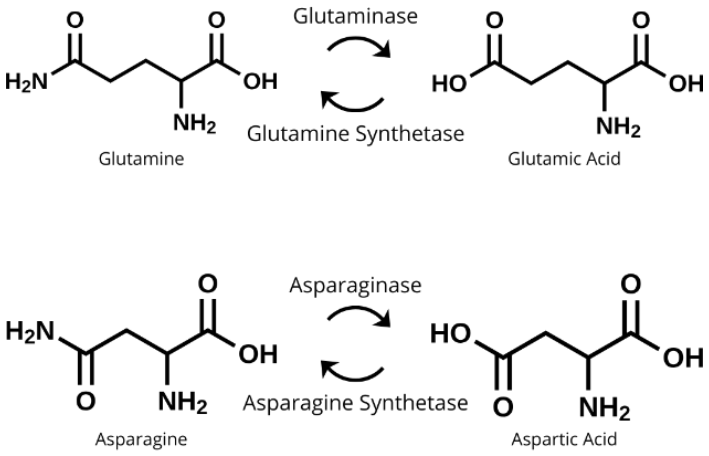


Figure 1.20: Enzymatic scrambling of Gln, Glu, Asn and Asp. The enzymes are present in E. coli lysates, which makes labelling these residues more difficult than other amino acids.

Several solutions to this problem have been proposed. It has been shown that exchanging the buffer with potassium acetate (KOAc) will allow for the isotopic labelling of Gln/Glu residues (229,258). Alternatively, it has been proposed to exchange the natural L-form of KGlu buffer with the D-form. This change shows a higher expression yield than that obtained when using KOAc buffer, however, labelling Gln/Glu residues in the presence of D-form KGlu has been reported to show mixed results, suggesting that complete labelling has to be confirmed

(258,259). Inhibitors have also been investigated for their anti-scrambling effect with reagents such as amino-oxyacetate and L-methionine S-sulfoximine (260–262). Different lysate strains have also been developed in an attempt to circumvent the scrambling effect and, while significant improvements were observed, an important result showed that even in strains with the suspected scrambling enzymes eliminated, the CF expression could still produce proteins without a specific amino acid added during the mixture (263). This suggests that even if the scrambling effect is blocked, CF lysates can retain the ability of endogenous amino acid biosynthesis of precursors. In the context of the present project, where we aimed at specifically deuterating Gln, reducing scrambling processes has been very important (described below in chapter 3).

2 OBJECTIVES

The thesis project had the following objectives:

1) Developing a robust strategy to produce amino-acid specific deuterated proteins for its general application to the SANS measurement of low-complexity proteins

2) Developing a robust computational approach to optimally integrate SAXS profiles and SANS data on amino-acid specifically deuterated protein samples to derive realistic ensemble models of disordered proteins

3) Characterizing the structural features of Htt Exon-1 and evaluation of the overall structural perturbations exerted when increasing the size of the poly-Q tract beyond the pathological threshold

3 SEGMENTAL LABELLING

The LCR (Poly-Q and PRR) of Htt provides an advantageous target for segmental labelling (deuteration), focusing on the structural analysis of the biomedically relevant Poly-Q region. Taking the Htt architecture into consideration, eight different labelling patterns can be designed (Figure 3.1).

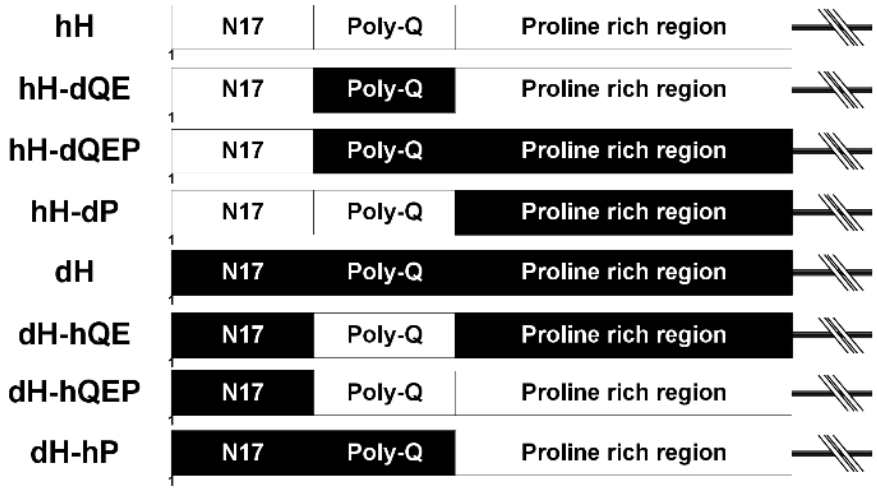


Figure 3.1: Labelling patterns of Htt. By specifically labelling the Q, P or remaining residues, eight deuteration patterns can be produced. Note that deuteration is indicated in black. The same nomenclature will be used along this thesis.

It was hypothesized that the SANS profiles would vary between the different labelling patterns and that this variation would allow an additional level of information to be gleaned from the SANS measurements, which was subsequently demonstrated with computational data (chapter 4). While fully protonated (hH) and deuterated (dH) samples are unlikely to provide additional information, compared to SAXS measurements, they provide an important complementary information to our SANS measurements of segmentally labelled samples.

The construct naming convention combined four elements: 1) The deuteration state of the main protein, 2) the Poly-Q length of the construct, 3) the specific labelling, and 4) the level of D_2O of the solution. As an example: hH16-dQE-80 signifies protonated Htt Exon1 construct with 16 glutamines, labelled with deuterated glutamine and glutamic acid amino acids, and calculated or measured for a solution with 80% of D_2O (Figure 3.2).

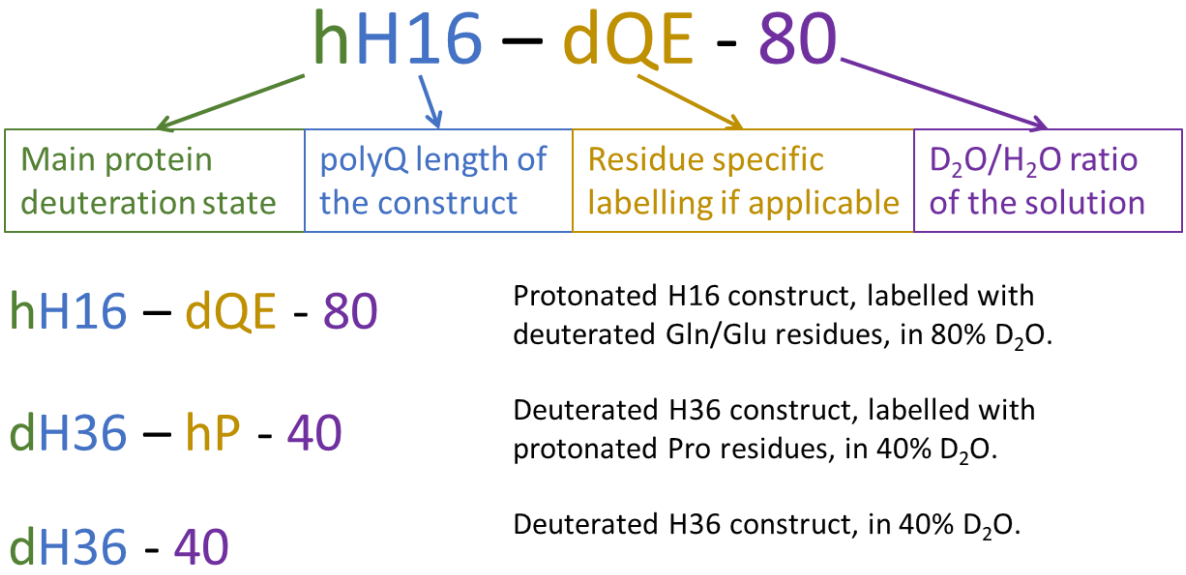


Figure 3.2: Naming convention of samples. The four parts are deuteration/protonation of main protein, Poly-Q length of the construct, amino acid specific deuterium/hydrogen labelling if applicable, and the D_2O/H_2O ratio of the solution.

By deuterating the amino acids, the SLD of the segment and that of the overall protein are modified and they will be further impacted by the variation of the solution level of D_2O. It was hypothesized that the structural features of the Poly-Q could be investigated by combining SAXS measurements with SANS data of segmentally labelled samples of Htt. This approach requires developing a novel method to produce segmentally labelled proteins adapted to the needs of the project.

3.1 Optimization of the labelling procedure

The CF expression technique previously employed by the group was modified in order to produce the labelled samples (91,215). Note that the CF procedure is based on the protocol developed by the group of G. Otting (The Australian National University) (264). The addition of deuterated amino acids into the protein production is relatively simple in CF. As an open system, CF allows for the direct manipulation of the ingredients of the mixture, after which the reaction mixture can be scaled, depending on the required amount of sample.

The reaction recipe (Materials & Methods, Figure 8.4) was split into four steps to ease sample preparation and optimization. The 10x reaction mixture contained components that were not optimized, while the master mixture contained all components, including those optimized, except for the reaction starting cell lysate and target sequence plasmid. Labelled amino acids, magnesium acetate, and tRNA were added as part of the master mixture at this stage. The

concentration of these three aforementioned components is often varied to optimize protein expression. Finally, the plasmid coding for the protein and the *E. coli* lysate was added after which the CF reactions could be incubated, thereby producing the samples. It was noted that in between the addition of each component, the solution had to be mixed, as omitting the step would result in an inhibited protein expression.

Each lysate production was tested for the optimal magnesium acetate concentration depending on protein construct by monitoring the protein expression over four hours in 50 µL samples. Once these parameters were optimized, the same conditions were maintained for all productions using the same *E. coli* lysate.

3.2 Selective Incorporation of deuterated glutamine and glutamic acid

3.2.1 Buffer optimization

Supplying the reaction mixture with the deuterated amino acid will incorporate it into the protein by the translation machinery of the lysate. Although this approach is valid for the majority of canonical amino acids, the incorporation of deuterated glutamine is challenging. As described in the introduction, our CF expression system still contains the transaminase that transforms glutamine into glutamic acid and *vice versa*. As our standard CF reaction buffer contains potassium glutamate, the glutamate from the buffer is converted into glutamine during the reaction. Therefore, throughout the protein synthesis, the level of deuterated glutamine would be depleted (converted into deuterated glutamate) and the protonated glutamate would be converted into protonated glutamine and incorporated into the expression, resulting in an infinite number of protein isotopologues with varying levels of deuterated glutamine and glutamate that could not be isolated.

In order to accomplish the specific labelling schemes, the CF mixture had to be modified. The two main solutions were:

1) Inhibit the enzymatic scrambling reaction.
2) Change the reaction conditions.

It has been shown that the inhibition of the enzymatic pathway can be achieved with β-chloro-L-alanine (265). That being said, the efficacy of this inhibitor has been shown to be partial, resulting in the production of multiple isotopologues in the labelled sample (258).

Consequently, the challenge was addressed by modifying the CF reaction conditions. One way to reduce the negative effect of scrambling is to maintain the same deuteration form for both glutamine and glutamic acid in all produced samples. Note that if the expression mixture was depleted for other sources of glutamine and glutamic acid, this would result in a homogeneously labelled sample. However, this approach would be only successful if an alternative buffer, different than the KGlu, could be used. While untested, another method to prevent scrambling would be to exchange the KGlu with deuterated KGlu, which would ensure that only deuterated Glu and Gln residues were incorporated, but this method again would be a significant expense due to the high concentration of deuterated buffer needed (100 mM).

The performance of the two alternative buffers, KOAc and the D-form of potassium glutamate (D-KGlu), was evaluated (258,259). KOAc is commercially available and can be readily prepared from the salt. Conversely, the D-KGlu solution has to be prepared from commercial D-glutamic acid. This was achieved by re-suspending the D-glutamic acid in 1 M HCl and titrated with 8 M KOH to reach a pH of 7.2. At this pH, the remaining glutamic acid was solubilized, and the clear solution was used as a buffer during CF expression tests. An equivalent buffer was prepared using the same procedure from the L-glutamate as a control.

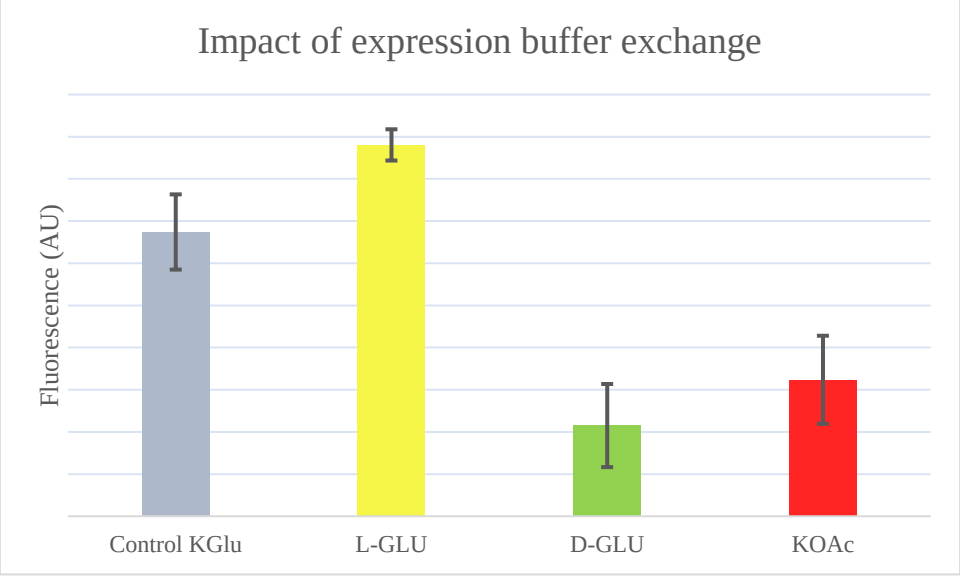


Figure 3.3: Average CF expression levels of unlabelled Htt using different buffer monitored by fluorescence. Control KGlu and KOAc buffers were produced from commercial compounds while the L-Glu and D-Glu buffers were prepared from commercial L- and D-glutamic acid, respectively. Each test was an average of three independent 4-hour CF reactions in 50 µL.

The expression level of hH16 was significantly lower, when KOAc and the D-form glutamate were used as buffer, while the control commercial potassium glutamate and titrated L-form glutamate buffers both allowed for higher expression (Figure 3.3). The comparable levels of

expression between the KOAc and the D-Glu buffer as well as the ease of preparation of the former, lead to the decision to focus on optimizing the use of acetate substitution during expression. The expression level being halved by the use of these buffers compared to the control buffer had a large impact on sample preparation during the project. Every sample containing deuterated Gln/Glu residues would require double the CF volume compared to that of unlabelled samples. This would significantly increase the reagent cost per labelled SANS sample.

3.2.2 Addition of Deuterated Glutamine

The impact on the yield of the level of glutamine supplied to the CF mixture of samples produced with acetate-based buffer was investigated. Samples of 50 µL CF mixture were prepared with increasing concentration of Gln (from the initial concentration of 1 mM) and incubated for four hours at 23 ℃.

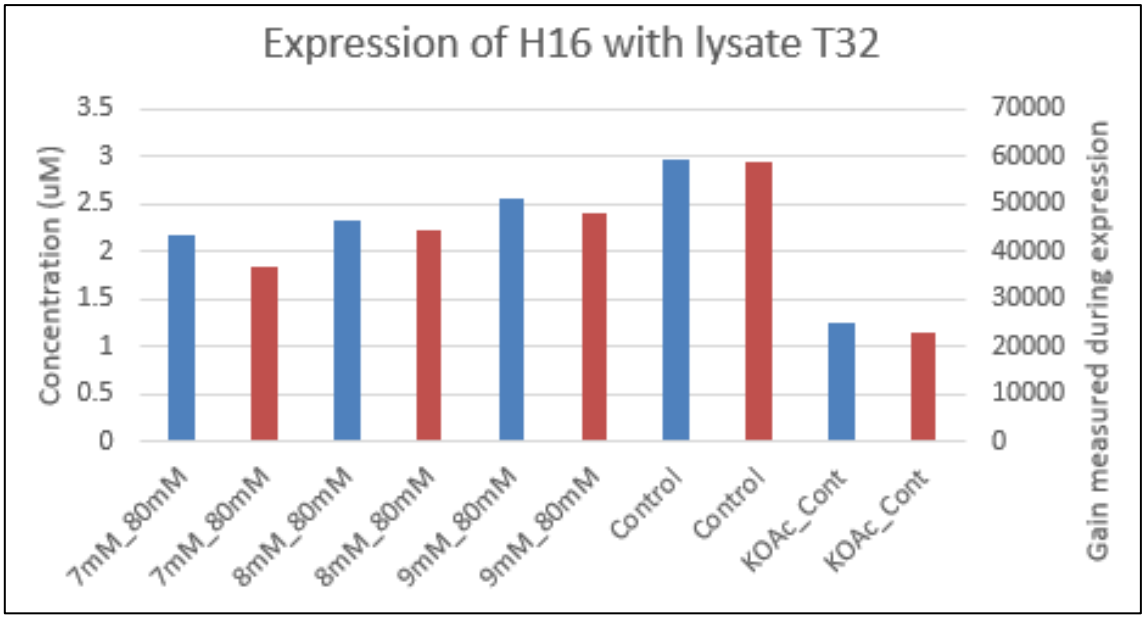


Figure 3.4: Increased concentration of glutamine significantly improved the yield of the CF reaction. The KOAc control and the three increased glutamine samples all had 80 mM KOAc as buffer and the Control was using KGlu buffer. These data were found to be specific to the lysate T32 with following lysates showing precipitation when 9 mM Glutamine was added to the CF mixture.

The experiment showed that increasing the concentrations had a positive effect on the yield of reaction (Figure 3.4). At 9 mM glutamine, the yield almost reached that of the KGlu-buffer control. The increased yield could make labelled samples easier to produce as the sample volume would be decreased, although the cost of labelled amino acid would increase significantly per sample due to the higher concentration needed. This result was unreproducible with later lysates produced in-house.

While it had a positive impact on the T32 lysate, tests using lysates T34 and T35 showed significant precipitation during CF incubation. Protein samples produced using subsequent lysates would only incorporate 1 mM Gln, as it had the most reproducible result.

3.3 PROTEIN PURIFICATION

The samples produced by CF were purified in two steps. The first step was an immobilized metal-ion affinity chromatography (IMAC) and the second a size-exclusion chromatography (SEC). The metal-ion affinity chromatography was done using either a 5 mL HisTrap™ High Performance column or 2 mL nickel-ion nitriloacetic acid (Ni-NTA) resin. The fractions collected from the IMAC purification were evaluated by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) (Figure 3.5).

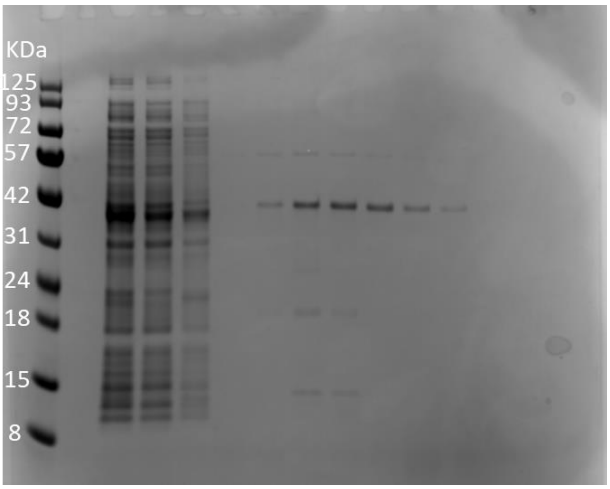


Figure 3.5: SDS-PAGE gel of IMAC fractions collected of a hH16 CF expression. Left most column was filled with a PageRuler™ Unstained Protein Ladder. Fractions primarily containing proteins around the 42 KDa mark was selected for further purification

The collected fractions were buffer exchanged overnight to a lower salt-concentration buffer (20 mM BisTris, 150 mM NaCl, pH 6.5) and then purified again by the SEC. The HiLoad® H16/600 Superdex® column was used for each purification and the protein eluded at ~88 mL (Figure 3.6).

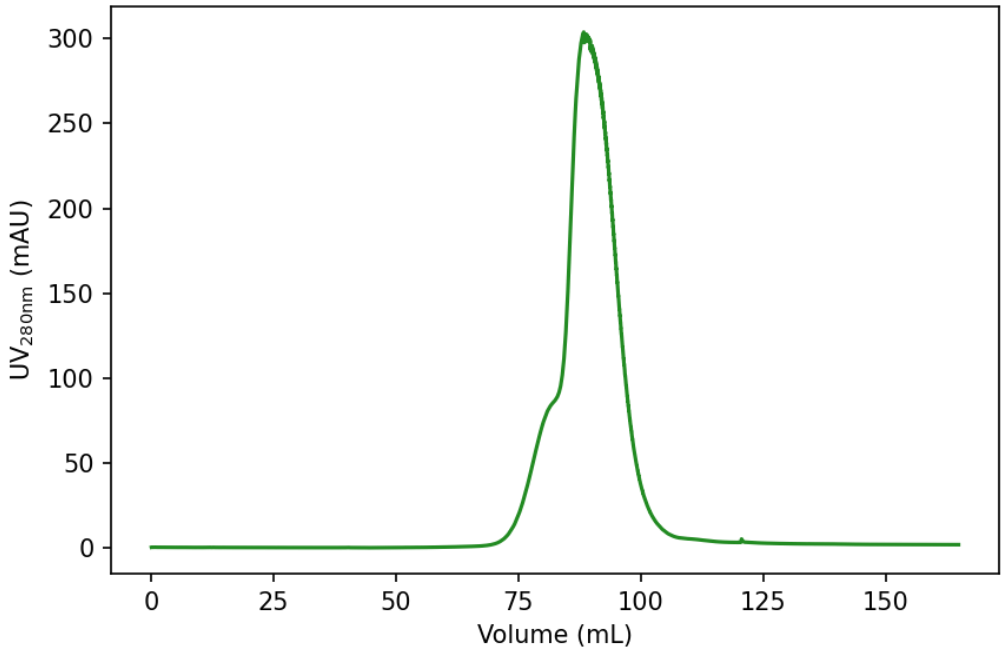


Figure 3.6:SEC FPLC elution profile of H16 protein sample. Fractions of 2 mL were collected between 50 mL and 120 mL.

Fractions of 2 mL were collected by the FPLC system between 50-120 mL and the UV traces showed a peak with a slight shouldering to the left indicating the presence of larger protein isomers prior to purification. The fractions surrounding the peak were tested by SDS-PAGE to confirm protein elution. According to the SDS-PAGE, the fractions just before the peak contained a weak protein band between 57-72 kDa, which was thought to be an impurity, but the weak band was not present in the fractions collected of hH16 suggesting that the sample was pure.

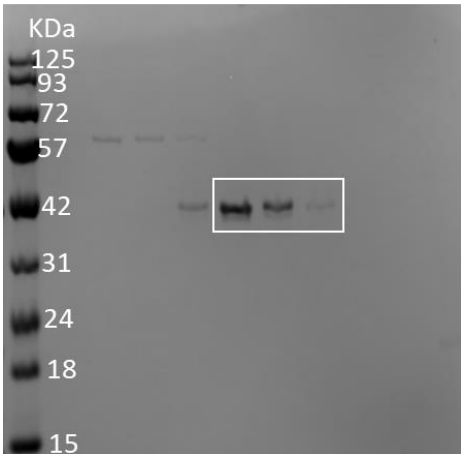


Figure 3.7: SDS-PAGE gel picture of hH16 sample after SEC-purification. The three fractions within the white-square were chosen as the final protein sample.

The selected fractions (Figure 3.7) were collected, flash frozen in liquid nitrogen, and stored at -80°C prior to scattering experiments. All protein samples of Htt produced showed a slight shoulder during SEC purification, suggesting the presence of larger species and confirming the need of SEC separation during scattering experiments.

3.4 EXPRESSION OF LABELLING SCHEMES

Depending on the chosen deuteration scheme (Figure 3.1) the expression mixture was adjusted to allow expression without significant scrambling. In general, the changes between each of the samples was confined to the buffer, amino acid mixtures, and separate deuterated amino acids. The volumes were all corrected with H_2O, to keep the concentrations constant.

Samples were separated into two groups depending on whether the labelling needed to account for the scrambling effect. hH, hH-dP, dH-hQE, and dH-hQEP samples were produced using KGlu buffer, while hH-dQE, hH-dQEP, dH, and dH-hP samples were produced using KOAc buffer. Gln, Glu, and Pro residues were added separately according to the labelling scheme. Depending on the buffer utilised for the specific labelling scheme, there would be differences in CF reaction volume. The samples produced with KGlu buffer were produced from 20 mL of CF mixture, while samples reliant on KOAc buffer were produced from 48 mL of CF mixture.

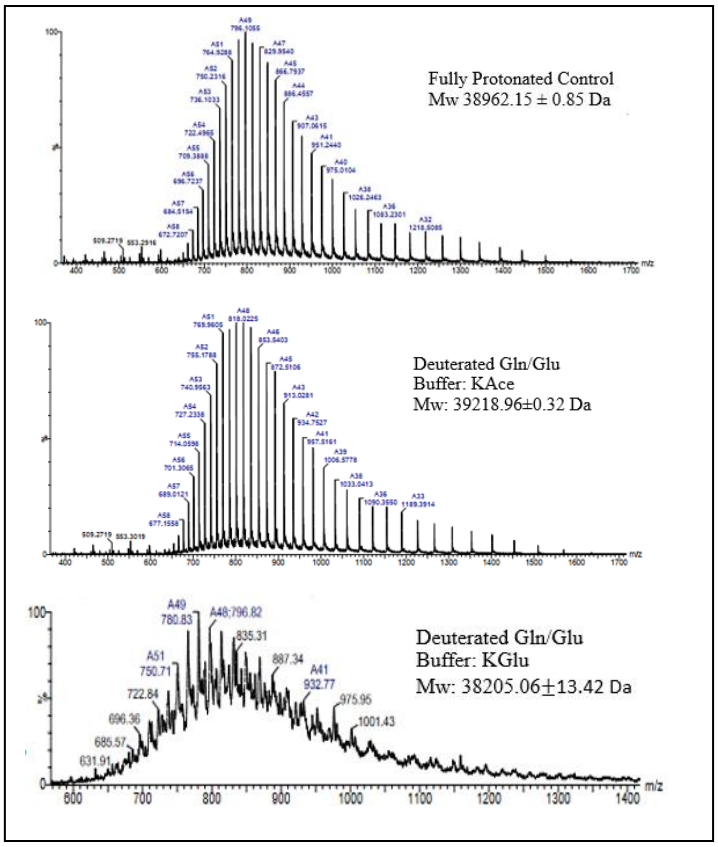


Figure 3.8: MS Spectra obtained from three protein samples, to evaluate homogeneity of the labelled Htt16 protein. Top: unlabelled control protein. Middle: Optimized CF labelling using deuterated Gln/Glu residues and KOAc buffer to suppress scrambling. Bottom: Unoptimized CF labelling using deuterated Gln/Glu residues and KGlu buffer. The bottom MS spectra shows how residue scrambling can introduce a multitude of isotopologues.

In order to verify the labelling of samples containing deuterated glutamine (dGln) and glutamic acid (dGlu), samples were produced with both KGlu and KOAc buffer. Mass spectrometry was then performed on the purified protein samples (Figure 3.8).

The three samples were measured at the IGF Montpellier (Functional Proteomics Platform) using the liquid chromatography mass spectrometry technique (LC/MS). It was visibly evident that the sample containing deuterated Gln/Glu animo acids, produced in KGlu buffer has undergone residue scrambling. The broadening of the peaks caused the weight estimation to become unreliable and suggested the presence of a multitude of isotopologues in the sample. In contrast to the poor quality of this sample, the sample produced in KOAc buffer was very similar to the control sample with a much higher peak separation and a weight consistent with the theoretically estimated molecular weight (Theoretical weight: 39229.6 Da).

Several samples of H16 and H36 were subsequently characterized by MS after scattering experiments. These MS measurements were performed by the MS Platform of Integrated Structural Biology Grenoble (IBSG, at IBS, Grenoble) and were used to confirm labelling success (Table 3.1). These experiments were done using the matrix assisted laser desorption/ionization time-of-flight mass spectrometry technique (MALDI TOF-MS)

Table 3.1: Mw measurements from MS experiments of samples measured at the IBSG. While most samples showed an excellent agreement to the theoretical mass, two samples diverted. dH36-hQE was lower than the expected Mw while hH36-dQEP showed two separate MW values neither of which matched the label.

Sample Name	Theoretical Mw	MS Mw	
hH16	38962.9	38966.5	
hH16-dQEP	39532.5	39522.0	
hH16-dP	39265.8	39264.6	
hH36-dQE	41892.8	41832.0	
hH36-dQEP	42195.7	42108	43015
hH36-dP	41828.4	41829	
dH36-hQE	43415.4	43140	

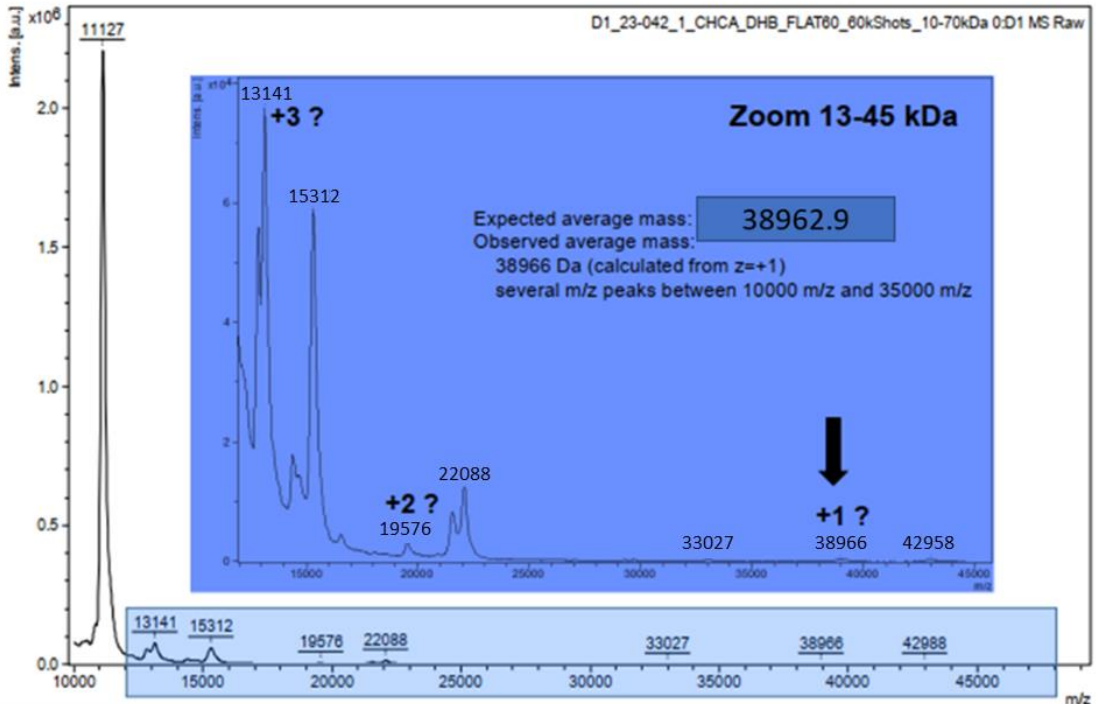


Figure 3.9: MALDI TOF-MS profile obtained of the most recent protonated hH16 sample. The observed average mass matches well that of the expected weight (~3 Da difference).

The MS experiment was missing for some samples where protein could not be recovered from the SANS experiments due to problems with the fraction collector of the SEC-SANS setup. Most of the tested samples showed a good correspondence with the theoretical MW, which confirmed that the labelling process had been a success (Figure 3.9).

The most significant deviation was that of the hH36-dQEP sample (Figure 3.11) for which the MS analysis indicated two separate molecular weights, a slightly lower and significantly higher (88 Da lower and 819 Da higher) suggesting that not only had the labelling process not succeeded, but two significantly different isomers were present in the sample. Note that SANS data measured for this sample was not used for the structural analysis.

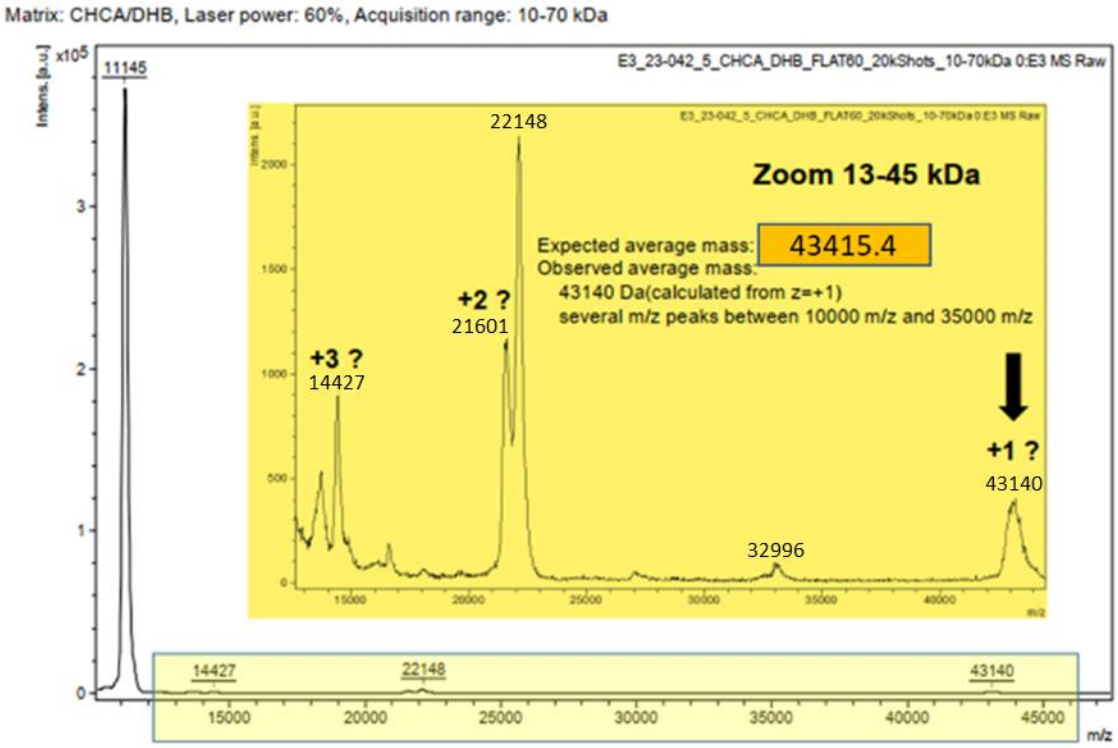


Figure 3.10: The MALDI TOF-MS spectrum obtained for dH36-hQE. The measured average mass is lower than that of the expected MW.

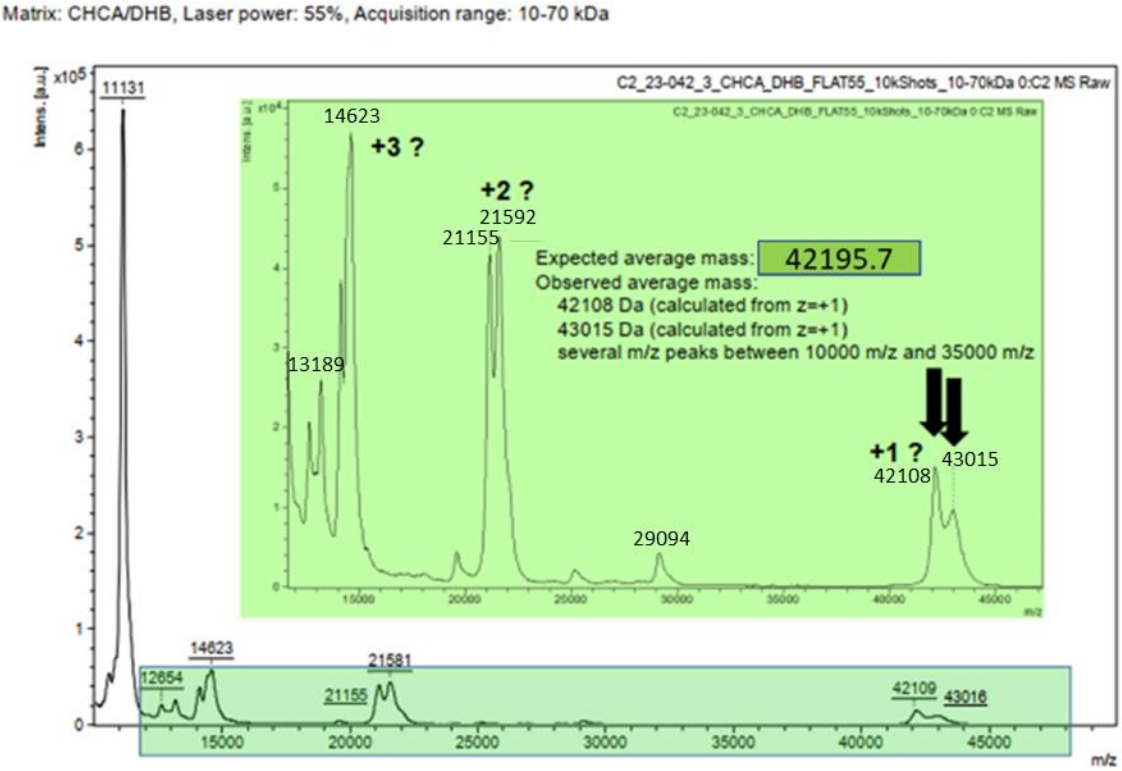


Figure 3.11: The MALDI TOF-MS spectrum obtained for hH36-dQEP. The spectrum shows two different molecular masses, neither of which matches that of the expe

4 COMPUTATIONAL APPROACHES

4.1 CONSTRUCTION OF SPECIFICALLY DEUTERATED HTTEXON-1 ENSEMBLES

In order to analyse the SAS data, it was necessary to base the simulations and ensemble optimisation on a very large set of plausible Htt-Exon-1 conformations. In this chapter, the ensemble generation is explained.

The generation of atomic ensembles to produce an adequate and realistic structural ensemble for analysis of experimental SAXS and SANS data is described in detail in the material and methods section (9.10); an ensemble of Htt-Exon-1 was initially built from a tripeptide-fragment database (165) derived from SCOPe (266), which compiles experimentally determined high-resolution 3D structures of proteins. As detailed in the original study (165), the procedure consisted in concatenating tripeptide fragments extracted from the database, ensuring the construction of conformations founded on structural data. Subsequently, the Htt-Exon-1 fragments of H16 and H36 were linked to a 3C linker [Seq: EASLE VLFQG PGSH], then connected to the sfGFP structure containing a C-terminal His-tag (see methods section).

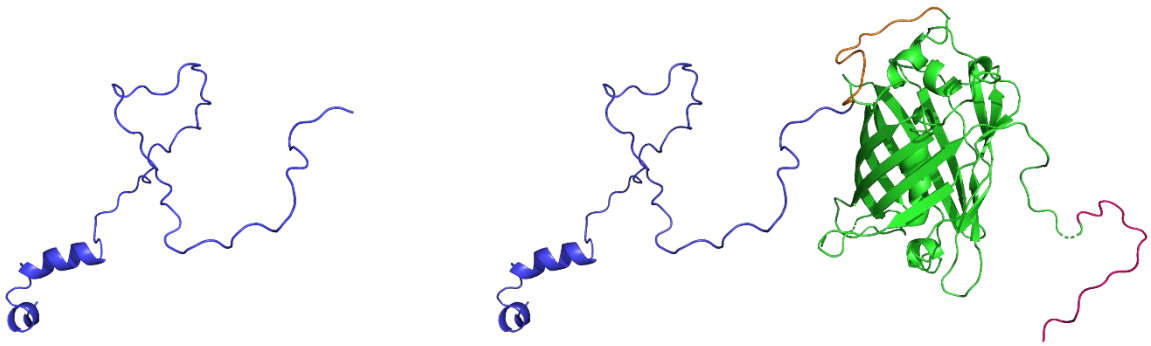


Figure 4.1: Left – Example of an initial HTTExon-1 peptide created using the tripeptide database. Right – Fully assempled conformation of H16 containing linker (Orange), sfGFP (Green) and His6-tag (pink)

Note that this computational model matched the samples produced using CF expression (Figure 4.1). While Ranch (174) and Pulchra (267) served to link and optimize the structures to avoid steric clashes, an important step was renaming the fragments and the atom numbers of the PDB files to obtain a single, continuous chain in order to do subsequent analyses. The python package pdb-tools (268) functions *reres* and *reatom* were used for renumbering the PDB files. Finally, an ensemble of 5,000 different conformations was generated for each of the HTTExon-1 constructs, H16 and H36.

The second step was to selectively deuterate the built structures according to each of the previously described deuteration patterns. For this, a Python code was written, enabling the residue specific deuteration that can be achieved when using the CF protein synthesis (Figure 4.2). A list of amino acids could be specified in the script by the user, enabling the program to read the pdb-file, find the specified amino acid, and substitute both labile and non-labile hydrogen atoms with deuterium atoms. Similarly, the inverse labelling could be obtained by specifying which amino acids were to be kept hydrogenated, while deuterating all other residues. With this approach an ensemble of 5,000 structures was prepared for each of the eight different labelling schemes, totalling 40,000 structures. Importantly, this Python script was of general use and could be applied to all experimentally possible deuteration patterns (full script: Appendix 11.1).

```
for file in os.listdir(directory):
    if file.endswith(".pdb"):
        filefrompath = os.path.basename(file)
        filename = (os.path.splitext(filefrompath)[0])
        ppdb = PandasPdb().read_pdb(directory + file)
        ppdb.df['ATOM'].head()
                        #Choose residues to be exchanged
        listofresidues = ['GLN', 'GLU', 'PRO']
                        #Choose Mask or inverted mask depending on mode of deuteration
        # invertedmask = (ppdb.df['ATOM']['residue_name'].isin(listofresidues)) &  (ppdb.df['ATOM']['element_symbol'] == 'H'
        # mask = ~invertedmask
        mask = (ppdb.df['ATOM']['residue_name'].isin(listofresidues)) &  (ppdb.df['ATOM']['element_symbol'] == 'H')
                        #ppdb.df function will exchange specified atoms from Hydrogen to Deuterium
        ppdb.df['ATOM'].loc[mask,['atom_name']] = ppdb.df['ATOM'].loc[mask,'atom_name'].str.replace("H", "D")
                        #Important for further use of files to change the element symbol of the PDB file
        ppdb.df['ATOM'].loc[mask,['element_symbol']] = ppdb.df['ATOM'].loc[mask,'element_symbol'].str.replace("H", "D")
                        #Specify output name below
        ppdb.to_pdb(outputpath+'Deuterated_HPro_'+filename+'.pdb')
    else:
        continue
```

Figure 4.2: Python Script for the selective deuteration of amino acids. Residues specified in the list "listofresidues" will have all hydrogen atoms exchanged to deuterium atoms. In this case, non-exchangeable hydrogen atoms of Glutamine, glutamic acid and proline residues would be transformed in deuterium atoms.

4.2 INCORPORATION OF H/D EXCHANGE INTO STRUCTURES AND CALCULATION OF THE ASSOCIATED SCATTERING PROFILES

In order to compute the SANS and SAXS patterns for each individual conformation of the ensembles, CRYSON (173) and CRYSOL (172) were used, respectively. The number of harmonics (*-lm 30*), number of datapoints (*-ns 200*), and the option of using explicit hydrogens (*-eh*) were kept consistent between the two softwares. Additionally, for the SANS calculations, CRYSON required the solution deuteration level to be specified (*-D2O* 0 specifies 0% D_2O in the buffer solution).

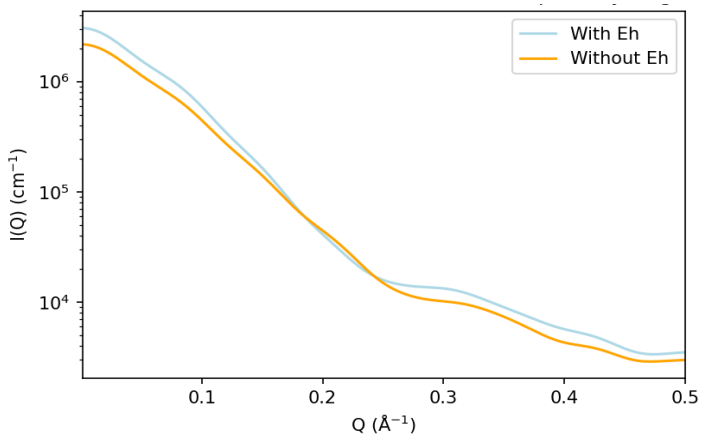


Figure 4.3: Computed profiles for H16 in 100% D_2O show that CRYSON does not exchange labile hydrogens when the explicit hydrogen option (-eh) is added. Concretely, the overall intensity (contrast) is higher when the explicit hydrogen option is enabled.

The SANS scattering profiles of fully hydrogenated protein could be calculated without the explicit hydrogen option by varying the *-D2O* parameter. However, a problem was discovered when the scattering profile was calculated for structures with specific deuteration. Concretely, while CRYSON could vary the level of solution D_2O, it was done by applying an average deuteration value to each residue, which influenced its scattering properties. Indeed, this meant that any specific deuterium labelling would have been lost if the *-eh* was omitted from the CRYSON command, because the specific H or D atoms were not considered. Therefore, the function *-eh* forced the software to recognize hydrogen and deuterium positions within the protein. However, while the labelling pattern of the structure would have been conserved when using *-eh* option, the labile hydrogen exchange depending on the D_2O level was not applied (see below).

As shown in figure 4.3, the scattering pattern of a fully hydrogenated structure varies when simulated in 100% D_2O solution with or without the explicit hydrogen option during profile calculation. In theory, the labile hydrogens of each residue should have been exchanged for deuterium, but when the *-eh* function was applied, the scattering intensity was slightly higher than that obtained without applying this function. This suggests that the content of hydrogen was higher, increasing the contrast between the theoretical solution and the simulated structure. While other scattering calculation software, such as PepsiSANS(118), could have been used to account for labile hydrogens during profile calculation, the ensemble production pipeline had already been optimized for the use of CRYSON and CRYSOL.

A solution to the problem of the incorporation of the H/D exchange phenomenon in the structures would be to account for labile exchange before calculating the scattering intensities and then not allowing the calculation software to adjust the hydrogen contrast (i.e. using the explicit hydrogen option of CRYSON). A Python script was produced to explicitly exchange labile hydrogens throughout a PDB structure. In contrast to CRYSON, which uses an average SLD value for each hydrogen position, this Python script randomly exchanged labile atoms. The python module *random* was used to randomly exchange each labile hydrogen to deuterium at a given rate, using the python pymol module (Figure 4.4). For instance, for a sample measured at 40% of D_2O, each exchangeable hydrogen (those bound to O, N or S) would have a 40% chance to be a deuterium and 60% chance to be a hydrogen. Note that this procedure can be done regardless of specific deuteration pattern used. In this procedure, it was not distinguished between globular and disordered sections of the same protein, and both of them present the same exchange rate. The python script could eventually be improved in order to provide different exchange rates depending on the solvent accessibility.

```
19          #The function will randomly deuterate hydrogen depending on the given rate
20      def myfunc(model,index):
21          change = random.random() < rate
22          if (change) :
23              cmd.alter('%s and index %d'%(model,index),'elem = \'D\'; name = \'D\'')
```

Figure 4.4: Random function defined in the labile exchange script.

While the random module will yield an average random deuteration at a given percentage, the variation of this rate within the ensemble, induced by the use of a random selection, had to be considered. The labelled ensembles should allow small variation in deuteration of labile atoms around the theoretical D_2O rate, but outliers should be identified and recalculated. After performing a computational exchange for a given conformation, a check was introduced to evaluate its overall deuteration. A variance of 10% of the overall deuteration level was allowed to accommodate experimental equipment uncertainties. For instance, if the deuteration target is 40%, a range of 36%-44% would be allowed by the algorithm. Any structure exceeding these limits would be recalculated and a new deuteration would be proposed (Figure 4.5) (full script: Appendix 11.2).

```
71              Upper = cut * 1.1
72              lower = cut * 0.9
73
74              if check > lower and check < Upper :
75                  # print(str(check)+' sample '+name_deut+' correct - contienuing')
76                  cmd.save(ratename+'/'+name_deut+'.pdb')
77                  cmd.delete('all')
78                  break
79              if check < lower or check > Upper :
80                  # print(str(check)+' sample '+name_deut+' should be resimulated')
81                  cmd.delete('all')
82                  e = e + 1
83              if e > 200:
84                  print('Script cannot compute correct deuteration pattern. Errorcount: '+str(e))
85                  cmd.delete('all')
86                  sys.exit(0)
```

Figure 4.5: Deuteration check to avoid outliers of labile exchange percentage. The three conditions allow (i) accepting a structure, (ii) recalculating the structure and (iii) in cases where the script can not deuterate the structure within the rate, the script will abort.

4.3 Effect of random deuteration on the scattering properties of proteins

In addition to controlling the exchange rate variation, the variability of the scattering patterns resulting from the random H/D exchanged structures would also need to be considered. Note that when the script is run multiple times on the same conformation, the number and positions of the added deuterium atoms could be different (Figure 4.6). Therefore, the variation between repeatedly exchanged structures might have led to differences in the calculated scattering intensity.

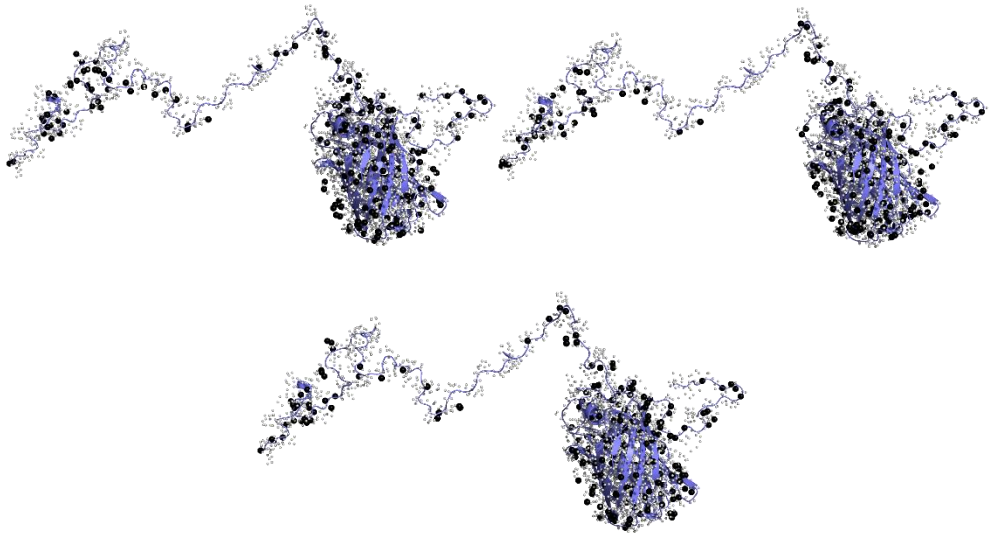


Figure 4.6: Representation of random labile atom exchange for a fully protonated H16 conformation. Exchanging the deuteration level to 40% will exchange hydrogen (white spheres) for deuterium (black spheres) in random positions throughout the molecule.

To address this point, the labile exchange script was tested by repeatedly exchanging the same conformation of fully protonated hH16 ten times at 40% D_2O and the scattering intensities were calculated using CRYSON with the explicit hydrogens option (see above). The ten individual curves and the average profile were plotted and compared (Figure 4.7). The only visible difference between the curves was a slight variability beyond 0.3 $Å^{-1}$. This difference can be attributed to the distinct number and positions of the deuterium atoms in the molecule, which induce slight perturbations in contrast to those that are only visible at high angles. Note, however, that the error margin of the experimental intensity of low concentration (1-5 mg/mL) biological SANS samples in partly deuterated solutions in this *Q*-range are expected to largely outweigh the variance observed in this theoretical simulation. Having validated this protocol of atomic model labelling exchange, it could be applied to the theoretical structure ensemble. The resulting ensemble pipeline afforded the project the ability to produce and compare profiles with specific labelling patterns in realistic deuteration states depending on the solution D_2O level.

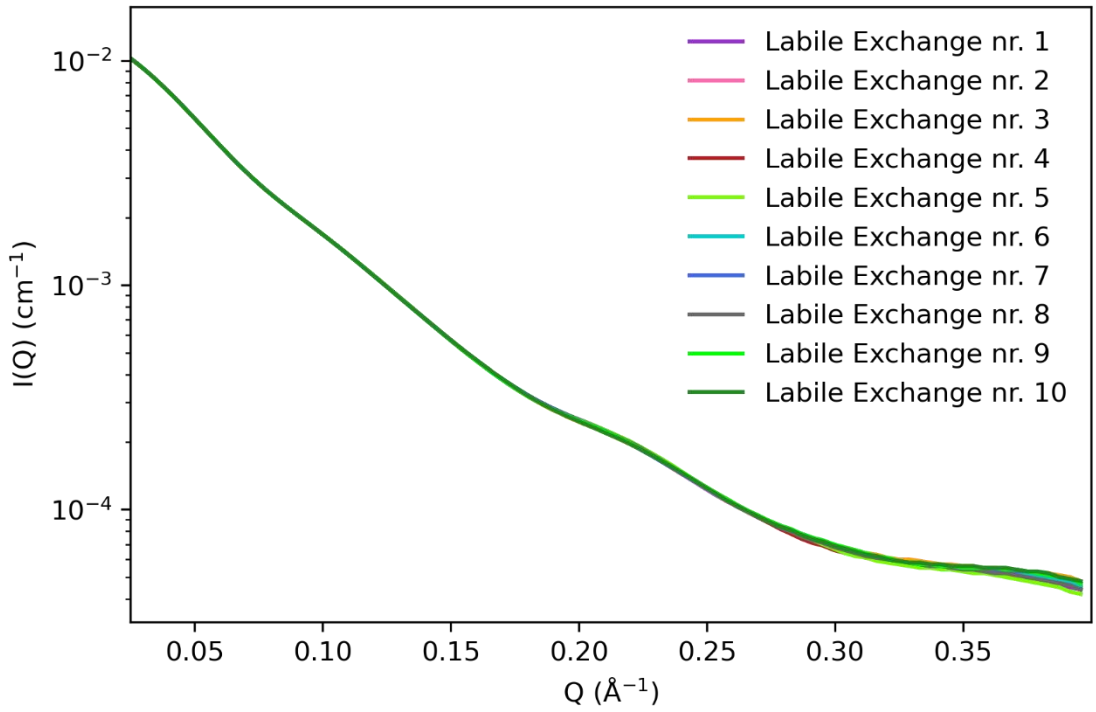


Figure 4.7: SANS profiles computed for the same conformation of the hH16 in 40% D_2O with ten different (random) labile H/D exchange and compared with their average profile, which is virtually equivalent to the 10 individual profiles..

4.4 THEORETICAL SCATTERING ENSEMBLES

An initial ensemble of 5,000 structures was chosen for both H16 and H36. The pipeline displayed in Figure 4.8 for an individual conformation was implemented to create the 48 ensembles (8 deuteration patterns in 6 D_2O levels), resulting in 240,000 structures per construct. While the calculation of the scattering intensities with both CRYSON and CRYSOL were fast, their use for such a large number of conformations proved slow, and the software *parallel* (269) was implemented to run multiple calculations simultaneously.

Example of CRYSOL command:

*ls *.pdb | parallel -j 12 'crysol {} -lm 30 -ns 200 -eh'*

Example of CRYSON command

*ls *.pdb | parallel -j 12 'cryson {} -lm 30 -ns 200 -eh -D2O 0'*

These calculations were performed on the Virtual Infrastructure for Scientific Analysis (VISA) server at the ILL.

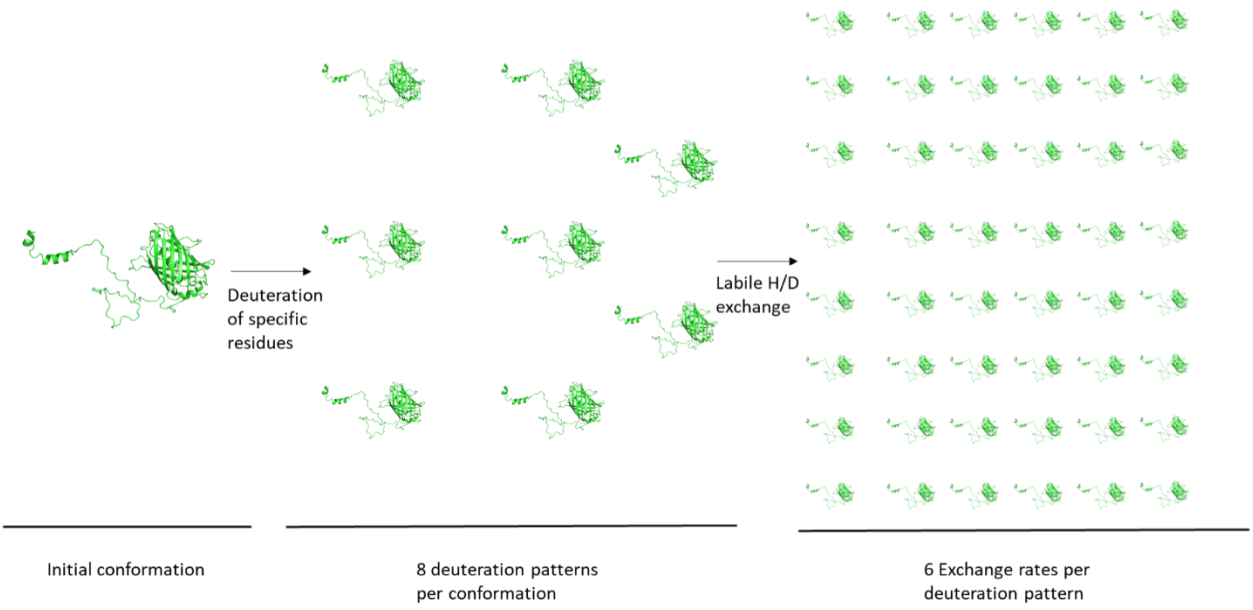


Figure 4.8: Visualization of the data generation for an individual conformation from the initial unlabelled structure to the 48 labelled and exchanged isotopologues, by creating eight differently labelled patterns and exchanging the labile hydrogens of each labelling pattern to six deuteration levels.

The average SANS profiles for the 48 ensembles for both constructs, H16 and H36, were calculated and visualized to show how the labelling and the deuteration pattern impact the resulting scattering profiles and the amount of information that could be derived from these

samples. The *I(0)* was extracted from the Cryson calculated scattering profiles. By plotting the square root of *I(0)* for each averaged profile (correcting for the contrast sign) as a function of the D_2O percentage, I have calculated the theoretical match point for each deuteration pattern of H16 and H36 (Figure 4.9 and table 4.1).

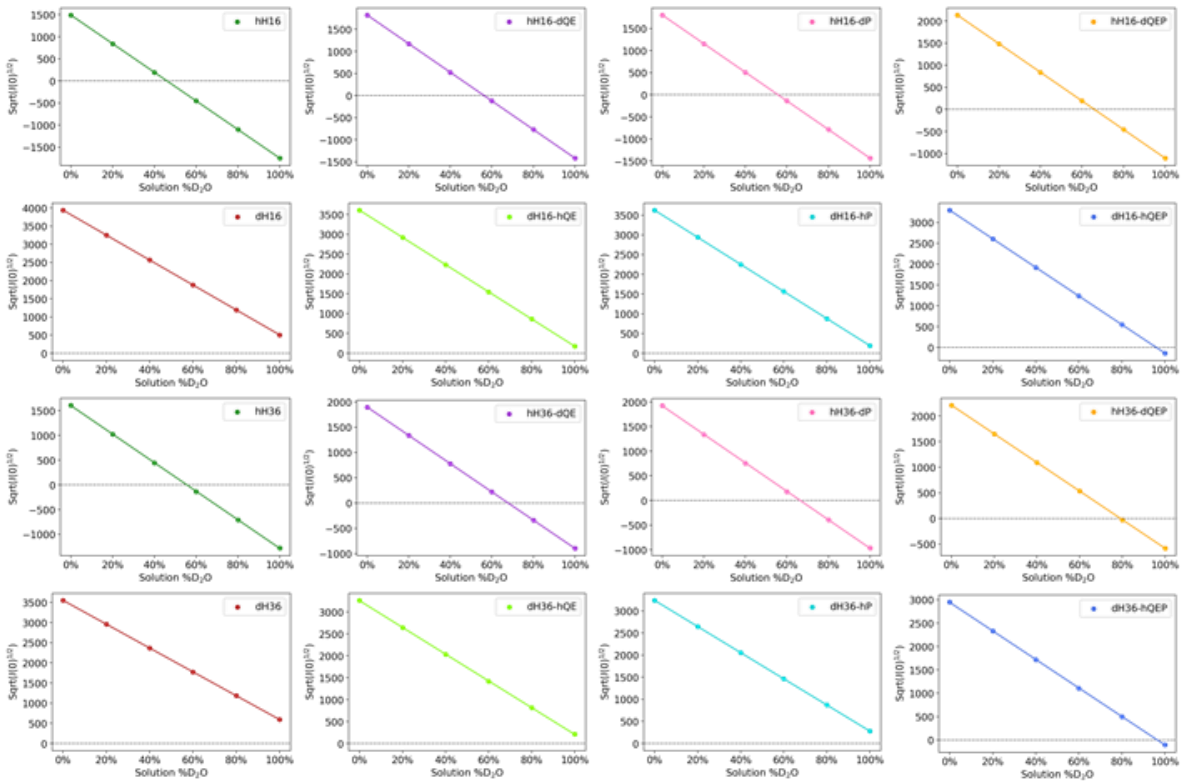


Figure 4.9: square root of I(0) computed for the ensemble of each of the eight deuteration patterns of H16 (top) and H36 (bottom) as a function of the D_2O level. Match points are defined as the D_2O level in which I(0) = 0 and they were obtained by linear regression (Table 4.1).

The matching points listed in table 4.1 show the impact of the deuteration pattern on the resulting match point. The fully protonated proteins of H16 and H36 have matching points higher than the average protein matching point of 42% D_2O. The increase of the unlabelled matching point to 46% D_2O (H16) and 55.5% D_2O (H36) was caused by the compositional bias of the constructs with a large number of Glu and Pro residues. For deuterium labelled construct patterns, the match point increased with the amount of deuterium atoms incorporated with the selective labelling, reaching 65.9 and 79.7% of D_2O level for H16 and H36, respectively. Interestingly, match points obtained for all deuterated patterns were equal to 100% of D_2O or above. Note that the match point could not be experimentally achieved when this percentage was above 100%. The amino acid composition also affected the match point of the protein, in particular the number of glutamines, with H36 displaying systematically a larger match point value than H16 with the same deuteration profile. Deuterated huntingtin with

protonated QEP was the exception given the larger amount of hydrogens in H36 than in H16. Match points obtained in this analysis will be helpful to interpret subsequent analyses.

Table 4.1: Table of matching points calculated from the slope of the averages of the simulated profiles when they are plotted by I(0)-values. Values above 100% means they can never be matched by a D_2O solution.

	Hydrogenated	dQE	dQEP	dP	Deuterated	hQE	hQEP	hP
H16	46.0%	56.2%	65.9%	55.7%	114.7%	105.1%	96.0%	105.6%
H36	55.5%	67.9%	79.1%	66.4%	120.0%	106.7%	96.4%	109.4%

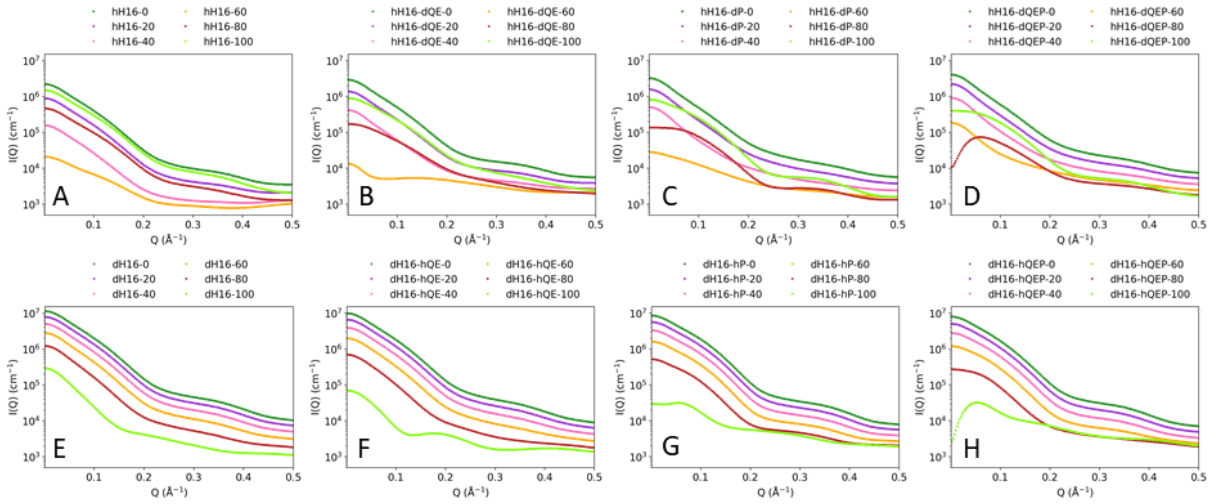


Figure 4.10: Average scattering profiles for the eight H16 deuteration patterns depending on the D_2O level (0%: dark green, 20%: purple, 40%: pink, 60%: yellow, 80%: red, 100%: light green).

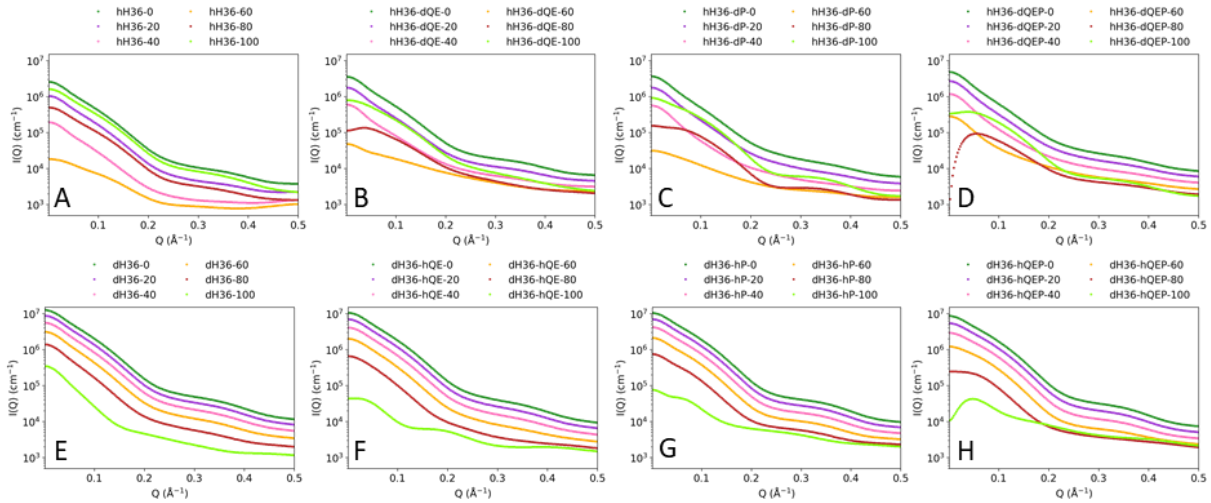


Figure 4.11: Average scattering profiles for the eight H36 deuteration patterns depending on the D_2O level (0%, 20%, 40%, 60%, 80%, 100%).

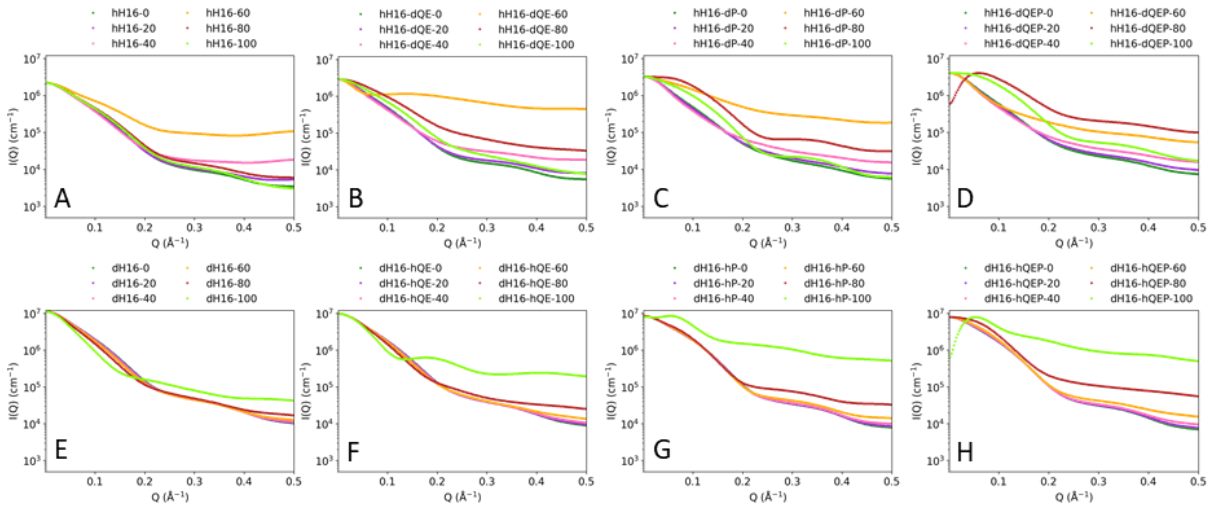


Figure 4.12: Scaled average scattering profiles for the eight H16 deuteration patterns depending on the D_2O level (0%, 20%, 40%, 60%, 80%, 100%).

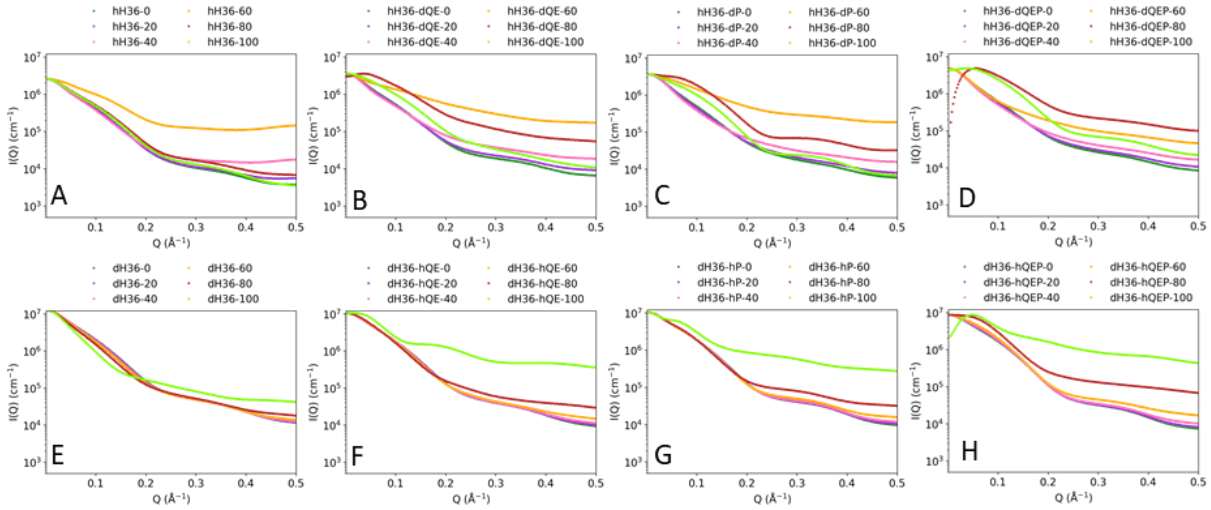


Figure 4.13: Scaled average scattering profiles for the eight H36 deuteration patterns depending on the D_2O level (0%, 20%, 40%, 60%, 80%, 100%).

The calculated average scattering profiles for H16 and H36 were grouped by deuteration pattern, plotted in absolute values, and scaled by maximum values in order to better compare and differentiate them (Figure 4.10 – 4.13). An initial observation is that the D_2O level changed the shape of the scattering pattern. This can be attributed to two factors:

1) the overall contrast between the protein and the buffer

2) the local contrasts of labelled and unlabelled regions.

When the profiles were normalized to the highest intensity value of a given curve to minimize their mutual discrepancy, some of the profiles become similar. An example of this was the fully protonated and fully deuterated H16 profiles (Figure 4.12, A & E), which showed very little

discrepancy when changing the D_2O level. However, both panels displayed an outlier, which corresponded to a D_2O level close to the match point. Note that, at or around the match point the profiles will differ more to the other conditions due to the low global scattering intensity that is hard to differentiate from the background scattering. This is an example of the effect of the overall contrast, and such an effect will not importantly increase the structural information that can be extracted from data. The second factor was evident in specifically labelled samples. hH16-dQEP had a matching point of 65.9% D_2O, but when the profiles were examined (4.10 C & 4.12 C) a difference was observed in the slope of the profiles between high (80-100% D_2O) and low (0-40% D_2O) levels of D_2O in the buffer, which originated from the labelling pattern. As the slope changed depending on the D_2O level of the buffer, these profiles would encompass additional structural information. The clustering of deuterated amino acids in a specific section of the protein enhanced these differences. While several other deuterated samples showed differences (Such as hH-dQE, hH-dQEP, hH-dP and dH-hQEP - 4.11 B, C, D, G & 4.13 B, C, D, G), there are a few, where these differences were negligible, such as the dH16-hP and dH36-hP (4.11 H and 4.13 H). This similarity suggests that the informational value of these samples would be low, and they should not be prioritized if the number of samples to be produced is limited.

In theory, the information content of partially deuterated samples would be the highest near the matching point of the hydrogenated protein as the contrast of the specifically deuterated regions would be higher compared to the non-labelled protein. However, the scattering intensity of such samples would be seriously compromised due to the low overall contrast, as observed in the previously mentioned outlier profiles. While these conditions can be produced and simulated, obtaining experimental data of such samples is technically very challenging as it would require high concentration and/or long exposure time, both affecting the monodispersity of the protein. Indeed, from SANS experiments described in Chapter 5, it was found that samples of protonated hH16 in 20% D_2O could be measured, albeit with very low signal to noise. This suggests that profiles with theoretical *I(0)* values below 10^6 cm^{-1} when using CRYSOL (Figure 4.10 & 4.11) would be difficult to experimentally obtain a good SANS profile, and samples at or below 10^5 cm^{-1} would likely not provide any helpful information. Note that these values were interpreted in the context of this study, where proteins could not be highly concentrated and in which the SEC-SANS mode was utilised. Other molecules measured in batch provide informative SANS data in experimental conditions close to the match point.

When comparing the H16 and H36 averages, several observations could be made. Profiles computed from deuterated ensembles were very similar. Theoretically, the incorporation of deuterated glutamine should have shown the biggest difference, but the introduction of 20 additional hydrogenated residues seemed to make very little difference when comparing the entire ensembles. Comparing the H16 and H36 ensembles with dQE and dQEP patterns to those of the unlabelled ensembles could provide several observations:

1. The contrast at 80-100% D_2O (Orange & Light Green profiles) was lower because of the higher % of deuterated residues.
2. Similarly, the 0-20% (Dark Green & Purple) D_2O profiles showed slightly higher *I(0)* average profiles confirming the effect of the additional deuterated amino acids.
3. Because the averaged ensembles were representing all conformations of the two poly-Q lengths, the slope of the averaged ensembles did not show significant differences when comparing hH16-dQE and hH36-dQE.

While the unscaled profiles showed a visual difference, the impact of the labelling schemes became more apparent if the averaged profiles were scaled (Figure 4.11). An example of this would be hH16-dQE which showed a difference between 0-40% and 80-100% D_2O solution simulations, and the scaled average profiles revealed an important difference between the ensembles. Apart from the high *Q* values (> 0.3 $Å^{-1}$) showing a difference, it was observed that there was a difference of the slope observed around 0.1 $Å^{-1}$ and this could be of value when combining samples of different labelling patterns for analysis.

This difference in slope was visible in all six deuteration levels and supports the theory of different labelling patterns providing slightly different structural information. While slight changes in overall intensity could be explained by contrast alone, the differences observed between the slopes of the profiles suggests differences of overall shape. While not exclusive, the labelled glutamine and proline residues were primarily located in the Htt-Exon1 domain of our H16 constructs. Theoretically, labelling the disordered part of the construct impacted the size and shape that can be observed from the SANS experiment.

When comparing the profiles by labelling pattern, the interest of the segmental deuteration was manifested (Figures 4.12 & 4.13). Profiles of hH16-dQE, hH16-dQEP, hH16-dP and dH16-hP showed differences at intermediate *Q*-range (~0.1 $Å^{-1}$), which further indicated that not only depending on the labelling pattern, but also on the individual sample conditions, impact on the scattering of a sample could be observed. Similar results were observed for the simulated

profiles of H36. In contrast, fully protonated/deuterated samples of both H16 and H36 showed only slight variations when changing the D_2O level. The protein samples were not measured at multiple different concentrations, so direct intermolecular impact could not be evaluated. Because of the low protein concentrations, the structure factor was assumed to be equal to one and not contaminated by the intermolecular interactions.

4.5 ENSEMBLE COMPARISON

In order to help to rationalize the above observations the R_g values of the individual conformations with the different deuteration patterns and H/D exchange levels computed with CRYSOL and CRYSON were compared. By plotting the calculated R_g values derived from SAXS calculations from CRYSOL to those of the labelled samples in different conditions, the impact of deuteration could be visualized on this low-resolution structural parameter (Figure 4.14 & 4.15). When correlating the two corresponding R_g values for each conformation, the scatter plot would diverge from the SAXS values in cases where the deuteration had an important impact on the overall size of the particle obtained by scattering. In this case, an important gain in structural information with respect to the SAXS data would be obtained by measuring this dataset.

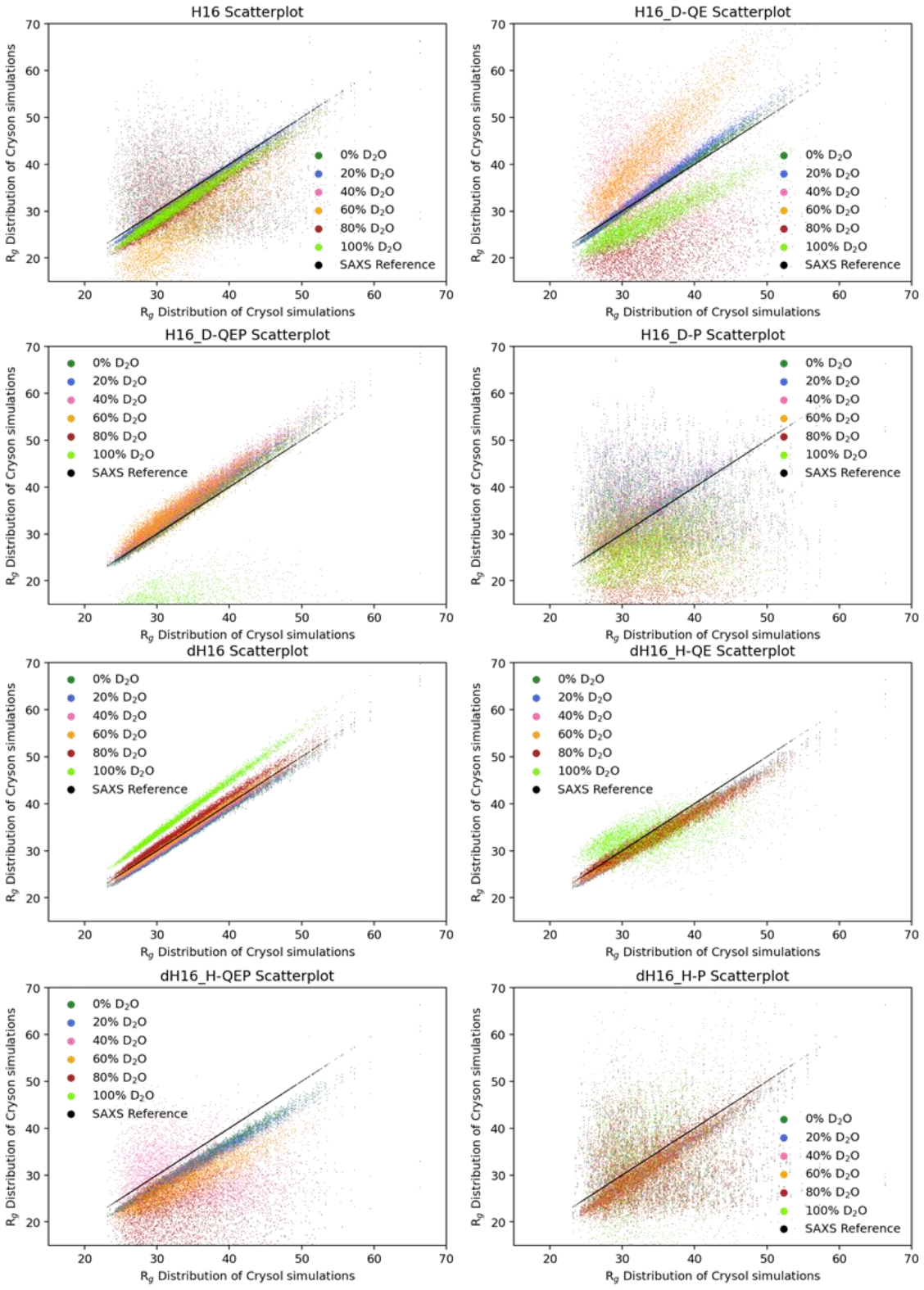


Figure 4.14: Scatter plots comparing the SAXS calculated R_g with the calculated SANS R_g of each deuteration pattern and solution D_2O level. The two major factors impacting these scatterplots are the shape and size of the observed structure and the overall match point of the given labelling pattern.

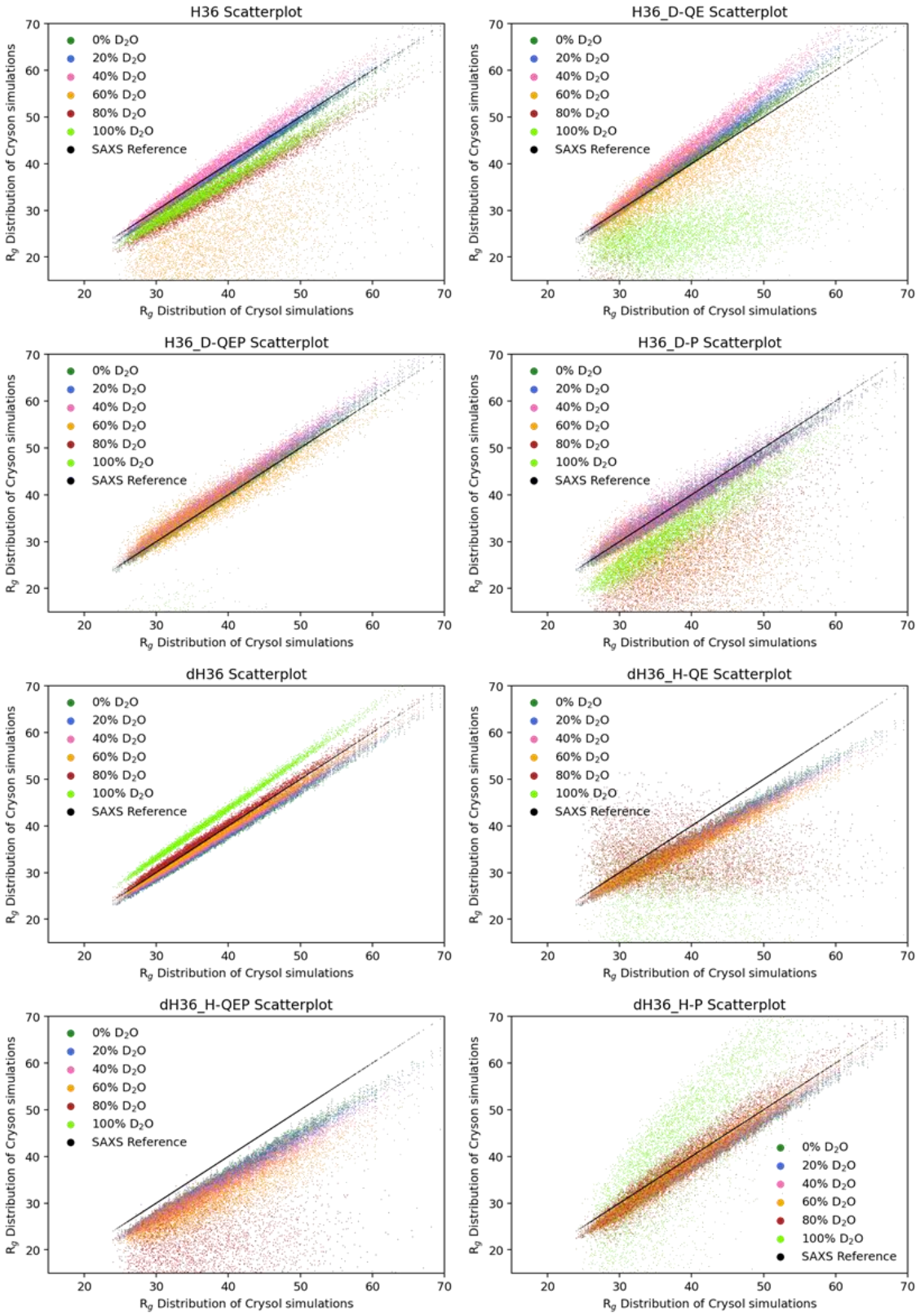


Figure 4.15: Scatterplots comparing the SAXS calculated R_g with the calculated SANS R_g of each labelling pattern and solution D_2O level. Calculated for the H36 construct ensembles

It was important to identify the origin of the distortions of the plots. When the protein was labelled, its SANS contrast became inhomogeneous, and this would cause the information

content to shift depending on the specific deuteration pattern. In the case of the labelling patterns presented in this project, the specific labelling was mainly located in the disordered part of the protein. The observed scattering from a sample would depend on both the deuterated and protonated part of the protein. The scattering pattern in an 80% D_2O solution was a combination of the positive scattering contrast of deuterium and the negative contrast of hydrogen. If the sample condition approached a matching point of the protonated part of the protein, the relative contribution of the deuterated part to the signal would increase, and this region could be highlighted by SANS. Notice that this scenario is similar to the match out experiments where one of the different components of a biomolecular complex was rendered invisible by scattering. In the case of Huntingtin, this exact approach became unrealistic due to the small number of deuterated residues and the low/moderate concentrations that could be achieved with this protein.

In theory, the closer to the protonated match point, the larger amount of specific structural information of the poly-Q could be extracted. This could be observed, for instance, for the hH16-dQE scatterplot, which showed the greatest deviation from the SAXS reference in conditions around the overall protein match point of 46.0% D_2O. In these cases, the majority of the scattering signal would arise from the poly-Q. The crux of this sample, however, was the realistic applicability of the experimental conditions. In the case of H16, the construct incorporates 30 glutamines and 23 glutamic acid residues. With only 53 deuterated residues and the relatively low protein concentrations (1-5 mg/mL), the scattering intensity from a partially matched sample, would be too low compared to that of the incoherent background scattering of a sample measured in 46.0% D_2O.

In samples such as hH16-dQE, the deviation from the SAXS reference was the largest between 40-60% D_2O (pink & orange), while the maximum deviation in the dH16-hQEP sample was in 100% D_2O (light green). In hH16-dQE, the deviation from the SAXS R_g of the 100% D_2O condition was less prominent than the 40-60% D_2O, but importantly this condition was further from the match point and therefore a more convenient sample to examine experimentally. Similarly, dH16-hQEP measured in 20% D_2O displayed slight variation compared to the reference ensemble. Importantly, this was a measurable sample despite the expected incoherent background scattering due to the presence of a large amount of hydrogens present in solution.

The H36 scatterplots showed similar trends in the R_g scatterplots than the H16 (Figure 4.15). The most significant scatterplots were these for hH36-dQE, hH36-dP, dH36-hQE and dH36-

hQEP. While this was generally echoed by the averaged scattering profiles (Figure 4.13), the difference of the R_g scatterplot of dH36-hQE in 80% D_2O suggested the sample would contain additional structural information even though the averaged profiles of this sample seemed not to be strongly impacted by the D_2O level.

When we compared the scatterplots of hH16-dQE and hH36-dQE (figure 4.16), a notable difference could be seen at 100% D_2O. Due to the higher number of glutamine residues, hH36-dQE scatterplot presented larger deviations from the SAXS R_g values than hH16-dQE. Interestingly, due to the higher D_2O percentage to match out hH36-dQE with respect to hH16-dQE, the deviations of hH36-dQE at 40% D_2O were less pronounced than for hH16-dQE (pink dots). This effect was systematically observed for all D_2O levels when comparing the scatterplots for both proteins.

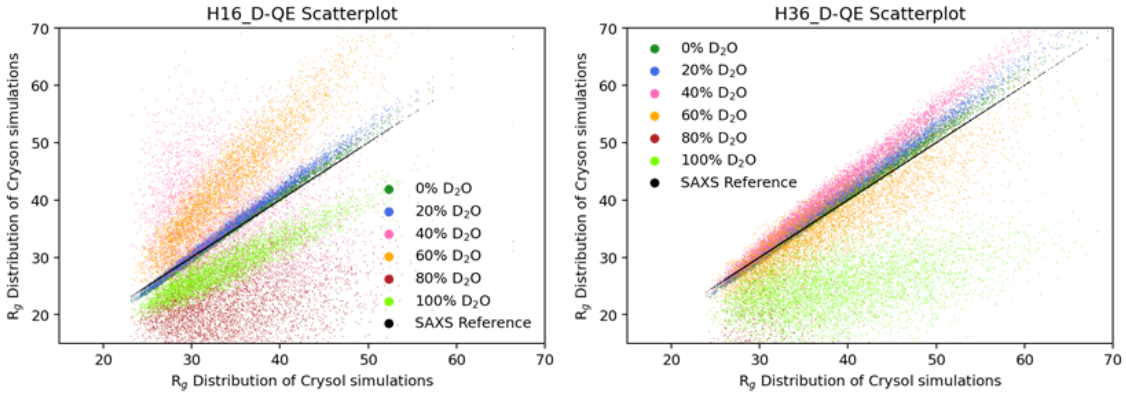


Figure 4.16: Comparison between scatterplots of the hH16-dQE (left) and hH36-dQE (right) ensembles. H16 show greater variation at 40-60% D_2O while the H36 ensembles shows a higher difference at 80-100% D_2O.

4.6 Optimal experimental conditions for SANS studies of huntingtin

From the scattering calculations, the optimal conditions for data collection could be proposed. Indeed, the possibility of computing SANS and SAXS profiles from atomistic models enabled a rational design of samples to be produced. Note that, in contrast with classical SANS studies, the project's approach did not necessarily search for contrast match points, but for experimental conditions providing a maximum of information. Furthermore, the stability of the protein and the concentration attainable would strongly modulate the choice of the optimal experimental conditions. Indeed, some highly informative conditions could be inherently low intensity, precluding their measurement in SEC-SANS.

While samples such as the hH-dQE and hH-dP constructs both showed a high difference compared to the SAXS dataset in conditions of 40-60% D_2O. The samples were, however,

unfeasible sample conditions for SEC-SANS experiments of Huntingtin samples produced by cell-free. The relatively small protein size (H16: 39.1 kDa, H36: 41.7 kDa) and limited labelling that was applied (1-3 residues, 12-27% of mass) would require high protein concentrations to achieve good signal-to-noise in SANS experiments. The aggregating nature of the H16 and H36 constructed and the low yield of CF expression both limited these experimental conditions. Therefore, protonated protein with deuterated labelling would be unsuitable for measurements at 40-60% D_2O. The inverse effect was true for deuterated protein samples containing protonated labelling. These labelling patterns would provide very low scattering intensity in 80-100% D_2O. Optimal SANS experiments should therefore have been focused on producing samples that were expected to yield measurable scattering profiles to subsequently combine them in the analysis to extract a maximum of structural information retained from the distinct labelling schemes. hH-dQE 100%, hH-dP 100% and dH-hQEP 0-20% should have been prioritized, as both profile differences from unlabelled samples and higher variation from the SAXS references according to the scatter plot have shown. These samples should have been combined with reference samples with high signal-to-noise ratios, such as fully protonated in 100% D_2O, and fully deuterated in 0% D_2O. Note that very different deuteration patterns could have been selected as optimal conditions for other proteins with a different amino acid composition and stability.

4.7 SMEARING OF THEORETICAL PROFILES

A final quality check of the theoretical ensembles was calculating the impact of the experimental momentum transfer (*Q*) resolution. The wavelength selection by neutron velocity, the pixel size, as well as the beam size and divergence result in data smearing. As a consequence, the intensity at each *Q*-value is smeared by a gaussian distribution whose standard deviation (σ) is called experimental resolution (Figure 4.17) (270). This effect, which is inherent for all scattering techniques, has an impact on the measured profiles and, consequently, on the structural interpretation of the data. Importantly, the experimental resolution in which data are measured depends on the beamline setup and, therefore, the impact of smearing will be different for distinct instruments. The smearing is calculated from the aperture (Source and Sample), collimation length, beam intensity and pixel size (270,271). This allows for the resolution of an instrument to be approximated from these values alone for theoretical approaches. To account for this phenomenon and to avoid eventual misinterpretation of the data, two solutions have been proposed by researchers:

(1) Desmearing the experimental data by deconvoluting them from the resolution gaussian. This approach is based on the assumption that, at the extremities of the curve, intensities are constant in *Q*.

(2) Smearing the theoretical models in order to appropriately accommodate the experimental bias. This approach is not limited by the Q-range and does not require approximations for calculation beyond min/max Q-values.

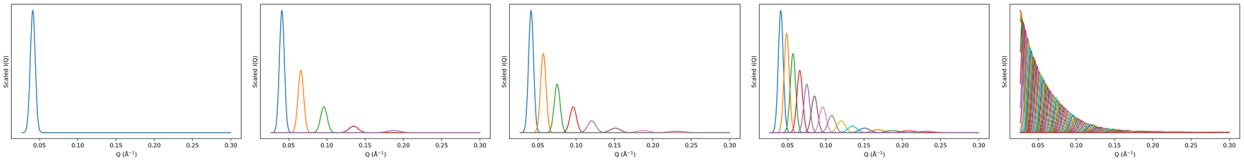


Figure 4.17: The resolution of SANS profiles is the resolution of the Q-angle calculated from the instrument geometry. If a single datapoint is considered, it can be represented as a gaussian (left). When multiple data points are considered, the Gaussians start overlapping (right). This overlap gives rise to the smearing effect of the experimental data.

In this thesis, the effect of smearing theoretical models was tested. For this, the experimental resolution profile from data measured at D22@ILL, was calculated by GRASP (270). GRASP used the following parameters for the calculation: Source aperture of 55x40 mm^2, Sample aperture of 11 mm diameter, collimation length of 5.6 m, source wavelength of 6 Å ± 10% and pixel size of 8x8 mm^2.

Using the "Pinhole smearing function" of SASview (https://www.sasview.org/docs/user/qtgui/Perspectives/Fitting/resolution.*html*), this smearing was applied to each of the theoretical scattering curves of the hH16-dQEP-100 ensemble. The greatest challenge of smearing the theoretical data is to extrapolate the resolution data at the low- and high-angle ends of the curves. This is where the Pinhole smearing function of SASview improved the calculation by forcing a positive theoretical σ-value in order to compute the resolution. The individually smeared profiles showed only minor changes, with a slight smoothing observed in profiles with the most prominent form factors (Figure 4.18).

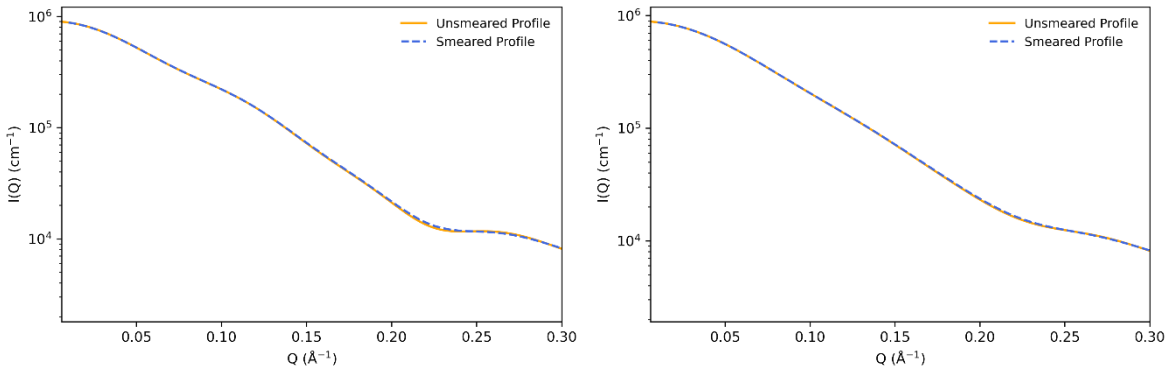


Figure 4.18: Individually smeared profiles of hH16-dQEP. The smeared profile (Blue) shows a slight smoothing of sharp features of a profile, which in the case of H16 is almost invisible with only a tiny offset seen at 0.22 Å^{-1} of the left profile.

The average of the smeared profiles of the ensemble was calculated and compared to that of the unsmeared ensemble. This was done for two different labelling patterns, hH16-dQE-100 and dH16-hQE-40, for which we had obtained the experimental resolution data (Figure 4.19). The smeared and unsmeared average theoretical profiles were virtually equivalent. Indeed, this observation was expected as our protein yields very smooth profiles.

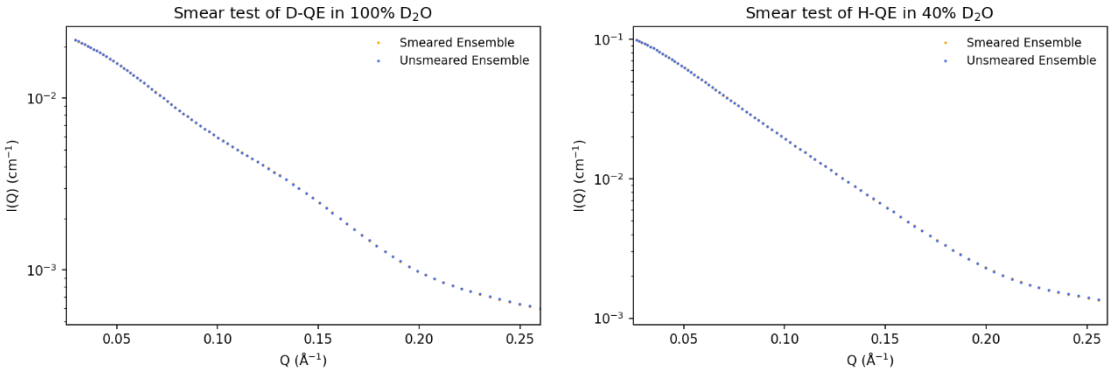


Figure 4.19: Two ensembles (hH16-dQE-100, left and dH16-hQE-40, right) were smeared and the average profiles were calculated and compared to that of the un-smeared ensemble.

To further validate the similarity of the smeared and unsmeared ensembles, I used EOM to fit the ensembles of both experimental SANS profiles (hH16-dQE-100 and dH16-hQE-40) using the smeared and unsmeared theoretical profiles (Figure 4.20). The main fitting parameter (χ^2) showed only very minor differences when using the smeared and unsmeared ensembles. Smearing the ensemble of hH16-dQE-100 changed the χ^2 of the fit from 2.07 to 2.12, and the fit of dH16-hQE-40 changed from 2.21 to 2.32.

The results demonstrated that the use of smeared curves was not significantly affecting our data analysis procedures and, as a consequence, the unsmeared profiles would be used for all subsequent analyses.

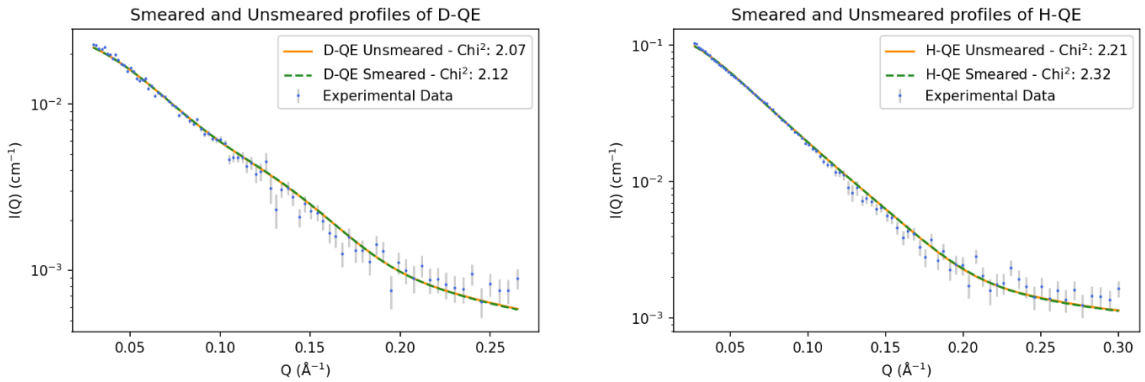


Figure 4.20: EOM Fit to the experimental profiles obtained at the D22 beamline at ILL using the smeared and unsmeared theoretical structural ensembles. The hH16-dQE-100 (left) sample was measured in 100% D_2O, while the dH16-hQE-40 (Right) sample was measured in 40% D_2O. As observed, the fits were not significantly improved when the smearing of the theoretical profiles.

5 EXPERIMENTAL MEASUREMENTS

During the project, SAXS data (Table 5.1) were measured at the Swing Beamline at Soleil (Paris, France) and SANS data (Table 5.3) were measured at the D22 beamline at ILL (Grenoble, France). All SAXS samples and the majority of SANS samples were measured using Size-Exclusion Chromatography coupled to SAXS/SANS (SEC-SAXS / SEC-SANS).

5.1 SAXS MEASUREMENTS

Table 5.1: SAXS samples collected at the Swing Beamline, Soleil.

SAXS Sample	% D_2O	R_g	Conc.	Experiment
hH16	0%	34.9 ± 0.2 Å	3.1 mg/ml	12/06-2021
	100%	35.9 ± 0.4 Å	2.9 mg/mL	Superdex 200, 5/150
hH36	0%	36.0 ± 0.3 Å	2.4 mg/mL	07/10-2021
				Superdex 200, 5/150
hH16	0%	33.7 ± 0.2 Å	2.6 mg/mL	09/02-2022
	100%	33.3 ± 0.2 Å	2.4 mg/mL	Superdex 200, 5/150
hH36	0%	33.3 ± 0.2 Å	3.3 mg/mL	17/07-2022
	100%	33.4 ± 0.3 Å	2.4 mg/mL	Superdex 200, 5/150

The SAXS profiles were collected in both H_2O and D_2O to accomplish two major goals. Firstly, the SAXS profiles of hH16 and hH36 were incorporated in the multi-dataset fitting done by EOM. Secondly, it allowed us to test the impact on the structure of using D_2O as a solvent compared to H_2O. The slight variation in the strength of bonds involving hydrogen or deuterium and the modifications in the hydration shell could affect the stability of secondary structures (272,273). Moreover, the different solvent properties between both isotopes could induce protein aggregation (274). The use of D_2O was not standard in synchrotrons and had to be discussed with the beamline responsible before experiments.

hH16 in H_2O and D_2O were measured twice. The first sample of hH16 protein was produced in *E. coli*, while the second one was produced by CF expression. Note that, although equivalent results were obtained from both hH16 samples, only data measured from the CF ones are discussed in this thesis. Once produced, the sample was split in two. One of them was buffer exchanged into a 100% D_2O buffer before injection into the column, which was previously equilibrated with the same buffer. The comparison of the SAXS profiles of hH16 obtained in H_2O and D_2O showed that both conditions provided virtually identical scattering profiles (Figure 5.1).

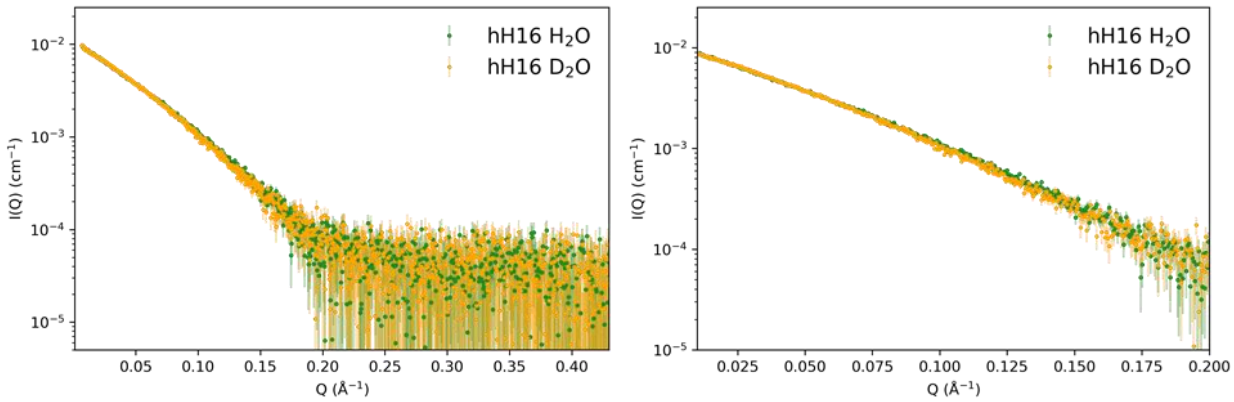


Figure 5.1: SAXS profiles of hH16 obtained in both H_2O and D_2O buffers at the Swing beamline, Soleil. The right figure showed the zoomed profiles, which highlighted that the low-resolution region of the two profiles was similar.

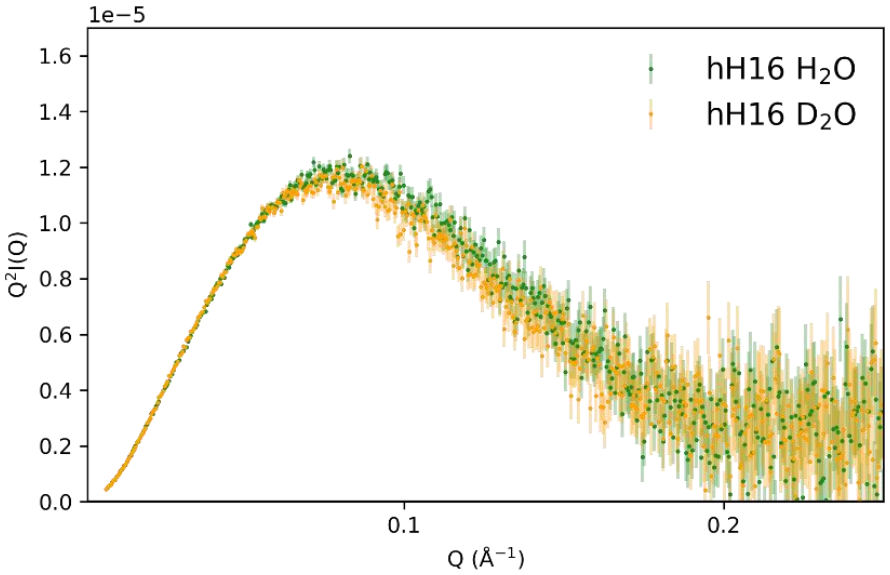


Figure 5.2: Kratky plot of hH16 measured in H_2O and D_2O buffers. Both profiles show a similar tendency towards primarily structured conformations due to the highly structured sfGFP-domain with some flexibility. This is concluded based on the initial bell-shape of the profile that does not returns to zero.

The initial points of the two profiles showed a sharp increase, which could be attributed to several factors such as over/under subtraction, radiation damage or the presence of larger species of the protein. This initial spike only affected a few points that were removed from subsequent analyses. The Kratky plots of the two samples were similar and showed a primarily structured protein (Figure 5.2). The sfGFP in the construct is globular, explaining the degree of structure observed in the Kratky representation. The R_g calculated from the two datasets were 33.7 ± 0.2 Å and 33.3 ± 0.2 Å for the H_2O and D_2O buffers, respectively, which were very similar to that of the averaged theoretical ensemble calculated by CRYSOL of 33.2 Å. The pair-wise distance distribution, *p(r)*, of both samples showed very similar D_{max} values (H16 H_2O: 110 Å and H16 D_2O: 107 Å), indicating an important degree of flexibility (Figure 5.3).

The *p(r)* function displayed a steep decay towards their D_{max} and the values were consistent with these from other samples.

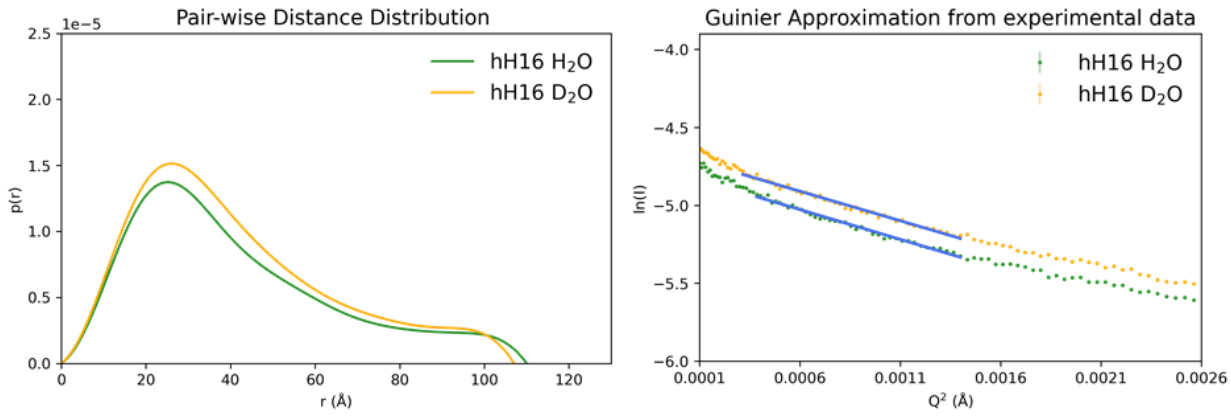


Figure 5.3: Left: Pair-wise distance distribution of the hH16 SEC-SAXS samples obtained in H_2O (Green) and D_2O (Orange). Right: Guinier approximation visualized on the experimental data from H_2O (Green) and D_2O (Orange).

The same measurements were performed for hH36, a pathogenic construct of huntingtin. Note that constructs with longer poly-Q tracts have shown a higher tendency towards aggregation in previous studies (66), and the impact of D_2O could be higher in such conditions.

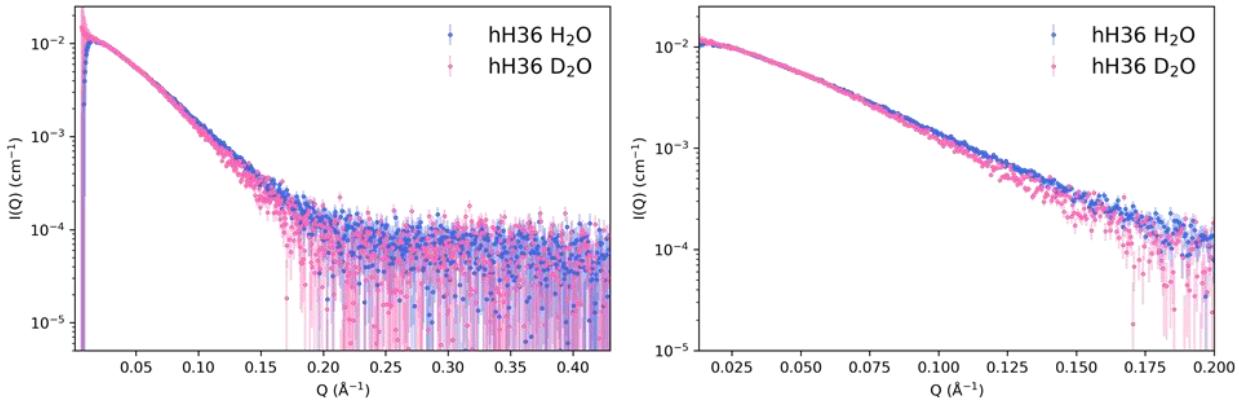


Figure 5.4: SAXS scattering profiles of hH36 measured in buffers containing H_2O and D_2O. The two scattering profiles are very similar and when the initial points are omitted from the Guinier plot, the derived R_g values are almost identical.

The hH36 profiles in H_2O and D_2O showed signs of low data quality at low Q-range (<0.01 $Å^{-1}$). The hH36 in D_2O show slightly higher deviation at lower Q-range, but this was likely a problem of buffer subtraction, due to very few appropriate buffer frames. When the initial points of the dataset were ignored (Figure 5.4, right), the slope of the scattering patterns was very similar. In the Q-range 0.075 – 0.125 $Å^{-1}$ the slope was slightly different, but beyond 0.175 $Å^{-1}$ the scattering data fully overlapped. The Kratky plot was similar to that of the hH16 construct, showing a structured part with some flexibility (bell shaped Kratky curve that did

not return to zero) (Figure 5.5). Both the slight upturn at low Q of the scattering profile and the lower area-under-the-curve of the Kratky plot of the D_2O sample compared to that of the H_2O sample were consistent with the presence of a small fraction of aggregates in the D_2O sample. D_2O buffer could have had a small effect on the stability of the H36 protein samples. The shape of the curves (scattering and Kratky) were still generally the same, so the impact was low.

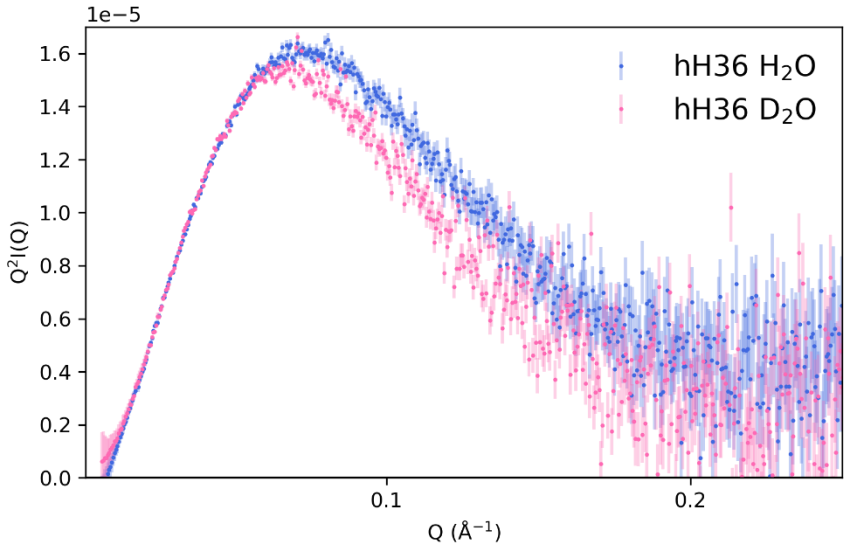


Figure 5.5: Kratky plot of the two hH36 SEC-SAXS datasets. Both show a structured protein encompassing flexibility. This is evident from the initial bell-shape which then did not return to zero. This is consistent with the flexible Huntingtin exon-1 connected to structured sfGFP.

The R_g calculated from the two profiles, showed very similar values (R_g - hH36 H_2O: 33.3 $\pm$ 0.2 Å and hH36 D_2O: 33.4 $\pm$ 0.2 Å). Also, the D_{max} values obtained from the pair-wise distance distribution function of the scattering patterns were very similar (D_{max} - hH36 H_2O: 126 Å and hH36 D_2O: 123 Å) (Figure 5.6 Left: *p(r)*-distribution and Right: Guinier region).

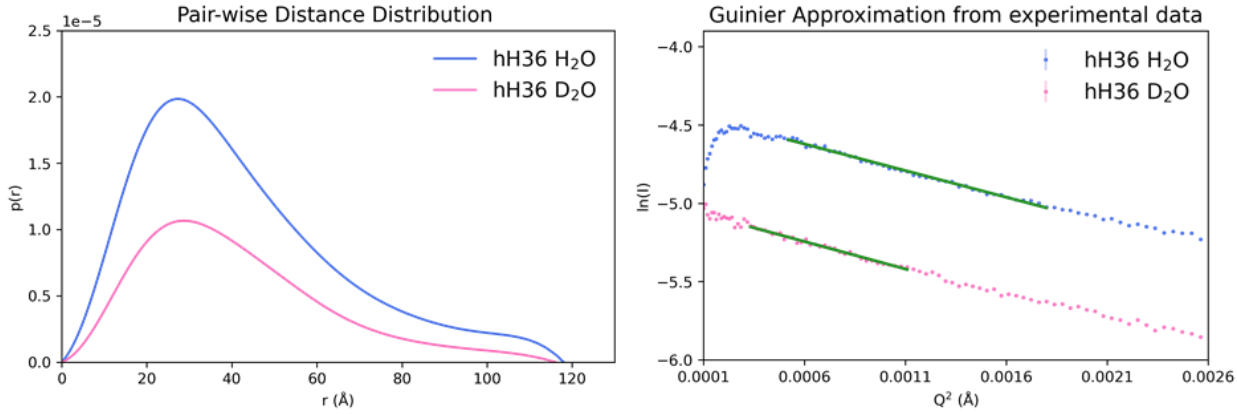


Figure 5.6: Visualization of the pair-wise distance distribution (Left) and Guinier region (Right) of hH36-0 and hH36-100 SAXS data. Both showed similar profiles between the two samples with only a slight intensity difference

The two samples showed similar R_g, D_{max}, Kratky plot, and slope of scattering data suggesting that the hH36 protein constructs could be measured in D_2O with only a slight impact of the buffer. Because of the low data quality at the low Q-range of both datasets, the samples were not used for ensemble fitting.

In October 2021, the initial hH36 dataset in H_2O was measured at Soleil. The experimental profile had a better data quality, especially at low Q-range compared to that of the above-mentioned measurement (Figure 5.7). The R_g and D_{max} calculated from this dataset were 36.0±0.3 Å and 136 Å, respectively. These values were slightly larger than previously described hH36 SEC-SAXS samples, likely caused by the better data quality at low Q-range (higher signal-to-noise and no sharp increases or decreases), and this measurement was subsequently used for structural analyses of hH36 (chapter 6.3).

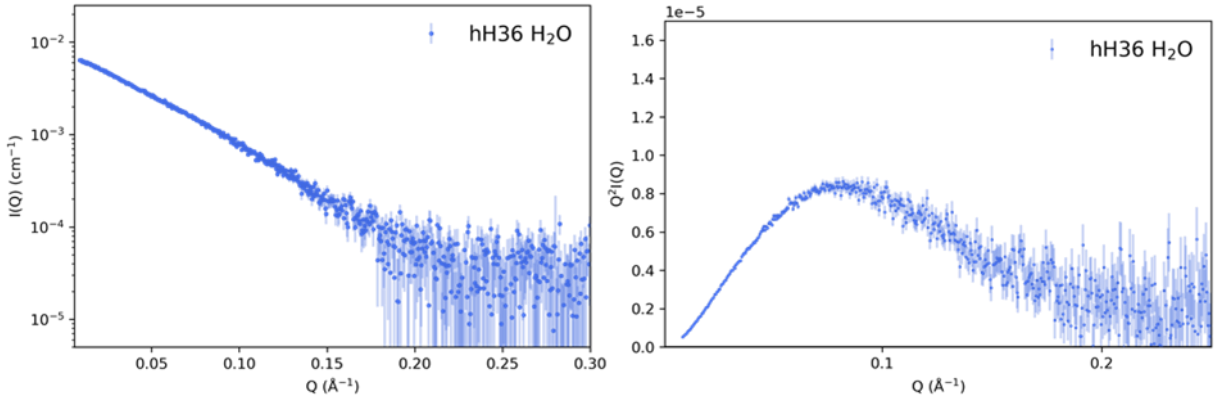


Figure 5.7: SEC-SAXS profile of hH36 in H_2O measured at the Swing beamline, Soleil. The structural parameters derived from this profile are very similar to these of the previous measurements. These data was used for the structural analyses.

The ensemble of the results showed that the structure and the oligomeric state (monomer) of hH16 and hH36 were not affected by the presence of D_2O, at least at the resolution of SAS and when measured using SEC-SAXS. These results were key, as they enabled the subsequent simultaneous analysis of the SAXS and SANS data (Chapter 6).

Experimental details of the two chosen experimental datasets of hH16 and hH36 used for structural analyses are outlined in table 5.2 using the SAS template for publications (275). The quality of the data was validated by modelling them with the EOM (183). Ensemble fitting the SAXS data of hH16 and hH36 (Figure 5.8, Left) showed that both SAXS datasets could be described by the structures built with our modelling strategy (chapter 4) with χ^2 values close to 1 (hH16 χ^2: 0.98 and hH36 χ^2: 0.93).

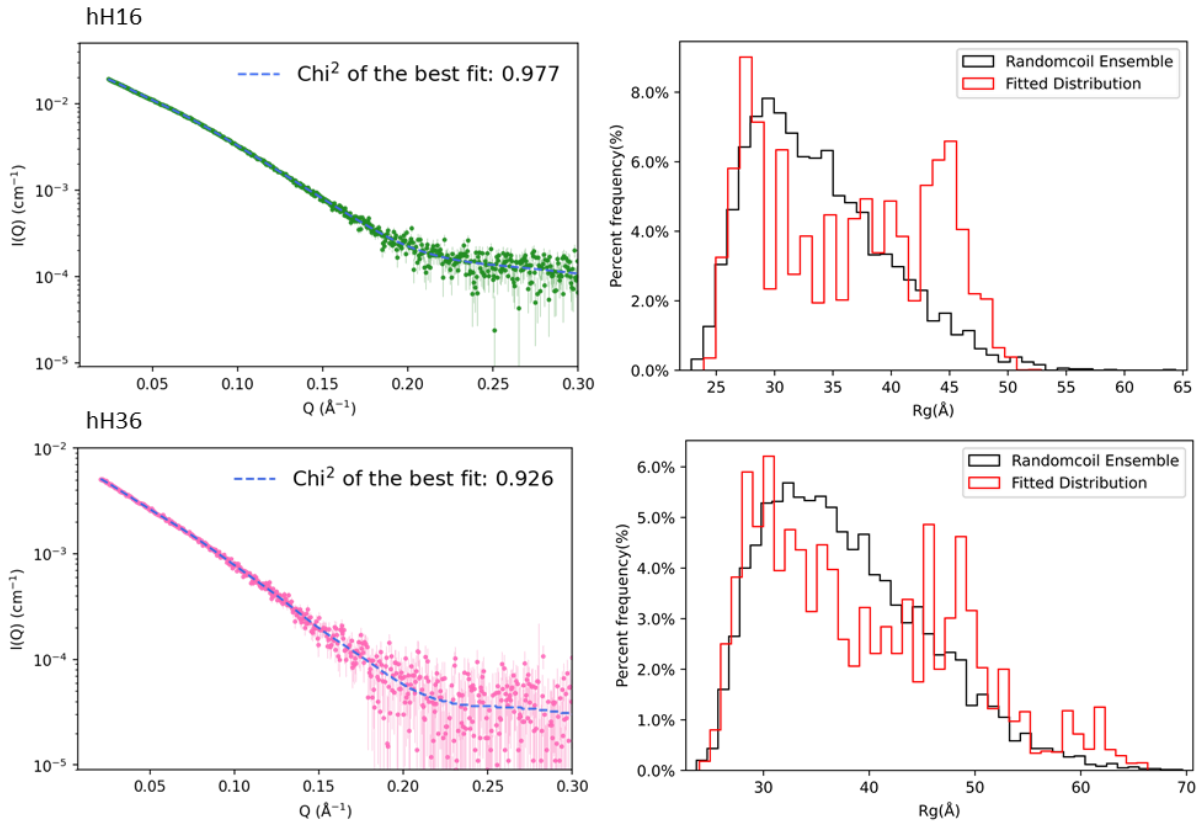


Figure 5.8: Left - EOM Fit of SAXS data of hH16 and hH36. Right – R_g distribution figures of the chosen structures from the EOM fits. The red distribution was obtained from the chosen structures, while the black distribution was the overall distribution from all structures of the ensemble.

The R_g distribution of the sub-ensemble (Figure 5.8, Right), chosen by the fitting routine, was extremely broad indicating that both proteins are highly flexible particles in solution (139). While the fit was excellent, the structural analysis of SAXS data, did not grant any high-resolution structural insight, into the nature of the disordered Huntingtin Exon-1 in their pathogenic and non-pathogenic forms.

Table 5.2: SAXS data, data-collection and analysis presented as outlined in the 2023 template for reporting of biomolecular SAS data (275).

(a) Sample details.

Organism	Human	
Source	Cell-free protein expression (Lysate: *E. coli BL21 star (DE3)::RF1-CBD3*)	
Scattering particle composition	H16-sfGFP-His6	H36-sfGFP-His6
Protein	hH16	hH36
DNA/RNA		
Stoichiometry of components	n.a.	n.a.
Sample environment/configuration	SEC-SAXS	SEC-SAXS
Solvent composition	20 mM BisTris, 150 mM NaCl, pH 6.5	

Sample temperature (℃)	20	20
In-beam sample cell	Horizontal Ø1,5 mm open quartz capillary	
Sample concentration (s) (mg/mL)	2.6	2.4

(b) SAS data collection

Data-acquisition/reduction software	Foxtrot Software
Source/instrument description	Size-exclusion Chromatography SAXS, EigerX4M Detector (in-vacuum)
Measured Q-range (Q_{min} - Q_{max})($Å^{-1}$)	0.004 – 0.550
Method for scaling intensities	Absolute scaling (cm^{-1}) referenced to water
Exposure time(s), No. of exposures	1s frames recorded over sample elution. 1s x 900 frames
Additional relevant details	Protein elution peak was determined using Chromix and buffer frames were subtracted from the protein frames. ~15-20 sample frames were selected per sample depending on statistics.

(c) SAS-derived structural parameters

Method(s)/software	Primus, AUTORG and GNOM (ATSAS 3.0.2: (113))	
	hH16	hH36
Guinier analysis		
I(0) ± σ (cm^{-1})	$0.0082 \pm 3e^{-5}$	$0.0062 \pm 3e^{-5}$
R_g ± σ(Å)	33.7 ± 0.2	36.0 ± 0.3
qR_g range (datapoint range)	0.68 – 1.29 (36-76)	0.57 – 1.30 (22-66)
Linear fit assessment (AUTORG fidelity)	0.00	0.11
PDDF/p(r) analysis		
I(0) ± σ (cm^{-1})	$0.0080 \pm 2.9e^{-4}$	$0.0063 \pm 3.3e^{-4}$
R_g ± σ(Å)	33.8 ± 0.01	39.2 ± 0.01
D_{max} (Å)	110	136
Q-range ($Å^{-1}$)	0.020 – 0.237	0.016 – 0.221

(d) Scattering particle size.

Method(s)/software	Primus (113,276)	
Volume Estimates ($Å^3$)		
Porod volume V_P (ratio to M)	63053 (1.61)	65144 (1.56)
Molecular mass M estimates (Da)		
From chemical composition	39112	41675

From c-independent method (Bayesian inference, range with % confidence)	41500 – 45200 (93%)	43300 – 47150 (94%)
From I(0)/c (ratio to expected)	11355 (0.29)	8586 (0.21)
Partial specific volume v(cm^3/g)	0.743	0.743
Contrast Δp (10^{10} cm^{-2})	2.809	2.809
From SAS-independent measure	n.a.	n.a.

(e) Data and Model Deposition.

	hH16	hH36
	TBD	TBD

5.2 Optimization of SANS Experimental setup

The initial neutron experiments served to explore and optimize the sample conditions and experimental setup. Batch samples of hH16-0, hH16-100, hH16-dP-100, hH16-dQEP-20 and hH16-dQEP-100 were measured to probe the data quality. During the initial batch tests, the D22 beamtime was equipped with only 1 detector and the samples were measured at 2, 4 and 8 meters from the detector. As a consequence, the buffers were measured separately at the same detector distances. Only data measured at 2 and 8 m distances were used, reduced and subtracted separately and then merged to form a dataset with a Q-range from 0.006 to 0.46 $Å^{-1}$. Unfortunately, all the samples showed evident problems at the low-Q range (Figure 5.9). Concretely, the hH16-0 and the hH16-dP-100 samples showed sharp upturns in this region.

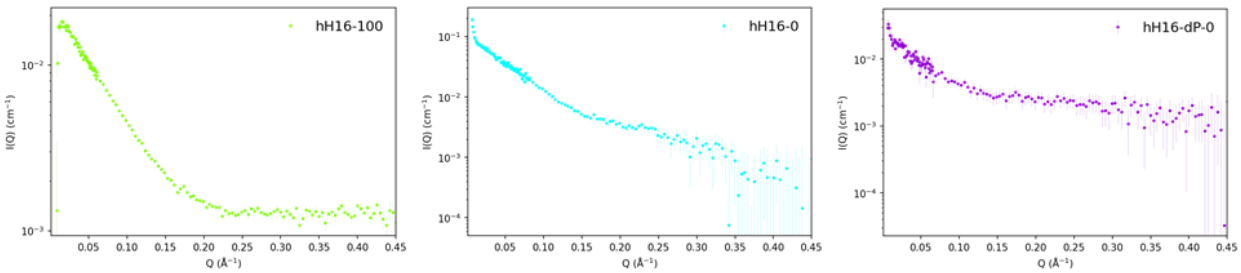


Figure 5.9: SANS data recorded in batch mode. All three samples show problems at low Q-range.

Conversely, the hH16-100 showed a sharp decrease at low Q-values. This decrease was most probably caused by a buffer subtraction problem.

In the same beamtime, SEC-SANS was attempted, using the fully protonated hH16 in 100% D_2O, to test whether the aggregation and/or buffer subtraction problem could be alleviated when using this alternative measuring mode. As described in the Materials and Methods (chapter 9.8), the sample elution of the SEC column was paused during the measurement, when the UV intensity reached a peak maximum in UV intensity. For the initial SEC-SANS test the detector was also manually moved between 2 m and 8 m. The two datasets were reduced and buffer-subtracted before finally merging them (Figure 5.10).

*Table 5.3: SANS samples measured at the D22 beamline at ILL, Grenoble in chronological order. *: Data quality was low and they were not used for subsequent modelling. **: indicate a protein behavior that could not be explained by our synthetic ensemble. (2nd) & (3rd). indicating that the sample was remeasured. H16 and H36 samples are coloured in white and blue, respectively. Orange fields indicate samples measured in Batch mode. Dilution factor of protein samples during SEC-SANS was x7 when using Superdex 10/300 columns and x2 when using Superdex 5/150 columns.*

SANS Sample	% D_2O	R_g	Conc.	Exposure time	Experiment
hH16 Batch	0%* 100%*	Aggregation Aggregation	1.5 mg/mL 0.4 mg/mL	1 hour 30 minutes 1 hour 30 minutes	Test_3129 (22/09-20)
hH16-dP Batch	100%*	Aggregation	0.2 mg/mL	1 hour 30 minutes	Test_3129 (22/09-20)
hH16	100%*	31.5 ± 1.1 Å	1.5 mg/mL	50 minutes	Test_3129 (22/09-20) (Superdex 200, 10/300)
hH16 (2nd)	20% * 100%	21.7 ± 5.8 Å 35.1 ± 1.2 Å	6.0 mg/mL	2 hours 1 hour 25 minutes	Test_8-03-1020 (04/02-21) (Superdex 75, 10/300)
hH16-dP	100% **	26.1 ± 0.4 Å	2.3 mg/mL	6 hours	Test_8-03-1020 (04/02-21) (Superdex 75, 10/300)
hH16-dQEP Batch	20%* 100%*	47.9 Å 30.3 Å	0.8 mg/mL	2 hours 30 minutes	Test_8-03-1020 (04/02-21)
hH16-dQE	0% * 40% 100%	39.7 ± 1.6 Å No data 30.8 ± 0.7 Å	4.6 mg/mL	2 hours 31 minutes 2 hours 1 hour 5 minutes	Test_9-13-984 (22/06-21) (Superdex 75, 10/300)
dH16	0% 40%	40.2 ± 0.3 Å 39.2 ± 0.7 Å	4.8 mg/mL	2 hours 5 minutes 1 hour 8 minutes	Test_8-03-1050 (21/09-21) (Superdex 200, 5/150)
dH16-hQE	0% 40%	38.0 ± 0.4 Å 37.3 ± 0.6 Å	4.5 mg/mL	1 hour 7 minutes 1 hour 10 minutes	Test_8-03-1050 (21/09-21) (Superdex 200, 5/150)
hH36	100%	32.6 ± 0.5 Å	2.3 mg/mL	1 hour 21 minutes	Test_8-03-1050 (21/09-21) (Superdex 200, 5/150)
hH36-dQE	100% *	27.0 ± 1.1 Å	1.7 mg/mL	8 hours 3 minutes	Test_8-03-1050 (21/09-21) (Superdex 200, 5/150)
hH16 (3rd)	100%	28.5 ± 1.0 Å	4.2 mg/mL	1 hour 15 minutes	DIR_279 (11/04-2023) (Superdex 75, 5/150)
hH16-dQEP	100% **	24.0 ± 0.3 Å	2.2 mg/mL	2 hours 38 minutes	DIR_279 (11/04-2023) (Superdex 75, 5/150)
hH36-dQEP	100% **	28.1 ± 0.2 Å	6.0 mg/mL	2 hours 56 minutes	DIR_279 (11/04-2023) (Superdex 75, 5/150)
hH36-dQE (2nd)	100%	28.0 ± 0.9 Å	0.9 mg/mL	2 hours 42 minutes	DIR_279 (11/04-2023) (Superdex 75, 5/150)
dH36-hQE	0% 40% *	28.4 ± 0.8 Å No data	2.2 mg/mL	2 hours 39 minutes 2 hours 40 minutes	DIR_279 (11/04-2023) (Superdex 75, 5/150)
dH36	40%	28.6 ± 3.4 Å	2.5 mg/mL	7 hours	8-03-1075 (18/09-2023) (Superdex 75, 5/150)
hH16-dP (2nd)	80% ** 100% **	27.8 ± 0.4 Å 30.7 ± 0.2 Å	4.7 mg/mL	1 hour 59 minutes 2 hours 48 minutes	8-03-1075 (18/09-2023) (Superdex 75, 5/150)
hH36-dP	80% ** 100% **	35.3 ± 0.7 Å 34.5 ± 0.3 Å	5.5 mg/mL	1 hour 57 minutes 3 hours 5 minutes	8-03-1075 (18/09-2023) (Superdex 75, 5/150)

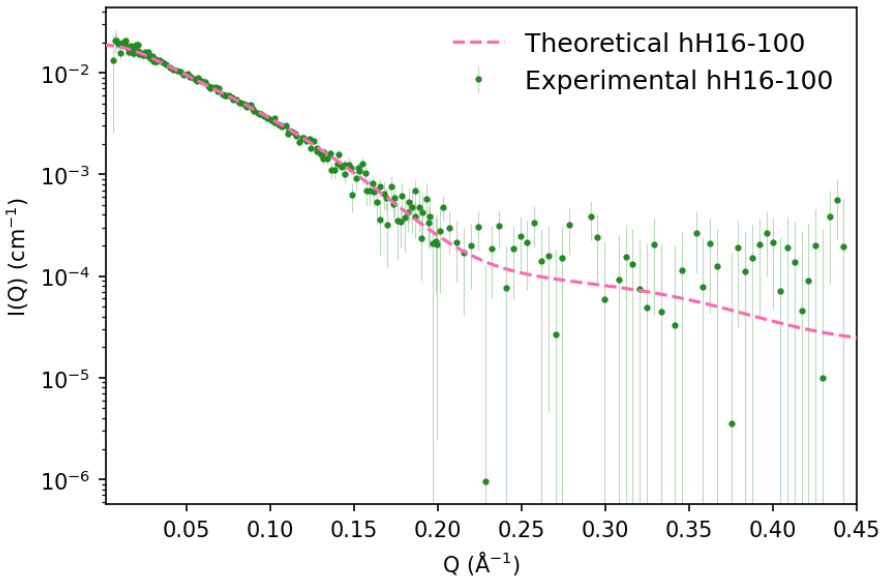


Figure 5.10: First SEC-SANS dataset for hH16 in 100% D_2O obtained at the D22 beamline, ILL. The buffer subtraction was better than that of the batch samples and showed no obvious signs of aggregation or other concentration effects. Additionally, it superimposes well with the averaged profile of the theoretical ensemble.

The scattering profile measured in the SEC-SANS mode presented a better defined small-angle region when compared with the one measured in batch, which suggested that SEC-SANS overcame the challenges of the aggregation-prone nature of the protein. The SEC-SANS profile was compared to that of the averaged theoretical ensemble of scattering profiles calculated for hH16-100 by scaling and superimposing them (Figure 5.10). Importantly, the two profiles were very similar in the complete *Q*-range. In addition, the R_g values were equivalent (Experimental R_g 31.5 $\pm$ 0.6 Å, theoretical R_g 32.7 Å). Due to the higher quality of the data obtained from the SEC-SANS profiles, this measuring mode was chosen for the subsequent SANS experiments. Note that the SEC-SANS mode was more time-consuming because of the lower concentration and column run.

The UV absorbance was followed in order to visualize possible changes in the sample during the experiment, such as diffusion of protein from the sample cell or precipitation during the experiment. After the flow-pump was paused to allow sample exposure, a small shift in the UV trace was observed (Figure 5.11, Left). This effect was even more pronounced when the pump was re-engaged: the UV absorbance spiked above 0.1 before returning to the UV-level of the buffer. This observation was attributed to the sudden change in pressure in the cuvette and was observed in all subsequent SEC-SANS profiles obtained with the D22 SEC-SANS setup. Importantly, during the sample exposure, the UV absorbance was relatively stable, although a slight decay was observed over the course of measurement. This slight decay could be caused

by: 1) an electronic drift of the measured spectrum during the long exposure time, or 2) the protein diffused out of the sample cuvette due to the release of the internal pressure of the cell.

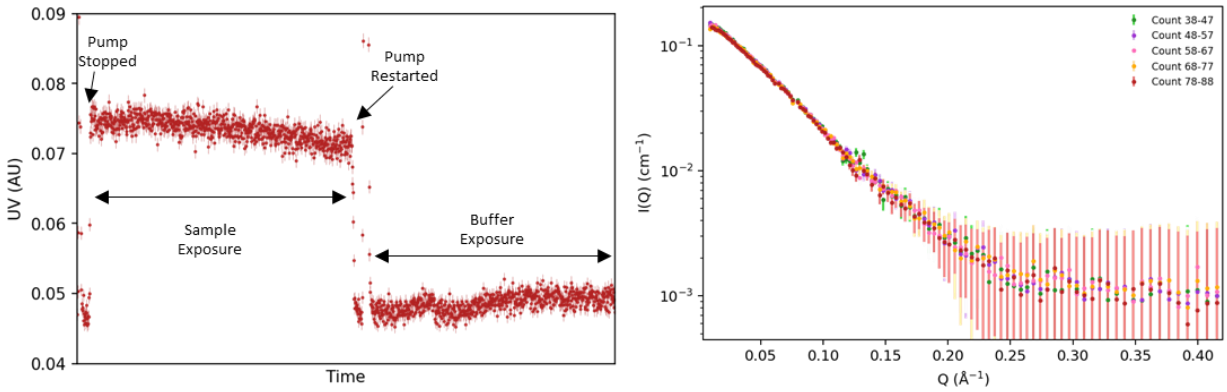


Figure 5.11: Left: Chromatogram of the UV absorbance obtained during the 2nd hH16 SEC-SANS experiment during the Test_8-03-1020. A slight decrease of UV absorbance was observed over sample exposure (85 min). Sharp spikes were observed when the flow was stopped and re-started. Right: 10 profiles averages (5 min) distributed over the 85 minute measurement period showing no structural changes of the protein sample from the beginning to the end of the exposure period. Flow was stopped just before count 38, where the maximum of the elution peak was observed.

In order to investigate if the protein sample evolved during the long neutron exposure time (>one-hour experiments), the experimental datasets were batched into smaller segments distributed over the whole exposure time. The smaller segments were buffer-subtracted averages of only 10 profiles of 30 seconds exposure each and the resulting scattering curves were compared (example of dataset seen in figure 5.11, Right). From these intermediate 5-minute exposures, the profiles did not show any systematic variation in slope or initial intensity (Figure 5.11, Right). Note that changes observed at 0.12-0-14 $Å^{-1}$ correspond to a low signal to noise region due to the overlap between the 2 m and 8 m detector-distance data merging.

Table 5.4: R_g and D_{max} values of the intermediate averages of the hH16 SEC-SANS sample.

Sample step	1	2	3	4	5
R_g	35.4±0.5 Å	33.1±0.5 Å	34.0±0.4 Å	33.4±0.6 Å	33.8±0.5 Å
D_{max}	102 Å	104 Å	105 Å	106 Å	105 Å

To further validate the sample's stability during measurements, both R_g and D_{max} were calculated for each of the 5-minute interval. The 5-minute profiles showed no systematic changes over the course of the experiment (Table 5.4). The small variance between the averages arose from the relatively low data quality obtained due to the short exposure considered. Therefore, the UV absorbance and the partial measurements lead to the conclusion that the huntingtin sample did not oligomerize or aggregate in the cell during the SEC-SANS experiment. After the analyses of the data measured during this beamtime, all samples were

measured in SEC-SANS mode. Moreover, the UV absorbance along the experiment was always monitored and the analysis of the consecutive short exposure profiles was systematically performed in order to identify potential aggregation or oligomerization of the huntingtin during experiments.

5.3 MEASURED SEC-SANS DATA.

During the project, 24 protein samples were measured using SEC-SANS (Figure 5.12 & 5.13). The datasets were recorded during the six beamtime slots granted at the ILL D22 beamline. Some of the 24 experimental datasets recorded were repeated measurements for the same labelling patterns, performed to improve data in specific conditions. The hH16-100 sample was measured three times, this to serve as an experimental control as well as to improve the final dataset. The hH16-dP-100 sample was measured twice to investigate the reproducibility of an artefact observed during analysis (see below). 19 datasets had high enough data quality to attempt ensemble fitting using EOM. A few datasets were discarded prior to fitting due to their low signal-to-noise ratio. The adapted table of datasets (table 5.5) lists the scattering profiles that were used in structural analyses. Each sample was described by:

- CF Expression observations of note (*i.e.* sample concentrations)
- Scattering profile characteristics (*i.e.* R_g, D_{max} etc.)
- EOM Fit (Fit and R_g distribution of the fitted sub-ensemble)
- Additional observations and adaptability of the dataset into the multiple fitting

The unfitted data portrayed the large differences that could be observed between experimental SANS datasets. Several factors impacted the quality of the datasets such as labelling pattern, protein concentration, experimental exposure time, contrast of the sample, and incoherent background scattering. These factors collectively contributed to the need for evaluating each dataset independently. In the following sections I present the SANS data measured according to their deuteration pattern.

Table 5.5: List of scattering profiles with adequate signal-to-noise ratio for subsequent structural analyses. *H16 and H36 samples are coloured white and blue, respectively.*

SANS Sample	% D_2O	R_g	Conc.	Exposure time
hH16	100%	28.5 ± 1.0 Å	4.2 mg/mL	1 hour 15 minutes
hH16	20%	21.7 ± 5.8 Å	6.0 mg/mL	2 hours
hH16-dQE	100%	30.8 ± 0.7 Å	4.6 mg/mL	1 hour 5 minutes
hH16-dQE	0%	39.7 ± 1.6 Å	4.6 mg/mL	2 hours 31 minutes
dH16	0%	40.2 ± 0.3 Å	4.8 mg/mL	2 hours 5 minutes
dH16	40%	39.2 ± 0.7 Å	4.8 mg/mL	1 hour 8 minutes
dH16-hQE	0%	38.0 ± 0.4 Å	4.5 mg/mL	1 hour 7 minutes
dH16-hQE	40%	37.3 ± 0.6 Å	4.5 mg/mL	1 hour 10 minutes
hH36	100%	32.6 ± 0.5 Å	2.3 mg/mL	1 hour 21 minutes
hH36-dQE	100%	28.0 ± 0.9 Å	0.9 mg/mL	2 hours 42 minutes
dH36	40%	28.6 ± 3.4 Å	2.5 mg/mL	7 hours
dH36-hQE	0%	28.4 ± 0.8 Å	2.2 mg/mL	2 hours 39 minutes

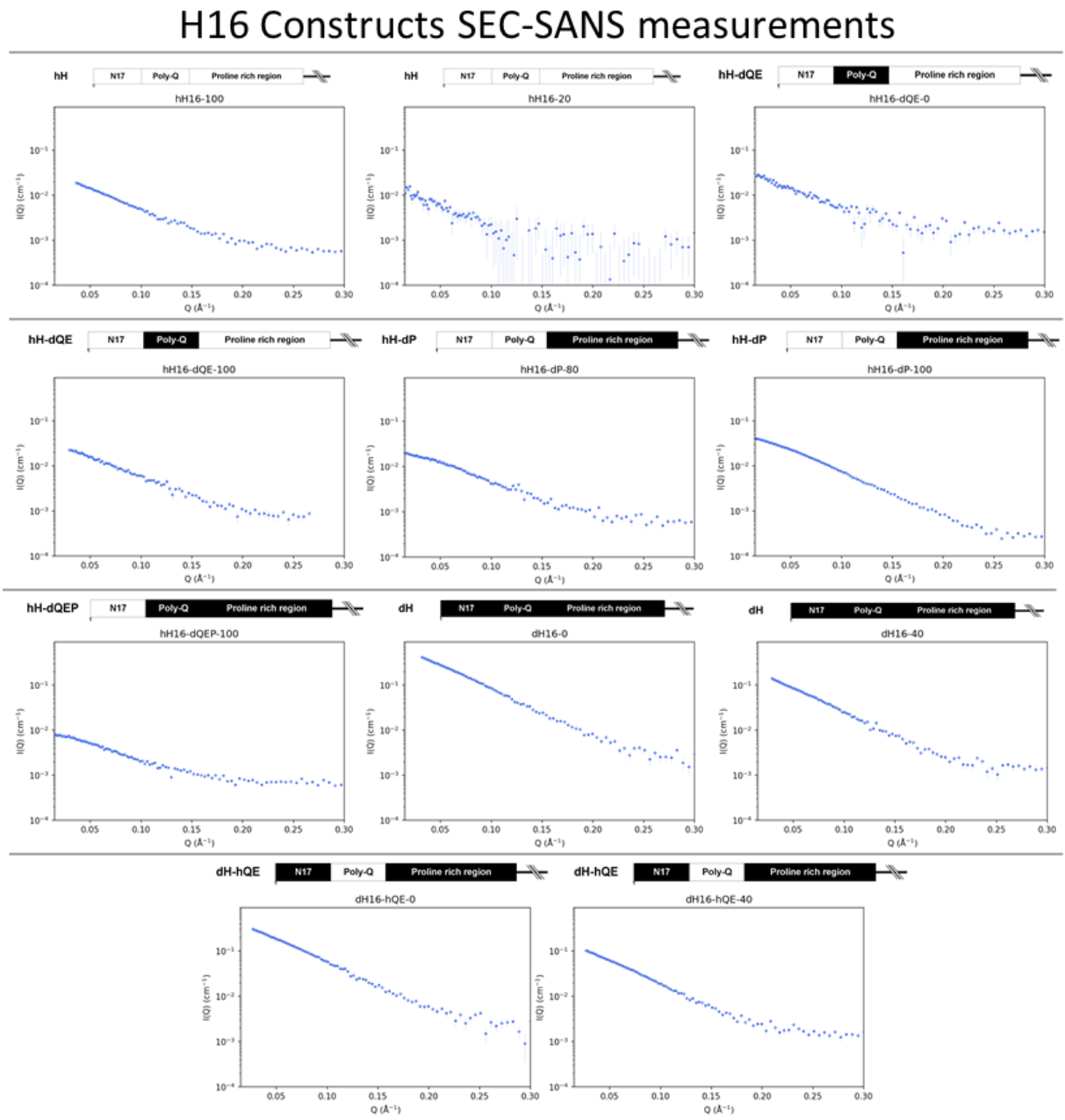


Figure 5.12: SEC-SANS scattering profiles obtained of H16 constructs. All neutron measurements were performed at the D22 beamline (ILL). Intensity is plotted at the same scale for all profiles, showing the difference in I(0) between datasets. The deuteration pattern is indicated as a cartoon on top of all the profiles with black regions indicating deuteration.

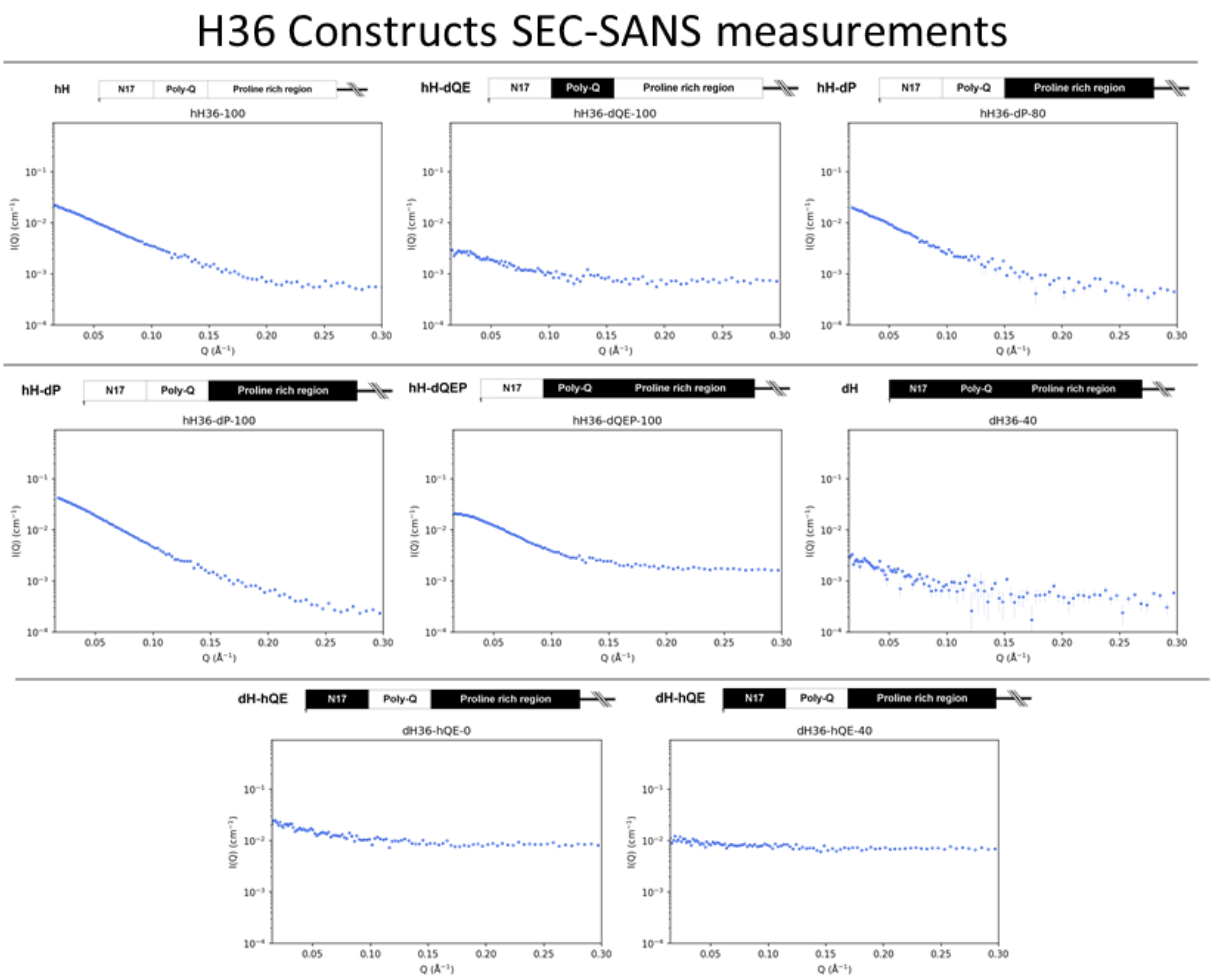


Figure 5.13: SEC-SANS scattering profiles obtained of H36 constructs. All neutron measurements were performed at the D22 beamline (ILL). Intensity is plotted at the same scale for all profiles, showing the difference in I(0) between datasets. The deuteration pattern is indicated as a cartoon on top of all the profiles with black regions indicating deuteration.

5.3.1 hH16

Since the fully protonated hH16 sample was used as sample control during measurements, it was measured several times. This sample was measured in two different D_2O levels; 100% and 20%. While the fully protonated protein could be produced using conventional *E. coli* expression, the samples were instead produced using the CF expression method to match the samples of labelled protein. The unlabelled protein was produced at a higher yield than samples requiring the Q/E deuteration (see chapter 3.2), yet sample concentration was limited to 4 – 6 mg/mL to match other samples of labelled protein. An additional concern in regards to the protein concentration was induced aggregation, since previous experience in the group had suggested that concentrating protein above 6 mg/mL could induce oligomerization/aggregation of Huntingtin exon1 samples, especially when increasing poly-Q length.

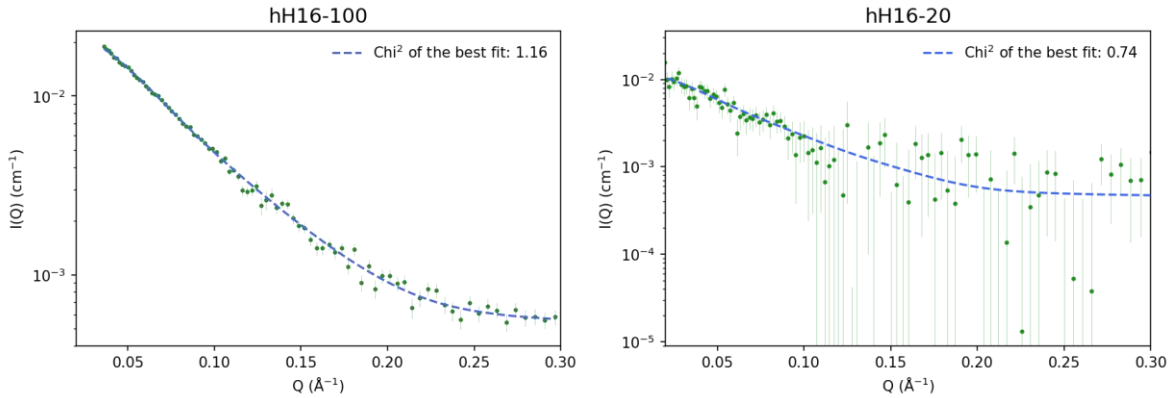


Figure 5.14: Left: EOM sub-ensemble fitted to hH16-100. An excellent agreement between the sub-ensemble and the experimental dataset was observed (χ^2= 1.16). Right: EOM sub-ensemble fitted to hH16-20. Very good fitting value (χ^2=0.74) observed while variation from the fitting line was present. The low χ^2 was likely caused by the large experimental error of the scattering pattern.

The R_g was calculated to 28.5 ± 1.0 Å from the Guinier region and the D_{max} was estimated to 110 Å from the *p(r)*. The latest dataset of hH16-100 (from the DIR_279 Experiment) was used to fit an ensemble of disordered atomistic structures of hH16 by EOM. This fitting of the ensemble to the experimental data displayed an excellent agreement with a χ^2 value of 1.16 (Figure 5.14, Left). Another visible way to estimate the data quality of an experimental dataset was to look at the R_g-distribution obtained from the sub-ensembles selected by EOM fitting. The R_g-distribution of the fitted sub-ensemble was relatively narrow and centered around a R_g of 28 Å, suggesting a more compact ensemble than that of the initial pool (Figure 5.15, Left).

When comparing the hH16-20 sample (Figure 5.14, Right) to the hH16-100 sample, it was clear that the signal-to-noise ratio of the 20%-D_2O data was significantly lower. Note that, while the hH16-100 sample was measured at 4.2 mg/mL and the hH16-20 sample was measured at 6.0 mg/mL. The signal-to-noise of the 100% D_2O data was significantly higher due to two factors: 1) The overall background scattering was lower due to the proportion of H_2O in the solution buffer; 2) The SLD difference between the sample and the buffer was higher (matching point of hH16: 46.0% D_2O). The proximity to the matching point dramatically reduces the contrast and the overall signal to noise, as shown in chapter 4 with the simulated data.

Not only was the variation of the intensity values much higher, but the low *Q* range exhibited fluctuations significantly larger than the error margin and as a result, the Guinier region of the dataset was poorly described and the resulting R_g derived from the profile was not very precise, 21.7 ± 5.8 Å. The EOM could fit the dataset with a significantly lower χ^2-value than other

SANS or even SAXS dataset. However, this very low χ^2 is due to the large error bars associated to this highly noisy dataset (Figure 5.14, Right). The R_g-distribution of the hH16-20 sub-ensemble (Figure 5.15, Right) also showed a broader distribution compared to the hH16-100 and with a clear bimodal shape (Figure 5.15, Left).

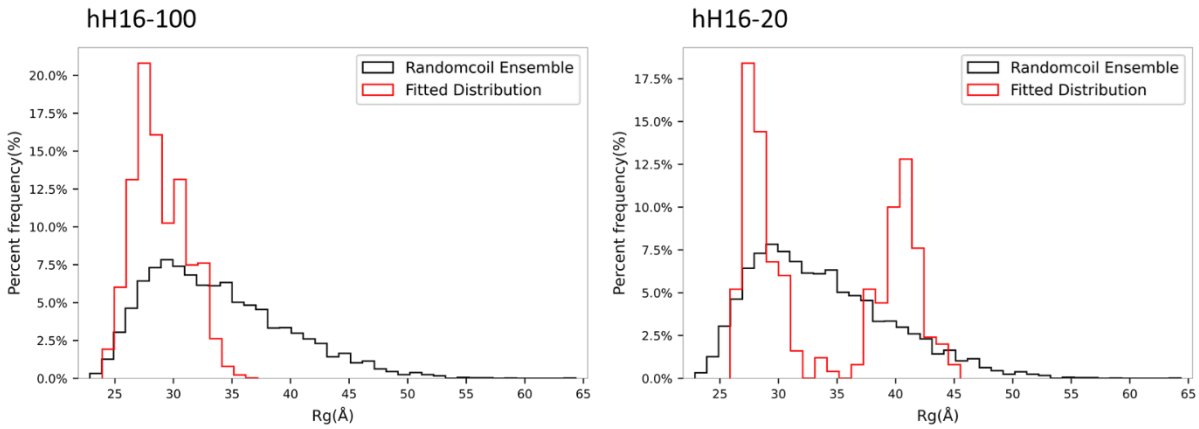


Figure 5.15: R_g distribution of the initial (pool) ensemble (Black) and the fitted sub-ensemble (Red). Left: The sub-ensemble of the hH16-100 fitting exhibits a significantly narrower distribution centered around a smaller R_g value (~28 Å) than the initial ensemble. Right: The R_g distribution of the fitted hH16-20 sub-ensemble. Broader distribution compared to that of the hH16-100 sample and with a bimodal shape.

It was important to observe the large experimental errors of lower D_2O/H_2O ratio measurements of protonated protein, before the expensive deuterium labelled samples were produced and measured. The difference between the simulated value, observed in chapter 4, and the realistic measurability of a sample had to be taken into consideration when planning SANS experiments.

5.3.2 hH16-dQE

The samples containing deuterated glutamine residues were these expected to be the richest in structural information. The protein yield was significantly lower, than that of the fully protonated sample, yet a protein concentration of 4.6 mg/mL was achieved by increasing the expression volume to 48 mL.

The hH16-dQE sample was measured in 0%, 40% and 100% D_2O solutions. The hH16-dQE-40 scattering pattern was indistinguishable from the one of the buffer solution (Figure 5.16), as this experimental condition was very close to the global match point of the protein (also called “zero averaged contrast”), which was 46.0%, as shown in chapter 4.4. We aimed at testing the signal-to-noise of this theoretical information rich sample. The zero averaged contrast is currently used for the study of partially deuterated polymers, but in the case of

biological samples it is practically unfeasible due to the lower signal-to-noise ratio and sample concentration (277).

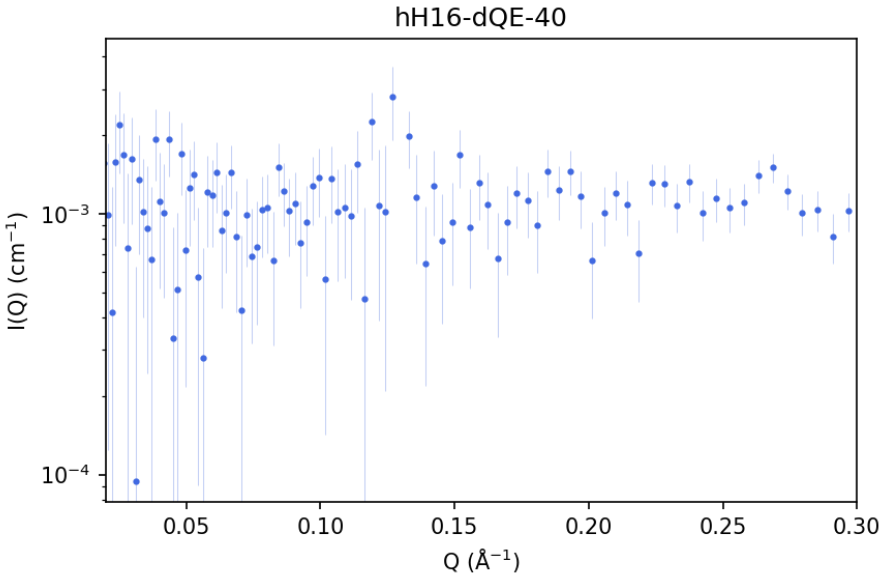


Figure 5.16: Subtracted scattering pattern of hH16-dQE-40. The signal of the labelled residues was too low to be isolated from the background scattering of the solution.

While the 40% D_2O sample yielded no useable data, the hH16-dQE-0 and hH16-dQE-100 samples provided data with reasonable to good signal-to-noise ratio (Figure 5.17). The R_g-value calculated from the hH16-dQE-0 scattering data was 39.7 ± 1.6 Å, while the R_g-value of the higher signal-to-noise profile of hH16-dQE-100 exhibited a lower size with a R_g-value of 30.8 ± 0.7 Å. The hH16-dQE-100 had, in addition to a better signal-to-noise ratio, a much better defined Guinier region and the R_g was more consistent with other H16 constructs. The D_{max}, as estimated by the *p(r)*, were 120 Å and 108 Å, respectively. The higher R_g and D_{max} of the hH16-dQE-0 dataset suggested that the sample could contain a portion of larger species of oligomerized protein.

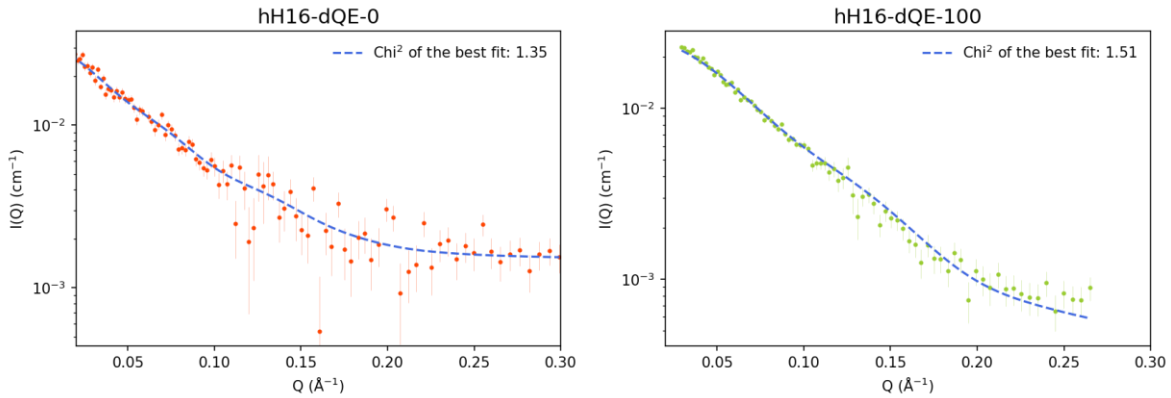


Figure 5.17: SANS scattering profile of hH16-dQE obtained in solutions of 0% (Left) and 100% (Right) D_2O.

A sub-ensemble was fitted with EOM to the hH16-dQE-0 and hH16-dQE-100 curves with χ^2 values of 1.35 and 1.51, respectively (Figure 5.17). The lower signal-to-noise ratio of the 0% D_2O data compared to the 100% D_2O data, mainly due to the incoherent background, explains the slight difference in χ^2 in both samples.

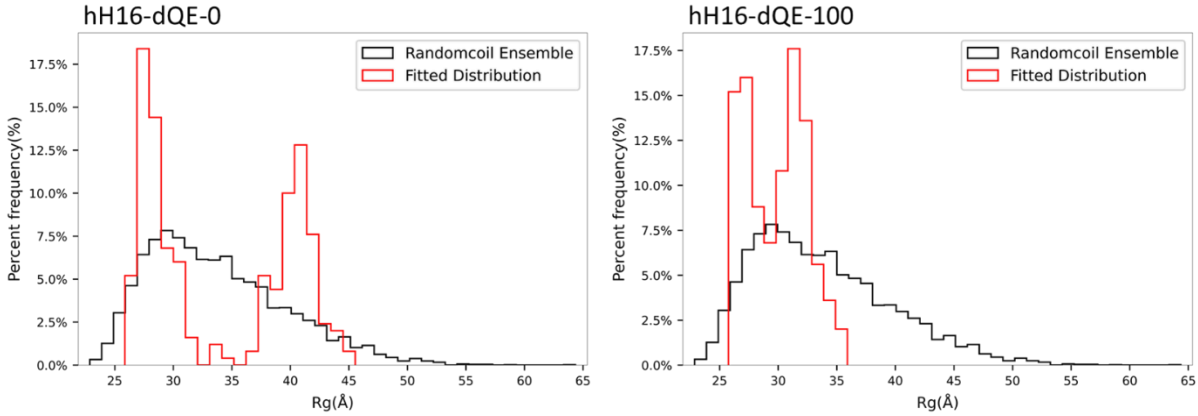


*Figure 5.18: R_g distributions of hH16-dQE-0 (Left) and hH16-dQE-100 (Right). The **black** distribution contains the R_g of all structures in the theoretical ensemble and the **red** distribution was the selected sub-ensemble from the EOM fit.*

From figure 5.18, it is clear that the R_g distribution of hH16-dQE-100 (Figure 5.18, Right) had a narrower distribution centered around 30 Å, while the hH16-dQE-0 dataset (Figure 5.18, Left) showed a bimodal distribution centered around 30 and 43 Å. Under the assumption that D_2O had no impact on the protein structure, these differences could be attributed to the sensitivity of the data to the deuteration pattern and/or to their distinct signal to noise level. Both samples were incorporated into the fitting of multiple dataset fitting (see below).

5.3.3 dH16

The fully deuterated sample served two purposes.

(1) The fully deuterated protein would show if the labelling had a structural effect.

(2) The full deuteration of the protein would provide an insight into the CF yield of fully labelled samples

The perdeuterated protein was produced using the same expression method as the hH16-dQE sample, with the amino acid mixture being exclusively deuterated amino acids. The sample could have been produced at a cheaper reagent cost using deuterated ISOGRO® with the four missing amino acids (Gln, Asn, Cys, and Trp) added as singular deuterated amino acids. However, in order to match the expression method with the rest of the samples, the CF was conducted using the combination of 20 individually deuterated amino acids. To avoid residue scrambling with the buffer, KOAc buffer was used for the deuterated sample. The yield of the expression was low, so 48 mL of CF expression was used to produce an adequate sample.

The dH16 sample was measured in both 0% and 40% D_2O at 4.8 mg/mL. The R_g of the protein samples were calculated to be 40.2 ± 0.3 Å in 0% D_2O and 39.2 ± 0.7 Å in 40% D_2O. The initial datapoints showed no sign of aggregation and also showed a similar radius of gyration as the fully protonated hH16 sample. Therefore, the full deuteration did not promote the oligomerization or aggregation in the experimental conditions tested. The D_{max} was estimated from the *p(r)* to be 115 Å and 116 Å, respectively, for the 0% and 40% data. The similar D_{max}-values suggested that the two protein samples were unaffected by the %D_2O of the solution.

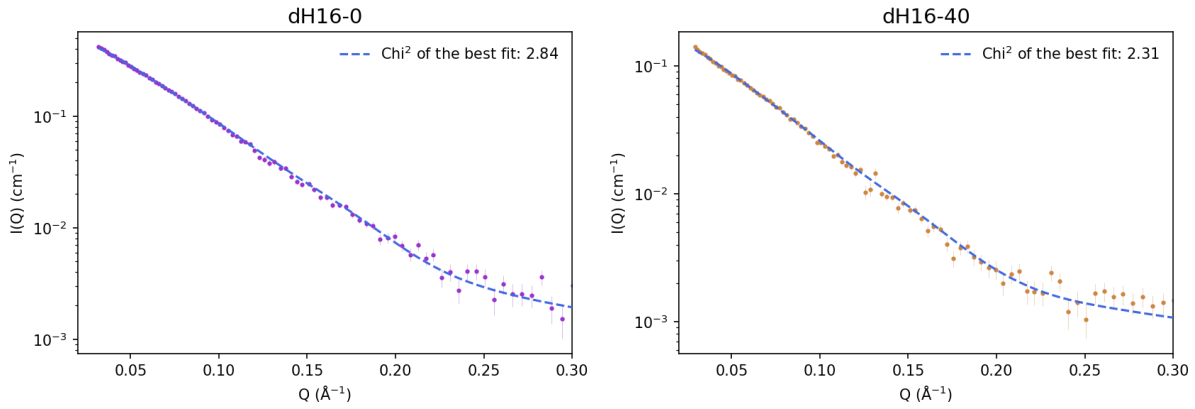


Figure 5.19: Fitted sub-ensembles to the dH16-0 (Left) and dH16-40 (right) datasets.

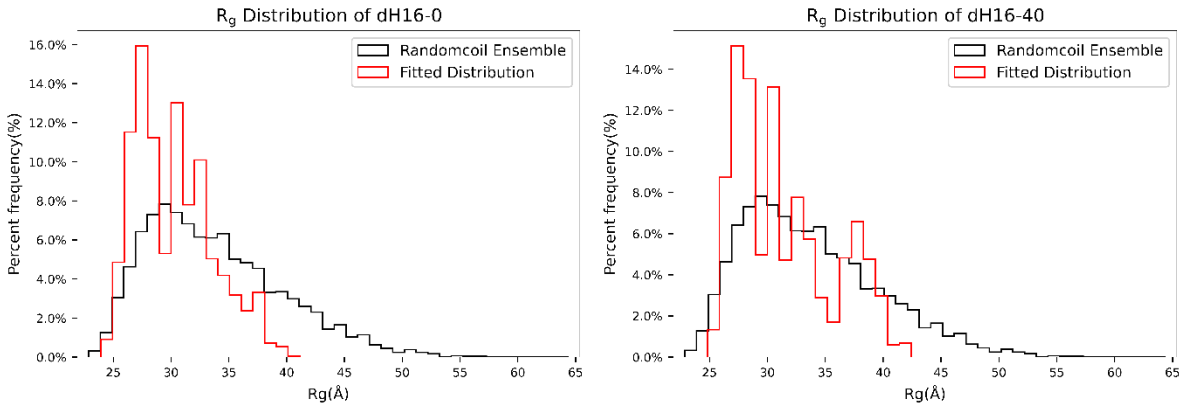


Figure 5.20: R_g distributions of dH16-0 and dH16-40. Both profiles showed a slightly narrower distribution compared to that of the full ensemble.

A theoretical sub-ensemble was fitted to the two datasets using EOM (Figure 5.19). The R_g distribution of both samples showed broad distributions, although slightly narrower than the theoretical ensembles. In this context, these distributions were similar to these obtained for the fully protonated hH16 measured using SEC-SANS and SEC-SAXS (Figure 5.20).

From the analyses of both the dH16-0 and dH16-40, it was concluded that the samples could be used for multiple SANS/SAXS fitting and that, importantly, the CF protein expression could produce deuterated protein in a monomeric state and in enough quantity for their structural investigation.

5.3.4 dH16-hQE

The deuterated protein with protonated glutamine/glutamic acid residues were easier to produce than the dH16 sample. The CF mixture was prepared using glutamate buffer allowing for a greater yield of protein synthesis (see chapter 3.2). The protein sample was measured at 4.5 mg/mL in 0% and 40% of D_2O (Figure 5.21).

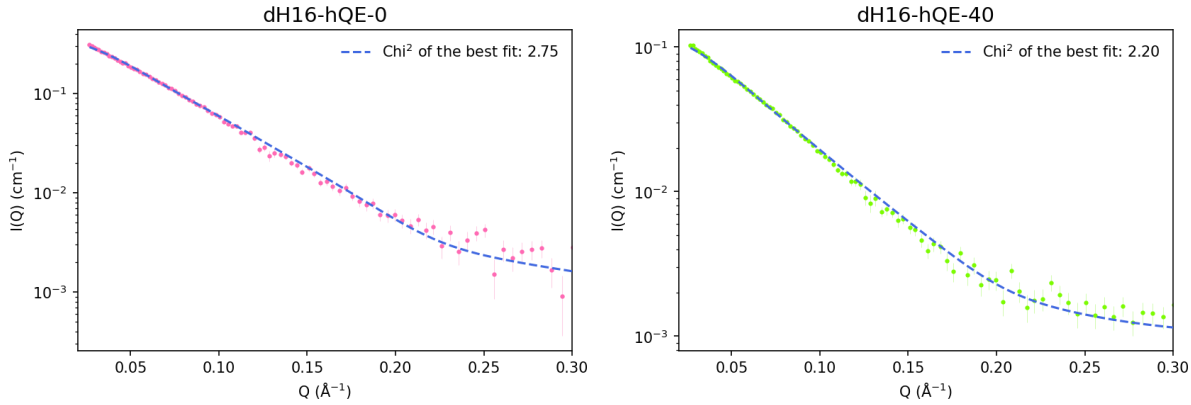


Figure 5.21: EOM sub-ensemble fit to the experimental scattering data for both dH16-hQE sampled, at 0% (Left) and 40% (Right).

The two experimental datasets did not show visual signs of aggregation and the initial analysis supported this observation. The R_g calculated from the Guinier regions were 38.0 $\pm$ 0.4 Å (dH16-hQE-0) and 37.3 $\pm$ 0.6 Å (hH16-dQE-40), which were very consistent between the two measurements. The D_{max} was estimated to 115 Å and 117 Å from the *p(r)*-distributions, respectively. The R_g-distribution of the two samples were similar with both distributions centered around 31 Å (Figure 5.22).

The χ^2-value of the of the fitted sub-ensemble was slightly higher for the 0% D_2O sample. The higher incoherent background scattering of the H_2O solution decreased the signal-to-noise ratio of the dH16-hQE-0 sample which decreased the quality of the fit. This difference between the two conditions was also observed for the dH16-0 and dH16-40 datasets (Figure 5.19).

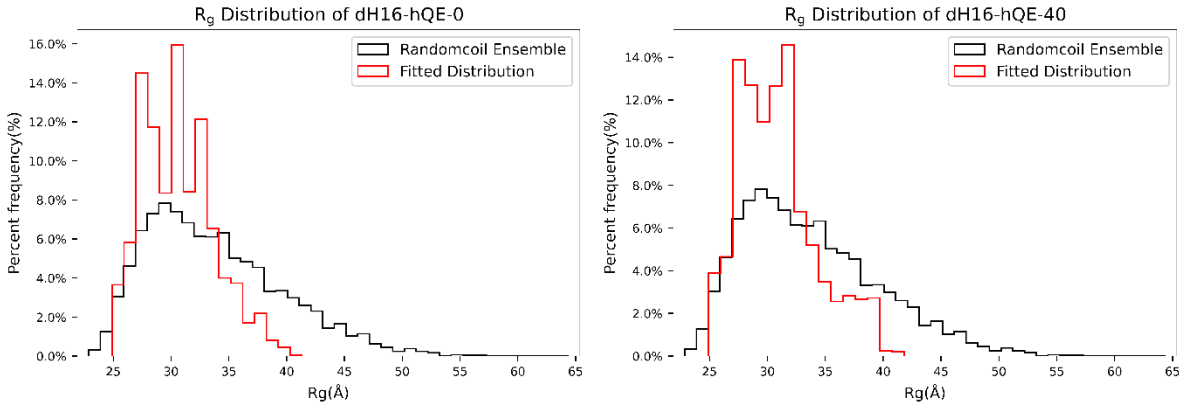


Figure 5.22: The R_g distribution derived with EOM for the dH16-hQE-0 (left) and dH16-hQE-40 (right) profiles.

5.3.5 hH36

The first sample of the H36 construct measured was the fully protonated hH36 sample. The protein was expressed by CF and measured for 1 hour and 21 minutes. The hH36 was produced at a slightly lower yield than that of the hH16 sample. This decrease was attributed to the more extended poly-Q tract, which could increase the amount of oligomerized/aggregated protein removed during purification. The scattering profile showed a very high signal-to-noise. From the scattering profile, the R_g and D_{max} were estimated to 32.6 $\pm$ 0.5 Å and 130 Å, respectively. The fitted sub-ensemble fit the overall profile decently with a χ^2 of 1.9, although a slight deviation could be observed between 0.10 – 0.17 $Å^{-1}$ (Figure 5.23, left) and at very low Q. The R_g distribution of the fitted sub-ensemble showed a narrower distribution compared to that of the full theoretical ensemble with a maximum around 32 Å (Figure 5.23, right).

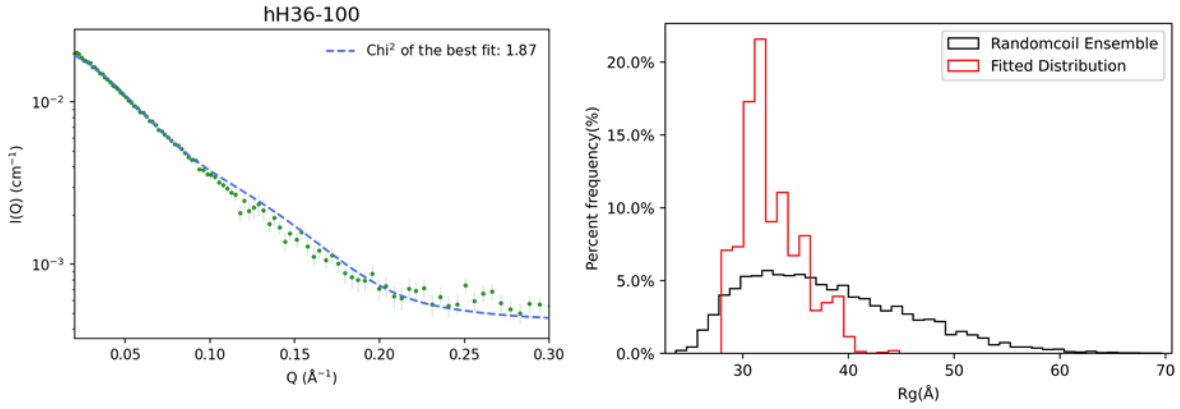


Figure 5.23: Left: Average profile of the fitted sub-ensemble compared to the experimental dataset of hH36-100. Right: R_g-distribution of the fitted sub-ensemble.

Concerning the slight deviation of the fit from the experimental data observed around 0.15 $Å^{-1}$, the fit passed within the error bars for most of the experimental datapoints. The overall fit quality and the tight R_g distribution of the sub-ensemble suggested that the hH36 dataset could be well integrated into the multiple-fit analysis.

5.3.6 hH36-dQE

The dQE labelled hH36 sample was produced at a lower yield compared to the control hH36 sample. Note that this deuteration pattern requires the use of a KOAc buffer, which notably reduces the yield of the CF (see chapter 3.2). Due to these limitations, the sample only reached a concentration of 0.9 mg/mL before injection onto the column, which further diluted the sample by about a factor of two. As a consequence, a curve with very limited data was obtained in a 100% D_2O buffer. R_g was calculated to 28.0 ± 0.9 Å, although the Guinier fit was of poor quality. The D_{max} could not be accurately estimated from the *p(r)*-distribution.

The averaged scattering profile of the EOM fitted sub-ensemble explained the experimental data well, indicating that, despite the high noise of the data, the theoretical ensemble could explain the experimental curve (Figure 5.24, Left). The EOM analysis yielded a narrow R_g-distribution centered around 38 Å, with a smaller population of very compact conformations. The overall agreement with the theoretical ensemble suggested that a higher sample concentration could have provided a significantly better dataset with rich structural information. While the data had a generally low data-signal-to-noise ratio, the curve was included in the multiple fit analysis of H36 (see below).

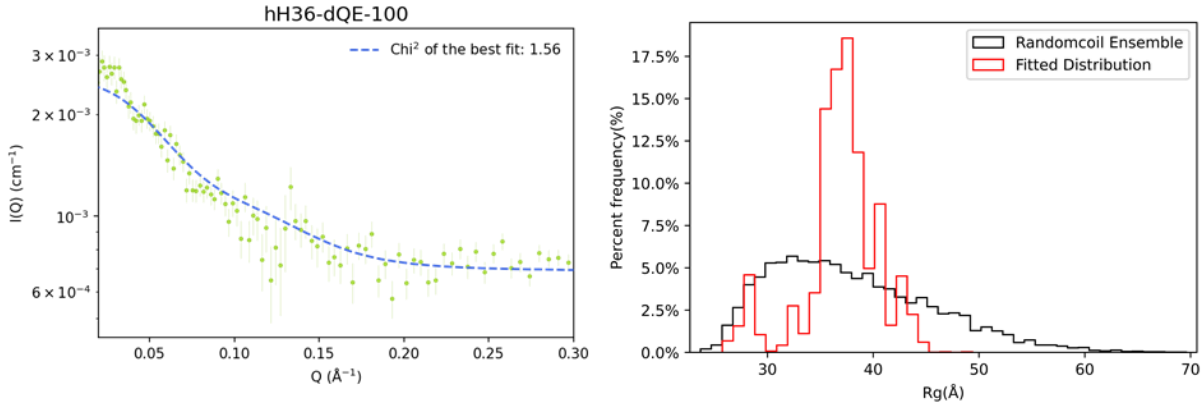


Figure 5.24: Left: Average profile of the fitted sub-ensemble compared to the experimental dataset of hH36-dQE-100. Right: R_g-distribution of the fitted sub-ensemble.

5.3.7 dH36

Fully deuterated dH36 was produced by CF using the KOAc buffer and the full 20 deuterated amino acids, and concentrated to 2.5 mg/mL before the SEC-SANS measurement. The CF expression yield was low and only one sample for SEC-SANS experiments could be prepared. The experimental plan was to recover the flow through of the column to concentrate and rerun the sample at a different %D_2O. Unfortunately, an instrumental error occurred during the experiment, resulting in the fractionator not running correctly, which in turn caused the protein to not be collected to measure it again with a different D_2O level. The sample was measured in 40% D_2O overnight. The signal-to-noise ratio was low over the Guinier range and the calculated R_g was 28.6 $\pm$ 3.4 Å. Similar to the hH36-dQE-100 sample, the *p(r)* was of very low quality and the D_{max} could not be accurately estimated.

The EOM fitted sub-ensemble provided a good fit with a low χ^2 of 0.97, albeit this low value could be attributed to the low signal to noise (Figure 5.25). The R_g-distribution showed a bimodal distribution with structures ranging from 29 – 56 Å. The dataset was incorporated to the multiple fitting analysis of H36.

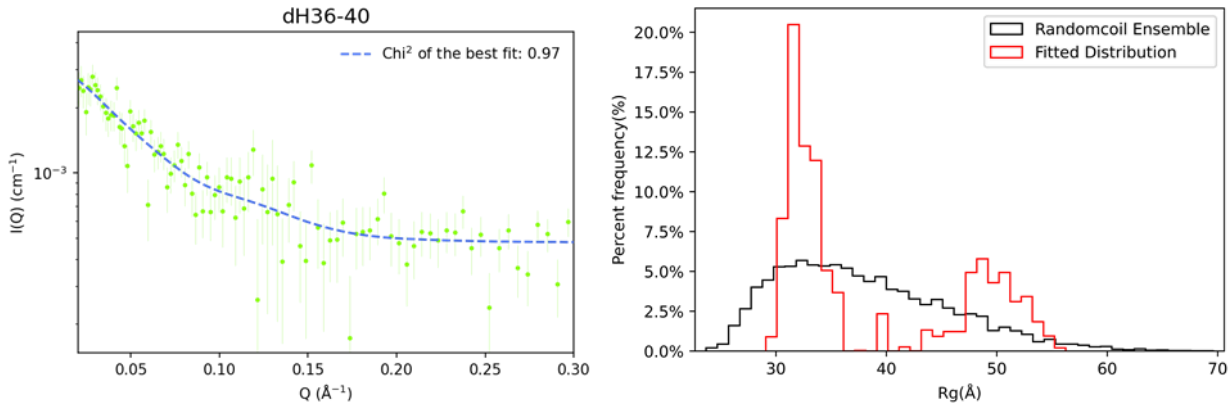


Figure 5.25: Left: Average profile of the EOM-fitted sub-ensemble compared to the experimental curve. Right: R_g-distribution of the sub-ensemble compared with that of the initial pool. The R_g-distribution obtained displayed a bimodal behavior.

5.3.8 dH36-hQE

The dH36-hQE sample was easily produced because the labelling scheme allowed for KGlu buffer to be used during expression. The CF expression yielded enough protein for two samples, which were concentrated to 2.2 mg/mL and measured in both 0% and 40% D_2O. The 0% D_2O sample yielded a low signal-to-noise scattering profile, while the 40% D_2O scattering data was indistinguishable from background noise, and was not further investigated. The R_g was calculated to 28.4 ± 0.8 Å, but the Guinier region had a poor signal-to-noise ratio. The poor signal-to-noise ratio also resulted in a low-quality *p(r)*-distribution and D_{max} could not be accurately estimated.

EOM fitted a sub-ensemble to the dH36-hQE-0 scattering profile with a χ^2 of 0.96. The fit was similar to that of the dH36 sample. The R_g-distribution of the EOM fit (Figure 5.26) showed a broad distribution of conformations. This dataset was included in the multiple fitting analysis of H36.

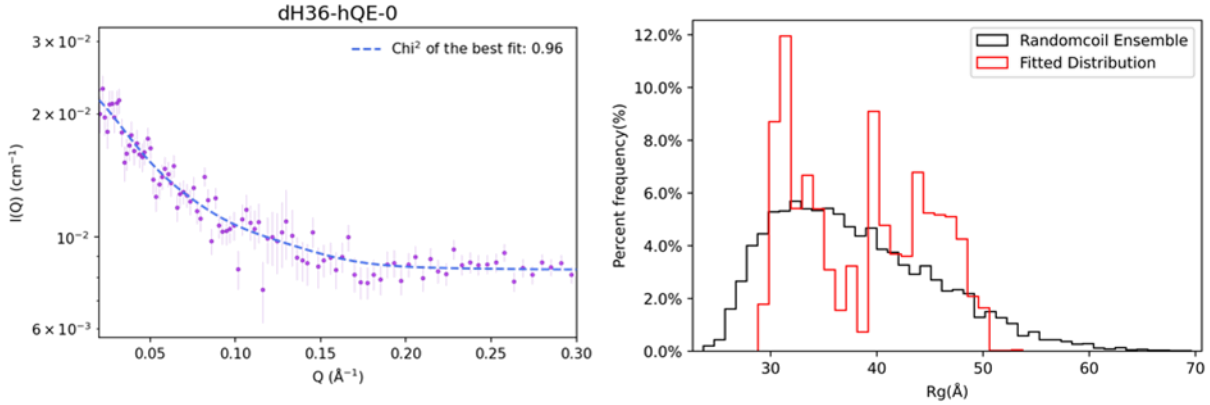


Figure 5.26: Left: Average profile of the EOM fitted sub-ensemble compared to the experimental dataset of dH36-hQE-0. Right: R_g-distribution of the resulting EOM sub-ensemble.

5.4 DEUTERATED PROLINES ALTER HUNTINGTIN SAMPLES

In my initial plans, I wanted to explore the specific labelling of the prolines in order to investigate the structural features of the poly-proline tracts present in huntingtin. However, during the analyses, all datasets of protonated protein with deuterated proline residues showed problems. The initial sample of hH16-dP, measured September 2020, could not be fitted with EOM. This inability to fit the experimental data was initially hypothesized to be an artefact of the sample preparation or from the batch-type measurement. Therefore, the hH16-dP, hH16-dQEP, hH36-dP, and hH36-dQEP were all measured during subsequent beamtimes using the SEC-SANS mode. Importantly, although most of these datasets were obtained with good signal-to-noise ratios, unfortunately they did not provide a good fit when using EOM (Figure 5.27).

The chemical composition of the samples was validated by MS as a potential origin of the problem. Indeed, for unknown reasons, the hH36-dQEP sample was found to contain two separate isotopologues when analyzed by MS (Chapter 3.4, Table 3.1), suggesting that the sample was not correctly produced. The incorrect labelling pattern would make the result of the EOM unusable due to the atomistic nature of the structural ensembles. However, this origin of the problem was not applicable to the other samples. Indeed, the hH16-dP, hH16-dQEP, and hH36-dP presented molecular weights that were in agreement with the theoretical ones.

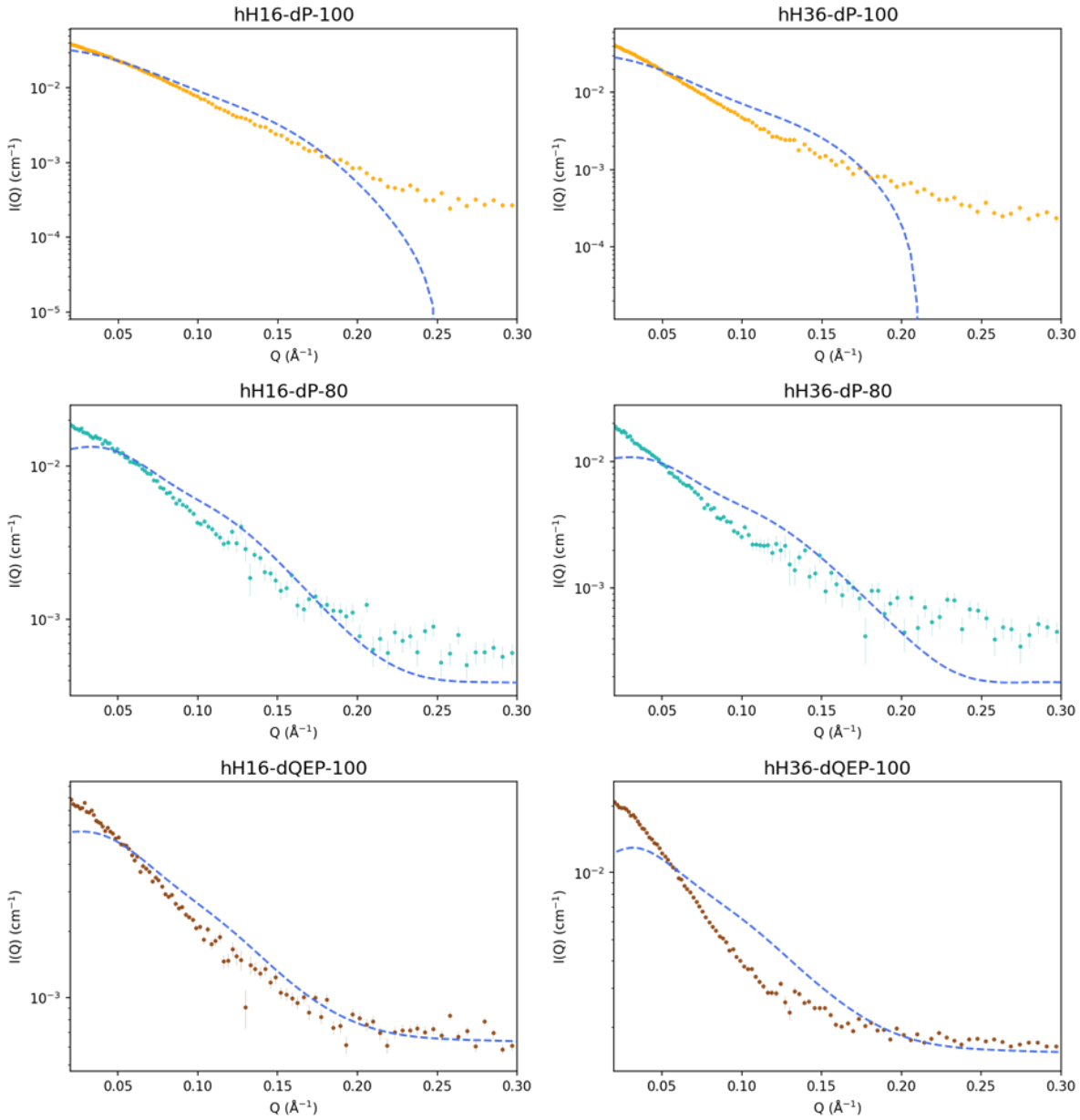


Figure 5.27: EOM Fitted sub-ensembles to samples containing protonated protein with deuterated proline. All 6 profiles show significant deviations from the experimental data with similar tendencies.

The two labelling patterns (dP and dQEP) showed the same trends in both H16 and H36 constructs and in 80% and 100% D_2O solutions. Importantly, when repeating the measurements for a new hH16-dP sample, the profile showed the same features and deviations from the EOM fitted sub-ensemble. All the fittings obtained sub-ensembles with a lower *I(0)* and a clear deviation in slope around 0.10 $Å^{-1}$. This was further validated by comparing each of the 5,000 theoretical profiles of the hH16-dP 100% D_2O to that of the experimental data. Plotting the minimum and maximum values for each momentum transfer and scaling them to the experimental data enabled a visual perspective of the problem (figure 5.28). Indeed, the

experimental profile could not be fitted within the range of profiles provided by the theoretical ensemble.

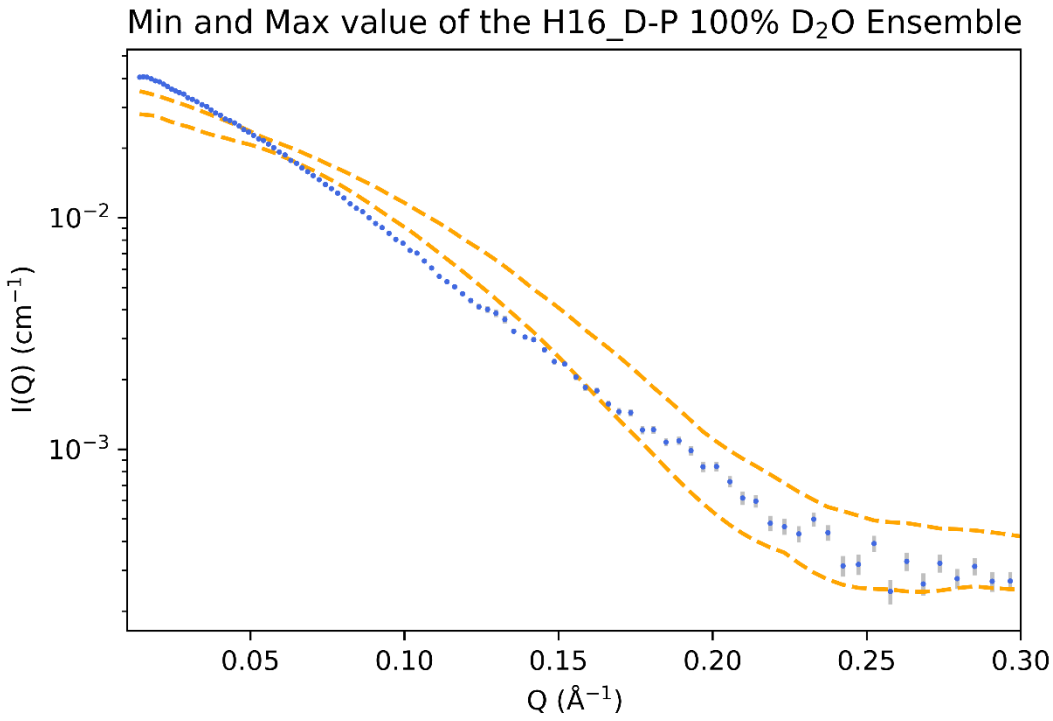


Figure 5.28: Minimum and maximum intensity values of each momentum transfer derived from the theoretical ensemble plotted together with the experimental dataset. The slope of the SANS scattering pattern could not be fully encompassed by the interval values

These analyses showed that the particle formed by the deuterated-proline versions of huntingtin in solution were larger than expected from the monomeric state of the protein. Therefore, the addition of deuterated proline in a mainly protonated huntingtin exon-1 had an effect on the structure and/or the oligomerization state of the protein that could not be explained by the theoretical ensemble of monomers. This problem was not explored further during the project, but experiments could be designed to further investigate the effect of deuteration of prolines in poly-proline containing proteins in relation to the oligomerization process. For instance, differences of the aggregation kinetics of deuterated and protonated proteins could bring some insights into this phenomenon. Although the increase of the nuclear weight of deuterium and its consequences are well known, their impact in the structure and stability of proteins remains poorly described (192,196). The capacity of our CF approach to specifically deuterate amino acids at wish could provide novel insights in this fundamental question.

From a practical point of view regarding the structural analysis of huntingtin, the SANS data measured for hydrogenated samples with deuterated prolines had to be excluded from the multiple fitting analysis despite their high signal-to-noise profiles and excellent statistics.

6 SIMULTANEOUS STRUCTURAL ANALYSES OF SAXS AND SANS DATA

With the goal to combine datasets of multiple experiments in mind, the SAXS and SANS data described in the previous chapter had to be evaluated for their compatibility. Although extensively applied in other techniques such as NMR (278), cross-validation has rarely been performed in SAS due to the absence of multiple datasets measured for the same protein. In the presented thesis work, the availability of distinct deuteration patterns for the same protein allowed for this possibility.

The beamtime used during the project yielded significantly more datasets for H16 than for H36. Due to this reason, the multiple dataset fitting analysis focused on H16 to take advantage of the better signal-to-noise ratio of these samples. Then, the same approach was applied to H36, although the reduced number of datasets limited the quality of the resulting conclusions. Before the simultaneous multiple curve analysis, each dataset was incorporated into the cross-validation test to confirm that the data could be complementarily combined.

6.1 HOW WAS CROSS-VALIDATION OF SAS DATA APPLIED?

In general, cross-validation is a tool for testing consistency between models and experimental data, and to prevent overfitting (278). In the present project, cross-validation was done by assessing the capacity of sub-ensembles refined by EOM from a given set of experiments to properly describe scattering profiles not used in the initial fitting.

To cross-validate the datasets, it was assumed that the deuterium labelling and exchange does not significantly change how the protein behave in solution. The theoretical ensemble of atomistic models was conformationally identical between each labelling pattern and solution D_2O levels with the exception of the H/D exchange.

It was proposed that when cross-validation was performed, data quality and compatibility could be discerned. Outlier datasets showing incompatible structural tendencies or low impact data, which would not be discriminative, could be identified from the cross-validation. The identified outliers could subsequently be discarded from the multiple dataset analysis to optimize the sub-ensemble of atomic structures proposed to collectively describe all the (compatible) experimental datasets. As described in chapter 1.4.6, the combination of multiple datasets during the EOM fitting would result in a more constrained fit due to the increased number of

degrees of freedom. This constraint normally yields a higher χ^2-value of a fit compared to the single dataset fitting. As a result of the assumptions and nature of multiple dataset fitting by EOM, a moderate increase of χ^2-values is expected. Similarly, any dataset that showed no change of the fitting value would not comprise any discriminative data, suggesting the data would have little to no impact on the multiple dataset fitting and therefore discarded from the multiple curve analysis.

The cross-validation was split into two χ^2-values: χ^2-work and χ^2-free.

- **χ^2-work:** The χ^2-work values were the EOM fitting values used throughout the project: they compare experimental datasets with the scattering curves calculated from the conformational ensembles refined from these datasets. These values would be impacted by the increased number of constraints when multiple curves were combined, meaning they would increase when successive datasets are incorporated.
- **χ^2-free:** The χ^2-free value assesses the capacity of a fitted sub-ensemble to describe datasets not used in the fitting.

The cross-validation was performed by three steps (Figure 6.1): 1) The structures of the EOM fitted sub-ensemble, which provides the χ^2-work, were translated from the experimental sample to another deuteration condition. 2) The synthetic scattering pattern of the new condition was then re-generated by averaging the simulated scattering patterns of the translated sub-ensemble. 3) the re-generated scattering profile was compared to the experimental scattering profile of the new condition to calculate a χ^2-free value.

The translation of the EOM sub-ensemble was done by taking advantage of the nature of the conformational ensemble that I have built (see chapter 4.1). Each individual structure of the initial ensemble was labelled and exchanged into all conditions and, in this way, structures #1 of each labelling pattern and solution condition were conformationally identical. This allowed for the selected structures of the EOM sub-ensemble to be combined in a different condition in a straightforward manner.

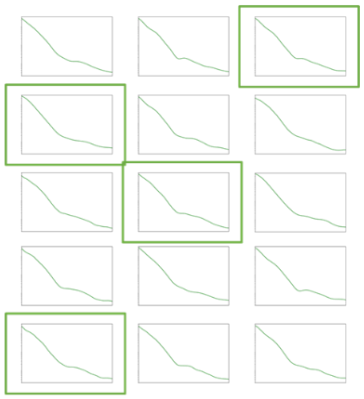
In order to compare the re-generated scattering profile with the experimental data, the profiles were re-gridded and, to match the experimental background, scattering constant subtraction was applied to find the optimal comparison.

The cross-validation was systematically performed for all samples. Each sample was submitted to three cross-validation tests: 1) Single experimental dataset. 2) SAXS and one SANS profile

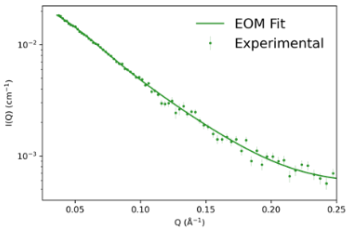
combined. 3) SAXS and two SANS profiles combined. This approach evaluated both the direct compatibility between labelling patterns and sample conditions, and the impact of combining SAXS and SANS data on the resulting ensemble. In other words, if the combination of scattering patterns is complementary, then additional structural information can be gathered from the combination of the scattering data (Full script for cross-validation: Appendix 11.3).

Experimental EOM Fit

Sub-ensemble fit to experimental dataset.



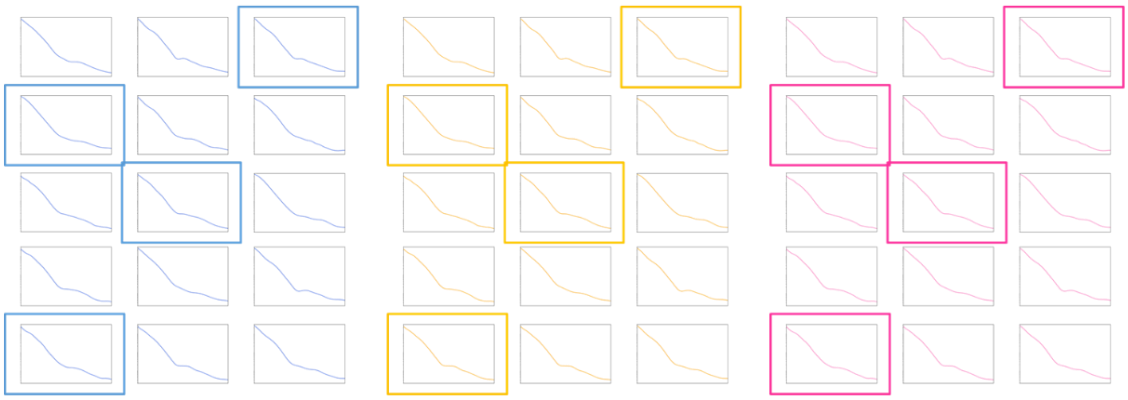
hH16-100



χ^2-work is obtained from the ensemble fitting by EOM

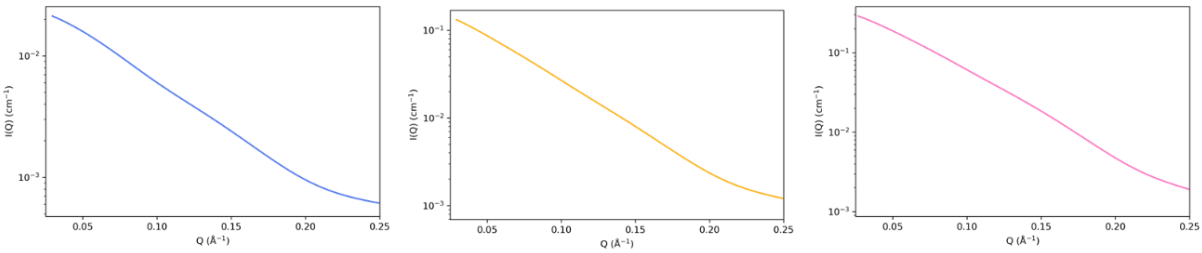
Regenerated Fits

Sub-ensemble of experimental fit to hH16-100 translated to other experiments.



hH16-dQE-100 dH16-40 dH16-hQE-0

Regeneration of translated sub-ensemble average for the new experimental condition



Comparison between regenerated scattering profile and experimental data (χ^2-free).

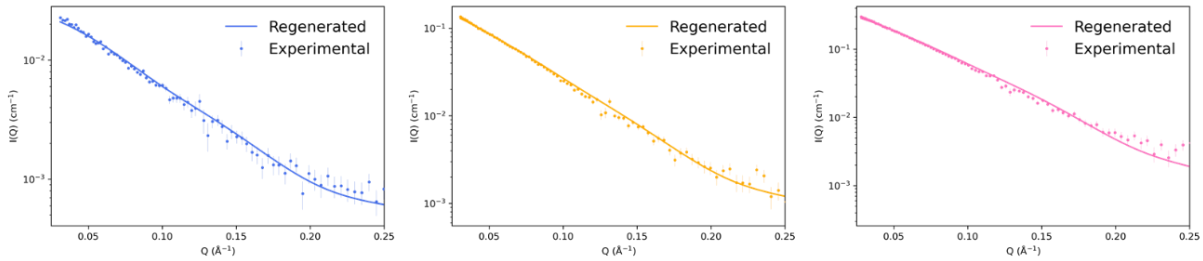


Figure 6.1: Visual representation of the cross-validation. χ^2-work was calculated by EOM during ensemble refinement. The structures of the EOM Sub-ensemble were translated to the labelling scheme and solution condition of another experiment. The fitted average was regenerated for the new sample and compared to the experimental scattering profile. The χ^2-free was the value obtained from the comparison.

6.2 CROSS-VALIDATION OF EXPERIMENTAL H16 DATA

The initial cross-validation was done with only one experimental dataset fitted at a time. Each fitted sub-ensemble was regenerated to each of the other labelling patterns yielding one χ^2-work and eight χ^2-free values per sample. The first conclusion that could be extracted from figure 6.2 was that all datasets can be properly fitted using the theoretical ensembles (χ^2-work). However, some of the resulting sub-ensembles were not able to explain the other datasets. For instance, the SAXS-derived ensemble could not explain the hH16-100, dH16-hQE-40, dH16-hQE-0, dH16-40, dH16-0 datasets (χ^2-free values above 9). Equivalently, the SANS datasets could not explain the SAXS curve. This could be due to the different sensitivity that structures had to the deuteration pattern or to the incompatibility of the datasets. When individual SANS-derived ensembles were cross-validated with the other curves, a general degradation of the χ^2-free was observed, although the values remain reasonable (< 4.0) in the vast majority of cases.

The analysis of the individual SANS datasets suggested a good complementarity between the sub-ensembles. The EOM-derived sub-ensembles for hH16-100, hH16-dQE-100, dH16-hQE-40 and dH16-40 datasets performed well when they were regenerated for the other SANS conditions. The dH16-0 sample performed well when the optimized ensemble was re-generated in other labelling schemes. However, when other datasets were re-calculated using the dH16-0 pattern, the χ^2-free value was poor.

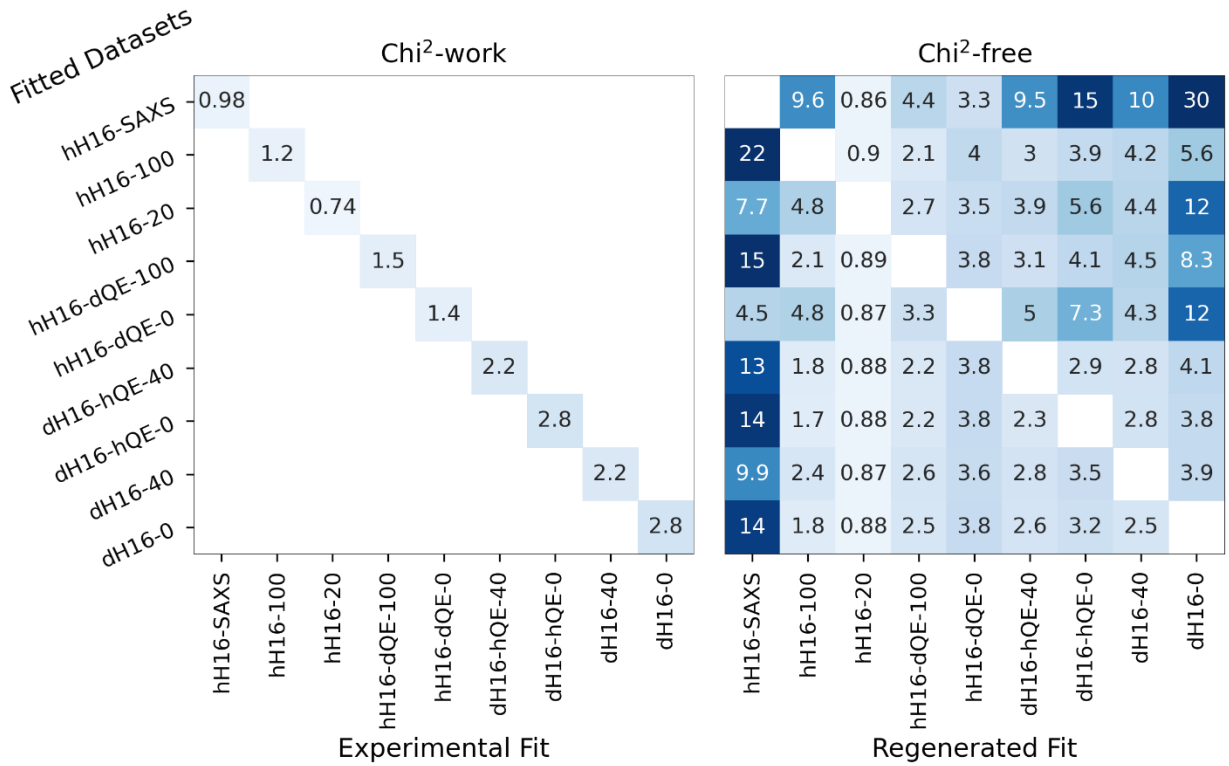


Figure 6.2: First cross-validation calculation. A single fitted dataset (χ^2-work) was used to calculate the agreement to each of the other labelling schemes (χ^2-free). hH16-100, hH16-dQE-100 and dH16-40 performed well when the fit was re-generated in other labelling schemes. The EOM-derived SAXS sub-ensemble performed poorly when compared to the SANS data.

A special case was the hH16-20 profile, which showed an excellent χ^2-free value of the re-generated profiles for all of the labelling schemes. This observation was due to the very low signal-to-noise-ratio of the experimental curve. The very low χ^2-free values and the stability of them (only 0.04 variation between the difference regenerated comparisons) suggested that including the hH16-20 data would not have a discriminative effect on the fitting, and it would not enrich the accuracy of the H16 ensemble when combined with other datasets. Therefore, this sample was unlikely to provide relevant information in the analysis.

6.3 MULTIPLE CURVE CROSS-VALIDATION OF H16 DATA

The next step was to combine two datasets and, for this, each of the SANS datasets was combined with the SAXS curve in order to incorporate the general information provided by SAXS with the more region-specific one coming from SANS (figure 6.3).

The first observation of note was that the combination of SAXS and SANS significantly improved the correlation between the two types of data. The χ^2-work values of most of the combinations of SAXS and SANS data stayed within a reasonable range (χ^2-work < 4). The

dH16-hQE-0 and dH16-0 samples exhibited the poorest EOM fit with χ^2-work values of 4.7 and 4.8 respectively. Specifically, the dH16-0 sample provided a poor fit when combined with SAXS, suggesting that the combination deteriorated the quality of the sub-ensemble. Similarly, the combination of hH16-SAXS and hH16-20 as well as hH16-SAXS and hH16-dQE-0 was shown to impact the χ^2-work value of the hH16-SAXS data very minimally. The very low impact of these two combinations suggested that the samples did not provide any discriminative information to the fit compared to the hH16-SAXS data.

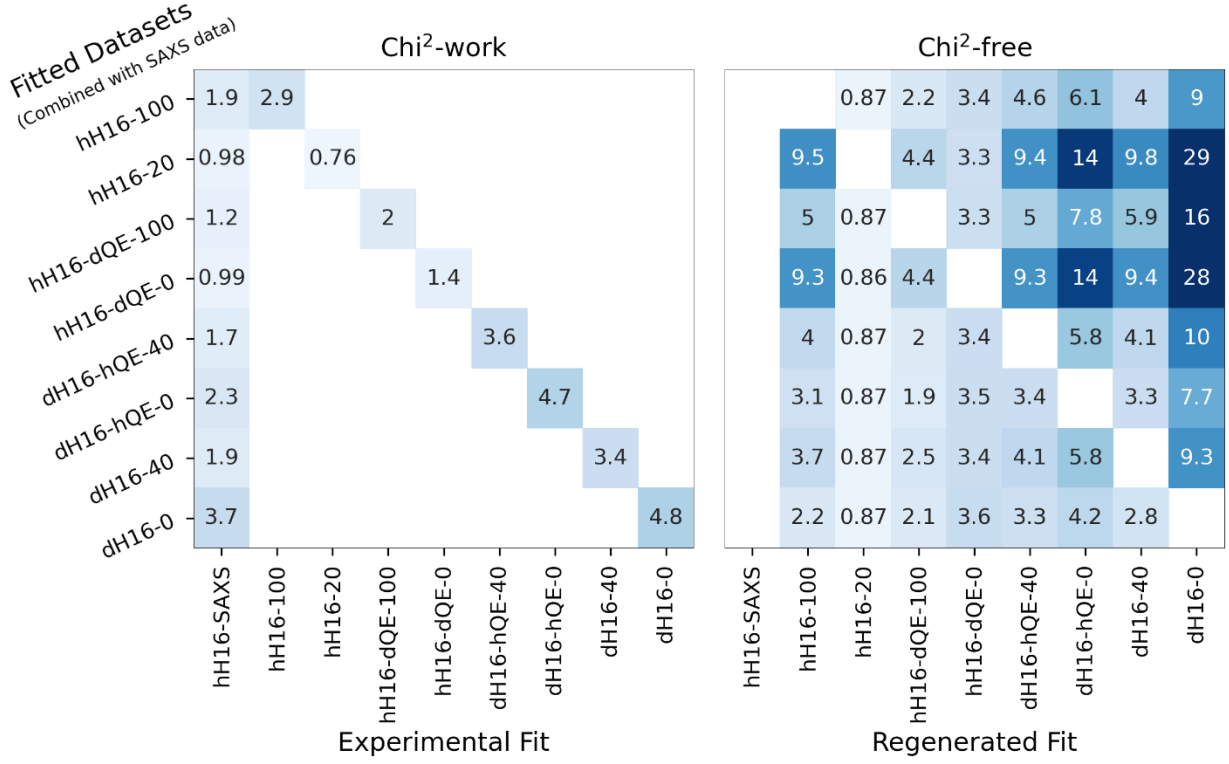


Figure 6.3: Two-dataset cross-validation by combining SAXS with SANS data of each labelling pattern. Similar to the One-dataset calculations, the unlabelled H16 in 100% D_2O, D-QE 100% D_2O dH16 40% D_2O performed well and in addition they fit the SAXS dataset significantly better than the previous re-generated fit. Both the unlabelled H16 in 20% D_2O and the D-QE 0% D_2O samples performed poorly when the fitted data was re-generated to other labelling patterns, with several χ^2-free values above 9.

The χ^2-work values of the SANS datasets all showed a slight increase. This observation was expected as multiple datasets were fitted and the chosen sub-ensemble had to accommodate both experimental datasets. Due to this effect, slightly higher χ^2-work values were accepted for multiple curve fitting.

The simultaneous fitting of both SAXS and SANS data was shown to generally allow for a decent comparison between the re-generated profile to other SANS labelling patterns (χ^2-free).

The exceptions to this behavior were the low signal-to-noise samples (hH16-20 and hH16-dQE-0), which both showed poor cross-validation from their experimental data. The poor χ^2-free values generated from these combinations were similar to the ones from the single fit of hH16-SAXS data (Figure 6.2) and further supported the observation that the two datasets were insensitive and would not improve the subsequent ensemble refinement. Additionally, the dH16-0 sample showed a poor regenerated comparison (χ^2-free) with all other experimental data combinations. In contrast to hH16-20, the dH16-0 sample had a higher signal-to-noise ratio and the combination with hH16-SAXS significantly deteriorated the χ^2-work, while simultaneously showing a poor cross-comparison. This effect suggested that the dataset was not compatible with the other SANS data.

Finally, the combination of two SANS datasets together with the hH16-SAXS scattering profile was tested. As in the double curve fitting described above, it was observed that the χ^2-work values of the SAXS and SANS profiles were not significantly increased upon the addition of a new SANS dataset, regardless of the specific dataset combination (figure 6.4). Indeed, several combinations showed an improved fitting when compared to the two-dataset fitting. It was also observed from the χ^2-work values that the hH16-SAXS fit was slightly worse than that of the two previous combinations (single and two-dataset fitting). Again, this behavior was expected when additional datasets were included in the fitting, due to the increased degrees of freedom.

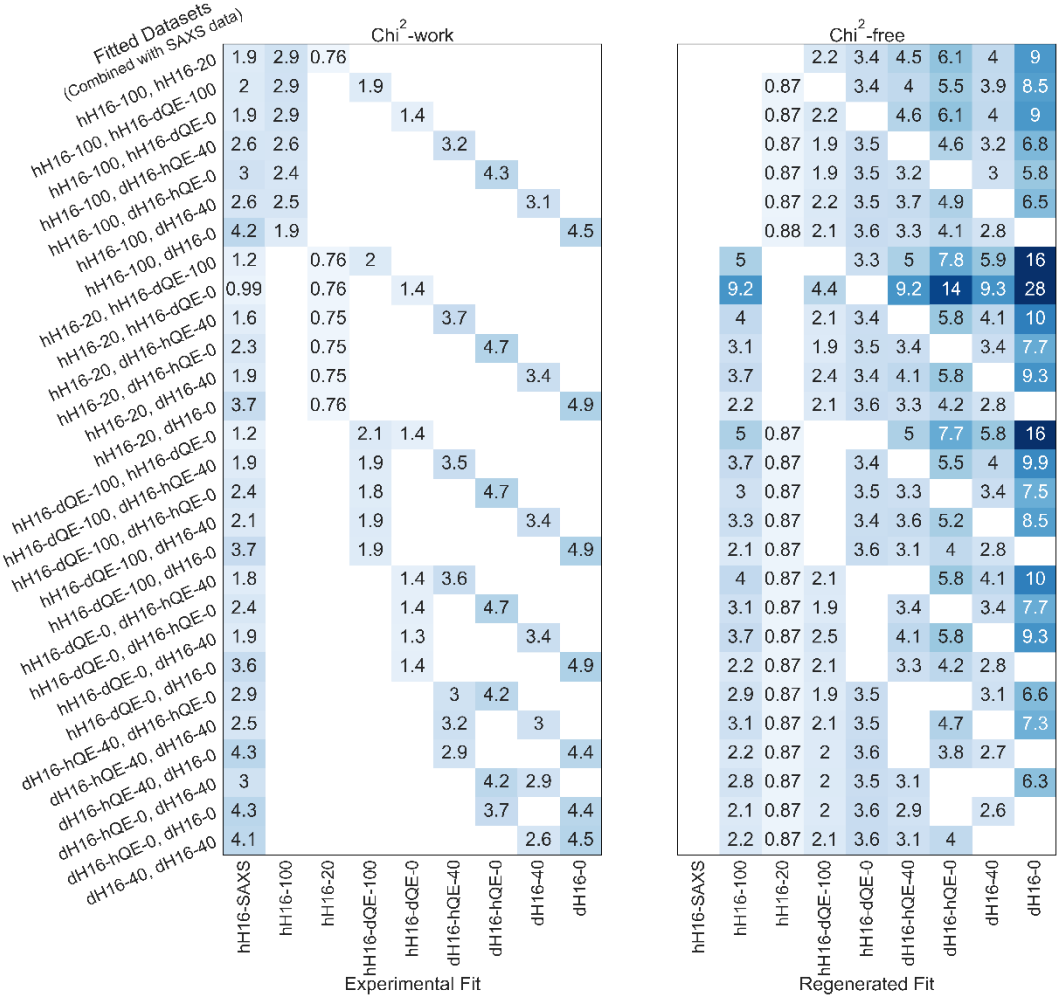


Figure 6.4: Triple dataset fitting of SAXS and SANS datasets. The combination improved the SANS fitting parameters (χ^2-work) of several of the SANS datasets, while slightly decreasing the fit of the SAXS data.

The results of the cross-validation analysis of the other SANS profiles were similar to these of the two-dataset fits. The χ^2-free stayed fairly stable for hH16-100, hH16-dQE-100, dH16-hQE-40 and dH16-40, with some improvement in the hH16-SAXS, hH16-100, and dH16-hQE-40 datasets. Similar to the two-set combinations, the fitting of the hH16-20 and hH16-dQE-0 datasets showed decreased χ^2-free re-generated comparisons, suggesting they would not improve the ensemble refinement. When the dH16-0 sample was introduced into the calculation, it systematically deteriorated the hH16-SAXS χ^2-work. Additionally, the χ^2-free values of dH16-0 never reached below 5.8. These two observations combined led to the conclusion that this sample was incompatible with the other scattering data.

The ensemble of the cross-validation analyses led to three conclusions: 1) several SAXS/SANS datasets could be simultaneously fitted without notably deteriorating the overall fitting. 2) The combination of SAXS and SANS data did not degrade the fitting of the non-used datasets. 3)

From the two- and three-curve fits, problematic (incompatible) datasets could be identified and taken out of the simultaneous analyses (see below). Concretely, the datasets chosen for the final multiple-curve fitting analysis were: hH16-SAXS, hH16-100, hH16-dQE-100, dH16-hQE-40, dH16-hQE-0, and dH16-40.

6.4 MULTIPLE CURVE ANALYSES OF H16 DATASETS

Fitting of multiple datasets was theorized to provide a more accurate description of the conformational ensemble of H16 in solution of that obtained from analyzing of a single dataset. By combining the datasets, the informational value of the ensemble was expected to increase. From the combined fitting of hH16-SAXS, hH16-100, and hH16-dQE-100 (Figure 6.5, top), as well as hH16-SAXS, hH16-dQE-100, and dH16-hQE-40 (Figure 6.6, top) it was observed that the three datasets could be fitted simultaneously with acceptable χ^2-values (1.8 - 3.5). Note that a slight increase of χ^2 values was expected when incorporating information from all three samples.

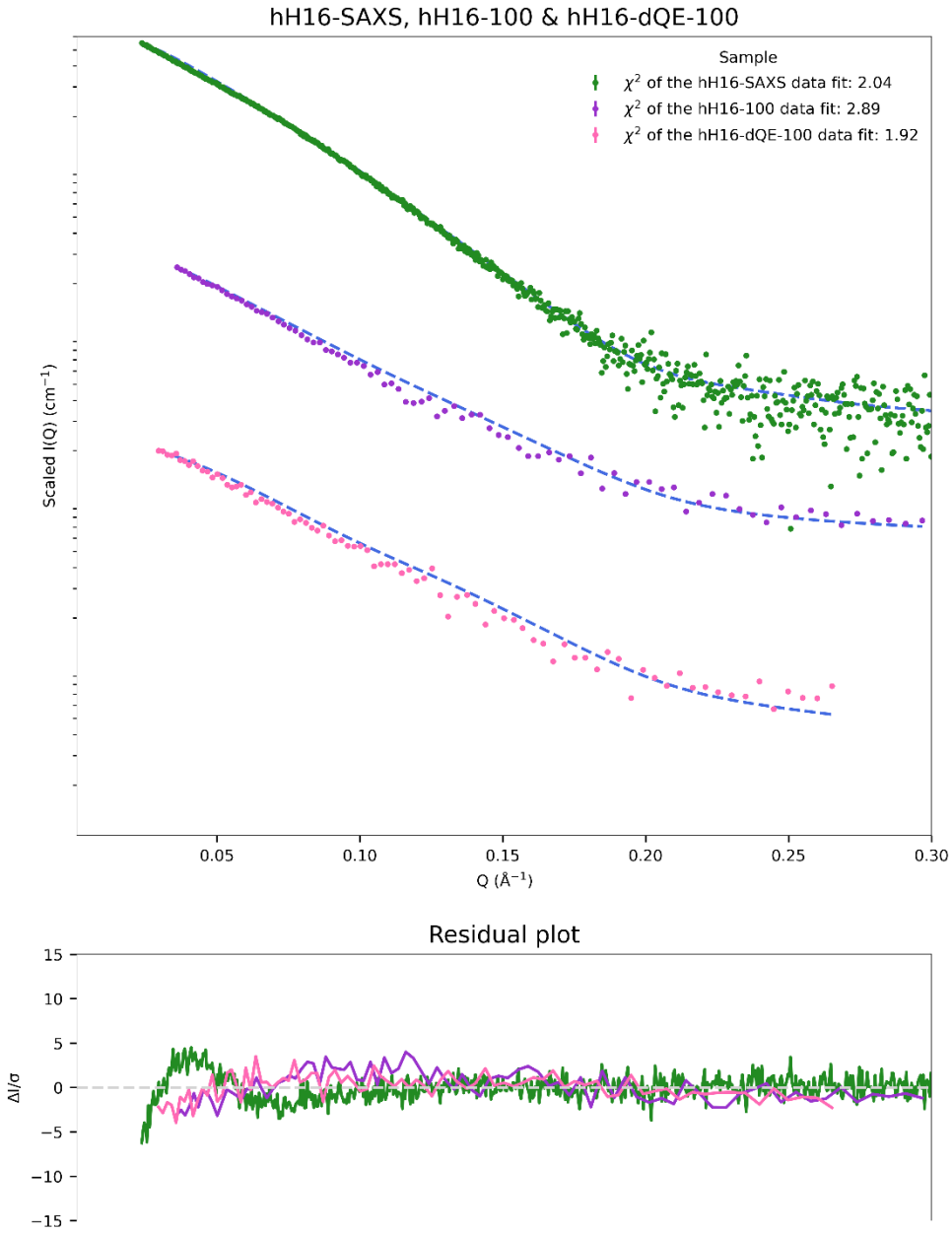


Figure 6.5: Multiple fit of H16 data. Combined fitting of hH16-SAXS, hH16-100, and hH16-dQE-100. The fit could accommodate the three datasets but, as expected, a slight decrease in the fitting quality was observed. The residual plot showed that the SAXS data (Green) has a slight deviation of the initial datapoints, while the SANS profiles fit the initial part of the curve better.

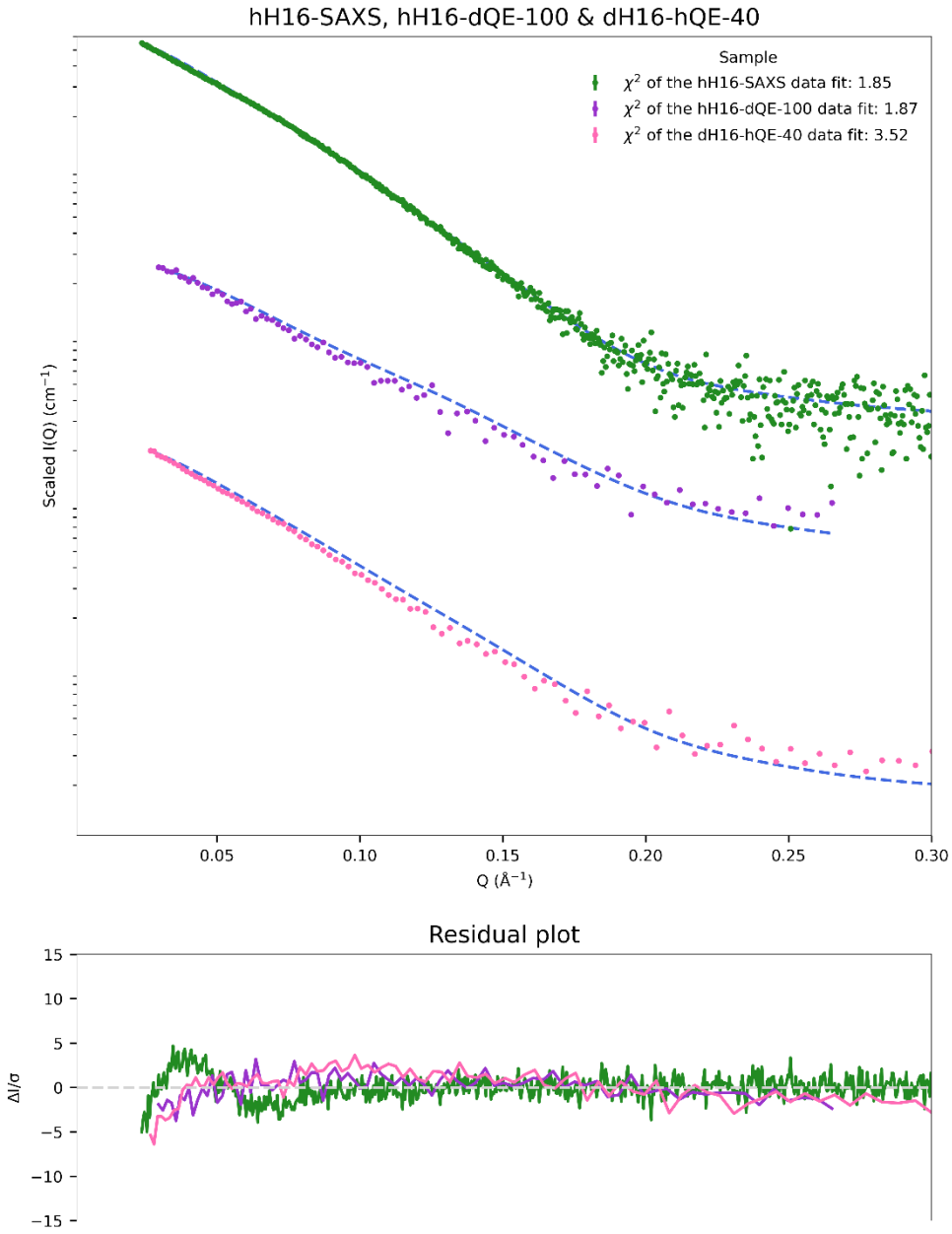


Figure 6.6: Multiple fit of H16 data. Combined fitting of hH16-SAXS, hH16-dQE-100, and dH16-hQE-40. The fit could accommodate the three datasets, but, as expected, a slight decrease in the fitting quality was observed. The residual plot showed that the SAXS data (Green) has a slight deviation of the initial datapoints.

In both analyses, the residual profiles showed a general good agreement with fluctuations around the 0 value, indicating that the resulting ensemble properly describes the three profiles. Only slight deviations for the SAXS dataset were observed at small angles.

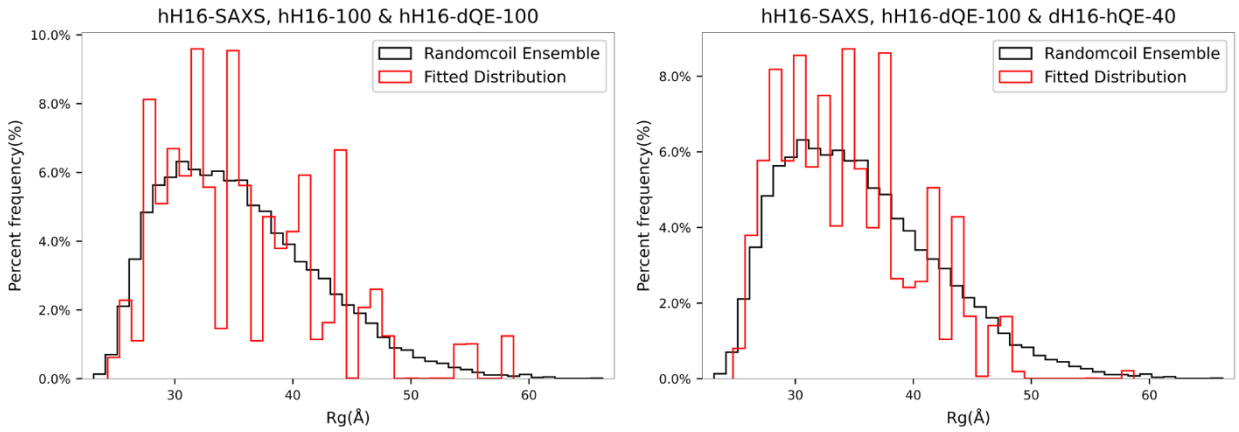


Figure 6.7: The R_g distributions of two triple dataset fittings. Both distributions showed a broad distribution of R_g values.

From the R_g distributions obtained from combining the three datasets, a broad distribution of R_g-values was observed. The triple dataset combinations did not significantly narrow the R_g distribution of the chosen sub-ensembles (figure 6.7), and both distributions used most of the R_g range of the theoretical ensemble.

To further investigate the impact of multiple dataset fitting, more datasets were combined. With the cross-fitting results in mind, six datasets were combined in one simultaneous analysis.

hH16-SAXS	hH16-100	hH16-dQE-100
dH16-hQE-40	dH16-hQE-0	dH16-40

The fitting value of the five SANS datasets remained stable when the six experimental datasets were simultaneously fitted, which suggested that the SANS data could be combined well (figure 6.8). The EOM χ^2-value of the hH16-SAXS profile was 4.3, which suggested that the influence of the SAXS profile decreased when more datasets were added to the EOM fit. The fit of the hH16-100 data improved from the three-dataset fitting to the six-dataset fitting. This

suggested that the SANS datasets were compatible with each other and improved the information value when combined.

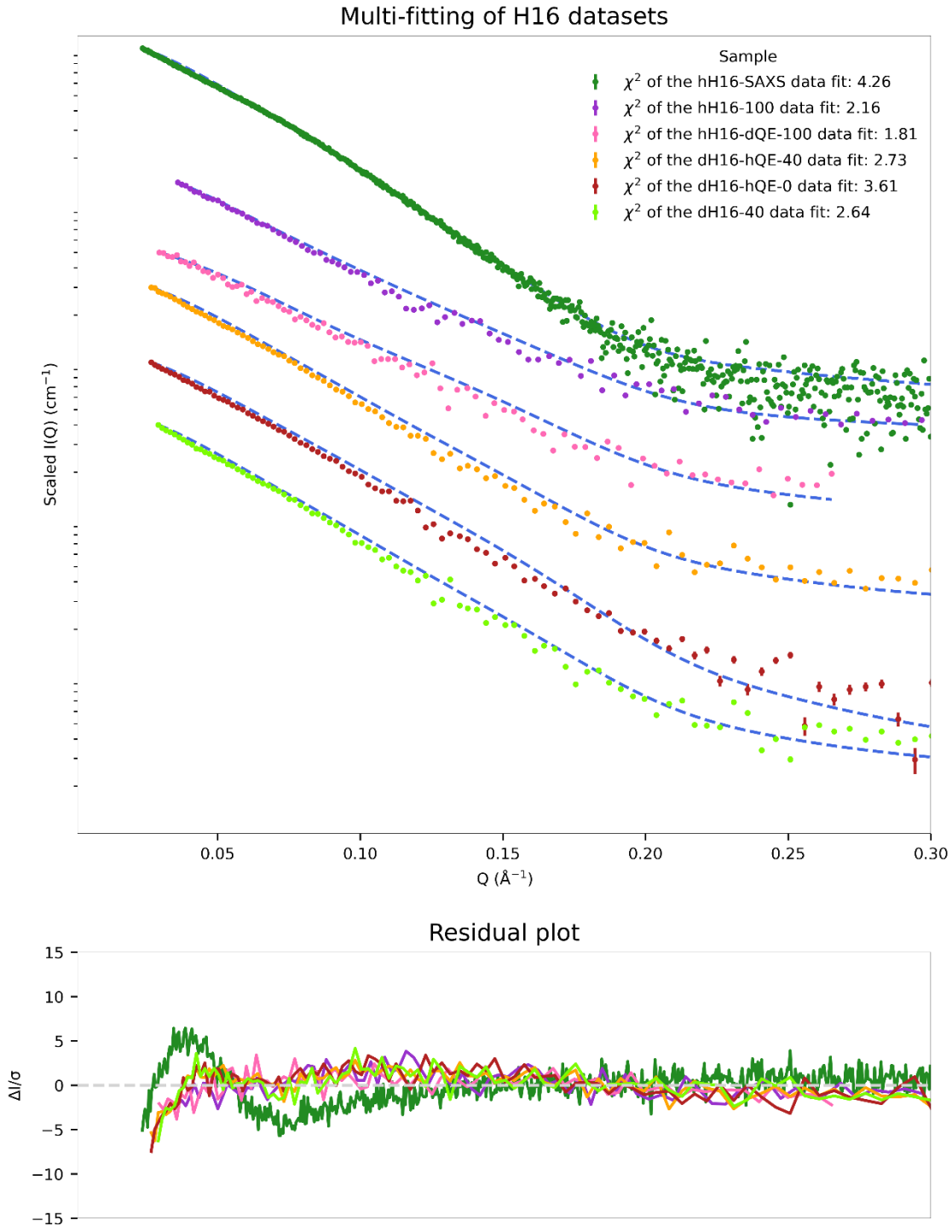


Figure 6.8: Multiple fit of six datasets simultaneously. The fitting χ^2-value of SANS datasets remained stable under addition of datasets, while the SAXS dataset fitting was slightly worsened. The initial points of the residual plot showed slight deviation, which was most pronounced in the SAXS data.

The R_g distribution of the combined dataset showed a similar trend to that of the triple dataset combinations: a broader distribution that did not significantly narrow the protein size of the sub-ensemble. The distribution was slightly smoother than that of the two described triple datasets.

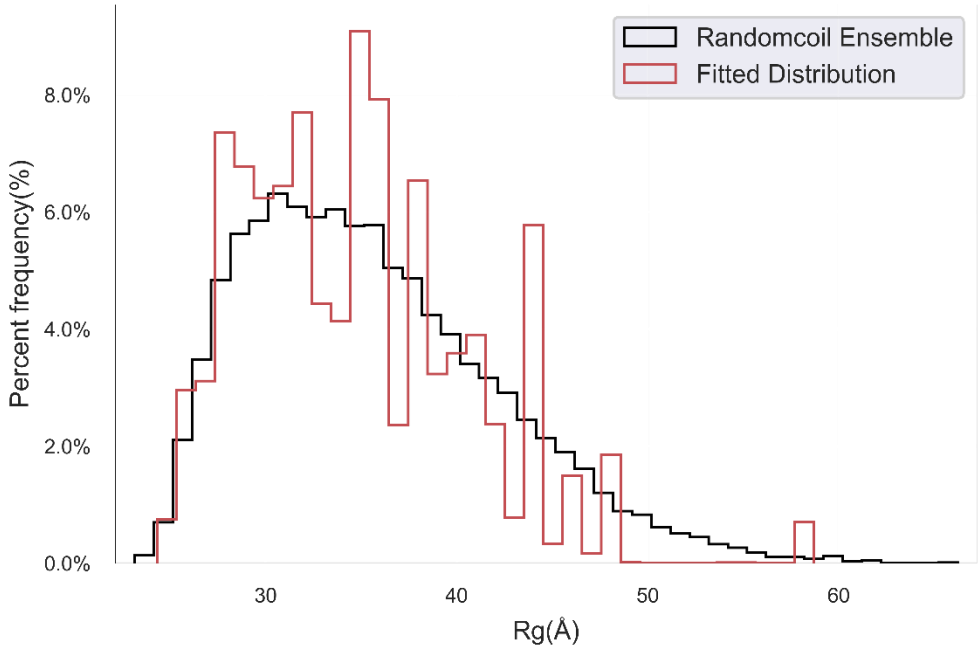


Figure 6.9: The R_g distribution of the sub-ensemble fitted to six experimental datasets simultaneously.

To further evaluate the chosen sub-ensemble of the six-dataset fitting, the 50 atomic structures were overlaid to visualize any specific tendencies (figure 6.10). From the figure it was observed that a mixture of extended structures with some alpha-helical content in the poly-Q region was present in several of the chosen structures. The large R_g-range, which was identified by the distribution (figure 6.9), was also visible in the overlaid structures. Due to the broad distribution of structures it was difficult to draw conclusions in regard to the behaviour of the poly-Q by analysis of the H16 data alone.

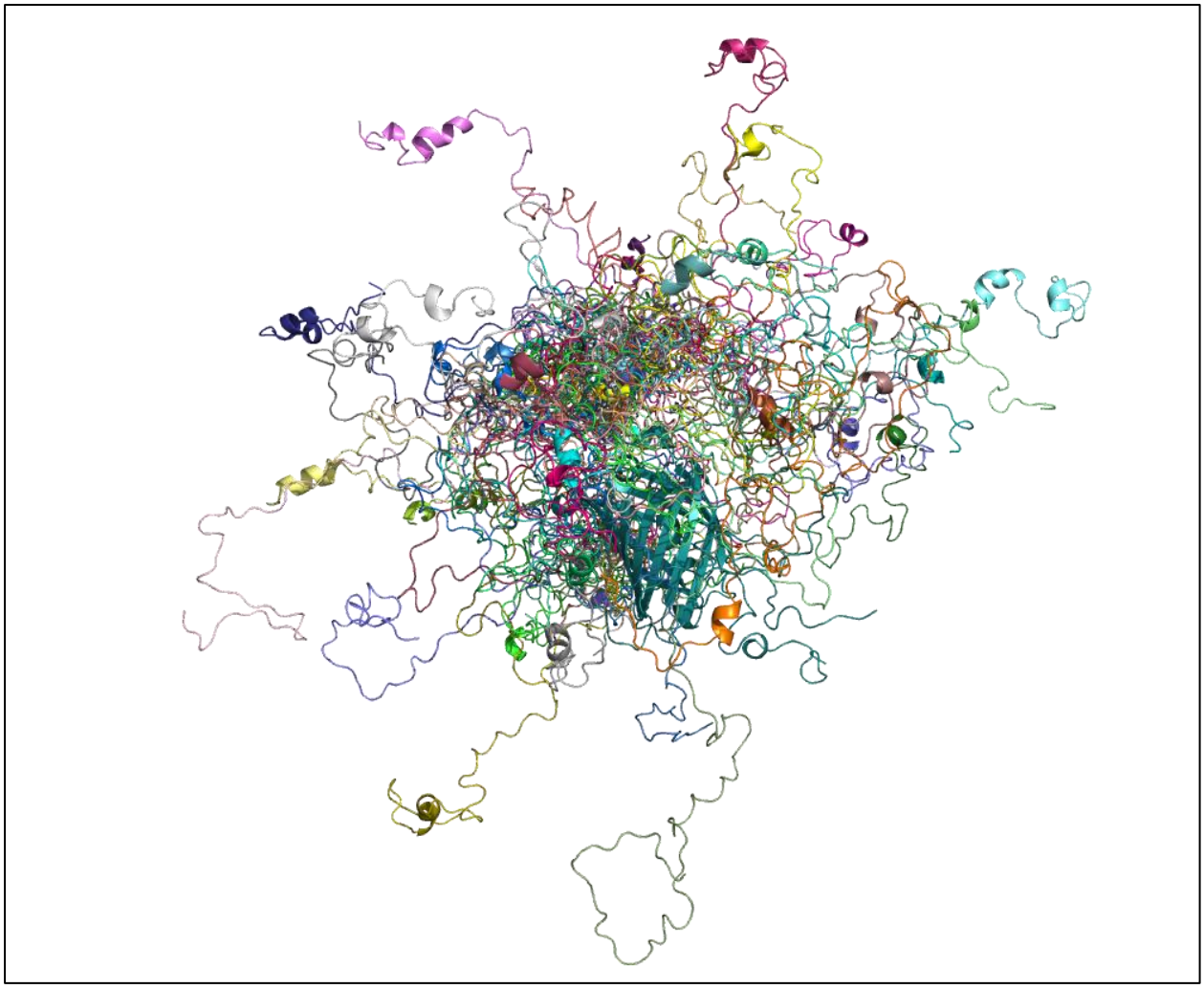


Figure 6.10: The 50 atomic structures of H16 chosen as the best possible sub-ensemble to fit the six datasets simultaneously. While no specific tendency of the entire sub-ensemble was identified, the group contained several extended profiles with little to no alpha-helical content in the poly-Q region.

In an attempt to explore the combination of the SANS profiles, the five best performing SANS scattering profiles were fitted simultaneously without the SAXS data. The chosen sub-ensemble fit the experimental datasets with a good fitting coefficient (figure 6.11).

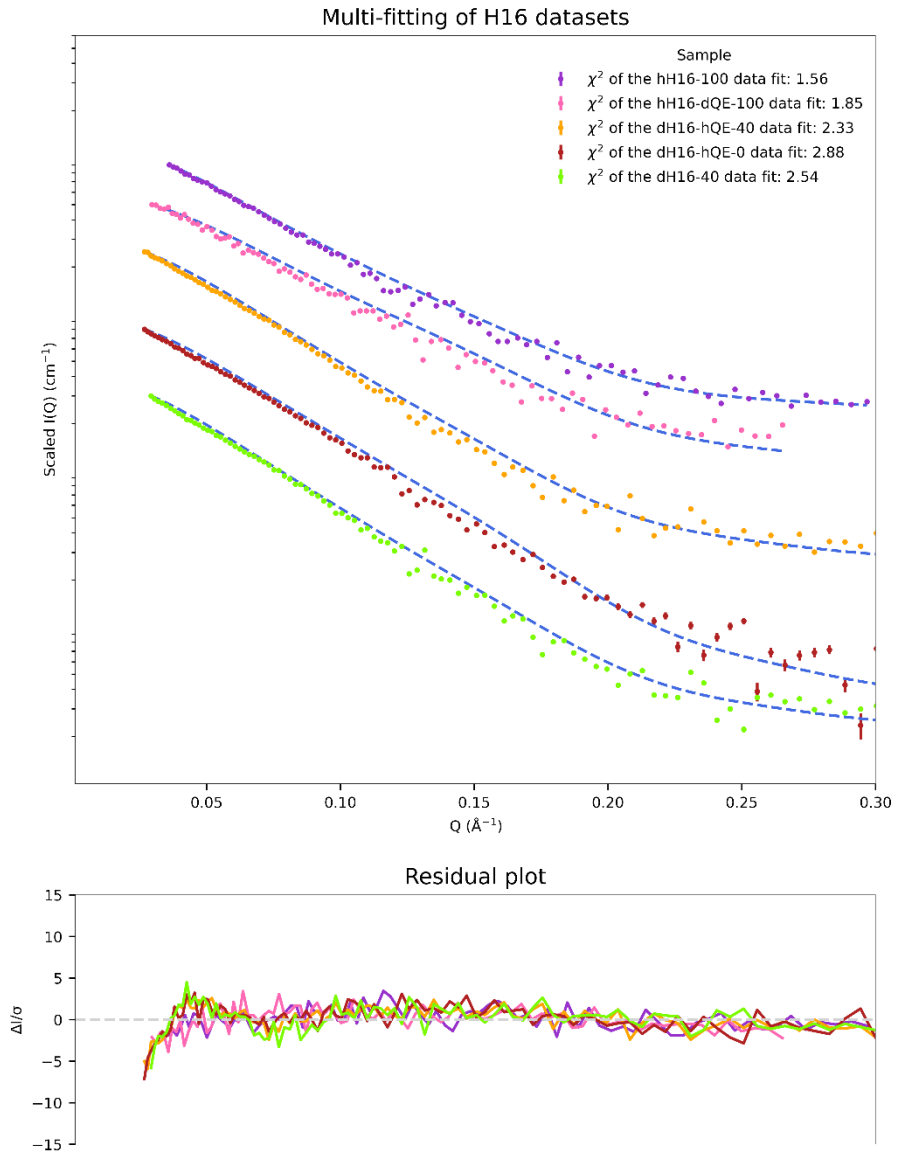


Figure 6.11: Five SANS datasets fitted simultaneously. The fitting coefficient (χ^2-values) of each fit was better than those previously observed when the fit included SAXS data.

The fitting values were all slightly better than those of the fit combined with SAXS data and the residual plot of the fit showed an excellent distribution around 0 with a slight deviation at the lowest angles, as observed when combined with the SAXS data. This slight improvement did not significantly change the size distribution of the chosen sub-ensemble (figure 6.12). The R_g distribution showed a smooth distribution of structures, which encompassed the entire range of sizes of the theoretical ensemble.

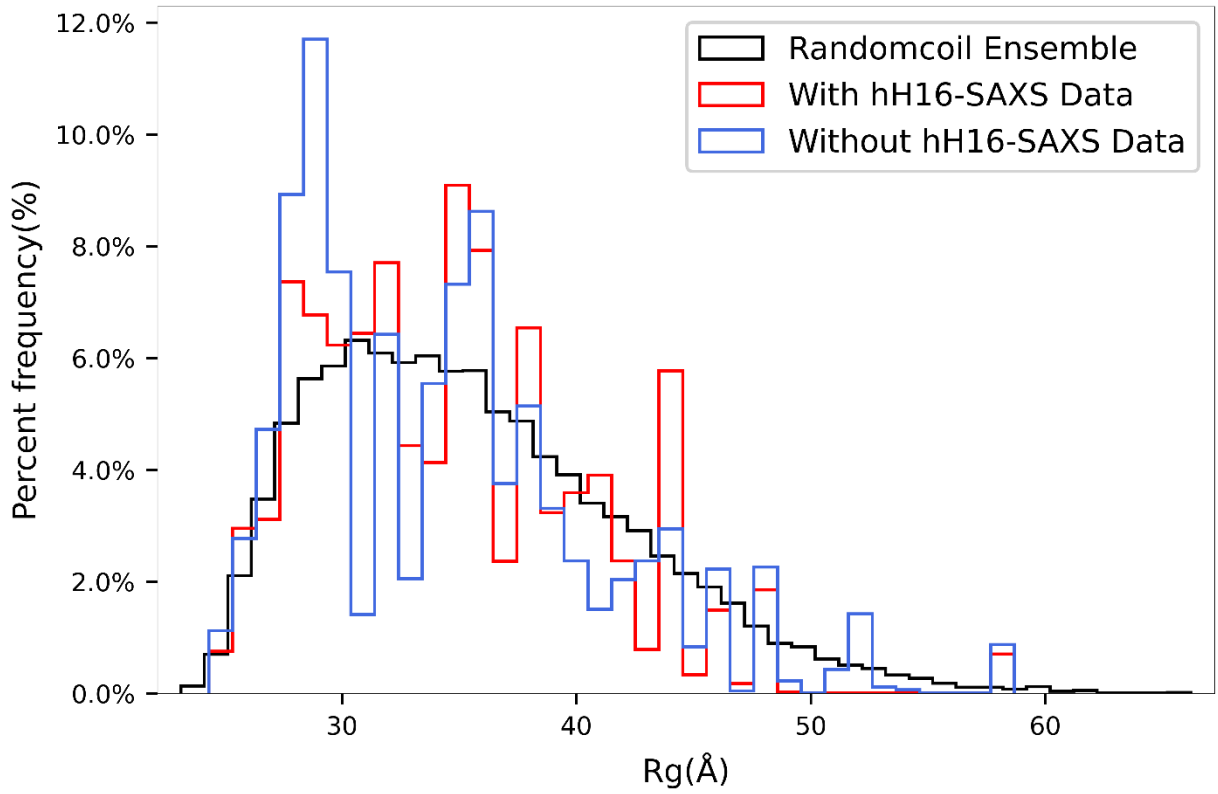


Figure 6.12: R_g distribution of the fit combining five SANS datasets (Blue) compared to both the full ensemble (Black) and the sub-ensemble refined combining SAXS and SANS data (Red). The smooth distribution encompassed the entire size range of the theoretical ensemble showing no significant size bias.

6.5 CROSS-VALIDATION OF EXPERIMENTAL H36 DATA

The number of H36 datasets available was significantly smaller, which limited the scope of the cross-validation and subsequent multiple-curve analyses. Additionally, the majority of the H36 data exhibited a lower signal-to-noise ratio than that of the H16 data.

The EOM analyses of the H36 datasets were run identically to those of the H16 datasets. Generally, the EOM fit (χ^2-work) and regenerated comparison values (χ^2-free) were low, which suggested a good correlation between the experimental datasets. The low χ^2-work values of the individual fit of SANS scattering data suggested that all four datasets could be explained by the structures of the ensemble, but, as mentioned in the previous chapter (chapter 5), most of the H36 samples had a low signal-to-noise ratio, which could also be the reason for the good fit. In other words, while the ensemble refinement fitted the experimental data well (low χ^2-work), it might have been because the data were not discriminatory enough to differentiate between the sub-ensembles of atomistic structures.

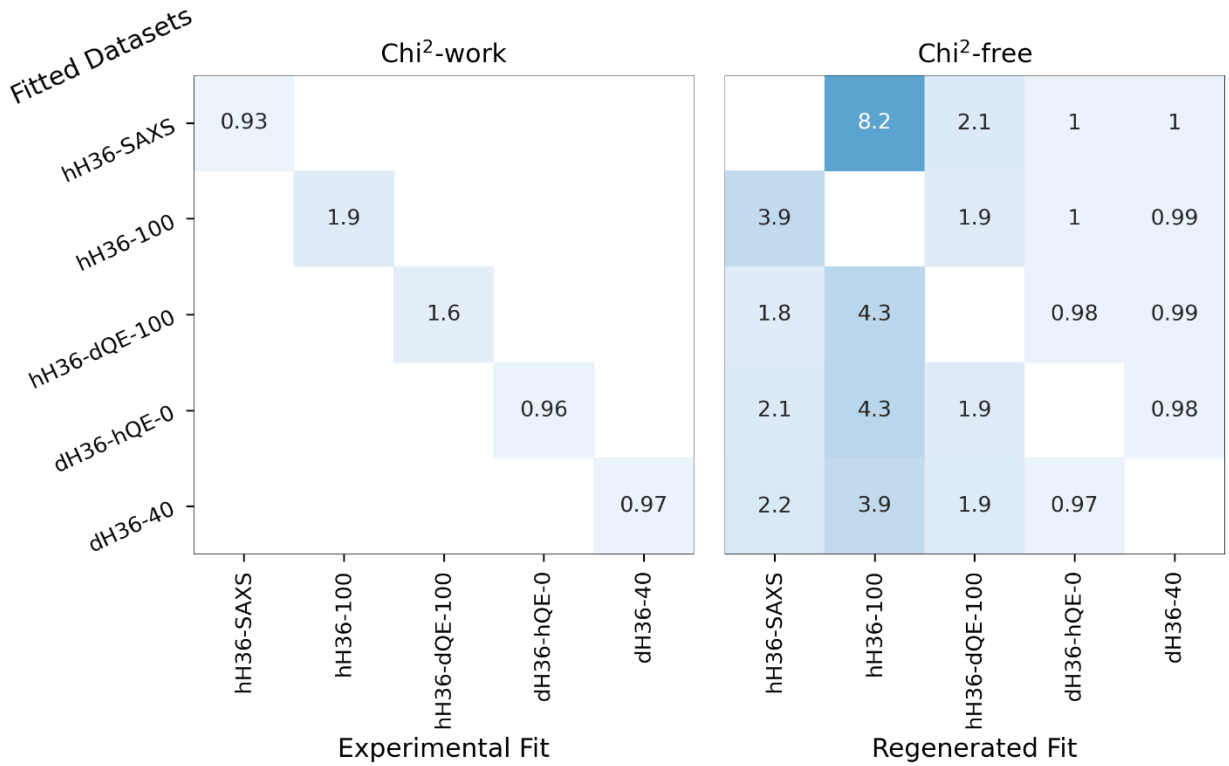


Figure 6.13: Cross-validation analysis of single dataset fittings of H36. The regenerated χ^2-free of dH36-hQE-40 and dH36-0 were both very low (<1.0) and stable, suggesting the structural ensemble was not significantly improved by the addition of these data.

A clear observation was that the χ^2-free of hH36-100 was poor when the regenerated fit was based on the hH36-SAXS data (χ^2-free = 8.2). This was very similar to the previously described effect between the hH16-SAXS and hH16-100 samples (figure 6.2).

The hH36-dQE-100 showed low variation between the χ^2-work and χ^2-free values. This could be perceived as the dataset only having a slight discriminative effect on the structures selected by the sub-ensemble. The low signal-to-noise ratio of the hH36-dQE-100 sample was expected to be the cause of the low information impact of the data.

The very low and stable χ^2-work and χ^2-free values of the dH36-hQE-0 and dH36-0 samples suggested that these samples did not provide any structural information to the analysis, which was likely due to the very low signal-to-noise ratio of the scattering profiles. Similar tendencies were observed in both two (figure 6.14) and three (figure 6.15) dataset cross-validation analyses. Furthermore, the hH36-SAXS data fit (χ^2-work) was not impacted by the addition of these two datasets, yielding χ^2-values below 1.0. Again, this feature suggested that the two datasets in question did not afford any additional structural information.

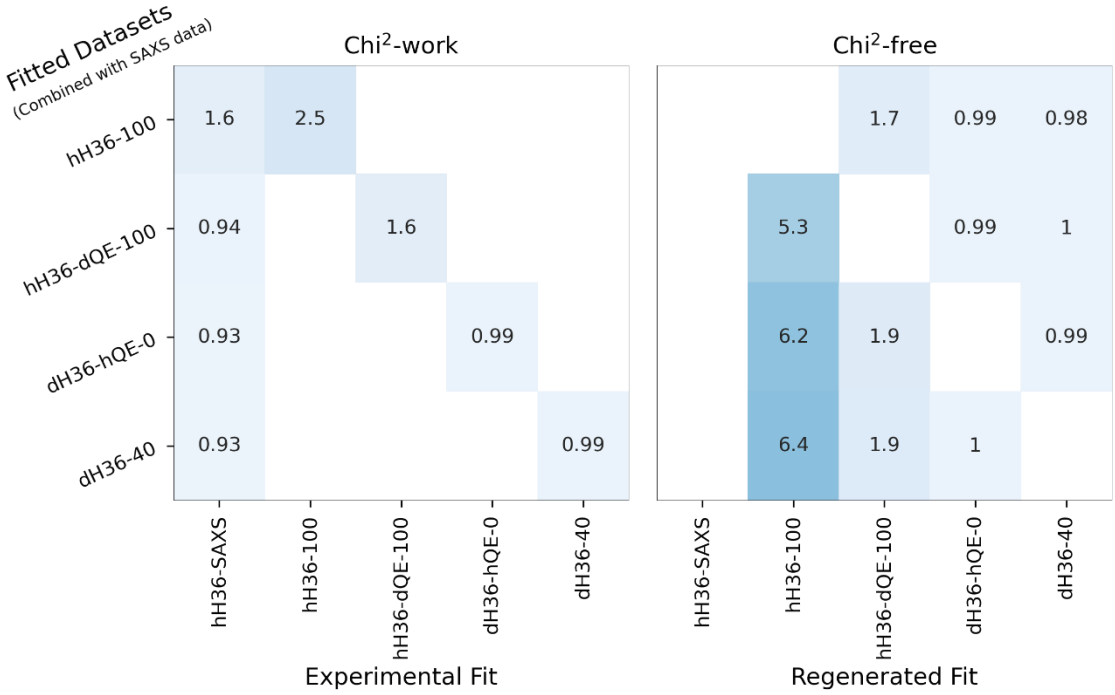


Figure 6.14: Two-dataset cross-validation analyses of H36. A very similar tendency as single dataset fitting was observed with unreasonably low χ^2-free values calculated for the dH36-hQE-40 and dH36-0 data.

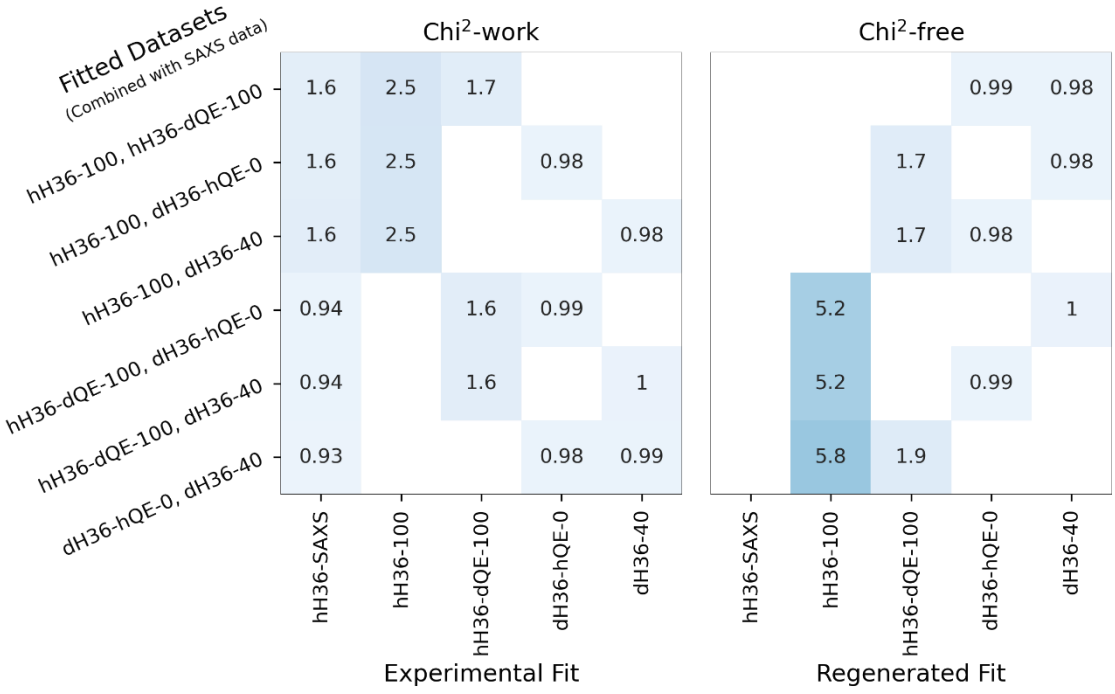


Figure 6.15: Three-dataset cross-validation analyses of H36. Results similar to those of the single and double curve cross-validation analyses of the H36 were observed.

The combination of the hH36-SAXS and hH36-100 samples showed a similar tendency as for H16. The EOM fit was constrained by both datasets slightly increasing the fitting values (χ^2-

work) from the single-curve fits, but providing a combination of the datasets with a correct simultaneous description. The combination between hH36-SAXS and hH36-dQE-100 showed very little impact on the X-ray scattering χ^2-work value, yet the combination improved the χ^2-free value when compared to other SANS data (the χ^2-free of the hH36-100 sample decreased from 8.2 to 5.3 when the SAXS data was not fitted alone). While the χ^2-work and χ^2-free values of hH36-dQE-100 were both fairly stable (maximum differences of 1.6 to 2.1), the slight impact on the regenerated fit suggested that the sample did encompass some structural information and provided some discriminative effect on the structures of the sub-ensemble. From the cross-validation analyses the following three conclusions could be derived:

1. hH36-SAXS and hH36-100 could be combined complementarily.
2. The present hH36-dQE-100 sample did provide a very slight impact on the refinement, but a higher quality dataset (higher signal-to-noise ratio) could contain important structural information.
3. Due to the low signal-to-noise ratio, the dH36-hQE-0 and dH36-40 samples did not provide any structural information to the ensemble refinement of H36.

The cross-validation analyses left only three datasets of H36 (hH36-SAXS, hH36-100, and hH36-dQE-100) for the combined multiple curve analysis.

6.6 Multiple Curve Fitting of H36 datasets

The combination of the three datasets during the EOM fitting provided correct χ^2 values, but also portrayed problematic fitting tendencies (figure 6.16). The SAXS dataset was fitted well, but the fit of hH36-100 showed deviations in the 0.1 - 0.2 $Å^{-1}$ range. The fit to the hH36-dQE-100 scattering profile showed a clear offset almost for the entire Q-range, which suggested that the EOM analysis cannot properly accommodate this profile. This disagreement was likely caused by the low signal-to-noise, which does not restraint the conformational ensemble.

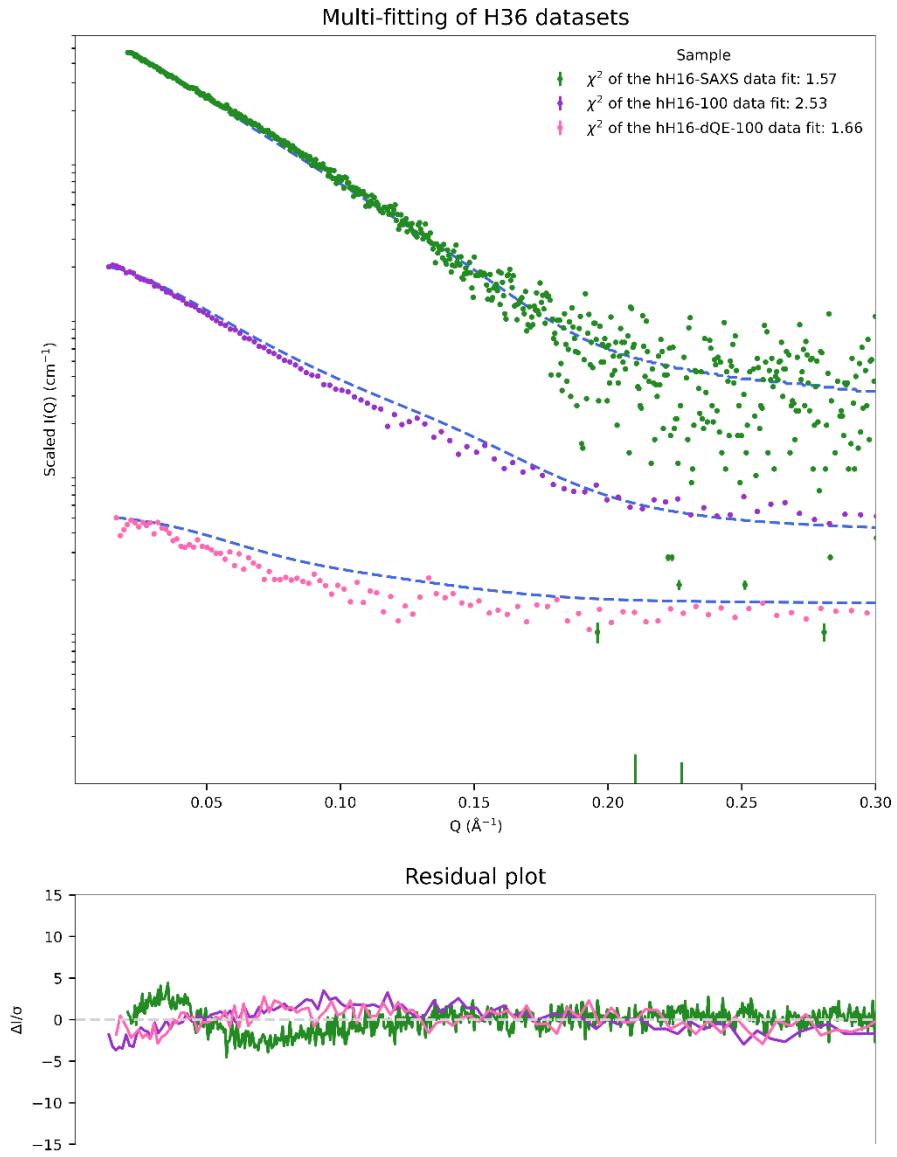


Figure 6.16: The multiple fit of H36 data combining three experimental datasets. Although the agreement was numerically correct, the hH36-100 fit showed a significant deviation around 0.15 Å^{-1} and the fit of the hH36-dQE-100 sample was generally above the scattering profile.

The fit of the three datasets of H36 data suggested that the scattering data could be combined despite the low sample quality of some of the H36 samples. The R_g-distribution of the simultaneous H36 dataset fitting was broad (figure 6.17), but still slightly more confined than that of the hH36-SAXS fit alone (Chapter 5, figure 5.8). The surprising observation from the R_g-distribution was the bimodal nature of the distribution with two maxima around 30 Å and 38 Å, which seems unphysical in the context of a highly disordered protein.

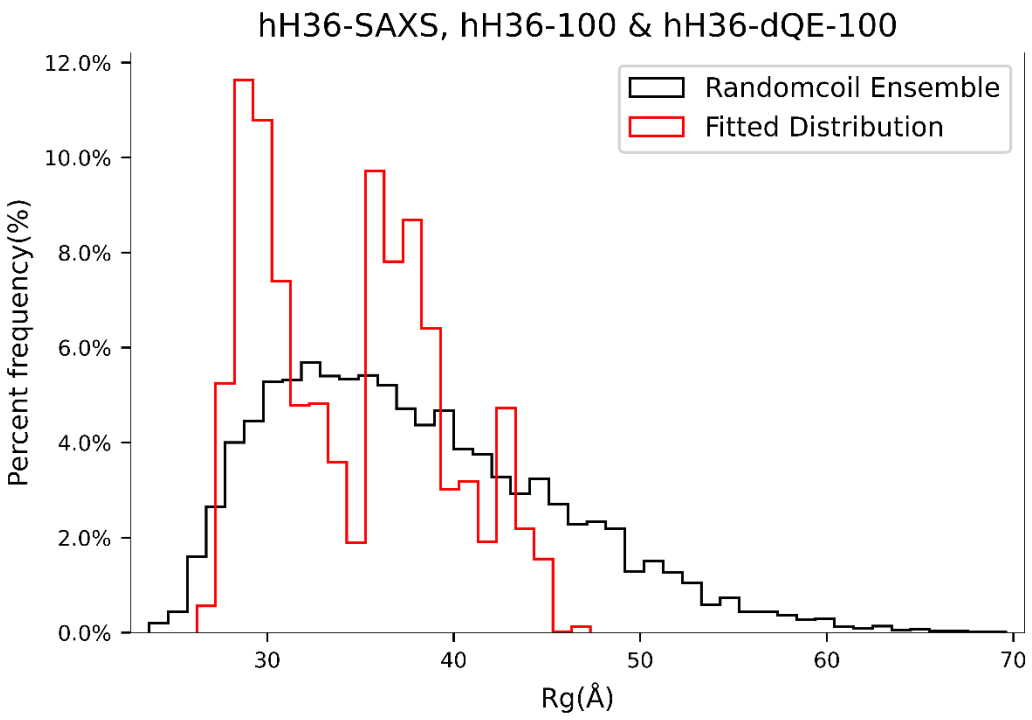


Figure 6.17: R_g-distribution of the sub-ensemble of structures refined by multiple fitting of three H36 scattering dataset. The fitted sub-ensemble (Red) was slightly more compact than that of the full initial ensemble (Black). The sub-ensemble showed a bimodal distribution.

This analysis offers, however interesting perspectives. If the sample concentration and scattering intensity of the H36 samples had been increased, the multiple fitting analysis could have likely provided useful structural information for refining the structure ensemble of H36. This was hinted at by the low impact changes caused by the incorporation of the hH36-dQE-100 sample, a dataset with very low signal-to-noise ratio and a poorly defined Guinier region, which still improved the cross-validation to the hH36-100 sample.

6.7 CONCLUSION OF SIMULTANEOUS STRUCTURAL ANALYSES

It was difficult to compare the results of the H16 and H36 multiple SAXS/SANS curve fittings. For H36, a narrower distribution of R_g-values than for H16 was observed, but this could have been caused by the lower number of scattering datasets of H36 and the hH36-dQE-100 sample being not discriminative enough. The H16 samples provided a broader distribution when fitted simultaneously, yet the compatibility of the datasets was better and the profiles generally had a higher quality (better signal-to-noise, linear Guinier regions, and meaningful *p(r)* functions to estimate robust D_{max}-values). The exclusion of hH16-SAXS from the multiple fitting did not significantly modified the refined ensemble. This showed that the SAXS data did not dominate the multiple SAXS/SANS curve fitting with EOM.

The H16 multiple fitting suggested that structural information could be gathered from the combination of SAXS and SANS datasets, but that it requires high-quality data. In the context of the flexible and disordered Huntingtin Exon-1 fused to structured sfGFP, structural tendencies are more difficult to be extracted. However, more importantly, the presented strategy of specific labelling, atomistic modelling and simultaneous SANS/SAXS data analysis was theorized to be highly informative for other biomolecular systems, which would not present the same limitations as the Huntingtin protein constructs.

7 DISCUSSION

7.1 SUMMARY OF THE KEY RESULTS

The main aim of the project was to investigate the structural features of huntingtin exon1, specifically the differences between the pathogenic H36 and non-pathogenic H16 constructs (3). The strategy centred on residue-specific deuteration of the low-complexity regions of the exon1 protein: the poly-Q tract and the proline-rich region. The deuteration was done using cell-free expression to control the specific amino acid composition during expression (215,258). The cell-free (CF) expression method was expected to allow specific homogeneous labelling of the protein, while producing adequate expression yields to subsequently perform small angle scattering (SAS) experiments (215,231,232,236,257).

To analyse the experimental scattering data, an exhaustive atomistic structure ensemble was produced to fit the scattering patterns using the ensemble optimization method (EOM) (183). To investigate the structural tendencies of the two protein constructs, multiple SANS and SAXS samples were measured and analysed simultaneously using the theoretical structure ensemble.

The project showed that CF expression was an adequate method to produce samples with the desired labelling. The samples were focused around deuteration of glutamine, which was one of the problematic residues in terms of the expression due to amino acid scrambling (229,258). Yet, the method proved to be applicable with only minor changes to the expression recipe: 1) both glutamine and glutamic acid were simultaneously deuterated; 2) the pH-buffering component of the CF mixture was exchanged from potassium glutamate (KGlu) to potassium acetate (KOAc). The trade-off was a lower expression yield, but the labelling pattern of the produced protein was confirmed by MS. Similarly, samples containing deuterated proline and nearly perdeuterated samples were produced using the same expression method with the buffer component chosen depending on the protonation/deuteration of glutamine.

The atomistic ensemble of H16 and H36 was produced with the labelling patterns being the focus of the calculations (165). Because the software used could not differentiate between the specific labelling pattern and the equilibrium of labile deuterium in a solution (173), a script was produced to manually produce the labelled protein structures to a given D_2O buffer percentage. This allowed for the production of 5,000 initial structures times 8 labelling patterns times 6 D_2O solution level equalling 240,000 structures for both H16 and H36. The ensemble

approach has the advantage that a single structure could be tracked between each of the labelling patterns and buffer solutions. This allowed to calculate EOM fits not only of multiple datasets of the same protein, but of multiple datasets of separate labelling patterns or experimental conditions (183). From the theoretical data, it was shown that combining multiple datasets could improve the structural information that could be gathered by comparing the H16 and H36 data.

High signal-to-noise datasets were obtained from SEC-SAXS and showed minimal impact of the addition of D_2O to the solution buffer (119,134). The SEC-SAXS profiles were used as controls during the multiple dataset fitting. The initial SANS measurements provided reasoning for the need of using the SEC-SANS setup of the D22 beamline (137); the reason being to avoid oligomers and/or aggregation during the measurements. Note that huntingtin exon-1 is an aggregation-prone protein, especially in its pathogenic forms (23,65).

The H16 protein with several labelling patterns could be measured with a good signal-to-noise ratio, showing the ability to produce homogeneously labelled samples. Some of the specific experimental conditions identified as valuable during the theoretical calculations could not be measured in a realistic experiment due to small differences between expected signal and background scattering of the experimental setup (primarily H_2O incoherent scattering of the solution) (93,123,127,131). The measured profiles of labelled proteins suggested that different information was revealed from different labelling, which was the expected outcome. Production and subsequent SANS measurements of H36 yielded fewer datasets with a significant lower signal-to-noise ratio due to the majority of the project focusing on getting a robust analysis of H16 with the limited experimental beamtime available.

The multiple fitting analysis proved the theory that datasets across labelling schemes could be combined and compared between the experiments by the theoretical ensemble bridging the labelling patterns (183). Unfortunately, the aim of comparing the H16 and H36 could not be accomplished due to the low quality of the H36 data.

The discussion of this dissertation will cover each of the four major chapters individually before integrating the findings into a wider perspective.

7.2 INTERPRETATION OF THE THEORETICAL AND EXPERIMENTAL RESULTS

7.2.1 Cell-free expression

From the cell-free expression performed, several labelling patterns of H16 and H36 exon1 were successfully produced. In contrast to previous projects performed in our group, the samples were not used for NMR studies (65,91,92,225), but aimed at recording small-angle neutron scattering data from the protein. This increased the requirements to both quantity and purity of the protein samples and, while the labelling process was fairly simple to implement, the yield was a constant challenge throughout the project (122,137).

As mentioned in chapter 3.1, the lysates produced during the project were tested for the optimal magnesium acetate level. This test did, however, also result in the observation that lysates grown from the same bacterial stock behaved slightly differently with regards to yield, when different proteins were expressed (215). The ratio of yield between H16, H36 and unrelated proteins differed slightly between each batch of lysate suggesting an unknown connection between the production of lysate and the subsequent expression yields, which is extremely challenging to elucidate. This observation could have been delved further into if the project had centred around the expression of different proteins using this method. Despite the slight differences in yield, the lysates could be used to produce the protein samples during the project.

For the development of the project the use of commercial lysates would have resulted in extremely expensive samples. Take note that commercial CF kits, such as Expressway™, costs 678 EUR per 1 mL of reaction mixture. Consequently, a 24 mL CF reaction, as used for some of the samples of this project, would have cost above 16.000 EUR, depending on the kit. Additionally, deuterium labelled amino acids have to be added to the mixture depending on the labelling pattern at a concentration of 1 - 2 mM of the reaction mixture. Therefore, the CF expression method was only viable for SANS studies, such as the one presented here, when the *E. coli* lysate is produced in-house, as expression volumes varied between 20 - 48 mL CF mixture depending on the sample, requiring from 4.5 up to 10.8 mL of cell lysate.

While almost full perdeuteration could be accomplished using algal growth media such as ISOGRO® (contains 16 deuterated amino acids), the more controlled samples required single deuterated amino acids to be incorporated into the protein homogeneously (215,279). Additionally, ISOGRO® contains deuterated glutamic acid, which would have caused problems during the project due to the isotope scrambling between glutamic acid and glutamine (258). ISOGRO® or similar algal extracts could have been used to produce the perdeuterated

samples of Htt Exon1 at a cheaper material cost by manually adding the four remaining amino acids to the CF expression medium, but would also have required optimization (280).

Despite the low yields, samples of several labelled H16 and H36 constructs were produced with only a single (H36 D-QEP) sample showing a significant problem (two separate MW) when verified by MS, which was attributed to the combination of two different protein expressions to reach a measurable protein concentration (122,123). For the rest of the samples, the MS experiments confirmed that the samples were produced with a molecular weight matching the calculated weights after a successful deuterium labelling. In addition to the labelling pattern, the MS also meant that isotope scrambling was avoided. This result suggested that any pattern of labelled Htt Exon1 could be produced with an adequate purity to measure scattering data. The choice of samples produced could then be based on simulations of theoretical data ahead of experimental beamtimes.

Because CF allowed for the production of the disordered low-complexity protein with the correct labelling and homogeneity, the method could be used to produce other proteins with similar labelling (215,221,257). Additionally, if Gln/Glu and Asn/Asp labelling is avoided and the protein purification is optimized, SANS samples could be produced from as little as 20 mL CF mixture (258). This could allow for projects focusing around optimization of SLD with regards to regions of low complexity or internal protein structure. While the method would not be cost-effective for classic match-out SANS studies (202,224), biasing scattering data by selective labelling could be explored for large and/or structured proteins produced by CF (206,209,210).

7.2.2 Theoretical ensembles of atomistic structures

The theoretical ensembles calculated during the project focused on accommodating the specific labelling patterns in a single pipeline in order to ensure that a single structure could be followed in several labelling patterns and across solution D_2O levels. This was accomplished via the scripts written during the project in combination with the scattering pattern calculations of the ATSAS software Crysol and Cryson. The list of patterns and levels of D_2O solutions limited the number of different ensembles produced because a single ensemble of 5,000 structures ended up containing 240,000 scattering patterns when all conditions were calculated and took considerable time to complete.

The theoretical scattering patterns provided insight into the effect of deuteration of the low-complexity regions of the Htt Exon1 protein. This was visibly discernible from both the average

profiles of each condition (labelling pattern and solution) and from scatterplots comparing the theoretical SANS profiles to the theoretical SAXS profile of the same structure. These results showed that the SANS profiles from samples of labelled protein calculated close to the matching point of either the protonated or deuterated protein (129), would provide the most structural information. This is a logical and expected result as the labelling becomes most visible when the unlabelled protein signal decreases. The difficulty comes when transferring the theoretical data to the planning of experiments. Although the protein could be produced in any labelling scheme, the scattering signal of a low signal-to-noise sample could not be separated from that of the incoherent background scattering, making the resulting curve unusable for structural purposes.

The choice of probing different initial theoretical ensembles could be an interesting perspective of the present study. The produced ensemble consisted of structures with only a very short imposed α-helical conformation (residue 10-14 of the exon1). A combination of structures with additional forced alpha helical content in the poly-Q region could provide a broader scope of movement of the exon-1 (65). Similarly, an ensemble already refined with other data, such as NMR, could be used as the base ensemble in which case the fitting would focus more on adding additional value on top of the already refined ensemble (65,182).

The scripts written during the project could be used for probing structural information from other protein systems, such as multi-domain proteins containing flexible linkers. In order to expand the methodology beyond LCR proteins, which has clear targets related to the strategy of this project due to the compositional bias, the scripts could be automated to test the impact of selective residue labelling in high-complexity regions. By linking the labelling and exchange scripts (Appendix 11.1 & 11.2) to a scattering profile calculation (such as CRYSON), conditions, which differed significantly from unlabelled samples, could be sought. Additionally, the solution D_2O level could be used as an optimization parameter with an automated screening of all percentages, instead of only implementing the specific intervals (0, 20, 40, 60, 80, 100) utilised during this project. The major limitation for automation would be the ensemble size needed for disordered proteins, although the need of a large ensemble to have predictive profiles remains to be explored. In summary, multi-domain structures with flexible elements could potentially be screened with regards to residue deuteration and solution contrast to find optimal conditions for improving the structural information coded in SANS data.

7.2.3 SAXS and SANS scattering experiments

SAXS data recorded at the Swing beamline (Soleil Synchrotron, Paris) provided two important observations; 1) SAXS data could be recorded in both H_2O and D_2O with minimal difference suggesting that the presence of D_2O in the buffer did not significantly induce structural changes or aggregation; 2) high signal-to-noise ratio SAXS data could be measured by SEC-SAXS without signs of oligomerization. While the SAXS and SANS methods have significant differences (i.e. exposure time, degree of sample degradation, etc.) the measurements provided an important insight into the feasibility of the subsequent, planned neutron experiments (204).

The experimental setup of the D22 instrument proved capable of measuring the otherwise difficult samples of Huntingtin Exon1 (122). As described during the introduction, the flexibility of the Exon1 domain of the Htt protein was a known limitation of structural studies of poly-Q tracts. The flexible protein was additionally known to form oligomers, especially in the case of pathogenic poly-Q lengths, which could be a problem during measurements (23,66,281). The size-exclusion chromatography SANS environment built at the D22 beamline represents a unique setup for the measurements of samples of both non-pathogenic and pathogenic Htt without samples showing any sign of aggregation (122,137).

The scattering profiles of H16 with different labelling schemes suggested that data could be recorded for each of the labelling patterns, providing that the protein concentration and contrast was sufficient. The protein samples with a low concentration ended up needing a long exposure time in order to provide good statistics of the scattering profile (127). This prolonged exposure could cause aggregation during the measurement, as the protein was kept in the sample cuvette during the experiment. This was not a problem in any of the experiments recorded using the SEC-SANS method, which suggested that the purification step just prior to the sample exposure was sufficient to measure H16 and H36 samples. The length of exposure could be shortened if the protein concentration of the samples was improved. Due to factors such as limited beamtime availability and challenges with protein production yield, H16 protein constructs were subject to more measurements than the H36 protein.

The H16 D-P and H36 D-P samples were all prepared with a good protein concentration and all showed a high signal-to-noise, but the data could not be subsequently fitted. This was not a problem of the SANS measurements, but rather a structural effect resulting from the deuterated proline incorporation. The effect was present in both H16 and H36, showing similar profiles between the two constructs, and when measurements of H16 was repeated, ruling out a

punctual protein production error. The impact of proline deuteration could, theoretically, be linked to the small differences between both isotopes affecting the proline rich region (273,274). The PRR of Huntingtin is generally described as a rigid region (65), meaning any bond changes could have an impact on the structure of the exon-1. This unexplained phenomenon had the unfortunate effect of removing a significant portion of the measured experimental scattering data from our further structural analysis.

7.2.4 Multiple Fitting Analysis

Cross-validation of small angle scattering data through EOM provided unique insights into the compatibility and structural information content of the different dataset. The 9 datasets (1 SAXS and 8 SANS) measured for H16 allowed for the comparison of separate and simultaneous fitting of multiple combinations. The χ^2 of the fit ("χ^2-work") increased slightly when multiple datasets were fitted, compared to singular dataset fitting. The slightly increased χ^2-work values were attributed to the simultaneous fitting of several datasets having to accommodate several scattering patterns and thus increasing the number of degrees of freedom. As each dataset was weighted equally based on their experimental error none of the datasets were preferentially fitted by the genetic algorithm (183,282). Importantly, if each dataset had yielded the same sub-ensembles, the multiple curve fitting analysis would not have provided any additional structural information compared to that of the single dataset fittings. The theory of each labelling pattern being able to provide different structural information implies that combining experimental data during the fitting process changed the χ^2 value.

The cross-validation against the non-used datasets (χ^2-free) showed an interesting evolution when moving from 1 to 3 simultaneous fitted datasets (278). The χ^2-free values were not improved when the SAXS and one SANS dataset were fitted simultaneously. Fitting three datasets instead of two did, however, improve the calculated χ^2-free. Ensembles fitted with the SAXS data exhibited a poor capacity to describe the SANS data, but the X-ray data was able to be fitted properly in combination with the SANS data. Comparing the H16 and H36 cross-validation showed the impact of structurally informative datasets in the multiple fitting analysis, with H36 showing very stable and excessively low χ^2-free values, which has been attributed to the poor signal-to-noise ratio of most H36 experimental datasets and the samples providing very little discriminative information.

The simultaneous fitting of five and six H16 datasets with EOM resulted in very broad R_g distributions compared to those of the single SANS dataset. Although a broader R_g distribution

was expected, the resulting sub-ensemble suggested the co-existence of conformations sampling a broad size range. All five labelling patterns of SANS data incorporated into the multiple analysis showed slimmer sub-ensemble size distributions when fitted individually, while the SAXS data resulted in a similarly broad distribution.

The structural sub-ensemble refined by EOM did not show any clear tendencies towards a certain type of structure, and no further insight into the specific structure and dynamical behaviour of H16 was gained. While the specific structure of H16 was not clearly elucidated from the analysis, the methodology of combining SANS and SAXS data resulted in a better complementarity between the two types of experiments according to the cross-validation. Hints that improved sample quality and stability could result in improved structural informational content.

The H36 samples provided very few useful scattering profiles and as such could not be compared to that of the H16 datasets. The low impact of the hH36-dQE-100 sample on the cross-validation suggested that even low signal-to-noise data could encompass some useful information for ensemble refinement. This could be improved with higher quality samples/scattering experiments.

7.3 Findings Across Chapters

The CF expression method was paired well with the theoretically labelled atomistic structure ensembles. The optimized expression conditions suppressed scrambling and allowed for the expression of protein samples with Q/E and P labelling. The use of theoretical ensembles could be used to optimize the deuteration pattern to increase the information content and reduce production costs. The difficulty of the presented project came from translating the information gathered from the theoretical ensembles into experimentally measured data.

The combination of poor sample stability and tendency towards protein aggregation, which forced the incorporation of sfGFP to the protein construct, played a significant role in decreasing the inherent information of the SAXS and SANS data, due to the highly structured, globular protein subunit linked to the flexible exon1 domain. This was inferred from the Kratky plots of the SEC-SAXS data obtained during the project, which showed relatively globular proteins, even if the exon1 was known to be intrinsically disordered. In the event that Htt Exon1 could be produced in stable and monomeric form, the impact of the labelling schemes would be increased because the deuterated amino acids would be further restricted to the labelled

regions of the protein domain. The small size would, however, require an even higher protein concentration to perform the SANS experiments. Therefore, for future applications of our technology, the size, the purification yield and the amino acid composition will have to be considered.

For structured samples one could automate the labelling and exchange scripts to optimize the informational content available from a sample. For the comparison of scattering patterns calculated for different labelling patterns and solution conditions, the EOM multiple data analysis algorithm could be modified and incorporated into the pipeline. This could allow projects to probe which labelling patterns, in which conditions, and how many combinations would offer a maximum of theoretical structural information. This would be especially applicable in regards to structured protein samples due to the ease of sampling single-structures or stable LCR proteins, which would exhibit compositional bias that could be exploited by this approach.

7.4 STRENGTH AND LIMITATIONS OF THE METHOD

The strategy outlined during the project allowed for the planning, preparation, and measurements of deuterated proteins. The calculation of theoretical ensembles of labelled protein yielded an insight into the value of a given labelling scheme, which allowed protein expression to focus on samples with a richer structural information. The use of CF expression system has led to the production of pure, homogeneously deuterated protein. The combination enabled the sample preparation of SANS samples with the chosen labelling scheme in sufficient purity to perform small angle X-ray and neutron scattering experiments.

The use of SANS to measure the amino-acid specifically labelled samples expanded on the conventional use of contrast matching. Both theoretical and experimental SANS profiles exhibited differences depending on the labelling pattern and suggested that high quality protein samples measured with different labelling combinations could be used to increase the information obtained about a given system.

The limitations of the system stemmed primarily from two main factors; 1) the low expression yield of protein samples labelled with deuterated glutamine resulted in SANS measurements with low signal-to-noise ratio, and 2) the protein stability of the Huntingtin Exon1 samples forced certain restrictions to the samples, which further impacted the signal quality of the neutron experiments.

The restrictions included:

- linking the target protein to sfGFP to improve solution stability and decrease oligomerization.
- Utilization of the SEC-SANS sample environment to purify samples of larger oligomeric states before neutron exposure.
- Limiting protein concentrations (≤6 mg/mL) to avoid aggregation during protein preparation.

If the presented methodology framework was applied to a protein system which did not exhibit all of the described constraints, the information value of the structural analysis would be improved.

8 CONCLUDING REMARKS AND FUTURE PERSPECTIVES

The presented PhD dissertation has provided results regarding the three of the major objectives of the project: 1) A robust methodology for sample production and measurements of selective amino acid labelled protein was described for low-complexity samples, using Cell-free expression, SEC purification, and the combined sample environment of SEC-SANS. 2) A strategy to create realistic, atomistic ensembles of deuterium labelled samples and the ability to perform cross-validation of SAS data was developed. 3) The labelled protein samples of H16 were measured and cross-validation suggested that complementary combination of SAXS and SANS data was achievable with good sample quality (high signal-to-noise ratio and overall signal intensity).

The first aim was establishing the strategy for expression and purification of deuterium labelled protein. Control of the amino acid composition during expression was required for residue specific labelling and the cell-free batch expression method allowed for the production of segmentally labelled sampled. Combining the addition of deuterated Gln and Glu residues to the CF mixture and exchanging the buffer from KGlu to KOAc allowed for the expression of homogeneously labelled protein samples and accounted for the Gln/Glu residue scrambling. The labelled protein samples were verified by MS to fit the theoretical molecular weights.

The second aim of developing an approach to optimally integrate SAXS and labelled SANS data to refine ensemble models of the low-complexity protein was done with a combination of new python scripts and EOM. An atomistic ensemble of the Huntingtin Exon-1 fused to sfGFP was prepared from combining generated exon-1 structures to the crystal structure of sfGFP. Labelling with deuterium and the labile H/D exchange was conducted by scripts written during the project and the ensemble refinement was done using an old version EOM. The nature of the project allowed for a structure to be translated between different labelling patterns and sample conditions and thus allowed for the use of cross-validation to compare the quality and complementary value of the SANS and SAXS data.

The third aim was to combine the two approaches to characterize the structural tendencies of Huntingtin Exon-1. The combination of SAXS and SANS data of the non-pathogenic H16 constructs proved that scattering data could be combined. This combination also proved that simultaneous analysis of the SANS/SAXS data improved the agreement between the two methods, when the atomistic ensembles were refined by a combination of datasets. The explicit structural tendencies of the poly-Q tract could not be described, although the methodology

described in the dissertation was theorized to enable higher resolution information with a less complicated system than huntingtin Exon-1.

The use of the complete strategy, from simulation to cross-validation of experimental data, is thought to be applicable to any protein system. While especially LCR proteins are opportune targets of segmental labelling, rigid multi-domain proteins with flexible linkers could be other interesting targets. A project aimed at implementing the labelling, exchange, and measurement strategy to structured protein systems has been initiated as a collaboration between the CBS and ILL. This might allow for future automation of the labelling simulation pipeline.

9 MATERIAL AND METHODS

9.1 BUFFERS

Buffers used during experiments were **A** – 50 mM 2,2-Bis(hydroxymethyl)-2,2′,2″-nitrilotriethanol (BisTris), 1000 mM NaCl, pH 7.5. **B** – 50 mM BisTris, 150 mM NaCl, 250 mM Imidazole, pH 7.5. **N** – 20 mM BisTris, 150 mM NaCl, pH 6.5.

9.2 HUNTINGTIN CONSTRUCTS OF EXON-1 H16 AND H36

All plasmids were produced according to instructions of the Maxiprep kit. Synthetic genes were cloned into pIVEX plasmids using In-Fusion® (Clontech) yielding constructs of pIVEX-(H16/H36)-3C-sfGFP-His6.

9.3 BL21 STAR (DE3)::RF1-CBD_3

The cell-line used to produce lysate for cell-free expression is a mutated strain of BL21, where the genomic release factor 1 (RF1) is tagged with three Chitin Binding Domains (CBD_3). The lysate expression was started with a pre-culture in Z-medium supplemented with kanamycin (50 μg/mL) inoculated with a culture stored at -80 °C. The pre-culture was incubated at 37°C for about one and a half hour until the OD_{600} had reached ~1. During incubation the 5 L fermenter was prepared. The fermenter with Z-medium was autoclaved the day before expression and to finish the preparation, additives (110 mM glucose, 10 mg/L thiamine, 1 mM $MgSO_4$ and 50 mg/mL) were added, stirring and oxygen pumps were started, and sensors were placed into the fermenter. Once the pre-culture had reached an OD_{600nm} of 1.0 the preculture was added to the fermenter. The fermenter was left to incubate until the OD_{600nm} reached 1.0 after which the fermenter was induced using 5 mL 1M IPTG. The expression medium was incubated until the oxygen levels stabilized (typically around 25% O_2). The bacterial cells were harvested by centrifugation, supernatant was discarded, and the pellets were washed with buffer and homogenized. The pellets were frozen at -80°C until further purification steps. The frozen pellets were thawed, and the cells were lysed using a French press. The cell lysate was subsequently centrifuged, washed, and dialyzed. The last step is a heat treatment at 45°C after which the cells were dialyzed in fresh buffer followed by flash freezing and storage at -80°C.

9.4 CELL-FREE PROTEIN EXPRESSION

For cell-free protein expression, cell lysate and the target plasmid were added to a solution of cell-free mixture. The cell-free expression mixture contains amino acids, ATP, GTP/UTP/CTP, cAMP, folinic acid, HEPES, DTT, NH_4OAc, Creatine phosphate (CP) and creatine kinase (CK), tRNA mixture, $Mg(OAc)_2$ and potassium glutamate (KGlu). The expression solution was incubated at 24℃ for 3-4 hours with agitation. After incubation the expression medium was centrifuged at 15.000 rpm for fifteen minutes to precipitate expression components. Samples containing deuterated glutamine required the expression mixture to also incorporate deuterated glutamic acid and exchange KGlu with potassium acetate (KOAc). The concentration of both KGlu and KOAc was decided after titration expression tests to optimize the protein expression. The optimal concentration of $Mg(OAc)_2$ was determined for each individually produced lysate. The recipe (Figure 9.1) was prepared before each expression to ensure the correct volumes were added.

DATE	30/08/2023		
Number of reactions	**24**		
Volume of reaction (uL)	**1000**		
Volume of 10 x reaction mix	100		
Plasmid Concentration (ng/uL)	2101		
Lysate number	T37		
Optimal Mg(Oac)2 Conc. (mM)	12.5		
	Stock C.(mM)	CF C.(mM)	Volume(uL)
amino acid mix			
water soluble aa each	50	15	493.2
acid soluble aa each	50	15	493.2
base soluble aa each	50	15	493.2
milliQ (sterile)			164.4
TOTAL			**1643.8**
10 x reaction mix	Stock C.(mM)	CF C.(mM)	Volume(uL)
25 mM rNTP each (0.8 mM)	25	0.8	768.00
2.0 M HEPES (55 mM)	2000	55	660.00
96 mM ATP (1.2 mM)	96	1.2	300.00
10 mM Folinic Acid (68 uM)	10	0.068	163.20
100 mM cyclic AMP (0.64 mM)	100	0.64	153.60
500 mM DTT (1.7 mM)	500	1.7	81.60
9.2 M NH4OAc (27.5 mM)	9200	27.5	71.74
milliQ (sterile)			201.86
TOTAL			**2400.00**
master mix	Stock C.(mM)	CF C.(mM)	Volume(uL)
10 x reaction mix			2400.00
1M Creatine phosphate (CP; 80 mM)	1000	80	1920.00
amino acid mix; 15 mM each (1 mM)*	15	1	1600.00
100 mM KGlu (4 mM)	4000.00	100	600.00
1.07 M Mg(OAc)2 (12.5 mM)	1070	12.5	280.37
17.5 mg/ml tRNA (0.175 mg/ml)	17.5	0.175	240.00
10 mg/ml Creatine kinase (250 ug/ml)	10000	250	600.00
Additive 25X anti-proteases EDTA free (1X)	25	1	960.00
TOTAL			**8600.37**
per reaction	**Volume (uL)**		
master mix	8600.37		
S30 extract T37	5400.00		
vector (16 ug/ml) Htt	182.77		
milliQ (sterile)	9816.86		
TOTAL	**24000**		

Figure 9.1: CF reaction recipe. Mixing is split into four stages. AA Mixture of unlabelled residues, 10x reaction mixture containing most of the system components, Master mix containing all components except for the cell lysate and plasmid and the final reaction mix

9.5 PROTEIN PURIFICATION OF HUNTINGTIN

The initial purification step was done by his-tag purification. This was done using either nickel(II)nitriloacetic acid (Ni-NTA) resin or HisTrap™ HP 5mL Columns. For Ni-NTA purification the protein was incubated with the resin at 4°C for 50-60 minutes before being packed in a PD-10 column, washed with 2 column volumes (CV) of buffer A and eluded with the 6-8 mL buffer B. HisTrap™ columns were connected to Fast protein liquid chromatography (FPLC) system. After being loaded into the column, the protein was eluded by increasing the imidazole on an increasing gradient, until the protein was eluded. In order to avoid protein loss,

the flowrate was reduced to 0.5 mL/min during loading, which made the purification time consuming.

The second purification step was size-exclusion chromatography (SEC) using a HiLoad™ 16/600 superdex™ 75 column. The protein sample was concentrated to ~2 mL using 6mL 5 K MWCO Pierce™ protein concentrators prior to injection on the SEC column. The gel filtration purification separated proteins depending both on size and shape with larger molecules passing through the column, while smaller molecules were retained due to the pore size of the gel resin. While protein samples visibly coloured the elute, the fractions collected were tested by SDS gel electrophoresis to select the purest fractions. Fractions were collected and stored at -20°C.

9.6 MASS SPECTROMETRY

The Mass Spectrometry samples were examined at two different platforms with different techniques. The two techniques were:

1) Liquid Chromatography Mass Spectrometry (LC/MS)
2) Matrix-assisted laser desorption/ionization – time of flight Mass Spectrometry (MALDI TOF-MS)

The initial mass verification was done using LC/MS, where the sample was passed through a size-exclusion column and the elute was passed into the electrospray interface of the mass spectrometry (283).

The verification of SANS samples was done using MALDI TOF-MS with the matrix consisting of 2,5-dihydroxybenzoic acid (DHB) and α-cyano-4-hydroxycinnamic acid (CHCA), which have been commonly used for analysis of peptides and proteins (284). The laser intensity was set to 25-60% with an acquisition range of 5-45 or 10-70 kDa. In MALDI MS, the dilute target solute was deposited together with the highly concentrated matrix components. Once dry, the sample could be desorpted/ionized using a laser, which facilitated the transition from solid-phase to gas-phase and subsequent MS analysis.

9.7 SEC-SAXS MEASUREMENTS

SAXS data was measured at the Swing beamline at synchrotron Soleil (France). Protein samples in buffer N were concentrated until 4 – 6 mg/mL (100 – 150 μM). 45μL was injected per elution. Size-exclusion chromatography was done using a Superdex 75 5/150 GL column and the elute of the column is passed through a 1.5 mm quartz capillary in the sample position. The size-exclusion step will theoretically allow separation of larger aggregates or smaller degradation products peaks from the monomeric protein scattering signal. Scattering frames were collected from injection for the duration of the elution continuously, without closing the shutter. Buffer subtraction was done using the software Chromixs (285), where the buffer baseline and sample peak can be selected and an average experimental profile can be calculated. R_g's were calculated for each frame (0.4s) across the sample elution peak, in order to select the frames corresponding to the monomeric protein.

9.8 SEC-SANS MEASUREMENTS

Depending on the yield of the cell-free expression, samples were concentrated to 1-6 mg/mL (~25-150 μM). The column was varied, with Superdex 75i 10/300 used for the first two experiments and Superdex 75i 5/150 GL used for succeeding experiments. The injection volume was 200 μL and 45 μL respectively. Data acquisition, started at sample injection and elution, was followed by UV absorbance. Once the protein peak was observed, the flow of the HPLC pump was stopped. The sample was exposed for one to three hours in order to obtain sufficient statistics for the dataset. The pump was subsequently restarted, and the buffer background was measured for an equal amount of time after the UV trace returned to the baseline.

A notable difference between SEC-SANS and SEC-SAXS is that the elute from the column is led into a circular 1mm quartz cuvette mounted in the sample position. This cuvette has a volume of 150 μL and is filled from the bottom, with the exit tube connecting the top of the cell to the fraction collector. The cell has a diameter of 13 mm while the beam is defined by a 11mm diameter aperture (Chapter 1.4, Figure 1.10 & 1.11). During the counting, while the flow is stopped, the pressure slightly drops within the sample cell and a slight drop in the UV-trace is also observed. This jumps up again, once the pump is restarted. Compared to SAXS measurements, protein samples measured during a SANS experiment do not undergo radiation-induced damage and can be recovered and re-concentrated for other experiments. The sample

environment was encased in a thermalized box which can hold the temperature around 11℃ with a constant flow of dry air.

The data was recorded on beamline D22 at the ILL. D22 has a dual detector setup with a static, slanted (20-degree tilt) detector at 1.4 m and a moveable detector, which for our experiments was kept at 5.6 m. This sped up experiment time significantly compared to previous detector setup, which required moving the detector and count at both Q-ranges individually. The wavelength of the beam was 6 Å ±10% and the source dimension was 40 mm x 55 mm. The intensity calibration was done using the direct beam intensity. A semi-transparent beam-stop is located in the position of the neutron beam to shield the detector. The semi-transparent beam-stop allows the sample and empty cell transmissions to be measured at the same time as the scattering and calculated by comparison to the direct beam measurement (122). Experimental data was reduced using Grasp (286). The scattering intensity of the empty cell is subtracted from the average scattering intensity of the buffer. Subsequently, the buffer intensity is subtracted from the average scattering intensity obtained from the sample, yielding the scattering pattern.

9.9 PREPARATION OF THEORETICAL ENSEMBLES

Ensembles of Huntingtin structures were modelled using an algorithm previously published (165). Several ensembles were produced by modifying the degree of structural influence propagated from either the initial N17 or from the proline-rich region resulting in 34 different ensembles. The algorithm uses a database of three-residue peptides (SCOPe) (266) to model random conformation. The database is based off of experimental protein structures. Both unrefined and NMR refined ensembles were prepared. The NMR refined ensemble has previously been reported in an earlier publication (287) and used Cα and Cβ chemical shift values calculated using SPARTA+(288) to fit the experimental data obtained by the group. For the unrefined ensemble, one set of structures which had no structural influence from the flanking regions was chosen.

To prepare the ensembles for EOM fitting first the Htt fragments were connected with a linker peptide, containing a 3C cleavage site, sfGFP and a His-tag using Ranch from the Atsas Softeware package (113). The python package pdb-tools (268) reres was used to renumber residues to correct the pdb files. The software Pulchra (289) was used to correct missing sidechains from the sfGFP crystal structure followed by reduce (290) to add hydrogens. The pdb-tools function reatom was used to renumber every atom in the pdb's. To produce

deuterated labelled ensembles, a python script was produced, which masked the chosen residues and exchanged hydrogen with deuterium. Finally, to correct for labile deuteration depending on the level of D_2O in a given sample, a python script was produced that could randomly exchange hydrogen with deuterium atoms to a given ratio. This script used the function:

$def\ myfunc(model, index)$:

$change\ =\ random.random() < rate$

$if\ (change)$:

$cmd.alter('\%s\ and\ index\ \%d'\%(model, index), 'elem\ = \'D\'; \ name\ = \'D\'')$

In order to verify the deuteration of labile hydrogens, a step was implemented that controls keeping the deuteration of labile hydrogens within 10% of the targeted deuteration level. The ensembles were calculated for 0%, 20%, 40%, 60%, 80% and 100% D_2O for each of the labelling patterns resulting in 48 ensembles for both Htt-Q16 and Htt-Q36.

To prepare the theoretical intensity files Cryso[(172) was run for the fully protonated ensemble and Cryson (173) was run for each deuteration level of each labelling scheme. Crysol was run with a maximum order of harmonics of 30 (-lm 30) and 200 data points (-ns 200). Similarly, Cryson was run with both options with the addition of the option to use explicit hydrogens from the pdb-files and the D2O level matching the given ensemble (-eh -D2O 0-1). The explicit hydrogen mode allows us to keep our labelling pattern during the calculation of the theoretical intensity files.

9.10 ENSEMBLE FITTING

The fitting of experimental data to ensembles of theoretical profiles was done using the software EOM (183) of the Atsas Package. The version used was the first unpublished iteration written by Pau Bernado, which allows for multiple datasets to be used as input for the fitting. This allows simultaneous fitting of both SAXS and SANS data as well as fitting multiple SANS datasets at once. The fitting uses a genetic algorithm to choose sub-ensembles, which is then scaled to the experimental data. A χ^2 value of the fit is then calculated and the next cycle is started. For the genetic algorithm 50 genes of 50 theoretical intensity profiles are randomly selected from the full ensemble. The algorithm subsequently creates 50 new genes by mutating up to 10 profiles for each of the selected genes with new profiles from the full ensemble pool.

Another 50 genes are created by crossing profiles between the initial 50 genes, totalling 150 genes for one generation, and a new round of mutation and crossing is repeated for 1000 generations. The sub-ensemble with the lowest χ^2 value when compared to the experimental data is saved and the next cycle is started from a new initial selection of profiles. Importantly the mutation and crossings are not allowing repetition of individual profiles within a single gene. Once the algorithm had run for 200 cycles, the best gene of each cycle were compared, and an average profile was generated of the sub-ensemble. Furthermore, the profile identifier of each theoretical intensity file selected, for each of the 200 cycles, were saved in order to analyse the resulting selections.

To fit multiple datasets at a time, the pool of structures is supplied for each of the experimental samples. The identifiers for each of the profiles match that of the original pdb-structure between the labelling patterns, which allows a structure to be fit to several experimental data sets at once. Each gene will be compared to all experimental datasets and the sub-ensemble that fits them all collectively the best will be reported for the cycle. χ^2-values are reported both as cumulative values and as individual fits of each experimental dataset.

10 REFERENCES

1. Huntington G. On Chorea. J Neuropsychiatry Clin Neurosci. 2003 Feb 1;15(1):109–12.

2. Zuccato C, Valenza M, Cattaneo E. Molecular Mechanisms and Potential Therapeutical Targets in Huntington's Disease. Physiol Rev. 2010 July 1;90(3):905–81.

3. Saudou F, Humbert S. The Biology of Huntingtin. Neuron. 2016 Mar 2;89(5):910–26.

4. Walker FO. Huntington's disease. The Lancet. 2007 Jan 20;369(9557):218–28.

5. Ross CA, Tabrizi SJ. Huntington's disease: from molecular pathogenesis to clinical treatment. Lancet Neurol. 2011 Jan 1;10(1):83–98.

6. Thomas B Stoker, Sarah L Mason, Julia C Greenland, Simon T Holden, Helen Santini, Roger A Barker. Huntington's disease: diagnosis and management. Pract Neurol. 2022 Feb 1;22(1):32.

7. Kremer B, Goldberg P, Andrew SE, Theilmann J, Telenius H, Zeisler J, et al. A Worldwide Study of the Huntington's Disease Mutation: The Sensitivity and Specificity of Measuring CAG Repeats. N Engl J Med. 1994 May 19;330(20):1401–6.

8. Langbehn D, Brinkman R, Falush D, Paulsen J, Hayden M, on behalf of an International Huntington's Disease Collaborative Group. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. Clin Genet. 2004 Apr 1;65(4):267–77.

9. Bruland O, Almqvist EW, Goldberg YP, Boman H, Hayden MR, Knappskog PM. Accurate determination of the number of CAG repeats in the Huntington disease gene using a sequence-specific internal DNA standard. Clin Genet. 1999 Mar 1;55(3):198–202.

10. Brinkman RR, Mezei MM, Theilmann J, Almqvist E, Hayden MR. The likelihood of being affected with Huntington disease by a particular age, for a specific CAG size. Am J Hum Genet. 1997 May;60(5):1202–10.

11. Rubinsztein DC, Leggo J, Coles R, Almqvist E, Biancalana V, Cassiman JJ, et al. Phenotypic characterization of individuals with 30-40 CAG repeats in the Huntington disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36-39 repeats. Am J Hum Genet. 1996 July;59(1):16–22.

12. Wexler NS, Lorimer J, Porter J, Gomez F, Moskowitz C, Shackell E, et al. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. Proc Natl Acad Sci U S A. 2004 Mar 9;101(10):3498–503.

13. Quarrell O, O'Donovan KL, Bandmann O, Strong M. The Prevalence of Juvenile Huntington's Disease: A Review of the Literature and Meta-Analysis. PLoS Curr. 2012 July 20;4:e4f8606b742ef3.

14. van der Burg JM, Björkqvist M, Brundin P. Beyond the brain: widespread pathology in Huntington's disease. Lancet Neurol. 2009 Aug 1;8(8):765–74.

15. Jiang A, Handley RR, Lehnert K, Snell RG. From Pathogenesis to Therapeutics: A Review of 150 Years of Huntington's Disease Research. Int J Mol Sci. 2023;24(16).

16. Morrison PJ. Accurate prevalence and uptake of testing for Huntington's disease. Lancet Neurol. 2010 Dec 1;9(12):1147.

17. Wexler A. Stigma, history, and Huntington's disease. The Lancet. 2010 July 3;376(9734):18–9.

18. MacDonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, Srinidhi L, et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell. 1993 Mar 26;72(6):971–83.

19. Gusella JF, MacDonald ME. Molecular genetics: unmasking polyglutamine triggers in neurodegenerative disease. Nat Rev Neurosci. 2000 Nov;1(2):109–15.

20. Nakamura K, Jeong SY, Uchihara T, Anno M, Nagashima K, Nagashima T, et al. SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. Hum Mol Genet. 2001 July 1;10(14):1441–8.

21. Hoffner G, Island ML, Djian P. Purification of neuronal inclusions of patients with Huntington's disease reveals a broad range of N-terminal fragments of expanded huntingtin and insoluble polymers. J Neurochem. 2005 Oct;95(1):125–36.

22. Cooper JK, Schilling G, Peters MF, Herring WJ, Sharp AH, Kaminsky Z, et al. Truncated N-terminal fragments of huntingtin with expanded glutamine repeats form nuclear and cytoplasmic aggregates in cell culture. Hum Mol Genet. 1998 May;7(5):783–90.

23. DiFiglia M, Sapp E, Chase KO, Davies SW, Bates GP, Vonsattel JP, et al. Aggregation of Huntingtin in Neuronal Intranuclear Inclusions and Dystrophic Neurites in Brain. Science. 1997 Sept 26;277(5334):1990–3.

24. Ehrlich ME. Huntington's Disease and the Striatal Medium Spiny Neuron: Cell-Autonomous and Non-Cell-Autonomous Mechanisms of Disease. Neurotherapeutics. 2012 Apr 1;9(2):270–84.

25. Arrasate M, Mitra S, Schweitzer ES, Segal MR, Finkbeiner S. Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. Nature. 2004 Oct 14;431(7010):805–10.

26. Saudou F, Finkbeiner S, Devys D, Greenberg ME. Huntingtin acts in the nucleus to induce apoptosis but death does not correlate with the formation of intranuclear inclusions. Cell. 1998 Oct 2;95(1):55–66.

27. Miguez A, Gomis C, Vila C, Monguió-Tortajada M, Fernández-García S, Bombau G, et al. Soluble mutant huntingtin drives early human pathogenesis in Huntington's disease. Cell Mol Life Sci CMLS. 2023 Aug 3;80(8):238.

28. Tabrizi SJ, Flower MD, Ross CA, Wild EJ. Huntington disease: new insights into molecular pathogenesis and therapeutic opportunities. Nat Rev Neurol. 2020 Oct 1;16(10):529–46.

29. Yang W, Dunlap JR, Andrews RB, Wetzel R. Aggregated polyglutamine peptides delivered to nuclei are toxic to mammalian cells. Hum Mol Genet. 2002 Nov 1;11(23):2905–17.

30. Monsellier E, Bousset L, Melki R. α-Synuclein and huntingtin exon 1 amyloid fibrils bind laterally to the cellular membrane. Sci Rep. 2016 Jan 13;6(1):19180.

31. Costanzo M, Abounit S, Marzo L, Danckaert A, Chamoun Z, Roux P, et al. Transfer of polyglutamine aggregates in neuronal cells occurs in tunneling nanotubes. J Cell Sci. 2013 Aug 15;126(Pt 16):3678–85.

32. Herrera F, Tenreiro S, Miller-Fleming L, Outeiro TF. Visualization of cell-to-cell transmission of mutant huntingtin oligomers. PLoS Curr. 2011 Feb 11;3:RRN1210.

33. Lin JT, Chang WC, Chen HM, Lai HL, Chen CY, Tao MH, et al. Regulation of feedback between protein kinase A and the proteasome system worsens Huntington's disease. Mol Cell Biol. 2013 Mar;33(5):1073–84.

34. Cortes CJ, La Spada AR. The many faces of autophagy dysfunction in Huntington's disease: from mechanism to therapy. Drug Discov Today. 2014 July;19(7):963–71.

35. Sánchez I, Mahlke C, Yuan J. Pivotal role of oligomerization in expanded polyglutamine neurodegenerative disorders. Nature. 2003 Jan 1;421(6921):373–9.

36. Albin RL, Young AB, Penney JB. The functional anatomy of basal ganglia disorders. Trends Neurosci. 1989 Oct;12(10):366–75.

37. Plotkin JL, Surmeier DJ. Corticostriatal synaptic adaptations in Huntington's disease. Curr Opin Neurobiol. 2015 Aug;33:53–62.

38. Faideau M, Kim J, Cormier K, Gilmore R, Welch M, Auregan G, et al. In vivo expression of polyglutamine-expanded huntingtin by mouse striatal astrocytes impairs glutamate transport: a correlation with Huntington's disease subjects. Hum Mol Genet. 2010 Aug 1;19(15):3053–67.

39. Dong X xia, Wang Y, Qin Z hong. Molecular mechanisms of excitotoxicity and their relevance to pathogenesis of neurodegenerative diseases. Acta Pharmacol Sin. 2009 Apr;30(4):379–87.

40. Cepeda C, Murphy KPS, Parent M, Levine MS. The role of dopamine in Huntington's disease. Prog Brain Res. 2014;211:235–54.

41. Beal MF, Brouillet E, Jenkins BG, Ferrante RJ, Kowall NW, Miller JM, et al. Neurochemical and histologic characterization of striatal excitotoxic lesions produced by the mitochondrial toxin 3-nitropropionic acid. J Neurosci Off J Soc Neurosci. 1993 Oct;13(10):4181–92.

42. Brouillet E, Hantraye P, Ferrante RJ, Dolan R, Leroy-Willig A, Kowall NW, et al. Chronic mitochondrial energy impairment produces selective striatal degeneration and abnormal choreiform movements in primates. Proc Natl Acad Sci U S A. 1995 July 18;92(15):7105–9.

43. Crotti A, Glass CK. The choreography of neuroinflammation in Huntington's disease. Trends Immunol. 2015 June;36(6):364–73.

44. Estrada-Sánchez AM, Rebec GV. Role of cerebral cortex in the neuropathology of Huntington's disease. Front Neural Circuits. 2013;7:19.

45. Tabrizi SJ, Langbehn DR, Leavitt BR, Roos RA, Durr A, Craufurd D, et al. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. Lancet Neurol. 2009 Sept;8(9):791–801.

46. Paulsen JS, Langbehn DR, Stout JC, Aylward E, Ross CA, Nance M, et al. Detection of Huntington's disease decades before diagnosis: the Predict-HD study. J Neurol Neurosurg Psychiatry. 2008 Aug;79(8):874–80.

47. Constantinescu R, Romer M, Oakes D, Rosengren L, Kieburtz K. Levels of the light subunit of neurofilament triplet protein in cerebrospinal fluid in Huntington's disease. Parkinsonism Relat Disord. 2009 Mar;15(3):245–8.

48. Vinther-Jensen T, Börnsen L, Budtz-Jørgensen E, Ammitzbøll C, Larsen IU, Hjermind LE, et al. Selected CSF biomarkers indicate no evidence of early neuroinflammation in Huntington disease. Neurol Neuroimmunol Neuroinflammation. 2016 Dec;3(6):e287.

49. Southwell AL, Smith SEP, Davis TR, Caron NS, Villanueva EB, Xie Y, et al. Ultrasensitive measurement of huntingtin protein in cerebrospinal fluid demonstrates increase with Huntington disease stage and decrease following brain huntingtin suppression. Sci Rep. 2015 July 15;5:12166.

50. Huntington Study Group. Tetrabenazine as antichorea therapy in Huntington disease: a randomized controlled trial. Neurology. 2006 Feb 14;66(3):366–72.

51. Huntington Study Group, Frank S, Testa CM, Stamler D, Kayson E, Davis C, et al. Effect of Deutetrabenazine on Chorea Among Patients With Huntington Disease: A Randomized Clinical Trial. JAMA. 2016 July 5;316(1):40–50.

52. Furr Stimming E, Claassen DO, Kayson E, Goldstein J, Mehanna R, Zhang H, et al. Safety and efficacy of valbenazine for the treatment of chorea associated with Huntington's disease (KINECT-HD): a phase 3, randomised, double-blind, placebo-controlled trial. Lancet Neurol. 2023 June 1;22(6):494–504.

53. Wang H, Chen X, Li Y, Tang TS, Bezprozvanny I. Tetrabenazine is neuroprotective in Huntington's disease mice. Mol Neurodegener. 2010 Apr 26;5(1):18.

54. Eisenstein M. CRISPR takes on Huntington's disease. Nature. 2018 May;557(7707):S42–3.

55. Morelli KH, Wu Q, Gosztyla ML, Liu H, Yao M, Zhang C, et al. An RNA-targeting CRISPR–Cas13d system alleviates disease-related phenotypes in Huntington's disease models. Nat Neurosci. 2023 Jan 1;26(1):27–38.

56. Guo Q, Bin Huang, Cheng J, Seefelder M, Engler T, Pfeifer G, et al. The cryo-electron microscopy structure of huntingtin. Nature. 2018 Mar 1;555(7694):117–20.

57. Andrade MA, Bork P. HEAT repeats in the Huntington's disease protein. Nat Genet. 1995 Oct 1;11(2):115–6.

58. Huang B, Guo Q, Niedermeier ML, Cheng J, Engler T, Maurer M, et al. Pathological polyQ expansion does not alter the conformation of the Huntingtin-HAP40 complex. Structure. 2021 Aug 5;29(8):804-809.e5.

59. Harding RJ, Deme JC, Hevler JF, Tamara S, Lemak A, Cantle JP, et al. Huntingtin structure is orchestrated by HAP40 and shows a polyglutamine expansion-specific interaction with exon 1. Commun Biol. 2021 Dec 8;4(1):1374.

60. Jung T, Shin B, Tamo G, Kim H, Vijayvargia R, Leitner A, et al. The Polyglutamine Expansion at the N-Terminal of Huntingtin Protein Modulates the Dynamic Configuration and Phosphorylation of the C-Terminal HEAT Domain. Structure. 2020 Sept 1;28(9):1035-1050.e8.

61. Alteen MG, Deme JC, Alvarez CP, Loppnau P, Hutchinson A, Seitova A, et al. Delineation of functional subdomains of Huntingtin protein and their interaction with HAP40. Structure. 2023 Sept 7;31(9):1121-1131.e6.

62. Azzaz F, Fantini J. The epigenetic dimension of protein structure. Biomol Concepts. 2022 Feb 21;13(1):55–60.

63. Barbosa Pereira PJ, Manso JA, Macedo-Ribeiro S. The structural plasticity of polyglutamine repeats. Curr Opin Struct Biol. 2023 June 1;80:102607.

64. Zhemkov VA, Kulminskaya AA, Bezprozvanny IB, Kim M. The 2.2-Angstrom resolution crystal structure of the carboxy-terminal region of ataxin-3. FEBS Open Bio. 2016 Mar 1;6(3):168–78.

65. Urbanek A, Popovic M, Morató A, Estaña A, Elena-Real CA, Mier P, et al. Flanking Regions Determine the Structure of the Poly-Glutamine in Huntingtin through Mechanisms Common among Glutamine-Rich Human Proteins. Structure. 2020 July 7;28(7):733-746.e5.

66. Elena-Real CA, Sagar A, Urbanek A, Popovic M, Morató A, Estaña A, et al. The structure of pathogenic huntingtin exon 1 defines the bases of its aggregation propensity. Nat Struct Mol Biol. 2023 Mar 1;30(3):309–20.

67. Urbanek A, Popovic M, Elena-Real CA, Morató A, Estaña A, Fournet A, et al. Evidence of the Reduced Abundance of Proline cis Conformation in Protein Poly Proline Tracts. J Am Chem Soc. 2020 Apr 29;142(17):7976–86.

68. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol. 2015 Jan;16(1):18–29.

69. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J. 2005 Oct;272(20):5129–48.

70. Galea CA, Wang Y, Sivakolundu SG, Kriwacki RW. Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. Biochemistry. 2008 July 22;47(29):7598–609.

71. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z. Intrinsic Disorder and Protein Function. Biochemistry. 2002 May 1;41(21):6573–82.

72. Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK. Intrinsic disorder in transcription factors. Biochemistry. 2006 June 6;45(22):6873–88.

73. Gsponer J, Babu MM. The rules of disorder or why disorder rules. Prog Biophys Mol Biol. 2009;99(2–3):94–103.

74. Wang Z, Jin L, Yuan Z, Węgrzyn G, Węgrzyn A. Classification of plasmid vectors using replication origin, selection marker and promoter as criteria. Plasmid. 2009 Jan 1;61(1):47–51.

75. Elena-Real CA, Mier P, Sibille N, Andrade-Navarro MA, Bernadó P. Structure–function relationships in protein homorepeats. Curr Opin Struct Biol. 2023 Dec 1;83:102726.

76. Mier P, Paladin L, Tamana S, Petrosian S, Hajdu-Soltész B, Urbanek A, et al. Disentangling the complexity of low complexity proteins. Brief Bioinform. 2020 Mar 23;21(2):458–72.

77. Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem. 1994 Sept;18(3):269–85.

78. Lobanov MYu, Galzitskaya OV. Occurrence of disordered patterns and homorepeats in eukaryotic and bacterial proteomes. Mol Biosyst. 2012;8(1):327–37.

79. Chavali S, Singh AK, Santhanam B, Babu MM. Amino acid homorepeats in proteins. Nat Rev Chem. 2020 Aug;4(8):420–34.

80. Chandra S, Shao J, Li JX, Li M, Longo FM, Diamond MI. A common motif targets huntingtin and the androgen receptor to the proteasome. J Biol Chem. 2008 Aug 29;283(35):23950–5.

81. Mier P, Elena-Real CA, Cortés J, Bernadó P, Andrade-Navarro MA. The sequence context in poly-alanine regions: structure, function and conservation. Bioinforma Oxf Engl. 2022 Oct 31;38(21):4851–8.

82. Ramazzotti M, Monsellier E, Kamoun C, Degl'Innocenti D, Melki R. Polyglutamine repeats are associated to specific sequence biases that are conserved among eukaryotes. PloS One. 2012;7(2):e30824.

83. Gonçalves-Kulik M, Mier P, Kastano K, Cortés J, Bernadó P, Schmid F, et al. Low Complexity Induces Structure in Protein Regions Predicted as Intrinsically Disordered. Biomolecules. 2022;12(8).

84. Estaña A, Barozet A, Mouhand A, Vaisset M, Zanon C, Fauret P, et al. Predicting Secondary Structure Propensities in IDPs Using Simple Statistics from Three-Residue Fragments. J Mol Biol. 2020 Sept 4;432(19):5447–59.

85. Darling AL, Uversky VN. Intrinsic Disorder in Proteins with Pathogenic Repeat Expansions. Mol Basel Switz. 2017 Nov 24;22(12):2027.

86. Babu MM, van der Lee R, de Groot NS, Gsponer J. Intrinsically disordered proteins: regulation and disease. Curr Opin Struct Biol. 2011 June;21(3):432–40.

87. Amiel J, Trochet D, Clément-Ziza M, Munnich A, Lyonnet S. Polyalanine expansions in human. Hum Mol Genet. 2004 Oct 1;13 Spec No 2:R235-243.

88. Schaefer MH, Wanker EE, Andrade-Navarro MA. Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks. Nucleic Acids Res. 2012 May;40(10):4273–87.

89. Stoyas CA, La Spada AR. The CAG-polyglutamine repeat diseases: a clinical, molecular, genetic, and pathophysiologic nosology. Handb Clin Neurol. 2018;147:143–70.

90. Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, Garcia de la Banda M, et al. Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res. 2005 Apr;15(4):537–51.

91. Urbanek A, Morató A, Allemand F, Delaforge E, Fournet A, Popovic M, et al. A General Strategy to Access Structural Information at Atomic Resolution in Polyglutamine Homorepeats. Angew Chem Int Ed Engl. 2018/03/07 ed. 2018 Mar 26;57(14):3598–601.

92. Urbanek A, Elena-Real CA, Popovic M, Morató A, Fournet A, Allemand F, et al. Site-Specific Isotopic Labeling (SSIL): Access to High-Resolution Structural and Dynamic Information in Low-Complexity Proteins. ChemBioChem. 2020 Mar 16;21(6):769–75.

93. Svergun DI, Koch MHJ. Small-angle scattering studies of biological macromolecules in solution. Rep Prog Phys. 2003 Sept;66(10):1735.

94. Jeffries CM, Ilavsky J, Martel A, Hinrichs S, Meyer A, Pedersen JS, et al. Small-angle X-ray and neutron scattering. Nat Rev Methods Primer. 2021 Oct 12;1(1):70.

95. Mühlbauer S, Honecker D, Périgo ÉA, Bergner F, Disch S, Heinemann A, et al. Magnetic small-angle neutron scattering. Rev Mod Phys. 2019 Mar;91(1):015004.

96. Fratzl P. Small-angle scattering in materials science - a short review of applications in alloys, ceramics and composite materials. J Appl Crystallogr. 2003 June 1;36(3–1):397–404.

97. Radlinski AP, Hinde AL. Small angle neutron scattering and petroleum geology. Neutron News. 2002 Jan 1;13(2):10–4.

98. Cabral JT, Higgins JS. Small Angle Neutron Scattering from the Highly Interacting Polymer Mixture TMPC/PSd: No Evidence of Spatially Dependent χ Parameter. Macromolecules. 2009 Dec 22;42(24):9528–36.

99. Svergun DI, Koch MHJ, Timmins PA, May RP. Small Angle X-Ray and Neutron Scattering from Solutions of Biological Macromolecules [Internet]. Oxford University Press; 2013 [cited 2024 Dec 13]. Available from: https://doi.org/10.1093/acprof:oso/9780199639533.001.0001

100. Koch MHJ, Vachette P, Svergun DI. Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. Q Rev Biophys. 2003/10/23 ed. 2003;36(2):147–227.

101. Guinier A. La diffraction des rayons X aux très petits angles : application à l'étude de phénomènes ultramicroscopiques. Ann Phys. 1939;11(12):161–237.

102. Guinier A, Fournet G. Small-Angle Scattering of X-Rays. John Wiley & Sons, Inc.; 1955.

103. Ibel K, Stuhrmann HB. Comparison of neutron and X-ray scattering of dilute myoglobin solutions. J Mol Biol. 1975;93(2):255–65.

104. Engelman DM, Moore PB. A New Method for the Determination of Biological Quarternary Structure by Neutron Scattering. Proc Natl Acad Sci. 1972;69(8):1997–9.

105. Stuhrmann HB. Neutron small-angle scattering of biological macromolecules in solution. J Appl Crystallogr. 1974 Apr 1;7(2):173–8.

106. Skou S, Rodrigues S, Hoghoj P, Bossan F. High-throughput biological solution SAXS instrumentation for the home laboratory. Vol. 73, Acta Crystallographica Section A. 2017. p. C689.

107. Seeger PA, Hjelm Jnr RP. Small-angle neutron scattering at pulsed spallation sources. J Appl Crystallogr. 1991 Oct 1;24(5):467–78.

108. Doligez X, Bouneau S, David S, Ernoult M, Zakari-Issoufou AA, Thiollière N, et al. Fundamentals of reactor physics with a view to the (possible) futures of nuclear energy. Demain L'énergie. 2017 Sept 1;18(7):372–80.

109. Joyce M. Chapter 6 - Moderation. In: Joyce M, editor. Nuclear Engineering [Internet]. Butterworth-Heinemann; 2018. p. 111–27. Available from: https://www.sciencedirect.com/science/article/pii/B9780081009628000068

110. Fomin N, Fry J, Pattie RW, Greene GL. Fundamental Neutron Physics at Spallation Sources. Annu Rev Nucl Part Sci. 2022 Sept 26;72(1):151–76.

111. Graewert MA, Svergun DI. Impact and progress in small and wide angle X-ray scattering (SAXS and WAXS). Curr Opin Struct Biol. 2013;23(5):748–54.

112. Trewhella J. Recent advances in small-angle scattering and its expanding impact in structural biology. Structure. 2022 Jan 6;30(1):15–23.

113. Manalastas-Cantos K, Konarev PV, Hajizadeh NR, Kikhney AG, Petoukhov MV, Molodenskiy DS, et al. ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. Vol. 54, Journal of Applied Crystallography. 2021. p. 343–55.

114. Bizien T, Durand D, Roblina P, Thureau A, Vachette P, Pérez J. A Brief Survey of State-of-the-Art BioSAXS. Protein Pept Lett. 2016 Mar 1;23(3):217–31.

115. Hansen S. BayesApp: a web site for indirect transformation of small-angle scattering data. J Appl Crystallogr. 2012 June;45(3):566–7.

116. Receveur-Brechot V. AlphaFold, small-angle X-ray scattering and ensemble modelling: a winning combination for intrinsically disordered proteins. Vol. 56, Journal of Applied Crystallography. 2023. p. 1313–4.

117. Spinozzi F, Ferrero C, Ortore MG, De Maria Antolinos A, Mariani P. GENFIT: software for the analysis of small-angle X-ray and neutron scattering data of macromolecules in solution. Vol. 47, Journal of Applied Crystallography. 2014. p. 1132–9.

118. Grudinin S, Garkavenko M, Kazennov A. Pepsi-SAXS : an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. Acta Crystallogr D Biol Crystallogr. 2017 May 1;D73:449.

119. Mathew E, Mirza A, Menhart N. Liquid-chromatography-coupled SAXS for accurate sizing of aggregating proteins. Vol. 11, Journal of Synchrotron Radiation. 2004. p. 314–8.

120. Wright GSA, Lee HC, Schulze-Briese C, Grossmann JG, Strange RW, Hasnain SS. The application of hybrid pixel detectors for in-house SAXS instrumentation with a view to combined chromatographic operation. Vol. 20, Journal of Synchrotron Radiation. 2013. p. 383–5.

121. Ryan TM, Trewhella J, Murphy JM, Keown JR, Casey L, Pearce FG, et al. An optimized SEC-SAXS system enabling high X-ray dose for rapid SAXS assessment with correlated UV measurements for biomolecular structure analysis. Vol. 51, Journal of Applied Crystallography. 2018. p. 97–111.

122. Martel A, Cocho C, Caporaletti F, Jacques M, El Aazzouzi A, Lapeyre F, et al. Upgraded D22 SEC-SANS setup dedicated to the biology community. Vol. 56, Journal of Applied Crystallography. 2023. p. 994–1001.

123. Jordan A, Jacques M, Merrick C, Devos J, Forsyth VT, Porcar L, et al. SEC-SANS: size exclusion chromatography combined in situ with small-angle neutron scatteringThis article will form part of a virtual special issue of the journal, presenting some highlights of the 16th International Conference on Small-Angle Scattering (SAS2015). Vol. 49, Journal of Applied Crystallography. 2016. p. 2015–20.

124. Bernadó P, Shimizu N, Zaccai G, Kamikubo H, Sugiyama M. Solution scattering approaches to dynamical ordering in biomolecular systems. Biophys Explor Dyn Ordering Biomol Syst. 2018 Feb 1;1862(2):253–74.

125. Buffet, Jean-Claude, Cristiglio, Viviana, Cuccaro, Sylvain, Demé, Bruno, Guérard, Bruno, Marchal, Julien, et al. Development of a large-area curved Trench-MWPC 3He detector for the D16 neutron diffractometer at the ILL. EPJ Web Conf. 2023;286:03010.

126. Sears VF. Neutron scattering lengths and cross sections. Neutron News. 1992;3(3):29–37.

127. Jeffries CM, Pietras Z, Svergun DI. The basics of small-angle neutron scattering (SANS for new users of structural biology). EPJ Web Conf [Internet]. 2020;236. Available from: https://doi.org/10.1051/epjconf/202023603001

128. Langer JA, Engelman DM, Moore PB. Neutron-scattering studies of the ribosome of Escherichia coli: A provisional map of the locations of proteins S3, S4, S5, S7, S8 and S9 in the 30 S subunit. J Mol Biol. 1978 Mar 15;119(4):463–85.

129. Jacrot B. The study of biological structures by neutron scattering from solution. Rep Prog Phys. 1976 Oct 1;39(10):911–53.

130. Schneider R, Mayer A, Schmatz W, Kaiser B, Scherm R. Neutron small-angle scattering from aqueous solutions of oxy- and deoxyhaemoglobin. J Mol Biol. 1969 Apr 28;41(2):231–5.

131. Mahieu E, Ibrahim Z, Moulin M, Härtlein M, Franzetti B, Martel A, et al. The power of SANS, combined with deuteration and contrast variation, for structural studies of functional and dynamic biomacromolecular systems in solution. EPJ Web Conf [Internet]. 2020;236. Available from: https://doi.org/10.1051/epjconf/202023603002

132. Heller WT. Small-angle neutron scattering and contrast variation: a powerful combination for studying biological structures. Acta Crystallogr Sect D. 2010;66(11):1213–7.

133. Svergun DI. Restoring Low Resolution Structure of Biological Macromolecules from Solution Scattering Using Simulated Annealing. Biophys J. 1999 June 1;76(6):2879–86.

134. Watanabe Y, Inoko Y. Size-exclusion chromatography combined with small-angle X-ray scattering optics. 33rd Int Symp High Perform Liq Phase Sep Relat Tech. 2009 Oct 30;1216(44):7461–5.

135. David G, Pérez J. Combined sampler robot and high-performance liquid chromatography: a fully automated system for biological small-angle X-ray scattering experiments at the Synchrotron SOLEIL SWING beamline. J Appl Crystallogr. 2009 Oct 1;42(5):892–900.

136. Blanchet CE, Spilotros A, Schwemmer F, Graewert MA, Kikhney A, Jeffries CM, et al. Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY). J Appl Crystallogr. 2015 Apr;48(2):431–43.

137. Johansen NT, Pedersen MC, Porcar L, Martel A, Arleth L. Introducing SEC–SANS for studies of complex self-organized biological systems. Acta Crystallogr Sect D. 2018 Dec;74(12):1178–91.

138. Hopkins JB, Thorne RE. Quantifying radiation damage in biomolecular small-angle X-ray scattering. J Appl Crystallogr. 2016 June;49(3):880–90.

139. Bernadó P, Svergun DI. Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. Mol Biosyst. 2012 Jan;8(1):151–67.

140. Cordeiro TN, Herranz-Trillo F, Urbanek A, Estaña A, Cortés J, Sibille N, et al. Small-angle scattering studies of intrinsically disordered proteins and their complexes. Curr Opin Struct Biol. 2017 Feb;42:15–23.

141. Mebert AM, Villanueva ME, Tovar GI, Perez Bravo JJ, Copello GJ. Chapter 10 - Small-angle scattering techniques for biomolecular structure and dynamics. In: Saudagar P, Tripathi T, editors. Advanced Spectroscopic Methods to Study Biomolecular Structure and Dynamics

[Internet]. Academic Press; 2023. p. 271–307. Available from: https://www.sciencedirect.com/science/article/pii/B9780323991278000155

142. Calmettes P, Durand D, Desmadril M, Minard P, Receveur V, Smith JC. How random is a highly denatured protein? Biophys Chem. 1994 Dec 1;53(1):105–13.

143. Pérez J, Vachette P, Russo D, Desmadril M, Durand D. Heat-induced unfolding of neocarzinostatin, a small all-β protein investigated by small-angle X-ray scattering11Edited by M. F. Moody. J Mol Biol. 2001;308(4):721–43.

144. Heller WT. Influence of multiple well defined conformations on small-angle scattering of proteins in solution. Acta Crystallogr D Biol Crystallogr. 2005 Jan;61(Pt 1):33–44.

145. Bernadó P. Effect of interdomain dynamics on the structure determination of modular proteins by small-angle scattering. Eur Biophys J EBJ. 2010 Apr;39(5):769–80.

146. Ellis PJ, Cohen AE, Soltis SM. Beamstop with integrated X-ray sensor. J Synchrotron Radiat. 2003 May;10(3):287–8.

147. Lyngsø J, Pedersen JS. A high-flux automated laboratory small-angle X-ray scattering instrument optimized for solution scattering. J Appl Crystallogr. 2021 Feb;54(1):295–305.

148. Wu H, Li Z. A new dual-thickness semi-transparent beamstop for small-angle X-ray scattering. J Synchrotron Radiat. 2024 Sept 1;31(Pt 5):1197–208.

149. Receveur-Brechot V, Durand D. How random are intrinsically disordered proteins? A small angle scattering perspective. Curr Protein Pept Sci. 2012 Feb;13(1):55–75.

150. Doniach S. Changes in biomolecular conformation seen by small angle X-ray scattering. Chem Rev. 2001 June;101(6):1763–78.

151. Durand D, Vivès C, Cannella D, Pérez J, Pebay-Peyroula E, Vachette P, et al. NADPH oxidase activator p67phox behaves in solution as a multidomain protein with semi-flexible linkers. J Struct Biol. 2010 Jan 1;169(1):45–53.

152. Stuhrmann HB. Ein neues Verfahren zur Bestimmung der Oberflächenform und der inneren Struktur von gelösten globulären Proteinen aus Röntgenkleinwinkelmessungen. 1970;72(4_6):177–84.

153. Svergun DI, Volkov VV, Kozin MB, Stuhrmann HB. New Developments in Direct Shape Determination from Small-Angle Scattering. 2. Uniqueness. Acta Crystallogr Sect A. 1996 May;52(3):419–26.

154. Chacón P, Morán F, Díaz JF, Pantos E, Andreu JM. Low-Resolution Structures of Proteins in Solution Retrieved from X-Ray Scattering with a Genetic Algorithm. Biophys J. 1998 June 1;74(6):2760–75.

155. Svergun DI. Restoring Low Resolution Structure of Biological Macromolecules from Solution Scattering Using Simulated Annealing. Biophys J. 1999 June 1;76(6):2879–86.

156. Franke D, Svergun DI. DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. Vol. 42, Journal of Applied Crystallography. 2009. p. 342–6.

157. Tuukkanen AT, Kleywegt GJ, Svergun DI. Resolution of ıt ab initio shapes determined from small-angle scattering. IUCrJ. 2016 Nov;3(6):440–7.

158. Petoukhov MV, Svergun DI. Global Rigid Body Modeling of Macromolecular Complexes against Small-Angle Scattering Data. Biophys J. 2005 Aug 1;89(2):1237–50.

159. Schneidman-Duhovny D, Hammel M, Sali A. Macromolecular docking restrained by a small angle X-ray scattering profile. Comb Comput Model Sparse Low-Resolut Data. 2011 Mar 1;173(3):461–71.

160. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. Nucleic Acids Res. 2016 July 8;44(W1):W424–9.

161. Grishaev A, Wu J, Trewhella J, Bax A. Refinement of Multidomain Protein Structures by Combination of Solution Small-Angle X-ray Scattering and NMR Data. J Am Chem Soc. 2005 Nov 30;127(47):16621–8.

162. Bernadó P. Effect of interdomain dynamics on the structure determination of modular proteins by small-angle scattering. Eur Biophys J. 2010 Apr 1;39(5):769–80.

163. Clerc I, Sagar A, Barducci A, Sibille N, Bernadó P, Cortés J. The diversity of molecular interactions involving intrinsically disordered proteins: A molecular modeling perspective. Comput Struct Biotechnol J. 2021;19:3817–28.

164. Borges-Araújo L, Pereira GP, Valério M, Souza PCT. Assessing the Martini 3 protein model: A review of its path and potential. Biochim Biophys Acta Proteins Proteomics. 2024 July 1;1872(4):141014.

165. Estaña A, Sibille N, Delaforge E, Vaisset M, Cortés J, Bernadó P. Realistic Ensemble Models of Intrinsically Disordered Proteins Using a Structure-Encoding Coil Database. Structure. 2019 Feb 5;27(2):381-391.e2.

166. Milles S, Salvi N, Blackledge M, Jensen MR. Characterization of intrinsically disordered proteins and their dynamic complexes: From in vitro to cell-like environments. Prog Nucl Magn Reson Spectrosc. 2018 Dec;109:79–100.

167. Holmstrom ED, Holla A, Zheng W, Nettels D, Best RB, Schuler B. Accurate Transfer Efficiencies, Distance Distributions, and Ensembles of Unfolded and Intrinsically Disordered Proteins From Single-Molecule FRET. Methods Enzymol. 2018;611:287–325.

168. Yang YI, Shao Q, Zhang J, Yang L, Gao YQ. Enhanced sampling in molecular dynamics. J Chem Phys. 2019 Aug 21;151(7):070902.

169. Bernadó P, Blanchard L, Timmins P, Marion D, Ruigrok RWH, Blackledge M. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. Proc Natl Acad Sci U S A. 2005 Nov 22;102(47):17002–7.

170. Feldman HJ, Hogue CWV. Probabilistic sampling of protein conformations: new hope for brute force? Proteins. 2002 Jan 1;46(1):8–23.

171. Liu ZH, Teixeira JMC, Zhang O, Tsangaris TE, Li J, Gradinaru CC, et al. Local Disordered Region Sampling (LDRS) for ensemble modeling of proteins with experimentally undetermined or low confidence prediction segments. Bioinforma Oxf Engl. 2023 Dec 1;39(12):btad739.

172. Svergun D, Barberato C, Koch MHJ. CRYSOL – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. J Appl Crystallogr. 1995 Dec;28(6):768–73.

173. Svergun DI, Richard S, Koch MHJ, Sayers Z, Kuprin S, Zaccai G. Protein hydration in solution: Experimental observation by x-ray and neutron scattering. Proc Natl Acad Sci. 1998 Mar 3;95(5):2267.

174. Franke D, Petoukhov MV, Konarev PV, Panjkovich A, Tuukkanen A, Mertens HDT, et al. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. J Appl Crystallogr. 2017 Aug;50(4):1212–25.

175. Chen PC, Hub JS. Validating solution ensembles from molecular dynamics simulation by wide-angle X-ray scattering data. Biophys J. 2014 July 15;107(2):435–47.

176. Linse JB, Hub JS. Scrutinizing the protein hydration shell from molecular dynamics simulations against consensus small-angle scattering data. Commun Chem. 2023 Dec 12;6(1):272.

177. Cordeiro TN, Sibille N, Germain P, Barthe P, Boulahtouf A, Allemand F, et al. Interplay of Protein Disorder in Retinoic Acid Receptor Heterodimer and Its Corepressor Regulates Gene Expression. Struct Lond Engl 1993. 2019 Aug 6;27(8):1270-1285.e6.

178. Cordeiro TN, Chen PC, De Biasio A, Sibille N, Blanco FJ, Hub JS, et al. Disentangling polydispersity in the PCNA-p15PAF complex, a disordered, transient and multivalent macromolecular assembly. Nucleic Acids Res. 2017 Feb 17;45(3):1501–15.

179. Fatafta H, Samantray S, Sayyed-Ahmad A, Coskuner-Weber O, Strodel B. Molecular simulations of IDPs: From ensemble generation to IDP interactions leading to disorder-to-order transitions. Prog Mol Biol Transl Sci. 2021;183:135–85.

180. Chen J, Liu X, Chen J. Targeting Intrinsically Disordered Proteins through Dynamic Interactions. Biomolecules. 2020 May 11;10(5):743.

181. Ahmed MC, Skaanning LK, Jussupow A, Newcombe EA, Kragelund BB, Camilloni C, et al. Refinement of α-Synuclein Ensembles Against SAXS Data: Comparison of Force Fields and Methods. Front Mol Biosci. 2021;8:654333.

182. Sagar A, Svergun D, Bernadó P. Structural Analyses of Intrinsically Disordered Proteins by Small-Angle X-Ray Scattering. Methods Mol Biol Clifton NJ. 2020;2141:249–69.

183. Bernadó P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI. Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. J Am Chem Soc. 2007 May 1;129(17):5656–64.

184. Tria G, Mertens HDT, Kachala M, Svergun DI. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. IUCrJ. 2015 Mar 1;2(Pt 2):207–17.

185. Pelikan M, Hura GL, Hammel M. Structure and flexibility within proteins as identified through small angle X-ray scattering. Gen Physiol Biophys. 2009 June;28(2):174–89.

186. Hermann MR, Hub JS. SAXS-Restrained Ensemble Simulations of Intrinsically Disordered Proteins with Commitment to the Principle of Maximum Entropy. J Chem Theory Comput. 2019 Sept 10;15(9):5103–15.

187. Różycki B, Kim YC, Hummer G. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. Struct Lond Engl 1993. 2011 Jan 12;19(1):109–16.

188. Bottaro S, Bengtsen T, Lindorff-Larsen K. Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy Reweighting Approach. Methods Mol Biol Clifton NJ. 2020;2112:219–40.

189. Yabukarski F, Leyrat C, Martinez N, Communie G, Ivanov I, Ribeiro EA, et al. Ensemble Structure of the Highly Flexible Complex Formed between Vesicular Stomatitis Virus Unassembled Nucleoprotein and its Phosphoprotein Chaperone. J Mol Biol. 2016 July 3;428(13):2671–94.

190. Mylonas E, Hascher A, Bernadó P, Blackledge M, Mandelkow E, Svergun DI. Domain conformation of tau protein studied by solution small-angle X-ray scattering. Biochemistry. 2008 Sept 30;47(39):10345–53.

191. Oksanen E, Chen JCH, Fisher SZ. Neutron Crystallography for the Study of Hydrogen Bonds in Macromolecules. Mol Basel Switz. 2017 Apr 7;22(4):596.

192. Gajdos L, Blakeley MP, Haertlein M, Forsyth VT, Devos JM, Imberty A. Neutron crystallography reveals mechanisms used by Pseudomonas aeruginosa for host-cell binding. Nat Commun. 2022 Jan 11;13(1):194.

193. Capel MS, Engelman DM, Freeborn BR, Kjeldgaard M, Langer JA, Ramakrishnan V, et al. A Complete Mapping of the Proteins in the Small Ribosomal Subunit of Escherichia coli. Science. 1987 Dec 4;238(4832):1403–6.

194. Callow P, Sukhodub A, Taylor JEN, Kneale GG. Shape and subunit organisation of the DNA methyltransferase M.AhdI by small-angle neutron scattering. J Mol Biol. 2007 May 25;369(1):177–85.

195. King WA, Stone DB, Timmins PA, Narayanan T, von Brasch AAM, Mendelson RA, et al. Solution structure of the chicken skeletal muscle troponin complex via small-angle neutron and X-ray scattering. J Mol Biol. 2005 Jan 28;345(4):797–815.

196. Haertlein M, Moulin M, Devos JM, Laux V, Dunne O, Trevor Forsyth V. Chapter Five - Biomolecular Deuteration for Neutron Structural Biology and Dynamics. In: Kelman Z, editor. Methods in Enzymology [Internet]. Academic Press; 2016. p. 113–57. Available from: https://www.sciencedirect.com/science/article/pii/S0076687915006370

197. Jensen KJ, Shelton PT, Pedersen SL. Peptide synthesis and applications. Springer; 2013.

198. Da'san MM J, Al Musaimi O, Albericio F. Advances in solid-phase peptide synthesis in aqueous media (ASPPS). Green Chem. 2022;24(17):6360–72.

199. Kent SB. Chemical synthesis of peptides and proteins. Annu Rev Biochem. 1988;57(1):957–89.

200. Katz JJ, Crespi HL. Deuterated Organisms: Cultivation and Uses. Science. 1966 Mar 11;151(3715):1187–94.

201. Carlstedt BC, Crespi HL, Blake MI, Katz JJ. Biosynthesis of Deuterated Benzylpenicillins I: Solvent Deuterium Oxide Participation. J Pharm Sci. 1970 Oct 1;59(10):1456–60.

202. Dunne O, Weidenhaupt M, Callow P, Martel A, Moulin M, Perkins SJ, et al. Matchout deuterium labelling of proteins for small-angle neutron scattering studies using prokaryotic and eukaryotic expression systems and high cell-density cultures. Eur Biophys J EBJ. 2017 July;46(5):425–32.

203. Compton ELR, Page K, Findlay HE, Haertlein M, Moulin M, Zachariae U, et al. Conserved structure and domain organization among bacterial Slc26 transporters. Biochem J. 2014 Oct 15;463(2):297–307.

204. Appolaire A, Girard E, Colombo M, Durá MA, Moulin M, Härtlein M, et al. Small-angle neutron scattering reveals the assembly mode and oligomeric architecture of TET, a large, dodecameric aminopeptidase. Acta Crystallogr D Biol Crystallogr. 2014 Nov;70(Pt 11):2983–93.

205. David R, Richter MPO, Beck-Sickinger AG. Expressed protein ligation. Eur J Biochem. 2004 Feb 1;271(4):663–77.

206. Skrisovska L, Schubert M, Allain FHT. Recent advances in segmental isotope labeling of proteins: NMR applications to large proteins and glycoproteins. J Biomol NMR. 2010 Jan 1;46(1):51–65.

207. Freiburger L, Sonntag M, Hennig J, Li J, Zou P, Sattler M. Efficient segmental isotope labeling of multi-domain proteins using Sortase A. J Biomol NMR. 2015 Sept 1;63(1):1–8.

208. Williams FP, Milbradt AG, Embrey KJ, Bobby R. Segmental Isotope Labelling of an Individual Bromodomain of a Tandem Domain BRD4 Using Sortase A. PLOS ONE. 2016 Apr 29;11(4):e0154607.

209. Sonntag M, Jagtap PKA, Simon B, Appavou MS, Geerlof A, Stehle R, et al. Segmental, Domain-Selective Perdeuteration and Small-Angle Neutron Scattering for Structural Analysis of Multi-Domain Proteins. Angew Chem Int Ed. 2017 Aug 1;56(32):9322–5.

210. Laux V, Callow P, Svergun DI, Timmins PA, Forsyth VT, Haertlein M. Selective deuteration of tryptophan and methionine residues in maltose binding protein: a model system for neutron scattering. Eur Biophys J. 2008 July 1;37(6):815–22.

211. Pronk Jack T. Auxotrophic Yeast Strains in Fundamental and Applied Research. Appl Environ Microbiol. 2002 May 1;68(5):2095–100.

212. Sreenath HK, Bingman CA, Buchan BW, Seder KD, Burns BT, Geetha HV, et al. Protocols for production of selenomethionine-labeled proteins in 2-L polyethylene terephthalate bottles using auto-induction medium. Protein Expr Purif. 2005 Apr 1;40(2):256–67.

213. Tawani A, Qian S, Sparks SE, Sashi P, Cahill S, Rout M, et al. Characterizing interactions in the nuclear pore complex transporter using novel site-specific deuteration and SANS [Internet]. 2024 [cited 2024 July 18]. Available from: http://biorxiv.org/lookup/doi/10.1101/2024.06.20.599953

214. Silverman AD, Karim AS, Jewett MC. Cell-free gene expression: an expanded repertoire of applications. Nat Rev Genet. 2020 Mar 1;21(3):151–70.

215. Morató A, Elena-Real CA, Popovic M, Fournet A, Zhang K, Allemand F, et al. Robust Cell-Free Expression of Sub-Pathological and Pathological Huntingtin Exon-1 for NMR Studies. General Approaches for the Isotopic Labeling of Low-Complexity Proteins. Biomolecules. 2020;10(10).

216. Carlson ED, Gan R, Hodgman CE, Jewett MC. Cell-free protein synthesis: Applications come of age. Biotechnol Adv. 2012 Sept 1;30(5):1185–94.

217. Khambhati K, Bhattacharjee G, Gohil N, Braddick D, Kulkarni V, Singh V. Exploring the Potential of Cell-Free Protein Synthesis for Extending the Abilities of Biological Systems. Front

Bioeng Biotechnol [Internet]. 2019;7. Available from: https://www.frontiersin.org/articles/10.3389/fbioe.2019.00248

218. Buchner E, Rapp R. Alkoholische Gährung ohne Hefezellen. Berichte Dtsch Chem Ges. 1897 Sept 1;30(3):2668–78.

219. Nirenberg MW, Matthaei JH. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci. 1961 Oct 1;47(10):1588–602.

220. Fuller RS, Kaguni JM, Kornberg A. Enzymatic replication of the origin of the Escherichia coli chromosome. Proc Natl Acad Sci. 1981 Dec 1;78(12):7370–4.

221. Bernhard F, Tozawa Y. Cell-free expression—making a mark. New Contructs Expr Proteins Seq Topol. 2013 June 1;23(3):374–80.

222. Hodgman CE, Jewett MC. Cell-free synthetic biology: Thinking outside the cell. Synth Biol New Methodol Appl Metab Eng. 2012 May 1;14(3):261–9.

223. Yang HJ, Lee KH, Lim HJ, Kim DM. Tandem Cell-Free Protein Synthesis as a Tool for Rapid Screening of Optimal Molecular Chaperones. Biotechnol J. 2019 May 1;14(5):1800523.

224. Cappuccio JA, Hinz AK, Kuhn EA, Fletcher JE, Arroyo ES, Henderson PT, et al. Cell-Free Expression for Nanolipoprotein Particles: Building a High-Throughput Membrane Protein Solubility Platform. In: Doyle SA, editor. High Throughput Protein Expression and Purification: Methods and Protocols [Internet]. Totowa, NJ: Humana Press; 2009. p. 273–95. Available from: https://doi.org/10.1007/978-1-59745-196-3_18

225. Elena-Real CA, Urbanek A, Lund XL, Morató A, Sagar A, Fournet A, et al. Multi-site-specific isotopic labeling accelerates high-resolution structural investigations of pathogenic huntingtin exon-1. Structure. 2023 June 1;31(6):644-650.e5.

226. Catherine C, Oh SJ, Lee KH, Min SE, Won JI, Yun H, et al. Engineering thermal properties of elastin-like polypeptides by incorporation of unnatural amino acids in a cell-free protein synthesis system. Biotechnol Bioprocess Eng. 2015 June 1;20(3):417–22.

227. Ge X, Luo D, Xu J. Cell-Free Protein Expression under Macromolecular Crowding Conditions. PLOS ONE. 2011 Dec 8;6(12):e28707.

228. Yamaguchi H, Miyazaki M. Refolding Techniques for Recovering Biologically Active Recombinant Proteins from Inclusion Bodies. Biomolecules. 2014;4(1):235–51.

229. Yokoyama J, Matsuda T, Koshiba S, Tochio N, Kigawa T. A practical method for cell-free protein synthesis to avoid stable isotope scrambling and dilution. Anal Biochem. 2011;411(2):223–9.

230. Kigawa T, Yabuki T, Yoshida Y, Tsutsui M, Ito Y, Shibata T, et al. Cell-free production and stable-isotope labeling of milligram quantities of proteins. FEBS Lett. 1999 Jan 8;442(1):15–9.

231. Wu Y, Wang Z, Qiao X, Li J, Shu X, Qi H. Emerging Methods for Efficient and Extensive Incorporation of Non-canonical Amino Acids Using Cell-Free Systems. Front Bioeng Biotechnol [Internet]. 2020;8. Available from: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00863

232. Hong SH, Kwon YC, Jewett MC. Non-standard amino acid incorporation into proteins using Escherichia coli cell-free protein synthesis. Front Chem. 2014 June 10;2:34–34.

233. Liu CC, Schultz PG. Adding new chemistries to the genetic code. Annu Rev Biochem. 2010;79:413–44.

234. Liu Y, Davis RG, Thomas PM, Kelleher NL, Jewett MC. In vitro-Constructed Ribosomes Enable Multi-site Incorporation of Noncanonical Amino Acids into Proteins. Biochemistry. 2021 Jan 26;60(3):161–9.

235. Martin RW, Des Soye BJ, Kwon YC, Kay J, Davis RG, Thomas PM, et al. Cell-free protein synthesis from genomically recoded bacteria enables multisite incorporation of noncanonical amino acids. Nat Commun. 2018 Mar 23;9(1):1203.

236. Klammt C, Löhr F, Schäfer B, Haase W, Dötsch V, Rüterjans H, et al. High level cell-free expression and specific labeling of integral membrane proteins. Eur J Biochem. 2004 Feb 1;271(3):568–80.

237. Swartz J. Developing cell-free biology for industrial applications. J Ind Microbiol Biotechnol. 2006 July 1;33(7):476–85.

238. Dopp JL, Rothstein SM, Mansell TJ, Reuel NF. Rapid prototyping of proteins: Mail order gene fragments to assayable proteins within 24 hours. Biotechnol Bioeng. 2019 Mar 1;116(3):667–76.

239. Jewett MC, Calhoun KA, Voloshin A, Wuu JJ, Swartz JR. An integrated cell-free metabolic platform for protein production and synthetic biology. Mol Syst Biol. 2008 Jan 1;4(1):220.

240. Jewett MC, Swartz JR. Mimicking the Escherichia coli cytoplasmic environment activates long-lived and efficient cell-free protein synthesis. Biotechnol Bioeng. 2004 Apr 5;86(1):19–26.

241. Zhang YHP. Production of biocommodities and bioelectricity by cell-free synthetic enzymatic pathway biotransformations: Challenges and opportunities. Biotechnol Bioeng. 2010 Mar 1;105(4):663–77.

242. Jung GY, Stephanopoulos G. A Functional Protein Chip for Pathway Optimization and in Vitro Metabolic Engineering. Science. 2004 Apr 16;304(5669):428–31.

243. Kuruma Y, Ueda T. The PURE system for the cell-free synthesis of membrane proteins. Nat Protoc. 2015 Sept 1;10(9):1328–44.

244. Cui Y, Chen X, Wang Z, Lu Y. Cell-Free PURE System: Evolution and Achievements. Biodesign Res. 2022;2022:9847014.

245. Zawada JF, Yin G, Steiner AR, Yang J, Naresh A, Roy SM, et al. Microscale to manufacturing scale-up of cell-free cytokine production—a new approach for shortening protein production development timelines. Biotechnol Bioeng. 2011 July 1;108(7):1570–8.

246. Gill DR, Pringle IA, Hyde SC. Progress and Prospects: The design and production of plasmid vectors. Gene Ther. 2009 Feb 1;16(2):165–71.

247. Studier FW, Moffatt BA. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J Mol Biol. 1986 May 5;189(1):113–30.

248. Stryer L. Biochemistry. 7th ed. New York: W. H. Freeman and Company; 2012.

249. Calhoun KA, Swartz JR. Energy Systems for ATP Regeneration in Cell-Free Protein Synthesis Reactions. In: Grandi G, editor. In Vitro Transcription and Translation Protocols

[Internet]. Totowa, NJ: Humana Press; 2007. p. 3–17. Available from: https://doi.org/10.1007/978-1-59745-388-2_1

250. Garenne D, Haines MC, Romantseva EF, Freemont P, Strychalski EA, Noireaux V. Cell-free gene expression. Nat Rev Methods Primer. 2021 July 15;1(1):49.

251. Spirin AS, Baranov VI, Ryabova LA, Ovodov S, Alakhov YB. A Continuous Cell-Free Translation System Capable of Producing Polypeptides in High Yield. Science. 1988 Nov 25;242(4882):1162–4.

252. Kigawa T, Yokoyama S. A Continuous Cell-Free Protein Synthesis System for Coupled Transcription-Translation1. J Biochem (Tokyo). 1991 Aug 1;110(2):166–8.

253. Endo Y, Otsuzuki S, Ito K, Miura K ichiro. Production of an enzymatic active protein using a continuous flow cell-free translation system. J Biotechnol. 1992 Sept 1;25(3):221–30.

254. Zubay G. IN VITRO SYNTHESIS OF PROTEIN IN MICROBIAL SYSTEMS. Annu Rev Genet. 1973 Dec 1;7(1):267–87.

255. Gregorio NE, Levine MZ, Oza JP. A User's Guide to Cell-Free Protein Synthesis. Methods Protoc. 2019;2(1).

256. Dopp JL, Jo YR, Reuel NF. Methods to reduce variability in E. Coli-based cell-free protein expression experiments. Synth Syst Biotechnol. 2019 Dec 1;4(4):204–11.

257. Kigawa T, Muto Y, Yokoyama S. Cell-free synthesis and amino acid-selective stable isotope labeling of proteins for NMR analysis. J Biomol NMR. 1995 Sept 1;6(2):129–34.

258. Ozawa K, Headlam MJ, Schaeffer PM, Henderson BR, Dixon NE, Otting G. Optimization of an Escherichia coli system for cell-free synthesis of selectively 15N-labelled proteins for rapid analysis by NMR spectroscopy. Eur J Biochem. 2004 Oct 1;271(20):4084–93.

259. Matsuda T, Koshiba S, Tochio N, Seki E, Iwasaki N, Yabuki T, et al. Improving cell-free protein synthesis for stable-isotope labeling. J Biomol NMR. 2007 Mar 1;37(3):225–9.

260. Koizumi M, Hiratake J, Nakatsu T, Kato H, Oda J. A Potent Transition-State Analogue Inhibitor of Escherichia coli Asparagine Synthetase A. J Am Chem Soc. 1999 June 1;121(24):5799–800.

261. Manning JM, Moore S, Rowe WB, Meister A. Identification of L-methionine S-sulfoximine as the diastereoisomer of L-methionine SR-sulfoximine that inhibits glutamine synthetase. Biochemistry. 1969 June;8(6):2681–5.

262. John RA, Charteris A. The reaction of amino-oxyacetate with pyridoxal phosphate-dependent enzymes. Biochem J. 1978 June;171(3):771–9.

263. Levin R, Löhr F, Karakoc B, Lichtenecker R, Dötsch V, Bernhard F. E. coli "Stablelabel" S30 lysate for optimized cell-free NMR sample preparation. J Biomol NMR. 2023 Aug;77(4):131–47.

264. Loscha KV, Herlt AJ, Qi R, Huber T, Ozawa K, Otting G. Multiple-site labeling of proteins with unnatural amino acids. Angew Chem Int Ed Engl. 2012 Feb 27;51(9):2243–6.

265. Kato K, Matsunaga C, Igarashi T, Kim H, Odaka A, Shimada I, et al. Complete assignment of the methionyl carbonyl carbon resonances in switch variant anti-dansyl antibodies labeled with [1-13C]methionine. Biochemistry. 1991 Jan 8;30(1):270–8.

266. Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 2014 Jan 1;42(D1):D304–9.

267. Rotkiewicz P, Skolnick J. Fast procedure for reconstruction of full-atom protein models from reduced representations. J Comput Chem. 2008 July 15;29(9):1460–5.

268. Rodrigues J, Teixeira J, Trellet M, Bonvin A. pdb-tools: a swiss army knife for molecular structures [version 1; peer review: 2 approved]. F1000Research [Internet]. 2018;7(1961). Available from: https://f1000research.com/articles/7-1961/v1

269. Tange O. GNU Parallel - The Command-Line Power Tool. Login USENIX Mag. 2011 Feb;36:42–7.

270. Dewhurst CD. Graphical reduction and analysis small-angle neutron scattering program: ıt GRASP. J Appl Crystallogr. 2023 Oct;56(5):1595–609.

271. Pedersen JS, Posselt D, Mortensen K. Analytical treatment of the resolution function for small-angle scattering. J Appl Crystallogr. 1990 Aug;23(4):321–33.

272. Cioni P, Strambini GB. Effect of heavy water on protein flexibility. Biophys J. 2002 June;82(6):3246–53.

273. Haidar Y, Konermann L. Effects of Hydrogen/Deuterium Exchange on Protein Stability in Solution and in the Gas Phase. J Am Soc Mass Spectrom. 2023 July 5;34(7):1447–58.

274. Cummings DL, Wood JL. The strength of the deuterium bond. J Mol Struct. 1974 Oct 1;23(1):103–12.

275. Trewhella J, Jeffries CM, Whitten AE. 2023 update of template tables for reporting biomolecular structural modelling of small-angle scattering data. Acta Crystallogr Sect Struct Biol. 2023 Feb 1;79(Pt 2):122–32.

276. Konarev PV, Volkov VV, Sokolova AV, Koch MHJ, Svergun DI. PRIMUS: a Windows PC-based system for small-angle scattering data analysis. J Appl Crystallogr. 2003 Oct;36(5):1277–82.

277. Cotton JP. Variations on contrast in SANS: determination of self and distinct correlation functions. Adv Colloid Interface Sci. 1996 Dec 1;69(1):1–29.

278. Brünger AT, Clore GM, Gronenborn AM, Saffrich R, Nilges M. Assessing the quality of solution nuclear magnetic resonance structures by complete cross-validation. Science. 1993 July 16;261(5119):328–31.

279. O'Brien ES, Lin DW, Fuglestad B, Stetz MA, Gosse T, Tommos C, et al. Improving yields of deuterated, methyl labeled protein by growing in H2O. J Biomol NMR. 2018 Aug;71(4):263–73.

280. Tzeng SR, Pai MT, Kalodimos CG. NMR studies of large protein systems. Methods Mol Biol Clifton NJ. 2012;831:133–40.

281. Barton S, Jacak R, Khare SD, Ding F, Dokholyan NV. The Length Dependence of the PolyQ-mediated Protein Aggregation *. J Biol Chem. 2007 Aug 31;282(35):25487–92.

282. Larsen AH. Optimal weights and priors in simultaneous fitting of multiple small-angle scattering datasets. J Appl Crystallogr. 2025 June 1;58(Pt 3):934–47.

283. Huang EC, Henion JD. LC/MS and LC/MS/MS determination of protein tryptic digests. J Am Soc Mass Spectrom. 1990 Apr 1;1(2):158–65.

284. Laugesen S, Roepstorff P. Combination of two matrices results in improved performance of maldi ms for peptide mass mapping and protein analysis. J Am Soc Mass Spectrom. 2003 Sept 1;14(9):992–1002.

285. Panjkovich A, Svergun DI. CHROMIXS: automatic and interactive analysis of chromatography-coupled small-angle X-ray scattering data. Bioinformatics. 2018 June 1;34(11):1944–6.

286. Dewhurst C. Grasp [Internet]. Institut Laue Langevin; 2023. Available from: https://www.ill.eu/users/support-labs-infrastructure/software-scientific-tools/grasp

287. Urbanek A, Popovic M, Morató A, Estaña A, Elena-Real CA, Mier P, et al. Flanking Regions Determine the Structure of the Poly-Glutamine in Huntingtin through Mechanisms Common among Glutamine-Rich Human Proteins. Structure. 2020 July 7;28(7):733-746.e5.

288. Shen Y, Bax A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. J Biomol NMR. 2010 Sept 1;48(1):13–22.

289. Badaczewska-Dawid AE, Kolinski A, Kmiecik S. Computational reconstruction of atomistic protein structures from coarse-grained models. Comput Struct Biotechnol J. 2020 Jan 1;18:162–76.

290. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-chain Amide Orientation. J Mol Biol. 1999;285:1735–47.

11 APPENDIX

11.1 SELECTIVE DEUTERATION SCRIPT

```
# -*- coding: utf-8 -*-

"""

@author: Amin & Xamuel

"""

from biopandas.pdb import PandasPdb

import os

import fileinput

# Path to unlabelled ensemble pdb-files & path to directory of the labelled ensemble. Does not create the folder automaticly

directory = 'Structures/Unlabelled_Ensemble/'

outputpath = 'Structures/Labelled_Ensemble/'

for file in os.listdir(directory):

    if file.endswith(".pdb"):

        filefrompath = os.path.basename(file)

        filename = (os.path.splitext(filefrompath)[0])

        ppdb = PandasPdb().read_pdb(directory + file)

        ppdb.df['ATOM'].head()

            # Specify which aminoacids to deuterate in listofresidues.

        listofresidues = ['PRO']

        mask = (ppdb.df['ATOM']['residue_name'].isin(listofresidues)) &  (ppdb.df['ATOM']['element_symbol'] == 'H')

            # In case of inverted labelling (deuterated protein with protonated labelling), comment out mask

            # and uncomment the two following lines.

        #invertedmask = (ppdb.df['ATOM']['residue_name'].isin(listofresidues)) &  (ppdb.df['ATOM']['element_symbol'] == 'H')

        #mask = ~invertedmask

        ppdb.df['ATOM'].loc[mask,['atom_name']] = ppdb.df['ATOM'].loc[mask,'atom_name'].str.replace("H", "D")

        ppdb.df['ATOM'].loc[mask,['element_symbol']] = ppdb.df['ATOM'].loc[mask,'element_symbol'].str.replace("H", "D")

            # Remember to change output name below

        ppdb.to_pdb(outputpath+filename+'_Label.pdb')

    else:

        continue
```

11.2 EXCHANGE OF LABILE HYDROGENS SCRIPT

```
# -*- coding: utf-8 -*-

#   Inspired by DeuterationLabily.py by Sergei Grudinin.

#   Adapted by Xamuel L. Lund to accept multiple files and deuteration levels.

#   Additionally it now accepts partially deuterated input files and will exchange

#    labile deuterium atoms as well.
```

```
#   With all modules installed run the script in the folder of your structure files
#   (PDB files) and it will create subfolders for each deuteration level in the
#   list "rates".
"""
@author: Xamuel
"""
import os,sys
import random
from glob import glob
import pymol
from pymol import cmd
        #-cq option makes sure pymol GUI doesn't open each file
pymol.finish_launching(['pymol','-cq'])


        #The function will randomly deuterate hydrogen depending on the given rate
def myfunc(model,index):
    change = random.random() < rate
    # print('%s`%s/%s %s %d' % (resn ,resi, name,element, change))
    if (change) :
        cmd.alter('%s and index %d'%(model,index),'elem = \'D\'; name = \'D\'')


        #Change the rates-list to include the % D2O of your solution
rates = [0.0, 0.2, 0.4, 0.6, 0.8, 1.0]


for rate in rates:
    ratename = (str(int(rate*100))+'pD2O')
    try:
        # Attempts to create directory for the current rate of D2O
        os.makedirs(ratename)
    except FileExistsError:
        # directory already exists
        pass


    listoffiles = glob('*.pdb') #Uses all .pdb files in the working directory
    for file in listoffiles:
        e = 0
        while True:
            cmd.load(file)
            name = os.path.splitext(file)[0] #Extracts the name of the file without file extension
            name_deut = name+'_'+ratename
```

```
            #First step (cmd.alter) changes all labile D-atoms to H
        cmd.select("nolabileD","element D and neighbor  element C")
        cmd.select("labileD","element D and not nolabileD")
        cmd.alter('LabileD','elem = \'H\'; name = \'H\'')

            #Second step is to select all labile hydrogens and randomly deuterate with the given rate
        cmd.select("nolabile","element H and neighbor  element C")
        cmd.select("labile","element H and not nolabile")
        total = cmd.count_atoms('Labile')
        cut = rate * total
        myspace = {'myfunc': myfunc}
        cmd.iterate('(labile)', 'myfunc(model,index)', space=myspace)
        cmd.deselect()
        if rate == 0 or rate == 1:
            cmd.save(ratename+'/'+name_deut+'.pdb')
            cmd.delete('all')
            break
            #This step will count the number of labile deuterium and compare it to the rate
            #If the nr. of deuterium is too high or low it reruns the deuteration
        cmd.select("nolabileNew","element D and neighbor  element C")
        cmd.select("labileNew","element D and not nolabileNew")
        check = 0
        check = cmd.count_atoms('labileNew')
        Upper = cut * 1.1
        lower = cut * 0.9

        if check > lower and check < Upper :
            # print(str(check)+' sample '+name_deut+' correct - contienuing')
            cmd.save(ratename+'/'+name_deut+'.pdb')
            cmd.delete('all')
            break
        if check < lower or check > Upper :
            # print(str(check)+' sample '+name_deut+' should be resimulated')
            cmd.delete('all')
            e = e + 1
        if e > 50:
            print('Script cannot compute correct deuteration pattern. Errorcount: '+str(e))
            cmd.delete('all')
            sys.exit(0)
    print(str(rate)+' have been simulated')
#If sys.exit(0) causes a traceback (not an error just the traceback)
```

```
#Python is running the -i option. Attempt to run without it or ignore the traceback
sys.exit(0)
```

11.3 SCRIPT FOR CALCULATING THE REGENERATED FIT AND χ^2-VALUES

```
# -*- coding: utf-8 -*-
"""
@author: Xamuel
@ Amin Sagar assisted in scaling and fitting functions
"""
import numpy as np
from scipy.optimize import minimize
from scipy.interpolate import griddata
# Suffix of the individual pdb structures calculated for each labelling pattern
strucnames = ['_SAXS', '_hHtt_100p', '_hHtt_20p','_D-QE_100p', '_dQE_0p','_H-QE_40p', '_H-QE_0p', '_dHtt_40p', '_dHtt_0p']
# Name of each experimental datafile, used to calculate error profiles.
expnames =['H16-SAXS.dat', 'H16_100.dat', 'H16_20.dat','D-QE_100p.dat', 'D-QE_0p.dat','H-QE_40p.dat',
                                                    'H-QE_0p.dat', 'dHtt_40p.dat', 'dHtt_0p.dat']
# Fill out the six datapoints below and launch the script from the command line
Resultpath ='../Profiles/2310_Results/'      #Path to the initial fit of experimental data
Prefix = 'Multi'                             #Prefix of the fitting data
D2O = 'Trio15'                               #Suffix of the fitting data
struc_path='../Profiles/Multi-1/'            #Path to all atomistic structures (with suffix from strucnames)
run = '_'+D2O+'_RegFit.int'                  #Name to save the calculated profile
runpath = 'Regenerated_Fits'                 #Name of or path to folder to save regenerated fits
cycle = 200                                  #Number of cycles of the initial EOM fitting

def get_scale(ExpProfile, TheoProfile):
                ExperimentalErrorsquare = np.square(ExpProfile[:,2])
                f1 = sum( ExpProfile[:,1]*TheoProfile[:,1] / ExperimentalErrorsquare )
                f2 = sum( TheoProfile[:,1] * TheoProfile[:,1]/ ExperimentalErrorsquare )
                if( f2 == 0 ):
                        return 0.0
                return f1 / f2

def fit_chi (scaleCnst, ExpProfile,TheoProfile):
    # Function to fit two profiles by optimizing scale and constant subtraction
    chisquare = np.sum(np.square((ExpProfile[:,1]-((TheoProfile[:,1]*scaleCnst[0])+scaleCnst[1]))/ExpProfile[:,2]))/(len(ExpProfile)-1)
    return chisquare

def get_chi (ExpProfile,TheoProfile):
    # Function to fit two profiles by optimizing scale and constant subtraction
```

```
    chisquare = np.sum(np.square((ExpProfile[:,1]-(TheoProfile[:,1]))/ExpProfile[:,2]))/(len(ExpProfile)-1)
    return chisquare


def regenerate_fit(par,theoscaled):
    # Function to regenerate fitted profiles from the output of minimizer
    fit = np.zeros(np.shape(theoscaled))
    fit[:,0] = theoscaled[:,0]
    fit[:,1] = (theoscaled[:,1]*par[0])+par[1]
    return fit
for i in range(len(strucnames)):
            name = strucnames[i]
            Exp_Data = expnames[i]
            CHIcheck = np.loadtxt(Resultpath+Prefix+'_'+D2O+'_profiles.fit', skiprows=cycle, usecols=(1), max_rows=1)
            cyclerow = 55*(CHIcheck-1)+3
            Result = np.loadtxt(Resultpath+Prefix+'_'+D2O+'_best_genes.txt', usecols=(0,1))
            RepStruc = []
            for rep in range(0,len(Result[:,0])):
                RepStruc.extend([Result[rep,0]]*int(Result[rep,1]))

            Qval = np.loadtxt(struc_path+str(int(Result[0,0]))+name+'.int', usecols=(0), skiprows=1)
            Matrix = np.zeros((200,len(RepStruc)))
            for i in range(0,len(RepStruc)):
                Matrix[:,i] = np.loadtxt(struc_path+str(int(RepStruc[i]))+name+'.int', usecols=(1), skiprows=1, max_rows=200)
            newI = np.mean(Matrix, axis = 1)
            Average = np.transpose(np.vstack((Qval,newI)))
            ExpProfile = np.loadtxt(struc_path+Exp_Data, usecols=(0,1,2))

                #Regridding and scaling averaged data
            grid_z0 = griddata(Average[:,0], Average[:,1], ExpProfile[:,0], method='cubic')
            NewAverage = np.transpose(np.vstack((ExpProfile[:,0],grid_z0)))
            Scale = get_scale(ExpProfile,NewAverage)
            ScaledAverage = np.transpose(np.vstack((ExpProfile[:,0],Scale * NewAverage[:,1])))

                #Fitting averaged data to experimental profile
            bnds = ((-5,5),(-1*ExpProfile[1,1]*0.02, ExpProfile[0,1]*0.02))
            scaleConst = minimize(fit_chi,[1,0], args=(ExpProfile, ScaledAverage), bounds=bnds)
            fitprofile = regenerate_fit(scaleConst.x, ScaledAverage)
            chi = get_chi(ExpProfile,fitprofile)
            np.savetxt(runpath+'/Fitted'+name+run, fitprofile, header= 'Chi2 of averaged profile fit to experimental data: '
+str(round(float(chi),4)))
            print (name+' chi2 = '+str(chi))
```

11.4 H16 Protein Sequence

Htt Q16 Exon 1 – Linker – sfGFP – His6-tag

MATLEKLMKA FESLKSFQQQ QQQQQQQQQQ QQQPPPPPPP PPPPQLPQPP PQAQPLLPQP QPPPPPPPPP PGPAVAEEPL HRPEASLEVL FQGPGSHMAS KGEELFTGVV PILVELDGDV NGHKFSVRGE GEGDATNGKL TLKFICTTGK LPVPWPTLVT TLTYGVQCFS RYPDHMKRHD FFKSAMPEGY VQERTISFKD DGTYKTRAEV KFEGDTLVNR IELKGIDFKE DGNILGHKLE YNFNSHNVYI TADKQKNGIK ANFKIRHNVE DGSVQLADHY QQNTPIGDGP VLLPDNHYLS TQSVLSKDPN EKRDHMVLLE FVTAAGITHG MDELYKLERA PGGGSHHHHH H

11.5 H36 Protein Sequence

Htt Q36 Exon 1 – Linker – sfGFP – His6-tag

MATLEKLMKA FESLKSFQQQ QQQQQQQQQQ QQQQQQQQQQ QQQPPPPPPP PPPPQLPQPP PQAQPLLPQP QPPPPPPPPP PGPAVAEEPL HRPEASLEVL FQGPGSHMAS KGEELFTGVV PILVELDGDV NGHKFSVRGE GEGDATNGKL TLKFICTTGK LPVPWPTLVT TLTYGVQCFS RYPDHMKRHD FFKSAMPEGY VQERTISFKD DGTYKTRAEV KFEGDTLVNR IELKGIDFKE DGNILGHKLE YNFNSHNVYI TADKQKNGIK ANFKIRHNVE DGSVQLADHY QQNTPIGDGP VLLPDNHYLS TQSVLSKDPN EKRDHMVLLE FVTAAGITHG MDELYKLERA PGGGSHHHHH H

12 RÉSUMÉ SUBSTANTIEL EN FRANÇAIS

La maladie de Huntington (HD) est une maladie neurodégénérative génétique dont les symptômes caractéristiques comprennent des effets moteurs, tels que des secousses musculaires et des troubles de la marche, avec, à un stade avancé, une bradykinésie, une rigidité ou une perte de la capacité de marcher, ainsi que des effets psychiatriques (la dépression, l'apathie et l'irritabilité), et des effets cognitifs, qui peuvent affecter la mémoire, l'attention et la flexibilité mentale, et peuvent entraîner une démence et une inhibition cognitive à un stade avancé de la maladie. La maladie est causée par une mutation du gène codant pour la protéine huntingtine (Htt), situé sur le locus chromosomique 4p16.3. Cette mutation entraîne une augmentation du nombre de trinucléotides CAG dans le premier exon de la protéine qui, après traduction, augmente le nombre de glutamines dans le tractus poly-glutamine (Poly-Q) de la région N-terminale intrinsèquement désordonnée de la protéine. Les symptômes de la HD ne se manifestent que chez les personnes présentant un tractus Poly-Q de plus de 35 glutamines consécutives, ce qui correspond au seuil pathologique. Il est important de noter que la longueur du tractus Poly-Q au-delà de ce seuil est corrélée à la précocité d'apparition et à la gravité de la pathologie. L'allongement anormal d'un tractus poly-Q est caractéristique des troubles poly-Q, qui comprennent, en plus de la HD, le syndrome de Haw River, l'atrophie musculaire spinale et bulbaire (SBMA) et l'ataxie spinocérébelleuse (SCA).

La protéine Htt complète contient environ ~3.142 acides aminés pour un poids moléculaire de ~347,6 kDa. L'exon 1 de la Htt ne contient que 83 acides aminés N-terminaux, ou plus selon la longueur du tractus poly-Q. Cette région N-terminale est une région de faible complexité (LCR) contenant trois domaines: une région N-terminale de 17 résidus (N17), le tractus poly-Q et une région riche en proline (PRR). Le tractus poly-Q et la PRR présentent chacun un biais de composition vers un seul résidu: la Glutamine (Q) ou la proline (P). La pathogénicité de la HD a été associée à des corps d'inclusion découverts dans le cytoplasme et le noyau des neurones striataux, composés de fragments du tractus poly-Q allongé de l'Htt. On pense que les agrégats solubles de fragments Htt provoquent l'apoptose des cellules nerveuses et on a observé que des fragments de la protéine pathogène sont absorbés par les cellules adjacentes, induisant également leur mort. Les niveaux de Htt soluble dans le liquide céphalo-rachidien ont été identifiés comme un biomarqueur de la HD. De plus, l'addition de l'exon 1 de Htt suffit à reproduire les effets cellulaires de la maladie. C'est pourquoi ce fragment est généralement utilisé pour les études mécanistiques et biophysiques de la HD. Dans cette étude, ce fragment, dans sa version pathogène (36 glutamines dans le tractus poly-Q) et non-pathogène (16

glutamines), est fusionné à la sfGFP (version très stable « superfold » de la protéine verte fluorescente) pour générer les constructions nommées H16 et H36.

La thèse de doctorat présentée ici comprend trois objectifs principaux :

1) Développer une stratégie d'expression robuste pour la production de protéines spécifiquement deutérées aux niveaux des homorépétitions.

2) Développer une approche intégrant de manière optimale les données de diffusion aux petits angles de rayons X (SAXS) et de neutrons (SANS), y compris avec variation de contraste et marquage aminoacide-spécifique, afin de générer des ensembles conformationnels plausibles de protéine à régions de faible complexité.

3) Élucider les différences conformationnelles entre les versions pathogène et non-pathogène de l'exon1 de Htt à l'aide d'une combinaison de mesures SAXS et SANS d'échantillons deutérés de façon amino-acide spécifique.

Le SAXS et le SANS sont des méthodes utilisées pour analyser la structure nanométrique d'échantillons aussi diverses que des polymères, des échantillons biologiques, des alliages métalliques ou des nanoparticules, et permettent de suivre son evolution au cours d'une réaction. Ici, la diffusion aux petits angles (SAS) est utilisée pour l'analyse de macromolécules biologiques en solution. Le SAS offre notamment des informations structurelles à basse résolution sur des protéines flexibles ou de grands complexes biomoléculaires, complémentant parfaitement des techniques à haute résolution telles que la cristallographie des protéines ou la microscopie électronique cryogénique (cryoEM). Le SAXS est plus couramment utilisé que le SANS en raison de la facilité d'accès aux installations. Plusieurs logiciels ont été publiés au fil des ans pour faciliter la simulation, l'ajustement et la validation des données SAS. D'autre part, des développements dans les environnements échantillon, tels que la chromatographie d'exclusion stérique couplée au SAXS (SEC-SAXS), ont également permis d'augmenter le type d'échantillons pouvant être mesurés. Le SEC-SANS a récemment été développé sur l'instrument D22 de l'ILL. Le SEC-SAS est idéal pour l'étude d'échantillons susceptibles de s'agréger, qui étaient auparavant difficiles (voire impossibles) à mesurer.

L'interprétation des données SAS par optimisation d'ensemble conformationnel permet l'analyse de protéines flexibles, telles que les protéines intrinsèquement désordonnées, et a été la principale méthode utilisée dans le présent projet. Contrairement aux solutions à structure unique, l'optimisation d'ensemble conformationnel (EOM) fournit un ensemble de structures

qui décrivent collectivement les données expérimentales. Les ensembles de structures peuvent être générés à partir de modèles atomistiques ou de modèles à gros grains. Une fois qu'un ensemble exhaustif de structures a été généré, la courbe de diffusion de chaque conformation doit être calculée puis un sous-ensemble de courbes, et donc de conformations, est sélectionné de façon itérative pour sa propension à décrire des données expérimentales.

La deutération est un outil puissant dans la diffusion neutronique pour augmenter les informations structurelles et dynamiques recueillies lors d'une expérience. Les échantillons biologiques contiennent un grand nombre d'atomes d'hydrogène. Par exemple, dans les proteines etudiees ici, 49,4 % des atomes sont des hydrogènes et, parmis eux 20,8 % sont non-labiles, c'est-à-dire qu'ils ne s'échangent pas avec les hydrogènes du tampon. Ainsi, ils peuvent être remplacés de façon pérenne par des deutériums au moment de la synthèse protéique. En raison de la différence de densité de longueur de diffusion (SLD) des neutrons par l'hydrogène (H1) et son isotope le deutérium (D ou H2), ce remplacement a un impact majeur sur la SLD mesurée des échantillons. En cristallographie neutronique, la diffraction de neutrons par des cristaux de protéines deutérées a permis aux chercheurs de recueillir des informations qui ne sont généralement pas disponibles dans les expériences de cristallographie aux rayons X, telles que l'emplacement des atomes d'hydrogène, les liaisons hydrogène et l'orientation des molécules d'eau dans les sites de liaison. Les échantillons deutérés ont également permis d'étudier plus en détail les complexes protéiques à l'aide du SANS en modifiant la SLD de composants individuels d'un complexe protéine-protéine et de la solution tampon. Le composant dont la SLD correspondant à celle de la solution tampon voit sa contribution à la diffusion annulée, et les données de diffusion mesurées du complexe ne décrivent donc que ses partenaires. Cette stratégie expérimentale est appelée « contrast matching ».

L'expression protéique sans cellules (CF) est une technique in vitro qui utilise les mécanismes de transcription-traduction purifiés à partir d'extraits cellulaires pour produire des protéines recombinantes. Contrairement à l'expression in vivo, la CF est un système ouvert qui permet la manipulation directe de l'expression des protéines et donc leur deutération. L'experimentateur peut notamment introduire des chaperones, des nano-disques, des acides aminés non naturels ou des agents d'encombrement stérique. Un autre avantage de la synthèse CF est qu'elle permet de produire des protéines qui seraient toxiques en conditions recombinantes in vivo, sujettes à l'agrégation ou susceptibles d'être exprimées dans des corps d'inclusion. Dans le contexte de cette thèse, le principal avantage de la synthèse CF est qu'elle permet l'insertion d'un certain nombre d'acides aminés deutéres afin d'obtenir des échantillons

de protéines marquées de façon amino-acide spécifique. Cependant, un phénomène naturel vient limiter cette stratégie : l'effet de brouillage (ou scrambling). Certains résidus, comme la glutamine et l'acide glutamique, peuvent être enzymatiquement transformés l'un en l'autre. Il n'est donc pas possible de les marquer de façon indépendante. Dans cette étude, cet effet a été résolu en deutérant les deux résidus (Gln & Glu) et en optimisant le milieu de synthèse CF afin de ne pas inclure d'autres sources de ces acides aminés, c'est-à-dire en changent le tampon de potassium-acide glutamique pour du potassium-acide acétique. La précision du processus de marquage et l'homogénéité des échantillons ont été confirmées par spectrométrie de masse. Ainsi, les constructions H16 et H36 ont été produites selon diffèrents marquages amino-acide spécifiques afin d'être mesurées par SAS dans des tampons aux niveaux de deutération variés, les données étant ensuite analysées par optimisation d'ensembles conformationnels.

L'ensemble exhaustif des conformations des deux versions de l'exons 1 de Htt (avec 16 ou 36 glutamines) ont été construites à partir d'une base de données de fragments tripeptidiques (SCOPe) afin de garantir leur conformité avec des données structurelles existantes. Les conformations de H16 et H36 ont ensuite été construites à partir de la combinaison de toutes ces conformations possibles de l'exon 1 avec la structure cristalline de la sfGFP comprenant une étiquette Histidine en C-terminale, à l'aide d'un linker 3C. Des scripts ont été écrits pour marquer virtuellement des résidus spécifiques avec du deutérium, puis échanger des hydrogènes labiles à un pourcentage donné afin d'imiter l'échange dans les solutions tampons H2O/D2O.

À partir de l'ensemble théorique initial, huit modèles spécifiques de marquage au deutérium et six taux d'échange des hydrogènes labiles ont été produits pour H16 et H36. Le marquage amino-acide spécifique était axé sur les résidus Gln, Glu et Pro afin d'exploiter le biais de composition de l'exon 1 (tractus poly-Q et PRR). Les profils SANS théoriques pour toutes les conformations et tous les états de deutération ont été calculés par CRYSON en vue d'analyses ultérieures. Les ensembles théoriques ont montré un enrichissement en informations structurelles en fonction du modèle de marquage et du pourcentage de D2O dans la solution tampon. L'analyse de ces données synthétiques a également guidé la sélection des échantillons optimaux à produire et à mesurer par SAS.

Les données de SEC-SANS collectées sur D22 (ILL) et les données de SEC-SAXS mesurées à la ligne de lumière Swing du Synchrotron Soleil ont été intégrées aux modèles atomistiques

mentionnés précédemment à l'aide de la méthode d'optimisation d'ensemble (EOM). 28 profils expérimentaux SANS pour les constructions pathogènes (H36) et non-pathogènes (H16), ainsi que des données SAXS pour H16 et H36 dans 0 % et 100 % de D2O ont été collectés. Les expériences aux rayons X ont montré que la deutération du tampon n'avait qu'un impact négligeable sur les échantillons de protéines. Les échantillons de protéines H16 avec des résidus Gln et Glu deutérés ont été mesurés, et leurs profiles montrent une bonne concordance avec les profils de diffusion théoriques en ce qui concerne le rayon de gyration (Rg), la distance maximale de la particule (Dmax) et l'intensité à angle nul (I(0)). Le marquage avec des résidus P deutérés a modifié la structure de l'exon 1 de Htt dans les échantillons de protéines H16 et H36, et laisse entrevoir un impact inexploré du marquage spécifique de la proline. Les six ensembles de données intégrant le PRR deutéré n'étaient pas compatibles avec les ensembles structurels produits au cours du projet.

Afin de valider la complémentarité des données SAS mesurées au cours du projet, une validation croisée a été mise au point pour comparer les données de diffusion entre elles. La validation croisée a confirmé que les données SAXS et SANS pouvaient être combinées pour mieux définir l'ensemble conformationnel adopté par H16. La validation croisée des données SANS a montré que des données à rapport signal/bruit élevé et aux marquages variés pouvaient être combinées, validant l'objectif d'étudier les protéines à LCR par marquage segmentaire. Elle a en outre montré l'importance de la qualité des données pour ce type d'analyse : deux courbes, de H16 et H36, n'apportent aucun effet discriminant à l'optimisation d'ensemble conformationnel car leur rapport signal/bruit et leur intensité de diffusion sont trop faibles.

Des combinaisons comprenant jusqu'à cinq ensembles de données H16 SANS et SAXS ont été réalisées. Les ensembles obtenus, compatibles avec tous les ensembles de données expérimentales, indiquent que l'exon 1 est une protéine allongée avec une structuration partielle dans le tractus Poly-Q. Les échantillons des constructions protéiques pathogènes H36 n'étaient pas comparables aux ensembles de données H16, en raison de la faible quantité de données SAS ayant un rapport signal/bruit suffisant.

En résumé, j'ai développé une méthode originale qui, en exploitant le marquage isotopique aminoacide-spécifique permis par la synthèse CF, le positionnement précis du deutérium dans les modèles atomistiques et la capacité à ajuster simultanément plusieurs ensembles de données, améliore le contenu informatif structurel des données SAS. Cette stratégie ouvre la

voie à l'étude des régions à faible complexité, une famille de protéines restée largement hors de portée de la biologie structurale en raison de l'absence de méthodologies appropriées.