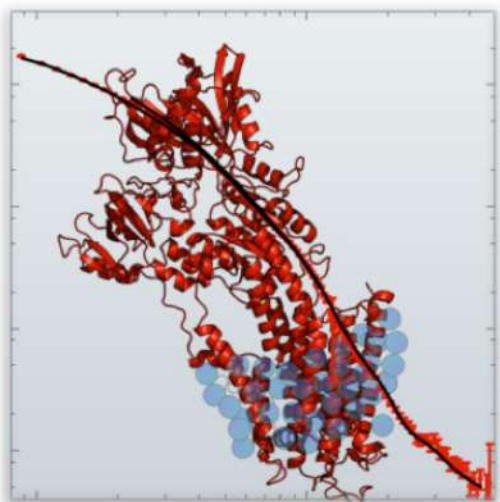




Thesis for the degree of Philosophiae Doctor

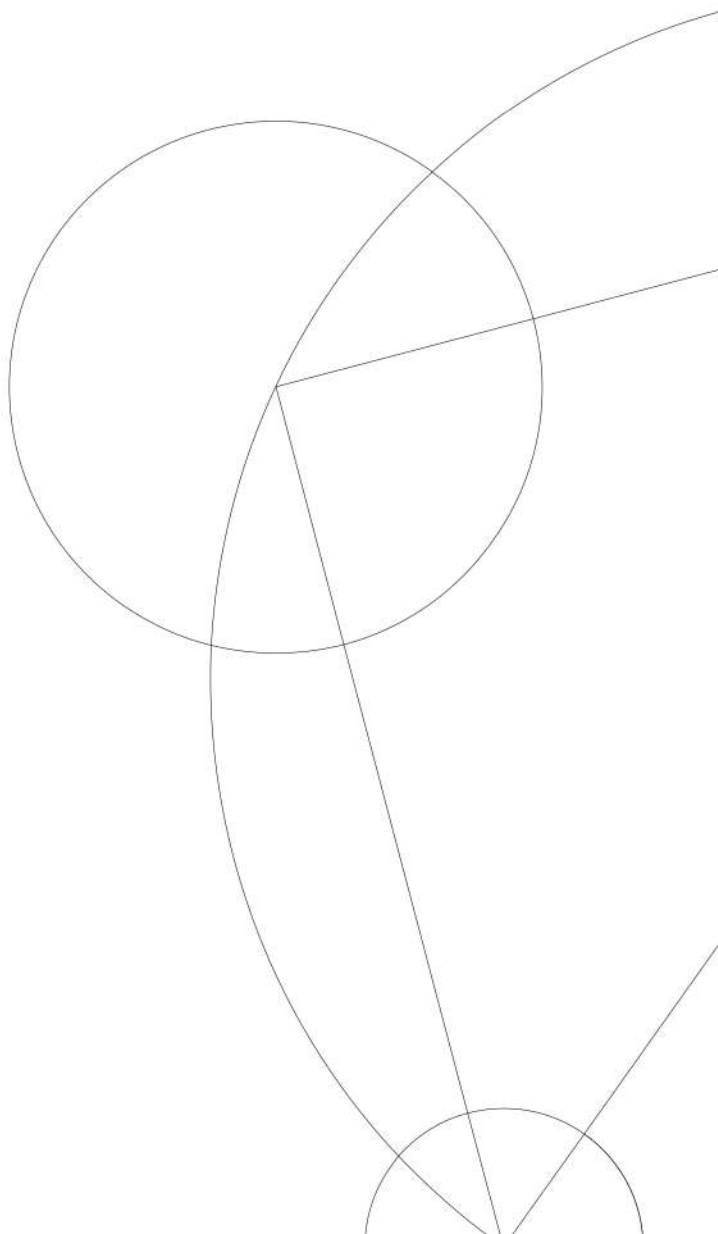
# Analytical tools for structure determination of protein complexes with small-angle scattering

Andreas Haahr Larsen



## Supervised by

Professor Lise Arleth,  
Niels Bohr Institute,  
University of Copenhagen



PhD thesis in biophysics

For the degree of Philosophiae Doctor

©2018 Andreas Haahr Larsen

andreas.larsen@nbi.ku.dk

Niels Bohr Institute,

University of Copenhagen,

1st edition, September 3, 2018

*Dedicated to Erik Sode, my physics teacher at Løgumkloster Efterskole,  
whose teaching inspired me to venture into the stunning world of physics,  
and to my parents, Gurli & Henning, for support in all things.*

The picture on the cover is designed by Nicholas Skar-Gislinge, and shows the crystal structure of the SERCA pump with an invisible detergent corona. The background is the small-angle neutron scattering data of the protein, with corresponding fit. Data were collected and analyzed in Midtgaard *et al.* (2018), Paper III.

## Acknowledgments

There are a few people I would like to acknowledge. First of all, my supervisor Lise Arleth for facilitating the project, for giving me much freedom to pursue my interests, and for leading a very ambitious and yet social and balanced research group. Thanks for valuable supervision in scientific matters, and for much help and time in administrative and bureaucratic matters, which are also part of a PhD. In that context, thanks to our secretary Gitte - you are a hero! I thank University of Copenhagen and CoNeXT for co-funding the project. During my time in the group, I have performed experiments at 9 different SAXS and SANS beamlines in 5 different countries. So thanks to the beamline scientists who maintain the instruments and I would like to thank for great support. I appreciate the collaboration I have had with Steen Hansen and thank for enjoyable and interesting discussions. Thanks to Søren Roi Midtgaard for supervision and collaboration through many years in the group. Also great thanks to Søren and Lise for involving me in the development of a very fruitful method using "invisible" detergents. As will be evident from the thesis, these turned out to play a central part in my PhD. Thanks to current and former colleagues. Especially to Martin Cramer Pedersen, Grethe Vestergaard Jensen, Nicholas Skar-Gislinge, Pie Huda, and Søren Kynde who were PhD students and post docs in the group in my first years in the group and from whom I have learned a lot. Thanks to Jens Bæk Simonsen for many good questions and discussions. I will also say thanks to external collaborators. Special thanks to Bente Vestergaard and Jette Kastrup and their respective group members. Finally, thanks to my friends, to my girlfriend Amalie, and to my parents, for great support.

## Abstract

Proteins perform a wide range of vital physiological tasks in a complex interplay with other biological components, such as signaling molecules, nucleotides and lipids. To better understand the role of the proteins, their structure must be surveyed, as their function and structure are strongly coupled. Advanced experimental techniques are vital to be able to probe such biological nanostructures. A key to better understanding therefore lies in the development of these techniques. Small-angle scattering (SAS) is one of these techniques and is successful at determining the low-resolution structure of proteins and protein complexes in solution. The current thesis deals with some of the recent challenges in biological SAS. One challenge is the investigation of membrane proteins. *In vitro* studies of membrane proteins require a system for solubilization of the proteins, where detergent is the most common. The scattering contribution from the detergents can be suppressed with contrast variation in small-angle neutron scattering (SANS) by use of specially synthesized "invisible" detergents as developed in our research group. These have zero scattering contribution in the full  $q$ -range when measuring in a D<sub>2</sub>O-based buffer. I have developed tools for fully exploring this method. One challenge was to correctly include a layer of densely packed water around the proteins without adding water at the region of the detergents. This and many other features is implemented in the program CaPP, developed during my PhD. CaPP also calculates the theoretical pair distance distribution function,  $p(r)$ , as well as the scattering for protein structures in the protein data bank (PDB) format. I show that the calculations in CaPP are rapid and accurate. Another issue we had to deal with when using the "invisible" detergents was protein aggregation. Aggregation may hinder correct structure determination from the data. We therefore applied and refined a method to take aggregation into account using analytical structure factors. It is essential to be able to assess if one hypothesized model describes data significantly better than others. The F-test was applied and proved useful in that context. Aimed with these new tools and the "invisible" detergents, we studied three different membrane protein complexes: the AMPA-type glutamate receptor 2 (GluA2), the sarco/endoplasmic reticulum calcium ATPase (SERCA), and the holo-translocon (HTL). Both GluA2 and SERCA are key players in neurological diseases, a field that is still poorly understood. GluA2 was investigated in solution in different ligand-induced conformational states. Some of the investigated states had been solved at high resolution, and we verified that these compact forms were also the solution structures. Moreover, we discovered a more open form, resembling that of a previously found electron microscopy structure. SERCA was investigated in a state with unknown structure. Our SANS data provided experimental evidence that SERCA was in an equilibrium state between two known forms. For HTL, it was established, that the protein complex contained a lipid core. Moreover, we provided evidence for flexibility in the SecDF domain of HTL. A fourth protein system, and a key player in neurodegenerative diseases,  $\alpha$ -synuclein ( $\alpha$ SN), was also studied. Under the right conditions,  $\alpha$ SN forms fibrils and we used SANS for dynamic studies of a hypothesized exchange between  $\alpha$ SN monomers in solution and monomers in the fibrils. The SANS data moreover confirmed the existence of a layer of densely packed water around the fibrils, but also showed that it was not more dense or extended than water layer formed around other proteins.

Finally, we developed a statistical tool that utilizes Bayesian statistics to include prior information about the investigated system in analytical modelling. The method was too immature to be applied to any of the scientific cases, but we showed that the method is very promising. The method e.g. automatically determines the most probable value for the regularization parameter that weighs the prior knowledge and new SAS data. The Bayesian method also provides a good measure for the information content in data. In conclusion, the thesis expands the borders of what can be "seen" with SAS by the development of new analytical and statistical tools as exemplified with four challenging scientific cases of biologically relevant protein complexes.



## Abstract in Danish (Resumé på dansk)

Proteiner udfører en lang række vitale fysiologiske opgaver i et kompliceret samspil med andre biologiske enheder såsom signal molekyler, nukleotider og lipider. For at opnå en dybere forståelse for disse komplekse systemer må deres struktur kortlægges, eftersom deres struktur og funktion er nært koblede. Avancerede eksperimentielle teknikker er nødvendige for at kunne undersøge sådanne biologiske nanostrukturer. En nøgle til bedre forståelse ligger derfor udviklingen af disse teknikker. Småvinkelspredning (SAS) er en af disse teknikker og har vist sig at være velegnet til at bestemme den lav-opløste struktur af proteiner og proteinkomplekser i vandig opløsning. Denne afhandling omhandler nogle af de aktuelle udfordringer i biologisk SAS. En af disse udfordringer er undersøgelsen af membranproteiner. *In vitro* studier af membranproteiner kræver et system til at holde proteinerne i opløsning, hvoraf detergenter er det mest almindelige system. Spredningsbidraget fra detergenter kan nedtones med kontrastvariation i småvinkel neutron spredning (SANS) ved brug af specielt syntetiserede "usynlige" detergenter, udviklet i vores forskningsgruppe. Disse har intet spredningsbidrag i hele  $q$  området, hvis det er i  $D_2O$ -baseret buffer. Jeg har udviklet metoder så detergent metoden kan blive fuldt udnyttet. En udfordring var på korrekt i vis, når den teoretiske spredning skulle beregnes, at inkludere et lag af tætpakket vand omkring proteinerne uden at tilføje vand i regionen hvor detergenterne er. Denne og mange andre features er implementeret i programmet CaPP, udviklet under min ph.d. CaPP udregner også den teoretiske par-afstandsfordeling,  $p(r)$ , samt spredningen for proteinstrukturer i protein data bank (PDB) formatet. Jeg viser at CaPP udregner disse hurtigt og præcist. Et andet problem vi måtte håndtere ved brug af de "usynlige" detergenter var protein aggregering. Aggregering forhindrer korrekt bestemmelse af proteinstrukturen. Vi anvendte og forfinede derfor en metode til at tage aggregering med i beregningerne ved brug af analytiske strukturfaktorer. Det er essentielt at være i stand til at vurdere om en foreslået model beskriver data væsentligt bedre end andre. F-testen blev anvendt og fundet nyttig i denne sammenhæng.

Bevæbnet med disse nye redskaber og de "usynlige" detergenter studerede vi tre forskellige membranproteinkomplekser: AMPA-type glutamat receptor 2 (GluA2), sarcoplasmisk retikulum kalcium ATPas (SERCA1a) og holo-tranlokatoren (HTL). Både GluA2 og SERCA spiller en væsentlig rolle i neurologiske sygdomme. Nogle af de undersøgte strukturelle tilstande var blevet løst til høj opløsning, og vi bekræftede at disse kompakte tilstande også var de strukturelle tilstande i opløsning. Derudover opdagede vi en mere åben tilstand, der lignede en struktur fra en tidligere elektronmikroskopi-undersøgelse. SERCA blev undersøgt i en strukturelt set ukendt tilstand. Vores SANS studier sandsynliggjorde at SERCA var i en ligevægtstilstand mellem to kendte strukturelle former. For HTL blev det fastslået at protein komplekset indeholdt en lipid kerne. Derudover sandsynliggjorde vi, at der er fleksibilitet i SecDF domænet af HTL. Et fjerde protein, og en væsentlig spiller i neurodegenerative sygdomme,  $\alpha$ -synuclein ( $\alpha$ SN), blev også studeret. Vi brugte SANS til dynamiske undersøgelser af en foreslået udveksling af monomerer mellem  $\alpha$ SN fibriller og monomerer i opløsning. Derudover bekræftede SANS data eksistensen af et fotættet vandlag omkring  $\alpha$ SN fibrillerne, men viste samtidig at laget ikke var tættere eller mere udstrakt end vandlaget omkring andre proteiner.

Endelig udviklede vi et statistisk værktøj som udnytter Bayesiansk statistik til at inkludere forhåndsviden omkring det studerede system ved analytisk modelleringen. Metoden var ikke moden til at blive brugt på de videnskabelige cases her i afhandlingen, men metoden er meget lovende. Fx giver metoden en automatisk måde til at bestemme den mest sandsynlige værdi af regulariseringsparameteren, som vægter forhåndsviden mod nyt SAS data. Den Bayesianske metode giver også et godt mål for informationsindholdet i data.

For at sammenfatte, så udvider afhandlingen grænserne for hvad der kan "ses" med SAS ved at udvikle nye analytiske og statistiske metoder, som det er eksemplificeret ved fire biologisk relevante protein komplekser.

## Preface: second part of a 4 years PhD program

This PhD thesis is the second part of a 4 years combined master's and PhD program at the Niels Bohr Institute (NBI), University of Copenhagen (UCPH). I have been in the structural biophysics group at the section for Neutron and X-ray Science, with Lise Arleth as my supervisor. The first part was handed in August 2016 and formally constituted my master's thesis (Larsen 2016).

The overall theme in my PhD has been to explore and refine methods to retrieve structural information about proteins and other biological macromolecules using small-angle X-ray and neutron scattering (SAXS and SANS). I have therefore had the chance to investigate a range of different systems in collaboration with many different groups and people. These will be mentioned as they appear in the thesis. My background is in physics and the emphasis is thus on modelling, statistics, and method development.

Since this work is a continuation of the studies reported in my master's thesis, I will briefly describe its content. The master's thesis contained three major parts. The first was a structural study of nanodiscs (Fig. 2 in Paper II). Nanodiscs are cell membrane mimicking particles composed of a small patch of lipid bilayer with a diameter of about 10 nm. The lipids are surrounded by two  $\alpha$ -helical proteins, so-called "belt proteins". Membrane proteins need to be in membranes, or in a membrane mimicking systems to be stable and active. Nanodiscs are therefore used for *in vitro* structural and functional studies of membrane proteins. I investigated a new type of nanodisc with peptides forming the belt, so-called "beltides". We studied the formation and structure of these beltide nanodiscs with SAXS and SANS combined with coarse-grained molecular dynamics computer simulations, and several complementary experimental techniques. The study was published in Soft Matter (Larsen *et al.* 2016). The second part of the thesis was a structural study of another type of nanodiscs, with apolipoprotein E (ApoE) as belt protein. ApoE is one of the major constituents in high density lipoprotein (HDL) in the central nervous system and the cardiovascular system, and is therefore physiological relevant. This study is still in progress and now includes high-quality SAXS and electron microscopy data. We hope to submit a paper soon (not included here). In the third part, I outlined how Bayesian statistics can be used in the analysis of SAS. I have continued this work quite extensively, and most of the results are reported in Paper II. Furthermore, the work during the first part of my PhD lead to a more general investigation of the statistical tools used in the analysis of SAS data. These statistical considerations is part of the present thesis.

The work has become rather extensive. The first about 100 pages are made for this thesis alone, and the next 160 pages are papers or paper drafts. I will give ongoing recommendation on when to read what, and I will encourage the reader by telling that the thesis only deals with highly interesting topics, which I have sincerely enjoyed spending four years working with.

**Andreas Haahr Larsen,**  
**University of Copenhagen,**  
**September 2018**

# Contents

<b>1</b>	<b>Proteins and Small-Angle Scattering</b>	<b>9</b>
1.1	Small-angle scattering and other complementary techniques in structural biology . . . . .	10
1.2	Historical evolution of small-angle scattering in structural biology and the struggle to obtain real-space information . . . . .	13
1.3	Overview of the thesis . . . . .	15
<b>2</b>	<b>The Basics of Small-Angle Scattering</b>	<b>17</b>
2.1	The scattering from point scatterers in vacuum . . . . .	17
2.2	Scattering from molecules in solution . . . . .	20
2.3	The scattering length weighted histogram of distances . . . . .	22
2.3.1	Pseudo Fourier mates . . . . .	22
2.3.2	Indirect Fourier transformation . . . . .	23
2.3.3	The relation between $p(\mathbf{r})$ and $h(\mathbf{r})$ . . . . .	23
2.3.4	What quantity is measured? . . . . .	24
2.3.5	The scattering invariant $Q$ and its relation to the molecular volume . . . . .	25
2.4	Model-free determination of the oligomeric state . . . . .	25
2.5	Incoherent scattering . . . . .	28
2.6	A few closing remarks . . . . .	29
<b>3</b>	<b>Modelling Tools for Membrane Proteins</b>	<b>31</b>
3.1	The hydrophobic effect and water layer around proteins . . . . .	31
3.2	Computer program CaPP . . . . .	32
3.2.1	The architecture of the program . . . . .	32
3.2.2	Features of CaPP . . . . .	32
3.2.3	Computational speed . . . . .	36
3.2.4	Approximations utilized to calculate the intensity in CaPP . . . . .	37
3.2.5	Scientific impact . . . . .	37
3.3	Analytical treatment of partly aggregated samples . . . . .	38
3.3.1	" <i>In silico</i> sample purification" . . . . .	38
3.4	New tools, new possibilities . . . . .	39
<b>4</b>	<b>Statistical Methods for the Analysis of Small-Angle Scattering Data</b>	<b>43</b>
4.1	Goodness of fit and generalizability . . . . .	43
4.2	The aim of the current chapter . . . . .	44
4.3	The frequentist versus the Bayesian approach . . . . .	44
4.4	C1: Evaluating a model . . . . .	45
4.4.1	Evaluating a model using $\chi^2$ statistics . . . . .	45

4.4.2	Error bar independent evaluation of the goodness of fit . . . . .	47
4.4.3	Evaluating a model with Bayesian statistics . . . . .	49
4.4.4	Predictive/cross-validating methods. . . . .	49
4.5	C2: Comparing alternative models . . . . .	49
4.6	C3: Correcting wrongly estimated error bars . . . . .	51
4.7	C4: Combining data with prior knowledge . . . . .	54
4.8	C5: The number of degrees of freedom and the information content in data . . . . .	56
4.9	Significant achievements in this chapter . . . . .	61
<b>5</b>	<b>Protein Complexes Studies with SANS Contrast Variation</b>	<b>63</b>
5.1	Structural investigation of membrane proteins in detergents with SAXS and SANS . . . . .	64
5.1.1	Elimination of free micelle scattering contribution . . . . .	64
5.1.2	SANS contrast variation with deuterated detergents . . . . .	65
5.1.3	Three protein complexes studied with novel "invisible" detergents and SANS contrast variation . . . . .	68
5.2	Studying $\alpha$ -synuclein structure and dynamics with SANS contrast variation . . . . .	72
5.3	New insight into challenging protein complexes made possible by new tools . . . . .	77
<b>6</b>	<b>Conclusion and Final Remarks</b>	<b>79</b>
<b>7</b>	<b>References</b>	<b>81</b>
<b>8</b>	<b>Appendices</b>	<b>87</b>
8.1	Appendix A: Scattering intensity in the continuous limit and the form factor . . . . .	87
8.2	Appendix B: Experimental report from the study of $\alpha$ -synuclein . . . . .	89
<b>9</b>	<b>Publications</b>	<b>101</b>
9.1	Paper I: Single-particle structure refinement from small-angle scattering data of partially aggregated protein samples . . . . .	102
9.2	Paper II: Analysis of small-angle scattering data using model fitting and Bayesian regularization	138
9.3	Paper III: Invisible detergents for structure determination of membrane proteins by small-angle neutron scattering . . . . .	150
9.4	Paper IV: Small-angle neutron scattering studies on the AMPA receptor GluA2 in the resting, AMPA and GYKI-53655 bound states . . . . .	177
9.5	Paper V: Structure, dynamics and function of a lipid pool at the centre of the bacterial holo-translocon . . . . .	227

# Chapter 1

## Proteins and Small-Angle Scattering

*"Trying to determine the structure of a protein by UV spectroscopy  
was like trying to determine the structure of a piano by listening  
to the sound it made while being dropped down a flight of stairs."  
- Francis Crick*

One of the fundamental questions in life is the understanding of ourselves. What are humans made of and how do we talk, think, and act? Part of the answer can be found at the nanoscale, where biological molecules interact to drive all living organisms. We have reached far in that field and now have a good understanding of genetics, metabolism, the immune system, the nervous system, and many other vital biological systems. We know that proteins are key players in all of these processes and have unveiled the structure of thousands of proteins to atomic resolution. Much of this success stems from development of experimental techniques to investigate biological matter from the macro level and down to the nanoscale (Fig. 1.1). New techniques can lead to major leaps in understanding and so can development and refinement of existing techniques. Much focus recent years has been on combining several biophysical techniques (integrated structural biology) to push the limit for what we can "see" even further. The current thesis is a 260-pages (sorry!) long attempt to explore and expand the limits of one of these techniques, namely small-angle scattering (SAS), and to discuss how SAS can be combined with complementary experimental data. By pushing the limits of SAS and other experimental techniques, and by combining them, we hope to be able to cast light on some of the biological systems that are still poorly understood. Some of these systems are proteins involved in neurological disorders, and part of the thesis deals with such systems.

**Correlation between protein structure and function** Proteins are characterized by a unique sequence of amino acids, the so-called primary structure of the protein. A mutation of a single amino acid in the sequence can perturb the function of the protein. This is due to the close correlation between protein primary structure and its function. The function of the proteins is also strongly related to the 3-dimensional fold of the amino acid chain, i.e. to the secondary and tertiary protein structure. Finally, proteins can form multi-chain complexes and this quaternary structure is also closely connected to the functionality of these fascinating biological complexes. Over time, a range of techniques to investigate protein structure have been developed, which makes life easier than in the early days of Francis Crick (quotation), one of the co-discoverers of the structure of DNA. This discovery was possible only due to the invention of X-ray crystallography.

## 1.1 Small-angle scattering and other complementary techniques in structural biology

Along with many other techniques in structural biology (Fig. 1.1), small-angle scattering (SAS) has aided the understanding of protein structures. SAS is suitable for studying structural features ranging from a few nm to about 100 nm. Thus, SAS can give low-resolution structural information about proteins, protein complexes, and other biological macromolecules. It is important to know the limitations of the technique, and these are defined partly by other available techniques in structural biology, as many structural questions are better answered by other techniques than SAS. Therefore, I will in the following present some of the closest related experimental techniques, and discuss how they complement, and are complemented by, SAS.

**X-ray and neutron diffraction.** If a protein crystal can be obtained, X-ray diffraction can be used to solve the protein structure to near-atomic ( $> 5 \text{ \AA}$ ), atomic ( $1\text{-}2 \text{ \AA}$ ), or even sub-atomic ( $<1 \text{ \AA}$ ) resolution (Blakeley *et al.* 2015). A hydrogen atom has a diameter of about  $1 \text{ \AA}$ , so  $\text{\AA}$  is a good unit for describing high-resolution protein structures. It is impossible to obtain near-atomic resolution with SAS, where the typical resolution is about  $10 \text{ \AA}$  or lower. Therefore SAS is often denoted a low-resolution technique. 3D structural models can be determined "directly" from SAS data. These so-called *ab initio* models typically have a resolution of  $20\text{-}30 \text{ \AA}$  (Tuukkanen *et al.* 2016). An *ab initio* model envelope overlaid on a high-resolution structure is shown in Fig. 3.6 on page 42. With high quality X-ray diffraction data, it is possible to obtain an electron density map representing the protein, with so high resolutions that the single amino acids, which are known from the primary structure, can be fitted into the density. Thus, a model can be obtained that shows the position of every atom in the protein. The higher resolution, the better is the certainty of the atom positions. X-ray crystallography has therefore led to major leaps in the understanding of protein and RNA/DNA structure since the first protein structure was solved 60 years ago (Kendrew *et al.* 1958). Automatization of structure refinement from crystallographic data has allowed non-experts to use the technique routinely, and thousands of protein structures have been solved by X-ray crystallization. These structures are deposited in the protein data bank (PDB). The PDB also contains structures solved with other techniques (see below), but the dominant technique is, without comparison, X-ray crystallography. Neutron crystallography can be used complementarity as this technique is sensitive to hydrogens in the structure, relevant e.g. for ligand docking or enzymatic protonation. Crystal structures obtained by X-ray or neutron crystallography may however not represent the native structure, and it may not be unique but represent only the structure that best crystallizes. Moreover, many proteins do not crystallize, in particular proteins with disordered domains as well as membrane proteins. Therefore, SAS is a good complementary technique to diffraction, as crystal structures can be verified, and proteins that do not crystallize can be investigated. Recently, several free electron laser (XFEL) facilities have opened (Doerr 2018). An XFEL uses a coherent X-ray beam and thus exploits positive interference to obtain flux that are several orders of magnitude higher than the highest flux at any synchrotron. This coherent and bright beam brings about interesting new possibilities within X-ray crystallography. With XFELs it is possible to obtain near-atomic resolution structures with smaller crystals, nanocrystals, that do not diffract well enough for a "conventional" synchrotron X-ray experiment. A synchrotron is a super advanced X-ray source, hence the quotation marks around "conventional". Moreover, the experiments can be done at ambient temperatures (Higgins & Lea, 2017) (before the sample is hit by the beam and destroyed). This technique will be interesting to follow, but is still rather immature.

**Electron microscopy.** With recent development in electron microscopy (EM) hardware and software it is now possible to solve protein structures with near-atomic resolution using cryo EM. That is, the

sample is frozen in a thin layer on a carbon grid, and then studied by irradiating it with an electron beam. The improvement of EM has been denoted the "EM revolution" as the number of near-atomic structures deposited in the electron microscopy data bank (EMDB) has grown rapidly (Egelman 2016). At least one structure has been solved to a resolution below 2 Å (PDB: 5K12; Merk *et al.* 2016) and several below 3 Å (Wlodawer *et al.* 2017). The vast majority of atomic protein structures are however still solved by X-ray crystallography (Higgins & Lea 2017) and EM is less suited for small molecules (e.g. proteins below about 100 kDa), so there is no doubt that both techniques will co-exist as complementary techniques (see also the review by Venien-Bryan *et al.* 2017). In negative stain EM, the molecules are fixed to a carbon grid by a staining agent, e.g. uranyl, instead of by cryogenics. Negative stain EM constitutes a cheaper (in instrumental price and processing power) and faster alternative to cryo EM, due to the high contrast provided by the electron-dense staining agent. The best resolution obtained with negative stain EM is about 20 Å, and it is therefore a low-resolution technique. Interestingly in this context, the resolution limit for negative stain EM is comparable to that of SAXS. The data are however easier to obtain, the samples need less purification, and are easier to interpret as data consist of real space images. With recent development in software, the data processing is also straight-forward. So it is relevant to question the role of SAS in the presence of negative stain EM. However, there are a range of areas where SAS and EM differs: First, the staining may affect the system by a pH change (Bradley 1962), and the stain also fixate the protein to a grid, meaning that the probed structure is not the solution structure, as it is in SAS. Note that in cryo EM there is no staining, and it is widely believed that the plunge freezing (very rapid freezing) has little or no effect on the protein structure. Second, EM struggles when the proteins are small, below  $\sim 100$  kDa (Merk *et al.* 2016), where SAS is still usable. Third, SAS is more sensitive to difference in contrast e.g. in lipid-protein complexes. In SAXS it is directly related to the electron density, see chapter 2 for a general introduction to contrast in SAS. In negative stain EM, the contrast comes mainly from the stain and the difference in contrast in the sample is therefore hardly measurable. On top of that, contrast variation is possible in SAS (see chapter 5), giving a range of possibilities. A final and important difference between EM and SAS is that SAS gives the average signal from the whole sample, whereas EM gives thousands of images, each with one particle, whereof only a subset is included in the final model(s). In EM processing, a selection process is made, partly automatic and partly manual (see e.g. Fernandez-Leiro & Scheres 2016), where e.g. aggregates, damaged particles and lumps of stain can be sorted out. Therefore, EM is less sensitive to aggregation. The particles can then be sorted into different 3D classes that represent unique conformational states. Thus, a range of states can in principle be found from a single dataset. This is a clear strength for EM, but also adds bias to the final model, as the selection process is partly manual. One advantage of SAS over to EM is therefore that the data is less biased. Another clear advantage is that models can be fitted directly to the full SAS data set using relatively few assumptions. Such direct hypothesis testing can not be done as easily, unbiased, and direct in EM, as a theoretical EM dataset can not be generated from a model and compared directly to the data. This last point implies that EM is very good at answering the open question "what do we have in this sample?" SAS, on the other hand, is better at answering the specific question "is the sample in this or that conformational state?".

**Nuclear magnetic resonance.** A third technique from which protein structures can be determined at atomic resolution is nuclear magnetic resonance (NMR) spectroscopy. NMR measures spin relaxation. The relaxation of an atomic spin depends on the local environment of that atom, meaning that an NMR signal contains local structural information, which can be used for structure determination. An atomic model is refined by the NMR data by relaxing its structure using constraints about local inter-atomic distances. The NMR constraints are supplemented by constraints about e.g. the relative positions of atoms that do not give any NMR signal. Solution NMR can solve protein structures to atomic resolution, but rely on

orientational averaging of the particles (tumbling). Due to reduced tumbling rate for large proteins, NMR is limited to relatively small proteins. Large proteins also lead to relaxation frequency degeneracies in the NMR spectrum. Solid state NMR provides an alternative for large proteins, and proteins that are fixed and can not tumble. This is e.g. the case for membrane proteins in large lipid bilayers. With solid state NMR, the whole sample is spun using so-called magic angle spinning to obtain orientation averaging and sharp peaks in the NMR spectrum (Opella 2015). Solid state NMR can also be used to study protein subunits fixed inside large protein complex structures, such as protein fibrils (Tuttle *et al.* 2016) as we shall return to in chapter 5. The degeneracy issue is however not solved by solid state NMR, but can to some extent be overcome by atom labeling. I.e. by controlling which part of the protein that is "visible" by NMR. An overview of the number of deposited NMR protein structures in the PDB (Frueh *et al.* 2013) shows that it gets increasingly difficult to resolve the structure when the size exceeds about 20 kDa and very few monomeric protein structures above 30 kDa have been solved. The complementarity between SAS and NMR is clear from Fig. 1.1, as the size range they can probe are about 1 Å to 10 nm for NMR, and a few nm to  $\sim 100$  nm for SAS, so they overlap and together they cover about three orders of magnitude. Moreover, SAS is far easier to process and interpret than NMR data.

**Molecular dynamics simulations.** The last technique I have included in the list is the only non-experimental technique. Increasing computer power have made it possible to do computer simulations for large proteins and over relatively large timescales (ms at most, but this is large timescales when compared to what was possible before). In particular molecular dynamics (MD) simulations is used extensively to investigate structural and dynamical properties of proteins. In the simulations, the forces between the atoms are described, and the structure is then "relaxed" by numerical integration to find the lowest energy state. Decades of benchmarking of these forcefields against data have ensured that the results are trustworthy and can be used to interpret and predict experimental findings. Due to limits in computational power, the size range is limited to single proteins and small complexes. However, with coarse-graining (CG), i.e. merging of several atoms into larger beads, larger systems can be investigated at longer time-scales. We used CG MD in the first part of my PhD to study the self-assembly of peptide-lipid particles (Larsen *et al.* 2016). Also, our collaborators did CG MD in Paper V to study the formation of a lipid core inside a protein complex. The price of coarse graining is reduced accuracy. Clearly, computer simulations are complementary and not stand-alone as the forcefields have to be constantly adjusted to agree with experimental results. But the simulations provide detailed interpretations of experimental results, as well as hypotheses that can be investigated experimentally. Moreover, MD simulations can probe timescales that are impossible or at least extremely difficult to access experimentally.

**Examples of complementarity** MD forcefields are used in the refinement of NMR structures, such that the refined structure is constrained simultaneously from the MD forcefield and NMR. MD has also been included in calculation of accurate SAXS patterns with explicit solvent, e.g. in the program WAXSiS (Knight & Hub 2015; Hub 2018). Another great achievement within complementarity is that SAXS has become a standard and streamlined complementary check for crystallographic data (Trewella *et al.* 2017). Also on the instrumental site, complementarity is acknowledged. Recently, the European synchrotron radiation facility (ESRF) has invested in high-performance cryo EM instruments. As the neutron facility Institut Laue-Langevin (ILL) with a high flux neutron reactor, as well as the Institut de Biologie Structurale (IBS) with high-performance NMR equipment are both located within 1 km from ESRF, users have access to both EM, SAXS, SANS, neutron and X-ray crystallography as well as NMR at the same location. Firstly, that is very convenient for the users (with the right passports, giving access to all the facilities), and secondly, such initiatives may very likely be necessary to push the limits of our current knowledge and



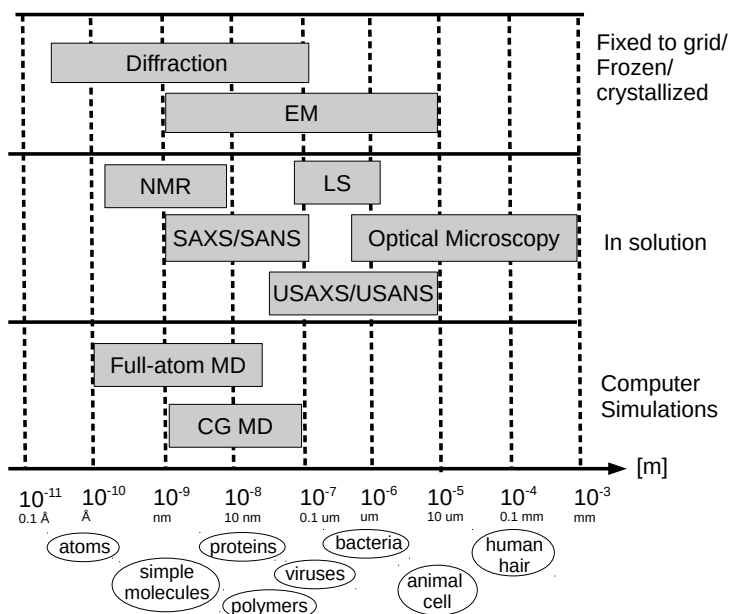


Figure 1.1: Different techniques in structural biology cover several orders of magnitude, from sub-atomic resolution, up to particles with macroscopic sizes. From top: Diffraction (including protein crystallography), electron microscopy (EM), including cryo EM and negative stain EM, nuclear magnetic resonance (NMR) including solution NMR and solid state NMR, light scattering (LS), small-angle X-ray and neutron scattering (SAXS/SANS) and ultra SAXS/SANS (USAXS/USANS) as well as optical microcopy. The overview also includes computer simulation techniques: full-atom molecular dynamics (MD) simulations, and coarse-grained (GC) MD.

to be able to understand even more challenging and complex systems.

**Techniques not included.** The list is incomplete as a range of other techniques exists, that can be used to probe overall structural features. These include, e.g., dynamic light scattering (DLS), spin-echo SANS (seSANS), atomic force microscopy (ATM), and mass spectroscopy (MS). Nor does the list include infrared (IR) spectroscopy and circular dichroism (CD) that can reveal the secondary structure of a protein. The list does however cover the most widely used techniques for solving the overall structure of proteins and protein complexes.

## 1.2 Historical evolution of small-angle scattering in structural biology and the struggle to obtain real-space information

In this section we dwell a second in the history, before we continue with the work done in the current thesis. I will outline historical development of SAS including recent development of biological SAS. This is all to provide the historical context for the current work.

**Early development.** In the late 1800s Thomson discovered a particle with a particular large charge with respect to its mass. Thomson's particle was the electron, and during the coming decade he described how incoming electromagnetic waves could interact with free electrons, and make them oscillate such that new electromagnetic waves were radiated. The reradiated waves had the same frequency as the

incoming. This elastic Thomson scattering is still the fundamental concept that most scattering theory and experiments build upon. The discovery of X-rays by Röntgen in 1895 and subsequent theoretical breakthroughs made it possible to study crystals at the atomic level. The theoretical breakthroughs included Laue's fundamental discovery from 1910 of X-ray diffraction from crystals. Soon after Bragg (father and son) formulated Bragg's law. These concepts were used in the following years to solve the structure of a range of salt compounds (Bragg 1913). In the beginning of the 20<sup>th</sup> century, quantum mechanics evolved with the notation of light as discrete energy packages, behaving like particles, as proposed by Einstein in 1905. Whether light were particles was debated until Compton described how photons could "bump" into electrons and deposit part of their energy. This Compton scattering was incompatible with the classical wave picture of light, which was therefore replaced by the particle-wave duality concept. The neutron was discovered in 1932 by Chadwick, and it was readily realized, that the particle-wave duality made it possible to treat the neutrons as waves. This is the basis for scattering experiments with neutrons.

**SAS theory emerges in the fifties** A series of theoretical achievements within development of SAS theory happened during the fifties. They allowed information to be retrieved from scattering patterns without any Bragg-peaks. Instead, the scattering at small scattering angles was analyzed. Kratky and Porod established the Kratky plot, which is used to determine if a protein is folded (Kratky & Porod 1949) and described the 4th order decay at the large scattering angles. They also described the scattering invariant,  $Q$ , used to determine what is nowadays called the Porod volume (Porod 1951), and which is used to determine the molecular weight of proteins from SAS data. Guinier formulated what is today called the Guinier approximation to determine the radius of gyration,  $R_g$ , of a particle from the data at the smallest scattering angles (Guinier & Fournet 1955), and Debye formulated his famous equation (chapter 2, [eqn. 2.6]) used to describe scattering from an ensemble of point scatterers (Debye & Brumberger 1957).

**Getting information in real-space.** SAS data is not easily interpreted, as the scattering gives information in inverse space. That is, intensity as function of the scattering vector (see chapter 2), which has units of 1/length. The intensity can therefore not directly be understood in terms of real space coordinates, and due to loss of phase-information, orientational averaging of the investigated molecules, and the noise of data as well as the limited measured  $q$ -range, data cannot be transformed directly to give the real space 3D structure.

A few real-space parameters can be gained directly from the data. They include the  $R_g$  as can be obtained by the Guinier approximation. Also, the volume and molecular weight can be obtained, as discussed in section 2.3.5. Besides that, there are two strategies to gain real-space information about the data. The first is to transform data into real space by indirect Fourier transformation. This was introduced in SAS by Glatter (1977). Glatter showed how IFT could be used to obtain the pair distance distribution function,  $p(r)$ , which is a real-space 1D representation of the scattering data (derived in chapter 2). Glatter also described how the  $p(r)$  could be used to get an intuitive idea about the shape of the investigated particle (Glatter 1977). See also Fig. 2.5 on page 2.5. The other approach describes a model in real space, e.g. a geometrical model, and Fourier transform this model in the inverse space ("scattering" space) to obtain the so-called form factor. This is shown for a sphere in Appendix A. The model parameters can then be refined by a fit to data. Many form factors have been derived (Pedersen 1997), and a range of programs have been developed for the fitting.

A major leap in the analysis of proteins with SAS was made by Svergun and co-workers at the EMBL Hamburg group as they developed the ATSAS program suite. It started with an implementation of IFT in the program GNOM (Svergun 1992). Later came CRY SOL (Svergun *et al.* 1995) that allowed a direct comparison of crystal structures deposited in the PDB with SAS data, by calculating the form factors for

the proteins. A range of different programs have later been added to the ATSAS software package, and the support and accessibility have dramatically extended the use of SAS for structural biology by non-experts. In 1996, the program SASHA made it possible to obtain an *ab initio* envelope structure of a protein (Svergun *et al.* 1996). That is, a real space 3D model of the investigated protein could be obtained from the data, with no input from the user. In other words, it could be obtained *emphab initio*. The approach was later improved by to the widely used DAMMIN (Svergun 1999) and DAMMIF (Franke & Svergun 2009) bead modelling tools. It is an approach of the second type, in that a real-space bead model is constructed, the form factor calculated and fitted to data. The parameters are the coordinates of the beads and/or the scattering length of each bead. As no input or prior information is needed from the user, the programs are very popular. The methods appears to be a transformation of the 1D scattering curve to the 3D real space structure. That is, IFT taken to the next level. This "transformation" is however highly underdetermined, as I will come back to in chapter 4.

**Future of SAS.** What is the future role of SAS in structural biology? With the recent development in EM, low-resolution *ab initio* models can easily be obtained with negative stain EM. A fully monodisperse sample is not even needed, as oligomers and aggregates can be sorted out. So SAS *ab initio* modelling might decrease in popularity in the coming decade due to the EM alternative. As described under each technique, SAS is however a strong complementary technique for static structure analysis. Another interesting area is time-resolved SAS studies, where timescales unreachable for EM and crystallography can be probed by time-resolved SAXS (trSAXS) to study e.g. unfolding or conformational change *in situ*. The combination of trSAXS and MD simulations is a powerful tool for dynamic studies of proteins. SAS is also a strong technique for studying oligomerization and aggregation processes. In Paper I we review a range of models that can be used in that context. These are inverse-space descriptions of aggregates, called structure factors. Another unique role for SAS is in the structural studies of intrinsically disordered proteins (IDPs), which have a, more or less, random walk-like structure. IDPs do not crystallize, and average out in standard EM image processing, but can be probed with SAS. As we shall see in the current thesis, SANS contrast variation also provides a range of unique possibilities. This is e.g. shown in chapter 5, where SANS is used to study the exchange of monomeric proteins in solution with the monomers in protein fibrils of  $\alpha$ -synuclein. Another example is in Paper V, where SANS contrast variation is used to highlight a lipid core in a protein complex. So I believe the future for SAS is bright.

## 1.3 Overview of the thesis

Chapter 2 is an introduction to the theory of SAS. Some core concepts are introduced and will be referred to in the following chapters.

Chapter 3 presents some of the analytical tools I have developed for the analysis of SAS data, in particular from samples of membrane protein complexes. Most of the tools were implemented in the program CaPP. Chapter 3 also includes a description of how analytical structure factors, as reviewed in Paper I, can be used to analyze data of aggregated samples.

Chapter 4 is also of methodological nature, as it contains a thorough discussion of the statistical tools most commonly used in the analysis of small-angle scattering data. Furthermore, it contains a presentation of newly developed tools based on Bayesian statistics, which is described more thoroughly in Paper II.

After the methodological chapters, the reader will in Chapter 5 see some of the methods and the theory applied to actual scientific cases. The chapter will however first introduce yet another method, namely SANS contrast variation with specially synthesized "invisible" detergents. Studies of three different membrane protein complexes are presented, whereof two play central roles in neurological disorders. These proteins

were investigated for conformational changes upon ligand binding. The chapter also includes a study of a fibrillating proteins system,  $\alpha$ -synuclein ( $\alpha$ SN). Among other things, it was investigated whether the hydration shell of densely packed water molecules around the protein fibrils was different from the hydration layer found around other proteins.

The thesis also contains two appendices. Appendix A is a supplement to the theory section. Appendix B is an experimental report from the  $\alpha$ -synuclein study. I wrote it as a report in hope that this will ease the further progression of the project. I will refer to the report in the relevant section (last part of chapter 5). The last part of the thesis consists of five papers. Three of these are published or accepted for publication (Paper II, III, and IV) and the remaining two are drafts to be published. I will try to guide the reader with recommendations on when to read the papers.

## Chapter 2

# The Basics of Small-Angle Scattering

*"Any fool can know. The point is to understand."*

*- Albert Einstein*

In this section I will give a brief introduction to the basic theoretical concepts of small-angle scattering (SAS). I will *not* give any overall introduction to the applications of SAS or overview of a typical experiment or the like. Therefore, for readers not familiar with SAS, I highly recommend spending four minutes watching the very nice introductory video by WeNMR (search for "WeNMR Small Angle X-ray Scattering Animation"; <https://www.youtube.com/watch?v=2HS4SOdxbS8>). For the reader familiar with SAS, I will still recommend watching it, as it visualizes fibril formation which I have studied (chapter 5 and Appendix B). It also illustrates *ab initio* bead modelling, which I have already mentioned in the introduction, and will discuss in context of statistical information content, in chapter 4. Also, the video has some very nice animations of the PetraII X-ray storage ring at the Deutsches Elektronen-Synchrotron (DESY) and the sample robot at the P12 bioSAXS beamline, where I did a three month internship as part of the PhD.

I will in the following introduce the pair distance distribution function,  $p(r)$ , which I have used extensively in the program CaPP (section 3). Also, the Porod analysis, which is used in Paper IV to assess the oligomeric state of the investigated proteins, will be introduced. I will shortly introduce incoherent scattering in SANS, as it will be referred to when discussing the design of the SANS studies in chapter 5. I have done some unconventional choices in the derivations. In the first part, I have limited the use of integrals and instead used summation, as they can be directly implemented in numerical computing. I have also avoided vectors in the introduction of  $q$ , as I only work with isotropic scattering from randomly oriented samples. Furthermore, I have introduced  $h(r)$ , which is similar to, but, as we shall see, not identical to  $p(r)$ .

### 2.1 The scattering from point scatterers in vacuum

In a scattering experiment, a beam of neutrons or high-energy photons hits a sample (Fig. 2.1). The photons will interact with the electrons in the sample, and the neutrons with the nuclei. These are collectively called scatterers. The key interaction in SAS is the elastic scattering event, where the incoming neutron/photon changes direction, or is reradiated, with conserved energy (Thomson scattering). The events can be described with wave equations. The intensity of an incoming plane wave is given in terms of the wave amplitude  $A$  and the complex phase  $i\phi$ :

$$I_{\text{inc}} = |A \exp(-i\phi)|^2 = A \exp(-i\phi) \cdot A \exp(i\phi) = A^2. \quad (2.1)$$

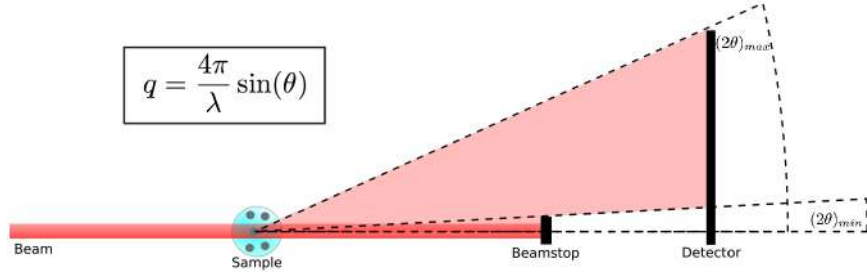


Figure 2.1: The measured range of scattering angle,  $2\theta$ , is limited by the dimensions of the detector and the distance from sample to detector. The lowest measured angle depends on the size of the so-called beamstop. The beamstop hinders that the direct beam damages the detector.  $q$  is the magnitude of the scattering angle as derived in the current chapter, and  $\lambda$  is the wavelength of the beam. Typically,  $q_{\max} \sim 0.5$ , so the largest measured angles are  $\sim 2^\circ$  in SAXS ( $\lambda \sim 1 \text{ \AA}$ ) and  $\sim 10^\circ$  in SANS ( $\lambda \sim 5 \text{ \AA}$ ). Figure from my master thesis (Larsen 2016).

As the intensity is given as a conjugate product, the phase information is lost when measuring. Due to this loss of phase, shape reconstruction of structures from scattering data will inevitably be an ill-posed problem, where several structures can describe the same measured scattering signal. Other factors add to the loss of information, including that the molecules in the sample are randomly oriented, so the measured signal is the orientational average. It is the same in all directions, and thus the information is 1-dimensional. Moreover, the intensity is measured only in a limited range of angles (Fig. 2.1). Finally, the sparse data that we are left with may be noisy due to low statistics or poor signal-to-noise ratio. The thesis includes plenty examples of such noisy 1-dimensional SAS intensity curves.

When encountering a scatterer, part of the wave will be scattered with a certain probability. This probability is given by the scattering length  $b_i$  of the atom. At small angles, the wave can be assumed to scatter equally in all (measured) directions (Als-Nielsen & McMorrow 2011):

$$I_{\text{scat}} = |b \cdot A \exp(-i\phi)|^2 = |bA|^2 = \sigma A^2, \quad (2.2)$$

where  $\sigma$  is the scattering cross section. The scattered intensity is directly proportional to  $\sigma$ , giving rise to a quite intuitive analogue. The probability that an archer hits a target is about proportional to the cross section of the target (at least if the archer is not very skilled). Therefore, it is stupid to ever let an archer shoot toward an apple on your head, as the cross section of the head is much larger than the that of the apple (Fig. 2.2).

Having more scatterers gives a far more interesting situation due to interference between the scattered waves. The measured intensity from  $N_s$  scatterers can be calculated as the sum of all the scattered waves. I will omit  $A$  in the following and also divide by the irradiated sample volume  $V$  to obtain the normalized intensity<sup>1</sup>:

$$\begin{aligned} I_{\text{scat}} &= \frac{1}{V} \left| \sum_{j=1}^{N_s} b_j \cdot \exp(-i\phi_j) \right|^2 \\ &= \frac{1}{V} \sum_{j,k=1}^{N_s} b_j b_k \cdot \exp(-i\Delta\phi_{jk}), \end{aligned} \quad (2.3)$$

<sup>1</sup>The normalized scattered intensity is more precisely denoted "the differential scattering cross section per unit volume" or  $d\Sigma/d\Omega(q)$ . I will, however just use "intensity".



Figure 2.2: A very lucky kid. Photograph by © Mike Mols, shutterstock.

where  $\Delta\phi_{jk} = \phi_j - \phi_k$  is the phase difference. This difference depends on the extra distance,  $d$ , traveled by a wave scattered by the scatterer  $P_k$ , compared to a wave scattered by  $P_j$  (Fig. 2.3).

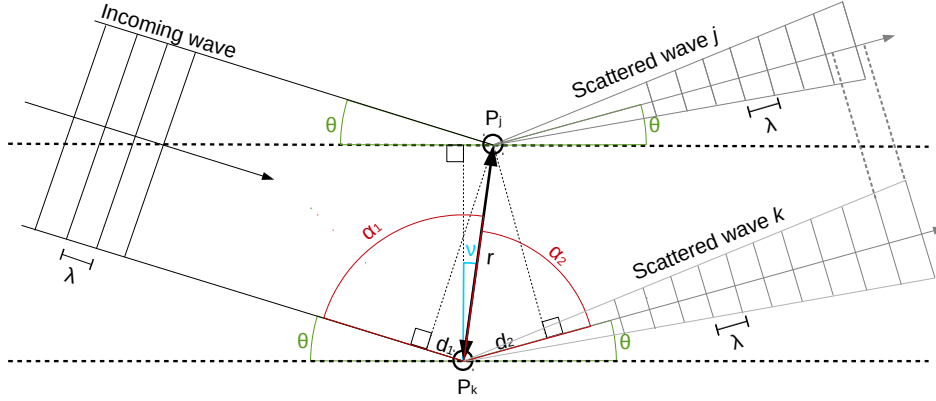


Figure 2.3: A schematic drawing of a scattering event. An incoming plane wave is scattered at  $P_j$  and  $P_k$  at a scattering angle  $2\theta$ .  $r$  is the distance between  $P_j$  and  $P_k$ ,  $\nu$  is the angle between  $r$  and the normal to the scattering plane (dashed line).  $\alpha_1$  and  $\alpha_2$  are the angles between  $r$  and the incoming and scattered waves respectively.  $d = d_1 + d_2$  is the extra length traveled by the wave scattered from  $P_k$  compared to the wave scattered from  $P_j$ . The lines on the incoming wave mark the position of wave crests with distance  $\lambda$ . Likewise, do the lines at the scattered wave mark the crests of the near-plane part of the outgoing waves. When the crests align, there is full constructive interference. In this example, wave  $j$  and  $k$  are slightly out of phase.

The distance between the scattering pair  $P_j$  and  $P_k$  is denoted  $r$ . Generally,  $d = d_1 + d_2 = r \cos(\alpha_1) + r \cos(\alpha_2)$ , where  $\alpha_1$  and  $\alpha_2$  are the angles between  $r$  and the direction of the incoming and scattered wave respectively.  $\nu$  is the angular difference between  $r$  and the normal to the the scattering plane. From Fig. 2.3, we see that  $\alpha_1 + \theta - \nu = \pi/2$  and  $\alpha_2 + \theta + \nu = \pi/2$ , where  $2\theta$  is the scattered angle. Using those relations together with the trigonometric identities  $\cos(\pi/2 - a) = \sin(a)$  and  $\sin(a \pm b) = \sin(a) \cos(b) \pm \cos(a) \sin(b)$ , we can obtain a simple expression for  $d$ :

$$d = 2r \sin(\theta) \cos(\nu). \quad (2.4)$$

The phase difference, in radians, accumulated over the distance  $d$  is  $\Delta\phi_{jk} = 2\pi d/\lambda$ . We insert this into

equation (2.3):

$$I(q) = \frac{1}{V} \sum_{j,k=1}^{N_s} b_j b_k \cdot \exp(-iqr \cos \nu), \quad (2.5)$$

where  $q = 4\pi \sin(\theta)/\lambda$ . As  $\mathbf{q}$  and  $\mathbf{r}$  are vectors, we see that  $\nu$  is the angle between them. However, as long as we stay in the isotropic regime, the vector notation is redundant for the derivation, despite arguably more simple than the trigonometric approach presented here.

In SAS on isotropic samples, we always measure the orientational averaged intensity  $\langle I(q) \rangle$ , and we can therefore exploit that  $\langle \exp(-iqr \cos \nu) \rangle = \text{sinc}(qr)$ , where  $\text{sinc}(x) = \sin(x)/x$ . We have thus derived a very fundamental equation in SAS, the Debye equation, describing the intensity for point scatterers in vacuum:

$$I(q) = \frac{1}{V} \sum_{j,k=1}^{N_s} b_j b_k \cdot \text{sinc}(qr). \quad (2.6)$$

## 2.2 Scattering from molecules in solution

Atoms can form macromolecules. Using equation (2.6) for all atoms in a particle provides the theoretical scattering for that particle in vacuum. However, biochemists are rarely happy if you tell them to deliver a sample of particles in vacuum. So we better consider a sample in a solvent. I will sometimes write "solvent" and sometimes "buffer" as the solvent for biological samples is usually pH controlled. The solvent can be assumed to consist of evenly distributed and identical scatterers with scattering length  $b_s$ . Calculating the elastic scattering from the solvent alone gives  $I(q) = Mb_s^2/V$ , where  $M$  is the number of solvent atoms. The  $q$ -dependency vanishes because all terms with  $\text{sinc}(qr)$  for  $r \neq 0$  cancel out. This happens because any wave will have a counter wave with opposite phase summing up to a total of zero (Fig. 2.4A). The only scattering left is therefore the  $q$ -independent scattering from the self-terms.

To understand scattering from a sample, it is useful to understand how lack of scattering can result in a scattering signal. If a small vacuum bubble somehow emerges in a solvent, it will give rise to a measurable  $q$ -dependent scattering signal, despite that the vacuum itself does not scatter. The measured signal is due to the *absence* of scattering. As the scattered wave from a solvent scatterer is missing its counterpart, the total scattering is non-zero (Fig. 2.4B). The resulting total scattered wave is identical to that from a particle in vacuum with the same shape as the vacuum bubble and with a scattering length of  $b_p = b_s$ , except for a phase shift of  $\pi$  (Fig. 2.4D). Identical scattering signals can also be obtained by a particle in solvent with the same shape but with  $b_p = 2b_s$  (Fig. 2.4C). Other situations with the same measured outcome could be constructed by varying  $b_p$  and  $b_s$ . In situation B, C and D the common factor is the shape and the excess scattering length  $\Delta b = b_p - b_s$ . This is a general scattering concept as realized in the late 1800s by Jacques Babinet and hence called Babinet's principle. It is a tremendously convenient result, as we can now calculate the expected theoretical coherent scattering from a sample by summing only over the scatterers in the macromolecules. This is done by replacing  $b_j$  with  $\Delta b_j = b_i - b_s$ :

$$I(q) = \frac{1}{V} \sum_{j,k=1}^{N_s} \Delta b_j \Delta b_k \cdot \text{sinc}(qr). \quad (2.7)$$

There are usually much fewer scatterers in the macromolecules than in the solvent, so  $N_s$  is much smaller than in equation (2.6).  $\text{sinc}(qr) \rightarrow 0$  as  $r \rightarrow \infty$ , so for a dilute system where the macromolecules are far apart, the inter-molecular terms can be neglected. We can therefore reduce  $N_s$  to be only the number of scatterers in a single macromolecule, and express the intensity in terms of the number density  $n = N_m/V$ ,



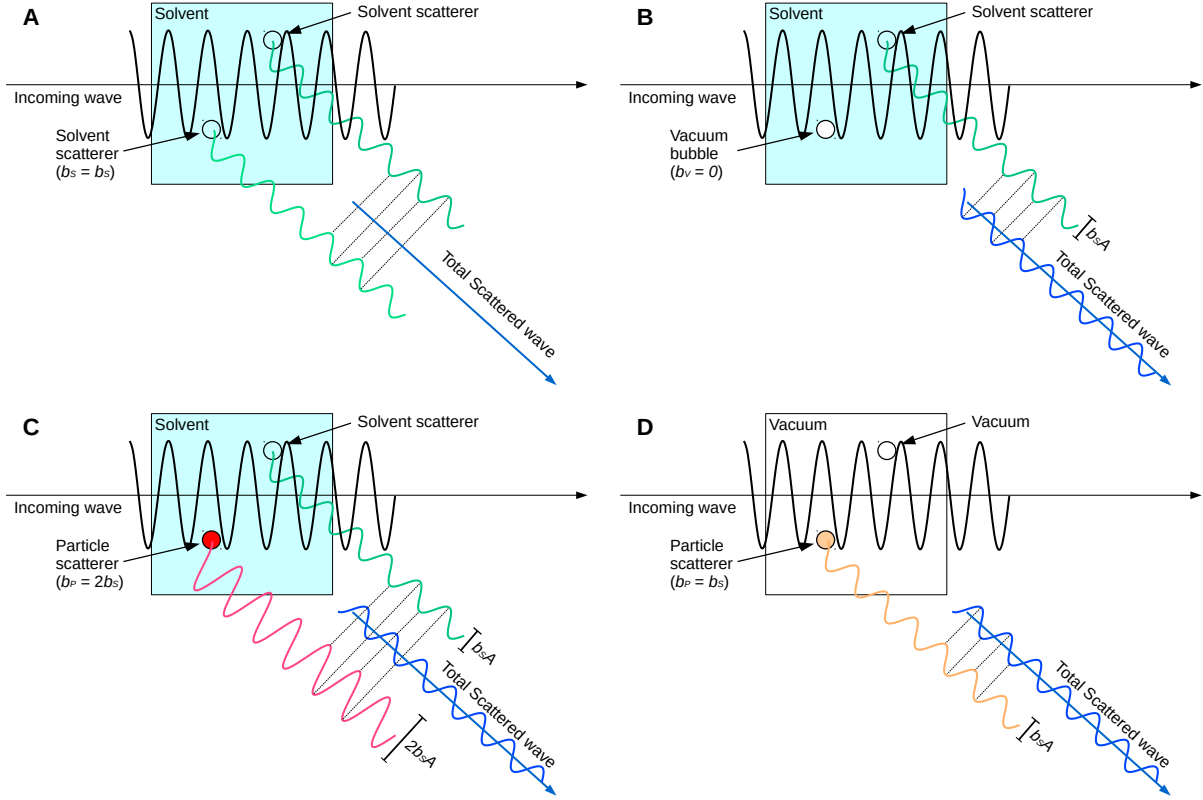


Figure 2.4: Four different scattering events. (A) The scattering from a solvent scatterers is canceled out by another solvent scatterer with a phase difference of  $\pi$ , resulting in a total (coherent) scattering of zero. (B) A vacuum bubble does not scatter, resulting in the scattering from the solvent scatterer not being canceled out. (C) A scatterer with a scattering length  $b_p$  twice that of the solvent scattering length  $b_s$  results in total coherent scattering defined by the difference in scattering length. (D) A scatterer in vacuum with a scattering length  $b_p = b_s$  results in total scattering. There is no scattering from (A). The total scattered wave for (C) and (D) are identical, and the scattered wave from (B) only differ in sign, so the measured intensity is the same for (B), (C), and (D).

where  $N_m$  is the number of macromolecules in the irradiated part of the sample:

$$I(q) = n \sum_{j,k=1}^{N_s} \Delta b_j \Delta b_k \cdot \text{sinc}(qr). \quad (2.8)$$

The number density can be determined experimentally, and is usually given as a molar concentration  $n_m = n/N_A$ , where  $N_A$  is Avogadro's number  $6.022 \times 10^{23} \text{ mol}^{-1}$ , in units of  $M = \text{mol/liter}$ . Or it may be given as a weight concentration,  $c = M_W n_m$ , where  $M_W$  is the molar weight, in units of  $\text{mg/ml}^2$ . When working with self-assembling particles, and colloids, where the number of particles vary, it is more convenient to work with volume fractions  $\phi = nV_m$ , where  $V_m$  is the volume of the macromolecule.

<sup>2</sup>The notation of  $\text{mg/ml}$  is convenient in a lab, where  $\text{ml}$  and  $\text{mg}$  are naturally occurring units. Therefore it is used instead of, from a physicist's point of view, more logical units such as  $\text{g/l}$  or  $\text{kg/cm}^3$ .

## 2.3 The scattering length weighted histogram of distances

By binning the scattering pairs after distance  $r$ , equation (2.8) can be reduced to a single sum:

$$I(q) = n \sum_{i=1}^{N_{bin}} h_i \cdot \text{sinc}(qr_i), \quad (2.9)$$

where  $h_i$  is the  $i^{\text{th}}$  bin, counting scatter pairs with distance  $r_i \pm dr/2$ , where  $dr$  is the bin size. The pairs are weighted with the product of their excess scattering lengths:

$$h_i = \sum_{j,k=1}^{N_s} f(r_{jk}) \cdot \Delta b_j \Delta b_k, \quad \text{where} \quad f(r_{jk}) = \begin{cases} 1 & \text{if } (i-1) \cdot dr \leq r_{jk} < i \cdot dr, \\ 0 & \text{otherwise,} \end{cases} \quad (2.10)$$

and  $r_{jk} = |r_j - r_k|$ . The bin size defines the spacial resolution of the histogram. The  $i^{\text{th}}$  distance,  $r_i$ , belonging to  $h_i$  is given by  $(i-1/2) \cdot dr$ . This histogram is very closely related to the pair distance distribution function,  $p(r)$ , which is used extensively in the interpretation of SAS data. I will in the following make a leap to the continuous limit, most frequently used to derive the expression for  $p(r)$  (e.g. Porod 1982; Spalla 2002), before I discuss the discrete pair distance distribution function with bins  $p_i$  and how  $p_i$  relates to  $h_i$ .

### 2.3.1 Pseudo Fourier mates

A molecule with many scatterers may be approximated as a continuous volume with a position-dependent scattering length density  $\rho(\mathbf{r})$ . The Patterson function describes the distribution of scatterers in the molecule:

$$\tilde{\rho}^2(\mathbf{r}) = \int_{V_m} \rho(\mathbf{r}') \rho(\mathbf{r}' + \mathbf{r}) dV_m, \quad (2.11)$$

where  $\rho(\mathbf{r}')$  is the scattering length density at position  $\mathbf{r}'$ .  $\mathbf{r}$  is a distance between two volume elements  $dV_m$ , and  $V_m$  is the molecular volume. The scattered intensity  $I(\mathbf{q})$  and  $\tilde{\rho}^2(\mathbf{r})$  are related by Fourier transformation:

$$I(\mathbf{q}) = n \int_{V_m} \tilde{\rho}^2(\mathbf{r}) e^{-i\mathbf{q}\mathbf{r}} dV_m, \quad \tilde{\rho}^2(\mathbf{r}) = \left( \frac{1}{2\pi n} \right)^3 \int_{q\text{-space}} I_1(\mathbf{q}) e^{i\mathbf{q}\mathbf{r}} d\mathbf{q}. \quad (2.12)$$

They are so-called Fourier mates. See also (Porod 1982) for a good and clear derivation<sup>3</sup>.

Assuming isotropic scattering, the vector relation can be reduced to a scalar relation between the one-dimensional pair distance distribution function  $p(r)$  and the intensity  $I(q)$ :

$$I(q) = 4\pi n \int_0^{D_{\max}} p(r) \text{sinc}(qr) dr, \quad p(r) = \frac{1}{2\pi^2 n} \int_0^\infty (qr)^2 I(q) \text{sinc}(qr) dq, \quad (2.13)$$

where  $r$  is the distance between pairs of scatterers in the molecules and  $D_{\max}$  is the largest distance.  $I(q)$  and  $p(r)$  are, like  $I(\mathbf{q})$  and  $\tilde{\rho}^2(\mathbf{r})$ , often referred to as Fourier mates. Truly,  $I(q)$  can be found by Fourier transformation of  $p(r)$ , but  $p(r)$  is, strictly speaking, not a Fourier transformation of  $I(q)$ , due to the factor  $(qr)^2$ . "Pseudo Fourier mates" might thus be a better notation for  $I(q)$  and  $p(r)$ . But their relation is directly derived from the Fourier transformations in equation (2.12).

<sup>3</sup>To directly compare the with Porod's derivations, it should be noted that Porod defines  $I(\mathbf{q})$  as the intensity from a single molecule, and as a unit less quantity (number of electrons) and  $\tilde{\rho}^2(\mathbf{r})$  has units of  $1/V$ . I use a slightly different notation, with  $I(\mathbf{q})$  being the normalized intensity from all irradiated particles, with units of  $1/\text{length}$  (squared scattering length per volume).  $\tilde{\rho}^2(\mathbf{r})$  likewise has units of  $1/\text{length}$  in my notation.

### 2.3.2 Indirect Fourier transformation

A scattering experiment may be considered a physical Fourier transformation of the real-space molecule, as represented by  $\tilde{\rho}^2(\mathbf{r})$  [eqn. (2.12)]. So an inverse Fourier transformation should bring the data back to real space and give information about spacial coordinates. The one-dimensional  $p(r)$  can in principle be derived directly from  $I(q)$  [eqn. (2.13)]. However, the integration in (2.13) goes to infinity, and  $I(q)$  is only known in a finite  $q$ -range (Fig. 2.1). On top of that,  $I(q)$  may be noisy, especially at large values of  $q$  as the intensity drops as  $q^{-4}$  (Porod, 1982). So the integral can not be solved directly. A solution to this problem is an indirect Fourier transformation (IFT) as introduced in the field of SAS by Glatter (1977). Here, the transformation is constraint by a smoothness criterion for  $p(r)$ . An alternative to Glatter's method was introduced by Moore (1980), and other methods exists, as discussed by Hansen & Pedersen (1991). The IFT method was later improved in BayesApp (Hansen 2000 and 2014) and autoGNOM (Petoukhov *et al.* 2007) such that the transformation algorithms are now fast, automatic and robust.  $p(r)$  gives direct real-space information about the investigated particles. The maximal distance,  $D_{\max}$ , can be determined as the largest non-zero value of  $p(r)$ , and the radius of gyration,  $R_g$ , can be calculated by:

$$R_g^2 = \frac{1}{2} \frac{\int_0^{D_{\max}} r^2 p(r) dr}{\int_0^{D_{\max}} p(r) dr}. \quad (2.14)$$

Moreover, the shape of  $p(r)$  relates directly and intuitively to the shape of the molecule (Glatter 1977). This is exemplified in Fig. 2.5, where the  $p(r)$  is shown for different simulated structures.

As an example of the intuitive interpretation of the  $p(r)$ , we can consider the sphere (blue), which has a completely symmetric  $p(r)$  with most pair distance at  $r = R$ , where  $R$  is the radius of the sphere. Another example is the cylinder (yellow) having a peak around its cross-sectional radius  $R = 0.1$  and a long tail ending at  $D_{\max}$ , which is approximately equal to (but slightly larger than) the rod length,  $L = 1$ . The simulated structures were generated with the simulation tool available at BayesApp (Hansen 2014; <http://www.bayesapp.org/simtest/>).

### 2.3.3 The relation between $p(r)$ and $h(r)$

I will in the following describe the relation between the discrete histogram and the continuous  $p(r)$ . Taking the continuous limit of equation (2.9), yields:

$$I(q) = n \int_0^{D_{\max}} h(r) \sin(qr) dr \quad (2.15)$$

Comparing with equation (2.13) shows that  $h(r) = 4\pi p(r)$ . That is,  $p(r)$  is a scattering-length weighted histogram. That also implies that  $h(r)$  can be calculated from  $I(q)$  by the inverse Fourier transformation:

$$h(r) = \frac{2}{\pi n} \int_0^\infty (qr)^2 I(q) \text{sinc}(qr) dq. \quad (2.16)$$

The expression for  $h(r)$  in equation (2.10) may also be written in the continuous limit, using the Dirac delta function  $\delta(\mathbf{r} - \mathbf{r}')$ , which is zero except at  $\delta(0)$ :

$$h(r) = \int_{V_m} \int_{V'_m} \Delta\rho(\mathbf{r}) \Delta\rho(\mathbf{r}') \delta(\mathbf{r} - \mathbf{r}') dV_m dV'_m. \quad (2.17)$$

For a homogeneous sample, i.e. when  $\Delta\rho(\mathbf{r})$  is constant, then equation (2.17) can be simplified to:

$$h(r) = \Delta\rho^2 \int_{V_m} \int_{V'_m} \delta(\mathbf{r} - \mathbf{r}') dV_m dV'_m. \quad (2.18)$$

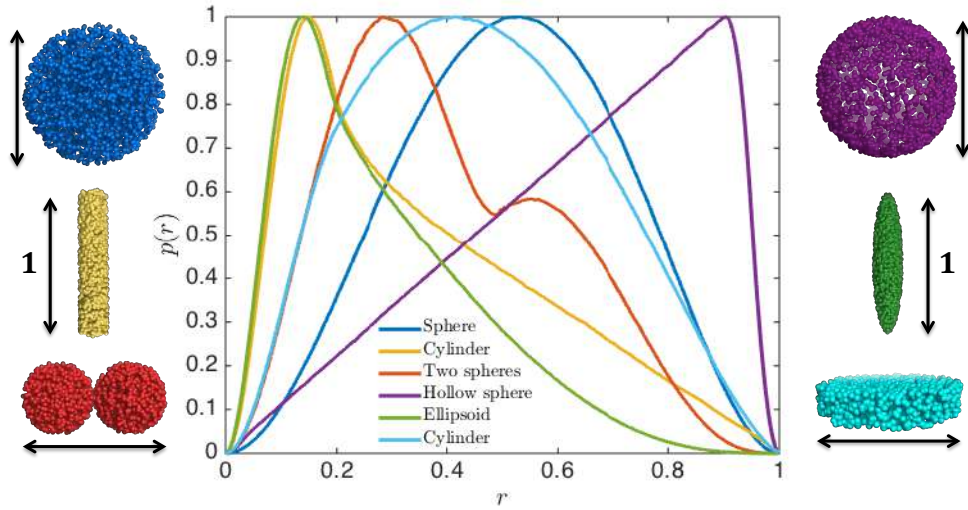


Figure 2.5: The  $p(r)$  function calculated for different shapes: sphere with  $R = 1$  (blue), rod with  $R = 0.1$  and  $L = 1$  (yellow), two touching spheres with  $R = 0.25$  shifted respectively  $+R$  and  $-R$  from their common center (red), hollow sphere with  $R_{\text{inner}} = 0.45$  and  $R_{\text{outer}} = 0.50$  (purple), ellipsoid of revolution with semi axes  $A = 0.5$  and  $B = 0.1$  (green), cylinder with  $R = 0.5$  and  $H = 0.2$  (cyan). They all have  $D_{\text{max}} \approx 1$ , but the shape of the corresponding  $p(r)$  as well as the  $R_g$  [eqn. (2.14)] varies a lot.  $R_{g,\text{sph}} = 0.39$ ,  $R_{g,\text{rod}} = 0.30$ ,  $R_{g,2\text{ sph}} = 0.32$ ,  $R_{g,\text{hlw sph}} = 0.48$ ,  $R_{g,\text{elip}} = 0.23$ ,  $R_{g,\text{cyl}} = 0.36$ . Note that  $R_{g,\text{hlw sph}} \approx (R_{\text{outer}} + R_{\text{inner}})/2$ .

In the limit  $r = \mathbf{r} - \mathbf{r}' \rightarrow 0$ , we get:

$$h(r \rightarrow 0) = \Delta\rho^2 4\pi r^2 \int_{V_m} dV_m = 4\pi r^2 \Delta\rho^2 V_m = 0 \quad (2.19)$$

Or, expressed in terms of  $p(r)$ :

$$p(r \rightarrow 0) = r^2 \Delta\rho^2 V_m = 0 \quad (2.20)$$

I will use these limit values in the next section. Interestingly, the discrete histogram does have a finite value at  $r \rightarrow 0$ , namely the sum of the self-terms ( $i = j$ ) [eqn. (2.10)]. As  $\text{sinc}(qr) = 0$  for  $r \rightarrow 0$ , these self-terms are independent on  $q$  and simply adds a constant to the scattering. In Fig 2.6, I have plotted  $h(r)$  with and without self-terms, denoted  $h^+(r)$  and  $h^-(r)$  respectively, as calculated for the large membrane protein GluA2 (AMPA type ionotropic glutamate receptor 2; PDB: 3KG2), and for the small soluble protein lyz (lysozyme; PDB: 1LYZ).  $h^+(r)$  has a peak in the first bin, which is not present when the self-terms are not included. When calculating  $I(q)$ , the self-terms add a  $q$ -independent constant to the scattering. As there are  $N$  self-terms and  $N^2 - N$  other terms in the Debye sum, the relative self-term contribution will be vanishing for large proteins, but considerable for smaller proteins (Fig. 2.6).

### 2.3.4 What quantity is measured?

When doing an experiment, one needs to measure both the sample of macromolecules in solvent and the solvent itself. The measurement of the solvent is often called the background measurement. There will be some coherent scattering in both measurements from the sample holder, air scattering etc. As this contribution is identical for sample and solvent, it vanishes upon subtracting. It will in the following be

assumed that there is no coherent scattering from the solvent in the measured  $q$ -range of a SAS experiment. In reality, the solvent may be slightly inhomogeneous and have structure correlation resulting in  $q$ -dependent scattering. Scattering from water in SAXS, for example, has some  $q$ -dependency (Huang *et al.* 2009). Even when neglecting this, the solvent still give self-term scattering. Note that the sample measurement does not have the self-term contribution from the solvent excluded by the macromolecule ( $K_{(\text{excl solv})}$ ). The subtracted scattering, assuming identical transmissions, is thus:

$$\begin{aligned} I_{\text{meas}}(q) &= I_{(\text{mol in solv} + \text{bg})}(q) - I_{(\text{solv} + \text{bg})}(q) \\ &= [I_{\text{mol}}(q) + I_{\text{bg}}(q) + K_{\text{solv}} - K_{(\text{excl solv})}] - [I_{\text{bg}}(q) - K_{\text{solv}}] \\ &= I_{\text{mol}}(q) + K_{\text{mol}} - K_{(\text{excl solv})}. \end{aligned} \quad (2.21)$$

The first term in the last line is the theoretical  $q$ -dependent scattering, as given in equation (2.13). The constant contributions are self-term contributions. Thus, there will always be a constant background that must be added/subtracted to get a measurement of the theoretical scattering. This is an important point for the next section.

### 2.3.5 The scattering invariant $Q$ and its relation to the molecular volume

In the work with GluA2 in Paper IV, we assessed the oligomeric state using the scattering invariant  $Q$ . Here, I will introduce the theoretical background.

We define  $\gamma(r) \equiv p(r)/r^2$ . For a homogeneous particle ( $\Delta\rho(\mathbf{r}) = \Delta\rho$ ), then  $\gamma(0) = \Delta\rho^2 V_m$  [eqn. (2.20)]. Taking the  $r \rightarrow 0$  limit of  $\gamma(r)$  [eqn. (2.13)], where  $\text{sinc}(qr \rightarrow 0) = 1$ , yields:

$$\gamma(r \rightarrow 0) = \frac{Q}{2\pi^2 n}, \quad \text{where} \quad Q = \int_0^\infty q^2 I(q) dq. \quad (2.22)$$

$Q$  is an "invariant" as it is independent on the particle shape (Porod 1951 and 1982). By comparing the two expressions for  $\gamma(0)$  we obtain a relation between  $Q$  and  $V_m$ :

$$Q = 2\pi^2 n \Delta\rho^2 V_m. \quad (2.23)$$

Using  $I(0) = n\Delta\rho^2 V_m^2$  (see appendix A), we can write an expression for the particle volume, which can, in principle, be found directly from the measured data, prior to any modeling:

$$V_m = \frac{2\pi^2 I(0)}{Q}. \quad (2.24)$$

This is the Porod volume, and can be found for data on arbitrary scale. As was the case for the inverse Fourier transformation, we have to deal with the limitation that  $I(q)$  is only known in a limited interval, so  $Q$  [eqn. (2.22)] cannot be solved exact. This issue will be discussed in the next section.

## 2.4 Model-free determination of the oligomeric state

For a protein, the  $M_W$  can be calculated directly from the amino acid sequence (e.g. by ExPASy ProtParam, <https://web.expasy.org/protparam/>). If the  $M_W$  can be determined experimentally, the oligomeric state can thus be determined by comparing with the theoretical value. There are two strategies to determine  $M_W$ . One is based on a determination of  $I(0)$  and the concentration, and one make use of the invariant  $Q$ . In the first method, it is exploited that  $I(0) = n\Delta\rho^2 V_m^2$ .  $V_m$  can be rewritten as  $M_W/\rho_W$ , where  $\rho_W$  is the molar weight density  $M_W/V_m$ , such that  $I(0) = n\Delta\rho^2 (M_W/\rho_W)^2$ , and:

$$M_W^2 = \frac{I(0)}{n} \left( \frac{\rho_W}{\Delta\rho} \right)^2 \quad \text{or} \quad M_W = \frac{I(0)}{c} \left( \frac{\rho_W}{\sqrt{N_A} \Delta\rho} \right)^2. \quad (2.25)$$

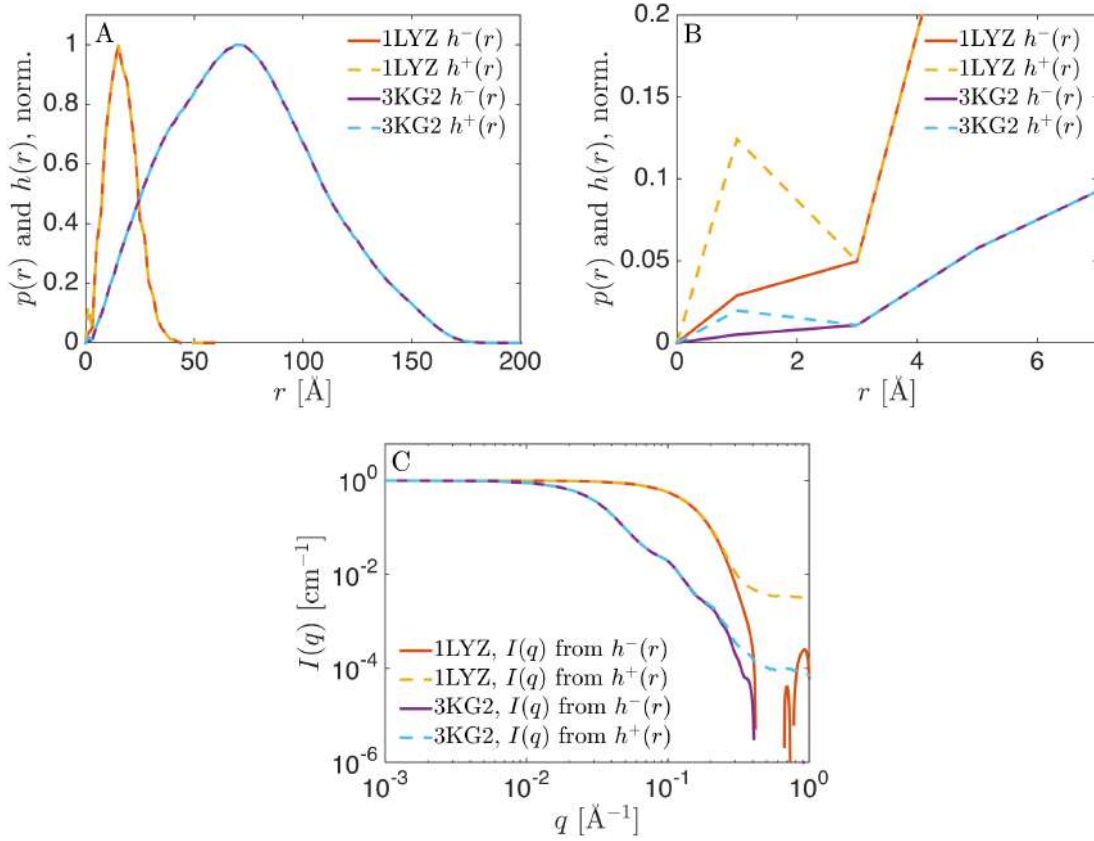


Figure 2.6: (A) Normalized  $h(r)$  without self-term  $h^-(r)$  (full line) and with self-term  $h^+(r)$  (dashed line) for lyz (PDB: 1LYZ, red and yellow) and the GluA2 (PDB: 3KG2, purple and blue). (B) Low- $r$  part, showing the first bin. (C) Scattering intensity for the protein structures, calculated with  $h^-(r)$  and  $h^+(r)$  respectively. They differ by a constant. Calculated with CaPP (chapter 3).

where  $c$  is the weight concentration,  $c = nM_W/N_A$ , and  $N_A$  is Avogadro's number.  $M_W$  can be determined by calculating  $\Delta\rho$  for the protein, e.g. by MULCh (Whitten *et al.* 2008) and using the average protein density  $\rho_W = 0.83$  kDa/nm<sup>3</sup> (Squire & Himmel 1979). Alternatively,  $\Delta\rho$  can be estimated by an average value for proteins. Mylonas & Svergun (2007) compared  $M_W$  calculated with different methods as tested on 14 structures. The  $\Delta\rho$  for these proteins range from  $2.9$  to  $3.2 \times 10^{-11}$  cm/nm<sup>3</sup>, with an average value of  $\Delta\rho = 3.0 \times 10^{-11}$  cm/nm<sup>3</sup>. That is,  $M_W$  can be approximated by:

$$M_W^2 = (3.1 \times 10^{10} \text{ kDa/cm})^2 \cdot \frac{I(0)}{n} \quad \text{or} \quad M_W = \left( \frac{1.3 \times 10^3 \text{ kDa}}{\text{cm}^{-1}/(\text{mg/ml})} \right) \frac{I(0)}{c}. \quad (2.26)$$

It is however recommendable to calculate  $\Delta\rho$  for each protein, as it may vary from protein to protein. E.g. the magnesium transporter protein CorA (PDB: 4I0U), has a  $\Delta\rho$  of  $2.7 \times 10^{-11}$  cm/nm<sup>3</sup>, and using the average value would thus introduce a 10% error (which is unnecessary as  $\Delta\rho$  is easily calculated).

As  $\rho_W$  and  $\Delta\rho$  are similar for different proteins,  $M_W$  can also be determined by measuring a standard protein with known molecular weight  $M_{W, \text{std}}$ :

$$\frac{M_W^2}{M_{W, \text{std}}^2} = \frac{I(0)/n}{I_{\text{std}}(0)/n_{\text{std}}} \quad \text{or} \quad \frac{M_W}{M_{W, \text{std}}} = \frac{I(0)/c}{I_{\text{std}}(0)/c_{\text{std}}}. \quad (2.27)$$

This method has the advantage that the intensity can be on arbitrary scale as long as the standard protein is measured under the exact same conditions.

These were all variation of the first method, that demands a measurement of the protein concentration.

The second method exploits the invariant  $Q$ , and has the advantage that a concentration measurement is not needed. The integral over  $q^2 I(q)$  [eqn. (2.22)] has to be solved by an approximation as  $I(q)$  is only measured in a limited  $q$ -range. The data has first to be expanded to  $q = 0$  by linear extrapolation or more sophisticated methods. The next problem is the limits in the positive direction.

At higher values of  $q$ , the data is poorly determined due to low signal-to-noise ratio. So the integral is truncated at a maximal value  $q_m$  to obtain  $Q_t$ :

$$Q_t = \int_0^{q_m} q^2 [I_{\text{exp}}(q) - K] dq, \quad (2.28)$$

where  $q_m$  is typically set to  $8/R_g$  (Petoukhov *et al.* 2012), but to my knowledge, this value for  $q_m$  has no theoretical justification. Note the constant  $K$ , as discussed in section 2.3.4. It ensures a correct background estimation, as the theory assumes that there is no constant  $q$ -independent contribution in the scattering. In SANS it also subtracts incoherent scattering. The estimation of  $K$  has a large impact on the final determination of  $M_W$ , so it is important to reliably estimate its value. In theory,  $I(q) \propto q^{-4}$  for large values of  $q$  if the surfaces of the protein is smooth (Porod 1982). The  $q^{-4}$  slope can be checked by a Porod plot (Fig. 2.7C),  $q^4 I(q)$  vs.  $q$ , that should approach a constant value at large values of  $q$  (Petoukhov *et al.* 2012; Rambo & Tainer 2011).

A "truncated" volume  $V_t$  can be determined from  $Q_t$  via equation (2.24):

$$V_t = \frac{2\pi^2 I(0)}{Q_t}. \quad (2.29)$$

This volume has no physical interpretation, and  $V_t > V_m$ . There are two approaches to obtain  $M_W$  from  $V_t$ , a first and a second-order approach. In the first-order approach,  $M_W$  is determined directly from  $V_t$  using  $M_W = 0.625 \text{ kDa/nm}^3 V_t$  (Petoukhov *et al.* 2012). This empirical constant was determined by applying the method to a range of structures from the PDB. The method is implemented in ATSAS (Franke *et al.* 2017) and has a reported uncertainty of about 20%. In the second-order approach (Fischer *et al.* 2010),  $V_m$  is determined from  $V_t$  using linear coefficients  $A$  and  $B$ :

$$V_m = A(q_m) \cdot V_t + B(q_m). \quad (2.30)$$

Fischer *et al.* showed that the best values of the linear coefficients depends on  $q_m$  and thus on the size of the protein (via.  $q_m = 8/R_g$ ).  $M_W$  is then determined from  $V_m$  using the average protein density of  $0.83 \text{ kDa/nm}^3$  (Squire & Himmel 1979). The method has a reported uncertainty of 12-16 %, depending on size, with larger proteins (small  $q_m$ ) having larger uncertainties. The method is implemented in SAXSMoW (Fischer *et al.* 2010; <http://saxs.ifsc.usp.br/>). We used my own implementation of the method in Paper IV, where the  $M_W$  for GluA2 was calculated to assess its oligomeric state.

Rambo & Tainer (2013) proposed an alternative approach for concentration and model independent determination of  $M_W$ . The invariant  $Q$  relies on an integration over the Kratky plot, which converge to zero for large values of  $q$  if and only if the protein is well-folded. For unfolded proteins, including IDPs, the Kratky plot however diverges. Rambo & Tainer exploited that  $qI(q)$  vs.  $q$  converges for large values of  $q$  also for unfolded proteins. By screening over 9000 structures from the PDB, they found the empirical relation:

$$M_W = \left( k_1 \cdot \frac{I^2(0)}{[\int_0^{q_m} q I(q) dq]^2} \right)^{k_2} \quad (2.31)$$

where  $k_1$  and  $k_2$  are empirical constants, which values depend on the type of sample. For proteins,  $k_1 = 8.12$  and  $k_2 = 1.0$  and for RNA  $k_1 = 107.06$  and  $k_2 = 0.808$  when  $q_m = 0.5 \text{ \AA}^{-1}$ . The reported error on the average mass is  $4.0 \pm 3.6\%$  for data with  $q_m = 0.5 \text{ \AA}^{-1}$ , and 4.6 for data with  $q_m = 0.3 \text{ \AA}^{-1}$ . In line with the Fischer method (Fischer *et al.* 2010),  $k_1$  and  $k_2$  depends on the  $q_m$ . The method is still incompletely

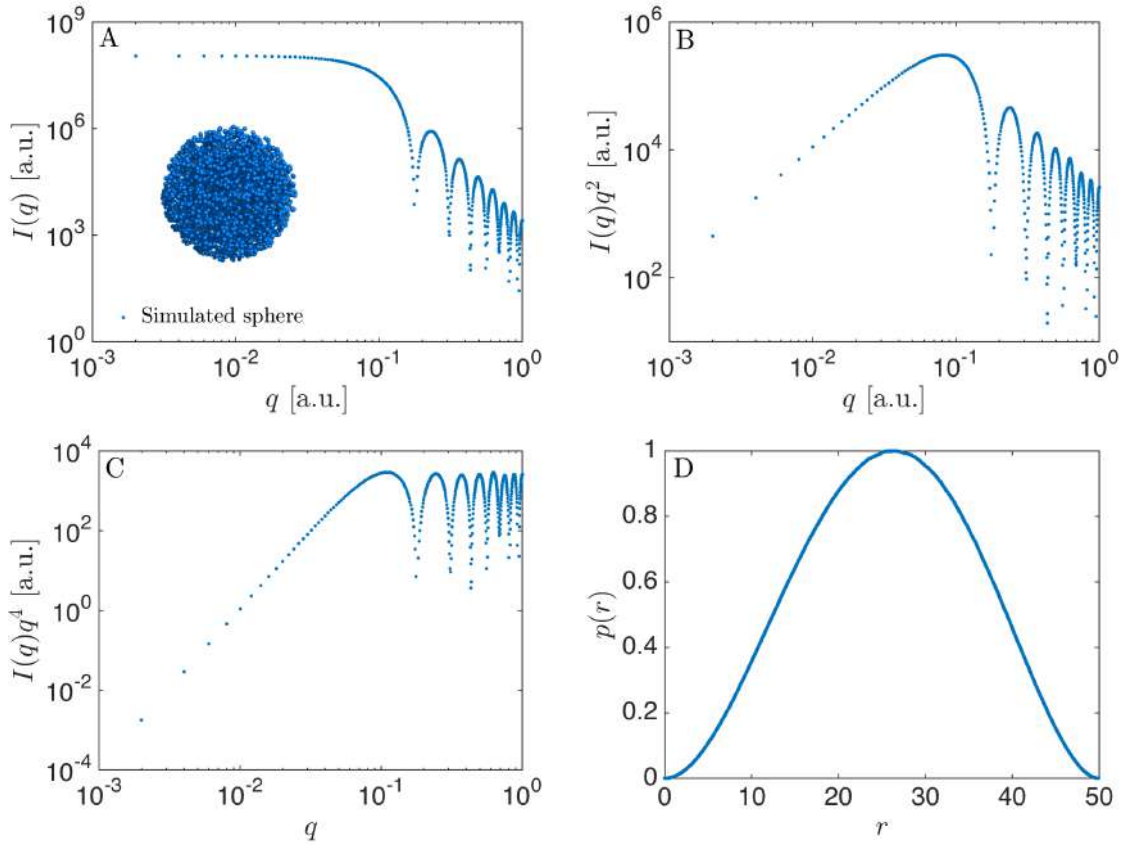


Figure 2.7: Calculated SAXS intensity for a simulated sphere (calculated without self-terms). (A)  $I(q)$  vs.  $q$ . The insert shows the simulated sphere. (B) Kratky plot,  $I(q)q^2$  vs.  $q$ , with a shape characteristic for a folded structure. (C) Porod plot,  $I(q)q^4$  vs.  $q$ , which reaches a linear plateau at high- $q$  assuming a smooth surface. (D) Pair distance distribution function,  $p(r)$  vs.  $r$ .

described as  $k_1$  and  $k_2$  are only reported for protein and RNA with  $q_m = 0.5 \text{ \AA}^{-1}$ . The small reported error is however promising, as well as its application in studying unfolded proteins. The relation between the integral over  $qI(q)$  and  $M_W$  is also interesting, and might contain some interesting fundamental scattering theory. The method is implemented in SCATTER (<http://www.bioisis.net/tutorial/16>)<sup>4</sup>.

## 2.5 Incoherent scattering

In this last section of the chapter I will discuss incoherent scattering in SANS. Decreasing the incoherent scattering is an essential part of designing SANS experiments, as those described in chapter 5, in order to ensure a sufficiently good signal-to-noise ratio.

Until now, I have only discussed the coherent scattering contribution from elastic scattering. What is meant by "coherent"? In fully coherent light, the photons has the same wavelength, and the wave sources have synchronized phases, such that full constructive interference is achieved. A laser is a light source of coherent light, both when talking about a 5 cm laser pointer or a 5 km X-ray free electron laser (XFEL). An incoming beam of light or neutrons can be more or less coherent, depending on the optics of the beamline. The scattered light is also divided into coherent and incoherent scattering. The wave sources are the scatterers in the investigated molecule. Are these identical, then the scattering will be coherent. But

<sup>4</sup>Clearly, the person choosing the name "SCATTER" was not Scandinavian :-)



there might be random differences leading to incoherence. As neutrons are sensitive to spin, the scattering length,  $b$ , differs for the same atom, depending on the spin state. A hydrogen nuclei consists of a proton. As protons are spin  $1/2$  fermions, they have spin  $+1/2$  ("up") or spin  $-1/2$  ("down"), and there is equal probability of being in each state. Importantly,  $b_{up} \neq b_{down}$ . Deuterium consist of a proton and a neutron, and therefore has spin  $+1$  (parallel up) spin  $0$  (antiparallel) or spin  $-1$  (down parallel). However, as can be shown with quantum chromo dynamics (QCD), which is far out of scope for this thesis, the spin  $+1$  is energetically very favorable, so there is only minor spin heterogeneity in deuterium. Consequently, there is little incoherent scattering. The coherent and incoherent scattering cross sections are given by:

$$\begin{aligned}\sigma_{\text{coh}} &\propto \langle b^2 \rangle, \\ \sigma_{\text{incoh}} &\propto \langle b^2 \rangle - \langle b \rangle^2,\end{aligned}\tag{2.32}$$

where  $\langle x \rangle$  is the average of  $x$ . The numerical values for  $b_{\text{coh}}$  and  $b_{\text{incoh}}$  are listed for hydrogen, deuterium and oxygen in Table 2.1. The incoherent scattering is independent on the scattering angle and the position of atoms with respect to each other. It solely depends on the type of atoms in the radiated sample. So incoherent scattering merely contributes with a constant background in SAS and carries no structural information. For that reason, it is favorable to use  $\text{D}_2\text{O}$  instead of  $\text{H}_2\text{O}$  in the solvent in SANS, to limit the incoherent background, which lowers the signal-to-noise ratio without adding any structural information. The incoherent scattering does contain interesting information, e.g. about vibrational states, but that is out of scope of the current thesis.

	$b_{\text{coh}}$ [fm]	$b_{\text{incoh}}$ [fm]
H	-3.7	25.3
D	6.7	4.0
O	5.8	0.0

Table 2.1: Coherent and incoherent scattering lengths for hydrogen, deuterium and oxygen. Negative sign indicate a phase shift,  $\pi$ , of the scattered wave with respect to the incoming wave.

## 2.6 I few closing remarks

None of the theory is new. However, the presentation is slightly different from how it is usually presented.  $p(r)$  is rarely derived as a histogram, and I have not, so far, seen any direct comparison and discussion of  $h(r)$  and  $p(r)$ , nor of their discrete counterparts. Also, I have only seen very few derivations of the invariant  $Q$ . Usually it is introduced without any derivation, and Porod's original derivations (Porod 1951 and 1982) are, in my opinion, not very clear (the first of the two mainly because it is in German).

The theory will be implemented and approximated in chapter 3, and applied in chapter 5.



## Chapter 3

# Modelling Tools for Membrane Proteins

*"Any code of your own that you haven't looked at  
for six or more months might as well have been  
written by someone else."  
- Eagleson's Law*

This chapter presents different tools for modelling SAS data with membrane proteins. Software for analysis of soluble proteins have been developed extensively over the last 2-3 decades, especially through the ATSAS software package (Franke *et al.* 2017). However, these tools are not always applicable for membrane proteins due to the transmembrane domain that must be solubilized by a cell membrane mimicking system. Secondly, some membrane proteins are prone to aggregate, which must be taken into account as well. First, I will introduce the hydrophobic effect and that explains why a layer of densely packed water tends to form around proteins. This water layer was included in the analysis of all proteins and it was addressed how to add a correct water layer to membrane proteins. After that, I present a computer program, CaPP, in which the inclusion of water layer for membrane proteins has been implemented. The program has a range of other features that will be described. Finally, I present a method to take into account protein aggregation in challenging samples of membrane proteins.

### 3.1 The hydrophobic effect and water layer around proteins

Bulk water is structured in a network of hydrogen bonds. A hydrophobic molecule dispersed in water will perturb this network, leading to increase in enthalpy,  $H$ . At the same time, the entropy,  $S$ , will decrease, as the water molecules close to the dispersed molecule are more constrained in their movement. The total free energy,  $G = H - TS$ , where  $T$  is the temperature, therefore increases due to both effects. To reach a lower energy state, proteins folds and bury the hydrophobic patches in a hydrophobic core, leaving only the hydrophilic parts to face the water. This effect (which is not limited to proteins) is called the hydrophobic effect. It is the driving (pseudo)force in much self-assembly, e.g. of detergents.

The polar (hydrophilic) parts of the proteins, which faces the water in a folded protein, also perturb the water network, but to a lesser extent than the nonpolar (hydrophobic) parts. This gives rise to a layer of water in the vicinity of the proteins that has different properties than bulk water. SAS is sensitive to changes in water density, and it has been experimentally shown for a selection of soluble proteins that the water layer is about 10% more dense than bulk water (Svergun *et al.* 1998). Early MD simulations for lysozyme estimated the water layer to be even denser, namely 15% more dense than bulk water (Merzel & Smith, 2002). The magnitude of the density difference was however debated, and recent improved MD

simulations indicate that the water layer is only about 6 % more dense than bulk water (Persson *et al.*, 2018). In a recent study (Henriques *et al.*, 2018) the scattering was calculated for four different proteins using the solvent-implicit approach in CRY SOL (Svergun *et al.*, 1995) and compared with the calculated scattering using WAXSiS (Knight & Hub, 2015). WAXSiS uses MD simulations to include water explicitly in the calculations of the scattering. Henriques *et al.* showed that the density only increased by 2-6% for globular folded proteins, and even less, 2-3%, for intrinsically disordered proteins.

## 3.2 Computer program CaPP

During my PhD, I have written the program CaPP (Calculating Pair distance distribution functions for Proteins in the PDB format), which is open source and freely available online (<https://github.com/Niels-Bohr-Institute-XNS-StructBiophys/CaPP>). The main motivation for writing the program was to be able to calculate the theoretical  $p(r)$  functions for proteins, whose structure had been solved and deposited in the protein data bank (PDB). Surprisingly, no tools were available for doing that (to my knowledge), even in the extensive ATSAS software suite (Franke *et al.*, 2017). The scattering had to be calculated first, and an approximate  $p(r)$  could then be obtained by IFT of the calculated scattering. That is, one needed to take a detour through reciprocal space to obtain the theoretical  $p(r)$ . The direct approach used in CaPP, gives accurate and unambiguous solutions. The second motivation was to correctly include the water layer for membrane proteins. We had studied membrane proteins in novel "invisible" detergents (Paper III), such that the scattering came only from the protein and the water layer. However, as the detergents cover the transmembrane region of the membrane protein, the water layer should be excluded from this region. No tools were available for doing that (to my knowledge). The third motivation for writing the program was, that it is convenient to have a flexible program to quickly calculate the SAXS and SANS scattering from PDB files. Having an open-source program provides the possibility to modify and tailor the program depending on specific needs. The program is designed to be fast, intuitive and accurate, and its architecture makes it easily adjustable to specific needs.

### 3.2.1 The architecture of the program

An overview of the program architecture is given in Fig. 3.1. Initially, the program was made for calculating  $p(r)$  in a command line interface (CLI) mode, and was purely written in c. I decided to build on top a Python GUI to increase accessibility. The Python script behind the GUI also calculates the form factors  $P(q)$  from the  $p(r)$  [eqn. (2.12)], and plots the results for quick evaluation. Recently, I have added an option to fit the theoretical curve to data, which is also done in the Python-based part of the program. The GUI creates an input file that is executed by the c-part, meaning that the c-part is independent and can be used separately. It can thus easily be incorporated as a module in other programs, and can be run in batch mode.

CaPP has three overall user levels (see Fig. 3.1). At level 1, CaPP is run from the Python graphical user interface (GUI). At level 2, CaPP is run from a command line interface (CLI) and can be included in other programs or run in batch mode. The third level is the developer level, where the underlying c-programs and Python scripts are modified.

### 3.2.2 Features of CaPP

In the following, I will introduce a range of features of the program and explain how they are implemented

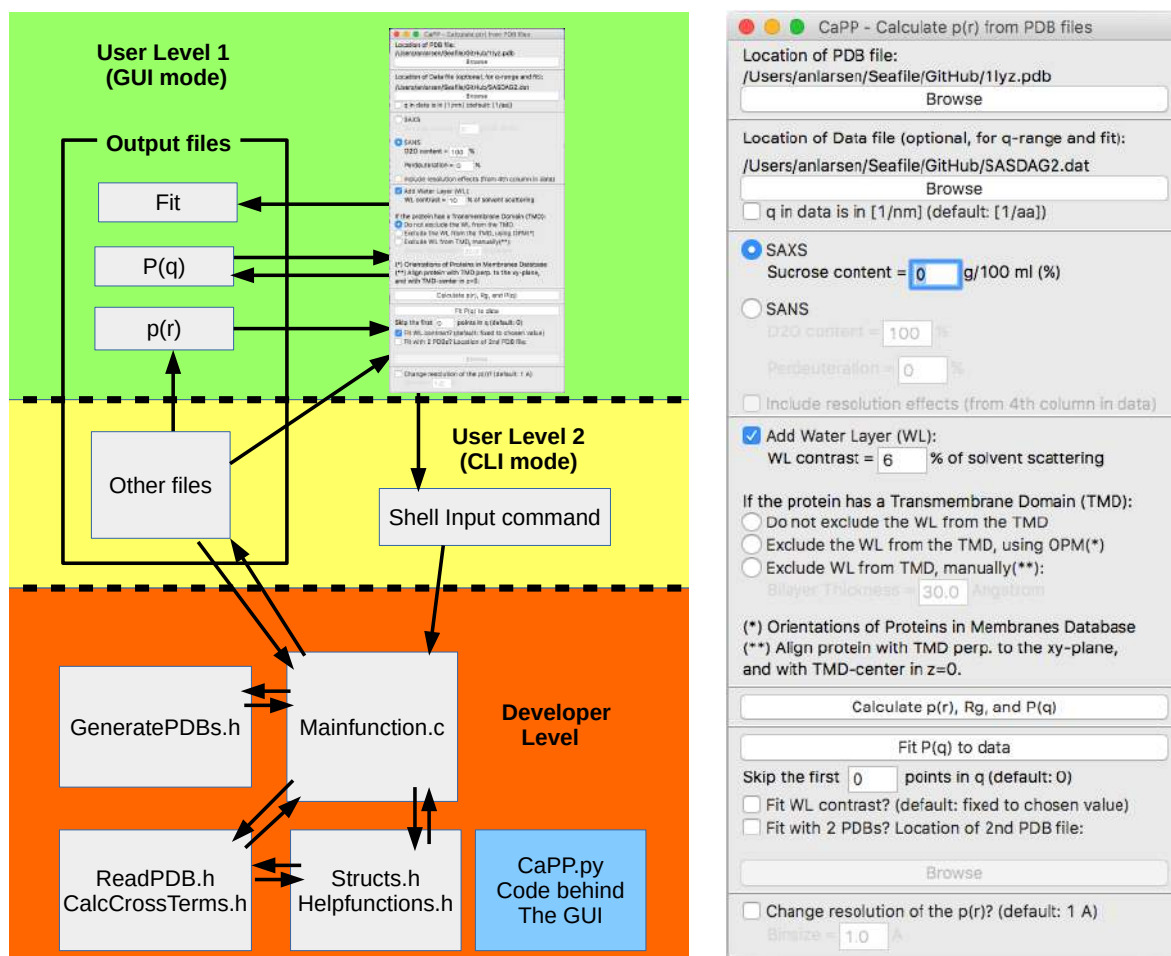


Figure 3.1: The Architecture of CaPP, Right: the CaPP GUI, enlarged. Level 1 (green) is the GUI mode. Level 2 (yellow) is the CLI mode and batch mode. Level 3 (orange/red) is the developer level. The arrows indicate how the different components of the program communicate.

## Inclusion of water layer

The water layer is included in CaPP by an algorithm made by Kynde (Kynde, 2014). Briefly, the algorithm analyzes if each amino acid is in the core or at the edge of the protein. By the edge-amino acids, a water bead is placed on the outside of the protein. A thickness of 3 Å is assumed, the surface area estimated, and each water bead then corresponds to 4.13 water molecules. The positions of the water beads are written into a separate file in the PDB format, which is read by CaPP when calculating the  $p(r)$  and the scattering. The file can also be used for visualization of the water layer, e.g. with PyMOL. I have compared CaPP with FoXS (Schneidman-Duhovny *et al.* 2010) and Crysol (Svergun *et al.* 1995), which are the most widely used tools for calculating the SAS scattering from a PDB. Crysol adds water by a continuous approach, and FoXS adds water in a way similar to CaPP, but at the position of the surface atom instead of on the outside of it (Schneidman-Duhovny *et al.* 2013). The CaPP/FoXS method has the advantage of being able to place water in cavities (Fig. 3.2). CaPP has the advantage over FoXS that the water expands the dimensions of the protein + water layer, e.g. resulting in larger  $D_{\max}$ , and is in that sense more correct. The density of the water layer can be adjusted in CaPP via the GUI or in the CLI by option `-c`. For membrane proteins, the water layer should be excluded in the transmembrane domain (TMD). This can be done in CaPP either by a manual option, where the protein is placed with the center of the TMD at  $z = 0$ , and the TMD placed orthogonal to the  $xy$ -plane, and then define the thickness of the hydrophobic part of

the bilayer, where no water beads should be placed. Alternatively, the structure can be downloaded from the Orientation of Proteins in Membranes (OPM) database (Lomize *et al.* 2012), which orients membrane protein structures from the PDB in a simplified lipid bilayer and estimate the thickness of the bilayer. If selected, CaPP uses the information from the OPM database to exclude water from the TMD (Fig. 3.2C).

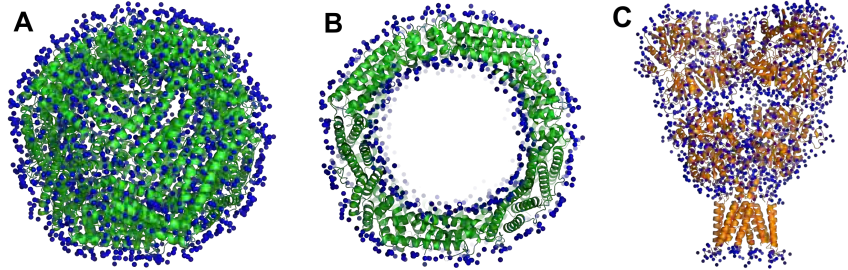


Figure 3.2: (A) The crystal structure of ferritin (PDB: 1MFR; green cartoon) with water layer added by CaPP (blue beads). (B) Cross section of the ferritin structure, showing the inner cavity. (C) The crystal structure of GluA2 (PDB: 3KG2; orange cartoon) with water layer added on the surface, except at the TMD.

### Fitting in reciprocal space

The calculated  $P(q)$  can be fitted to a dataset. If a dataset is given to the program, then the theoretical scattering will be calculated at the  $q$ -values of the data. A background and a scale parameter are fitted using a build-in Python least square fitting algorithm "curve\_fit", where the error bars are included as weight in the fit. The contrast of the water layer can also be fitted. This is however included as a separate option, as the  $p(r)$  function is altered by a change in contrast of the water layer and has to be recalculated. For large proteins, it is time-consuming to calculate the  $p(r)$ . Therefore, the  $p(r)$  is separated into three terms:

$$p(r) = p_{\text{prot}}(r) + p_{\text{wl}}(r) + p_{\text{cross}}(r). \quad (3.1)$$

The  $p(r)$  for the protein, for the water layer, and one with all the cross terms between water layer and protein. Only  $p_{\text{wl}}(r)$  and  $p_{\text{cross}}(r)$  need to be recalculated when the water layer contrast is changed. With that modification, it is feasible to fit the water layer contrast, as most terms are in the  $p_{\text{prot}}(r)$ . This has been implemented with a so-called golden section search, assuming a water layer excess scattering length (contrast) between -30% and +30% of the bulk water scattering length. For each tested value of the WL, the background and scale parameters are optimized with the fast built-in Python fitting algorithm. A fit of the lysozyme structure (PDB: 1LYZ) to a synchrotron SAXS dataset of lysozyme is shown in Fig. 3.3. An option for inclusion of two different structures (PDB files) has been included as well. The program finds the distribution of the two structures that best describes data. This parameter is fitted with the build-in Python fitting algorithm.

### Inclusion of resolution effects for SANS data

In SANS, the beam on the sample is relatively large, often around  $1 \times 1 \text{ cm}^2$ . Moreover, it is divergent, and has a considerable energy spread of about  $\pm 10 \%$  (FWHM). So there is a significant uncertainty on the values of  $q$  assigned to each pixel (Pedersen *et al.* 1990). These effects are collectively denoted smearing

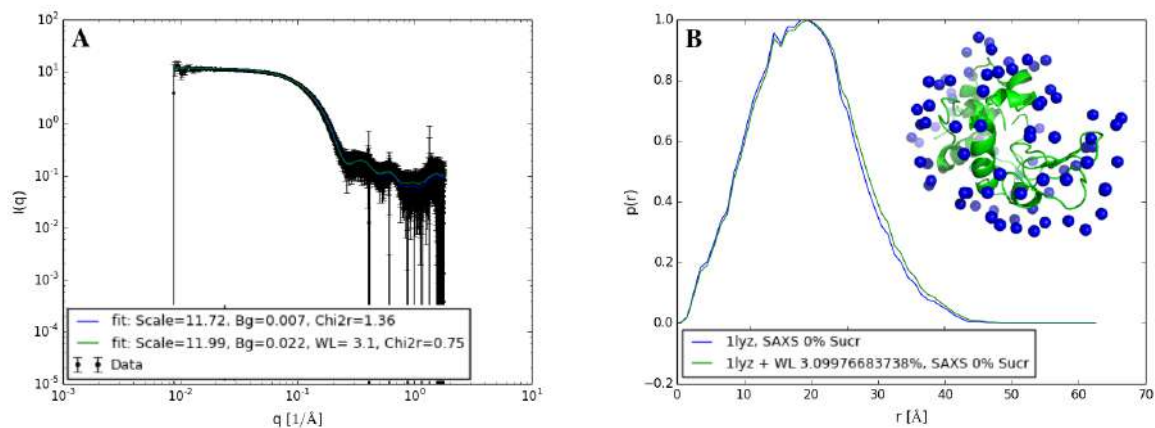


Figure 3.3: Direct output from CaPP except for inserted structure and enlarged legends. (A) SAXS dataset of a sample of monodisperse lysozyme from the SAS biological data bank (black, SASBDB: SASDAG2), fitted with the crystal structure shown in (B). Two models were fitted to data, one without a water layer (blue,  $\chi_r^2 = 1.36$ ), and one with a water layer (green,  $\chi_r^2 = 0.75$ ). (B) Theoretical  $p(r)$  for the two models. Insert is the fitted lysozyme crystal structure (PDB: 1LYZ) in green cartoon with blue water beads.

effects or resolution effects, as they smear the beam and lower the effective resolution obtainable by the data. The resolution effects are most evident for Bragg peaks and other sharp features with large values of  $\frac{dI}{dq}(q)$ , as these features are smeared. Most SANS beamlines provides a fourth column in data with the uncertainty on  $q$ , called  $\sigma_q$ , which is obtained either by calculations (Pedersen *et al.* 1990), by simulations, by measurements of well-known standard samples or by a combination of these. The resolution effect are then included by:

$$I(q) = \int_{-\infty}^{\infty} I(q') \mathcal{N}(q, \sigma_q) dq', \quad (3.2)$$

where the integrand  $I(q')$  is weighted by the normal distribution  $\mathcal{N}(q, \sigma_q)$  with mean  $q$  and standard deviation  $\sigma_q$ . In practice, the integral is approximated by a sum, and the limits changed from  $[-\infty, \infty]$  to  $[-3\sigma_q, 3\sigma_q]$ . SAXS is usually so well-collimated and monochromatic that the resolution effects are negligible. At home-source instruments, with less flux, the resolution effects may however be important, especially for data with Bragg-peaks or other sharp features.

### Contrast variation in SAXS and SANS

As explained in chapter 5, contrast variation is important in SAS, in particular in SANS. The contrast situation can be varied either by changing the solvent contrast by adjusting the D<sub>2</sub>O content, or by deuteration of the sample. Both parameters can be adjusted in CaPP. In SAXS, the contrast can be varied by changing the content of salt or other electron-dense molecules in the solvent, e.g. by addition of sucrose (Tokuda *et al.* 2016). The sucrose content in SAXS and the D<sub>2</sub>O contents and degree of protein deuteration can be adjusted easily in CaPP.

CaPP thus provides a tool for quick and easy comparison of theoretical curves for samples measured at different contrast. Examples are given in Fig. 3.4. The associated Python based plotting tool provides quick and easy comparison of the expected results, which is essential when preparing an experiment.

## Resolution of the $p(r)$

The binsize of  $r$ , denoted  $\Delta r$ , can be adjusted in CaPP. This corresponds to the "resolution" of  $p(r)$ . If  $\Delta r$  is too large, then the  $p(r)$  will be inaccurate, and the scattering calculated from  $p(r)$  will likewise be inaccurate. On the other hand, if a very small value is chosen for  $\Delta r$ , then  $p(r)$  will oscillate drastically, and can not be intuitively interpreted and easily compared with the experimentally obtained  $p(r)$ . I found  $\Delta r = 1.0 \text{ \AA}$  to be an appropriated value, but it can be adjusted directly in the GUI, and in the CLI by the option `-r`.

## Implicit hydrogen/deuterium

To increase speed and because the hydrogens are not written explicitly in a standard PDB file, the hydrogens/deuterium in the structures are accounted for implicitly, when calculating the  $p(r)$  with CaPP. The number of hydrogens bound to each atom in the amino acids are known exactly, so CaPP adds the volume and scattering length from the bound hydrogens to the heavy atom to which they are bound. This is the same method as applied in Crysol (Svergun *et al.* 1995). CaPP also contains a small library of the most common HETATM (hetero atoms), such as ligands, lipids, salts etc., and adds hydrogen/deuterium to these. If not in the library, no hydrogen will be added. In SAXS, hydrogen scattering is a relatively small contribution as the scattering length scales with the number of electrons, but in SANS the hydrogen/deuterium contribution is important to take into account. When using a SANS contrast, the labile hydrogens are exchanged with deuterium in a rate corresponding to the rate of  $D_2O$  in the solvent. NH is treated as a semi-labile group, with only 90 % of the NH groups being exchangeable. The same method is used to implement perdeuteration for the structures.

## Reading PDB files

A key to obtain accurate results with CaPP is to read the PDB files correctly. Examples of pitfalls is alternative positions, usually denoted A and B, whereof only one should be included. Another pitfall, that may lead to erroneous results if not accounted for is solved water (HOH) in the crystal. This should be given the same density as bulk water and thus do not contribute to the scattering. If not accounted for, it would be treated as oxygen with the same density as oxygen bound in proteins and thus contribute to the coherent scattering.

### 3.2.3 Computational speed

A typical protein has  $\sim 10^3$  atoms, so using a Debye sum to calculate the intensity, gives  $N_s^2 \approx 10^6$  terms to calculate [eqn. (2.6)], which is a problem as it is time consuming to compute  $\text{sinc}(qr)$ . Half of the terms are identical and the  $N_s$  self-terms can be calculated separately and with  $\text{sinc}(qr)=1$  since  $r_{jj} = 0$ . We are left with  $N_s + (N_s - 1) \cdot N_s/2 \approx 10^6/2$ . By binning data and calculating the scattering via.  $h(r)$  [eqn. (2.9)] the number of evaluations of  $\text{sinc}(qr)$  is drastically reduced from  $\sim 10^6$  to  $N_{bin}$ , where about 100-1000 bins is sufficient, depending on the size of the protein. The scattering length-weighted histogram  $h(r)$  is calculated by a double sum in  $N_s$  [eqn. (2.10)], so we still have to calculate  $10^6$  terms. But these have no evaluation of  $\text{sinc}(qr)$ , and are therefore computationally much less expensive. Using CaPP, a scattering curve can be calculated within a few seconds for a small protein ( $<150 \text{ kDa}$ ) and within less than a minute for larger proteins ( $150\text{-}500 \text{ kDa}$ ) on a standard laptop computer. Thus it is comparable to Crysol/Cryson in computational speed. The speed is also in part because this part of the code is written in c, which is much faster than higher level languages such as Python or MATLAB.



### 3.2.4 Approximations utilized to calculate the intensity in CaPP

The scattering from point scatterers can be calculated directly from the  $p(r)$  (or from  $h(r)$  as discussed in chapter 2 - the difference is a constant). In SANS, the scatterers are the nuclei, and these can be approximated as points. In SAXS however, the scatterers are the atomic electron clouds with empirically determined atomic form factors,  $f_a(q)$ . The  $q$ -dependency is lost when rebinning the atom pairs by distance. In CaPP, all form factors are therefore approximated by the form factor for carbon:

$$b_a \cdot f_a(q) \approx b_a f_C(q) \quad [\text{Approximation(1)}].$$

Following the methodology in Crysol and Cryson, the excluded water shape is approximated with a Gaussian sphere (Svergun *et al.*, 1995),  $g(V_a, q)$ . This sphere is unique for each atom, as its size depends on the atomic volume,  $V_a$ , as approximated by the Van der Waals volume. Again, the  $q$ -dependency is lost when generating the histogram. CaPP therefore finds the average atomic volume and uses this for all terms in order to take into account the excluded volume:

$$V_a \cdot g(V_a, q) \approx V_{av} g(V_{av}, q) \quad [\text{Approximation(2)}].$$

Using Approximation (1) and (2), the scattering can be calculated rapidly via the theoretical  $h(r)$ . To test the accuracy of the approximations, CaPP was benchmarked against Crysol and FoXS. The theoretical scattering for lysozyme (PDB: 1LYZ) was calculated without water layer (Fig. 3.5A) and with a water layer (Fig. 3.5B). The scattering calculated with CaPP matched well with that from the two other programs. The intensity calculated with FoXS did however differ from the two other curves by a constant, which was added prior to plotting and comparison. The theoretical scattering differed slightly when a water layer was included, due to the different implementations of this.

CaPP was also benchmarked by fitting the crystal structure to a high-quality dataset from the SAS biological data bank (SAXSBDB) of lysozyme monomers (SAXSBDB: SASDAG2). The number of harmonics in Crysol were increased from the default 15 harmonics to 30 in order to gain precise scattering predictions up to  $1.0 \text{ \AA}^{-1}$ . All programs generated a good fit up to about  $q = 0.7 \text{ \AA}^{-1}$  (Fig. 3.5C). Crysol and FoXS are hard-coded to stop fitting at  $q = 1.0 \text{ \AA}^{-1}$ . CaPP managed to fit the data with quite good accuracy up to about  $q = 1.5 \text{ \AA}^{-1}$ . Looking closer at the fits at  $q = 0.5 \text{ \AA}^{-1}$  (Fig. 3.5C, insert) shows that FoXS was struggling with the accuracy here, as compared to the two other methods that better capture the features of the data. The resulting  $\chi_r^2$  values were 0.92 for FoXS, 0.87 for Crysol, and 0.75 for CaPP. It should be noted that  $\chi_r^2$  for CaPP in the  $q$ -range 0 to  $1 \text{ \AA}^{-1}$  was 0.87, as this is the range in which the two other programs are limited to. The water layer density found with CaPP was about 3% larger than bulk water and with Crysol the water layer was estimated to be around 1% more dense than bulk water. FoXS gives a fitting value,  $c_2 = -0.47$  meaning that the water has less density than bulk water, and that there were about 1 water molecule for every second surface atom. The results from CaPP for the water layer matches best with the best theoretical predictions of a 6% denser water layer (Persson *et al.* 2018).

### 3.2.5 Scientific impact

CaPP provides an easy-to-use program for calculating the  $p(r)$  directly from PDB files, and has been used in Paper III to V. CaPP was also used in Paper I for calculating the  $A_0^0$  functions for the decoupling approximation (see the paper for details). The approximations applied to calculate the intensity appear to be good, as shown by benchmarking against Crysol and FoXS. The approximations can be used in other programs where computational speed is important, and accuracy is still needed. The implementation of the water layer is different from other methods in the literature and provides for the first time

automatic exclusion of the water at the transmembrane region. Moreover, it works very well for placing water in cavities (Fig. 3.2). CaPP is open source and freely available and can thus be modified to specific needs, as exemplified in Paper I, for which it was altered to calculate the so called  $A_0^0$  function. CaPP was also modified for Paper V to be able to calculate the scattering from lipids. Finally, I believe there is value in having several programs for doing the same, as long as they are developed independently, as these programs can be used as benchmarks for each other. Benchmarking CaPP towards Cryson e.g. lead to the correction of a bug in Cryson, which did not ignore explicit deuterium in the PDB files (<https://www.saxier.org/forum/viewtopic.php?t=3412>). The bug has been corrected in ATSAS 2.8.1.

### 3.3 Analytical treatment of partly aggregated samples

In the following, it is described how aggregations can be taken into account in the modelling of SAS data. The project is a collaboration with Jan Skov Pedersen from the department of chemistry at Aarhus University.

Protein aggregation is an important phenomenon in structural biology, and many proteins fibrillate in nature. Some proteins, like  $\alpha$ -synuclein (chapter 5 and Appendix B) form amyloids with a characteristic secondary cross- $\beta$  structure. In other aggregates, the local structure is conserved and the aggregates are thus oligomeric assemblies. Aggregation is however a major problem in structural analysis of single protein structures. In SAS, the intensity is proportional to the square of the particle volume (chapter 2), so even a minor fraction of fractal aggregates contributes significantly to the scattering, and may hinder correct conclusions to be drawn from data if not taken into account. This was the case for the SANS data on SERCA (Paper III), as well as for GluA2 (Paper IV). These aggregation contributions were taken into account by including a fraction of aggregates in the model. In those two cases, the aggregates were described in terms of a fractal structure factor. The usefulness of the method and the lack of a collected list of aggregation structure factors made us initiate a review on the topic (Paper I). As the approach is described in the paper, I will not describe it further here. I recommend reading the paper before continuing to the next section.

#### 3.3.1 "*In silico* sample purification"

The following is not included in Paper I, as it is a debatable strategy. I however think it deserves mentioning and is therefore included here. Biochemists in our group have jokingly used the phrase "*In silico* sample purification" about this method, because aggregation is taken care of in the analysis instead of in the lab. There is no doubt, that the samples should always be purified as well as possible before the experiment. However, in the following I will outline how "*In silico* purification" could be done using the aggregation descriptions from Paper I. I will use simulated data for the demonstration.

The scattering from a partly aggregated sample of monomeric protein is a sum of intensities from the proteins in monomeric and proteins in aggregated form:

$$I_{\text{meas}}(q) = I_{\text{mono}}(q) + I_{\text{aggr}}(q), \quad (3.3)$$

where  $I_{\text{meas}}(q)$  is the measured intensity, and  $I_{\text{mono}}(q)$  and  $I_{\text{aggr}}(q)$  are unknown. By fitting the model, the following quantities are found:

$$I_{\text{fit}}(q) = I_{\text{fit,mono}}(q) + I_{\text{fit,aggr}}(q). \quad (3.4)$$

The data can then be "*in silico* purified" to obtain a filtered dataset by:

$$I_{\text{filt}}(q) = I_{\text{meas}}(q) - I_{\text{fit,aggr}}(q). \quad (3.5)$$

The "filtered" intensity can then be used e.g. for *ab initio* structure determination, or Guinier analysis. To demonstrate the method, I used a simulated dataset of a sample of lysozyme with 10% 25-mer globular aggregates (Fig. 3.6A). The data were simulated as described in Paper I. The simulated data were then fitted using  $S_6(q)$  (see the paper) (Fig. 3.6B) to a  $\chi_r^2$  value of 0.92. The simulated data had a  $\chi_r^2$  of 0.87, so the fit was almost perfect. The aggregation part of the fit was then subtracted from the simulated data to obtain a filtered dataset [eqn. 3.5]. This filtered dataset is "purified" *in silico*, and can be treated as a monomeric dataset. For example,  $R_g$  can be determined by Guinier analysis (Fig. 3.6C). For the filtered data, an  $R_g$  of  $14.6 \pm 0.6$  was obtained, very close to the theoretical value of the crystal structure (14.5 Å), as calculated with CaPP. The data can also be fitted with hypothesized models in reciprocal space, as shown in Fig. 3.6D. Finally, *ab initio* modelling can be performed as shown in the bottom of Fig. 3.6. The danger is of course, that the filtered "data" is not real data, but relies on assumptions about both the structure of the single protein and the structure of the aggregates. So treating the filtered data as real data may lead to wrong conclusions. However, if used carefully and with precaution, it might be useful, and if compared to a simple truncation of data before further analysis, this filtering approach might constitute the a better option in some cases. There should be no doubt however, that optimal real life purification is by far the best option.

### 3.4 New tools, new possibilities

The tools presented in the current chapter have aided the study of membrane proteins with SANS. An automatic algorithm was implemented in Capp to add and fit a water layer around membrane proteins without adding water in the transmembrane region. CaPP also made it possible to calculate the  $p(r)$  directly from deposited protein structures in the PDB without any detour into reciprocal space. Finally, a method to take protein aggregation into account was outlined and tested, and a review with possible models for the aggregation was written (Paper I).

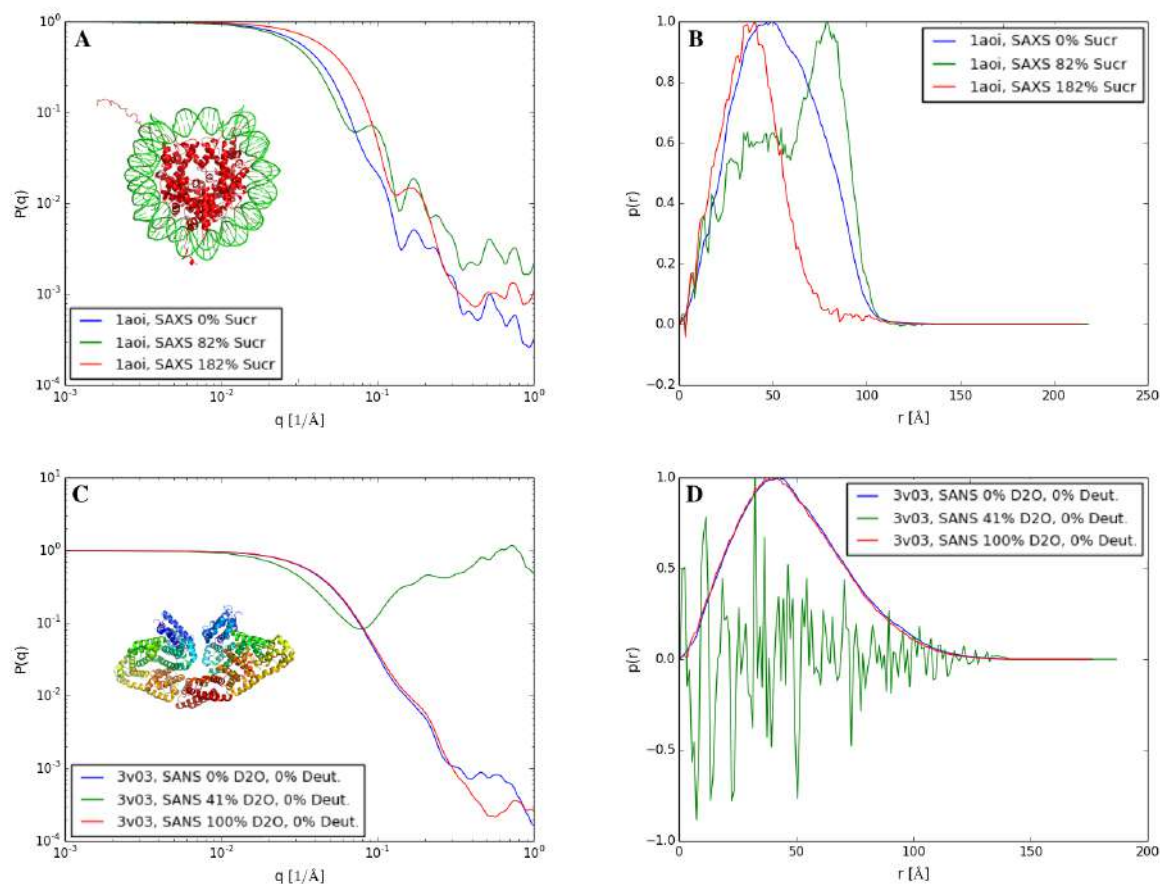


Figure 3.4: The plots are direct outputs from CaPP except for inserted structures and letters, and enlarged legends. (A) Calculated SAXS scattering and (B)  $p(r)$  for a protein-DNA complex (PDB: 1AOI) with different amounts of sucrose in the solvent ("%" meaning g/100ml). At 0 % sucrose (blue) both protein and DNA are "seen", at 82% sucrose (green) the protein is matched out, and at 182% sucrose (red), the DNA with higher electron density is matched out. The solvent is saturated at 200% sucrose, which is therefore an upper limit in the program. (C) Calculated SANS scattering and (D)  $p(r)$  for a dimer of bovine serum albumine (BSA; PDB: 3V03) at 0% D<sub>2</sub>O (blue), at 41% D<sub>2</sub>O (close to the match-point, green), and at 100% D<sub>2</sub>O. The curves are calculated for non-deuterated protein without a water layer.

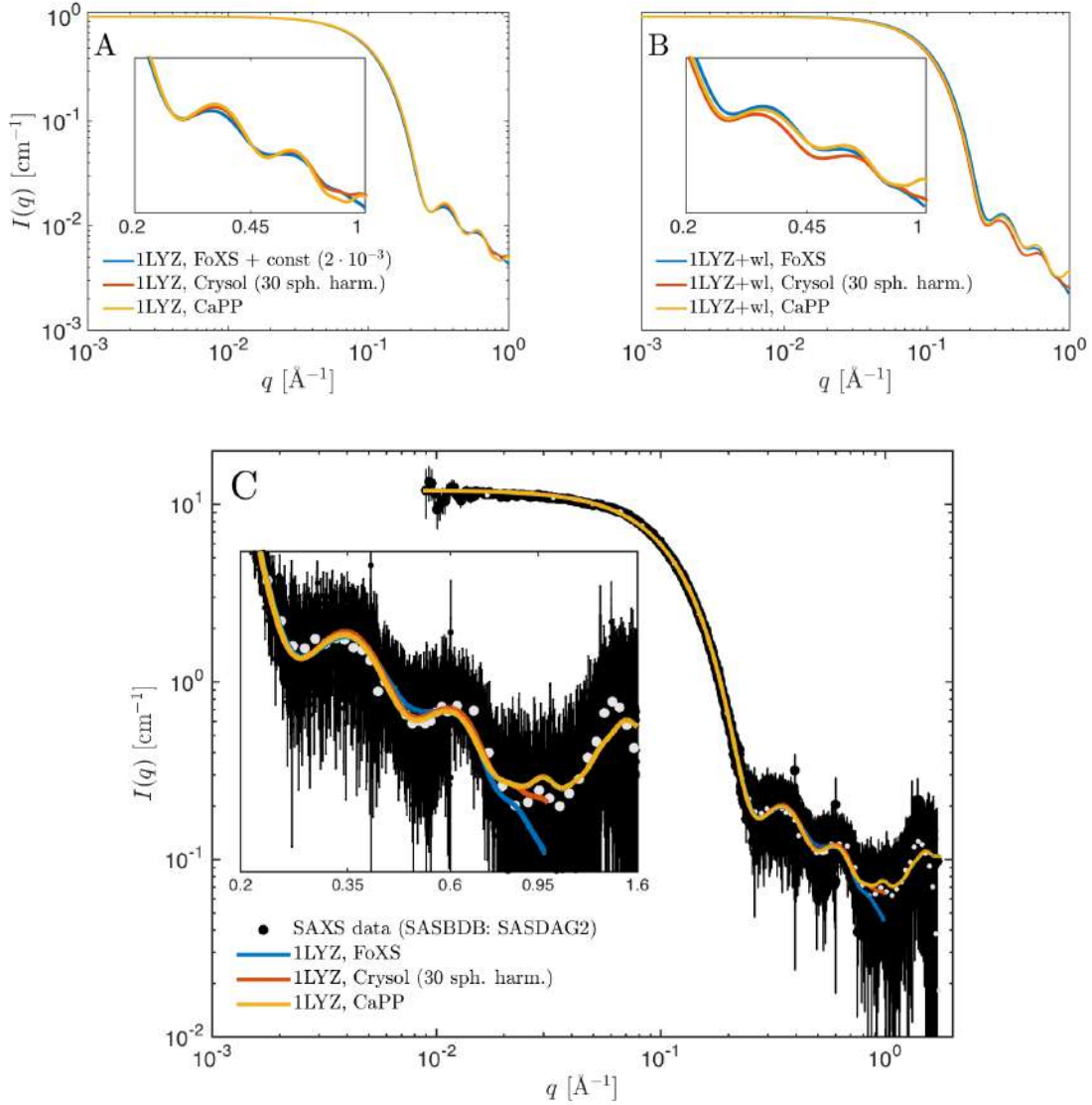


Figure 3.5: Theoretical scattering for lysozyme (PDB: 1LYZ) calculated with CaPP (yellow) as compared to the scattering calculated with Crysol (red) and FoXS (blue). (A) Theoretical scattering with default parameters, no water layer. (B) Theoretical scattering with water layer (wl), Crysol and CaPP with 10 % density increase, FoXS with  $c_2 = 1.0$ . (C) fit to SAXS data (SASBDB: SASDAG2; black). Rebinned data shown as white dots. Insert shows the data and fits between  $q = 0.2$  and  $1.6 \text{ \AA}^{-1}$ .

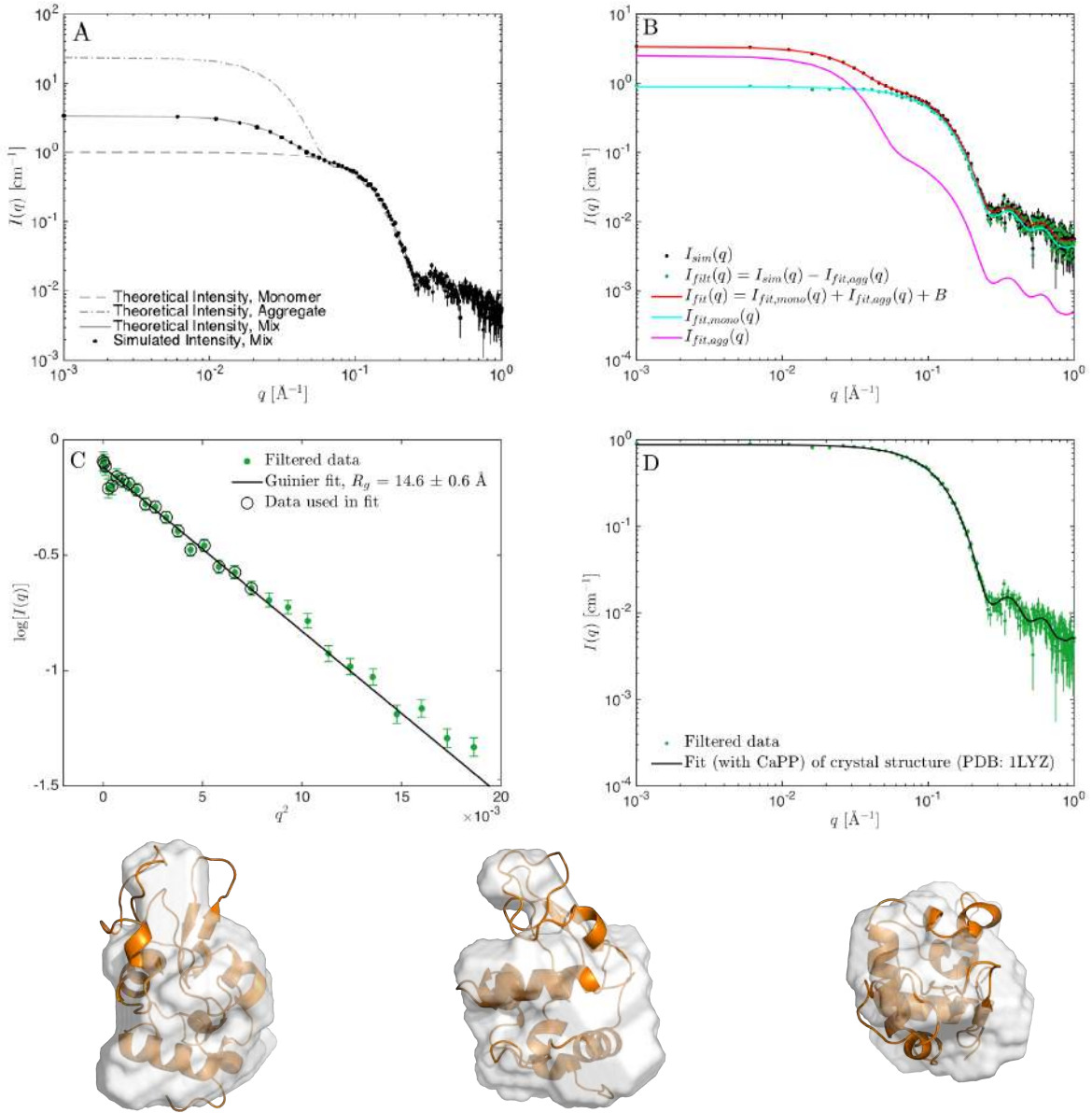


Figure 3.6: (A) Simulated data (black dots) with 90 % monomeric lysozyme (PDB: 1LYZ) and 10% globular aggregates. (B) Simulated data (black), fit (red) using the crystal structure and the structure factor  $S_6(q)$  (see Paper I for the aggregate description). The monomer component of the fit (cyan), the aggregate component of the fit (magenta), and the filtered data (green). (C) Guinier analysis on the filtered data, giving and  $R_g$  of  $14.6 \pm 0.6$  Å. The true  $R_g$  of lysozyme (PDB: 1LYZ) is 14.5 Å, as calculated with CaPP. (D) Fit of the filtered data with the crystal structure (using CaPP). Bottom: DAMMIF *ab initio* models generated from the filtered data. The filtered *ab initio* model, where "filtered" here refer to the model (Franke & Svergun 2009) is shown as a semi-transparent surface aligned to the crystal structure (PDB: 1LYZ).

## Chapter 4

# Statistical Methods for the Analysis of Small-Angle Scattering Data

*"Science may be described as the art of systematic over-simplification"*

*- Karl Popper*

Statistical methods serve to solve two main problems. Firstly, they should be used to evaluate if a given model/hypothesis is a good description of data given all available knowledge, and secondly, the methods should provide a measure to judge which of several alternative hypothesis/models is best.

This part of the thesis consists of the current chapter and Paper II, which describes how Bayesian statistics can be employed to include prior knowledge into the analysis of SAS data using analytical form factors. A short note on form factors can be found in Appendix A. I will give a notice when I recommend reading the paper.

### 4.1 Goodness of fit and generalizability

There are two main criteria upon which a model should be evaluated. One is the goodness of fit, i.e. how well does the model describe data. The second is the generalizability of the model, i.e. how well the model can describe a more general phenomena. The first is important to ensure that the model reflects the real world, and the second is essential if the model should aid our understanding and ability to navigate in the world.

The instrumentalist aspect of generalizability is easily understood for maps. What is a good map? The satellite maps provided by Google fit very well to data (the world). But how good is this model when we wish to take the train from Copenhagen central station to a party at a good friends place near Hundige station (a widely overseen train station southwest of Copenhagen)? We might get there at some point. But a simplistic overview of the local train stations provides a better model for this purpose (Fig. 4.1), despite that the goodness of fit is worse. This example represents the instrumentalist point of view: how we use a model to navigate.

The other aspect is the phenomenological perspective. How we understand the world. Fruits can provide a simple example. In order to describe a specific pear we might describe it very detailed. Give the exact size, in mm, the exact weight, in grams, describe the color at different positions of the surface, etc. This model would fit very well with observations of the fruit. We could however also describe the fruit with a more generic model. The model describes a rounded shape thicker in the button than in the top. The object is green and/or red, and it has a stem at the top. We will call this model "pear". The second

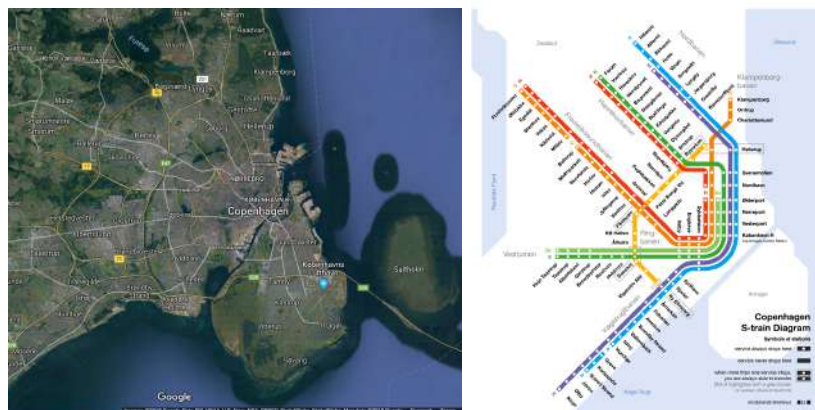


Figure 4.1: Two different representations of greater Copenhagen.

model has a worse goodness of fit for that particular dataset, but can describe the overall features of all (or at least most) pears in the world, whereas the first model only describes a specific pear. The second model may thus be a better model, despite having a worse goodness of fit. In some cases the model "pear" is however too simple, e.g. if one wishes to differentiate between pear varieties. Hence, there is an intrinsic conflict between simplicity on one side and accuracy on the other.

The goodness of fit can easily be quantified by comparing the model with the data, where closer resemblance gives a better goodness of fit. The demand for generalizability is similar to the Occam's razor (or "principle of parsimony"). That is, everything equal, the simplest model is preferred. A simple model may be a model with fewer parameters, or a model that is less controversial. Occam's razor is not as easily quantified as the goodness of fit.

## 4.2 The aim of the current chapter

In this chapter, I will discuss and suggest solutions to some core challenges in the analysis of SAS data:

- C1** How to evaluate whether a model is good? And how can Occam's razor and systematic deviations be taken into account?
- C2** How to compare models to find the most probable model?
- C3** In case of wrongly estimated experimental errors: How to correct them?
- C4** How can information from a dataset be combined with other knowledge, either from another experiments or from available prior knowledge?
- C5** How to determine the information content in SAS data?

I will discuss alternative solutions to these challenges based on frequentist and Bayesian statistics. Therefore, before I venture into discussing C1 to C5, I will give a very brief outline of the differences between Bayesian and Frequentist statistics.

## 4.3 The frequentist versus the Bayesian approach

Bayesian statistics and frequentist statistics are philosophically different. In a Bayesian perspective, the probability of an event describes a belief about that event. A new dataset can then update this belief. The



frequentists on the other hand, sees the probability as a frequency by which the event happens (Hamelryck 2012). As the frequentists ignore any influence of prior belief on the probability of the event, one can argue that this approach is the most objective. Supporters of a Bayesian approach can, on the other hand, argue that the belief should be included since the goal of scientific investigations is to update and refine our current belief.

**A short comment on the word "belief"** The word "belief" have associations to daily life, uncertain claims and to religion. Science strives for knowledge and certainty, and attempts to prove what is true or false. Real proofs however belong to the world of mathematics, and scientific knowledge always has a degree of uncertainty. So updated belief about the world is, in my opinion, what we obtain by scientific investigations. Probability can be increased, but true certainty can never be obtained. However, due to these associations, it is easily misunderstood when the word "belief" is used. For that reason, I have changed occurrences of "belief" and "believe" in Paper II, that discuss Bayesian statistics in the content of SAS, to "knowledge" and "know".

**A semi-Bayesian approach** A semi-Bayesian approach is often used in the analysis of SAS data, due to its ill-posed nature, where several solutions can explain data. Solutions that disagree with the prior belief are routinely discarded as unphysical, or limits are set up for the model parameters as explained in section 4.7. In some cases, parameters are even adjusted by hand to obtain a fit that is consistent with prior belief. As the discarded solutions and parameter limits are rarely (never?) reported, this gives a misleading impression that the solutions are objective and unique, despite being highly affected by the prior believe. I advocate for exposing the priors whenever they are used.

## 4.4 C1: Evaluating a model

This section discusses a range of methods to evaluate the goodness of fit of a model when compared to data. As shall be shown, some methods are complementary to the conventional methods as they highlight systematic errors or can be used when experimental errors are not available.

### 4.4.1 Evaluating a model using $\chi^2$ statistics

$\chi^2$  statistics is possibly the most widely used statistical tool in science. For the simple case of counts in a range of  $N$  bins,  $\chi^2$  is defined as:

$$\chi^2 = \sum_{i=1}^N \frac{(\mathbb{E}[M_{\text{theory},i}] - M_{\text{data},i})^2}{\mathbb{E}[M_{\text{theory},i}]}, \quad (4.1)$$

where  $\mathbb{E}[M_{\text{theory},i}]$  is the theoretical expectation value for number of counts in the  $i^{\text{th}}$  bin, given the underlying model.  $M_{\text{data},i}$  is the measured number of counts in the  $i^{\text{th}}$  bin. The  $\chi^2$  compares the measured data with the underlying theoretical model.  $\mathbb{E}[M_{\text{theory},i}]$  is unknown and is therefore approximated by experimentally determined values. In the nominator, it is approximated by the measured intensity:  $\mathbb{E}[M_{\text{theory},i}] \approx I_{\text{data},i}$ , and in the denominator by the experimental variance:  $\mathbb{E}[M_{\text{theory},i}] \approx \sigma_i^2$ . The variance,  $\sigma_i^2$ , is determined by counting statistics and error propagation. The altered  $\chi^2$  compares a fitted model with data, so  $M_{\text{data},i}$  [eqn. (4.1)] is replaced by  $I_{\text{fit},i}(\mathbf{p})$ :

$$\chi^2 \approx \sum_{i=1}^N \frac{(I_{\text{data},i} - I_{\text{fit},i}(\mathbf{p}))^2}{\sigma_i^2}, \quad (4.2)$$

where  $N$  is the number of data points. The fitted model  $I_{\text{fit}}(\mathbf{p})$  depends on  $K$  model parameters,  $\mathbf{p} = p_1, p_2, \dots, p_K$ .  $I_{\text{data}}$  and  $\sigma$  are the mean and standard deviation for the normally distributed estimate of the underlying intensity.

The best fit to data is found through minimization of  $\chi^2$  by varying the model parameters,  $\mathbf{p}$ . The reduced  $\chi^2$  gives a measure for the goodness of fit, and is defined as:

$$\chi_r^2 = \chi^2 / f \quad (4.3)$$

where  $f$  is the number of degrees of freedom. It follows the  $\chi_r^2$ -distribution with expectation value 1 and variance  $2/f$ . Hence, the  $\chi_r^2$  is expected to be close to unity if the model is good. The variation around the expectation value depends on  $f$ . Values significantly larger than unity indicate a wrong model, and values significantly below unity indicate that the model is overfitting data. The number of degrees of freedom is however not trivially determined, as discussed in section 4.8, but the conventional choice is  $f = N - K$ .

**Determining the probability for the model using  $\chi_r^2$**  In case the scientist has only one hypothesized model, it is important to be able to assess the validity of that model alone.  $\chi_r^2$  should be close to one, but how close? For example, if  $\chi_r^2 = 1.4$  is the model then "good enough"? This can be rephrased into a more precise question: Assuming that the model is true (the null-hypothesis), what is the probability of obtaining a  $\chi_r^2 \geq 1.4$ , i.e. a fit that is as bad or worse than the obtained? The  $\chi_r^2$  probability distribution is well-known, and depends on  $f$ , i.e. on  $N$  and  $K$ . Thus, the probability is expressed as  $P(\chi_r^2 \geq 1.4 | f)$ . We investigate case 1 where  $N_1 = 100$  and  $K = 5$  ( $f_1 = N - K = 95$ ) and case 2 where  $N_2 = 50$  and  $K = 5$  ( $f_2 = 45$ ). In Fig. 4.2, the probability distributions are shown for these cases. We will use a significance level of 1% for rejection of our null-hypothesis. The null hypothesis should be rejected in the first case ( $P(\chi_r^2 \geq 1.4 | f = 95) = 0.67\%$ ) but not in the second case ( $P(\chi_r^2 \geq 1.4 | f = 45) = 4.3\%$ ). Intriguingly, the evaluated probabilities are rarely given in research papers on SAS data. Usually, only the  $\chi_r^2$  are given (and are often denoted  $\chi^2$ ), despite the fact that the conclusions that can be drawn from data depends on  $f$ , as just demonstrated.

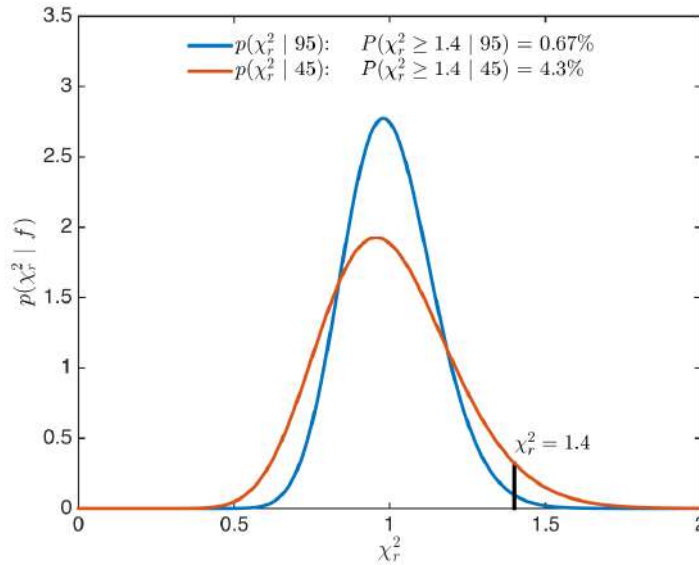


Figure 4.2:  $\chi_r^2$  probability distribution,  $p(\chi_r^2 | f)$ , for  $f = N - K = 100 - 5 = 95$  (blue) and  $f = N' - K = 50 - 5 = 45$  (red). The obtained  $\chi_r^2$  of 1.4 has been plotted as a vertical black line. The probabilities given  $\chi_r^2 = 1.4$  and  $f$  are listed in the legend.

**Residual plots** Residual plots can be used to find systematic errors in fits and it is recommended to plot these together with the fit (Trehwella *et al.* 2017). Systematic errors may in some cases be more evident than in the fit. The normalized residuals are given as:

$$(\Delta I/\sigma)_i = \frac{I_{\text{data},i} - I_{\text{fit},i}(\mathbf{p})}{\sigma_i}. \quad (4.4)$$

Each  $(\Delta I/\sigma)_i$  is expected to fall within a range of  $\pm 3$ , if the fitted model is true. Given the true model, only 3 out of 1000 data point would in general fall outside of this interval, in accordance with the normal distribution. It can also be visually checked, whether a range of fitted points are systematically above or below the data. The check for systematic errors using the residuals is a great tool, as systematic errors may not increase the  $\chi^2$  drastically if the deviation is small, but they are evident from the residual.

**When not to use  $\chi^2$**  In some situations, the  $\chi^2$  should not be used when comparing data and model. The first situation is when the data are not normal distributed. This is the basic assumption for using  $\chi^2$ . If the number of detector counts are small ( $<10$ ), then the normal distribution is a poor approximation for the underlying Poisson distribution, and maximum entropy should be used as the likelihood function. This is however rarely the case in SAS. In other techniques, such as triple axis spectroscopy it is however an issue. Also in health science, where  $N$  is often small, it is a relevant concern. The second situation is when the error bars are unknown or cannot be trusted, which I will discuss in the next section.

#### 4.4.2 Error bar independent evaluation of the goodness of fit

The error bars in SAS are derived from counting statistics, but propagated through the data reduction using a range of assumptions. This may in some cases lead to wrongly estimated errors. In case of wrongly estimated errors, the  $\chi_r^2$  can not be used to correctly evaluate the goodness of fit. Overestimated or underestimated errors can however be identified in several ways:

Overestimated errors can be identified prior to modelling, if the error bars are large compared with the fluctuations in data. IFT can likewise be used as a check. The final fit in inverse space should always fit data well, and go smoothly through the data points. As the data follow a normal distribution around the underlying model, every third of the error bars should *not* touch the fitted function. With overestimated errors, almost all error bars will however cross the fit. Moreover, the  $\chi_r^2$  obtained from the IFT should be close to unity. Much smaller values indicate underestimated errors, and *vice versa*. This will be discussed more thoroughly in section 4.6. Examining the normalized residuals for the IFT fit will give another clue. About 0.3% of the normalized residual points should, on average, have a magnitude larger than 3. Again, many points with magnitude larger than 3 indicate underestimated errors, and if the maximal magnitude is much less than 3, then the error bars are probably overestimated.

In the following I will outline and discuss some ways to evaluate the goodness of fit in the case of wrongly estimated error bars.

**Sign tests** If we assume that the fitted model is true, then the deviations between fit and model would stem only from random statistical variations. Each data point would have an equal probability of lying above or below the corresponding fitted value. We will use "+" for a point above and "-" for a point below. Having 10 points, we might e.g. get the following sequence of signs: - + - - - + + + -. There are 5 runs, i.e. 5 streaks of equal signs. The probability for the number of runs can be calculated and gives an error bar independent evaluation of the null-hypothesis that the model is true. This method is well-established and called "Wald-Wolfowitz runs test" (Wikipedia 2018; Barlow 1999). A related method was introduced in the so-called CorMap test by Franke *et al.* (2015) from the Svergun group (at EMBL Hamburg). The

CorMap test calculates the the probability of having a run equal to or longer than the longest run,  $C$ , given  $N$  data points,  $P(C' \geq C|N)$ . It assumes equal probability for  $+$  and  $-$ , and as the intensities are normal distributed, this is a good assumption. In the example above,  $C = 4$  and  $P(C \geq 4|10) = 0.24$  as can be derived from the binomial distribution (de Moivre 1738). The CorMap test has become a standard secondary test in SAS (Trehella *et al.* 2017). To my knowledge it is not established why it is preferred above the well-established and widely used runtest. The runtest is not even mentioned by Franke *et al.* (2015).

None of the sign tests are good "stand-alone" measures for the goodness of fit, as they ignore the distance between data and fit. However, they serve as a good measure for finding systematic errors. They can be considered as a way of summing the information from a visual inspection of the residuals into a single number. In that way they provide complementary information to  $\chi_r^2$  about the goodness of fit and are thus beneficial also when the error bars are correctly estimated.

**Coefficient of determination,  $R^2$**  Besides  $\chi^2$ , methods involving coefficient of determination,  $R^2$ , may be the most widely used tool for regression and evaluation of goodness of fit. It is e.g. implemented in Microsoft Excel.  $R^2$  is determined without any knowledge of experimental errors:

$$R^2 = 1 - \frac{\sum_{i=1}^N (I_{\text{data},i} - I_{\text{fit},i})^2}{\sum_{i=1}^N (I_{\text{data},i} - \langle I_{\text{data}} \rangle)^2} \quad (4.5)$$

where  $\langle I_{\text{data}} \rangle$  is the mean of the intensities. That is, if the fit goes through all data points,  $R^2 = 1$ , whereas if the fit is no better than the mean, then  $R^2 = 0$ . The output between 0 and 1 provides an appealing intuitive interpretation, analogous to  $\chi_r^2$ , namely that the model that fits best is the one with  $R^2$  closest to unity.  $R^2$  can be considered a comparison of the explained variance with the overall variance (deviation from the mean). When  $R^2 = 0$ , none of the variance is explained by the model, and when  $R^2=1$ , all of the variance is explained. In line with the inclusion of  $K$  in  $\chi_r^2$ , the adjusted  $R^2$ ,  $R_A^2$ , penalizes the use of many parameters:

$$\begin{aligned} R_A^2 &= 1 - \frac{\sum_{i=1}^N (I_{\text{data},i} - I_{\text{fit},i})^2 / (N - K)}{\sum_{i=1}^N (I_{\text{data},i} - \langle I_{\text{data}} \rangle)^2 / (N - 1)} \\ &= 1 - (1 - R^2) \frac{(N - 1)}{N - K}, \end{aligned} \quad (4.6)$$

where  $f = N - K$  is the degrees of freedom for the fitted model, and  $f = N - 1$  is the degrees of freedom for the mean. Thus,  $R_A^2$  has a build-in Occam factor, similar to  $\chi_r^2$ . The coefficient of determinant methods has some shortcomings. Firstly, noisy data will result in a poor  $R^2$  value despite fitting with the true, perfect model. The fit is penalized for not explaining the random variance. Secondly, and more importantly, when working with data spanning several orders of magnitude in intensity, the contribution to  $R^2$  from the high- $q$  data will be negligible as the absolute intensities are small. In  $\chi^2$  methods, the data are weighted with  $1/\sigma$ , and high- $q$  points with small absolute intensities have correspondingly small error bars (absolute, not relative). Therefore, when using  $\chi^2$ , the high- $q$  data will affect the goodness of fit despite of the inferior absolute intensities, as opposed to when  $R^2$  is used.

**The  $R^2$ -based F-test for evaluating a model** The so-called F-test can be used to compare the probability of two models. Here, we are interested in evaluating the probability of a single model. Therefore, following the logic of the  $R^2$ , the fitted model is compared to the mean,  $\langle I_{\text{data}} \rangle$ . In other words, the mean is the null-hypothesis, and we wish to evaluate the probability of obtaining an certain  $R^2$  value given that the mean describes the data well. If the probability is low (below some critical significance level), then the

model is a significantly better description than the mean and the null-hypothesis must be rejected. One must evaluate the  $F_0$  value for the model:

$$F_{0,R^2} = 1 - R_A^2. \quad (4.7)$$

Here the degrees of freedom is taken into account through the  $R_A$  value. The degrees of freedom for the model is  $f = N - K$  and the degrees of freedom for the mean is  $N - 1$ .  $F_0$  follows the  $F$ -distribution, which depends on the number of degrees of freedom for the to models, and by comparing  $F_0$  with the  $F$ -distribution, a probability for the model can be obtained.

### 4.4.3 Evaluating a model with Bayesian statistics

Bayesian statistics rely on Bayes theorem, which evaluates the probability of a model given data:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}, \quad (4.8)$$

where  $P(D|M)$  is the probability of the data given the model, also denoted the evidence,  $P(M)$  is the prior probability for the model, and  $P(D)$  is the probability for the data. The last term  $P(D)$  can only be determined for very simple examples, e.g. coin tosses, and is in practice a normalization constant to ensure correct normalization of the probabilities. Therefore, it is difficult to obtain a probability for the model in absolute units. In the Bayesian method given in Paper II, the  $\chi^2$  is used as the likelihood, i.e. as part of the expression for the evidence,  $P(D|M)$ . Therefore, the model can be evaluated using the usual methods based on  $\chi^2$  and  $\chi_r^2$ . I now recommend reading Paper II. The work in the paper was done in collaboration with Steen Hansen from our department. I also want to acknowledge discussions on the topic with Martin Cramer Pedersen.

### 4.4.4 Predictive/cross-validating methods.

No matter the measure for the evaluation of the goodness of fit, one can use prediction methods to ensure a more robust solution, and prevent overfitting. In the prediction methods, a part of the dataset is taken aside before any fitting, and the model is fitted to the remaining data. The model is then evaluated against the removed data to cross-validate the refined model. If the model overfits data, such than random noise is fitted, then the fit to the removed data will be poor, and the predicted goodness of fit will be much lower than the goodness of fit obtained in the usual manner. This method is a standard approach in many branches of science, e.g. in crystallography (Brünger 1992) and NMR (Brünger *et al.* 1993). Rambo & Tainer suggested a similar approach for SAS through the  $\chi_{free}^2$  (Rambo & Tainer 2013), which they showed was more robust for noisy data than the usual  $\chi_r^2$ . The predictive/cross-validating methods are very powerful to prevent misinterpretations and overfitting. However, predictive/cross-validation tools are cumbersome to implement in programs for fitting, and thus decrease the speed, so they are not widely used in SAS (yet?). Surprisingly, Franke *et al.* (2015) claim to have proved that there is no significant difference between  $\chi_r^2$  and  $\chi_{free}^2$ . I have not tested that myself, but as the method is widely accepted in a range of scientific fields, I strongly doubt that cross-validation has no effect for SAS data.

## 4.5 C2: Comparing alternative models

We have seen several methods to evaluate if a model is a good description of data. From methods based on  $\chi_r^2$ , over residual plots, to methods that can be used without knowledge of the experimental errors. The next question I will address is the assessment of competing models. How to find the most probable model.

**Occam's razor when evaluating models with  $\chi^2$  statistics** Inclusion of  $K$  in  $f$  ( $f = N - K$ ), penalizes models with many parameters. If two models fit a dataset equally well, say with  $\chi^2 = 100$  and  $N = 100$ , but Model A has 5 parameters and Model B has 10 parameters, then the  $\chi_r^2$  will differ and be 1.05 and 1.11 for Model A and B respectively. I.e. the simpler model is more probable. Thus Occam's razor is included in the  $\chi^2$  framework.

**The  $\chi^2$ -based F-test for comparison of models** If a complex model B fits data better than a simple model A, but only slightly, which is then the better? Say  $\chi^2 = 100$  and  $\chi_r^2 = 1.05$  for Model A and  $\chi^2 = 90$  and  $\chi_r^2 = 1.00$  for Model B respectively. The difference might be pure coincidence, and it might be significant. Luckily, the F-test can be used to answer that question. The null-hypothesis is, that the simpler model (Model A) is true. The quantity of evaluation is the ratio of the two  $\chi_r^2$  values,  $F_0$ :

$$F_0 = \frac{\chi_{r,1}^2}{\chi_{r,2}^2} \quad (4.9)$$

In this example,  $F_0 = 1.05/1.00 = 1.05$ . The value of  $F_0$  is  $F$ -distributed and depends on  $f$  for the two models. The output of the test is a probability of obtaining a value of  $F_0$  equal to or larger than the obtained, given that the models describe data equally well. In this example  $P = 28.5\%$ . It is thus very likely that the situation appears from pure coincidence, and the null-hypothesis cannot be rejected (with a significance level of e.g. 1 %). We used this test extensively in Paper II. The hypothesized models had different degrees of complexity as reflected in the number of parameters, and it was tested whether the more complex models fitted data significantly better than the simpler ones. If not, the simpler were believed. The  $F$ -test is unchanged upon a constant multiplied to the experimental errors. That is, the test will give the correct result despite that the errors are wrongly estimated (by a  $q$ -independent factor).

**The  $R^2$ -based F-test for comparison of models** In this context, we are interested in comparing two models with each other instead of comparing them with the mean:

$$F_{0,R^2} = \frac{1 - R_{A,1}}{1 - R_{A,2}^2} \quad (4.10)$$

By this operation, the mean of the intensity  $\langle I_{\text{data}} \rangle$  is removed from the equation, and we are left with a comparison of the residual for the two models, weighted with their respective degrees of freedom. This  $F_0$  can then be "translated" to a probability by comparison with the  $F$ -distribution for the given degrees of freedom. The  $F$ -test using the  $R^2$  corresponds to the  $F$ -test for the  $\chi^2$  but with all errors being unity. That is, as mentioned earlier, the low- $q$  data, with large absolute values are weighted much more than the high- $q$  data with low absolute magnitude.

**Comparing models with Bayes factor** Using Bayes formula [eqn. (4.8)], two models can be compared by the ratio of their probabilities:

$$R_{12} = \frac{P(D|M_1)P(M_1)}{P(D|M_2)P(M_2)} = \frac{P(D|M_1)}{P(D|M_2)} \times \frac{P(M_1)}{P(M_2)}, \quad (4.11)$$

where the first factor, the ratio of evidences, is called Bayes factor:

$$BF_{12} = \frac{P(D|M_1)}{P(D|M_2)}. \quad (4.12)$$

Bayes factor takes into the account the prior, as these are included in the evidence for each model. Intuitively, one would set the prior probability to be the same for the two models  $P(M_1) = P(M_2)$ , such that  $R_{12} = BF_{12}$ . It is however common that one model is less likely (more surprising) than the alternative. This is in line with the logic of significance levels. A significance level of 5% is analogous to a prior belief of  $P(M_1)/P(M_2) = 20$ , given that  $M_1$  is the null-hypothesis.

## 4.6 C3: Correcting wrongly estimated error bars

The goal of this section is to outline how to correct wrongly estimated errors. It is a severe problem having wrongly estimated error bars, as it leads to wrong  $\chi_r^2$  values and therefore wrong assessments of the fitted models. This can, to a certain degree, be compensated by use of error bar independent methods for evaluation of the goodness of fit, such as the  $R^2$  or sign tests, as discussed in section 4.4, but the assessment is more accurate if correct errors are available. Also, when combining data with prior information it is important that the errors on data are correctly estimated, in order to give the correct weight to data and prior respectively. As discussed in Paper II, underestimated experimental errors will give too much weight to the data, and too little to the prior, *vice versa*. Similarly, if several datasets are fitted simultaneously, it is crucial to have correctly estimated errors on all datasets in order to obtain the correct weighting between them.

**Correcting the error bars by redoing the data reduction.** The initial attempt should always be to find the reason for the wrongly estimated error bars and correct it. As the mistakes must lie in the error propagation through the data reduction, one should redo the reduction process, and consult the beamline scientist.

**Correcting the error bars by renormalization: generally a bad strategy** If a dataset is fitted with the true underlying model, then  $\chi_r^2 \approx 1$ . If the data had underestimated errors, say by a factor of 4, the  $\chi_r^2$  would, on average, be 2 times too large. It is therefore tempting to normalize the error bars with a factor  $\beta = \sqrt{\chi_r^2}$ , i.e.  $\sigma_{\text{new}} = \beta\sigma$ . The new error bars will look more reasonable. However, as described by Andrae (2010a), there are four reasons why this is not a good approach in general. These are given here (own formulation and numbering):

- A1** The experimental errors may not be Gaussian.
- A2** In non-linear models (most models used in SAS are non-linear), the number of degrees of freedom is not generally equal to  $N - K$ , since the model parameters may be correlated and thus not free (see section 4.8). That is, each parameter corresponds to less than 1 degree of freedom, and effectively  $f > N - K$ .
- A3** The model might not be the true model, even though visual inspection and residual plots indicate that it is. The model is generally unknown and the scope of the experiment is to test this model.
- A4** The  $\chi_r^2$  values follow a probability distribution (Fig 4.2) and it is therefore unlikely that  $\chi_r^2$  is exactly 1.0, even for a fit with the true underlying model (e.g. when fitting simulated data). By chance,  $\chi_r^2$  could be 1.4 or 0.9. Such deviations are not unusual, especially when  $N$  is small.

For these for reasons, correction of the error bars using  $\chi_r^2$  from a fit is generally a bad strategy.

**Normalization of error bars using Bayesian indirect Fourier transformation** In the following I will outline how the experimental errors can be corrected by a slightly adjusted normalization approach. In the Bayesian indirect Fourier transform (BIFT) algorithm implemented in BayesApp (Hansen 2012 and 2014) the scattering data are fitted to obtain the  $p(r)$ . Data are fitted under the constraint that  $p(r)$  is smooth (Glatter 1977; Hansen 2000). The fit has a resulting  $\chi_{r,B}^2$  ("B" for BIFT) which may be used for correction of the experimental errors. This was first proposed by Martin Pedersen in his thesis (Pedersen 2014):

$$\sigma_{c,i} = \sqrt{\chi_{r,B}^2} \sigma_i. \quad (4.13)$$

The method passes the first three criteria outlined by Andrae (2010a):

**Point A1:** SAS data are normal distributed, since  $N$  is sufficiently large.

**Point A2:** The model used in the BIFT is fully linear as it consist of a sum of points multiplied by each a scalar coefficient (Hansen 2000). However, the number of effective free parameters is smaller than the number of points. This is because the points are correlated via. the regularization term, such that each point cannot be chosen freely. The number of effective, free parameters (good parameters,  $N_g$ ) can however be determined in the Bayesian approach (Gull 1989; Vestergaard 2006), using the regularization parameter  $\alpha$ . The number of free parameters is then well-determined as  $f = N - N_g$ .

**Point A3:** The BIFT uses a generic model and is therefore (approximately) true for all SAS data.

The fourth criterion is however problematic:

**Point A4:** The statistical variance of  $\chi_r^2$  means that the value of  $\chi_r^2$  for a specific dataset rarely equals exactly unity.

**Renormalizing error bars with BIFT demonstrated with simulated data** I will in this example renormalize errors on simulated data on detergent micelles. A detergent micelle is a self-assembled particle (Fig. 4.3A). The single detergents form (more or less) spherical particles with the hydrophobic tail groups pointing towards the center and the hydrophilic headgroups facing the water. For the simulations I used a mathematical model for the micelles, with theoretical intensity  $I_{\text{mod}}(q)$ . They were described by a core-shell oblate ellipsoid. I simulated 1000 datasets with relative noise of 4% and absolute noise of 0.001 (Fig. 4.3A). That is,  $\sigma_{\text{sim}} = 0.04 \cdot I_{\text{mod}}(q) + 0.001$ , and the simulated data were sampled from a normal distribution with mean  $I_{\text{mod}}$  and standard deviation  $\sigma_{\text{sim}}$ . Each simulated dataset had a corresponding noise level, i.e. a value of  $\chi_{r,s}^2$  ("s" for simulated). This  $\chi_{r,s}^2$  was found by comparing  $I_{\text{sim}}(q)$  with  $I_{\text{mod}}(q)$  (with zero parameters, i.e.  $f = N$ ). The simulated data were indirect Fourier transformed to obtain the  $p(r)$  using the BIFT algorithm in BayesApp, and the resulting  $\chi_{r,B}^2$  were monitored and compared with  $\chi_{r,s}^2$  for each dataset. This comparison is seen in Fig. 4.3B. The values of  $\chi_{r,s}^2$  were reproduced within a few percent, when using  $f = N - N_g$  (green dots in Fig. 3.6B). This shows that  $N_g$  is a good estimate for the number of effective free parameters. Note the minor systematic discrepancy, that  $\langle \chi_{r,B}^2 \rangle = 0.97 \pm 0.01$  and  $\langle \chi_{r,s}^2 \rangle = 1.01 \pm 0.01$ , so there is a minor but significant discrepancy of about 4%.

This may be due to minor numerical inaccuracies in the program, but the origin is unknown so far. Fig. 4.3B clearly illustrates Andrae's last point (A4): if the error bars are scaled with  $\chi_{r,B}^2$ , then the datasets that by chance have  $\chi_{r,s}^2 > 1$  would erroneously get their error bars enlarged. Likewise, data with  $\chi_{r,s}^2 < 1$  would erroneously get reduced error bars. If fitted with the true model (in this case an oblate micelles) after rescaling of the error bars, we would consequently not expect the obtained  $\chi_{r,c}^2$  ("c" for corrected) to follow a  $\chi_r^2$  distribution, but generally be much closer to unity (Fig. 4.4), since:

$$\chi_{r,c}^2 = \chi_{r,s}^2 / \chi_{r,B}^2 \approx 1. \quad (4.14)$$

The corrected quantity  $\chi_{r,c}^2$  is thus not a true reduced  $\chi_r^2$ . But it shares an important property with  $\chi_r^2$ , namely that a good fit results in a value close to unity, and a less good fit results in a value larger than unity. Even though this would in practice be good enough for many applications, it is not satisfying that the probability for a hypothesis with respect to the null-hypothesis can not be calculated following the procedure outlined in section 4.4. Therefore, I outline how to empirically make a corrected  $\chi_{r,k}^2$  that approximately follow the  $\chi_r^2$  distribution. By comparing the distributions for  $\chi_{r,s}^2$  and  $\chi_{r,c}^2$  (Fig. 4.4), we see that they differ in width and in a slight shift for the mean value. The shift in mean value correspond to the above mentioned  $\sim 4\%$ . By correcting for the 4% with a correction constant  $k_1$ , the shift of the mean is corrected. By adding to  $\chi_{r,c}^2$  a number randomly sampled from a normal distribution with mean  $\mu = 0$



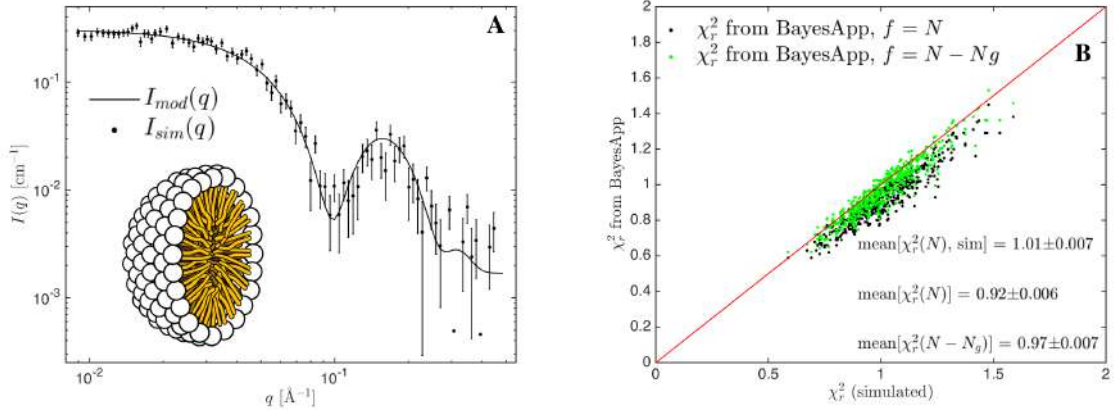


Figure 4.3: (A) Theoretical scattering for oblate micelles ( $I_{\text{mod}}(q)$ ; full line) and simulated data generated from the model ( $I_{\text{sim}}(q)$ ; error bars). Data with error bars exceeding values below zero are plotted as points. Insert shows a micelle (from Wikipedia commons). (B)  $\chi_r^2$  from BIFT, as implemented in BayesApp, with  $f = N$  (black) and  $f = N - N_g$  (green), plotted as function of the  $\chi_r^2$  values of the simulated data ( $f = N$ ).

and width  $\sigma = k_2$ , the distribution  $\chi_{r,k}^2$  is obtained, with corrected width. That is:

$$\chi_k^2 = \chi_{r,c}^2 \cdot k_1 + \mathcal{N}(0, k_2), \quad (4.15)$$

where  $\mathcal{N}(0, k_2)$  is the normal distribution with mean 0 and standard deviation  $k_2$ .  $k_1$  and  $k_2$  are empirical constants, and from the simulations I obtain:  $k_1 = 1.03835$  and  $k_2 = 0.131$ .

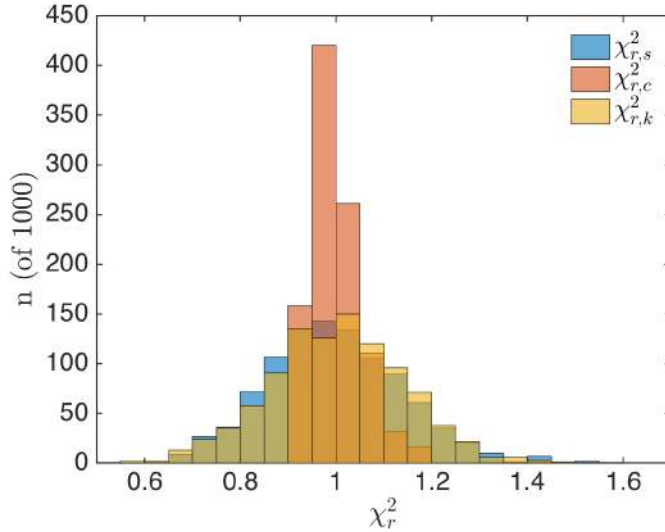


Figure 4.4: Histogram for  $\chi_{r,s}^2$  (from simulations; blue),  $\chi_{r,c}^2$  (corrected; red) and  $\chi_{r,k}^2$  (using correction constants  $k_1$  and  $k_2$ ; yellow).  $\chi_{r,s}^2$  follow a  $\chi_r^2$  distribution whereas  $\chi_{r,c}^2$  follow a narrower, unknown distribution.  $\chi_{r,k}^2$  is an empirical approximation of  $\chi_{r,s}^2$ .

As judged by visual inspection (Fig. 4.4),  $\chi_{r,k}^2$  is a good approximation of  $\chi_{r,s}^2$  and thus follow the  $\chi_r^2$  distribution. The errors should also be corrected with  $k_1$  and  $k_2$  by:

$$\sigma_k = \sqrt{\chi_{r,k}^2} \cdot \sigma_s. \quad (4.16)$$

If we only look at the effect of  $k_2$ , the error bars are perturbed slightly and randomly. This seems counter-intuitive, as there is no reason to believe that the new errors are closer to the true ones. The goal with  $k_2$  is therefore to obtain the correct distribution for  $\chi_r^2$  in order to draw the correct conclusions after statistical analysis, and avoid misinterpretations of data. The renormalization of the simulated micelle data shows that the BIFT algorithm quite exactly finds  $\chi_{r,s}^2$ , i.e.  $\chi_{r,B}^2 \approx \chi_{r,s}^2$ . So renormalization works well, when (by chance)  $\chi_{r,sim}^2 = 1$ . However, the example also illustrates the problem of point A4: data that, by chance, has a true  $\chi_r^2$  below or above 1 will be normalized incorrectly, and the corrected values  $\chi_{r,c}^2$  will not follow a  $\chi_r^2$  distribution (Fig. 4.4). It may be possible to determine general empirical values for  $k_1$  and  $k_2$  [eqn. (4.15)] to obtain a correct distribution for  $\chi_{r,k}^2$ .  $k_1$  is probably inherent for BayesApp. It may be possible to find the cause for the discrepancy  $k_1$  represents and correct it, thus getting rid of  $k_1$ .  $k_2$  also depends on the program. If BayesApp reproduced the  $\chi_{r,s}^2$  perfectly, then  $\chi_{r,c}^2$  would not be a distribution but be unity for all simulated data. The constants may also depend on the number of data points, the noise of data and the underlying model. This is still left to be investigated.

## 4.7 C4: Combining data with prior knowledge

Finding a 3-dimensional structure from SAS data is an ill-posed problem. This is because of loss of phase information, orientational averaging, noise of data, and a limited available  $q$ -range for the measured data (as discussed in chapter 2). Therefore, several models will fit to data (example 4.7.1). The prior knowledge of the system must be used to choose only the physical relevant model(s).

**Molecular constraints** Molecular constraints are constraints on the model parameters based on prior knowledge about the system. They can thus be used to include prior knowledge in the model. For example, a detergent micelle may be modelled by a core-shell particle (Fig. 4.3A) with a volume for the core and a volume for the shell. The core represents the detergent tails, and the shell the detergent heads. Both are related to the number of detergents in the micelle (the aggregation number,  $M_{agg}$ ). The head group volume, and the tail group volume of a single detergent has been found experimentally. Thus,  $M_{agg}$  can be fitted, and the volume of the core and shell determined from  $M_{agg}$ . Thereby, it is ensured that the model is physically meaningful, and in this example the number of parameters is also reduced. This is described in more detail and with relevant references in Paper II. Molecular constraints can also be introduced as limits on each parameter, such that the refined value is constrained to a certain interval. That is implemented in much software for analysis of SAS, as explained in Paper II. The limits can be introduced in a statistical framework as so-called prior distributions for each parameter. In the case of simple hard limits for the parameters, the prior is uniform in a limited interval and zero outside. For parameter  $\kappa$  constrained to the interval  $[a, b]$  the probability density  $p(\kappa)$  takes the form:

$$p(\kappa) = \begin{cases} s & \text{if } a \leq \kappa \leq b, \\ 0 & \text{otherwise,} \end{cases} \quad (4.17)$$

where  $s = (b - a)^{-1}$  such that the total probability is unity,  $P_{tot} = \int_{-\infty}^{\infty} p(\kappa) d\kappa = 1$ . Bayesian statistics is the proper statistical framework for including such priors, which leads to the next section.

### Example 4.7.1 The ill-posed nature of structure determination in SAS

CorA is a membrane proteins that transports magnesium. Its structure is debated, and Nicolai Johansen from our group has investigated its structure with SAXS and SANS.

We managed to obtain SANS data with the "invisible" detergents described in Paper III and in chapter 5, so the detergents in the sample did not contribute to the SANS signal (Fig 4.5). The crystal structure of CorA with  $\text{Mg}^{2+}$  did not fit data (magenta in Fig 4.5, inset). At that stage of the project, we tried several models to fit the data. The best of these models included four components: (1) the unperturbed crystal structure (magenta in Fig. 4.5, inset), (2) a derived structure with broken symmetry (green in Fig. 4.5, inset), (3) dimer of the crystal structure, (4) dimer of the derived structure. The 4-component model fitted fairly well to data as judged from visual inspection ( $\chi_r^2 = 18$ ; Fig. 4.5, red line). I also fitted data with a simple elliptical cylinder, which fitted data slightly better than the four-component model ( $\chi_r^2 = 16$ ; Fig. 4.5, inset and gray line). However, we know for sure that the CorA has not formed elliptical cylinders, and the cylinder model was therefore immediately rejected.

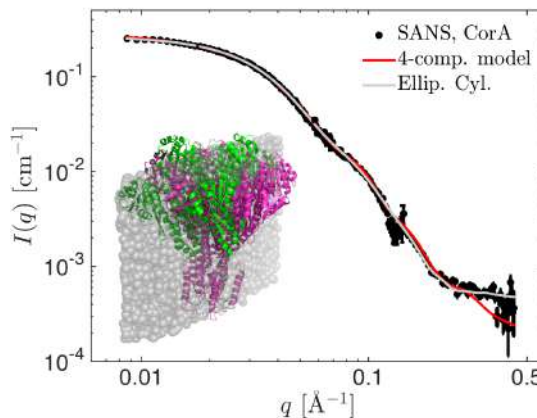


Figure 4.5: SANS data of a sample of CorA (black), fitted with a 4-component model of CorA structures (see text, red). and with a highly elliptical cylinder (gray). Inset shows the two monomer components of the 4-component model (magenta and green), and the cylinder model (gray).

**Inclusion of prior knowledge using Bayesian statistics** In Paper II, we describe how Bayesian statistics allows for direct inclusion of prior knowledge about the model parameters in the refinement process. The prior knowledge is given as a sum of probability distributions  $S$ , and the minimization of  $\chi^2$  is replaced by a minimization of:

$$Q = \chi^2 + \alpha S, \quad (4.18)$$

where  $\alpha$  is a regularization parameter that weights the data via  $\chi^2$  and the prior via  $S$ . The optimal value for  $\alpha$  is determined in an automatic and statistical sound way, such that any subjective choices for  $\alpha$  is avoided. In Paper II, we show for two experimental examples that the method gives a solution that balances the two terms well. We obtain a goodness of fit that is almost the same as that obtained without inclusion of the prior, but the refined parameter values are generally closer to the prior values. Moreover, the method gives more realistic error estimates on the refined values. By fitting to simulated noisy data it is shown how the regularization stabilizes the solution and ensures a meaningful set of refined parameters even for very noisy data.

**Combining the information from several datasets by simultaneous fitting** One of the major advantages of SANS is that the same system can be investigated in different contrast situations by exchange of the  $\text{D}_2\text{O}/\text{H}_2\text{O}$  content in the solvent. That is, the sample is ideally unchanged, but different part of the sample can be highlighted. A DNA-protein complex, for example, can be measured at  $\sim 40\%$

D2O where the protein is matched out, at  $\sim 70\%$  D2O where the DNA is matched out, and at  $\sim 100\%$  D2O where both DNA and protein have non-zero contrast (Fig. 3.4 in chapter 5). By measuring the same sample with X-rays, an additional contrast is obtained.

The focus here is on how to optimize the information gained from a simultaneous fit to a series of measurements. All datasets are fitted with the same model. A few parameters may change for the different contrasts, e.g. concentration and incoherent background. But the certainty about the common parameters describing the structure of the biomolecule increases for each new measured contrast. Presently, using frequentist statistics, the model is refined by minimization of the total  $\chi^2$ . With  $M$  datasets, it is simply given as:

$$\chi_{\text{tot}}^2 = \sum_{i=1}^M \chi_i^2, \quad (4.19)$$

where it is assumed that the experimental errors are correctly determined for all  $M$  datasets. Else they should be corrected, as previously discussed here and by Pedersen in his thesis (Pedersen 2014). In the Bayesian approach, we wish to include the prior as well. Moreover, regularization parameters,  $\alpha_i$  [eqn. (4.18)], should be included to ensure that the maximum amount of information is retrieved from the data. When weighing a total of  $N_f$  functions,  $N - 1$  regularization parameters should also be included. In this case,  $N_f = M + 1$ , i.e. the  $M$  datasets and the prior function. So  $M$  regularization parameters should be included in the minimization:

$$Q = \sum_{i=1}^M \alpha_i \chi_i^2 + S. \quad (4.20)$$

It is, in principle, possible to determine several  $\alpha$ -parameters. In BayesApp, the Powell algorithm (Powell 1964) is implemented to optimize 2-4 hyper parameters (Hansen 2000 and 2014), and the algorithm is able to handle more parameters (Powell 1964). This is however computationally expensive and may be unstable when many datasets are included measured, which is often the case in SANS (see e.g. Arleth 2001).

A practical solution is therefore to use the same regularization parameter for all datasets measured under the same conditions. E.g. for  $M - 1$  SANS dataset and a single SAXS dataset, the fitted model should be refined by minimizing:

$$Q = \alpha_n \sum_{i=1}^{M-1} \chi_{n,i}^2 + \alpha_x \chi_x^2 + S, \quad (4.21)$$

where subscript  $n$  and  $x$  indicate SANS and SAXS respectively. In that way, the number of hyperparameters is limited to a feasible number.

This method is yet to be tested, but is promising, as the inclusion of prior knowledge (via.  $S$ ) and optimal weighing of the data (via. the  $\alpha$  parameters) can help optimizing the information obtained from SAS data.

## 4.8 C5: The number of degrees of freedom and the information content in data

This last section deals with the number of degrees of freedom in SAS data and in models. The number of degrees of freedom are used to evaluate the information content in data, and they are considered when evaluating how well a model fits data (C1) and when finding the most probable out of several alternative models (C2).

The number of degrees of freedom are not always trivially determined, especially for non-linear models (Andrae 2010b). The most conventional choice is to use  $f = N - K$ . In SAS the approximations  $f = N$  and  $f = N - 1$  are frequently used, e.g. in Crysol (Svergun *et al.* 1995) and FoXS (Schneidman-Duhovny *et al.* 2010 and 2013). Even in a recent statistical paper (Franke *et al.* 2015),  $f = N - 1$  is used. A wrong choice of  $f$  however gives inaccurate values of  $\chi_r^2$ , and the expectation value of the resulting distribution does systematically differ from unity, as demonstrated in example 4.8.2. Tanner & Rambo (Tanner & Rambo 2013, supplemental) use  $f = N + 1 - K$  as they add one degree of freedom to the data due to the free choice of PDB model used for the simulations. I do not agree in that reasoning. Following the same reasoning, one should also add one degree of freedom to the model for choosing the same PDB again, and we are back to  $f = (N + 1) - (K + 1) = N - K$ .

As demonstrated in example 4.8.2, the expression for  $f$  is of little importance when  $N \gg K$ , as  $N - K \approx N - 1 \approx N$ . But when  $N$  and  $K$  are comparable in size, it is important to include  $K$  in the expression for  $f$  when evaluating  $\chi_r^2$ .

$N - K$  is a good approximation for  $f$  in example 4.8.2 because the parameters  $A$ ,  $B$ , and  $R$  are only weakly correlated. In a model with correlated parameters, the degrees of freedom for the model is significantly smaller than  $K$ . I will return to this issue in section 4.8.

Programs such as Crysol and FoXS have 3-5 free parameters, depending on the chosen options, and  $K$  could easily be included to get a better estimation of the  $\chi_r^2$ .

**The degrees of freedom of model versus the degrees of freedom of data** To give a good approximation for  $f$ , we must understand better what the degrees of freedom is. A dataset has a given number of degrees of freedom,  $f_d$ , which equals the number of data points,  $f_d = N$ . A model also has a number of degrees of freedom,  $f_m$ . For example, the linear model  $ax + b$  has two degrees of freedom, one for each parameter. In that case  $f_m = K$ , where  $K$  is the number of parameters. The total degrees of freedom of a fitting problem  $f$ , as given in the definition of the  $\chi_r^2$  is the number of degrees of freedom of the data minus the degrees of freedom in the model,  $f = N - f_m$ . The excess degrees of freedom, so to say. When a model is nonlinear, the model parameters are often more or less correlated, and the effective number of degrees of freedom for the model is less than the number of parameters in the model.

**Maximum information retrievable from data** There is a maximum of parameters that can be obtained from a given data set. The maximal information that can be retrieved from a signal is limited by the bandwidth, as derived by Shannon (1949). Due to the Fourier theory underlying SAS, this can be applied in the context of SAS, as done by Damashcun *et al.* (1968) and later by Taupin and Luzatti (1982). The maximal number of parameters deduced by a SAS dataset is given by the number of Shannon channels:

$$N_s = q_m D_{\max} / \pi, \quad (4.22)$$

where the maximum measured value of  $q$ ,  $q_m$  defined the bandwidth, and  $\pi/D_{\max}$  is the width of each Shannon channel. Thus, the first Shannon channel is at  $\pi/D_{\max}$ . It is assumed that  $q_m \leq \pi/D_{\max}$ , i.e. the first Shannon channel is part of the data. No matter how large  $N$  is, the dataset can be used only to refine  $N_s$  free parameters in a given model.

Equation (4.22) is the definition of  $N_s$  most commonly used, e.g. by Moore (1980), who used it in the context of IFT and Rambo & Tainer (2013) who used it to propose an improved measure for the goodness of fit, the  $\chi_{\text{free}}^2$  (see also section 4.4.4). Taupin & Luzatti (1982) however argue that  $q_{\max}$  should not be the largest measured value of the scattering vector, but the maximal value that can not be described by a simple Porod decay,  $I = Aq^{-4} + B$ , where  $A$  is a scaling parameter and  $B$  is a constant. Interestingly, Taupin & Luzatti (1982) therefore propose another definition of  $N_s$ . They denote it the number of degrees

of freedom (of the model)  $J$ :

$$J = q_m D_{max} / \pi + 2, \quad (4.23)$$

with two degrees of freedom of the Porod plot for large  $q$ -values.  $q_m$  is the maximum  $q$ , before the data can be adequately described by the Porod law.

#### Example 4.8.2 Using wrong values for the degrees of freedom

To demonstrate that a wrong choice of  $f$  may lead to wrongly estimated values of  $\chi_r^2$ , I simulated a SAS dataset of identical, monodisperse spheres, using the sphere form factor  $P_S(q, R)$  (Appendix A), with radius  $R = 50 \text{ \AA}$  (Fig 4.6).

A scaling parameter  $A = 0.5$  and a background  $B = 0.001$  were also included in the model:

$$I_{\text{mod}}(q) = A \cdot P_S(q, R) + B. \quad (4.24)$$

Noisy datasets were simulated with  $\sigma_{\text{sim}}(q) = 0.01 \cdot \sqrt{I_{\text{mod}}(q)} + 0.0001$ . The simulated data were sampled from a normal distribution with mean  $I_{\text{mod}}(q)$  and standard deviation  $\sigma_{\text{sim}}(q)$ . The  $\chi_r^2$  values of the simulated datasets were calculated by comparing  $I_{\text{mod}}(q)$  with  $I_{\text{sim}}(q)$  without any fitting,  $\chi_{r,s}^2$ . The data were then fitted with the true model, using the Levenberg-Marquardt minimization algorithm (Levenberg 1944; Marquardt 1963), with  $A, B$  and  $R$  as free parameters. The  $\chi_{r,f}^2$  values obtained from the fits were calculated using respectively  $f = N$ ,  $f = N - 1$  and  $f = N - K$ . 1000 datasets were simulated with respectively 10, 100 and 1000 data points ( $N$ ). To evaluate the choice of  $f$ ,  $\chi_{r,f}^2$  values were compared with  $\chi_{r,s}^2$ . These should be the same, for the correct choice of  $f$ .

When using  $f = N$  the differences between  $\chi_{r,f}^2$  and  $\chi_{r,s}^2$  were  $29.0 \pm 0.6\%$  for  $N = 10$ ,  $2.95 \pm 0.07\%$  for  $N = 100$ , and  $0.305 \pm 0.008\%$  for  $N = 1000$ . When using  $f = N - 1$  the differences were smaller, namely  $21.1 \pm 0.6\%$ ,  $1.97 \pm 0.07\%$ , and  $0.205 \pm 0.008\%$  respectively. Using Rambo & Taylor's  $f = N + 1 - K$  ( $f = N - 2$ ) gives even smaller discrepancies,  $13.5 \pm 0.7$  for  $N = 10$ ,  $0.96 \pm 0.07$  for  $N = 100$  and  $0.10 \pm 0.07$  for  $N = 1000$ . Finally, when using  $f = N - K$  ( $f = N - 3$ ), the differences were insignificant, namely  $1.4 \pm 0.8\%$ ,  $0.05 \pm 0.08\%$ , and  $0.005 \pm 0.008\%$  respectively. This demonstrates that  $f = N - K$  is a good approximation for the degrees of freedom in this case, in accordance with usual conventions (see e.g. Taylor 1997).

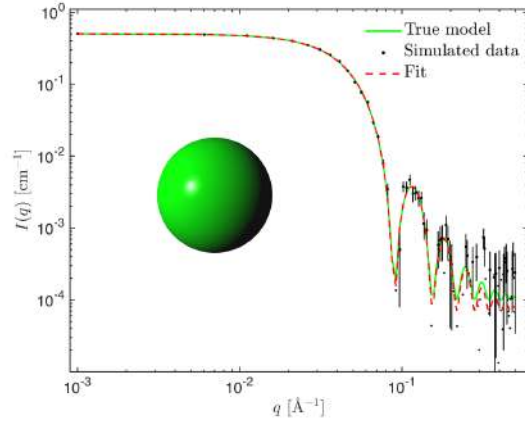


Figure 4.6: Simulated data (black) of diluted spheres with radius  $R$  (green), fitted with the same model (red). (Picture from <https://pixabay.com/en/sphere-ball-plastic-round-3d-953964/>, no copyright).

Konarev & Svergun (2015) suggested a similar approach. Here data is truncated at a certain  $q_m$ , where the data is too noisy to add any information. Thereby, the noise in data is taken into account. This measure is denoted  $M_S$ :

$$M_S = \pi q_m / D_{\text{max}}. \quad (4.25)$$

$M_S$  can be calculated with the program, Shannum, which is part of the ATSAS software package (<https://www.embl-hamburg.de/biosaxs/manuals/shanum.html>).

In the Bayesian context, another expression for the maximum number of retrievable parameters is given, namely the number of good parameters  $N_g$ .

**The maximum number of degrees of freedom found with the Bayesian method**  $N_g$  is explained in Paper II. But as it is only discussed briefly, and is not easily grasped, I will give an explanation here as well. I will try to give an intuitive explanation, that only demands vague memories of long passed linear algebra lessons. In order to determine  $N_g$ , it is examined how large the eigenvalues of the curvature matrix of  $Q$  is. This matrix is denoted  $\mathbf{C}$  (following the notation in Paper II).  $\mathbf{C}$  gives the correlation between different parameters in the model, such that element  $C_{i,j}$  is the correlation between parameters  $p_i$  and  $p_j$ , etc. The elements of  $\mathbf{C}$  are normalized with the prior widths to obtain unitless quantities  $C_{i,j} \rightarrow C_{i,j}/(\delta p_i \delta p_j)$ . To find the eigenvalues, the normalized  $\mathbf{C}$  is diagonalized,  $\mathbf{C} \rightarrow \tilde{\mathbf{C}}$ .  $\tilde{\mathbf{C}}$  is then the (unitless) covariance matrix for a new set of parameters that describe the model. These parameters,  $\tilde{p}_1, \tilde{p}_2, \dots, \tilde{p}_K$  are unphysical. The eigenvalues are the diagonal elements of  $\tilde{\mathbf{C}}$  and are called  $\lambda = \lambda_1, \lambda_2, \dots, \lambda_K$ . Some of the eigenvalues are large, compared to the regularization parameter,  $\alpha$ , and some are small. The interpretation is, that the large values have great impact on the model, whereas the small eigenvalues have little impact. Each  $\lambda_i$  adds to the total value of  $N_g$  depending on its magnitude relative to  $\alpha$ :

$$N_g = \sum_{i=1}^K \frac{\lambda_i}{\lambda_i + \alpha}. \quad (4.26)$$

Clearly, each eigenvalue can at add a maximum of one to the sum, so  $N_g \leq K$ , which makes sense as the correlation between parameters reduce the effective number of parameters in the model. If, and only if all parameters are fully uncorrelated, then  $N_g = K$ . If the parameters are fully correlated, then  $N_g = 1$ , e.g.  $y(x) = a \cdot b \cdot x$ , where  $K = 2$ , but  $N_g = 1$ . Moreover, if  $\alpha$  is very large, which is the case for noisy data, then  $N_g$  will also decrease, as shown in Paper II. Noise in data will also reduce  $N_g$ , as  $\alpha$  gets larger.

In Paper II we derive, for the first time,  $N_g$  in the context of models with analytical form factors.  $N_g$  was described about 30 years ago by Gull (1989) in the context of image reconstruction and introduced for SAXS in the context of BIFT by Hansen (2000). Vestergaard & Hansen (2006) showed how  $N_g$  gave a good measure for the information content in data, and Pedersen *et al.* (2014) demonstrated how  $N_g$  (in the context of IFT) could be used to optimize experimental choices about total exposure time (Fig. 2 and 3 in Pedersen *et al.* 2014), distribution of exposure time at different SANS settings (Fig. 4 in the paper) and percentage of D<sub>2</sub>O in the buffer (Fig. 5 in the paper).

With Paper II, we show that  $N_g$  is a general concept, not limited to BIFT. In the paper,  $N_g$  is found for a sample of nanodiscs and a sample of oblate micelles. The obtained  $N_g$  depends on the data, but also on the model and the prior knowledge. It makes sense, that the information content in data depends on what we know beforehand. This is illustrated in the Paper II, Fig. 10, where  $N_g$  is plotted as function of the prior widths. The information in data decreases as the prior knowledge increases (the prior widths are narrowed). This is, in my opinion, an important point, that is not taken into account by  $N_S$ ,  $J$ , or  $M_S$ , namely that information content is not an intrinsic property of the data. It strongly depends on the posed questions and the prior information.

$N_g$  obtained from BIFT uses very weak assumptions about the data, and thus provide a good estimate of the maximum number of retrievable parameters in data.  $N_g$  used with other models tells how much information (in terms of effective parameters) that was obtained in a given experiment, but this may not be the maximum amount of information in data, e.g. if a simple model with few parameters is fitted to

data.

Intriguingly,  $N_g$  can determine the information content in data, when fitting a model simultaneously to several datasets. This is not possible with any of the other methods, or with  $N_g$  obtained from BIFT (Pedersen *et al.* 2014).

**Oversampling** Oversampling is when the number of data points exceeds the maximum number of retrievable parameters in data, i.e.  $N > N_S$  (or  $J$  or  $M_S$  or  $N_g$ ). SAS data is almost always oversampled. This is however not a problem<sup>1</sup>. Oversampling of data can even lead to increase of the maximum number of parameters that can be refined from data. This maximal number of retrievable parameters is often denoted the information content of data. This increase of information content by oversampling means that the number of parameters derived from data may exceed  $N_S$ , as shown for  $M_S$  (Konarev & Svergun 2015) and for  $N_g$  (Vestergaard & Hansen 2006). The cause for the increase of information content is that data is so well-determined that it can be extrapolated outside of the measured  $q$  range. Note that the extra information is not gained by just by having more data points within the covered bandwidth. It comes from the expansion of the bandwidth. Only a few extra maximum number of parameters can however be gained this way. Note also, that there are no good reasons for rebinning the data heavily to obtain  $N = N_S$  (or  $J$  or  $M_S$  or  $N_g$ ), as proposed e.g. by Rambo & Tanner (2013). Rebinning can only reduce the information content. Rebinning should therefore only be done to speed up calculations and to ease the visual assessment of the quality of the fit.

**Assessing the information content of SAS *ab initio* structure determination** I shortly mentioned *ab initio* structure determination in the introduction. In *ab initio* modelling, a number of beads represent a structure and are moved around to fit data. Beads can also be added or removed. The position of the beads are constrained, e.g. by penalizing having few nearest neighbors. This connectivity constraint (Svergun 1999) ensures a physically meaningful model and favors compact structures. Thus is conceptually similar to IFT, as both approached apply some generic constraints to transform the scattering data from inverse space into real space. However, *ab initio* modelling also transform from 1D to 3D. In that sense, it is "taking IFT to the next level" (or dimension).

Clearly, obtaining a 3 dimensional structure from a noisy 1D scattering curve measured in a limited  $q$ -range is a highly underdetermined problem. So the obtained models are not unique (Petoukhov & Svergun 2015). But how underdetermined is the problem?

Briefly, the method uses a 3-dimensional grid of (approximate) dimensions  $M \times M \times M$ . Each grid point represents a dummy particle and can either be on or off. The degrees of freedom of the unconstrained model is therefore  $K = M^3$ . A realistic number for  $M$  is 50, so  $K \sim 10^5$ . That is  $K \gg N$ , as there are rarely more than 1000 points in a SAXS dataset, and even less in a SANS dataset. Thus the conventional  $f = N - K$  is negative, resulting in negative values for  $\chi_r^2$ . The *ad hoc* solution to this problem, used in DAMMIN (Svergun 1999) is to use  $f = N$ . This does however severely underestimate the  $\chi_r^2$ . By finding  $N_g$  for a given *ab initio* run, one can determine how many of the  $K$  parameters that are determined by the data and how many by the prior constraints. As  $N_S$  rarely exceeds 40, even for a very good synchrotron SAXS dataset (Konarev & Svergun 2015), the problem is for sure very underdetermined with almost all parameters determined by the prior,  $10^5 - 40 \approx 10^5$ .  $N_g$  will provide an even more realistic picture than

---

<sup>1</sup>Sometimes it is discussed whether oversampling can lead to underestimated error bars. It can not, as oversampling and error estimation are uncorrelated. The errors are derived from counting statistics and error propagation and are not affected by oversampling.



$N_s$ , showing how many parameters that are actually determined by data, which is, in most cases, a number well below 40 (Vestergaard & Hansen 2006). Moreover, using  $N_g$  can provide a more realistic value for  $\chi_r^2$ .

## 4.9 Significant achievements in this chapter

The most important contribution from the current part of the thesis is the demonstration of how Bayesian priors can be used in the analysis of SAS data with analytical form factors. This is described thoroughly in Paper II. Interestingly, when ignoring any philosophical aspects, the frequentist approach may be interpreted as a special case of the more general Bayesian approach. Because, when the prior knowledge is very limited, then  $Q \approx \chi^2$ .

The Bayesian approach is useful for more aspects in SAS analysis than those presented in Paper II. Some of these were discussed in the current chapter, including Bayes factor for comparison of models, correction of experimental errors using BIFT, and optimization of the information gained by fitting several datasets simultaneously.

The F-test provides a statistically sound way to compare alternative models, given the goodness of fit of each model. I have applied that method in SAS, as shown in Paper IV. The discussion about degrees of freedom and information content is of fundamental interest. It is also a helpful tool for deducing as much information from valuable data as possible (Pedersen *et al.* 2014). This is important as beamtimes at synchrotron and neutron facilities are limited, and because sample preparation is time-consuming and expensive. Even after great effort in the laboratory and after several beamtimes, the data quality may be of a quality, where optimal analysis is necessary to draw solid conclusions from the data. Especially when working with increasingly challenging systems and problems.



## Chapter 5

# Protein Complexes Studies with SANS Contrast Variation

*"It's not the daily increase but daily decrease.*

*Hack away at the unessential."*

*- Bruce Lee*

In this chapter, SANS and the methods that were introduced in chapter 3 are applied to deduce structural information about four different protein systems. This part of the thesis includes the current chapter as well as Paper III to V and the report in Appendix B.

The studies all make elaborate use of SANS contrast variation. As described in chapter 2, the scattered intensity from a macromolecule depends on its contrast with respect to the solvent. In SAXS, the contrast stems from difference in the electron density. In SANS, the contrast depends on the atomic composition of the nuclei in the sample and solvent, and on the isotope distribution, as isotopes of the same element give different contrasts. Contrast variation can be used to decrease the signal from parts of the sample in order to highlight other parts.

**Contrast variation in SANS.** Hydrogen (H) is the most abundant atom in biological matter and has a coherent neutron scattering length of -3.7 fm (Table 2.1), and its isotope, deuterium (D), has a neutron scattering length of 6.7 fm. The negative sign indicates a change of phase of  $\pi$  for the scattered wave with respect to the incoming wave. Despite having very different scattering lengths, the chemical properties of H and D are almost identical. Therefore, by exchanging H with D in the buffer and/or in the sample, the contrast can be tuned while chemical properties are conserved.

**Contrast variation in SAXS.** Contrast variation is also possible in SAXS, by increasing the electron density of the solvent (see e.g. Tokuda *et al.* 2016). The electron density can be increased by adding small soluble molecules or salt to the solvent. This added salt and/or molecules may however alter the chemical properties of the solvent, and this may affect the structure of the macromolecule or quench interparticular forces. Salt, usually NaCl, is e.g. used in solvents as counter ions, to screen electrostatic interactions between charged molecules, such that long-range order is avoided. On top of that, the molecules may absorb a considerable amount of the X-rays. Sucrose is the most commonly used molecule for SAXS contrast variation.

## 5.1 Structural investigation of membrane proteins in detergents with SAXS and SANS

Membrane proteins need a carrier system that mimics the membrane to be active and stable in solution. The simplest and most widely used system is detergent. A detergent is an amphipathic molecules with a hydrophobic hydro-carbon tail group and hydrophilic head group (Fig. 5.1). Many different natural and synthesized detergents exists, and for membrane proteins it is essential that they solubilize the protein without intruding into hydrophobic pockets, resulting in unfolding. Therefore, mild nonionic detergents are used for solubilization of membrane proteins. When mixed with membrane proteins, the detergent tail groups cover the hydrophobic transmembrane part of the protein, whereas the hydrophilic head groups are oriented away from the membrane proteins, towards the solvent. Detergents are widely used to solubilize and stabilize membrane proteins for functional and structural studies, since they are cheap, effective and easy to use. As mentioned in the preface, nanodiscs is another example of a system for solubilization of membrane proteins. Studying different nanodisc systems with SAXS and SANS were the topic of the first part of my PhD.

The detergents in a sample of membrane proteins are in an equilibrium state. Part of the detergents are dissolved as free molecules in the solution and the concentration of free detergent molecules depends on the critical micelle concentration (CMC). The rest of the detergents either form free micelles, or detergent coronas around the transmembrane part of the membrane protein (Fig. 5.1).

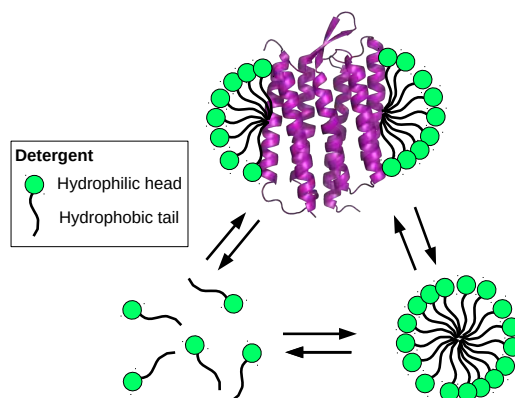


Figure 5.1: In a sample with detergents and membrane proteins, the detergents will be in an equilibrium between free form (lower left), free micelles (lower right), and coronas around the transmembrane part of the membrane proteins (top). The membrane proteins are represented by a crystal structure of bacteriorhodopsin (purple, cartoon representation, PDB: 1M0L).

A sample of membrane protein solubilized in detergents measured with SAS thus has scattering contributions from all three elements. The free detergent molecules are typically too small to be resolved in SAS, so they effectively just change the contrast of the solvent slightly. This is no different than other small molecules in the solvent, such as salt and buffer agents. Therefore, the contribution from the free detergent molecules can be eliminated by careful buffer subtraction [eqn. (2.21)].

### 5.1.1 Elimination of free micelle scattering contribution

The contribution from the free detergent micelles can, in principle, also be eliminated by buffer subtraction. This is however challenging for several reasons. Firstly, it is not trivial to deduce what the detergent

concentration should be in the buffer in order to have the same amount of free micelles as in the sample. Remember, part of the detergents form a detergent corona around the membrane proteins, so if the detergent concentration is the same in buffer and sample, then there will be more free micelles in the buffer. Finally, the sizes of the micelles, which affect the scattering, are variable and typically depends on concentration, temperature and ionic strength of the buffer. The sizes are also affected by the presence of the membrane protein.

The free micelles can be removed from the sample by lowering the detergent concentration to a critical concentration, where there are no free micelles in solution, but the detergent corona is still present. This is possible, as the affinity for corona formation is significantly larger than the affinity for micelle formation (Kaspersen *et al.* 2014). It is however not trivial to predict this critical concentration. At higher concentrations, there are free micelles present in the sample, and at lower concentrations the detergent corona will be starved (Kaspersen *et al.* 2014), which may affect protein stability. Kaspersen *et al.* (2014) showed that the low detergent concentration induced dimer formation. Firstly this is a sign that the membrane protein solubilization conditions are not optimal (the protein is not "happy"), and secondly, such dimerization complicates the structure determination, as more complicated models are needed.

An improved method to isolate the membrane proteins from the free micelles was proposed by Berthaud *et al.* (2012). Using combined size exclusion chromatography (SEC) and SAXS, the scattering signal from different components in a sample can be separated if they differ in size (Pérez & Nishino 2012). This technique was used to investigate the structure of the membrane protein aquaporin in detergent DDM micelles. Berthaud *et al.* (2012) were thus able to fully subtract the signal from the free micelles. The authors also developed a method for generating a coarse-grained model for the detergent corona, and were able to fit the data to high accuracy. Thus, the size and elliptical shape of the corona could be deduced. Software for modelling membrane proteins in detergent coronas have been developed further, such as MEMPROT (Pérez & Koutsioubas, 2015) and the software used by Kaspersen *et al.* (2014).

It is possible to obtain the low-resolution structure of a protein-detergent complex in an *ab initio* approach, using SANS (Koutsioubas 2017). To do that, data must be collected at (at least) two different contrast, and the aggregation number of the detergent corona must be determined. This method however has the disadvantage that a SANS contrast around the protein match-point ( $\sim 42\%$  D<sub>2</sub>O) is needed. This measurement has a large incoherent background, which must be compensated by a high protein concentration or very long exposure time to obtain a proper signal-to-noise ratio.

Methods for analyzing the detergent corona are needed for studies on the whole protein-detergent complex. For example when studying corona formation or detergent-protein interaction. In this thesis, however, the interest lies in the protein structure alone. In that case, the optimal solution is to fully eliminate the scattering contribution from the detergents, as can be done by SANS contrast variation and "invisible" detergents. This was done for the first time by our group, as described in Paper III. I recommend first reading this section (section 5.1.1), and then read the paper.

### 5.1.2 SANS contrast variation with deuterated detergents

I will in the following discuss some different approaches that aim at doing the same job, namely matching out the detergent scattering contribution from a sample of membrane proteins solubilized in detergent. I will argue why the method with "invisible" detergents is most optimal.

The SAS intensity scales with the square of the excess scattering length density,  $\Delta\rho^2$  (Appendix A). That is, if the solvent can be tuned to have the same  $\rho$  as the detergents, then the excess scattering length (contrast),  $\Delta\rho = \rho - \rho_{\text{solv}}$  of the detergents is zero, and the scattering from the detergents is likewise zero. The detergents are said to be "matched out". In SANS,  $\rho_{\text{solv}}$  can easily be tuned by changing the D<sub>2</sub>O

content.

The detergent tail groups consist of carbon and hydrogen, and the head groups contain heavier atoms such as oxygen and phosphor. The head and tail groups therefore have different scattering contrasts with respect to the solvent, and can not be matched out simultaneously. That is, no ratio of  $D_2O$  will result in zero scattering from both head and tail. The average  $\rho$  of the detergents can however be matched, thereby obtaining zero forward scattering,  $I(0) = 0$  (Appendix A). See example 5.1.3.

---

### Example 5.1.3 Contrast match-points for DDM.

The non-ionic detergent n-dodecyl- $\beta$ -Maltoside (DDM) has a head group (the aromatic rings in Fig. 5.2) that is matched out at 49%  $D_2O$  and a tail group (CH chain in Fig. 5.2) that is matched out at 2%  $D_2O$  (Breyton *et al* 2013).

The head group and tail group have an average contrast. For DDM the average contrast is zero at 22%  $D_2O$ , so the total forward scattering,  $I(0)$ , from DDM is zero at 22%  $D_2O$ . At larger values of  $q$ , the scattering from detergents micelles is however non-zero, as the contrast difference between tail groups (core of the micelle) and headgroups (shell of the micelle) is "seen".

Due to the large incoherent scattering from  $H_2O$ , it is not optimal to measure at 22%  $D_2O$ , especially for relatively weakly scattering samples of diluted membrane protein. The match-point (where the contrast is zero) can be altered by deuteration of the detergents (i.e. exchange of H with D).

Deuterated DDM is commercially available with the 25 hydrogens in the tail exchanged by deuterium (d25-DDM). The deuterated tails match out at 114%  $D_2O$  (theoretically), and the average contrast of d25-DDM is zero at 86%  $D_2O$  (Breyton *et al.* 2013), which is more optimal than 22%  $D_2O$  due to lower incoherent scattering from hydrogens in the solvent.

---

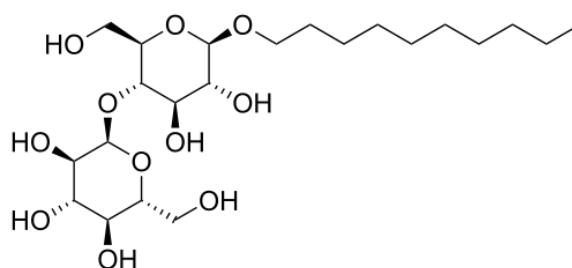


Figure 5.2: The structure of DDM. From Wikipedia common.

At the average match-point for the detergents (where the average detergent contrast is zero),  $I(0)$  stems solely from the membrane protein (corona and free micelles are matched out). Thus, the  $M_W$  can be determined and the oligomeric state evaluated using the methods from chapter 2. To limit the incoherent scattering from hydrogen, the detergent match-point should optimally be close to 100%  $D_2O$ . Several detergents are commercially available in tail-deuterated versions (see example 5.1.3), and the match-point may be shifted to more optimal values using them.

The next problem is that the scattering contribution from matched-out detergents is non-zero at larger  $q$ -values (Breyton *et al.* 2013), as finer structure is seen, and the contrast difference between head and tail is evident. This will affect the scattering, also at relatively low values of  $q$ , as shown by the simulations in Paper III (Fig. 3 in the paper).

**Retrieving overall structural parameters with d25-DDM for large proteins** The simulations in Paper III were done for a relative small membrane protein, bacteriorhodopsin. For larger proteins, the detergent contribution is negligible up to larger values of  $q$ , and some overall structural information can be deduced, such as  $R_g$  from the Guinier plot. We showed this experimentally at the QUAKKA beamline

at ANSTO, where the 363 kDa protein photosystem 1 (PS1) was measured in 0.25 mM d25-DDM at 86 % D<sub>2</sub>O (Fig. 5.3A). The CMC of DDM is 0.17 mM (VanAken *et al.* 1986), so the sample contained DDM coronas and possibly free micelles in the solution. The SANS data could be fitted well with the high-resolution structure of PS1, as deposited in the protein data bank (PDB: 4RKU). The signal-to-noise ratio was too low to deduce finer structure, but the  $R_g$  was determined to be 51 Å, which is in accordance with the theoretical  $R_g$  of 51 Å calculated from the crystal structure (using CaPP). As seen in Fig. 5.3A in the

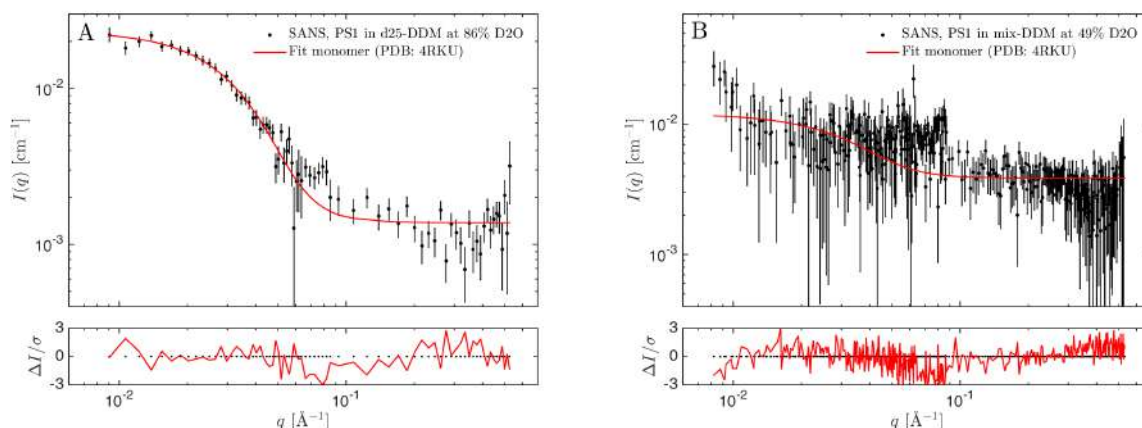


Figure 5.3: SANS data of a sample of PS1 (black) in detergents fitted with the crystal structure of PS1 (PDB: 4RKU, red). Data were measured at the QUAKEA beamline at ANSTO. (A) PS1 in d25-DDM. (B) PS1 in mix-DDM.

residuals, there is a weak signal from the detergents around  $q = 0.1 \text{ Å}^{-1}$  resulting in systematic deviation between data and model. This high- $q$  detergent signal hinders finer structural details to be deduced from the data even if better signal-to-noise ratio was obtained.

### Detergent micelles with homogeneous contrast to quench the scattering in the full $q$ -range

To solve this multi contrast problem, leading to high- $q$  scattering, Oliver *et al.* (2017) showed how to form micelles with an approximative homogeneous contrast that match out in the full measurable  $q$ -range by mixing non-deuterated DDM with d25-DDM. Using a mix of 43 molar percent d25-DDM and 57 molar percent non-deuterated DDM, both mixed head groups and mixed tail groups match out independently at 49% D<sub>2</sub>O. The mix is here denoted mix-DDM. Hence, when measuring at 49% D<sub>2</sub>O, the detergent scattering signal vanishes in the full  $q$ -range. Thus, at a sample of membrane proteins in mix-DDM measured in SANS at 49% D<sub>2</sub>O will scatter as if there were only protein in the sample. Clearly, the contrast situation for a non-deuterated membrane protein is poor, as protein match out around 42%, i.e. close to the needed 49% D<sub>2</sub>O. That is, the excess scattering length density,  $\Delta\rho$ , and hence the scattering intensity  $I(q)$  is low. The scattering signal from the protein therefore "drowns in noise". This is shown experimentally for PS1 in Fig. 5.3B. The contrast of the protein can be improved by exchange of H in the protein with D (protein deuteration), thus obtaining a much better contrast in the experiment. However, it is difficult and expensive to express proteins in deuterated conditions, as it has considerably lower yield than under usual hydrogenated conditions. Moreover, the incoherent scattering is very large at 49% D<sub>2</sub>O, so this technique is not optimal, even if deuterated protein is available.

**The optimal solution: "invisible" detergents** The optimal solution is thus to have detergents with a homogeneous contrast that is matched out in 100% D<sub>2</sub>O, where the incoherent scattering is low

and the contrast of the (non-deuterated) protein high. These "invisible" detergents are the subject of Paper III. The project was lead by Søren Roi Midtgaard (our group). I contributed to the analysis and method development. The project involved several collaborators who provided protein and did part of the analysis, as reflected by the extensive author list of the paper. Moreover, experiments were performed at three different neutron facilities (QUOKKA@ANSTO in Australia, D22@ILL in France and KWS-1@FRM2 in Germany) with support from respective beamline scientists. The the project was a collaboration with Tamin Darwish and co-workers from the national deuteration facility at the ANSTO, who synthesized the detergents. I recommend reading the paper now.

### 5.1.3 Three protein complexes studied with novel "invisible" detergents and SANS contrast variation

In the following, I will describe three scientific cases, where we used the novel detergents. Two of the proteins, SERCA1a and GluA2, are related to neurological diseases, which is a field that is still poorly understood. The SERCA1a studies were done in collaboration with Poul Nissen's group at the department of molecular biology and genetics at Aarhus University. This group has previously solved several atomic resolution structures of SERCA with X-ray crystallography. This work was published as part of Paper III about "invisible" detergents. The GluA2 studies were done in collaboration with Jette Kastrup's group at the department of drug design and pharmacology at University of Copenhagen. This work was part of Paper III and a more elaborate study was subsequently made, focusing on the system rather than on the method. This work is shown in Paper IV. The thirds system, HTL, contains lipids in the center, and the "invisible" detergents allowed for investigations of this lipid core. It is a project in collaboration with Ian Collinson's group at the biochemistry department at University of Bristol, in particular with Remy Martin who conceived the project together with Ian Collinson as well as Søren Roi Midtgaard and Lise Arleth (our group). I participated in the SANS data collection and was involved in the project to analyse the SANS data. This study is reported in Paper V.

**Case 1: SERCA1a** The sarcoplasmic reticulum (SR)  $\text{Ca}^{2+}$ -ATPase (SERCA1a) is an active transporter of calcium ions (front cover image). Energy from dephosphorization of ATP ( $\text{ATP} \rightarrow \text{ADP}$ ) is used to move the positively divalent calcium ions against the electrochemical gradient across the SR membrane. SERCA1a is abundant in skeleton muscles, as this process induces muscle relaxation. There exists a range of SERCA isoforms, all structurally similar (Møller *et al.* 2010), but important for very different physiological processes ranging from neurotransmission and antibody formation (Carafoli 2002) to heart function (Lipskaia *et al.* 2010).

SERCA1a undergoes a structural cycle while pumping (Fig. 5.4), and most of the structural states have been revealed by X-ray crystallography (Møller *et al.* 2010). In Paper III, we studied SERCA1a in a state without calcium and in the presence of the ATP inhibitor AMPPCP. Due to these additives, we expected the SERCA1a to be in either one of the E1 and E2 states (the two left structures in Fig. 5.4), or in a dynamic state between the two. With the SANS data, we could exclude that the protein was in the calcium bound E1 state (Fig. 5.4, top row, middle; PDB: 1T5S), thereby demonstrating that with the novel detergents and SANS we were able to differentiate rather subtle structural changes. The E1 state (PDB: 4H1W) and the E2 state (PDB: 4UU1<sup>1</sup>) fitted equally well with data. The goodness of fit was  $\chi_r^2 = 1.02$  for both the E1 and the E2 states. This should be compared with  $\chi_r^2 = 1.31$  for the calcium bound state. The  $F$ -test yields a probability of 5.4% for getting those  $\chi_r^2$  values given that the models with and without

---

<sup>1</sup>The crystal structure of SERCA1a in the E2 state shown in Fig. 5.4 (PDB: 2C88) is almost identical to the crystal structure of the E2 state we used in Paper III (PDB: 4UU1), and indistinguishable in SANS.



calcium are equally good. So the model could differentiate between states with and without bound calcium. However, the differences were subtle, and the difference would not be assessed significant if a significance level of e.g. 1% was applied. A linear combination of E1 and E2 with about 50% of each structure fitted better than any of the individual structures, with a  $\chi_r^2$  value of 0.94. Thus, the SANS data suggest that SERCA1a in the presence of AMPPCP is in a dynamic state between the E1 and the E2 state. However, when comparing that fit with the fit with either one of the calcium-bound structures using the F-test, a probability of 30% is obtained. That is, the data only very weakly favors the linear combination. As this linear combination is very probable *a priori*, it does constitute the most probable model, when collectively (and qualitatively) taking into account prior knowledge and the new data. It would be more surprising if the protein was permanently in either the E1 or the E2 state. The ratio between the states can however not be estimated reliably from data.

The sample was partly aggregated. To take this into account, we included a structure factor, accounting for this effect. This is thoroughly described in Paper II in general terms, and in the supplemental information of Paper I for SERCA specifically. I will not add anything to that discussion here.

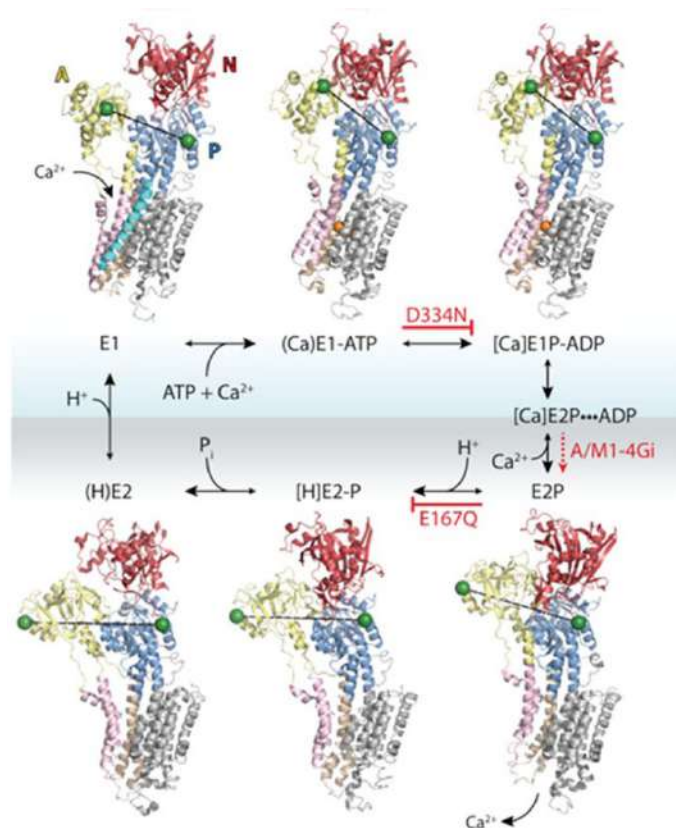


Figure 5.4: The pumping cycle of SERCA1a. PDB codes (clockwise, starting from E1): 4H1W, 1T5S, 1T5T, 3B9B, 3B9R, 2C88. The transmembrane domain, where the detergents sit (compare with the cover image) is shown in pink, brown and gray.  $\text{Ca}^{2+}$  is shown as orange spheres. The figure is adapted from a FRET study (Dyla *et al.* 2017) with permission, and the green spheres show the positions of the FRET labeling cites.

Future perspectives include optimization of the sample preparation to obtain a sample with less or none aggregation. Thereby, structural conclusions can be drawn with higher certainty, and details, e.g. about the ratio between the E1 and the E2 states, can be investigated. As the protein most probably is in a dynamic state between two states, it would be interesting to combine SANS with MD simulations, where

the dynamic transition, that may involve a range of intermediate states, can be investigated.

**Case 2: GluA2** The heterotetramer of the AMPA type glutamate receptor 2 (GluA2) is an ion channel found primarily in the outer cell membranes of neurons. The channel opens upon binding of the neurotransmitter glutamate and allow  $\text{Na}^+$  and  $\text{K}^+$  ions to pass through the cell membrane. In the post synaptic neuron, the flow of ions through GluA2 induces an action potential, and a nerve signal emanates. GluA2 is therefore vital for control of nerve signaling, and plays a role in a range of diseases related to abnormal nerve signal regulation (Bowie 2008).

There are four different subunits that can form glutamate receptor channels (GluA1, GluA2, GluA3 and GluA4). A range of physically relevant homotetramers and heterotetramers exists, e.g. GluA2/3 with two GluA2 units and two GluA3 units (PDB: 5IDF and 5IDE; Herguedas *et al.* 2016). GluA2 (used to refer to the heterotetramer of 4 GluA2 chains) consists of a transmembrane domain (TMD) a ligand binding domain (LBD), an amino terminal domain (ATD) and a cytosolic C-terminal domain (CTD) (Fig. 5.5A shows GluA2 without the CTD). The intrinsically disordered CTD is typically cleaved off before structural studies as it hinders crystallization. It is however of functional relevance and may be related to memory and learning (Zhou *et al.* 2018). The TMD-LBD-ATD protein is conventionally referred to as full-length GluA2, despite missing the CTD.

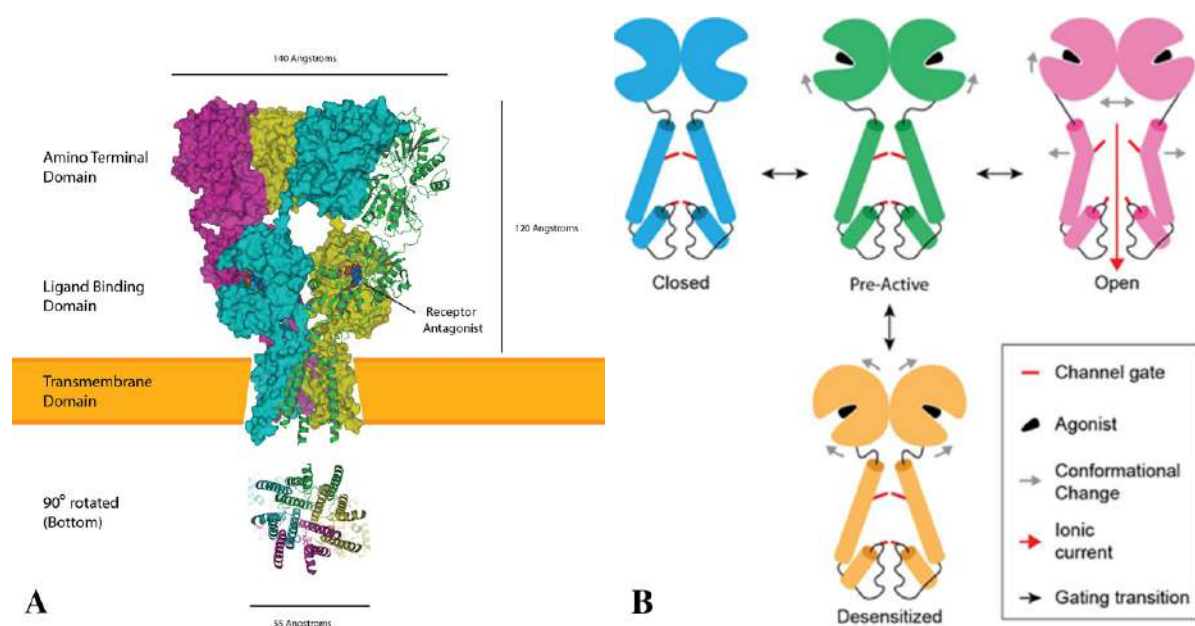


Figure 5.5: (A) The domains of the GluA2 homotetramer with the competitive antagonist ZK-200775 bound (PDB: 3KG2; Sobolevsky 2009). One chain is shown as green cartoon, the three other chains are shown with surface representation. Adopted from Wikipedia (Wikipedia: Glutamate receptor). (B) The major conformational states of the GluA2 channel (only TMD and LBD shown). Adapted from (Twomey & Sobolevsky 2018) with permission.

GluA2 undergoes conformational changes upon ligand binding (Fig. 5.5B). The major states are the closed state, the pre-active (closed) state, the desensitized (closed) state and the open state. Structural studies on GluA2 focus on unveiling these states by studying the protein with different ligands bound (see Table 5.1, page 78). Furthermore, studies from the Sobolevsky Lab focus on the modulation of GluA2 structural states upon binding of regulative TM proteins (Twomey *et al.* 2016 and 2017; Twomey & Sobolevsky 2018).

The first high-resolution structure of full-length GluA2 was obtained by X-ray crystallography (PDB:

3KG2; Sobolevsky *et al.* 2009; Fig. 5.5). GluA2 was in the closed state with the non-competitive antagonist ZK200775 bound to the LBD. A range of X-ray structures followed with GluA2 in different states (Durr *et al.* 2014; Chen *et al.* 2014; Yelshanskaya *et al.* 2014)<sup>2</sup>.

GluA2 is very well-suited for EM due to its impressive size (368 kDa), which was exploited first by Meyersen and co-workers (Meyersen *et al.* 2014). Notably, a loose low-resolution structure of GluA2 in the desensitized state was reported by Meyerson *et al.* (EM class3; EMD: 2688), with the ATD spread apart in a loose form (see Fig. 1B in Paper IV). This is in contrast to all the X-ray structures, that were more compact in the ATD. The development of hardware and software for cryo EM have made it possible to obtain electron density maps with resolution comparable to those obtained with X-ray crystallography. This was shown by Twomey *et al.* from the Sobolevsky lab, and a range of high-resolution EM structures were published during 2016 and 2017 with resolution as low as 4.2 Å (PDB: 5WEO; EMD: 8821; Twomey *et al.* 2017b)<sup>3</sup>. In our SANS study we measured GluA2 in solution, and could thereby test potential artifacts from crystallization of grid fixation (in EM). We also aimed at investigating the desensitized state, i.e. whether it was compact (PDB: 5VHZ; Twomey *et al.* 2017b), or loose (EMD: 2688; Meyersen 2014).

We therefore studied GluA2 in the presence of the full agonist AMPA (1 mM), bringing GluA2 into either the compact active (open) state, the compact resting (pre-active) or the compact or loose desensitized state. We also studied GluA2 in the presence of the non-competitive antagonist GYKI-53655 (1 mM). Intriguingly, GYKI has been tested as a therapeutic drug against epilepsy (Fritsch *et al.* 2010) and a similar non-competitive antagonist, Perampanel is on the market. A high-resolution crystal structure of GluA2 with GYKI was available, revealing yet a compact structure (PDB: 5L1H; Yelshanskaya *et al.* 2016). As a reference, we studied GluA2 in the apo form, expecting it to be in the compact resting state. Finally, we studied GluA2 in the presence of 10 mM AMPA, with lowered pH from neutral pH 7.5 to acidic pH 5.5.

We could verify that the high-resolution structures from X-ray crystallography and EM were consistent with the data for apo GluA2, as well as for GluA2 in the AMPA bound state at pH 7.5 and in the GYKI-53655 bound state. It was likewise possible to exclude that the AMPA sample at pH 7.5 was in the loose desensitized EM class 3 form (EMD: 2688; Meyersen *et al.* 2014).

We could, however, not differentiate finer structural differences between the compact forms with the available resolution of our SANS data. Therefore, it could not be concluded whether the AMPA bound GluA2 sample at pH 7.5 was in the compact resting state (PDB: 4U2P; Durr *et al.* 2014), the compact active state (PDB: 5WEO; Twomey *et al.* 2017b) or the compact desensitized state (PDB: 5VHZ; Twomey *et al.* 2017b). The F-test was used to test whether the difference between the goodness of fits were significant.

Intriguingly, we found that in the presence of AMPA, and at acidic pH, GluA2 was in a loose state resembling the loose EM structure found by Meyersen *et al.* (2014). I recommend reading the paper now, where the results will be presented and discussed in more detail.

**The holo-translocon protein complex, HTL** The holo-translocon (HTL) is a protein complex that translocates nascent proteins across the cell membrane and embeds membrane proteins in the membrane. It is therefore essential for the folding of other membrane proteins, such as SERCA and GluA2. This connection exemplifies a tiny corner of the impressive biological clockwork of proteins, membranes, organelles, neurotransmitters etc. that all have to work together for living organisms to retain life.

HTL is itself a membrane proteins, with three major domains, the SecYEG domain (3 protein chains), the SecDF domain (2 chains) and the YidC domain (1 chain), as shown in Fig. 5.6. All domains have a large

<sup>2</sup>Interestingly, most structures are found by scientists from the Goaux Lab or the Sobolevsky Lab. Sobolevsky was postdoc in Goaux lab from 2004 to 2010.

<sup>3</sup>I can recommend having a glance at the EMD electron density map. Details such as the layer of detergent headgroups, and  $\alpha$ -helices in the TMD can be seen.

TM part and SecDF and YidC furthermore have substantial periplasmic parts. It has been suggested from X-ray crystallography that the periplasmic part of the SecDF domain can change formation and protrude into the periplasmic space (Fig. 5.6B).

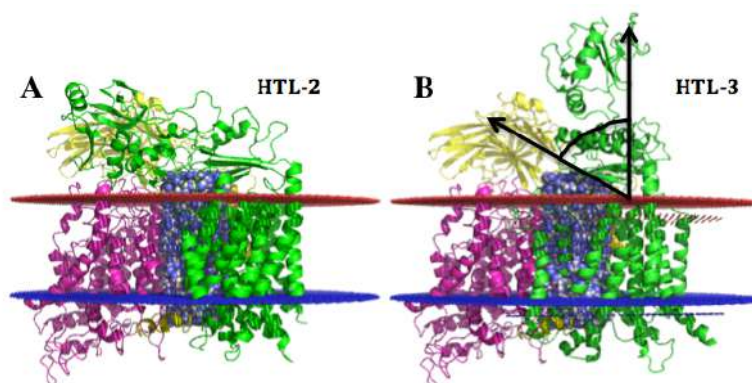


Figure 5.6: (A) The structure of HTL refined with EM (PDB: 5MG3; Botte *et al.* 2016). SecYEG domain in magenta, the SecDF domain in green and the YidC domain in yellow. A simplified lipid core is shown in the center (blue and white for C and H respectively), corresponding to 15 lipids. The blue and red planes show the position of the complex in a lipid bilayer as estimated with the OPM database (<http://opm.phar.umich.edu>). (B) HTL with part of the SecDF domain protruding into the periplasmic space (PDB: 5XAM; Furukawa *et al.* 2017)

It is believed that the HTL complex has a central cavity containing a minor lipid bilayer (Botte *et al.* 2016). This hypothesized lipid core play an important role for membrane protein insertion, but is difficult to probe experimentally.

With our SANS study (Paper V) we aimed at verifying or refute the existence of the lipid core, and investigate the flexibility of the SecDF domain. The HTL was solubilized in "invisible" detergents and the sample was measured in 100% D<sub>2</sub>O-based buffer, such that the scattering came solely from the HTL protein complex and the lipid core. The forward scattering was consistent with the existence of a lipid core of about 8 lipids. Fitting to the data refined the number of lipids to  $17 \pm 5$ . The existence of a lipid core was consistent with CGMD simulations performed by Remy Martin and reported in the paper.

The lipid core used for SANS was generated in a Monte Carlo approach using home-written software. With a certain probability, a carbon atom was placed on a grid. A hydrogen was then placed next to the carbon, in a random direction, and a second hydrogen was placed at random, but at a 110° angle with respect to the first hydrogen.

The SANS data suggested flexibility of the SecDF domain, as a combination of HTL-2 and HTL-3 gave the best fit to data. The study is described in more detail in Paper V, which I recommend reading now.

## 5.2 Studying $\alpha$ -synuclein structure and dynamics with SANS contrast variation

In this last scientific case, on  $\alpha$ -synuclein ( $\alpha$ SN), SANS contrast variations was also used.  $\alpha$ SN is very different from the three membrane protein systems, SERCA1a, GluA2 and HTL, and the novel detergents were not used. Like GluA2 and SERCA, the system is related to neurological disorders. This project was done in collaboration with Bente Vestergaard's group at the department for drug design and pharmacology at University of Copenhagen.

Isolated  $\alpha$ -synuclein ( $\alpha$ SN) is a small soluble and intrinsically disordered protein. It has a high tendency to fibrillate and form amyloids, characterized by a cross- $\beta$  secondary structure. Accumulation of  $\alpha$ SN fibrils is a hallmark of Parkinson's disease (Stefanis 2012), and its casual relation to the disease has been studied extensively. The fibrillation process includes formation of intermediate oligomers and the final sample is in an equilibrium between the monomeric, oligomeric and fibril form (Fig 5.7). The fibrillation pathway is complex, and several alternative pathways have been discovered (Marasini & Vestergaard 2017). In Fig. 5.7, a fibrillation curve is shown, and some of the monomeric, intermediate and fibril structural states are highlighting. Even monomeric  $\alpha$ SN can vary structurally on both secondary and tertiary level, and  $\alpha$ SN has therefore, quite telling, been named a "protein chameleon" in a review by Uversky (2003).

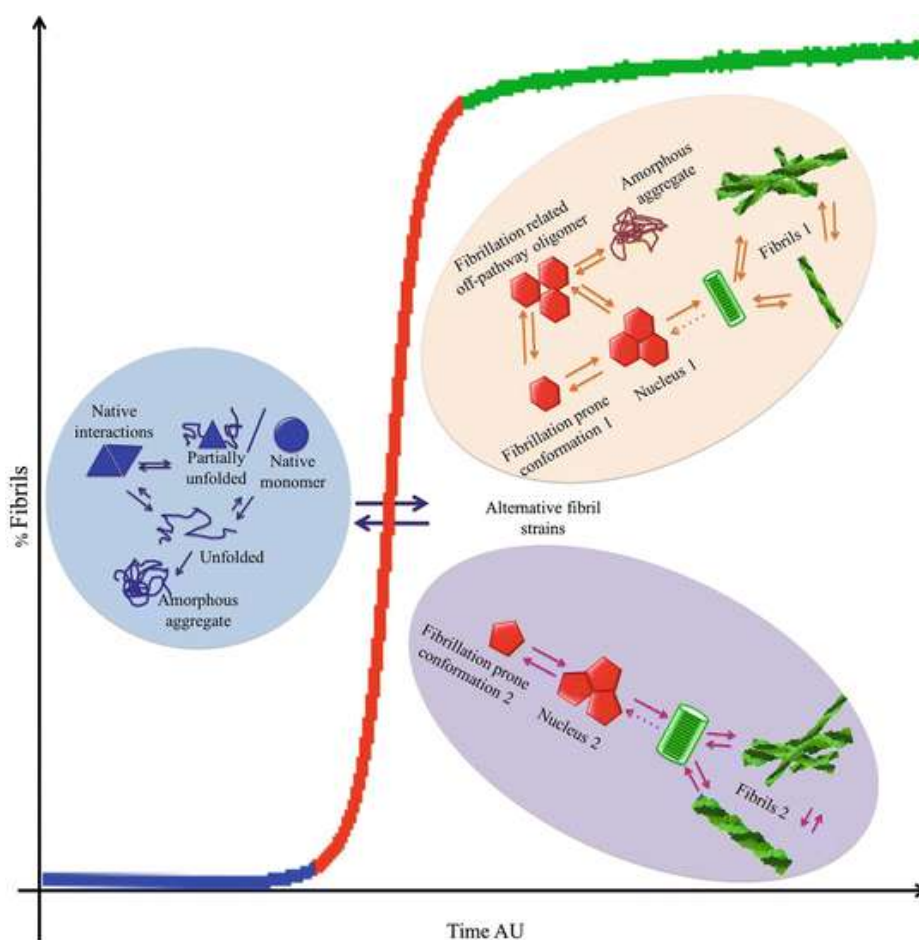


Figure 5.7:  $\alpha$ SN fibrillation over time. Before aggregation,  $\alpha$ SN is in monomeric form (blue). At some point, the protein forms oligomers (red), and aggregate further to form amyloid fibrils (green). There are many different forms of both monomer, oligomer and fibrils as indicated in the highlighted circles, and  $\alpha$ SN is at all times in an equilibrium between different states. Adapted from (Marasini & Vestergaard, 2017) with permission.

One of the key questions in the field of  $\alpha$ SN, is what structural states are toxic and have a causal connection to neurodegenerative diseases (Roberts & Brown 2015). Much evidence suggests that the oligomeric states and not the final fibrils are the toxic components (Stefanis 2012). To gain control over this complex system, it is important to be able to understand the fibrillation process. In our SANS contrast variation experiment, we aimed at studying three effects: water layer around the fibrils, subunit structure and monomer-fibril exchange. Amyloid fibrillation is a general structural state achievable for many proteins (Dobson 2004), so

the relevance of the results expands beyond the field of  $\alpha$ SN and is of more fundamental interest. I will give ongoing references to Figures in the experimental report in Appendix B.

**SANS data supports a conventional water layer** It is generally accepted that water is disrupted in the vicinity of proteins, as discussed in chapter 3. It was proposed by Nielsen *et al.* (2013) that an extended solvent phase formed around  $\alpha$ SN fibrils. This was revealed from combined SAXS and NMR data. The SAXS data were measured during fibrillation of wild type (WT)  $\alpha$ SN and mutant A30P  $\alpha$ SN. A change of the high- $q$  scattering was observed for both samples. Intriguingly, the authors observe that the high- $q$  SAXS scattering of WT  $\alpha$ SN decreases to a level below the signal from the buffer. The authors argue that this stems from an extended and dense water layer. It is speculated that this extended water layer may affect how  $\alpha$ SN interacts with other cellular components. This idea was pursued in our work, where we could highlight the scattering from the water layer by SANS contrast variation.

Assuming there is an extended water layer, then the  $\alpha$ SN would be a three-phase system: bulk water, water layer and protein, or in other words, the  $\alpha$ SN with water layer would effectively be a multi-contrast core-shell particle. At the match-point, a subtracted SAS curve from a particle with uniform contrast is zero in the full  $q$ -range. A multi-contrast particle, on the other hand, would only have zero scattering at  $q = 0$ , and non-zero scattering at higher values of  $q$ . By measuring  $\alpha$ SN at the match-point, we could investigate whether there were an extended and dense water layer, as this would result in non-zero scattering arising from the multi-contrast situation.

The scattered signal from a water layer is weak. Therefore, the protein had to be very concentrated in order to obtain sufficient signal over noise. Secondly, the match-point of proteins differs slightly around  $\sim 42\%$  and even a weak signal from the protein would easily dominate over the water layer signal, so the match-point had to be determined with high precision.

The match-point was determined to be at 39.4 % D<sub>2</sub>O (Appendix B, Fig. 1), which is below the typical protein match-point of 42%, indicating that the scattering length density is slightly smaller in  $\alpha$ SN than in an average protein. This match-point was used as basis for a long measurement of  $\alpha$ SN at matched-out conditions. Fig. 5 in Appendix B shows SANS data from  $\alpha$ SN fibrils measured in solvents with different D<sub>2</sub>O contents. The purple curve in the bottom is the measurement at the match-point. No (or very little) significant signal over background was observed. The sample was measured for almost 12 hours at the instrument KWS-1 at FRM2, and the concentration was  $\sim 10$  mg/ml. At higher concentrations, the sample became viscous and could not be measured in standard cuvettes.

However, by measuring a more samples close to the match point, and simultaneously fitting all the curves, as shown in Fig. 5 in Appendix B, the information content about the water layer from the data was maximized. By including these new measurements in the contrast variation series, the match-point was updated to 40.2% (Fig. 2 i Appendix B).

All measured SANS curves were fitted with a simplified model of  $\alpha$ SN, namely a core-shell cylinder with elliptical cross section (see insets in Fig. 4 in the appendix). The core represented the protein phase and the shell the dense water layer around the protein. The model was consistent with data up to a  $q$ -value of about  $0.05 \text{ \AA}^{-1}$ . At higher values of  $q$ , significant systematic errors were apparent, showing that the model failed to describe the finer structure of the fibrillating system. By minimized  $\chi^2$ , the thickness of the water layer was refined to a value of  $2.3 \pm 1.3 \text{ \AA}$  assuming 10% increased density with respect to bulk water (Svergun *et al.* 1998). That is consistent with a monolayer of compressed water molecules ( $\sim 3 \text{ \AA}$ ). Hence, the SANS data support an uncontroversial standard description of the perturbed water layer around fibrillated  $\alpha$ SN, similar to the water layer we would expect around other proteins. It can therefore not explain the SAXS an NMR data (Nielsen *et al.* 2013).



**The effect of compressibility** The SANS data seemingly contradicts the SAXS data by Nielsen *et al.* (2013). The authors claim that the change in high- $q$  (above  $0.23 \text{ \AA}^{-1}$ ) is not due to structural changes of the protein. There is however no theoretical support for that claim. Secondly, Nielsen *et al.* observe that the subtracted data for WT  $\alpha$ SN is negative at  $q$ -values above  $0.35 \text{ \AA}^{-1}$ . Assuming it is not an instrumental effect, it can only stem from the buffer having a larger incoherent scattering signal than the sample. The  $q$ -independent SAXS scattering comes from inelastic Compton scattering and from elastic Thompson scattering from density fluctuations in the sample. The contribution from Compton scattering can however be neglected in SAXS (Svergun *et al.* 2013), so the incoherent scattering signal is given as (Zemb *et al.* 2003):

$$I(q) = \rho^2 kT \chi_T \quad (5.1)$$

where  $\rho$  is the scattering length density,  $kT$  is the thermal energy and  $\chi_T$  is the isothermal compressibility. Thus, the incoherent SAXS scattering depends on the material properties  $\rho$  and  $\chi_T$ . A large  $\rho$  means that there are many scatterers, and a large  $\chi_T$  increase density fluctuations.

At high  $q$ , the scattering is dominated by the flat ( $q$ -independent) scattering, and negative scattering can emerge only if the flat scattering is smaller for the sample than for the buffer. Hence,  $\rho$  and/or  $\chi_T$  must decrease radically during fibrillation to account for the reported results.

Bulk water has an electron density of  $0.33 \text{ \AA}^{-3}$ , and protein typically has an electron density of about  $0.44 \text{ \AA}^{-3}$  (Gekko & Noguchi 1979), i.e. significantly higher than bulk water. A hydration layer is expected to have an electron density of  $0.33$  to  $0.36 \text{ \AA}^{-3}$  (Svergun *et al.* 1998; Persson *et al.* 2018). The compressibility of bulk water is  $\chi_T \approx 46 \text{ Mbar}^{-1}$  at ambient temperatures (Millero *et al.* 1969), whereas  $\chi_T$  for proteins is in the range  $10$ - $25 \text{ Mbar}^{-1}$  (Kharakoz 2000). Using these values, we can calculate the flat scattering for bulk water to be  $0.017 \text{ cm}^{-1}$  and for proteins to be in the range  $0.006$  to  $0.016 \text{ cm}^{-1}$ . This means that a sample of protein dispersed in water has a slightly smaller incoherent scattering contribution than the water itself. For a diluted sample of proteins this effect is minor. However, for fibrillated proteins of relatively high concentrations ( $12 \text{ mg/ml}$  in Nielsen *et al.* 2013), the proteins constitute a significant amount of the sample volume. Therefore, it is not very surprising that a small negative contribution can be obtained at large values of  $q$ . A closer packing of the water layer would in fact have the opposite effect than what is seen in Nielsen *et al.* (2013). As seen from equation (5.1), the intensity increases with increased  $\rho$ , meaning that negative data after subtraction is less likely.

The spread in the values of  $\chi_T$  for proteins show the diversity of the compressibility among proteins, and indicate that the change in incoherent scattering may come from a change of  $\chi_T$  or  $\rho$  of the protein during fibrillation. It has been reported that hen lysozyme gets more voluminous and compressible as it forms amyloids (Akasaka *et al.* 2007). Such change in the protein during fibrillation might explain the SAXS data. The measured compressibility is however also affected by the water layer, as shown by Persson & Halle (2018), so the SAXS data may be change of the compressibility of the water layer. Changes in the compressibility does not change the SANS contrast situation and are thus consistent with the SANS data. Nielsen *et al.* show how NMR studies reveal a decrease in water mobility during fibrillation and relate this to higher density of the surface water, which, as we have seen, is not consistent with the SANS data. An alternative explanation is that the density is conserved, but the compressibility of the surface water is decreased.

The change of  $\chi_T$  can be measured with ultrasonic velocimetry (Akasaka *et al.* 2007; Kharakoz 2000). Such measurement on  $\alpha$ SN during fibrillation may provide additional understanding of the system.

**Subunit structure** In the field of polymer science, it is a standard technique to measure the structure of a labeled polymers in a polymer matrix by SANS contrast variation (Cotton 1996). A sample is prepared

with a minor fraction of deuterated polymers, and the sample is then measured at the match-point for the hydrogenated polymers. In that way, the structure of the single polymers in the polymer matrix can be investigated. In a pilot experiment, we attempted to use the same technique to study the structure of single subunits of  $\alpha$ SN in a matrix of  $\alpha$ SN fibrils (Fig. 5.8A).

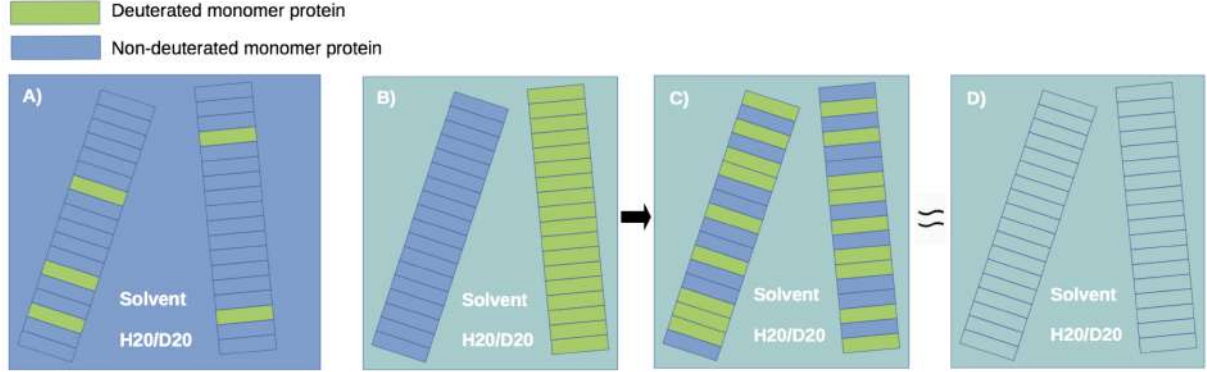


Figure 5.8: A) SANS on cofibrillated  $\alpha$ SN gives structural information on the monomer proteins. B) Deuterated fibrils and non-deuterated fibrils are mixed. C) After a time period with exchange of monomer proteins, the system has reached equilibrium. (D) At very small  $q$ , the scattering is well-approximated by fibrils with homogeneous contrast equal to the average contrast of the deuterated and non-deuterated protein. The solvent is chosen to match this average contrast, such that  $I(0)$  decreases with exchange.

The obtained SANS data did differ significantly from the SANS data for hydrogenated fibrils (Fig. 7 in Appendix B). The average size of the particle is smaller than for the fibrils as revealed by the Guinier region, which is linear as opposed to the sample of hydrogenated fibrils. The deduced  $p(r)$  however has the same shape for large  $r$ . So most likely, the hydrogenated part of the protein was not fully matched out. We chose not to continue that project after the initial pilot project. Optimization of sample preparation and experimental condition was needed to obtain a structure of the monomers isolated from the fibrils. This was hindered by expenses (in money and time) for the sample preparation. Secondly, and most importantly, the structure of  $\alpha$ SN was solved in an impressive solid-state NMR study about a year after our pilot experiment (Tuttle *et al.* 2016).

**Monomer-Amyloid Exchange in  $\alpha$ SN rejected** It has been discussed whether there is monomer-amyloid fibril exchange for  $\alpha$ SN or whether the monomers are firmly bound in the amyloid matrix. Due to the internal bonds in the amyloid fibril, one would not expect exchange. There are however experimental evidence for monomer-oligomer exchange in  $A\beta$  amyloids (Fawsi *et al.* 2010), which is structurally similar to  $\alpha$ SN.

We probed the possible exchange in  $\alpha$ SN by SANS contrast variation. Fibrillated non-deuterated  $\alpha$ SN were prepared along with fully deuterated fibrillated  $\alpha$ SN. The samples were then dissolved in each their buffer with different amounts of  $D_2O$ . The samples were mixed and the measured SANS signal monitored over time (Fig. 5.8). Exchange induces a different contrast situation, with all samples having a mix of deuterated and non-deuterated  $\alpha$ SN fibrils. The experiment was designed such that the average scattering length density of the mixed  $\alpha$ SN would be almost the same as that of the mixed buffer, such that the forward scattering would approach zero upon exchange. The method has been used previously to study the lipid exchange in nanodiscs (Nakano *et al.* 2009). The mixed sample was measured several times over a timespan of 48 hours at room temperature. All measured SANS curves were identical (Fig. 3 in Appendix B), meaning that no



exchange had occurred during those 48 hours.

There may still be exchange on a longer time-scale and/or at higher temperatures. But with the experiment, we could exclude that the suggested monomer-amyloid exchange took place within 2 days at ambient temperature.

**Future prospects** New important findings have been published in 2018 in the description of the water layer around proteins (Persson *et al.* 2018; Henriques *et al.* 2018), and this knowledge should be included in the analysis, that was made before these studies. The value for the density increase for the water layer was believed to be about 9-10% for many years due to the finding by Svergun *et al.* (1998), but from the recent studies, it seems like 6% is a better default value.

As amyloid fibrillation is a general property of many proteins, some of the investigated aspects could be investigated using other model proteins systems, which are simpler and better understood and therefore easier get and to handle. A candidate is  $\alpha$ -lactalbumin (see e.g. Goers *et al.* 2002).

### 5.3 New insight into challenging protein complexes made possible by new tools

In summary, we have explored how the novel "invisible" detergents made it possible to study the structure of membrane proteins complexes with SANS without having to take into account the detergent corona of free micelles in the analysis. The detergents matched out in the full  $q$ -range at 100% D<sub>2</sub>O, which allowed subtle structural details to be probed. The software CaPP was used to add a water layer to the membrane proteins except in the transmembrane region, and to calculate the theoretical  $p(r)$  functions for direct comparison with data. The tools for taking into account aggregate contributions, as described in Paper II, were used in all three scientific cases.

SERCA was investigated in a structurally unknown state in the presence of the ATP inhibitor AMPPCP and in absence of calcium. We could see the difference between calcium-bound and calcium-unbound states, and the data supported a suggested hypothesis that the protein was in an equilibrium between the E1 and E2 conformational states, which were known from X-ray crystallography.

We confirmed that GluA2 in solution and in the resting state were in a compact forms, as suggested by X-ray crystallography and cryo EM. The SANS data also verify the crystal structure of GluA2 in the presence of the allosteric modulator GYKI-53655 and in presence of AMPA at neutral pH. Finally, we found that GluA2 in the presence of AMPA and at acidic pH was in a loose form with a wide spread in the ATD. The F-test allowed for quantitative assessment of how well the different models fitted the data, so the most probable model could be found. Several tools were used in the study of HTL. First off all the "invisible" detergents allowed the lipid core signal to be separated from the signal from the detergent corona. Secondly, CaPP was used to calculate the theoretical  $p(r)$  and scattering intensities, and to add a water layer in the analysis. I added a simple, simulated lipid core to the HTL structure. By modification of CaPP, I could ensure that the scattering from this core was calculated correctly. I also took into account aggregates in the sample by inclusion of one of the structure factors from Paper II. Finally, I renormalized the experimental errors using BIFT as discussed in chapter 4.

It is my hope that the methods will be applied in many future SAS studies to aid the understanding of complex biological systems at the nanoscale.

Full length GluA2 structures												
#	Subtype	Reference	PI	PDB	Ligand(s)/peptides	EMD	State	Tech.	Res. [Å]	Form	Construct (mutation)	Detergent
1	AMPA 2 homotetramer	Sobolevsky et al, 2009	Gouaux	3KG2	ZK 200775	--	closed	X-ray	3.6		GluA2cryst	C11thio
2	AMPA 2 homotetramer	Durr et al, 2014	Gouaux	4U2P	--	--	rest	X-ray	3.2	form A	5M	DM
3	AMPA 2 homotetramer	Durr et al, 2014	Gouaux	4U1W	kainate, (R,R)-2b	--	part act	X-ray	3.3	form A	5M	C11thio
4	AMPA 2 homotetramer	Durr et al, 2014	Gouaux	4U1X	kainate, (R,R)-2b	--	part act	X-ray	3.3	form B	10M	C11thio
5	AMPA 2 homotetramer	Durr et al, 2014	Gouaux	4U1Y	FW, (R,R)-2b	--	part act	X-ray	3.9	form A	10Mdel	C11thio
6	AMPA 2 homotetramer	Durr et al, 2014	Gouaux	4U2Q	kainate	--	part act	X-ray	3.5	form A	5M	DDM
7	AMPA 2 homotetramer	Chen et al, 2014	Gouaux	4U5B	(R,R)-2b, kainate	--	part act	X-ray	3.5		GluA2cryst1 (A622T)	C11thio
8	AMPA 2 homotetramer	Chen et al, 2014	Gouaux	4U5C	FW, (R,R)-2b	--	part act	X-ray	3.7		GluA2cryst1	C11thio
9	AMPA 2 homotetramer	Chen et al, 2014	Gouaux	4U5D	kainate, (R,R)-2b	--	part act	X-ray	3.6		GluA2cryst1	C11thio
10	AMPA 2 homotetramer	Chen et al, 2014	Gouaux	4U5E	kainate, (R,R)-2b	--	part act	X-ray	3.5		GluA2cryst1 (T625G)	C11thio
11	AMPA 2 homotetramer	Chen et al, 2014	Gouaux	4U5F	kainate, (R,R)-2b	--	part act	X-ray	3.7		GluA2cryst2	C11thio
12	AMPA 2 homotetramer	Yelshanskaya et al, 2014	Sobolevsky	4U4F	(S)-5-Nitrowillardiine	--	part act	X-ray	4.8		GluA2*	DDM
13	AMPA 2 homotetramer	Yelshanskaya et al, 2014	Sobolevsky	4U4G	ZK200775	--	closed	X-ray	4.5		GluA2*	DDM
14	AMPA 2 homotetramer	Yelshanskaya et al, 2016	Sobolevsky	5L1B	--	--	rest	X-ray	4.0		GluA2del (C589A)	C11thio
15	AMPA 2 homotetramer	Yelshanskaya et al, 2016	Sobolevsky	5L1E	CP465022	--	closed	X-ray	4.4		GluA2del	C11thio
16	AMPA 2 homotetramer	Yelshanskaya et al, 2016	Sobolevsky	5L1F	Perampanel	--	closed	X-ray	4.0		GluA2del	C11thio
17	AMPA 2 homotetramer	Yelshanskaya et al, 2016	Sobolevsky	5L1G	GYKI-Br	--	closed	X-ray	4.5		GluA2del	C11thio
18	AMPA 2 homotetramer	Yelshanskaya et al, 2016	Sobolevsky	5L1H	gyki-53655	--	closed	X-ray	3.8		GluA2del	C11thio
19	AMPA 2 homotetramer	Twomey et al, 2016	Sobolevsky	5KBS	0x5TZ	8229	rest	EM	8.7		GluA2	DDM
20	AMPA 2 homotetramer	Twomey et al, 2016	Sobolevsky	5KBT	1x5TZ	8230	rest	EM	6.4		GluA2-STZ	DDM
21	AMPA 2 homotetramer	Twomey et al, 2016	Sobolevsky	5KBU	2x5TZ	8231	rest	EM	7.8		GluA2-STZ	DDM
22	AMPA 2 homotetramer	Twomey et al, 2016	Sobolevsky	5KBV	ZK200775	8232	closed	EM	6.8		GluA2	DDM
23	AMPA 2 homotetramer	Twomey et al, 2017a	Sobolevsky	5VHY	2xGSG1L, ZK	8687	closed	EM	4.6		GluA2-GSG1L	DDM
24	AMPA 2 homotetramer	Twomey et al, 2017a	Sobolevsky	5VHZ	2xGSG1L, L-quisqualate	8688	act/des/rest	EM	8.4		GluA2-GSG1L	DDM
25	AMPA 2 homotetramer	Twomey et al, 2017b	Sobolevsky	5WEK	ZK, GSG1L	8819	closed	EM	4.6	state 1	GluA2-GSG1L	Digitonin
26	AMPA 2 homotetramer	Twomey et al, 2017b	Sobolevsky	5WEL	ZK, GSG1L	8820	closed	EM	4.4	state 2	GluA2-GSG1L	Digitonin
27	AMPA 2 homotetramer	Twomey et al, 2017b	Sobolevsky	5WEM	GSG1L	8821	rest	EM	6.1	state 1	GluA2-GSG1L	Digitonin
28	AMPA 2 homotetramer	Twomey et al, 2017b	Sobolevsky	5WEO	Glu, CTZ and STZ	8821	open	EM	4.2	state 1	GluA2	Digitonin
29	AMPA 2 homotetramer	Meyerson et al, 2014	Subramaniam	4UQJ	ZK200775	2680	closed	EM	10.4		GluA2em	DDM
30	AMPA 2 homotetramer	Meyerson et al, 2014	Subramaniam	4UQ6	glutamate, LY451646	2684	act/des/rest	EM	12.8		GluA2em	DDM
31	AMPA 2 homotetramer	Meyerson et al, 2014	Subramaniam	4UQQ	2S,4R-4-methylglutamate	2685	act/des/rest	EM	7.6		GluA2em	DDM
32	AMPA 2 homotetramer	Meyerson et al, 2014	Subramaniam	--	quisqualate	2686	act/des/rest	EM	21.4	class 1	GluA2em	DDM
33	AMPA 2 homotetramer	Meyerson et al, 2014	Subramaniam	--	quisqualate	2687	act/des/rest	EM	25.9	class 2	GluA2em	DDM
34	AMPA 2 homotetramer	Meyerson et al, 2014	Subramaniam	--	quisqualate	2688	act/des/rest	EM	22.9	class 3	GluA2em	DDM
35	AMPA 2 homotetramer	Meyerson et al, 2014	Subramaniam	4UQK	quisqualate, LY451646	2689	act/des/rest	EM	16.4		GluA2em	DDM
Related structures												
#	Subtype	Reference	PI	PDB	Ligand(s)/peptides	EMD	State	Tech.	Res. [Å]	Form	Construct (mutation)	Detergent
36	Kainate 2 homotetramer	Meyerson et al, 2016	Subramaniam	5KUF	2S,4R-4-methylglutamate	8289	act/des/rest	EM	3.8		GluK2	DDM
37	Kainate 2 homotetramer	Meyerson et al, 2016	Subramaniam	5KUH	LY466195	8290	act/des/rest	EM	11.6		GluK2	DDM
38	AMPA 2/3 heterotetramer	Herguedas et al, 2016	Greger	5IDF	--	8090	rest	EM	8.3	model 1	GluA2/3	N/A
39	AMPA 2/3 heterotetramer	Herguedas et al, 2016	Greger	5IDE	--	8090	rest	EM	10.3	model 2	GluA2/3	N/A
40	AMPA 2 homotetramer, no ATD	Zhao et al, 2016	Gouaux	5KK2	TARP	8256	rest	EM	7.3		GluA2-tarp	Digitonin

Table 5.1: Available high-resolution GluA2 full-length structures. Rotated to obtain a (hopefully) readable font size. Three low-resolution EM structures were included in the table (#32-34), for completeness, and because structure #34 (EM class 3) is used in Paper IV. Some related structures are included in the table as they are referred to in Paper IV. PI: principle investigator, PDB: protein data bank, EMD: electron microscopy data bank, Tech.: experimental technique, Res.: resolution.

## Chapter 6

# Conclusion and Final Remarks

*"The journey of a thousand miles begins with one step"*

- Lao Tzu

The aim of my Phd was to aid the understanding of ourselves. More precisely to dig deeper into the biological complexes that allow us to think, act, and write a PhD thesis. This fundamental problem can partly be answered by zooming into the nanoscale. SAS is one of the experimental techniques capable of probing structures at this scale. I have developed analytical and statistical tools for SAS and discussed the limitations of the technique as well as its complementarity with other techniques. By this process, I have clarified the limitation of SAS, and most importantly expanded these limits, such that more complex biological nanoscale systems can be investigated.

More concretely, I have developed a computer program CaPP, as described in chapter 3 for calculation of theoretical SAS scattering and the pair distance distribution function for proteins, in particular for membrane proteins. I have also used and developed tools for inclusion of aggregates in the analysis of SAS, as described in chapter 3 and in Paper I. In chapter 4, I discussed how SAS can be analyzed in a statistical sound way, which is not always the case in widely used methods. In particular, I discussed the degrees of freedom and information content of SAS data. I demonstrated how a wrong choice of the degrees of freedom may lead to wrong conclusions from the data. Also I have introduced Bayesian statistics in the analysis of SAS with analytical models, as described in Paper II. This is essential for correct inclusion of prior information, and to combine several SAS datasets. The Bayesian method also added fundamental insight into the information content of SAS data, via the number of good parameters  $N_g$ . The information content of data and the degrees of freedom of SAS data was thoroughly discussed in the thesis. The Bayesian method is yet to be applied in actual experiments, as it was not mature enough to be applied in the scientific cases of the current thesis.

Finally, the F-test was introduced in SAS to assess whether one model gives a significantly better description of data than alternative models.

The methods were applied in the study of four different protein complexes, as described in chapter 5. The methods I developed were used in SANS contrast variation studies. Three of the scientific cases used "invisible" detergents developed in our group. With the new methods, we could obtain novel information about the solution structure of the glutamate receptor GluA2, which is vital for nerve signaling and thus involved in many neurological disorders as well as healthy processes such as learning and memory (Paper III and IV). With the methods we could also probe the structure of an unknown equilibrium state of SERCA (Paper III). SERCA is important for a wide range of processes from antibody formation to neurotransmis-

sion. The methods were also applied to study the HTL complex in order to investigate the inner lipid core that facilitates translocation of transmembrane  $\alpha$ -helices into or pass the lipid bilayer (Paper V). Finally,  $\alpha$ SN, which is known to play a central role in neurodegenerative diseases, was studied (Appendix B). The water layer around fibrillated  $\alpha$ SN was investigated and found to be similar to that of other proteins. A hypothesized exchange of  $\alpha$ SN monomers between the solution and the  $\alpha$ SN fibrils was rejected using dynamic SANS measurements and contrast variation.

In summary, the methods developed in the current thesis allow SAS to be used for investigations of more challenging systems and problems, e.g. of protein complexes related to neurological diseases.

I hope this work has brought the field of structural biology a tiny, but significant, step in the right direction. If you, the reader, have reached this far, I thank you for your attention and encourage you to share and discuss your thoughts about the work with me and others.

# Chapter 7

## References

- Als-Nielsen, J. & McMorrow, D. (2011). *Elements of Modern X-ray Physics* (2nd ed., Wiley & Sons, Chichester), chap 1, 1–28.
- Akasaka, K., Latif, A. R., Nakamura, A., Matsuo, K., Tachibana, H. & Gekko, K. (2007). *Biochemistry* **18**, 10444–10450.
- Andrae, R. (2010a). *arXiv*: 1009.2755.
- Andrae, R. (2010b). *arXiv*: 1012.3754.
- Arleth, L. & Pedersen, J. S. (2001). *Phys. Rev. E* **63**, 061406.
- Barlow, J. (1999). *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences* (John Wiley & Sons, Chichester), chap. 8, 153–155.
- Berthaud, A., Manzi, J., Pérez, J. & Mangenot, S. (2012). *JACS*, **134**, 10080–10088.
- Blakeley, M. P., Hasnain, S. S., & Antonyuk, S. V. (2015). *IUCrJ*, **2**, 464–474.
- Botte, M., Zaccai, N. R., Nijeholt, J. L. á., Martin, R., Knoops, K., Papai, G., Zou, J., Deniaud, A., Karuppasamy, M., Jiang, Q., Roy, A. S., Schulten, K., Schultz, P., Rappsilber, J. & Zaccai, G. (2016). *Sci. Rep.* **6**, 38399.
- Bowie, D. (2008). *CNS Neurol. Disord. Drug Targets* **7**, 129–143.
- Bradley, D. E. (1962). *J. Gen. Microbiol.* **29**, 503–516.
- Bragg, W. L. (1913). *Proc. Royal Soc. Lond. A* **89**, 248–277.
- Breyton, C., Gabel, F., Lethier, M., Flayhan, A., Durand, G., Jault, J.-M., Juillan-Binard, C., Imbert, L., Moulin, M., Ravaud, S., Härtlein, M. & Ebel, C. (2013). *Eur. Phys. J. E* **36**, 71.
- Brünger, A. T. (1992). *Nature*. **355**, 472–475.
- Brünger, A. T., Clore, G. M., Gronenborn, A. M., Saffrich, R. & Nilges, M. (1993). *Science* **261**, 328–331.
- Carafoli, E. (2002). *PNAS* **99**, 1115–1122.
- Chen, L., Dürr, K. L. & Gouaux, E. (2014). *Science* **345**, 1021–1026.
- Cotton, J. P. (1996). *Introduction to Neutron Scattering, Lecture Notes of the Introductory*. (Course 1st European Conference on Neutron Scattering at Interlaken, Switzerland, October 6–11, 1996, Furrer, A., ed.), 144–161. [http://www.iaea.org/inis/collection/NCLCollectionStore/\\_Public/28/024/28024536.pdf](http://www.iaea.org/inis/collection/NCLCollectionStore/_Public/28/024/28024536.pdf) (visited August 3rd, 2018).

- Damaschun, G., Müller, J. J. & Pürschel (1968). *Monatshefte für Chemie* **99**, 2343–2348.
- Debye, P., Anderson, H. R. & Brumberger, H. (1957). *J. Appl. Phys.* **28**, 679–683.
- de Moivre (1738). *The Doctrine of Chances* (2nd ed., London.)
- Dobson, C. M. (2004). *Semin. Cell Dev. Biol.* **15**, 3–16.
- Doer, A. (2018). *Nat. Methods.* **15**, 33.
- Dürr, K. L., Chen, L., Stein, R. A., De Zorzi, R., Folea, I. M., Walz, T., Mchaourab, H. S. & Gouaux, E. (2014). *Cell* **158**, 778–792.
- Dyla, M., Terry, D. S., Kjaergaard, M., Sørensen, T. L., Andersen, J. L., Andersen, J. P., Knudsen, C. R., Altman, R. B., Nissen, P. & Blanchard, S. C. (2017). *Nature* **551**, 346–351.
- Egelman, E. H. (2016). *Biophys. J.* **110**, 1008–1012.
- Fawzi, N. L., Ying, J., Torchia, D. A. & Clore, G. M. (2010). *JACS* **132**, 9948–9951.
- Fernandez-Leiro, R., & Scheres, S. H. W. (2016). *Acta Cryst. D* **73**, 496–502.
- Fischer, H., de Oliveira Neto, M., Napolitano, H. B., Polikarpov, I. & Craievich, A. F. (2010). *J. Appl. Cryst.* **43**, 101–109.
- Franke, D. & Svergun, D. I. (2009). *J. Appl. Cryst.* **42**, 342–346.
- Franke, D., Jeffries, C. M. & Svergun, D. I. (2015). *Nat. Methods* **12**, 419–422.
- Franke, D., Petoukhov, M. V., Konarev, P. V., Panjkovich, A., Tuukkanen, A., Mertens, H. D. T., Kikhney, A. G., Hajizadeh, N. R., Franklin, J. M., Jeffries, C. M. & Svergun, D. I. (2017). *J. Appl. Cryst.* **50**, 1212–1225.
- Fritsch, B., Stott, J. J., Donofrio, J. J., & Rogawski, M. A. (2010). *Epilepsia*, **51**, 108–117.
- Frueh, D. P., Goodrich, A., Mishra, S. & Nichols, S. (2013). *Curr. Opin. Struct. Biol.* **23**, 734–739.
- Furukawa, A., Yoshikaie, K., Mori, T., Mori, H., Morimoto, Y. V., Sugano Y., Iwaki, S., Minamino, T., Sugita, Y., Tanaka, Y. & Tsukazaki, T. (2017). *Cell Rep.* **19**, 895–901.
- Gekko, K. & Noguchi, H. (1979). *J. Phys. Chem.* **83**, 2706–2714.
- Glatte, O. (1977). *J. Appl. Cryst.* **10**, 415–421.
- Goers, J., Permyakov, S. E., Permyakov, E. A., Uversky, V. N. & Fink, A. L. (2002). *Biochemistry.* **41**, 12546–12551.
- Guinier, A. & Fournet, G. (1955). *Small-angle scattering of X-rays* (Guinier, A. & Fournet, G. (ed.), John Wiley & Sons, London), chap 2, 5–82.
- Gull, S. F. (1989). *Maximum Entropy and Bayesian Methods* (Skilling, J. (ed.), Kluwer Academic Publishers, Cambridge), 53–71.
- Hamelryck, T. (2012). *Bayesian Methods in Structural Bioinformatics* (Hamelryck, T., Mardia, K., Ferkinghoff-Borg, J. (ed.), Springer, Heidelberg), chap. 1, 3–48.
- Hansen, S. & Pedersen, J. S. (1991). *J. Appl. Cryst.* **24**, 541–548.
- Hansen, S. (2000). *J. Appl. Cryst.* **36**, 1190–1196.
- Hansen, S. (2012). *J. Appl. Cryst.* **45**, 566–567.

- Hansen, S. (2014). *J. Appl. Cryst.* **47**, 1469–1471.
- Henriques, J., Arleth, L., Lindorff-Larsen, K. & Skepö, M. (2018). *J. Mol. Biol.* **430**, 2521–2539.
- Herguedas, B., García-Nafria, J., Cais, O., Fernández-Leiro, R., Krieger, J., Ho, H. & Greger, I. H. (2016). *Science* **352**, aad3873.
- Higgins, M. K. & Lea, S. M. (2017). *Curr. Opin. Struc. Biol.* **46**, 95–101.
- Huang, C., Wikfeldt, K. T., Tokushima, T., Nordlund, D., Harada, Y., Bergmann, U., Niebuhr, M., Weiss, T. M., Horikawa, Y., Leetmaa, M., Ljungberg, M. P., Takahashi, O., Lenz, A., Ojamäe, L., Lyubartsev, A. P., Shin, S., Pettersson, L. G. & Nilsson A. (2009). *PNAS* **106**, 15214–15218.
- Hub, J. S. (2018). *Curr. Opin Struc. Biol.* **49**, 18–26.
- Kaspersen, J. D., Jessen, C. M., Stougaard, B. V., Sørensen, E. S., Andersen, K. K., Glasius, M., Oliveira, C. L. P., Otzen, D. & Pedersen, J. S. (2014). *Chembiochem* **15**, 2113–2124.
- Kendrew, J. C., Bodo, G., Dintziz, H. M., Parrish, R. G., Wyckoff, H. & Phillips, D. C. (1958). *Nature* **181**, 662–666.
- Kharakoz, D. P. (2000). *Biophys. J.* **79**, 511–525.
- Knight, C. J. & Hub, J. S. (2015). *Nucleic Acids Res.* **43**, W225–W230.
- Konarev, P. V. & Svergun, D. I. (2015). *IUCrJ* **2**, 352–360.
- Koutsioubas, A. (2017). *Biophys. J.* **113**, 2373–2382.
- Kynde, S. A. R. (2014). *Modelling small-angle scattering data from complex protein-lipid systems* (University of Copenhagen, PhD Thesis).
- Larsen, A. L. (2016). *Structural Analysis of Unconventional Nanodiscs* (University of Copenhagen, Master Thesis).
- Larsen, A. N., Sørensen, K. K., Johansen, N. T., Martel, A., Kirkensgaard, J. J. K., Jensen, K. J., Arleth, L. & Midtgaard, S. R. (2016). *Soft Matter*, **12**, 5937–5949.
- Levenberg, K. (1944). *Q. Appl. Math.* **2**, 164–168.
- Lipskaia L., Chemaly, E. R., Hadri, L., Lompre, A.-M., & Hajjar, R. J. (2010). *Expert Opin. Biol. Ther.* **10**, 24–41.
- Lomize, M. A, Pogozheva, I.D., Joo, H., Mosberg, H. I. & Lomize, A. L. (2012). *Nucleic Acids Res.* **40**, D370–D376.
- Marasini, C. & Vestergaard, B. (2017). *Biological Small Angle Scattering: Techniques, Strategies and Tips* (Chaudhuri B., Muñoz I., Qian S., Urban V. (ed.), Advances in Experimental Medicine and Biology, vol. 1009., Springer, Singapore), chap. 9, 149–165.
- Marquardt, D. W. (1963). *SIAM J. Appl. Math.* **11**, 431–441.
- Merk, A., Bartesaghi, A., Banerjee, S., Falconieri, V., Rao, P., Davis, M., Pragani, R., Boxer, M., Earl, L. A., Milne, J. L. S. & Subramaniam, S. (2016). *Cell*, **165**, 1698–1707.
- Merzel, F. & Smith, J. C. (2002). *PNAS* **99**, 5378–5383.
- Meyerson, J. R., Kumar, J., Chittori, S., Rao, P., Pierson, J., Bartesaghi, A., Mayer, M. L. & Subramaniam, S. (2014). *Nature* **514**, 328–334.

- Meyerson, J. R., Chittori, S., Merk, A., Rao, P., Han, T. H., Serpe, M., Mayer, M. L. & Subramaniam, S. (2016). *Nature* **537**, 567–571.
- Midtgaard, S. R., Darwish, T. A., Pedersen, M. C., Huda, P., Larsen, A. H., Jensen, G. V., Kynde, S. A. R., Skar-Gislinge, N., Nielsen, A. J. Z., Olesen, C., Blaise, M., Dorosz, J. J., Thorsen, T. S., Venskutonyte, R., Krintel, C., Møller, J. V., Friehlinghaus, H., Gilbert, E. P., Martel, A., Kastrup, J. S., Jensen, P. E., Nissen, P. & Arleth, L. (2018). *FEBS*, **283**, 357–371.
- Millero, F. J., Curry, R. W., & Drost-Hansen, W. (1969). *J. Chem. Eng. Data* **14**, 422–425.
- Moore, P. B. (1980). *J. Appl. Cryst.* **13**, 168–175.
- Møller, J. V., Olesen, C., Winther, A.-M. L. & Nissen, P. (2010). *Q. Rev. Biophys.*, **43**, 501–566.
- Mylonaz, E. & Svergun, D. I. (2007). *J. Appl. Cryst.* **40**, 245–249.
- Nakano, M., Fukuda, M., Kudo, T., Miyazaki, M., Wada, Y., Matsuzaki, N., Endo, H. & Handa, T. (2009). *JACS* **131**, 8308–8312.
- Nielsen, S. B., Macchi, F., Raccosta, S., Langkilde, A. E., Giehm, L., Kyrsting, A., Svane, A. S., Manno M., Christiansen, G., Nielsen, N.C., Oddershede, L., Vestergaard, B. & Otzen, D. E. (2013). *PLoS One* **8**: e67713.
- Oliver, R. C., Pingali, S. V. & Urban, V. S. (2017). *J. Phys. Chem. Lett.* **8**, 5041–5046.
- Opella, S. J. (2015). *J. Magn. Reson.* **253**, 129–137.
- Pedersen, J. S., Posselt, D. & Mortensen, K. (1990). *J. Appl. Cryst.* **23**, 321–333.
- Pedersen, J. S. (1997). *Adv. Colloid Interface Sci.* **70**, 171–210.
- Pedersen, M. C., Hansen, S. L., Markussen, B., Arleth, L. & Mortensen, K. (2014). *J. Appl. Cryst.* **47**, 2000–2010.
- Pedersen, M. C. (2014). *Mathematical, computational, and statistical aspects of model refinement from small-angle scattering data* (University of Copenhagen, PhD thesis), chap. 5, 70–72.
- Pérez, J. & Nishino, Y. (2012). *Curr. Opin. Struct. Biol.* **22**, 670–678.
- Pérez, J. & Koutsioubas, A. (2015). *Acta Cryst. D* **71**, 86–93.
- Persson, F., Söderhjelm, P. & Halle, B. (2018). *J. Chem. Phys.* **148**, 215101.
- Persson, F. & Halle, B. (2018). *J. Chem. Phys.* **148**, 215102.
- Petoukhov, M. V., Konarev, P. V., Kikhney, A. G. & Svergun, D. I. (2007). *J. Appl. Cryst.* **40**, 223–228.
- Petoukhov, M. V. & Svergun, D. I. (2015). *Acta Cryst. D* **71**, 1051–1058.
- Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., Gorba, C., Mertens, H. D. T., Konarev, P. V. & Svergun, D. I. (2012). *J. Appl. Cryst.* **45**, 342–350.
- Porod, G. (1951). *Kolloid-Z.* **124**, 5183–114.
- Porod, G. (1982). *Small Angle X-ray Scattering* (Glatter, O. & Kratky, O. (ed.), Academic Press, London), chap. 2, 17–52.
- Powell, M. J. D. (1964). *Comput. J.* **7**, 155–162.
- Rambo, R. & Tainer, J. A. (2011). *Biopolymers* **95**, 559–571.



- Rambo, R. & Tainer, J. A. (2013). *Nature* **496**, 477–481.
- Robert, H. L. & Brown, D. R. (2015). *Biomolecules* **5**, 282–305.
- Schneidman-Duhovny, D., Hammel, M. & Sali, A. (2010). *Nucleic Acids Research*, **38**, W540–W544.
- Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. (2013). *Biophys. J.* **105**, 962–974.
- Shannon, C. E. (1949). *Proc. IEEE* **86**, 447–457.
- Sobolevsky, A. I., Rosconi, M. P., & Gouaux, E. (2009). *Nature* **462**, 745–756.
- Spalla, O. (2002). *Neutrons, X-rays and Light: Scattering Methods Applied to Soft Condensed Matter* (Lindner, P. & Zemb, T. (ed.), Elsevier, Amsterdam), chap. 3, 49–72.
- Squire, P. G., & Himmel, M. E. (1979). *Arch. Biochem. Biophys.*, **196**, 165–177.
- Stefanis, L. (2012). *Cold Spring Harb Perspect Med.*, **2**, a009399.
- Svergun, D. I. (1992). *J. Appl. Cryst.* **25**, 495–503.
- Svergun, D. I., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.
- Svergun, D. I., Volkov, V. V., Kozin, M. B. & Stuhrmann, H. B. (1996). *Acta Cryst. A* **52**, 419–426.
- Svergun, D. I., Richard, S., Koch, M. H., Sayers Z., Kuprin, S. & Zaccai, G. (1998). *PNAS*, **95**, 2267–2272.
- Svergun, D. I. (1999). *Biophys. J.* **76**, 2879–2886.
- Svergun, D. I., Koch, M. H. J., Timmins, P. A., & May, R. P. (2013). *Small Angle X-Ray and Neutron Scattering from Solutions of Biological Macromolecules*. (Oxford University Press.), chap. 3, 65–66.
- Taupin, D. & Luzzati, V. (1982). *J. Appl. Cryst.* **15**, 289–300.
- Taylor, J. R. (1997). *An Introduction to Error Analysis: The study of uncertainties in physical measurements* (2nd ed., University Science Books, California), chap. 12, 245–273.
- Tokuda, J. M., Pabit, S. A. & Pollack, L. (2016). *Biophys. Rev.* **8**, 139–149.
- Trewhella, J., Duff, A.P., Durand, D., Gabel, F., Guss, J. M., Hendrickson, W. A., Hura, G. L., Jacques, D. A., Kirby, N. M., Kwan, A. H., Pérez, J., Pollack, L., Ryan, T. M., Sali, A., Schneidman-Duhovny, D., Schwede, T., Svergun, D. I., Sugiyama, M., Tainer, J. A., Vachette, P., Westbrook, J. & Whitten A. E. (2017). *Acta Cryst. D* **73**, 710–728.
- Tuttle, M. D., Comellas, G., Nieuwkoop, A. J., Covell, D. J., Berthold, D. A., Kloepper, K. D., Courtney, J. M., Kim, J. K., Barclay, A. M., Kendall, A., Wan, W., Stubbs, G., Schwieters, C. D., Lee, V. M., George, J. M. & Rienstra, C. M. (2016). *Nat. Struct. Mol. Biol.* **23**, 409–415.
- Tuukkanen, A. T., Kleywegt, G. J. & Svergun, D. I. (2016). *IUCrJ* **3**, 440–447.
- Twomey, E. C., Yelshanskaya, M. V., Grassucci, R. A., Frank, J., & Sobolevsky, A. I. (2016). *Science*, **353**, 83–86.
- Twomey, E., Yelshanskaya, M. V., Grassucci, R. A., Frank, J. & Sobolevsky, A. I. (2017a). *Neuron* **94**, 569–580.
- Twomey, E., Yelshanskaya, M. V., Grassucci, R. A., Frank, J. & Sobolevsky, A. I. (2017b). *Nature* **549**, 60–65.
- Twomey, E. & Sobolevsky, A. (2018). *Biochemistry*. **57**, 267–276.

- Uversky V. N. (2003). *J. Biomol. Struct. Dyn.* **21**, 211–34.
- VanAken, T., Foxall-VanAken, S., Castleman, S. & Ferguson-Miller, S. (1986). *Methods Enzymol.* **125**, 27–35.
- Vénien-Bryan, C., Li, Z., Vuillard, L. & Boutin, J. A. (2017). *Acta Cryst. F* **73**, 174–183.
- Vestergaard, B. & Hansen, S. (2006). *J. Appl. Cryst.* **39**, 797–804.
- Whitten, A. E., Caib, S. & Trehwella, J. (2008). *J. Appl. Cryst.* **41**, 222–226.
- Wikipedia (2018). *Glutamate receptor*. [https://en.wikipedia.org/wiki/Glutamate\\_receptor](https://en.wikipedia.org/wiki/Glutamate_receptor) (visited July 6th, 2018).
- Wikipedia (2018). *Wald–Wolfowitz runs test*. [https://en.wikipedia.org/wiki/Wald%E2%80%93Wolfowitz\\_runs\\_test](https://en.wikipedia.org/wiki/Wald%E2%80%93Wolfowitz_runs_test) (visited June 17th, 2018).
- Wlodawer A., Li, M. & Dauter, Z. (2017). *Structure.* **25**, 1589–1597.
- Yelshanskaya, M. V., Li, M. & Sobolevsky, A. I. (2014). *Science* **345**, 1070–1074.
- Yelshanskaya, M. V., Singh, A.K., Sampson, J. M., Narangoda, C., Kurnikova, M. & Sobolevsky, A. I. (2016). *Neuron* **91**, 1305–1315.
- Zemb, T., Tache, O., Né, F. & Spalla, O. (2003). *J. Appl. Cryst.* **36**, 800–805.
- Zhao, Y., Chen, S., Yoshioka, C, Bacongus, I. & Gouaux, E. (2016). *Nature* **536**, 108–111.
- Zhou, Z, Liu, A, Xia, S, Leung, C., Qi, J., Meng, Y., Xie, W., Park, P, Collingridge, G. L., Jia, Z. (2018). *Nat. Neurosci.* **21**, 50–62.

# Chapter 8

## Appendices

### 8.1 Appendix A: Scattering intensity in the continuous limit and the form factor

In this appendix, the scattering intensity is expressed in the continuous limit, and it is shown that  $I(0) = n\Delta\rho^2 V_m^2$  for homogeneous particles. Also, the form factor is introduced

We start by equation (2.8). That is, we have assumed that the sample consists of diluted, identical and monodisperse macromolecules with  $N_s$  point scatterers in each macromolecule. Then the intensity is given as:

$$I(q) = n \sum_{j,k=1}^{N_s} \Delta b_j \Delta b_k \cdot \text{sinc}(qd), \quad (8.1)$$

where  $n$  is the number density of the macromolecules,  $\Delta b_i$  is the excess scattering length of the  $i$ th scatterer,  $q$  is the magnitude of the scattering vector, and  $d$  is the distance between the  $i$ th and the  $j$ th scatterer. I have changed notation for the appendix from  $r$  to  $d$  as  $r$  will be used as a spacial coordinate. Equation (8.1) can be rewritten in continuous form by replacing  $\Delta b_j$  by  $\Delta\rho(\mathbf{r})dV_m$ , where  $\Delta\rho(\mathbf{r})$  is the excess scattering length density of an infinitesimal volume element  $dV_m$  at position  $\mathbf{r}$ . The double sum is replaced by a double integral over the volume of the macromolecule,  $V_m$ :

$$I(q) = n \int_{V_m} \int_{V'_m} \Delta\rho(\mathbf{r}) \Delta\rho(\mathbf{r}') \cdot \text{sinc}(qd) dV_m dV'_m, \quad (8.2)$$

where  $d = |\mathbf{r} - \mathbf{r}'|$ . This expressed the simple case of spherical macromolecules, where the exponential can be rewritten as  $\text{sinc}(qd)$ . When  $q \rightarrow 0$ , then  $\text{sinc}(qr)$  is unity, so the forward scattering,  $I(0)$  is given as:

$$I(0) = n \int_{V_m} \int_{V'_m} \Delta\rho(\mathbf{r}) \Delta\rho(\mathbf{r}') dV_m dV'_m \quad (8.3)$$

$$= n \left( \int_{V_m} \Delta\rho(\mathbf{r}) dV_m \right)^2 \quad (8.4)$$

$$= n \left( \frac{\int_{V_m} \Delta\rho(\mathbf{r}) dV_m}{\int_{V_m} dV_m} \cdot \int_{V_m} dV_m \right)^2 \quad (8.5)$$

$$= n \Delta\rho_{\text{mean}}^2 V_m^2. \quad (8.6)$$

For homogeneous particles  $\Delta\rho(\mathbf{r}) = \Delta\rho_{\text{mean}} = \Delta\rho$ , such that the forward scattering is given by:

$$I(0) = n \Delta\rho^2 V_m^2. \quad (8.7)$$

The full intensity (for a homogeneous particle) can be written as:

$$I(q) = n\Delta\rho^2 V_m^2 P(q), \quad (8.8)$$

where  $P(q)$  is the form factor:

$$\begin{aligned} P(q) &= \frac{\int_{V_m} \int_{V'_m} \text{sinc}(qd) dV_m dV'_m}{V_m^2} \\ &= \frac{\int_{V_m} \int_{V'_m} \text{sinc}(qr') \text{sinc}(qr) dV_m dV'_m}{V_m^2} \\ &= \frac{1}{V_m^2} \left( \int_{V_m} \text{sinc}(qr) dV_m \right)^2 \end{aligned} \quad (8.9)$$

The form factor depends on the the shape (but not the size) of the macromolecule, via the integration limits. Analytical expressions of  $P(q)$  have been derived for a range of geometrical objects (Pedersen 1997). The form factor has the limits  $P(q \rightarrow 0) = 1$  and  $P(q \rightarrow \infty) = 0$ . For a sphere, for example, the integral in (8.9) is given by:

$$\begin{aligned} \int_{V_m} \text{sinc}(qr) dV_m &= \int_0^R \int_0^\pi \int_0^{2\pi} \text{sinc}(qr) r^2 \sin(\phi) dr d\phi d\theta \\ &= 4\pi \int_0^{qR} x \sin(x) r^2 dx \\ &= 3V_m [\sin(x) - x \cos(x)]/x^3, \end{aligned} \quad (8.10)$$

where the substitution  $x = qr$  was used in the second step and the last step was solved using integration by parts.  $V_m$  is the volume of a sphere  $4\pi R^3/3$ , such that:

$$P(q) = (3[\sin(x) - x \cos(x)]/x^3)^2, \quad (8.11)$$

## 8.2 Appendix B: Experimental report from the study of $\alpha$ -synuclein

**Small-angle neutron scattering reveals the dimensions of the hydration layer around alpha-synuclein fibrils and shows that protein-exchange between fibrils is very limited**

**List of Authors** Andreas Haahr Larsen, Carlotta Marasini, Bente Vestergaard & Lise Arleth.

**Status** Report, to be used in further studies.

**Abstract** In this study, we used contrast variation in small-angle neutron scattering (SANS), to investigate  $\alpha$ -synuclein ( $\alpha$ SN) fibrils. Three aspects were studied. Firstly, the monomer-fibril exchange was studied by monitoring the change in SANS signal from a sample of deuterated and hydrogenated  $\alpha$ SN fibrils. The signal would change upon exchange. No measurable exchange was however observed within 48 hours. Secondly, the water layer around  $\alpha$ SN was investigated in order to find its density and extend. Previous SAXS studies (Nielsen et al, *PLOP ONE*, 8, e67713) indicated the presence of a particular dense and extended water layer in the vicinity of the  $\alpha$ SN fibrils. The present SANS data were however consistent with a conventional water layer, as observed around small soluble proteins extending less than 3 Å away from the protein and being about 10% more dense than bulk water. Finally, the structure of the structural subunits of  $\alpha$ SN fibrils were investigated by deuterating minor part of the fibril and observing them in a matrix of matched-out hydrogenated  $\alpha$ SN fibrils. As judged by the pair distance distribution functions, we conclude that the measured signal has contributions both from the subunit and from the nearly matched out fibril, so the structure of the fibril subunits could not be deduced from data.

**Contributions** AHL, LA, CM and BC collected the SANS data, which were analyzed by AHL. AHL wrote the report.

## Report

# Studies of alpha synuclein using small-angle neutron scattering and contrast variation

*Andreas Haahr Larsen, Carlotta Marassini, Ersoy Cholak, Bente Vestergaard & Lise Arleth*

## Abstract

In this study, we used contrast variation in small-angle neutron scattering (SANS), to investigate  $\alpha$ -synuclein ( $\alpha$ SN) fibrils. Three aspects were studied. Firstly, the monomer-fibril exchange was studied by monitoring the change in SANS signal from a sample of deuterated and hydrogenated  $\alpha$ SN fibrils. If there was exchange, then the amplitude of the forward scattering would change. No measurable change was however observed within 48 hours. Secondly, the water layer around  $\alpha$ SN was investigated in order to find its density and extent. Previous SAXS studies (Nielsen et al, PLOS ONE, 8, e67713) indicated the presence of a particular dense and extended water layer in the vicinity of the  $\alpha$ SN fibrils. The present SANS data were however consistent with a conventional water layer, as observed around small soluble proteins extending less than 3 Å away from the protein and being about 10% more dense than bulk water. Finally, the structure of the structural subunits of  $\alpha$ SN fibrils were investigated by deuterating a minor part of the fibril and observing them in a matrix of matched-out hydrogenated  $\alpha$ SN fibrils. As judged by the pair distance distribution functions, we conclude that the measured signal has contributions both from the subunit and from the nearly matched out fibril, so the structure of the fibril subunits could not be deduced from data.

Updated:

Friday, August 31, 2018

## Materials and Methods

### Protein production and purification

Performed by Carlotta Marasini and Ersoy Cholak, reported elsewhere

### Sample preparation for small-angle neutron scattering experiments

For SANS experiment 1, the samples were prepared in H<sub>2</sub>O based buffer and dialyzed to obtain the desired D<sub>2</sub>O/H<sub>2</sub>O content in the final buffer. For the subsequent experiments,  $\alpha$ SN was lyophilized and redispersed in the relevant buffer prior to the SANS measurements.

### Negative-stain electron microscopy (EM)

Carlotta Marasini, reported elsewhere. Few results reported here for comparison with the SANS data.

### Small-angle neutron scattering (SANS) data collection

SANS data were measured at three different instruments, at four independent beamtimes.

#### *SANS Experiment 1, PSI, May 2015*

At the first beamtime, in May 2015 at the SANS1 beamline at SINQ (PSI, Villigen), we measured a contrast-match series, a samples for studying the monomer-fibril exchange and two samples with deuterated  $\alpha$ SN in a matrix of non-deuterated  $\alpha$ SN.

A velocity selector was used to obtain a neutron beam with wavelength  $\lambda = 6.0 \text{ \AA} \pm 10\%$  (FWHM). The sample was measured in 1 mm sandwich cuvettes with quartz windows and a window diameter of 15 mm. The beam was collimated to a diameter of 10 mm. The cuvette was rotating with 12 turns per minute to avoid precipitation. The optimal rotation speed was established at the beamtime. Samples were measured at 15°C. Data were measured in two settings with a sample to detector distance and collimation length (SD/C) of 2m/2m and 8m/8m to obtain data in the  $q$ -range 0.009 to 0.3  $\text{\AA}^{-1}$ . Data were radially averaged, buffer subtracted and normalized with H<sub>2</sub>O as standard using the BerSANS software available at the beamline.

For the contrast match series,  $\alpha$ SN samples with a concentration of 10 mg/ml were measured at 20%, 40%, 50%, 60% and 100% D<sub>2</sub>O.

Exchange of monomers was probed by mixing equal amounts of deuterated and hydrogenated fibrils in 60% D<sub>2</sub>O based buffer and measure over time. The estimated degree of deuteration in the deuterated  $\alpha$ SN fibrils was 74%. The hydrogenated fibrils had negative contrast and the deuterated fibrils had positive contrast. We expected the scattering at low  $q$  to decrease if exchange took place, since a half deuterated and half hydrogenated fibril would be almost matched out on average at 60% D<sub>2</sub>O. The sample concentration was estimated to be 35 mg/ml.

Two samples with co-fibrillated  $\alpha$ SN (7% deuterated and 93% hydrogenated) were measured around the match point, at 40% and at 50% D<sub>2</sub>O. Both samples had a protein concentration of 75 mg/ml, meaning that the concentration of deuterated  $\alpha$ SN was 5.3 mg/ml. At these concentrations, the samples were gel-like, and macroscopic bubbles were present in the sample.

*SANS Experiment 2, PSI, May 2016*

The second beamtime, in May 2016, was also performed at the SANS-1 beamline at SINQ. The samples had been prepared using a new protocol (see sample purification) to gain better control over the protein concentration and D<sub>2</sub>O content in the samples.

The instrumental setup was the same as for the first experiment at SANS1, but the sample was measured at ambient temperatures.

The contrast match series was done again. Samples of hydrogenated  $\alpha$ SN fibrils were measured at 0%, 20%, 30%, 60%, 80% and 100% D<sub>2</sub>O at 10 mg/m.

Moreover, a sample of hydrogenated  $\alpha$ SN fibrils was measured at the match point, which was determined to be at 39%. This sample had a concentration of about 44 mg/ml.

A sample of coaggregated deuterated and hydrogenated  $\alpha$ SN was also measured at the match point, with a concentration of 77 mg/ml.

Microscopic air bubbles were present in the two highly concentrated samples. Due to bubbles, less sample was illuminated by the beam. This was accounted for by changing the sample thickness to an effective sample thickness below 1 mm, as estimated by the transmission of the sample. Possible reflections could however not be accounted for.

*SANS Experiment 3, MLZ, December 2016*

The third beamtime, in December 2016 was performed at the KWS-1 beamline at FRM2 (MLZ, Munich).

A velocity selector was used to obtain a wavelength of  $\lambda = 5.0 \text{ \AA} \pm 10 \%$  (FWHM). The sample was measured in a 1 mm square Hellma quartz cuvette at 15 °C. Data were measured in four settings with SD/C of 20m/20m, 8m/8m, 4m/4m and 1.5m/4m to obtain data in the q-range 0.003 to 0.35  $\text{\AA}^{-1}$ . Data were radially averaged, buffer subtracted and normalized with plexiglass as standard using the qiKWS software available at the beamline.

A sample of hydrogenated  $\alpha$ SN fibrils was measured at the match point, at 39% D<sub>2</sub>O. The sample had a concentration of 10 mg/ml. Due to precipitation, a slight concentration gradient emerged in the sample during the measurement. The concentration gradient could be observed visually due to ThT flourophores present in the sample, which emits green flouroscent light in the presence of fibrils. The gradient meant that the concentration of the illuminated (lower) part of the sample was slightly larger than the mean concentration in the sample (10 mg/ml).



*SANS Experiment 4, PSI, August 2017*

The fourth beamtime, in August 2017, was performed at the SANS-2 beamline at SINQ (PSI, Villigen), using  $\lambda = 4.8 \text{ \AA} \pm 10\%$  (FWHM). Samples were measured in rotating sandwich cuvettes with rotation speed of 12 rounds per minute. Two samples were measured in 1 mm cuvettes, and one in 2 mm cuvettes. The beam was collimated to a diameter of 12 mm. This increase from 10 mm to 12 mm gave some problems with flares as the beam hit the edge of the cuvette. Measurements with flares were remeasured. Two settings were used with S/D of 2.5m/3.0m and 6m/6m. Data were radially averaged, buffer subtracted and normalized with H<sub>2</sub>O using the BerSANS software available at the beamline.

Two samples were measured close to the match-point, fibrillated  $\alpha$ SN at 34% D<sub>2</sub>O and fibrillated  $\alpha$ SN at 43% D<sub>2</sub>O, and one sample was measured at the match-point, at 39% D<sub>2</sub>O.

**Determining the SANS contrast match-point**

To investigate the water layer, the match-point of the protein fibrils (with water layer), had to be determined. Samples were measured in a range of different H<sub>2</sub>O/D<sub>2</sub>O mixtures, and the normalized contrast  $\widetilde{\Delta\rho}$  plotted (Fig. 1), to find the match-point

$$\widetilde{\Delta\rho}_X (\alpha\text{SN in } X\% \text{ D}_2\text{O}) = \sqrt{\frac{I_X(q)/c_X}{I_{100}(q)/c_{100}}} = \sqrt{\frac{\Delta\rho_X^2}{\Delta\rho_{100}^2}} = \frac{\Delta\rho_X}{\Delta\rho_{100}}.$$

Where  $I_X(q)$ ,  $c_X$  and  $\Delta\rho_X$  are intensity, concentration and excess scattering length density (contrast) for  $\alpha$ SN fibrils in X% D<sub>2</sub>O. Below the match-point, the protein contrast is negative and positive above. It was assumed that the volume and form factor of the fibrils were independent of the D<sub>2</sub>O/H<sub>2</sub>O content.

**Pair distance distribution functions**

The pair distance distribution functions ( $p(r)$ ) were obtained with the Bayesian indirect Fourier transform algorithm implemented in BayesApp (Hansen, 2014).

**Guinier analysis**

Guinier analysis was performed using a home-written MATLAB script.

**Analytical modelling**

Data were compared with an analytical model using WillItFit (Pedersen et al, 2013).

## Results

### Contrast match series from experiment 1

Only the samples at 20% and 100% D<sub>2</sub>O gave significant signal over background. It was suspected that the sample concentrations were lower than anticipated due to loss of material during buffer exchange to deuterated buffers. The match-point was determined to be at 43.4% D<sub>2</sub>O. However, the experiment was repeated in experiment 2.

### Contrast match series from experiment 2

The match-point was determined to be at 39% D<sub>2</sub>O (Fig. 1).

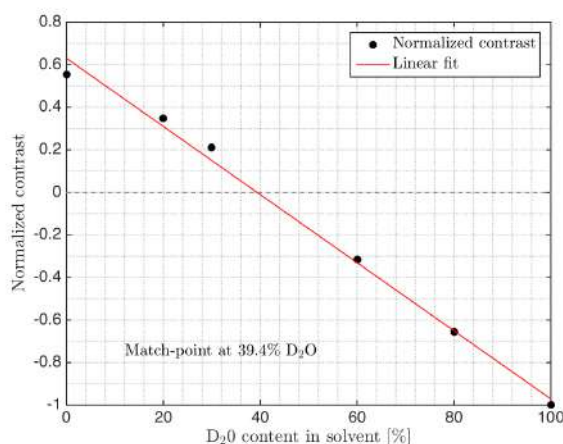


Figure 1. Contrast match series for fibrillated  $\alpha$ SN. The match-point was found at 39% D<sub>2</sub>O in the sample. The points at 0%, 20% and 30% deviated slightly from the linear tendency as a result of the relative low signal to noise ratio, due to incoherent scattering from the H<sub>2</sub>O-dominated solvent. All samples had a protein concentration of 10 mg/ml.

### Contrast match revisited

The measurements around the match point measured at KWS1@FRM2 (39% D<sub>2</sub>O) and at SANS2@SINQ (34%, 39% and 43% D<sub>2</sub>O) were added to the contrast match series. Including these points updated the best estimate of the match point to be at 40% D<sub>2</sub>O (Fig. 2)

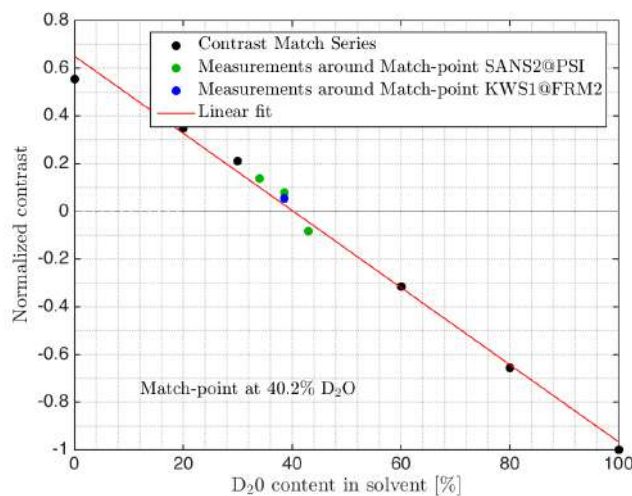


Figure 2. Contrast match series of  $\alpha$ SN fibrils. Including the measurements around the match-point in the contrast variation series shifts the predicted match-point from 39% to 40% D<sub>2</sub>O in the solvent.

### Exchange

No change was observed within 48 hours, as the measured SANS signal was unchanged (within the uncertainty) after 12 min, 24 min, 36 min, 15 hours and 48 hours (Fig. 3). Monomer-fibril exchange would result in a significant decrease of the forward scattering.

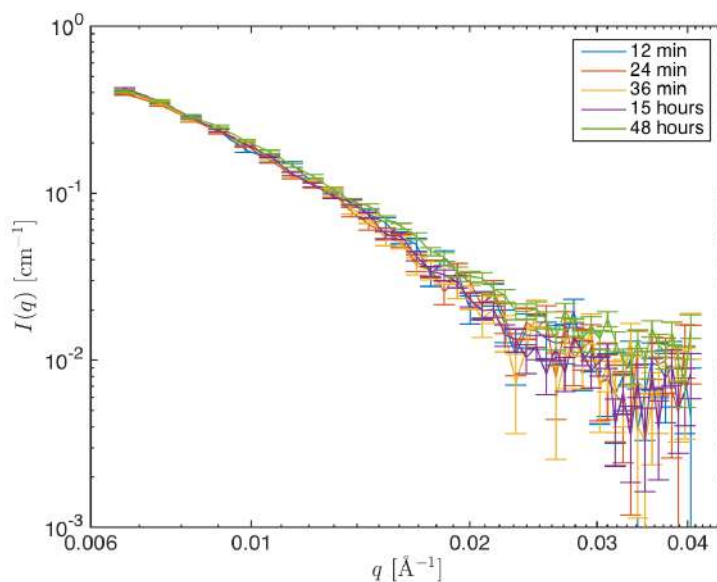


Figure 3. 75% perdeuterated fibrils and fully hydrogenated fibrils were mixed. The SANS signal after 12 min (blue), 24 min (red), 15 hours (purple), and 48 hours (green) after mixing.

### Probing the water layer thickness and density

If an extended measurable water layer existed, the  $\alpha$ SN fibrils would effectively resemble an elongated core-shell particles with a protein core and a shell of dense water (as illustrated in the inserts in Fig. 4). When matched out, such particle will have zero forward scattering, whereas the scattering at larger values of  $q$  is non-zero (purple curves in Fig. 4). The hydrogenated  $\alpha$ SN fibrils were measured at matched out conditions at KWS1 (Experiment 3). Despite measuring for almost 12 hours on a 10 mg/ml sample, no significant signal was observed at medium or large  $q$ -values (lower curve in Fig. 4). It thus deviated from the simulations with a large and dense water layer (Fig. 4D).

In an attempt to obtain observable evidence for the existence of a water layer, two measurements were collected around the match-point, at 34% D2O and at 43% D2O, where the scattering from the water layer has a relatively large effect on the total scattering as compared to the contribution from the protein. As seen from Fig. 5B and 5D, the scattering varies strongly on each side of the match-point, when the density is high (in those plots,  $\Delta\rho_{wl}$  is 40% that of the scattering length density of bulk water,  $\rho_w$ ). The difference is however small for a "normal" water layer (Svergun et al., 1998) with  $\Delta\rho_{wl}$  of 10% and a thickness of  $T=3$  Å (Fig. 4A).

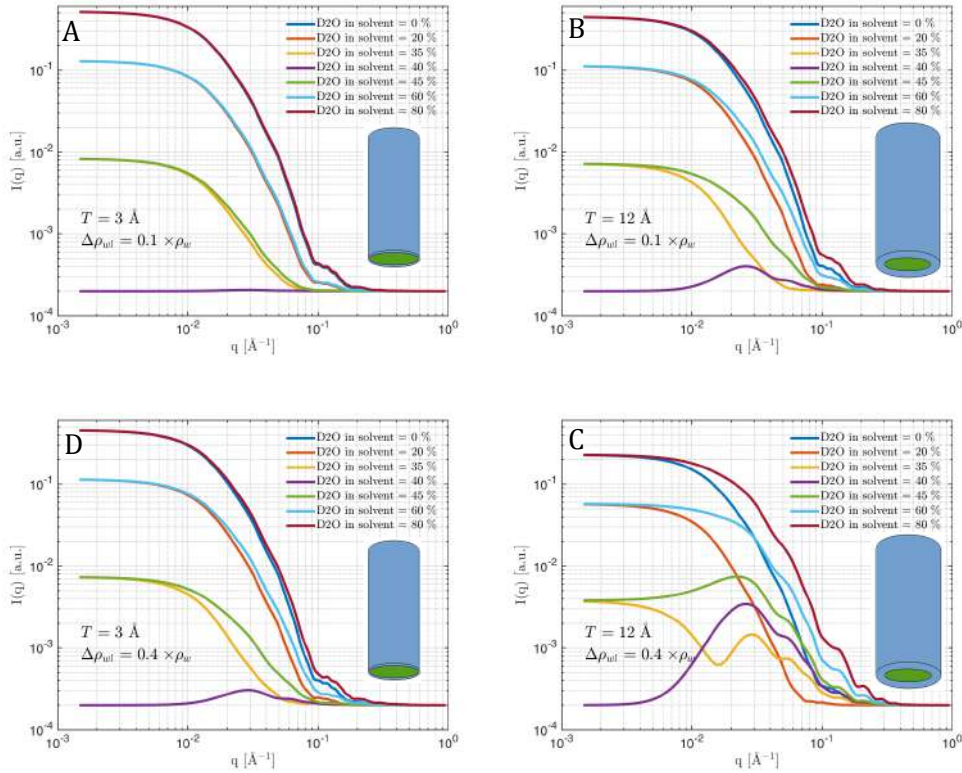


Figure 4. Theoretical curves for core-shell cylinders measured in SANS. The core represents the  $\alpha$ SN protein fibrils, and the shell represents the water layer. A) Water layer thickness,  $T$ , is 12 Å, and excess scattering length density, of the water layer is,  $\Delta\rho_{wl}$ , is 10% that of the scattering length density of bulk water,  $\rho_w$ , corresponding to a "normal" water layer (Svergun et al., 1998). B)  $T = 12$  Å and  $\Delta\rho_{wl} = 10\%$ . C)  $T = 3$  Å and  $\Delta\rho_{wl} = 40\%$ . D)  $T = 12$  Å and  $\Delta\rho_{wl} = 40\%$ .

To harvest the maximal amount of information from data about the water layer, all SANS data from the contrast match series, the measurement at the match-point and the two measurements close the match-point were fitted simultaneously. They were simultaneously fitted with a model of elongated core-shell cylinders with an elliptical cross section as shown in the inserts in Fig. 4. The core represents the fibrillated  $\alpha$ SN protein and the shell represents the water layer.

The thickness and density of the water layer are the parameter of interest. Effects of varying these can be seen in Fig. 4. Scaling parameters and backgrounds were fitted individually to all dataset. The backgrounds differed due to difference in the incoherent scattering and experimental conditions, and a scaling parameter was needed as the concentrations were not very well determined. The minor and major axis were shared fitting parameters for all dataset. The thickness and density of the water layer did however affects the scattering at all levels of D<sub>2</sub>O, as shown with simulations in Fig. 4. The value of  $\Delta\rho_{WL}$  is expected to be in the vicinity of  $0.1 \cdot \rho_W$  for protein (Svergun et al, 1998), but has been proposed to be larger for  $\alpha$ SN (Nielsen et al., 2013). Letting both parameters be free lead to an unphysical solutions. Therefore,  $\Delta\rho_{WL}$  was fixed to the expected value of  $0.1 \cdot \rho_W$  and  $T$  was then refined to a value of  $2.3 \pm 0.4 \text{ \AA}$ , which is consistent with prior knowledge.

The model fitted the data relatively well to data up to a  $q$ -value of about  $0.05 \text{ \AA}^{-1}$  (Fig. 5, reduced  $\chi^2 = 13.1$ ). It is not surprising that the simplistic elliptical model does not fit to data at higher values of  $q$ , representing the fine structure. However, the fit at low- $q$  should be sufficient for this purpose, as the most significant changes induced by the water layer for the model are in the intermediate- $q$  region (Fig. 4).

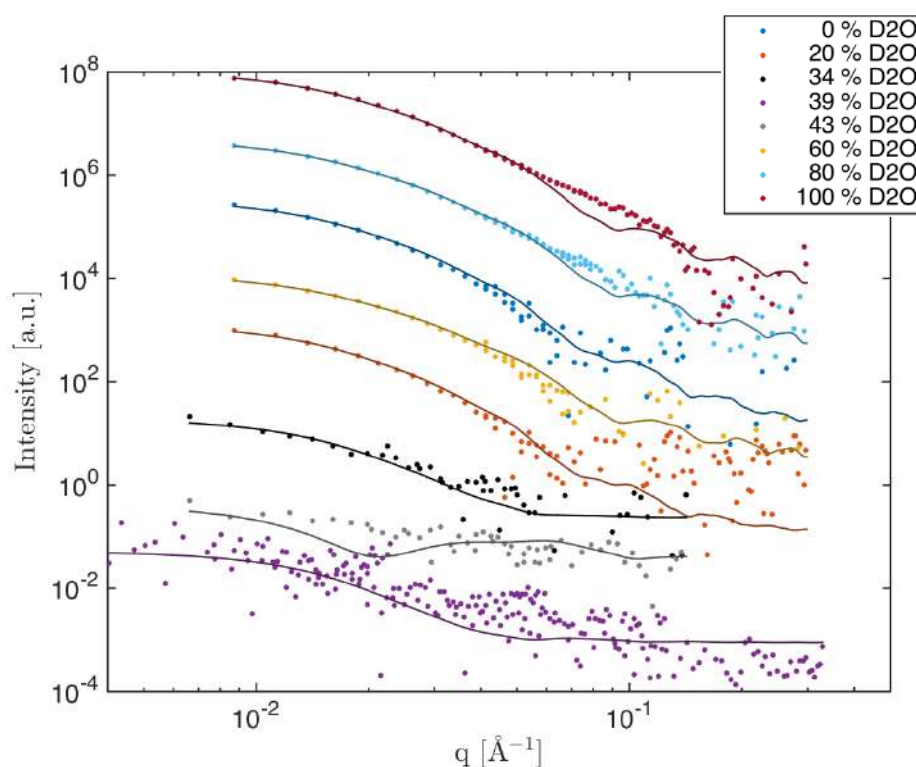


Figure 5. Simultaneous fit to all datasets with the model of elliptical cylinders with a water layer.

The refined values for the dimensions of the fibrils were  $44 \pm 4$  Å and  $115 \pm 7$  Å for the minor and major core axis respectively. These values were fairly consistent with negative stain electron microscopy (EM) data (Fig. 6), from which the minor and major axes were estimated to be around 50 Å and 110 Å respectively. The ellipticity comes from the double strand structure of  $\alpha$ SN fibrils (Giehm et al, 2011)

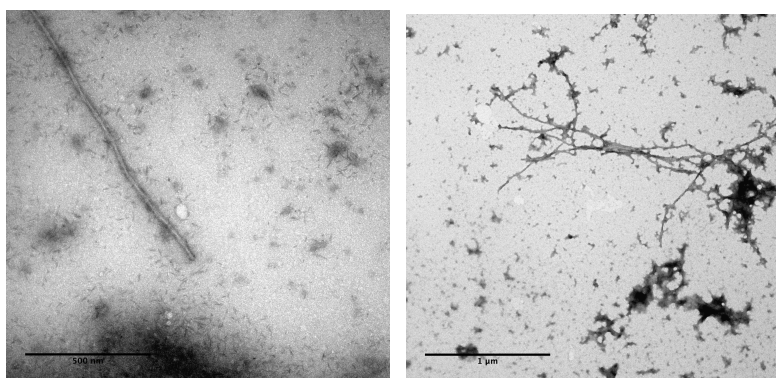


Figure 6. Negative stain TEM of hydrogenated  $\alpha$ SN fibrils. Scale bars are 500 nm (left) and 1  $\mu$ m (right).

The SANS data is consistent with a model of  $\alpha$ SN fibrils with a conventional water layer with about 10% higher density than bulk water and a thickness of about 3 Å. Such water layer is similar to what has been observed for small soluble proteins (Svergun et al, 1998). A change of the water layer density can thus not explain the SAXS results.

### Monomer structure in fibril matrix

Co-fibrillated  $\alpha$ SN were prepared with 7% deuterated  $\alpha$ SN (degree of deuteration was 75%), and 93% hydrogenated  $\alpha$ SN. The sample was measured at the match-point at 39% D2O (Experiment 2, Fig. 7). The SANS data (Fig. 7A) is clearly different for the co-fibrillated sample than for the reference sample of  $\alpha$ SN fibrils measured in 100% D2O. The co-fibrillated sample has a linear region in the Guinier plot (Fig. 7C) allowing for Guinier analysis and determination of  $R_g \approx 100$  Å. The reference sample did not have a flat region (Fig. 7D), and the  $R_g$  could not be determined. The  $p(r)$  of the co-fibrillated sample and the reference sample (Fig. 7B) are similar for large values of  $r$ . The  $p(r)$  functions were scaled to align in this area. The shape of  $p(r)$  functions at  $r < \sim 250$  Å are however very different, and this part of the  $p(r)$  for the co-fibrillated sample may reveal some characteristic structural properties of the repeated unit of which  $\alpha$ SN fibrils is build up. The co-fibrillated signal is presumably a mixed signal from the large and almost fully matched out fibrils, and from the small deuterated subunits. The peak at 100 Å in the  $p(r)$  is characteristic for the cross-sectional dimensions of  $\alpha$ SN fibrils, and has previously been observed in SAXS (Giehm et al, 2011).

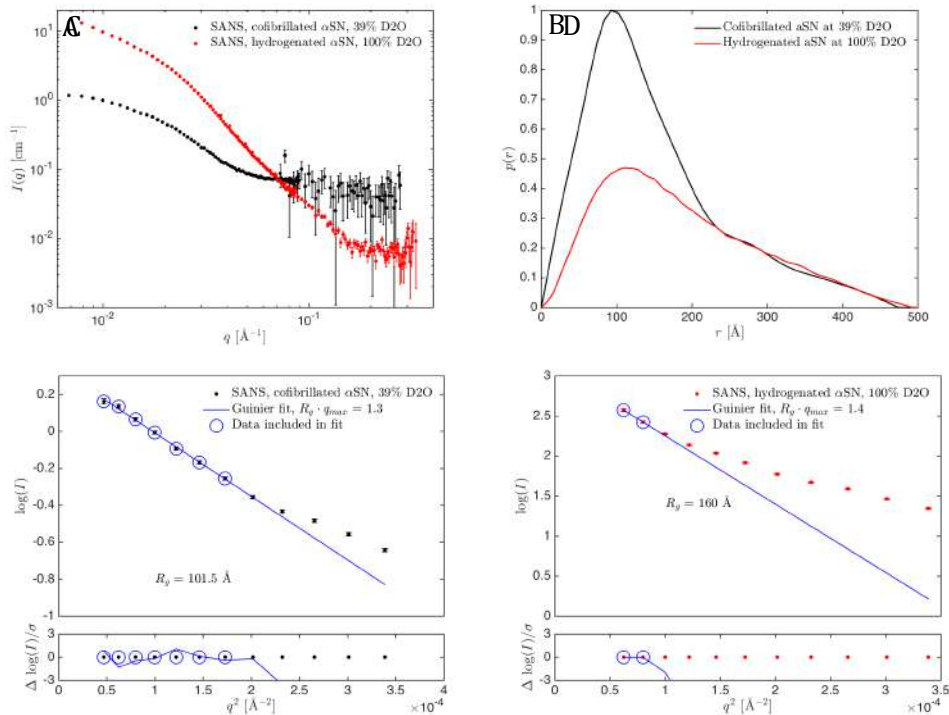


Figure 7. (A) SANS data on co-fibrillated  $\alpha$ SN in 39% D2O (black) and hydrogenated  $\alpha$ SN in 100% D2O (red). (B) Corresponding  $p(r)$  for the two



samples, which were scaled to align in the region  $r = 250$  to  $500 \text{ \AA}$ . (C-D) Guinier fits (blue) for the two samples (same colors).

## Conclusions

SANS exchange study shows that there is no exchange of monomers between the  $\alpha$ SN fibrils over a time span of 48 hours.

The data furthermore suggest the presence of a conventional water layer with a thickness of  $2\text{-}3 \text{ \AA}$  and a density about 10% higher than bulk water.

## References

- Giehm, L., Svergun, D. I., Otzen, D. E., & Vestergaard, B. (2011). PNAS, 108, 3246–3251.
- Hansen, S. (2014). J. Appl. Cryst., 47, 1469-1471.
- Nielsen, S. B., Macchi, F., Raccosta, S., Langkilde, A. E., Giehm, L., Kyrsting, A., Svane, A. S., Manno M., Christiansen, G., Nielsen, N.C., Oddershede, L., Vestergaard, B. & Otzen, D. E. (2013). PLoS One 8, e67713.
- Pedersen, M. C., Arleth, L. & Mortensen, K. (2013). J. Appl. Cryst. 46, 1894--1898.
- Svergun, D. I., Richard, S., Koch M. H., Sayers, Z., Kuprin S, Zaccai, G. (1998). PNAS, 95, 2267-2272.



## Chapter 9

# Publications

### **Paper I: Single-particle structure refinement from small-angle scattering data of partially aggregated protein samples**

Andreas Haahr Larsen, Jan Skov Pedersen and Lise Arleth

*planned submission to J. Appl. Cryst, not submitted yet.*

### **Paper II: Analysis of small-angle scattering data using model fitting and Bayesian regularization**

Andreas Haahr Larsen, Lise Arleth and Steen Hansen

*J. Appl. Cryst, 2018, 51, 1151-1161.*

### **Paper III: Invisible detergents for structure determination of membrane proteins by small-angle neutron scattering**

Søren Roi Midtgaard, Tamim A. Darwish, Martin Cramer Pedersen, Pie Huda, Andreas Haahr Larsen, Grethe Vestergaard Jensen, Søren Andreas Røssell Kynde, Nicholas Skar-Gislinge, Agnieszka Janina Zy-gadlo Nielsen, Claus Olesen, Mickael Blaise, Jerzy Jozef Dorosz, Thor Seneca Thorsen, Raminta Vensku-tonyte, Christian Krintel, Jesper V. Møller, Henrich Frielinghaus, Elliot Paul Gilbert, Anne Martel, Jette Sandholm Kastrup, Poul Erik Jensen, Poul Nissen and Lise Arleth

*FEBS, 2018, 285, 357-371.*

### **Paper IV: Small-angle neutron scattering studies on the AMPA receptor GluA2 in the resting, AMPA and GYKI-53655 bound states**

Andreas Haahr Larsen, Jerzy Dorosz, Thor Seneca Thorsen, Nicolai Tidemand Johansen, Tamim Darwish, Søren Roi Midtgaard, Lise Arleth and Jette Sandholm Kastrup

*iUCrJ, 2018, accepted.*

### **Paper V: Structure, dynamics and function of a lipid pool at the centre of the bacterial holo-translocon**

Remy Martin\*, Andreas Haahr Larsen\*, Robin A. Corey, Søren Roi Midtgaard, Nathan Zaccai, Marie-Sousai Appavou, Christiane Schaffitzel, Lise Arleth and Ian Collinson

*\*These authors contributed equally to this work.*

*not submitted yet.*

## 9.1 Paper I: Single-particle structure refinement from small-angle scattering data of partially aggregated protein samples

**List of Authors** Andreas Haahr Larsen, Jan Skov Pedersen and Lise Arleth

**Status** Draft.

**Abstract** Aggregation is an important process in much material science. Protein aggregation in particular is relevant for biological processes *in vivo* and for drug design. Aggregation of nanoparticles and proteins can be studied with small-angle scattering (SAS) and analytical descriptions of the aggregates are necessary tools for interpreting the data. When studying single proteins or isolated protein complexes with SAS, aggregation however constitute a problem as even a minor fraction of aggregates contributes significantly to the scattering. These contributions must be removed or taken into account to draw correct conclusions from data about the single protein structure. In the present paper, we review a list of structure factors describing different types of aggregates. We also derive one new for a spherical cluster with local hard sphere interaction,  $S_7(q)$ . The structure factors are compared and their characteristics discussed. They are moreover tested on simulated data of aggregated protein to demonstrate their usefulness in the description of aggregates. Moreover, two of the structure factors,  $S_4(q)$  (linear aggregate structure factor) and  $S_6(q)$  (spherical cluster structure factor) were used to take aggregation into account in simulated SAS data of monomeric lysozyme with a minor fraction of respectively linear and globular aggregates. We show how the correct monomeric protein structure, out of two alternative forms, could be determined if the structure factors were included in the analysis.

**Contributions by AHL** AHL and LA conceived the project. AHL did the simulations and implemented and tested the models. JSP derived the  $S_7(q)$  structure factor. AHL wrote the paper with contributions from JSP and LA.

## Single-particle structure refinement from small-angle scattering data of partially aggregated protein samples.

ANDREAS HAAHR LARSEN,<sup>a</sup> JAN SKOV PEDERSEN<sup>b</sup> AND LISE ARLETH <sup>a\*</sup>

<sup>a</sup>*Niels Bohr Institute, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark, and* <sup>b</sup>*Department of Chemistry, Aarhus University, Langelandsgade 140, 8000 Aarhus, Denmark. E-mail: arleth@nbi.ku.dk*

### Abstract

Aggregation is an important process in much material science. Protein aggregation in particular is relevant for biological processes *in vivo* and for drug design. Aggregation of nanoparticles and proteins can be studied with small-angle scattering (SAS) and analytical descriptions of the aggregates are necessary tools for interpreting the data. When studying single proteins or isolated protein complexes with SAS, aggregation however constitute a problem as even a minor fraction of aggregates contributes significantly to the scattering. These contributions must be removed or taken into account to draw correct conclusions from data about the single protein structure. In the present paper, we review a list of structure factors for the description of aggregates. We also derive one new structure factor for a spherical cluster with local hard sphere interaction,  $S_7(q)$ . The structure factors are compared and their characteristics discussed. They are moreover tested on simulated data of aggregated protein to demonstrate their usefulness in the description of aggregates. Two of the structure

factors, a linear aggregate structure factor,  $S_4(q)$ , and a spherical cluster structure factor,  $S_6(q)$ , were used to take aggregation into account in simulated SAS data of monomeric lysozyme with a minor fraction of respectively linear and globular aggregates. We show how the correct monomeric protein structure, out of two alternative forms with subtle structural differences, could be determined if the structure factors were included in the analysis.

### 1. Introduction

#### **INSERT SOMETHING ABOUT WHY AGGREGATION IS IMPORTANT**

#### **IN OTHER FIELDS THAN STRUCTURAL BIOLOGY - LISE/JAN.**

Protein aggregation is of biological relevance and constitute the main problem of interest in many studies (Krueger *et al.*, 2006), in part due to its role in neurodegenerative diseases (Stefani & Dobson, 2003). Small-angle X-ray and neutron scattering (SAXS and SANS) allows the study of the structure and formation of aggregates in solution. Particle sizes spanning from a few nm to hundreds of nanometers can be probed, and SAXS and SANS is therefore well suited for studying many aggregating systems. In order to study these aggregating structures, one needs a mathematical model for the scattering from the aggregates, in order to relate the measured intensity to a real-space coordinates. In the present paper we provide a list of structure factors that can be used in such aggregation studies.

In structural studies of single proteins, however, aggregation is problematic (Bondos & Bicknell, 2002) and should be avoided before SAS experiments, as most theory and software assumes identical and monomeric proteins. The problem is amplified by the fact that the intensity from a particle scales with the square of the particle volume, meaning that even a minor fraction of aggregates in a sample contributes significantly to the scattered intensity. This aggregate contribution can be detected as an upturn

at low scattering angles (Fig. 1), and as a non-linearity on the corresponding Guinier plot. Moreover, aggregates are revealed in the pair distance distribution function,  $p(r)$ , giving large values for the largest intraparticle distance,  $D_{max}$ , and for the radius of gyration,  $R_g$ . If the sample contains only protein, aggregation can also be detected by determining the molecular weight,  $M_W$ , from the data.  $M_W$  can be retrieved either from the forward scattering,  $I(0)$ , and the concentration, or via the Porod volume (Porod, 1951; Korasick & Tanner, 2018). By comparing the experimentally determined  $M_W$  with the theoretical value for the construct, the average oligomerization state can be assessed. For mixtures of monomeric proteins and oligomers, the average  $M_W$  will be systematically larger than the  $M_W$  of the monomer. These aggregates need to be taken into account in the analysis, and here the listed structure factors are useful. By including these in the analysis, correct conclusions can be drawn about the single protein structure.

### 1.1. Types of aggregates

In this paper we use the term protein aggregate to describe any composite particle build up of protein subunits. Many different aggregates and aggregation processes exist (Frieden, 2007). An important differentiation is between aggregates with and without conserved subunit structure. Aggregates in which the protein subunit preserves its folded structure are often denoted oligomers (note that the subunits may themselves be oligomeric protein complexes). Importantly, for structural analysis with SAS, the oligomeric aggregates contain information about the monomer structure and the scattering pattern of the monomer and the oligomer coincide at large values of the scattering vector  $q$  (Fig. 1). In the other class of aggregates, the protein subunits unfold partly or fully before aggregation and then rearrange in protein aggregates with a local structure different from that of the isolated monomer. This new structure

may be unspecific or it may form amyloid aggregates with a characteristic cross- $\beta$  secondary structure (Nelson *et al.*, 2005).

There is no clear consensus about these terms. In the field of amyloid protein formation, "oligomers" is e.g. often used to denote an intermediate structural state between the monomeric and the amyloid state, despite that the inner structure of these intermediate oligomers is altered with respect to the monomer (Carrotta *et al.*, 2001). In the present paper, however, an oligomeric aggregate will refer only to a protein aggregate with preserved protein subunit structure.

### *1.2. Reducing protein aggregates before the SAS experiment*

In most cases, aggregation can be avoided by refinement of the protein formulation, until a sample of monodisperse, identical particles is obtained. See e.g. Bondos & Bicknell (2002) who describe an effective way to find the best buffer condition, or Skou *et al.* (2014) describing ways to get rid of aggregates at the beamline. The latter include centrifugation of the sample and spin filtration to get rid of precipitation and large aggregates. Also, it is good practice to do size exclusion chromatography (SEC) to purify the sample prior to the experiment. A monodisperse sample results in a single, symmetric elution peak. Besides SEC, aggregates in the sample can be revealed before the SAS experiment by static or dynamic light scattering (SLS/DLS) and analytic ultracentrifugation (AUC) (see also Svergun *et al.*, 2013).

One of the more recent options for ensuring a monodisperse sample is the combination of SEC and synchrotron small-angle X-ray scattering (SEC-SAXS) as described e.g. in a review by Pérez & Nishino (2012). SEC-SAXS effectively circumvents scattering from time-dependent aggregation. Recently, this option also became available for small-angle neutron scattering (SEC-SANS; Jordan, 2016), and for home-source SAXS instruments (Bucciarelli *et al.*, 2018). However, if the amount of protein is limited, and

several conditions such as temperature, pH, and added ligands, need be tested, then it may not be feasible to to SEC-SAS. The aggregation process may also be too fast to be fully avoided by SEC-SANS, where the accumulation time is in the order of minutes.

Aggregation and other interparticle effects are often concentration effects, and can in those cases be avoided by dilution of the sample. High concentrations is however needed in some cases to obtain sufficient statistics, especially at the high- $q$  region, where the signal decreases with the magnitude of the form factor  $P(q)$ , which follow the  $q^{-4}$  Porod power law. High concentrations may also be needed to simulate physiological situations/problems. Therefore, as described e.g. by Svergun *et al.* (2013), a high-concentration and a low-concentration sample can be measured. If the concentration effects are only present in the high-concentration sample, the two dataset can be merged, thereby obtaining a combined dataset without concentration effect, but with good statistics in the full  $q$ -range. However, if the aggregation is not concentration dependent, but is caused e.g. by certain ligands, or is an inherent property of the sample, that strategy does not solve the problem.

### 1.3. Minimizing aggregate scattering contributions after the SAS experiment

The most simple, and probably the most common way to minimize the effect of aggregates in analysis of SAS data is to truncate the data at low- $q$ . As the form factor for large particles decreases rapidly and therefore mainly affects the first few data points, this can be rather effective. The same principle can be exploited to get rid of the contribution from concentration-dependent inter-molecular effects, as demonstrated by Müller & Glatter (1982) for SAXS data on latex spheres and by Pedersen *et al.* (1994) for SANS data on samples of insulin and insulin fibrils. The method however raises two issues: firstly, it must be decided where to truncate the data, and secondly,

the point of truncation may need be at so high  $q$ -values that valuable structural information about the particle is lost. Truncation is therefore a good solution only if the aggregates are much larger than the individual particles, since the scattering signal from the aggregates is then negligible already in the Guinier region of the individual particles. Thus, there is still a monomer Guinier region left after truncation, and the scattering signal from monomers and aggregates can be separated.

Alternatively, the scattering from the aggregates can be included in the model as a power law. This has been implemented in the generalized indirect Fourier transform (Bergmann *et al.*, 2000) and by a later and slightly altered approach (Oliveira *et al.*, 2009). The disadvantage is the need for assumptions about the aggregates that may be inaccurate and, in case the the study aims a solving the structure of the monomeric protein, of no particular interest. The advantage of including aggregates in the model is that information can be obtained from the full  $q$ -range, and problems regarding where to truncation data is avoided. It is not evident at all, where the aggregation signal stops, and in principle, there is a contribution in the whole  $q$ -region despite negligible at large values of the  $q$ -vector (Fig. 1). It can be argued that an approximate, but reasonable assumption about the aggregates is better than assuming their non-existence. As we shall see, I also have positive effects on the conclusions that can be drawn from data.

The aim of the present paper is to present and discuss analytical tools to handle the aggregation contribution after the experiment. The goal is to provide tools to find the best possible model for the single-particle structure despite the presence of a minor fraction of protein aggregates in the measured sample. Moreover, we will show how the analytical tools can ensure reasonable measures for the goodness of fit, thus allowing a better evaluation of the models and comparison of different hypothesized models. The presented methods will be demonstrated on simulated data of protein sample with a



minor fraction of aggregates, as generated from atomic protein structures from the protein data bank (PDB).

## 2. Theory and review of the use of structure factors for the description of aggregates

The differential scattering cross section per unit volume, colloquially denoted the intensity, from a sample of diluted, monodisperse, identical, and homogeneous particles is given as:

$$\frac{d\Sigma}{d\Omega}(q) = I(q) = cV^2\Delta\rho^2P(q), \quad (1)$$

where  $c$  is the number of particles per volume,  $V$  is the particle volume, and  $\Delta\rho$  is the excess scattering length density (contrast) of the particle with respect to the buffer.  $q$  is the magnitude of the scattering vector, given as  $4\pi\sin(\theta)/\lambda$ , where  $\lambda$  is the wavelength of the incoming beam and  $2\theta$  is the angle between the incoming and scattered beam.  $P(q)$  is the form factor, describing the shape of the identical particles. Together  $A = cV^2\Delta\rho^2$  constitutes a  $q$ -independent prefactor, which may as well be expressed as  $A = \phi V\Delta\rho^2$ , where  $\phi$  is the volume fraction. Using  $\phi = n/V$  is most sensible for self-assembling systems where  $\phi$  is conserved, and  $c$  vary. For protein systems, on the other hand, the molar concentration  $c_M = c/N_A$ , where  $N_A$  is Avogadro's number, is usually known, whereas  $\phi$  is unknown. We will in the following use  $A$  for the general prefactor.

The scattering from a sample of identical aggregated particles can be described with equation (1), but with a new form factor  $P(q) \rightarrow P_{agg}(q)$ , a new volume  $V \rightarrow V_{agg}$  and a new concentration  $c \rightarrow c_{agg}$ . Note that the contrast is unchanged. Assuming that the aggregates all consists of  $N$  particles, we have  $V_{agg} = NV$  and  $c_{agg} = c/N$ , where  $V$  and  $c$  are the volume and number density of the non-aggregated protein.  $\phi$

is conserved upon aggregation. That is:

$$I_{agg}(q) = N A P_{agg}(q). \quad (2)$$

For oligomeric aggregates of monodisperse, spherically symmetric particles, the intensity can be described as a product of the monomer form factor  $P(q)$ , and a structure factor  $S(q)$ :

$$I_{agg}(q) = A P(q) S(q), \quad (3)$$

such that  $S(q) = N P_{agg}/P(q)$ .  $S(q)$  may be considered a "super-form factor" more than a structure factor, as there is no long-range order, only local interactions between the subunits of the aggregated protein. We demand the usual normalization  $S(q \rightarrow 0) = N$  and  $S(q \rightarrow \infty) = 1$  to be able to fit on absolute scale.

For non-spherical particles, the expression is more complex, as the orientation of each particle has to be taken into account:

$$I(q) = A \left[ \sum_i \psi_i^2(q, \mathbf{e}_i) + \frac{1}{N} \sum_{j,i} \psi_i(q, \mathbf{e}_i) \psi_j(q, \mathbf{e}_j) [S_{i,j}(q, \mathbf{e}_i, \mathbf{e}_j) - 1] \right], \quad (4)$$

where  $\psi(q, \mathbf{e}_i)$  is the form factor amplitude ( $P(q) = |\psi(q)|^2$ ) for the  $i$ 'th particle with orientation given by the unit vector  $\mathbf{e}_i$ , and  $S_{i,j}(q, \mathbf{e}_i, \mathbf{e}_j)$  is the partial structure factor between the  $i$ 'th and the  $j$ 'th particle. The expression was first approximated in order to find the structure factor for polydisperse spheres, as polydispersity like asymmetry has the effect that the distance between the center of neighboring spheres differs from  $2R$ . The structure factor for monodisperse, hard spheres was derived by Percus & Yevick (1958) and the expression was later generalized to polydisperse spheres by Vrij (1979). In 1983, Kotlarchyk & Chen proposed the decoupling approximation, assuming that size and position of the spherical particles were uncorrelated, and the effective  $\tilde{S}(q)$  is given by:

$$\tilde{S}(q) = 1 + \beta [S(q) - 1], \quad (5)$$

where  $S(q)$  is the structure factor for a particles with spherical symmetry, and  $\beta(q) = \langle \psi(q) \rangle^2 / P(q)$ ,  $\psi(q)$  is the form factor amplitude and  $\langle \dots \rangle$  denote orientational averaging.  $\beta(q)$  can be estimated by direct fitting a simple spherical model to the low- $q$  part of the dataset, as described by Oliveira *et al.* (2009). Alternatively, as shown by Hoiberg-Nielsen *et al.* (2009)  $\langle \psi(q) \rangle = A_0^0(q)$ , with  $A_0^0(q)$  being the zeroth order spherical harmonic expansion of  $\psi(q)$  (Svergun *et al.*, 1995). For  $N$  atoms  $A_0^0(q)$  takes the form:

$$A_0^0(q) = \frac{\sum_{i=1}^N \Delta b_i \frac{\sin(qr_i)}{qr_i}}{\sum_{i=1}^N \Delta b_i}, \quad (6)$$

where  $r_i$  is the distance from the center of the particle to atom  $i$ , and  $\Delta b_i$  is the excess scattering length of atom  $i$ . For spherical particles,  $\tilde{S}(q) = S(q)$ , as the zeroth order spherical harmonics is a sphere. For non-spherical particles,  $\beta(q)$  decreases as  $q$  increases, since the structural difference from a sphere gets more and more pronounced at larger values of  $q$ .  $A_0^0(q)$  is given as output when running CRY SOL or CRYSON (Svergun *et al.*, 1995) (in the \*.alm binary files).  $A_0^0(q)$  and  $\beta(q)$  are likewise given as output in our home-written software CaPP (source code available at <https://github.com/Niels-Bohr-Institute-XNS-StructBiophys/CaPP>).

The intensity for aggregates of asymmetric particles is then given as:

$$I_{agg}(q) = cV^2 \Delta \rho^2 P(a) \tilde{S}(q). \quad (7)$$

The decoupling approximation was used to account for non-sphericity of the particles. This is a good approximation for particles that are near-globular. For very elongated particles, the approximation is however less accurate. The approximation is however simple and does not add any additional parameters to the model, and many proteins are near-globular, so the approximation will be valid and useful in most cases.

We will in the following consider the situation, where a fraction  $a$  of the particles are

10

in an aggregated form. The fraction  $a$  is given as  $n/m$ , where  $m$  is the total number of particles in the sample, and  $n$  is the number of particles that are part of an aggregate. The total intensity is then is given as:

$$I_{tot}(q) = A[(1 - a)P(q) + aNP_{agg}(q)], \quad (8)$$

or, using equation (7):

$$I_{tot}(q) = AP(q)[(1 - a) + a\tilde{S}(q)]. \quad (9)$$

Polydispersity can be included in the description of the aggregates by assuming a normal distribution in  $N$  with mean  $N$  and standard deviation  $\sigma_N$ :

$$S_{poly}(q) = \frac{1}{\sqrt{2\pi\sigma_N^2}} \int_{-\infty}^{+\infty} S(q, N') \exp\left(-\frac{(N' - N)^2}{2\sigma_N^2}\right) dN', \quad (10)$$

where the integration limits in practice can be replaced by  $\pm 3\sigma_N$ .

### 2.1. Descriptions of $P_{agg}(q)$

In some cases, it is possible to give an approximate description for the aggregate form factor,  $P_{agg}(q)$ . Firstly, the aggregates may be described by a geometrical shape, as in the study by Chatani *et al.* (2015), where the protein was known to form elliptical cylinder-shaped aggregates. In that case equation (2) can be used Secondly, if the protein of interest has been crystallized, then an oligomer from the protein crystal (collection of symmetry mates) may constitute a reasonable model for the aggregates. Thirdly, if the number of particles per aggregate is known (dimer, trimer, ..., N-mer), but their internal positioning is unknown, then rigid body modelling can be used to form one or more representative aggregates that fits the data. Tools such as EOM (Tria *et al.*, 2015) are available for that approach.

**Addition of exponential.** In many cases, however, the exact form factor of the aggregate is unknown, and a more generalized formulation is needed. A simple approach

is to use a form factor that approximate the scattering from the aggregates by a power law:

$$P_{agg}(q) = Bq^{-D}, \quad (11)$$

where  $B$  is a scaling factor and  $D$  is the decay exponent, given as the slope on a  $\log(I)$  vs  $\log(q)$  plot. For very large aggregates, the exponent is simply the Porod decay,  $D = 4$ , as used e.g. in Pedersen (1993). For smaller aggregates, the exponent gives the fractal dimensionality of the aggregate (Lin *et al.*, 1989; Beaucage, 1995), where  $D = 1$  corresponds to an elongated rod-like structure, and  $D = 3$  to a globular aggregate. The intensity takes the form  $I_{tot}(q) = A[(1 - a)P(q) + aNq^{-b}]$ . Clearly,  $P_{agg}(q) \rightarrow \infty$  as  $q \rightarrow 0$  and is unphysical, as it describes an infinitely large aggregate. This also means that data cannot be fitted on absolute scale. Inserting (11) into (8) gives an expression with three scaling parameters,  $A$ ,  $a$  and  $N$ . As data cannot be fitted on absolute scale, these can be reduced to two independent scaling parameters,  $A'$  and  $B'$ , that should be determined by fitting to data:

$$I_{tot}(q) = A'P(q)[1 + B'q^{-D}], \quad (12)$$

where  $A' = A(1 - a)$  and  $B' = NBa/(1 - a)$ .

## 2.2. Descriptions of $S(q)$

If the subunits of the aggregate are assumed to be identical to the isolated single protein structure, then equation (3) can be used. That is, when an assumption of oligomeric aggregates is applied. A list of alternative aggregate structure factors is given in the following. But first, we will give some general notes.

**Fractal aggregates.** Fractal theory is a successful tool to describe aggregates of nanoparticle (see e.g. Sørensen, 1997, 1999 and 2001).  $S_1(q)$ ,  $S_2(q)$  and  $S_3(q)$  are based on fractal theory, and are listed with increasing complexity.

**Characteristic radius of the particles.** Most of the structure factors include a parameter  $R$  for the radius of the particles in the aggregate, as the structure factors are derived for spherical protein subunits. This radius may be fixed to reduce the number of parameters.  $R$  can be approximated with the radius of a sphere having the same volume as the particle of interest, where the volume of the particle can be calculated as the sum of Van der Waals volumes for the atoms (Svergun *et al.*, 1995).

**Fisher-Burford structure factor.** The FB structure factor (Fisher & Burford, 1967; Sørensen, 2001) has only two parameters, the radius of gyration of the aggregates,  $R_g$ , and the aggregate fractal dimensionality,  $D$  (Sørensen & Wang, 1999).  $D$  should be between one (linear aggregate) and three (spherical aggregates), i.e.  $1 < D \leq 3$ .  $S(q)$  takes the form:

$$S_1(q) = 1 + \left( \frac{2R_g^2}{3D} \right)^{-D/2} q^{-D}, \quad (13)$$

where  $R_g$  can be related to  $N$  by the scaling relationship  $N = k(R_g/R)^D$ , and  $k$  is a prefactor of the order of unity, which depends on  $D$  (Sorensen & Roberts, 1997). Fitting empirically measured values for  $k$  and  $D$  gives the empirical relation  $k = -0.5836 \cdot D + 2.1739$  (Fig. 2)  $R$  can be fixed as described above.  $S_1(q)$  can be simplified by introducing the notation  $\gamma = qR_g$ :

$$S_1(q) = 1 + \left( \frac{2}{3D} \right)^{-D/2} \gamma^{-D}, \quad (14)$$

In the case  $D = 2$ , the FB structure factor takes the particular simple form:

$$S_{1,D=2}(q) = 1 + 3\gamma^{-2}. \quad (15)$$

$S_1(q)$  converges to unity as  $q$  increases. Likewise,  $S_1(q)$  is unity if  $R_g$  is very large, i.e. for very large and unmeasurable aggregates (Fig. 3).  $S_1(q)$  is thus an improvement of the simple exponential term, as the unity term as well as the multiplication with the

form factor, (equation (7)) ensures that the intensity from the aggregate is dominated by  $P(q)$  at large values of  $q$ . For  $q \rightarrow 0$   $S_1(q)$  diverges. The prefactor  $(2R_g^2/3D)$  therefore has no practical effect, as  $S_1$  diverges for  $q \rightarrow 0$  and therefore has to be fitted on arbitrary scale. So  $S_1(q)$  might as well, for all practical purposes, be simplified as

$$S_1(q) = 1 + Aq^{-D} \quad (16)$$

**Mass fractal structure factor.** For identical spheres with radius  $R$ , the structure factor can be written as (Teixeira, 1988):

$$S_F(q) = 1 + \frac{1}{(qR)^D} \frac{D\Gamma(D-1)}{[1 + (qC)^{-2}]^{(D-1)/2}} \sin \left[ (D-1) \tan^{-1}(qC) \right], \quad (17)$$

where  $\Gamma(x)$  is the gamma function of  $x$ ,  $C$  is the correlation length, which is directly related to  $R_g$  by  $R_g^2 = D(D+1)C^2/2$  (Teixeira, 1988). The high- $q$  limit is dominated by the unity term, so  $S_F(q \rightarrow \infty) = 1$ . In the low- $q$  limit, the second term is dominating, and  $S(q \rightarrow 0)$  is proportional to  $N$  via. the scaling relationship, and assuming  $N \gg 1$ . The proportionality constant depends on  $D$ . Thus, it is more physical than  $S_1(q)$  as it converges to a finite value for  $q \rightarrow 0$ , as would any real (not infinitely large) particle. The low- $q$  limit has been calculated (Teixeira, 1988), and to first order it is  $S_F(q \rightarrow 0) = \Gamma(D+1)(C/R)^D$ . Thus, in order to obtain the correct normalization, we divide by the limit value and multiply with  $(N-1)$ :

$$S_2(q) = 1 + \frac{ND\Gamma(D-1)}{\Gamma(D+1)(qCR^2)^D \cdot [1 + (qC)^{-2}]^{(D-1)/2}} \sin \left[ (D-1) \tan^{-1}(qC) \right]. \quad (18)$$

$S_2(q)$  is shown in Fig. 3 for  $D \rightarrow 1$ ,  $D = 2$  and  $D = 3$ .  $S_2(q)$  was used by Midtgaard *et al.* (2018) to fit SANS data measured on a partly aggregated sample of the membrane protein sarco/endoplasmic reticulum calcium ATPase (SERCA).  $S_2(q)$  was also used by Larsen *et al.* (2018) to describe SANS data measured on samples of the AMPA-type glutamate receptor 2 (GluA2) with a fraction of oligomeric aggregates. In this study, the aggregate dimensionality,  $D$ , was fixed to two, thereby reducing the number of

parameters to two,  $C$  and  $R$ . When  $D = 2$ ,  $R_g$  and  $C$  are related by  $R_g^2 = 3C^2$ . Using  $\sin(\tan^{-1} x) = x/\sqrt{(x^2 + 1)}$ , equation (18) can thus be simplified to:

$$S_{2,D=2}(q) = 1 + N(1 + 2\gamma^2/3 + \gamma^4/9)^{-1/2}, \quad (19)$$

with  $\gamma \equiv qR_g = qR\sqrt{N}$ . In order to introduce  $N$  explicitly, the scaling relationship was used,  $N = k(R_g/R)^D$ , where  $k \approx 1$  for  $D = 2$  (Fig. 2; Sorensen & Roberts, 1997).

**Fractal structure factor with nearest neighbor perturbation.** The fractal description of an aggregate is valid only at long distances. Locally, oligomeric aggregates has the same structure as the isolated monomeric protein. These local nearest-neighbor interactions are not accounted for in the mass fractal structure factor. Dimon *et al.* (1986) proposed another fractal structure factor taking the nearest-neighbor interaction into account. The structure factor successfully described a sample of aggregating gold particles measured with SAXS:

$$S_D(q) = 1 + z_1 \frac{\sin(2qR)}{2qR} + 4\pi \left( z_2 \int_{2R}^{4R} r^2 g_2(r) \text{sinc}(qr) dr + z_A \int_{4R}^{\infty} r^{D-1} e^{-r/C} \text{sinc}(qr) dr \right), \quad (20)$$

where  $z_1$  and  $z_2$  are the coordination numbers of the first and second shell of nearest neighbors, and  $z_A(r)$  is the amplitude of the fractal term.  $R$  is the radius of the gold particles,  $D$  is the fractal dimension and  $C$  is the fractal correlation length.  $g_2(r)$  is a second-order polynomial fit to simulated data describing the second shell of nearest neighbors. Dimon *et al.* note that the second shell term with  $g_2(r)$  provides only a minor improvement of the fit to data. By including only the first shell,  $S_D(q)$  simplifies to:

$$S_3(q) = 1 + z_1 \text{sinc}(2qR) + z_A \int_{2R}^{\infty} r^{D-1} e^{-r/C} \text{sinc}(qr) dr. \quad (21)$$

where  $4\pi$  was included in  $z_A$ .  $z_1$  reflects the number of particles in the first shell, which increases exponentially with the dimensionality  $D$ , and  $z_A$  ensures that  $S_3(q \rightarrow 0) =$



$N$ :

$$z_1(D) = 2^{D-1}, \quad z_A(D, N, C, R) = \frac{(N - z_1 - 1)}{\int_{2R}^{\infty} r^{D-1} e^{-r/C} dr}. \quad (22)$$

$S_3(q)$  is more physical than  $S_2(q)$  as it includes local interactions between neighboring particles. This is reflected in the calculated SAS curve via a correlation holes (Fig. 3, the first corr. hole is between  $q = 0.1 \text{ \AA}^{-1}$  and  $q = 0.2 \text{ \AA}^{-1}$ ).

**Linear aggregate structure factor.** In the case of a linear aggregate, the structure factor can be derived directly from the Debye formula:

$$S_{L,N}(q) = 1 + 2/N \sum_{j=1}^{N-1} (N-j) \text{sinc}(2jqR) \quad (23)$$

The linear aggregate structure factor has only two parameters, which are the interparticular distance  $2R$  and the number of particles  $N$ , whereof  $R$  can be fixed as described above. The mean of  $N$  may not be an integer, so the expression has to be altered to handle non-integer values of  $N$ . By  $N_b$  we denote the largest integer values below  $N$ , and  $w = N - N_b$ . The linear aggregate structure factor takes the form:

$$S_4(q) = (1 - w) \cdot S_{L,N_b}(q) + w \cdot S_{L,(N_b+1)}(q). \quad (24)$$

$S_3(q)$  is comparable to the fractal structure factors with  $D = 1$  and they have all been plotted in Fig. 3A.  $S_4(q)$  has the usual limits  $S_3(q \rightarrow 0) = N$  and  $S_3(q \rightarrow \infty) = 1$ . Like  $S_3(q)$ , the linear structure factor,  $S_4(q)$ , has a correlation hole. Note the sharp oscillations after the first correlations hole (Fig. 3). These sharp features are of small relative magnitude, but cause a significant discrepancy when fitted to simulated data (Fig. 4A,  $q \sim 0.2 \text{ \AA}^{-1}$ ).

**Random flight structure factor.** The aggregates can also be described by a random flight (Burchard & Kajiwara, 1970; Giehm *et al.*, 2010):

$$S_{RF,N}(q) = \frac{2}{1 + \text{sinc}(q2R)} - \frac{2 - 2\text{sinc}(q2R)^N}{N(1 - \text{sinc}(q2R))^2} \text{sinc}(q2R) - 1, \quad (25)$$

where  $2R$  is the step size, corresponding to the center-to-center distance between neighboring particles in the aggregate and  $N$  is the number of particles in the aggregate. Again, the mean value of  $N$  may not be an integer, so:

$$S_5(q) = (1 - w) \cdot S_{RF, N_b}(q) + w \cdot S_{RF, (N_b+1)}(q), \quad (26)$$

with  $w$  and  $N_b$  defined as before.  $S_5(q)$  has the usual limits and is comparable to the fractal structure factors with  $D = 2$  (Fig. 3B). It resembles the mass fractal structure factor  $S_2(q)$  at  $D = 2$ , but has a correlation hole for local interactions.

**Spherical cluster.** The spherical cluster assumes that the proteins subunits form a spherical aggregate with density defined by a volume fraction  $\phi$ , which is given in terms of the radius of the subunits,  $R$ , the mean radius of the spherical cluster  $\mu_r$ , and the number of particles in the cluster,  $N$ :

$$\phi = N \left( \frac{R}{\mu_r} \right)^3. \quad (27)$$

The spherical cluster aggregates are assumed to have a degree of polydispersity, described by a normal distribution with mean  $\mu_r$  and width  $\sigma_r$ . The form factor is then described by the usual form factor for a sphere:

$$P_C(q, \mu_r, \sigma_r) = \frac{1}{\sqrt{2\pi\sigma_r^2}} \int_{-\infty}^{+\infty} \exp[-(\mu_r - r')^2 / (2\sigma_r^2)] \frac{3[\sin(qr') - qr' \cos(qr')]}{(qr')^3} dr', \quad (28)$$

where the integral limits can be approximated by  $\pm 3\sigma_r$ . The structure factor is then given as:

$$S_6(q) = 1 + (N - 1)P_C(q, \mu_r, \sigma_r), \quad (29)$$

and has the usual limits,  $S_6(q \rightarrow \infty) = 1$  and  $S_6(q \rightarrow 0) = N$ . A model of spheres with different sizes were used by Niimura *et al.* (1995) to model a sample of aggregating lysozyme. In that study the polydispersity was introduced by adding the scattering

from an ensemble of spheres with the number and radius of each sphere manually chosen to match data.

$S_6(q)$  is one example out of a group of structure factors with the same format, which describes the aggregate via. a geometrical shape. The the polydisperse spherical form factor  $P_C(q, \mu_r, \sigma_r)$  can be replaced by any other form factor, e.g. for a cylinder or an ellipsoid. The form factor then reflects the overall geometry the oligomeric aggregates.

**Spherical cluster structure factor with local hard sphere interactions.** This extended spherical cluster structure factor with local hard sphere interactions gives an analytical description of the simulated results in the work by Genix & Oberdisse (2017) on polydisperse nanoparticle assemblies. It describes a cluster of circular and monodisperse particles inside an embedding sphere. As for  $S_6(q)$ , the spherical cluster is characterized by a volume fraction,  $\phi$ , a mean radius  $\mu_r$  and a polydispersity  $\sigma_r$ . The correlation of the particles inside the embedding sphere is described by the hard sphere structure factor  $S_{HS}(q, r, \phi)$  (Kinning & Thomas, 1984). The derivation of the extended spherical cluster structure factor is given in the appendix, and the final result is:

$$S_7(q) = S_{HS}(q, R, \phi) + (N - S_{HS}(q, R, \phi))P_L(q, \mu_r, \sigma_r), \quad (30)$$

where  $P_L(q, \mu_r, \sigma)$  is the form factor for the polydisperse embedding spheres (see appendix).  $S_7(q)$  has three free parameters:  $N$ ,  $\mu_r$  and  $\sigma_r$ , assuming the radius of the particles,  $R$ , is fixed as previously described. It is plotted in Fig. 3C together with the fractal structure factors  $S_1(q)$ ,  $S_2(q)$  and  $S_3(q)$  with  $D = 3$ . Note that  $S_7(q)$  has the same form as  $S_6(q)$ , with unity replaced by  $S_{HS}(q, R, \phi)$ .

### 3. Methods

#### 3.1. Artificial oligomeric aggregates

Artificial aggregates were generated with a Monte Carlo (MC) approach, and saved in the protein data bank format. A monomeric protein structure was first placed. A new monomer was then placed and rotated at random. Subsequently, the new monomer was translated randomly until there was no overlap between the two monomers. New monomers were added iteratively to obtain a random aggregate. Two types of aggregates were generated. The first was generated by placing each new monomer at the origin and doing random translation until there was no overlap with any previously placed monomers. That resulted in a spherical aggregate (Fig. 5A). The second aggregate was generated by placing every new monomer at the position of the previous and the MC translation steps were constrained to positive values. This resulted in an elongated aggregate (Fig. 5B). These aggregates represent dimensional extremities and real physical aggregates can have any dimensionality between the two.

#### 3.2. Demonstrating structure factors on simulated SAXS data

The program CaPP (source code available at <https://github.com/Niels-Bohr-Institute-XNS-StructBiophys/CaPP>) was used to directly calculate the scattering from the monomeric structures and artificial aggregates. The theoretical scattering from a mixture of monomers and aggregates were generated from the calculated form factors as  $I_{theory}(q) = (1 - a)P_{monomer}(q) + aNP_{agg}(q)$ , where  $N$  is the number of proteins per aggregate. Noise was simulated as  $\sigma = n_r \cdot \sqrt{I_{theory}(q)} + n_a$  with relative noise  $n_r = 0.02$  and absolute noise  $n_a = 0.0001$ . Simulated data were randomly sampled from a normal distribution with standard deviation  $\sigma$  and mean  $I_{theory}(q)$ . WillItFit (Pedersen *et al.*, 2013) was used to fit the simulated data using the protein subunit form factor obtained from CaPP and the presented analytical structure factors for the

oligomeric aggregates.

### 3.3. Assessment of the goodness of fit

The fits to the simulated data were evaluated using the reduced  $\chi^2$ , defined as  $\chi_r^2 = \chi^2/f$ , where  $f$  is the number of degrees of freedom, conventionally given as  $f = N - K$ , where  $N$  is the number of data points and  $K$  is the number of model parameters.  $\chi^2$  is given as:

$$\chi^2 = \sum_{i=1}^N \frac{(I_{fit,i} - I_{sim,i})^2}{\sigma_{sim,i}^2}, \quad (31)$$

where  $\sigma_{sim}$  is the standard deviation of the simulated data.

## 4. Example: lysozyme sample with monomers and linear aggregates

To demonstrate the methodology, an elongated 25-mer oligomeric aggregate was generated in a Monte Carlo approach (similar to the 10-mer in Fig. 5). The theoretical scattering from a lysozyme monomer (PDB 1LYZ) and lysozyme 25-mer were calculated using CaPP. Data was simulated with a mixture of monomeric and aggregated lysozyme with added noise. The simulated data had a  $\chi_r^2$  of 1.06 (compared to the underlying true model),  $N = 25$  and  $a = 3.0$  %. The data were fitted first with a model of lysozyme monomers (Fig. 6). The model resembled the data at medium and high- $q$ , but there were clear systematic discrepancies at low- $q$ , as evident from the fit and the normalized residuals. The fit had a  $\chi_r^2$ -value of 6.63. As there were 200 data points and two parameters (scale and background), the model had 198 degrees of freedom. Therefore, the 95% confidence interval for the  $\chi_r^2$  value was [0.76, 1.28]. The data were also fitted with a model of lysozyme monomers and linear aggregated, using  $S_7(q)$  and the decoupling approximation. The model fitted well with data and gave a  $\chi_r^2$  of 1.09, very close to true value of 1.08. This model has two less degrees of freedom due to the two extra model parameters, but within two decimals, the 95%

confidence interval for  $\chi_r^2$  is the same. So for this model, the  $\chi_r^2$  is within the interval. The radius was fixed to 17.0 Å, which is the radius of a sphere with the same volume as the sum of Van der Waals volumes of all atoms in lysozyme. The refined parameters were  $N = 21 \pm 20$ , and  $a = 3.6 \pm 3.0\%$ . The parameters are in close proximity to the true values from the simulated data. The uncertainty is however large, as  $N$  and  $a$  are correlated (equation (9)).

The inclusion of the aggregate structure factor ensures that the correct conclusion is drawn from the analysis, namely that the monomeric part of the sample has the expected structure (PDB 1LYZ).

To explore the consequences of this, a modified atomic structure was generated (Fig. 7A), with the C-terminal of the lysozyme structure (PDB 1LYZ) being rotated and translated. This was generated to investigate a structural change that may be investigated with SAXS. The simulated data were first fitted with a model of monomers of the modified structure, resulting in a  $\chi_r^2$  of 7.17. That is, it fitted worse than the monomer of the unmodified structure ( $\chi_r^2 = 6.63$ ), but both structures were outside the expected confidence interval, so it would be difficult to conclude that one structure fitted better than the other from the monomer fits alone. Comparing the fits with an F-test gives the result that there is a 31% chance to get these  $\chi_r^2$  values for equally good models. E.g. one model was not significantly better than the other as judged from the statistical analysis. A model including linear aggregates were also fitted to data with a resulting  $\chi_r^2$  of 1.31, i.e. just outside the 95% confidence interval. So by including the structure factor to account for the aggregates, the correct conclusion about the sample can be extracted using the statistical analysis, namely that the unmodified structure best explained data. An F-test gave a 10% chance of obtaining these  $\chi_r^2$ -values for equally good models.

### 5. Example: lysozyme sample with monomers and spherical aggregates

The methodology was also tested on a simulated sample with spherical aggregates. The sample was a mix of lysozyme monomers (PDB 1LYZ) and 25-mer spherical aggregates of the monomers (like the 10-mer aggregates shown in Fig. 5). Fitting without inclusion of aggregates gave a  $\chi_r$  of 12.4 for the real structure, and a  $\chi_r$  of 11.0 for the modified structure. That is, the immediate conclusion is that the modified structure best describes the data. Data were however also fitted with a model including the spherical cluster structure factor,  $S_6(q)$ . Fig. 7B shows the simulated data and fits.  $\chi_r$  for the model of lysozyme and spherical cluster aggregates was 1.13, i.e. well inside the confidence interval.  $\chi_r$  for the modified lysozyme structure and aggregates was 1.29, just outside the confidence interval. Thus, the method leads to the correct conclusions about the sample, namely that the unmodified structure best describes data. The true  $\chi_r^2$  for the simulated data was 1.18.

### 6. Which structure factor to use?

Obviously, the best structure factor for description of a sample depends on the structure of the aggregate. As illustrated in Fig. 3, the fractal aggregates structure factors  $S_2(q)$  and  $S_3(q)$  provide approximative descriptions for aggregates with dimensionality between 1 and 3. For completely linear aggregates,  $D = 1$ , the linear structure factor  $S_4(q)$  is a good description, whereas the random flight structure factor  $S_5(q)$  may be a good description if  $D$  is close to two, and the spherical cluster structure factors with or without hard sphere interaction potential provide approximative descriptions for globular aggregates.  $S_3(q)$ ,  $S_4(q)$ ,  $S_5(q)$ ,  $S_7(q)$  and  $S_6(q)$  all has a correlation hole, which is also seen for the simulations (Fig. 1), where the calculated scattering for the aggregates has lower intensity than the scattering from the monomers for intermediate  $q$ . For the simulated globular 25-mer aggregates, the calculated scattering

had oscillations (Fig. 1A), which qualitatively resembles these correlation holes. The degree of aggregations is also important for what structure factor that is best suited. If there is only little aggregation, the information in data about the aggregates is very limited, and it is thus preferable to use a model with few parameters, whereas highly aggregated samples needs a more accurate description.

## 7. Conclusion

The present paper discuss particle aggregation how to describe them in the analysis of SAS data using analytical structure factors. A list of structure factors were presented and compared. Some were renormalized, and one structure factor is new,  $S_7(q)$ . The structure factors can be used in studies where the aggregated particles are the main interest, as demonstrated for simulated protein aggregates in Fig. 4. The structure factors may also be useful when a protein sample with a minor fraction of aggregates is studies in order to retrieve information about the single protein structure. In that context, the structure factors are used to "filter out" the scattering contribution from the aggregates, which ensures that the correct conclusions about the structure of the single protein is drawn from the data, as demonstrated and shown in Fig. 7. This methodology has the advantage over simple truncation of data, that no subjective choice about the point of truncation has to be made. Furthermore, the aggregates contribute to the scattering in the whole  $q$ -range. For aggregates much larger than the protein of interest, the contribution is negligible before the Guinier region of the protein of interest and truncation can be done safely, but when the difference in size is smaller, then the effect is no longer negligible and should be taken into account in the analysis.



### 8. Appendix: derivation of $S_{clust}$

The calculations are inspired by the simulations by Genix and Oberdisse (2017).

Using the Debye formula for a collection of monodisperse spheres within a larger sphere, the form factor of the complete object can be calculated. There are  $N$  smaller spheres within the large sphere.  $P_s(qR)$  is the scattering form factor of the small spheres. Then:

$$\begin{aligned} P(q) &= P_s(qR) \sum_{j,i=1}^{N,N} \text{sinc}(qR_{ij}) \\ &= NP_s(qR) + P_s(qR) \sum_{i \neq j}^{N,N} \text{sinc}(qR_{ij}). \end{aligned} \quad (32)$$

The last term is essentially the cross term between the small spheres and the large embedding sphere with form factors  $P_L(qr)$ . Considering that there are  $N(N-1)$  cross terms in the sum, one gets:

$$\begin{aligned} P(q) &\approx NP_s(qR) + N(N-1)P_s(qR)P_L(qr) \\ &= NP_s(qR)(1 + (N-1)P_L(qr)), \end{aligned} \quad (33)$$

which has the normalization  $P(q=0) = N^2$  for the usual normalization of the sphere form factor:

$$P(x) = \left( \frac{3(\sin(x) - x \cos(x))}{x^3} \right)^2. \quad (34)$$

The expression for the cluster form factor neglects correlations between the small spheres, which must be present for a high density of small spheres in a compact cluster. Such correlations will be present in both terms, however, since the last term  $(N-1)P_L(qr)$  decays strongly at high  $q$  due to the Porod behaviour of the sphere form factor, the main influence is on the first term. The effects of the correlations can be described by a structure factor  $S(q)$ :

$$P(q) = NP_s(qR)(S(q) + (N-S(q))P_L(qr)), \quad (35)$$

where an additional modification of the last term has been done in order to preserve the normalization. In order for the small spheres to form the cluster, one has to allow that they get into contact. Therefore a hard-sphere structure factor with an interaction radius equal to the actual radius can be used for describing the correlations. The hard-sphere volume fraction is given by the number of spheres and the ratio between the radius of the small spheres and the embedding sphere:

$$\phi = N(R/r)^3. \quad (36)$$

The hard-sphere structure factor is formally written as  $S_{HS}(q, R, \phi)$  and

$$P(q) = NP_s(qR)(S_{HS}(q, R, \phi) + [N - S_{HS}(q, R, \phi)]P_L(qr)), \quad (37)$$

so that structure factor  $S_{clust}(q)$  with the usual normalization  $S_{clust}(q \rightarrow \infty) = 1$ :

$$S_{clust,mono}(q) = S_{HS}(q, R, \phi) + (N - S_{HS}(q, R, \phi))P_L(qr), \quad (38)$$

which in addition has the normalization  $S_{clust}(q \rightarrow 0) = N$ .

Polydispersity in the radius of the embedding spheres is expected and can be included by replacing  $r$  by a mean radius  $\mu_r$  and a radius spread  $\sigma_r$  and integrate over the radius:

$$S_{clust}(q) = \frac{1}{\sqrt{2\pi\sigma_r^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(\mu_r - r')^2/\sigma_r^2} (S_{HS}(q, R, \phi) + [N - S_{HS}(q, R, \phi)]P_L(qr')) dr'. \quad (39)$$

Notice that there is an implicit  $r'$  dependency in the hard sphere potential through the  $\phi(N, r, R)$ . To simplify the expression,  $\phi$  may be approximated by

$$\phi \approx N(R/\mu_r)^3. \quad (40)$$

Such that the polydispersity is separated to  $P_L$  alone, i.e.:

$$P_L(q, \mu_r, \sigma_r) = \frac{1}{\sqrt{2\pi\sigma_r^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(\mu_r - r')^2/\sigma_r^2} P_L(qr') dr'. \quad (41)$$

which can be inserted in (38) to get an approximative expression for  $S_{clust}(q)$ :

$$S_{clust}(q) \approx S(q, R, \phi) + (N - S(q, R, \phi))P_L(q, \mu_r, \sigma) \quad (42)$$

**Acknowledgements** The authors would like to thank CoNeXT, University of Copenhagen and Aarhus University for co-funding the project.

## References

- Beaucage, G. (1995). *J. Appl. Cryst.* **28**, 717–728.
- Bergmann, A., Fritz, G. & Glatter, O. (2000). *J. Appl. Cryst.* **33**, 1212–1216.
- Bondos, S. E. & Bicknell A. (2002). *Anal. Biochem.* **316**, 223–231.
- Burchard, W. & Kajiwara, K. (1970). *Proc. Roy. Soc. London Ser. A* **316**, 185–199.
- Bucciarelli et al, (2018). ACCEPTED.
- Carrotta, R., Bauer, R., Waninge, R. & Rischel, C. (2001). *Protein Sci.* **10**, 1312–1318.
- Chatani, E., Inoue, R., Imamura, H., Sugiyama, M., Kato, M., Yamamoto, M., Nishida, K. & Kanaya, T. (2015). *Sci. Rep.* **5**: 15485.
- Dimon, P., Sinha, S. K., Weitz, D. A., Safinya, C. R., Smith, G. S., Varady, W. A. & Lindsay, H. M. (1986). *Phys. Rev. Lett.* **57**, 595–598.
- Fisher, M. E., & Burford, R. J. (1967). *Phys. Rev. A* **156**, 583–622.
- Frieden, C. (2007). *Protein Sci.* **16**, 2334–2344.
- Gazzillo, D., Giacometti, A., Valle, R. G. D., Venuti, E. & Carsughi, F. (1999). *J. Chem. Phys.* **111**, 7636–7645.
- Giehm, L., Oliveira, C. L. P., Christiansen, G., Pedersen, J. S. & Otzen, D. (2010). *J. Mol. Biol.* **401**, 115–133.
- Hansen, S. (2013). *J. Appl. Cryst.* **46**, 1008–1016.
- Hoiberg-Nielsen, R., Westh, P., Skov, L. K. & Arleth, L. (2009). *Biophys. J.* **97**, 1445–1453.
- Jordan, A., Jacques, M. Merrick, C., Devos, J., Forsyth, V. T., Porcar, L. & Martel, A. (2016). *J. Appl. Cryst.* **49**, 2015–2020.
- Kinning, D. J. & Thomas, E. L. (1984). *Macromolecules* **17**, 1712–1718.
- Korasick, D. A. & Tanner, J. J. (2018). *Protein Science* **27**, 814–824.
- Kotlarchyk, M. & Chen, S.-H. (1983). *J. Chem. Phys.* **79**, 2461–2469.
- Krueger, S., Ho, D. & Tsai, A. (2006). *Misbehaving Proteins: Protein (Mis)Folding, Aggregation, and Stability*. (Murphy, R. M. & Tsai, A. M. (eds.), Springer, New York), 125–146.

- Larsen, A. H., Dorosz, J., Thorsen, T. S., Johansen, N. T., Darwish, T., Midtgaard, S. R., Arleth, L. & Kastrup, J. S. (2018). *iUCrJ*, ACCEPTED.
- Lin, M. Y., Klein, R., Lindsay, H. M., Weitz, D. A., Ball, R. C., Meakin, P. (1989). *J. Colloid Interface Sci.* **137**, 263–280.
- Müller, K. & Glatter O. (1982). *Makromol. Chem.* **183**, 465–479.
- Nelson, R., Sawaya, M. R., Balbirnie, M., Madsen, A. Ø., Riekel, C., Grothe, R., & Eisenberg, D. (2005). *Nature*, **435**, 773–778.
- Niimura, N., Minezaki, Y., Ataka, M. & Katsura, T. (1995). *J. Cryst. Growth* **154**, 136–144.
- Oliveira C. L., Behrens, M. A., Pedersen, J. S., Erlacher K., Otzen D., Pedersen J. S. (2009). *J. Mol. Biol.*, **387**, 147–161.
- Pedersen, J. S. (1993). *Phys. Rev. B* **47**, 657–665.
- Pedersen, J. S. (1997). *Adv. Colloid Interface Sci.* **70**, 171–210.
- Pedersen, M. C., Arleth, L. & Mortensen, K. (2013). *J. Appl. Cryst.* **46**, 1894–1898.
- Percus, J. K. & Yevick, G. J. (1958). *Phys. Rev.* **110**, 1–13.
- Pérez, J. & Nishino, Y. (2012). *Curr. Opin. Struct. Biol.* **22**, 670–678.
- Porod, G. (1982). *Small Angle X-ray Scattering* (Glatter, O. & Kratky, O. (eds.), Academic Press, London), chap. 2, 17–52.
- Skou, S., Gillilan, R. E. & Ando, N. (2014). *Nature Prot.* **9**, 1727–1739.
- Sorensen, C. M. & Roberts, G. C. (1997). *J. Colloid Interface Sci.* **186**, 447–452.
- Sorensen, C. M. & Wang, G. M. (1999). *Phys. Rev. E* **60**, 7143–7148.
- Sorensen, C. M. (2001). *Aerosol Sci. Technol.* **35**, 648–687.
- Stefani, M. & Dobson, C. M. (2003). *J. Mol. Med.* **81**, 678–699.
- Svergun, D. I., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.
- Svergun, D. I., Koch, M. H. J., Timmins, P. A., & May, R. P. (2013). *Small Angle X-Ray and Neutron Scattering from Solutions of Biological Macromolecules*. Oxford University Press.
- Teixeira, J. (1988). *J. Appl. Cryst.* (1988). **21**, 781–785.

- Tria, G., Mertens, H. D. T., Kachala, M. & Svergun, D. I. (2015). *IUCrJ* **2**, 207–217.
- Trehella, J., Duff, A.P., Durand, D., Gabel, F., Guss, J. M., Hendrickson, W. A., Hura, G. L., Jacques, D. A., Kirby, N. M., Kwan, A. H., Pérez, J., Pollack, L., Ryan, T. M., Sali, A., Schneidman-Duhovny, D., Schwede, T., Svergun, D. I., Sugiyama, M., Tainer, J. A., Vachette, P., Westbrook, J. & Whitten A. E. (2017). *Acta Cryst. D* **73**, 710–728.
- Vrij, A. (1979). *J. Chem. Phys.* **71**, 3267–3270.

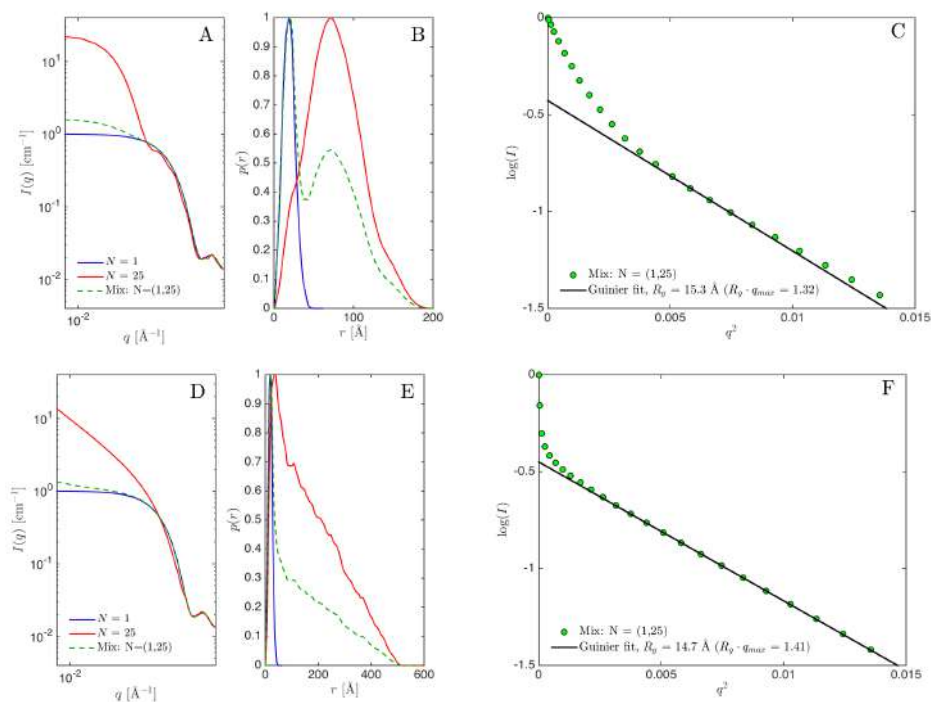


Fig. 1. Calculated scattering and pair distance distribution function for a lysozyme monomer (PDB 1LYZ,  $N = 1$ , blue full line) and lysozyme aggregates ( $N = 25$ , red full line) and for a mixture of the two (green dashed line). (A-B) Spherical aggregates. (D-E) elongated aggregates. (C and F) Guinier fits for the mix of monomers and aggregates. The true  $R_g$  for lysozyme (PDB 1LYZ) is 14.5 Å. The estimated  $R_g$  values from the Guinier fits are given in the legend.

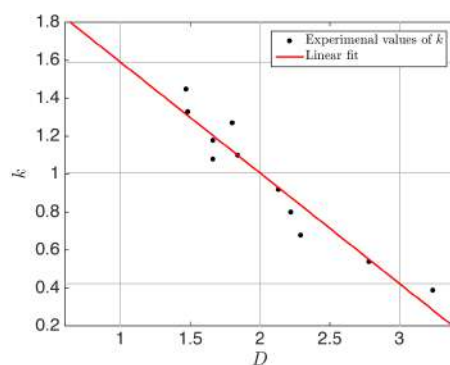


Fig. 2. Linear fit to obtain an empirical relation between the structure coefficient,  $k$ , and the dimensionality,  $D$ . Data collected by Sorensen & Roberts (1997).  $k(D = 1) \approx 1.6$ ,  $k(D = 1) \approx 1.0$  and  $k(D = 3) \approx 0.4$ .



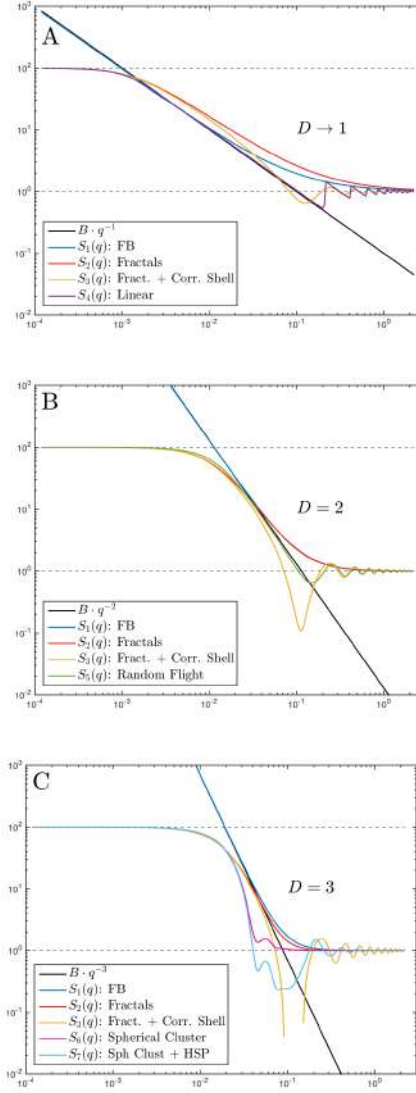


Fig. 3. All structure factors. For  $N = 100$ ,  $R = 15$ ,  $\phi = 0.3$  [ $S_6(q)$ ,  $S_7(q)$ ],  $\sigma_{r,relative} = 0.1$  [ $S_6(q)$ ,  $S_7(q)$ ].  $D$  were varied [ $S_1(q)$ ,  $S_2(q)$ ,  $S_3(q)$ ]. The exponential decay and the FB structure factor,  $S_1(q)$ , were scaled to align with the other structure factors. (A) Dimensionality  $D \rightarrow 1$ . Exp. decay (black), FB structure factor  $S_1(q)$  (blue), mass fractal structure factor  $S_2(q)$  (red), fractals with shell correlation  $S_3(q)$  (yellow), and linear aggregates  $S_4(q)$  (purple). Lower dashed line is 1 (high- $q$  limit) and the upper line is  $N = 100$  (low- $q$  limit). (B) Dimensionality  $D = 2$ . Exp. decay (black),  $S_1(q)$  (blue),  $S_2(q)$  (red),  $S_3(q)$  (yellow) and random flight  $S_5(q)$  (green). (C) Dimensionality  $D = 3$ . Exp. decay (black),  $S_1(q)$  (blue),  $S_2(q)$  (red),  $S_3(q)$  (yellow) and spherical cluster structure factor  $S_6(q)$  (light blue)

IUCr macros version 2.1.10: 2016/01/28

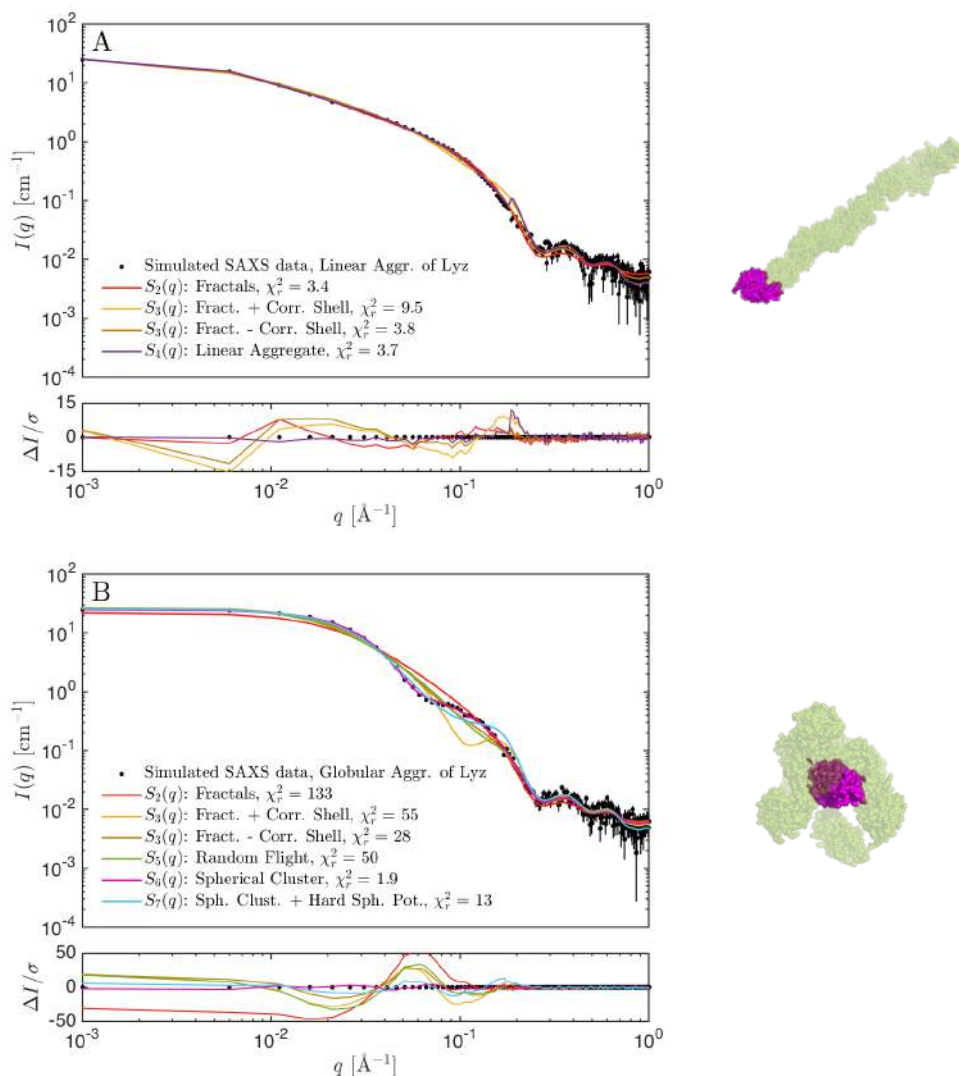


Fig. 4. Simulated data of 25mer oligomeric aggregates of lysozyme fitted with different structure factors. (A) Linear aggregates (black) fitted with fractal structure factor,  $S_2(q)$  (red), fractal structure factor with a correlation shell,  $S_3(q)$  (yellow), and linear aggregate structure factor,  $S_3(q)$  (purple). (B) Spherical aggregates (black) fitted with  $S_2(q)$  (red),  $S_3(q)$  (yellow), random flight structure factor,  $S_5(q)$  (green), spherical cluster structure factor,  $S_6(q)$  (pink), and spherical cluster structure factor with hard sphere potential describing the local interactions,  $S_7(q)$  (light blue).

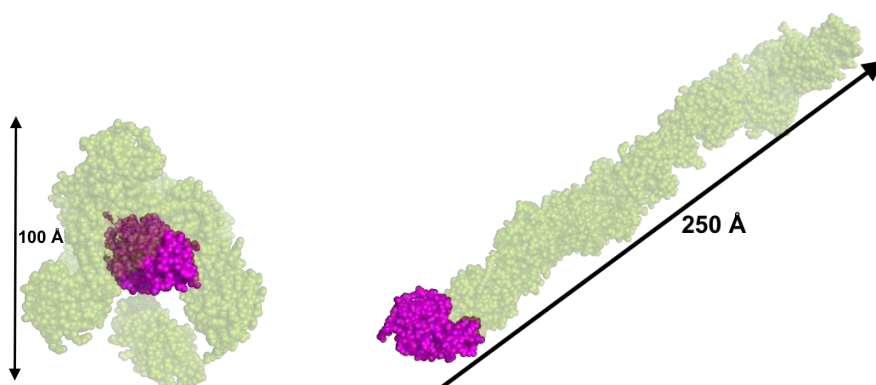


Fig. 5. 10mer aggregates of lysozyme (protein subunit of Lyz in magenta, rest of aggregate in green). (A) Globular aggregate. (B) Linear aggregate.

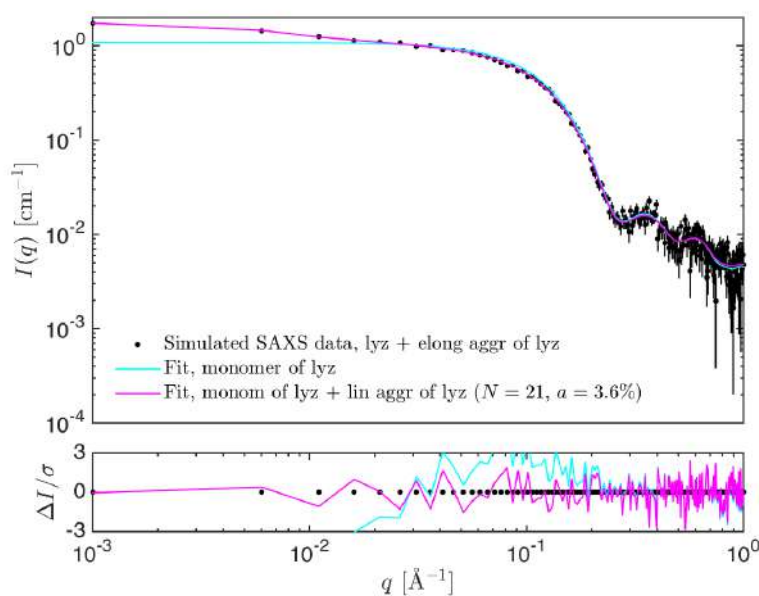


Fig. 6. Simulated SAXS data for a mix ( $a = 3.0\%$ ) of lysozyme (lyz) monomers and 25-mer linear oligomeric aggregates (black), similar to the 10-mers in Fig. 5B. The data were fitted with a model of monomeric lyz (light blue) and with a mix of monomeric lyz and linear aggregates of lyz, using  $S_4(q)$  (magenta).

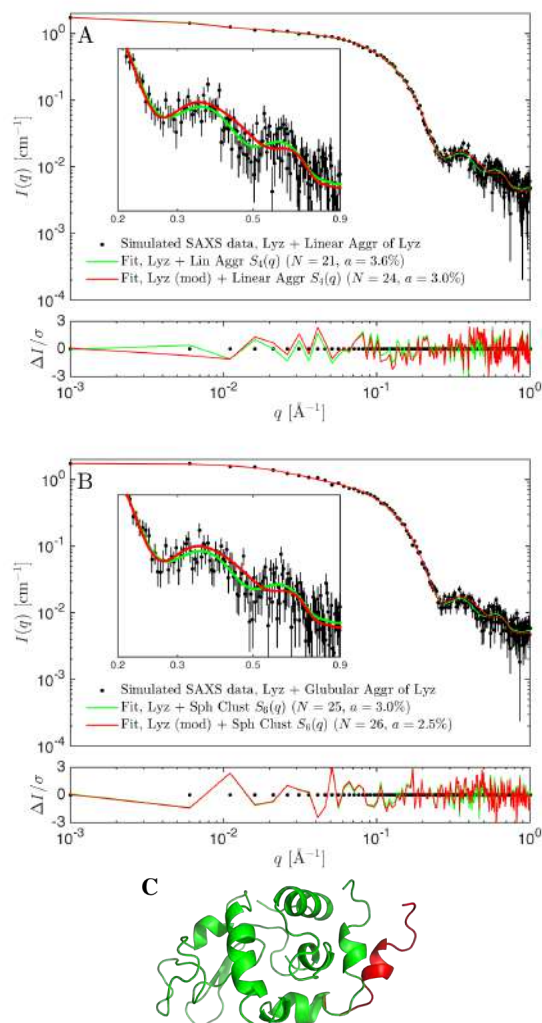


Fig. 7. (A) Simulated SAXS data for a mix of lysozyme (lyz) monomers and 25-mer linear oligomers (black). Simulated using a the crystal structure of lyz (PDB: 1LYZ). The data was fitted with a mix of lyz monomers and linear aggregates, using the spherical cluster structure factor,  $S_6(q)$ . The insert shows the fit at  $q = [0.2, 0.9] \text{Å}^{-1}$ . Two structures of lyz were used in the fit, the true crystal structure (green,  $\chi_r^2 = 1.09$ ), and a modified structure of lyz (red,  $\chi_r^2 = 1.31$ ). (B) As (A), but using globular aggregates in the simulated data, and the spherical cluster structure factor,  $S_6(q)$  for the fit. (C) Visualization of the true structure (green) and the modified structure (red). The modified structure (red) was generated in PyMOL with the true structure (PDB: 1LYZ) as basis. The structures are aligned such that the modified structure is only visible where the structures differ.

---

**Synopsis**

---

## 9.2 Paper II: Analysis of small-angle scattering data using model fitting and Bayesian regularization

**List of Authors** Andreas Haahr Larsen, Lise Arleth and Steen Hansen

**Status** Published in J. Appl. Cryst., 2018, 51, 1151-1161.

**Abstract** The structure of macromolecules can be studied by small-angle scattering (SAS), but as this is an ill-posed problem, prior knowledge about the sample must be included in the analysis. Regularization methods are used for this purpose, as already implemented in indirect Fourier transformation and bead-modeling-based analysis of SAS data, but not yet in the analysis of SAS data with analytical form factors. To fill this gap, a Bayesian regularization method was implemented, where the prior information was quantified as probability distributions for the model parameters and included via a functional  $S$ . The quantity  $Q = 2 + S$  was then minimized and the value of the regularization parameter determined by probability maximization. The method was tested on small-angle X-ray scattering data from a sample of nanodiscs and a sample of micelles. The parameters refined with the Bayesian regularization method were closer to the prior values as compared with conventional  $2$  minimization. Moreover, the errors on the refined parameters were generally smaller, owing to the inclusion of prior information. The Bayesian method stabilized the refined values of the fitted model upon addition of noise and can thus be used to retrieve information from data with low signal-to-noise ratio without risk of overfitting. Finally, the method provides a measure for the information content in data,  $N_g$ , which represents the effective number of retrievable parameters, taking into account the imposed prior knowledge as well as the noise level in data.

**Contributions by AHL** SH, AHL and LA developed the idea for the project, together with Martin Cramer Pedersen (in acknowledgments). AHL designed the study, generated all figures and wrote the paper. AHL developed and tested the software, which was written by SH. AHL parametrized the models and introduced molecular constraints together with LA, and the models were then implemented by SH. Theory developed by SH and AHL.



# Analysis of small-angle scattering data using model fitting and Bayesian regularization

Andreas Haahr Larsen,\* Lise Arleth and Steen Hansen

Niels Bohr Institute, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark. \*Correspondence e-mail: andreas.larsen@nbi.ku.dk

Received 7 February 2018  
Accepted 19 June 2018

Edited by D. I. Svergun, European Molecular Biology Laboratory, Hamburg, Germany

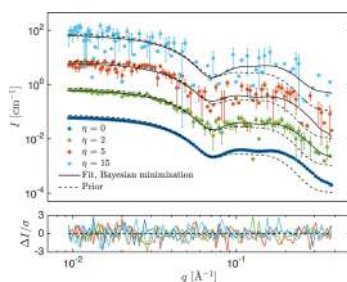
**Keywords:** small-angle scattering; Bayesian regularization; information content; molecular constraints.

The structure of macromolecules can be studied by small-angle scattering (SAS), but as this is an ill-posed problem, prior knowledge about the sample must be included in the analysis. Regularization methods are used for this purpose, as already implemented in indirect Fourier transformation and bead-modeling-based analysis of SAS data, but not yet in the analysis of SAS data with analytical form factors. To fill this gap, a Bayesian regularization method was implemented, where the prior information was quantified as probability distributions for the model parameters and included *via* a functional  $S$ . The quantity  $Q = \chi^2 + \alpha S$  was then minimized and the value of the regularization parameter  $\alpha$  determined by probability maximization. The method was tested on small-angle X-ray scattering data from a sample of nanodiscs and a sample of micelles. The parameters refined with the Bayesian regularization method were closer to the prior values as compared with conventional  $\chi^2$  minimization. Moreover, the errors on the refined parameters were generally smaller, owing to the inclusion of prior information. The Bayesian method stabilized the refined values of the fitted model upon addition of noise and can thus be used to retrieve information from data with low signal-to-noise ratio without risk of overfitting. Finally, the method provides a measure for the information content in data,  $N_g$ , which represents the effective number of retrievable parameters, taking into account the imposed prior knowledge as well as the noise level in data.

## 1. Introduction

Small-angle scattering (SAS) is widely used for investigating the low-resolution structure of macromolecules (Svergun & Koch, 2003; Svergun *et al.*, 2013). Physical quantities such as the radius of gyration and molecular weight can be obtained directly from the data, and the overall structure of the macromolecules can be probed indirectly by modeling.

Deducing a structure exclusively from SAS data is an ill-posed problem, meaning that several structures can explain the data. In SAS modeling with analytical form factors, a geometrical model that describes the scattering intensity in terms of a set of model parameters is tested against data (see *e.g.* Pedersen, 1997). Typical parameters include particle dimensions, excess scattering length densities, concentration *etc.* These parameters are then refined to obtain the values that provide the best fit to data. In order to circumvent the ill-posed nature of the problem and minimize the number of free parameters, Hayter & Penfold (1981) introduced molecular constraints in an early small-angle neutron scattering (SANS) study of SDS micelles. This allowed for explicit use of the information available about the SDS chemical structure, the partial specific molecular volumes and the sample concentration, such that the model could be reparametrized into a



© 2018 International Union of Crystallography

## research papers

minimal number of free parameters. The core-shell micelle model and associated interparticle structure factor were reparametrized into a particularly simple model with only two free parameters: the charge and aggregation number of the micelles. The approach of using molecular constraints has been generalized to various later and more complicated applications in SAS (e.g. by Cabane *et al.*, 1985; Arleth *et al.*, 1997; Kučerka *et al.*, 2004; Skar-Gislinge & Arleth, 2011). However, the approach may lead to an over-constrained fit where the experimental data cannot be fitted. This will often be the case if one or more of the fixed parameters are slightly wrong. At the same time, all information about the fixed parameters in the new data is ignored. To circumvent these problems, model parameters that, according to Hayter & Penfold (1981), should ideally be well known and kept fixed are instead taken as free parameters. This may, on the other hand, create a situation where the most optimal fit has unrealistic values for central parameters; for example, the fitted concentration could be incompatible with an independent concentration assessment, the shape of the particle unrealistic, or the fitted internal scattering length densities too far from the expected values. If the overall model is trusted, this creates a situation where the scientist has to make a choice: either the inconsistent parameters are fixed, thereby ignoring any information about those parameters in the new data and possibly having to accept a poor fit, or alternatively, the new refined values are trusted, thus effectively ignoring the prior knowledge. Clearly, none of these solutions are optimal and an improved framework for inclusion of the prior knowledge is required.

As will be shown in the following, regularized expressions provide such a framework and can be utilized to include prior knowledge directly in the data analysis. Regularization methods are already used extensively in the analysis of SAS data, for example in indirect Fourier transformation (Glatter, 1977; Svergun, 1992), where a smoothness constraint is imposed on the pair distance distribution function, in *ab initio* modeling (Svergun, 1999), where a compactness constraint is applied to the refined models, and in rigid-body modeling (Petoukhov & Svergun, 2005), where regularization terms prevent overlap of the rigid bodies and ensure that the solution does not diverge significantly from known residue distances. However, to the best of our knowledge they have not been used in the analysis of SAS data modeled with analytical form factors, as proposed in the present work.

In this paper, a regularization method that allows for inclusion of prior knowledge and avoids fixing parameter values is presented. The prior knowledge is quantified as probability distributions, so-called priors. The approach exploits Bayesian statistics, which provides an ideal framework for inclusion of priors in analysis of experimental data. Bayesian methods have been used for decades in the field of image processing (see e.g. Gull, 1989; Schultz & Stevenson, 1994) and more recently in the processing of electron microscopy images, as implemented, for example, in the program *RELION* (Scheres, 2012). Moreover, Bayesian statistics is used in the effort of effectively combining experimental data

with molecular dynamics simulations, as presented for instance in the recent paper by Shevchuk & Hub (2017).

The second issue treated in the present paper is the quantification of information in data. It is of fundamental interest to assess the information in experimental data and thus be able to optimize the information content under different experimental conditions that may be varied, such as concentration, exposure time and neutron contrast situation (Pedersen *et al.*, 2014), and it will be argued that the ‘number of good parameters’  $N_g$  constitutes a suitable measure for that purpose.  $N_g$ , as introduced by Gull (1989), has been discussed in relation to indirect Fourier transform of SAS data by Müller *et al.* (1996) and by Vestergaard & Hansen (2006), and in the present paper we show how it applies in the context of SAS data analysis using analytical form factors.

## 2. Theory

In conventional analysis of SAS data with analytical form factors, a mathematical model is hypothesized, which describes the theoretical intensity and can be tested against data (see e.g. Pedersen, 1997). The model is expressed in terms of a set of model parameters, for example the particle dimension, the contrast situation, the concentration or the polydispersity of the sample. These parameters are refined by minimizing the likelihood function,  $\chi^2$ , defined in terms of the theoretical intensities  $I^{\text{th}}$  and the experimentally measured intensities  $I^{\text{exp}}$  as

$$\chi^2(\mathbf{p}) = \sum_{i=1}^N \frac{[I_i^{\text{exp}} - I_i^{\text{th}}(\mathbf{p})]^2}{\sigma_i^2}. \quad (1)$$

Here,  $N$  is the number of data points and  $\sigma_i$  is the experimental standard deviation of data point  $i$ .  $I_i^{\text{th}}(\mathbf{p})$  is assumed to be a function of  $K$  model parameters  $\mathbf{p} = (p_1, \dots, p_K)$ . Both experimental and theoretical intensities are functions of the momentum transfer,  $q$ , given in terms of the wavelength of the incoming beam  $\lambda$  and the scattering angle  $2\theta$ ,  $q = 4\pi \sin(\theta)/\lambda$ . The detector image is azimuthally averaged and binned into discrete  $q$  values such that the intensity is also discretized, i.e.  $I_i = I(q_i)$ . The reduced  $\chi^2$  is used to assess the goodness of fit and is defined as  $\chi_r^2 = \chi^2/f$ , where  $f$  is the number of degrees of freedom, conventionally found as  $f = N - K$ . Residual plots are used to evaluate the goodness of fit visually and give the difference in intensity in units of  $\sigma$ , i.e.  $(\Delta I/\sigma)_i = (I_i^{\text{exp}} - I_i^{\text{th}})/\sigma_i$ .

In the Bayesian approach, the prior knowledge is directly incorporated in the minimization process through a functional,  $S(\mathbf{p})$ , that gives a penalty to solutions with parameter values far from the prior values. We will assume normally distributed priors with mean values  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$  and standard deviations  $\boldsymbol{\delta p} = (\delta p_1, \dots, \delta p_K)$ . Then  $S(\mathbf{p})$  takes the form

$$S(\mathbf{p}) = \sum_{k=1}^K \frac{(p_k - \mu_k)^2}{\delta p_k^2}. \quad (2)$$

$\mu_k$  and  $\delta p_k$  reflect the prior knowledge about the  $k$ th parameters. If this comes from a measurement, or a previous



## research papers

experiment, a mean and a standard deviation is usually available. If the prior, on the other hand, is based on general biophysical knowledge about the system, this knowledge must be expressed in terms of  $\mu_k$  and  $\delta p_k$ . If almost no knowledge is available, a mostly non-informative prior should be used, for example a uniform prior or a very wide normal distribution. The determination of priors is exemplified and explained for the two experimental examples in §3.  $\chi^2(\mathbf{p})$  is then replaced in the minimization routine by the expression

$$Q(\mathbf{p}) = \chi^2(\mathbf{p}) + \alpha S(\mathbf{p}), \quad (3)$$

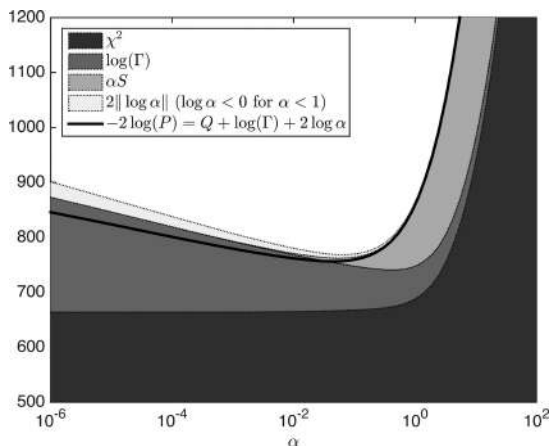
where  $\alpha$  is a regularization parameter, balancing the influence of the prior knowledge ( $S$ ) and the data ( $\chi^2$ ).

### 2.1. Determining $\alpha$ and introducing the Bayesian Occam term

The Bayesian method provides a consistent way of determining  $\alpha$ , the regularization parameter.  $\alpha$  is a so-called hyperparameter and must be determined by other means than the model parameters (MacKay, 1999; Hansen, 2000), namely by maximizing the probability for  $\alpha$  and the data  $D$  given the hypothesized model  $H$ . Using standard probability rules, we can express this probability as a product,

$$P(D, \alpha | H) = P(D | \alpha, H) P(\alpha), \quad (4)$$

where  $P(D | \alpha, H)$  is the evidence, describing the probability for the data set given both  $\alpha$  and the model. For a more elaborate introduction to the evidence and Bayesian probability theory see, for example, Bolstad (2007).  $P(\alpha)$  is the prior for  $\alpha$ . As  $\alpha$  is a so-called scale parameter, Jeffreys' prior,  $P(\alpha) = 1/\alpha$  (Jeffreys, 1946), is used in the following. Also, it is exploited that minimizing  $-2 \log[P(D, \alpha | H)]$  is analogous to maximizing  $P(D, \alpha | H)$ . Denoting by  $\mathbf{A}$ ,  $\mathbf{B}$  and  $\mathbf{C}$  the curvature matrices  $\mathbf{A} = \nabla \nabla \alpha S$ ,  $\mathbf{B} = \nabla \nabla \chi^2$  and  $\mathbf{C} = \nabla \nabla Q$ , and denoting by  $\Gamma$  the fraction  $\Gamma = \det(\mathbf{C})/\det(\mathbf{A})$ , it can be shown that (Hansen, 2000)



**Figure 1**  
Graphical representation of equation (5) for the nanodisc example. The optimal value of  $\alpha$  is found at the minimum ( $\alpha = 0.24$ ). Note that the lower y limit is 500, i.e.  $\chi^2$  constitutes the major contribution.

$$-2 \log[P(D, \alpha | H)] = Q(\mathbf{p}) + \log(\Gamma) + 2 \log(\alpha), \quad (5)$$

where  $Q(\mathbf{p})$  is defined in equation (3) and the third term is the Jeffreys prior for  $\alpha$ .  $\Gamma$  plays a significant role in the analysis: the determinant  $\det(\mathbf{A})$  is given as  $\alpha(\prod_{j=1}^K \delta p_j)^{-2}$ , i.e. it is inversely proportional to the squared product of the standard deviations of the priors for the model parameters. This product spans the volume in the parameter space where the solution is expected to exist *a priori*. The determinant  $\det(\mathbf{C})$  can be written as  $\det(\nabla \nabla \chi^2 + \alpha \nabla \nabla S)$ , where the curvature matrix  $\nabla \nabla \chi^2$  depends on the analytical model and must be found numerically. So the expression cannot be simplified any further in the general case. However,  $\det(\mathbf{C})$  is generally inversely proportional to the *a posteriori* solution volume. In summary,  $\Gamma \propto (\text{a priori volume})/(\text{a posteriori volume})$ .

In the simplest possible solution where the data contain no new information about the parameters ( $\nabla \nabla \chi^2 = 0$ ), the two volumes are identical, i.e. the prior knowledge is not altered, and  $\log(\Gamma)$  is zero. Otherwise, the term will be positive, since the *a priori* volume is generally larger than the *a posteriori* volume. Hence, the term favors simple solutions and will be denoted the Occam term (MacKay, 1992). The contributions of all terms of equation (5) are shown graphically for the nanodisc example in Fig. 1, and it is clearly seen how the Occam term ‘pushes’ the solution towards higher  $\alpha$  values, i.e. towards simpler solutions closer to the prior.

### 2.2. Quantifying the information content in data

Following the argumentation in previous work (Gull, 1989; Müller *et al.*, 1996; Vestergaard & Hansen, 2006), the information content can be quantified as the number of good parameters  $N_g$ , describing the effective number of free parameters retrievable by the data. It is defined in terms of  $\alpha$  and the eigenvalues  $\eta_i$  and  $\gamma_i$  of the diagonalized curvature matrices  $\mathbf{B}$  and  $\mathbf{C}$ , respectively. By change of units  $C_{ij} \rightarrow C_{ij} \delta p_i \delta p_j$ , the eigenvalues of  $\mathbf{C}$  can be written as  $\gamma_i = \alpha + \eta_i$ , and  $N_g$  can then be expressed simply in terms of  $\alpha$  and  $\eta_i$  as

$$N_g = \sum_{i=1}^K \frac{\eta_i}{\gamma_i} = \sum_{i=1}^K \frac{\eta_i}{\alpha + \eta_i}, \quad (6)$$

where  $K$  is the number of parameters in the model. The measure is similar in methodology to single value decomposition, i.e. the model is, so to say, redescribed in a new basis. The good parameters do not therefore correspond directly to parameters in the investigated model, but  $N_g$  is the minimum number of independent effective parameters retrievable from the data. The magnitude of  $\eta_i$  (eigenvalue  $i$  of  $\mathbf{B} = \nabla \nabla \chi^2$ ) expresses the significance of the  $i$ th effective parameter. All eigenvalues are positive, but some are very small compared with  $\alpha$ . If an eigenvalue is very large,  $\eta_i \gg \alpha$ , it will contribute 1 to  $N_g$ , and if  $\eta_i \ll \alpha$ , then  $\eta_i$  will not contribute to the sum at all. Thus  $N_g$  is between 0 and  $K$ . The information may be distributed evenly among the physical model parameters, but the data may also contain much information about some parameters and very limited information about others. This

## research papers

will be reflected in the difference between the prior and the posterior distribution for each parameter.

### 3. Methods

#### 3.1. Experimental examples

To test the method, we analyzed the experimental small-angle X-ray scattering (SAXS) data from two different macromolecular samples.

The first sample contained nanodiscs of 1,2-dilauroyl-*sn*-glycero-3-phosphocholine (DLPC) and the membrane scaffolding protein MSP1D1, measured at 293 K. The data set was previously obtained and analyzed by Skar-Gislinge *et al.* (2010). The nanodisc is a composite particle consisting of a phospholipid bilayer surrounded by two amphipathic and  $\alpha$ -helical scaffolding proteins that form a stabilizing belt around the hydrophobic edge of the bilayer (Fig. 2). Each belt protein has a protruding His tag with a tobacco etch virus (TEV) cleavage site, and these were modeled as random Gaussian coils. The nanodisc itself was modeled by combining analytical form factor amplitudes, as described and illustrated by Skar-Gislinge & Arleth (2011). In brief, the bilayer was described as stacked elliptical cylinders with different scattering length densities, and the two scaffolding proteins were collectively described as a homogeneous hollow cylinder with elliptical cross section. For the purpose of the present work, the model was parametrized to have 12 physically relevant parameters, as listed in Table 1. The parameters were background  $B$ , concentration  $c$ , molecular volume of the lipids  $V_l$ , molecular volume of the lipid tailgroups  $V_t$ , volume of the protein  $V_p$ , number of lipids per nanodisc  $N$ , number of water molecules per lipid headgroup  $n_w$ , thickness of the protein belt

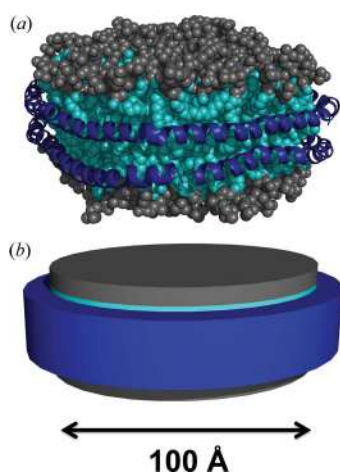


Figure 2

Illustration of a nanodisc. (a) All-atom structure from Shih *et al.* (2007), with a hydrophobic core of lipid tails (turquoise), caps of hydrophilic lipid headgroups (gray), and a surrounding ‘belt’ of two amphipathic and  $\alpha$ -helical proteins (blue). (b) Analytical nanodisc model with dimensions corresponding to the prior values in Table 1. The His tags with TEV sites were not included in the illustrations.

Table 1

Refined parameter values from the analysis of the nanodisc data set, comparing the Bayesian regularization method with conventional  $\chi^2$  minimization.

One standard deviation is given as error (in parentheses). The prior values are listed in the middle column in terms of the mean (and standard deviation) of the respective prior normal distributions. The goodness of the fits were evaluated with the reduced  $\chi^2$  and the Cmap test (Franke *et al.*, 2015).

Model parameter	$\chi^2$ minimization	Prior	Bayesian minimization
$N$	103 (22)	152.0 (10.0)	119 (7)
$\varepsilon$	1.3 (4.5)	1.40 (0.15)	1.33 (0.03)
$A$ ( $\text{\AA}^2$ )	76 (19)	61 (5)	70 (4)
$n_w$	18 (12)	8 (2)	10 (3)
$V_l$ ( $\text{\AA}^3$ )	996 (19)	985 (30)	1001 (3)
$V_t$ ( $\text{\AA}^3$ )	702 (111)	666 (20)	684 (22)
$V_p$ ( $\times 10^4$ $\text{\AA}^3$ )	5.3 (0.7)	5.7 (0.2)	5.4 (0.2)
$T$ ( $\text{\AA}$ )	11.4 (2.1)	10.0 (0.3)	10.2 (0.6)
$R_g$ ( $\text{\AA}$ )	13.8 (1.2)	12.5 (1.0)	14.1 (0.7)
$\sigma_R$ ( $\text{\AA}$ )	3.2 (1.0)	6.0 (1.0)	3.3 (0.5)
$c$ ( $\mu\text{M}$ )	22 (19)	22.6 (3.0)	23.3 (4.9)
$B$ ( $\times 10^{-4}$ $\text{cm}^{-1}$ )	-0.3 (2.1)	1.0 (10.0)	0.1 (1.1)
Goodness of fit, $\chi_r^2$	6.26	–	6.30
Goodness of fit, $C = 10$ , $N = 106$	–	–	$C = 10$ , $N = 106$
Cmap $P(C \geq 10   N) = 9.2\%$	–	–	$P(C \geq 10   N) = 9.2\%$

$T$ , surface roughness  $\sigma_R$  [implemented as in the work of Skar-Gislinge *et al.* (2010)], area per lipid  $A$ , ellipticity of the disk  $\varepsilon$  and radius of gyration of the random Gaussian coils  $R_g$ .  $V_l$  was determined by densitometry with an estimated 2% uncertainty and  $V_t$  was given by Tanford’s formula (Tanford, 1972), also with an estimated uncertainty of 2%, and from these, the volume of the lipid headgroups could be calculated as  $V_h = V_l - V_t$ .  $V_p$  was calculated by summing the atomic van der Waals volumes (Svergun *et al.*, 1995), assuming a relative error of 4%. Excess scattering length densities,  $\Delta\rho$ , were calculated from the molecular volumes and scattering lengths, with the latter calculated from the chemical composition of the relevant molecules.  $T$  was known approximately from the  $\alpha$ -helical structure of the protein belt, and the priors for  $A$  and  $n_w$  were estimated in accordance with the work of Kučerka *et al.* (2005). SAS experiments on similar systems (Midtgaard *et al.*, 2015; Kynde *et al.*, 2014) were used to estimate the prior for  $\varepsilon$ . Finally, the prior for  $R_g$  was estimated from molecular dynamics simulations of proteins with random coil structure by Fitzkee & Rose (2004).

The second example was a sample of self-assembled *N*-dodecyl- $\beta$ -maltoside (DDM) micelles, measured at room temperature. The micelles were modeled as core-shell ellipsoids (Pedersen, 1997), using seven parameters, as listed in Table 2. The seven parameters were constant background  $B$ , concentration  $c$ , scattering contrast of the detergent headgroups in the shell  $\Delta\rho_h$  and of the detergent tailgroups in the core  $\Delta\rho_t$ , number of detergents per micelle  $N$ , ellipticity  $\varepsilon$  of the micelle, and surface roughness  $\sigma_R$ . The form factor and parametrization are as described by Arleth *et al.* (1997), with a roughness term added, as in the nanodisc model. The partial specific molecular volumes used to determine the scattering

**Table 2**  
Refined parameter values for the micelle data set.

Notation as in Table 1, and  $b_e$  is the electron scattering length (2.82 fm).

Model parameter	$\chi^2$ minimization	Prior	Bayesian minimization
$N$	125.0 (0.3)	130 (15)	125.0 (0.3)
$\varepsilon$	0.5398 (0.0007)	1.00 (0.30)	0.5398 (0.0007)
$\Delta\rho_h$ ( $b_e \text{ \AA}^{-3}$ )	0.183 (0.033)	0.184 (0.013)	0.184 (0.006)
$\Delta\rho_t$ ( $b_e \text{ \AA}^{-3}$ )	-0.055 (0.010)	-0.056 (0.006)	-0.056 (0.002)
$\sigma_R$ ( $\text{\AA}$ )	5.41 (0.03)	6.0 (1.0)	5.41 (0.03)
$c$ (mM)	30.3 (11.0)	30.0 (3.0)	29.8 (1.9)
$B$ ( $\times 10^{-3} \text{ cm}^{-1}$ )	0.89 (0.01)	1.0 (10.0)	0.89 (0.01)
Goodness of fit, $\chi_r^2$	170	–	170
Goodness of fit, $C = 36, N = 90$	–	–	$C = 36, N = 90$
Cmap	$P(C \geq 10   N) \simeq 0\%$	–	$P(C \geq 10   N) \simeq 0\%$

contrasts,  $\Delta\rho_h$  and  $\Delta\rho_t$ , were found with densitometry and the volumes were assumed to have a relative uncertainty of 2% (supporting information of Midtgaard *et al.*, 2018). The priors for  $N$  were estimated according to Oliver *et al.* (2013), and the detergent concentration was determined by weighing the added detergent in the stock solution before making the samples, with an estimated uncertainty of 10%.

### 3.2. Implementation of the Bayesian optimization routine

The Bayesian fitting algorithm was implemented in Fortran 77 and the source code is freely available online (<https://github.com/Niels-Bohr-Institute-XNS-StructBiophys/BayesFit>). A Levenberg–Marquardt algorithm (Levenberg, 1944; Marquardt, 1963) was used to minimize  $Q(\mathbf{p})$ . It was implemented with minor modifications of the algorithm from *Numerical Recipes* (Press *et al.*, 1992) and with the parameters constrained to a range defined by the prior mean  $\mu_i$  and standard deviation  $\delta p_i$  such that  $\mu_i - 5\delta p_i < p_i < \mu_i + 5\delta p_i$ . A golden section search was used to determine the most probable  $\alpha$ , assuming that  $-10 < \log(\alpha) < 10$ . The CPU time for the refinement of the nanodisc model is about 20 min on a typical PC, searching 17  $\alpha$  values to determine the optimal  $\alpha$ . The CPU time for conventional  $\chi^2$  minimization is thus 17 times faster, *i.e.* approximately 1 min. The CPU time for the micelle model is only about 2 min with 19 steps in  $\alpha$  (*i.e.* less than 10 s for a  $\chi^2$  minimization). Parallelization has not been included in the present implementation but is in principal easy to implement, since the calculations for each  $q$  value are independent. With other  $\alpha$ -optimization algorithms, the  $\alpha$  calculations would also be independent and thus parallelizable, for example with grid search or random search (Bergstra & Bengio, 2012).

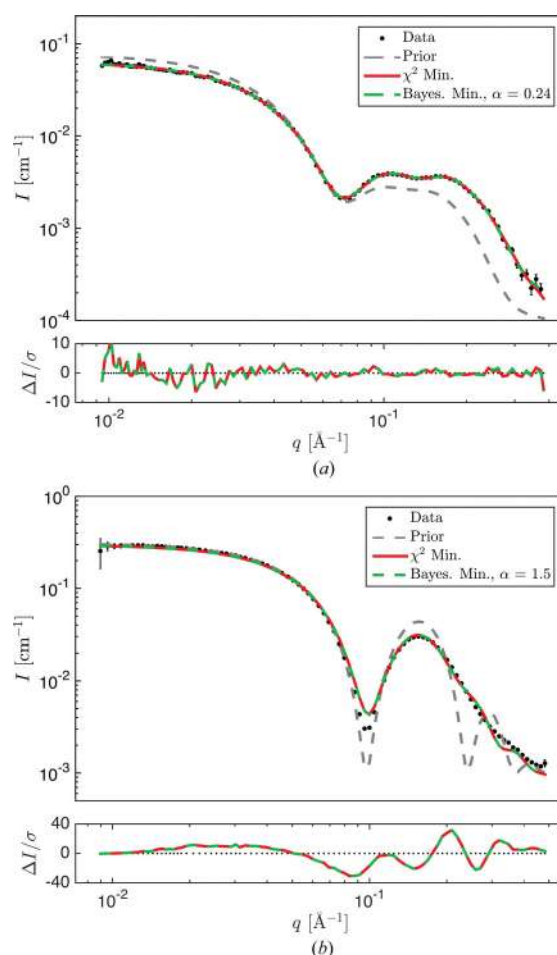
## 4. Results

### 4.1. Nanodiscs

The Bayesian approach was compared with conventional  $\chi^2$  minimization. As seen in Fig. 3(a), both methods found a solution that fitted the data well. The conventional method

varied the 12 parameters freely to minimize  $\chi^2$ , with the mean of the prior values used as the starting point for the fitting routine. In the Bayesian approach, the most probable  $\alpha$  was determined, and the parameters were refined as described in §§2 and 3. The optimal  $\alpha$  was found at 0.24. Moreover, to monitor the effect of  $\alpha$ , a minimization of  $Q$  [equation (3)] was performed for a range of logarithmically spaced values of  $\alpha$  from  $10^{-10}$  to  $10^{10}$ , and  $-2\log[P(D, \alpha | H)]$  [equation (5)] was calculated at each step.

The refined values of the fitting parameters obtained with both the Bayesian and the  $\chi^2$ -minimization methods are listed in Table 1. The parameters refined by the Bayesian approach are generally closer to the prior and have smaller uncertainties, as a consequence of including the regularization term. Notice, for example, that the area per lipid headgroup,  $A$ , was refined to  $70 \pm 4$  with the Bayesian method (prior value  $61 \pm 5$ ) as compared to  $76 \pm 19$  with  $\chi^2$  minimization, and  $N$  was



**Figure 3**

Analyzed examples of SAXS data sets for (a) a nanodisc sample and (b) a sample of detergent micelles. The data sets (black points with error bars) were fitted using conventional  $\chi^2$  minimization (red solid line) and Bayesian minimization (green dashed line). The gray dashed line is the prior. Residual plots are shown below, where  $\Delta I = I_{\text{exp}} - I_{\text{fit}}$  and  $\sigma$  is the experimental standard deviation.

## research papers

refined to, respectively,  $119 \pm 7$  and  $103 \pm 22$  with the Bayesian and the conventional methods (prior value  $152 \pm 10$ ). These two parameters have been plotted for a range of  $\alpha$

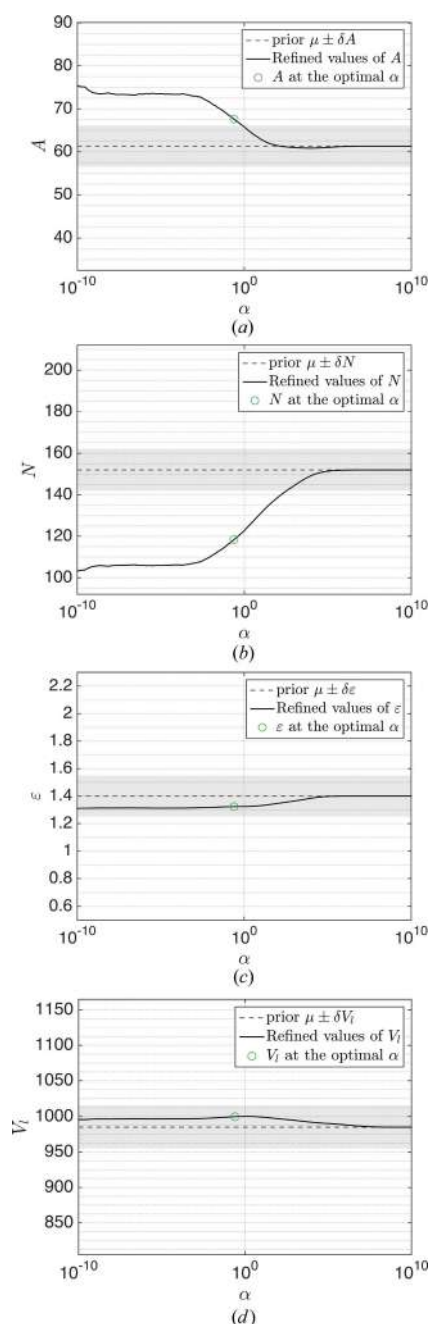


Figure 4

The refined value of four different parameters for the nanodisc model, as a function of  $\alpha$ . The refined parameter values for the optimal  $\alpha$  are marked by a green ring. The gray dashed line and the gray shaded area show, respectively, the prior mean and the prior standard deviation. Some parameters were significantly altered by the prior, e.g.  $A$  (a) and  $N$  (b), whereas other parameters were virtually unaffected, e.g.  $\varepsilon$  (c) and  $V_l$  (d).

values in Figs. 4(a) and 4(b), and they clearly approach the prior value as  $\alpha$  increases. The refined values were thus influenced concurrently by the SAXS data and the prior. In Fig. 5 the prior, likelihood and posterior distributions for  $N$  are plotted, clearly showing how the refined value for  $N$  using the Bayesian method (posterior distribution) is affected both by the prior and by the likelihood. Figs. 4(c) and 4(d) show the values of  $\varepsilon$  and  $V_l$ , which were not affected significantly by the prior at the optimal  $\alpha$ . Generally, parameters are mostly effected by the prior if, firstly, there is a large discrepancy between the prior mean value and the likelihood value (see Fig 5), secondly,  $\delta p$  (the prior width) is narrow, and, thirdly, the parameters have little effect on  $\chi^2$ .

#### 4.2. Detergent micelles

In the micelle example, both the  $\chi^2$  minimization and the Bayesian minimization found a solution that fitted the data relatively well as judged by visual inspection (Fig. 3b), and the regularization parameter,  $\alpha$ , was optimized to 1.5. The residual plot reveals some systematic discrepancies. This is verified by a correlation map (Cmap) test (Franke *et al.*, 2015), from which it can be concluded that the data are significantly different from the model [significance level 1%,  $C = 36$ ,  $P(C \geq 36 | N = 90) \simeq 0\%$ ]. The monodisperse prolate ellipsoidal model is thus not a perfect description of the physical micelles, but constitutes an approximate model. In the micelle example the prior had only a minor effect on the fitted results, as seen from Table 2. This means that the global minimum for  $\chi^2$  in the parameter space is physically meaningful and consistent with the prior. While the prior hardly affects the model parameters, it does lead to more reasonable errors (Table 2). Note that the concentration had a prior value of  $30.0 \pm 3.0$  mM. The error should decrease after taking the

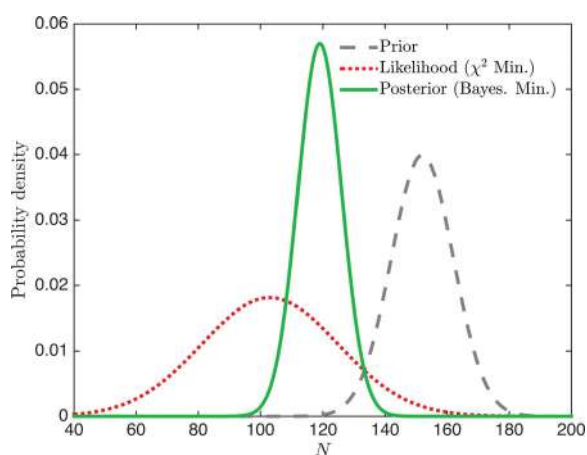


Figure 5

Probability distributions for  $N$  in a nanodisc sample.  $N$  was refined with  $\chi^2$  minimization to obtain the likelihood distribution (red dotted line) and with Bayesian minimization to obtain the posterior distribution (green solid line), which was regularized by the prior distribution (gray dashed line).



## research papers

SAS data into account, since these data refined the concentration to a value very close to the prior value (30.3 and 29.8 for the conventional and Bayesian methods, respectively). Thus, the error of  $\pm 1.9$  found with the Bayesian approach is more sensible than the error of  $\pm 11.0$  found with conventional  $\chi^2$  minimization. The same applies for the refined values of  $\Delta\rho_h$  and  $\Delta\rho_t$ .

#### 4.3. The regularization stabilizes the solution upon addition of noise

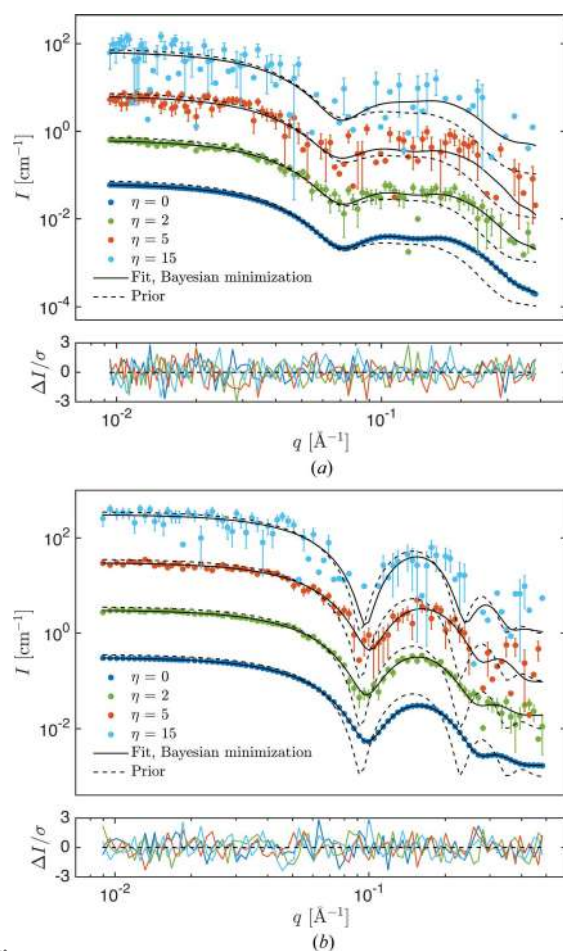
Noisy data were simulated with different noise levels to examine the influence of the Bayesian regularization on noisy data. The best fits for the nanodisc and the micelle data sets were used to generate respective simulated data sets. Standard deviations (error bars) were assigned to each point in  $q$  by  $\sigma(q) = \eta[I_{\text{fit}}(q)]^{1/2} + B$ , where  $I_{\text{fit}}(q)$  is the refined fit value found by the Bayesian approach,  $\eta$  is a relative noise parameter and  $B$  is a constant noise level, set to  $B = 10^{-5}$ . The

simulated intensities were randomly sampled from a normal distribution with mean  $\mu = I_{\text{fit}}(q)$  and standard deviation  $\sigma$ . The simulated data and corresponding fits for selected noise levels can be seen in Fig. 6. As in the experimental situation, the prior differs slightly from the simulated data, and it is also plotted in Fig. 6.

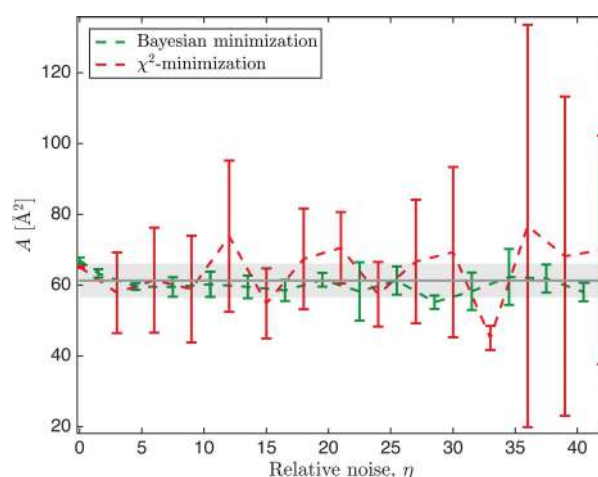
For each noise level, several data sets were generated by random sampling from the normal distribution and fitted with the model, so the variation in the refined parameter values could be evaluated. This is shown for  $A$  in Fig. 7, where each point is the mean value of five runs simulated with the same noise level and the error bars are standard deviations. The final refined value of  $A$  was stabilized considerably in the Bayesian method as compared to the conventional method, expressed by a nearly constant mean value for all noise levels and small standard deviations.

#### 4.4. The information content in data

The information content for the nanodisc SAXS data, according to equation (6) and given the prior, was  $N_g = 9.1$ , while the number of fitted parameters was 12: that is, 12 parameters were refined, but the information coming from the SAXS data corresponded to nine parameters. The rest of the information came from the prior. For the micelle data set, the information content from the SAXS data was  $N_g = 6.0$ , while the model had seven fitting parameters. Therefore, in both cases, the parameters were refined mainly from the SAXS data and to a lesser degree from the prior. However, when analyzing the simulated data with added noise, the prior played a greater role. In Fig. 8(b),  $N_g$  is plotted for an increasing value of the relative noise parameter  $\eta$ .  $N_g$  decreases from around 10 (nanodisc example) and 7 (micelle example) at  $\eta = 0$  to  $N_g < 3$  (both cases) at  $\eta = 40$ : that is, for noisy data sets, the refined parameters are mainly determined



**Figure 6**  
Simulated data with increasing relative noise,  $\eta = 0$  (blue),  $\eta = 2$  (green),  $\eta = 5$  (red) and  $\eta = 15$  (cyan). Fit with Bayesian minimization (solid line) and regularized with the prior (dashed line). (a) Simulated nanodisc data and (b) simulated micelle data.



**Figure 7**  
Refined value for the area per headgroup  $A$ , found by  $\chi^2$  minimization (red) and by Bayesian minimization (green), for increasing relative noise  $\eta$ . The prior, used in the Bayesian minimization, is shown with a gray line for the mean and a gray area for the standard deviation.

## research papers

by the prior. In accordance with our intuition, this shows that less information can be obtained from noisy data, but intriguingly, it also implies that, since the risk of fitting the noise in data is circumvented by the prior, some information can still be extracted with the Bayesian regularization method, even from very noisy data. This would not be possible with the conventional approach, owing to the large fluctuations of the refined parameter values, as exemplified in Fig. 7. The information content depends on the value of  $\alpha$ , *i.e.* on how the prior information is weighted with respect to the new data set. In Fig. 8(a), it is shown how  $N_g$  decreases as  $\alpha$  increases, from  $N_g \simeq K$  at  $\alpha = 10^{-10}$  ( $K = 12$  for the nanodisc example and 7 for the micelle example) to  $N_g \simeq 0$  for  $\alpha = 10^{10}$ . Large  $\alpha$  values give weight to the prior, resulting in a low estimated information content of the new data set.

After having introduced  $N_g$ , it is worth returning to the Occam term from equation (5). This term pushes the algo-

rithm towards solutions with higher  $\alpha$  values and closer to the prior parameter values (Fig. 1). Higher  $\alpha$  values also imply a smaller  $N_g$  (Fig. 8a), that is, fewer parameters can be retrieved from the data. Hence, the Occam term favors simpler solutions with fewer effective parameters.

## 5. Discussion

In SAS data analysis with analytical form factors, the prior knowledge can be included *via* molecular constraints as implemented in the parametrization of the hypothesized model. The remaining model parameters are then, in principle, free and can take any value. In practice, however, many parameter values cannot be accepted, owing to inconsistency with the prior knowledge about these parameters, for example from other experiments. This is often accounted for by fixing certain parameters or by setting up limits for the parameter values, *i.e.* not allowing the parameters to exceed a certain range. This is implemented in several commonly used programs for SAS data analysis with analytical form factors, for example *SasView* (<http://www.sasview.org>), *SASfit* (Breßler *et al.*, 2015), *Scatter* (Förster *et al.*, 2010) and *WillItFit* (Pedersen *et al.*, 2013). It can be argued that this practice corresponds to a Bayesian approach using uniform priors with a finite probability in a given interval and zero probability outside this interval. In the present paper we improve this conventional method by allowing for normally distributed priors that better represent the prior knowledge than uniform priors.

The Bayesian approach is similar to other optimization methods using regularized expressions, but the regularization parameter is here determined automatically and in a statistically sound way, such that a subjective choice of  $\alpha$  is avoided.

In a wider perspective, the presented method is a solution to a multi-objective problem (for details see *e.g.* Miettinen, 1998). The objectives are here quantified in terms of the likelihood and the prior functions ( $\chi^2$  and  $S$ ), and the wanted solution is a set of model parameters. The objective functions may be minimized by different sets of model parameters, and the goal is to find the most probable solution taking into account both functions. The  $\chi^2$  versus  $S$  solution space can be divided into two regions, as shown for the nanodisc example in Fig. 9. One region is unreachable since no set of parameters results in these combinations of  $\chi^2$  and  $S$  values. The other region is reachable, but most solutions here are non-optimal since there exists another set of parameters which is superior with respect to one of the objective functions without being inferior with respect to the others. The border between the regions is denoted the Pareto frontier (Miettinen, 1998). It contains all sets of model parameters that constitute an optimal solution for a given weight between the two objective functions (Pareto optimal sets). A scan over  $\alpha$  corresponds to a walk along the Pareto frontier, as indicated in Fig. 9. At  $\alpha = 0$ ,  $\chi^2$  is minimized and  $S$  takes a relatively high value. As  $\alpha$  increases,  $S$  converges towards 0 and  $\chi^2$  towards the  $\chi^2$  value for the prior solution. Intriguingly, the Pareto frontier is convex for the nanodisc example, meaning that a small

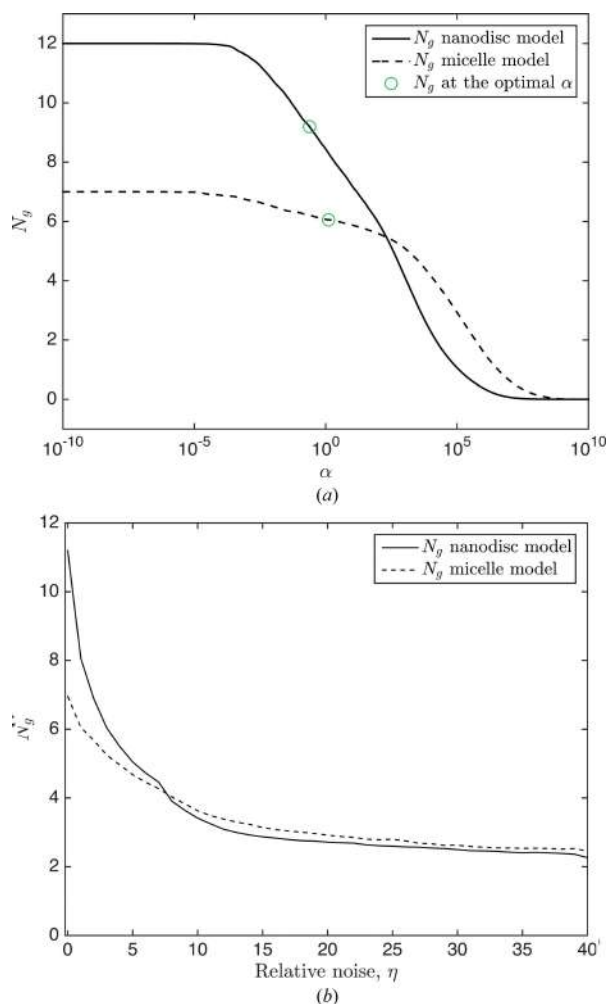


Figure 8  
(a)  $N_g$  as a function of  $\alpha$ , with the value for the optimal value of  $\alpha$  marked in green. (b)  $N_g$  for varying noise levels. Each point was a mean for a small range of subsequent values of  $\alpha$  (a) or  $\eta$  (b).

perturbation of  $\chi^2$  allows a large improvement of  $S$ , and *vice versa*. The Pareto frontier for the micelle example is almost single valued, since the same set of parameters minimizes both  $\chi^2$  and  $S$ . The present method is a so-called scalarization, transforming the multi-objective problem into a single-objective problem with only one solution, namely that for the most probable  $\alpha$ .

We have chosen to use Gaussian priors for all parameters, despite the fact that non-Gaussian priors may better represent the knowledge about some of the model parameters. Gaussian priors are, however, computationally economical and simpler to comprehend. The computational speed is relevant, because the Bayesian algorithm needs to refine the model for several values of  $\alpha$  to find the most probable solution, thus being 10–20 times slower than conventional  $\chi^2$  minimization (depending on the effectiveness of the  $\alpha$ -optimization algorithm). For a complex model with two (or more) numerical integrals, such as the nanodisc model, the CPU time can thus extend to 20 min on a standard PC (single core). Considerable speedup can, however, be obtained by parallelization in  $q$ .

An inherent problem of the presented method is that it relies on the principle that priors and experimental errors are correctly estimated. Priors may be wrongly estimated, for example because of an erroneous concentration measurement, or errors on refined parameters from previous experiments may be underestimated. A prior for a certain parameter can either be too wide, be too narrow or have a wrong mean value. If the prior is too wide, its effect on the refined value will be underestimated and the errors overestimated. If, on the other hand, a prior is too narrow, it will over-restrict the refined parameter, and the refined error will be underestimated. In the case of a wrong prior mean value, the data will pull the solution far away from this value. Large deviations are thus

apparent when comparing the prior with the refined result, so the method constitutes an evaluation of prior assumptions. Generally, a wrongly estimated prior for a given parameter will affect the solution the most if the new data contain relatively little information about that parameter, but will only have a minor effect if the parameter is well determined by the new data. Wrongly estimated priors should, of course, be avoided since inaccurate input will inevitably lead to inaccurate output.

The errors on SAS data may likewise be wrongly estimated, as discussed for example by Franke *et al.* (2015) and Rambo & Tainer (2013). In the nanodisc example the fit is good, as judged by visual inspection. However, the residuals (Fig. 3a) are expected to be within  $\pm 3\sigma$  for a good fit, but in this case reach up to  $\pm 10\sigma$ . In the same way,  $\chi_r^2$  is expected to be in the range [0.67, 1.43] (95% confidence interval), but a value of 6.26 was obtained. The size of the experimental errors can be evaluated by indirect Fourier transformation, since data are here fitted with a generic function that should result in a  $\chi_r^2$  value close to unity. However, a  $\chi_r^2$  value of 6.6 was obtained in the Bayesian indirect Fourier transformation, thereby indicating that the experimental errors are underestimated. With the Cmap test, the fit could be evaluated independently of the experimental errors. The Cmap test confirmed that the similarity of model and data could not be rejected [significance level of 1%,  $C = 10$ ,  $P(C \geq 10 | N = 106) = 9.2\%$ ] and hence confirmed that the experimentally determined error bars were underestimated.

Underestimation of the experimental errors will give too much weight to data (and too little to the prior), since the weight given to data is inversely proportional to the square of the experimental errors [equation (1)]. For a data set with severely over- or underestimated errors, an error correction could therefore be included either separately before the analysis or as an implicit part of the analysis to avoid the effect of erroneously determined experimental errors. We have not included that in the present work because we believe it deserves a more thorough discussion, and it is not a question related specifically to the Bayesian method presented here but affects all methods based on  $\chi^2$ .

The stabilization of the refined solution upon addition of noise, as exemplified in Fig. 7, shows that the Bayesian regularization method is especially relevant for data with a low signal-to-noise ratio: that is, when sample concentration is limited, for example for protein samples with low-yield expression and samples that are only stable at low concentrations, when exposure time is limited, for example in time-resolved studies, or when flux is limited, for example in SANS and in SAXS at home-source instruments.

The number of degrees of freedom in a SAS data set with  $q$  range  $q_{\max} - q_{\min}$  and maximum intraparticle distance  $D_{\max}$  has been described in terms of the number of Shannon channels (Shannon, 1949; Moore, 1980) as  $N_S = D_{\max}(q_{\max} - q_{\min})/\pi$ , provided that  $q_{\min} < \pi/D_{\max}$ .  $N_S$  is widely used to assess the information content in data (e.g. Grant *et al.*, 2015). As a measure for the information content, however,  $N_S$  has the obvious shortcoming that it does not take into account the

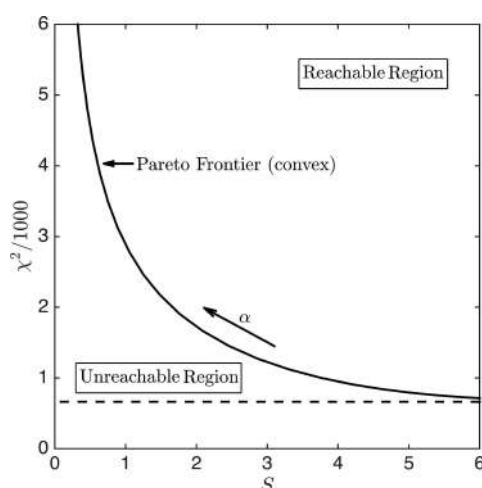


Figure 9

The  $\chi^2$  versus  $S$  space for the nanodisc example. The Pareto frontier (black line) separates the unreachable region and the reachable region. The minimum  $\chi^2$  value (dashed line;  $\alpha = 0$ ) and the direction of increasing  $\alpha$  are shown. The most probable solution was found at  $\alpha = 0.24$ ,  $S = 14$  and  $\chi^2 = 668$  (point not included).

## research papers

noise level of data. A solution was proposed by Konarev & Svergun (2015), who introduced an effective number of Shannon channels  $M_s$  by truncation of data at high  $q$  values with poor signal-to-noise ratio, thus taking into account the noise level of data.

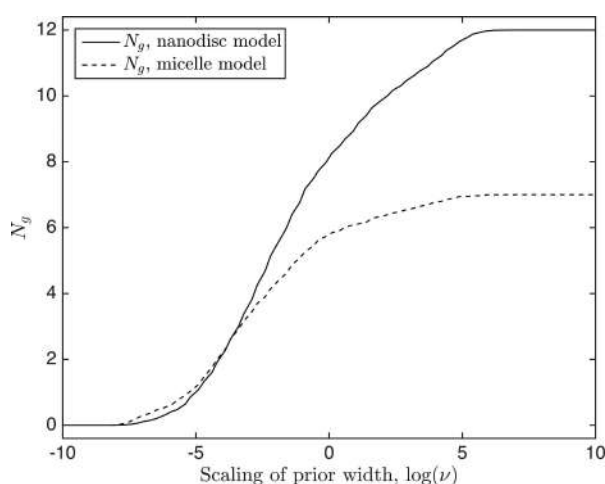
As shown here, and by Pedersen *et al.* (2014), the noise is also effectively taken into account by  $N_g$ . Moreover,  $N_g$  takes into account the included prior knowledge. Pedersen *et al.* (2014) and Vestergaard & Hansen (2006) used a generic prior, namely that  $p(r)$  is a smooth function. In fact, this is the same general information used to estimate  $M_s$ . We will in the following denote the number of good parameters obtained with the smoothness constraint by  $N_g^S$  (not to be confused with  $N_s$ ).  $N_g^S$  can be calculated with the indirect Fourier transform algorithm in *BayesApp* (<http://www.bayesapp.org>; Hansen, 2012). The  $N_g$  introduced in the present paper uses Gaussian priors for each parameter and will therefore be denoted  $N_g^G$ . For the micelle data set  $N_g^S = 8.8$  and  $N_g^G = 6.0$ , and for the nanodisc data set  $N_g^S = 7.3$  and  $N_g^G = 9.1$ : that is, the estimated information content varies with the prior. In the same way, if the Gaussian prior is altered, then  $N_g^G$  will change accordingly. To show this, the priors (Tables 1 and 2) were altered by rescaling the prior width with a scale factor  $\nu$ , *i.e.*  $\delta p \rightarrow \nu \delta p$ , corresponding to a change in the certainty about the priors.  $N_g^G$  increases asymptotically as the prior width increases (Fig. 10), *i.e.* when the *a priori* certainty about the parameters decreases. The dependence on prior knowledge is especially evident for repetition series. Here, the first measurement has a relatively high information content, but since that measurement will be included in the updated prior knowledge, the second measurement will contain less information, the third repetition even less, *etc.* At some point, no more measurements need to be taken, since the information content of succeeding measurements would effectively be zero. The prior knowledge has no effect on  $N_s$ , which is nevertheless widely used as a measure for the information in

data. Therefore, we propose to use  $N_g^S$  or  $M_s$  instead of  $N_s$  to assess the information content in a single SAS data set or a repetition series prior to modeling. After modeling,  $N_g^G$  can be used to evaluate the information obtained when SAS is combined with other experimental results and/or other available prior knowledge, as shown in the two examples.

## 6. Conclusion

A Bayesian regularization method for SAS data analysis was developed and tested on two data sets: a sample of nanodiscs described by a model with 12 parameters and a sample of detergent micelles described by a model with seven parameters. In both cases, the Bayesian regularization method found a set of model parameters that were physically meaningful without compromising the goodness of fit. The regularization method, furthermore, stabilized the solution when tested against simulated data with increasing noise, thereby preventing overfitting of random noise. This had the important advantage that information could be retrieved even from very noisy data. The method is founded upon probability theory and provides an automatic procedure for weighing the likelihood function  $\chi^2$  and the prior function  $S$  with respect to each other, by optimizing the regularization parameter  $\alpha$ . Moreover, the Bayesian method provides a measure for the information content in data, the number of good parameters  $N_g$ , which takes into account both the noise level of the data and the prior knowledge about each model parameter.

Bayesian regularization is generally applicable to inverse problems and is indeed widely applied in many other fields, as mentioned in §1. But, owing to the relatively low information content in SAS data combined with the use of models with multiple parameters, the Bayesian regularization method is of clear relevance for this field.



**Figure 10**  
 $N_g$  versus prior width. The prior width,  $\delta \mathbf{p} = (\delta p_1, \dots, \delta p_K)$ , where  $K$  is the number of model parameters, was scaled by  $\delta \mathbf{p} \rightarrow \nu \delta \mathbf{p}$ . Each point was a mean for five subsequent values of  $\log(\nu)$ .

## Acknowledgements

The authors would like to thank Per Hedegård and Martin Cramer Pedersen for fruitful discussions. Thanks to Pie Huda, Søren Roi Midtgaard and Nicholas Skar-Gislinge for providing the example data.

## Funding information

Thanks to CoNeXT and University of Copenhagen for co-funding the project.

## References

- Arleth, L., Posselt, D., Gazeau, D., Larpent, C., Zemb, T., Mortensen, K. & Pedersen, J. S. (1997). *Langmuir*, **13**, 1887–1896.
- Bergstra, J. & Bengio, Y. (2012). *J. Mach. Learn. Res.* **13**, 281–305.
- Bolstad, W. M. (2007). *Introduction to Bayesian Statistics*, pp. 121–330. Hoboken: John Wiley and Son.
- Breßler, I., Kohlbrecher, J. & Thünemann, A. F. (2015). *J. Appl. Cryst.* **48**, 1587–1598.
- Cabane, B., Duplessix, R. & Zemb, T. (1985). *J. Phys. Fr.* **46**, 2161–2178.



## research papers

- Fitzkee, N. C. & Rose, G. D. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 12497–12502.
- Förster, S., Apostol, L. & Bras, W. (2010). *J. Appl. Cryst.* **43**, 639–646.
- Franke, D., Jeffries, C. M. & Svergun, D. I. (2015). *Nat. Methods*, **12**, 419–422.
- Glatter, O. (1977). *J. Appl. Cryst.* **10**, 415–421.
- Grant, T. D., Luft, J. R., Carter, L. G., Matsui, T., Weiss, T. M., Martel, A. & Snell, E. H. (2015). *Acta Cryst.* **D71**, 45–56.
- Gull, S. F. (1989). *Maximum Entropy and Bayesian Methods*, edited by J. Skilling, pp. 53–71. Dordrecht: Springer.
- Hansen, S. (2000). *J. Appl. Cryst.* **33**, 1415–1421.
- Hansen, S. (2012). *J. Appl. Cryst.* **45**, 566–567.
- Hayter, J. B. & Penfold, J. (1981). *J. Chem. Soc. Faraday Trans. 1*, **77**, 1851–1863.
- Jeffreys, H. (1946). *Proc. R. Soc. London Ser. A*, **186**, 453–461.
- Konarev, P. V. & Svergun, D. I. (2015). *IUCrJ*, **2**, 352–360.
- Kučerka, N., Liu, Y., Chu, N., Petrache, H. I., Tristram-Nagle, S. & Nagle, J. F. (2005). *Biophys. J.* **88**, 2626–2637.
- Kučerka, N., Nagle, J. F., Feller, S. E. & Balgavý, P. (2004). *Phys. Rev. E*, **69**, 051903.
- Kynde, S. A. R., Skar-Gislinge, N., Pedersen, M. C., Midtgaard, S. R., Simonsen, J. B., Schweins, R., Mortensen, K. & Arleth, L. (2014). *Acta Cryst.* **D70**, 371–383.
- Levenberg, K. (1944). *Q. Appl. Math.* **2**, 164–168.
- MacKay, D. J. C. (1992). *Adv. Neural Inf. Process. Syst.* **4**, 839–846.
- MacKay, D. J. C. (1999). *Neural Comput.* **11**, 1035–1068.
- Marquardt, D. W. (1963). *J. Soc. Ind. Appl. Math.* **11**, 431–441.
- Midtgaard, S. R. *et al.* (2018). *FEBS J.* **285**, 357–371.
- Midtgaard, S. R., Pedersen, M. C. & Arleth, L. (2015). *Biophys. J.* **109**, 308–318.
- Miettinen, K. (1998). *Nonlinear Multiobjective Optimization*. Boston: Kluwer Academic Publishers.
- Moore, P. B. (1980). *J. Appl. Cryst.* **13**, 168–175.
- Müller, J. J., Hansen, S. & Pürschel, H.-V. (1996). *J. Appl. Cryst.* **29**, 547–554.
- Oliver, R. C., Lipfert, J., Fox, D. A., Lo, R. H., Doniach, S. & Columbus, L. (2013). *PLoS One*, **8**, e62488.
- Pedersen, J. S. (1997). *Adv. Colloid Interface Sci.* **70**, 171–210.
- Pedersen, M. C., Arleth, L. & Mortensen, K. (2013). *J. Appl. Cryst.* **46**, 1894–1898.
- Pedersen, M. C., Hansen, S. L., Markussen, B., Arleth, L. & Mortensen, K. (2014). *J. Appl. Cryst.* **47**, 2000–2010.
- Petoukhov, M. V. & Svergun, D. I. (2005). *Biophys. J.* **89**, 1237–1250.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes*, pp. 4–93. Cambridge University Press.
- Rambo, R. & Tainer, J. A. (2013). *Nature*, **496**, 477–481.
- Scheres, S. H. W. (2012). *J. Mol. Biol.* **415**, 406–418.
- Schultz, R. R. & Stevenson, R. L. (1994). *IEEE Trans. Image Process.* **3**, 233–242.
- Shannon, C. E. (1949). *Proc. IRE*, **37**, 10–21.
- Shevchuk, R. & Hub, J. S. (2017). *PLOS Comput. Biol.* **13**, e1005800.
- Shih, A. Y., Freddolino, P. L., Sligar, S. G. & Schulten, K. (2007). *Nano Lett.* **7**, 1692–1696.
- Skar-Gislinge, N. & Arleth, L. (2011). *Phys. Chem. Chem. Phys.* **13**, 3161–3170.
- Skar-Gislinge, N., Simonsen, J. B., Mortensen, K., Feidenhans'l, R., Sligar, S. G., Lindberg Møller, B., Bjørnholm, T. & Arleth, L. (2010). *J. Am. Chem. Soc.* **132**, 13713–13722.
- Svergun, D. I. (1992). *J. Appl. Cryst.* **25**, 495–503.
- Svergun, D. I. (1999). *Biophys. J.* **76**, 2879–2886.
- Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.
- Svergun, D. I. & Koch, M. H. J. (2003). *Rep. Prog. Phys.* **66**, 1735–1782.
- Svergun, D. I., Koch, M. H. J., Timmins, P. A. & May, R. P. (2013). *Small Angle X-ray and Neutron Scattering from Solutions of Biological Macromolecules*. Oxford University Press.
- Tanford, C. (1972). *J. Phys. Chem.* **76**, 3020–3024.
- Vestergaard, B. & Hansen, S. (2006). *J. Appl. Cryst.* **39**, 797–804.

### 9.3 Paper III: Invisible detergents for structure determination of membrane proteins by small-angle neutron scattering

**List of Authors** Søren Roi Midtgaard, Tamim A. Darwish, Martin Cramer Pedersen, Pie Huda, Andreas Haahr Larsen, Grethe Vestergaard Jensen, Søren Andreas Røssell Kynde, Nicholas Skar-Gislinge, Agnieszka Janina Zygałło Nielsen, Claus Olesen, Mickael Blaise, Jerzy Jozef Dorosz, Thor Seneca Thorsen, Raminta Venskutonyte, Christian Krintel, Jesper V. Møller, Henrich Frielinghaus, Elliot Paul Gilbert, Anne Martel, Jette Sandholm Kastrup, Poul Erik Jensen, Poul Nissen and Lise Arleth


**Status** FEBS, 2018, 285, 357-371.

**Abstract** A novel and generally applicable method for determining structures of membrane proteins in solution via small-angle neutron scattering (SANS) is presented. Common detergents for solubilizing membrane proteins were synthesized in isotope-substituted versions for utilizing the intrinsic neutron scattering length difference between hydrogen and deuterium. Individual hydrogen/deuterium levels of the detergent head and tail groups were achieved such that the formed micelles became effectively invisible in heavy water (D<sub>2</sub>O) when investigated by neutrons. This way, only the signal from the membrane protein remained in the SANS data. We demonstrate that the method is not only generally applicable on five very different membrane proteins but also reveals subtle structural details about the sarco/endoplasmatic reticulum Ca<sup>2+</sup> ATPase (SERCA). In all, the synthesis of isotope-substituted detergents makes solution structure determination of membrane proteins by SANS and subsequent data analysis available to nonspecialists.

**Contributions by AHL** AHL participated in the SANS data collection. AHL did the SANS analysis for GluA2 and SERCA and participated in streamlining the SANS analysis for all data. AHL wrote a piece of software (CaPP, see chapter 3) that was used for the analysis of several of the proteins. AHL wrote the part of supplemental related GluA2 and SERCA.

**Supporting Information** The part of the SI related to SANS is attached here. The full SI (40 pages) can be downloaded at: <https://febs.onlinelibrary.wiley.com/doi/abs/10.1111/febs.14345>

## Invisible detergents for structure determination of membrane proteins by small-angle neutron scattering

Søren Roi Midtgaard<sup>1</sup> , Tamim A. Darwish<sup>2</sup>, Martin Cramer Pedersen<sup>1,3</sup>, Pie Huda<sup>1</sup>, Andreas Haahr Larsen<sup>1</sup>, Grethe Vestergaard Jensen<sup>1</sup>, Søren Andreas Røssell Kynde<sup>1</sup>, Nicholas Skar-Gislinge<sup>1</sup>, Agnieszka Janina Zygodlo Nielsen<sup>4</sup>, Claus Olesen<sup>5</sup>, Mickael Blaise<sup>6,7</sup>, Jerzy Józef Dorosz<sup>8</sup>, Thor Seneca Thorsen<sup>8</sup>, Raminta Venskutonytė<sup>8</sup>, Christian Krintel<sup>8</sup>, Jesper V. Møller<sup>9,10</sup>, Henrich Frielinghaus<sup>11</sup>, Elliot Paul Gilbert<sup>12</sup>, Anne Martel<sup>13</sup>, Jette Sandholm Kastrup<sup>8</sup>, Poul Erik Jensen<sup>4</sup>, Poul Nissen<sup>10,14</sup> and Lise Arleth<sup>1</sup>

1 Structural Biophysics, X-ray and Neutron Science, The Niels Bohr Institute, University of Copenhagen, Denmark

2 National Deuteration Facility, Australian Nuclear Science and Technology Organization, Lucas Heights, Australia

3 Department of Applied Mathematics, Research School of Physics and Engineering, Australian National University, Canberra, Australia

4 Copenhagen Plant Science Center, University of Copenhagen, Denmark

5 Department of Biomedicine, Aarhus University, Denmark

6 Institut de Recherche en Infectiologie de Montpellier, CNRS, Université de Montpellier, France

7 Centre for Carbohydrate Recognition and Signaling, Department of Molecular Biology, Aarhus University, Denmark

8 Department of Drug Design and Pharmacology, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

9 Department of Biomedicine, Aarhus University, Denmark

10 Department of Molecular Biology and Genetics, Centre for Membrane Pumps in Cells and Disease – PUMPKin, Danish National Research Foundation, Aarhus University, Denmark

11 Forschungszentrum Jülich GmbH, TUM FRM-2, Garching, Germany

12 Australian Centre for Neutron Scattering, Australian Nuclear Science and Technology Organization, Lucas Heights, Australia

13 Institut Laue-Langevin, Grenoble, France

14 DANDRITE, Nordic-EMBL Partnership for Molecular Medicine, Aarhus University, Denmark

### Keywords

contrast matching; deuteration; membrane proteins; SANS; Small-angle neutron scattering

### Correspondence

S. R. Midtgaard and L. Arleth, Structural Biophysics, X-Ray and Neutron Science, Niels Bohr Institute, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen, Denmark

Emails: soromi@nbi.ku.dk (S.R.M.); arleth@nbi.ku.dk (L.A.)

and

T. Darwish, Australian Nuclear Science and Technology Organisation, Locked Bag 2001, Kirrawee DC, NSW 2232, Australia

Email: tde@ansto.gov.au

and

P. Nissen, Aarhus University, Gustav Wieds Vej 10, 8000, Aarhus C, Denmark

Email: pn@mbg.au.dk

(Received 10 July 2017, revised 20 October 2017, accepted 21 November 2017)

doi:10.1111/febs.14345

A novel and generally applicable method for determining structures of membrane proteins in solution via small-angle neutron scattering (SANS) is presented. Common detergents for solubilizing membrane proteins were synthesized in isotope-substituted versions for utilizing the intrinsic neutron scattering length difference between hydrogen and deuterium. Individual hydrogen/deuterium levels of the detergent head and tail groups were achieved such that the formed micelles became effectively invisible in heavy water (D<sub>2</sub>O) when investigated by neutrons. This way, only the signal from the membrane protein remained in the SANS data. We demonstrate that the method is not only generally applicable on five very different membrane proteins but also reveals subtle structural details about the sarco/endoplasmic reticulum Ca<sup>2+</sup> ATPase (SERCA). In all, the synthesis of isotope-substituted detergents makes solution structure determination of membrane proteins by SANS and subsequent data analysis available to nonspecialists.

## Introduction

Biological membranes form the barrier between internal and external environments of cells and create the separation of compartments and organelles within the cell. They are functionalized by a plethora of membrane proteins that include structural proteins, receptors, channels, active transporters and proteolytic enzymes. Studies of the structure and conformational changes related to the function of membrane proteins therefore represent very important problems in molecular cell biology. Structural studies usually require that the membrane protein of interest is removed from the native membrane, isolated and purified to homogeneity. This requires that the amphipathic membrane protein is stabilized in solution. Multiple systems, including traditional detergents [1], bicelles [2,3], peptide discs [4], amphiphilic polymers [5,6] and nanodiscs [7] are available for this purpose. While detergents and bicelles have been useful for Cryo-EM [8], NMR [9] and crystallization experiments [1], the comparatively monodisperse nanodisc particles are promising in combination with small-angle scattering and allows low-resolution structural information of the membrane protein to be obtained under solution conditions [10–12]. However, reconstitution of membrane proteins in nanodiscs remains a challenge with respect to obtaining sufficiently homogeneous samples for scattering experiments. In our attempt to find a more general and easily applicable solution to these problems, the use of detergents was re-examined and combined with our previously developed idea of a “stealth”-carrier for SANS studies of membrane proteins [10]. The detergents, octyl  $\beta$ -D-glucopyranoside (OG) and n-dodecyl- $\beta$ -D-maltopyranoside (DDM) that are both compatible with membrane protein reconstitution were deuterated through a new synthesis method that allowed for selective deuteration of the two detergents to different predetermined levels at the hydrophilic head groups and hydrophobic tails. To obtain the best possible SANS signal-to-noise of the membrane protein structure, the different parts of the detergents were deuterated such that they had the same neutron scattering length density as 100% deuterium oxide ( $D_2O$ ). This way, the obtained SANS data of membrane proteins stabilized in these detergents in  $D_2O$  only consists of the single contrast signal stemming from the purified membrane protein.

## Abbreviations

bR, bacteriorhodopsin; Cryo-EM, Cryo Electron Microscopy; DDM, n-dodecyl- $\beta$ -D-maltopyranoside;  $D_{max}$ , maximum distance; FWHM, full width half maximum; GluA2, ionotropic glutamate receptor A2; kDa, kilo dalton; LamB, maltoporin; MW, molecular weight; OG, octyl  $\beta$ -D-glucopyranoside; PSI, photosystem I;  $R_g$ , radius of gyration; SANS, small-angle neutron scattering; SAXS, small-angle X-ray scattering; SEC, size exclusion chromatography; SERCA, sarco/endoplasmic reticulum  $Ca^{2+}$  ATPase; SI, supplemental information.

The general idea to use the contrast match point of a detergent or surfactant to study membrane protein structures has previously been investigated [13–29]. However, a common feature for all previous studies is that they rely on commercially available hydrogenated detergents or fully deuterated detergents developed for different purposes. This is reflected in the resulting neutron scattering data, which include scattering cross-terms from the entire detergent–membrane protein complex due to the differences in excess scattering length density of the hydrophobic and hydrophilic parts of the detergents present on length scales of 10–20 Å. This effect is difficult to disentangle from the signal from the membrane protein and significantly limits the resolution that can be obtained. In addition, the overall match point of the commercially available detergents is generally different from that of 100%  $D_2O$ , hence yielding additional incoherent scattering background from the hydrogen atoms in the buffer. While the studies have provided overall shape parameters of studied membrane proteins [14,20,21,24,27,28], their finer structure could not be extracted from the data due to these factors.

In this study, two commonly used detergents, OG and DDM, were synthesized with varying deuteration levels in head- and tail groups such that they were fully match-out in  $D_2O$  buffer when investigated by SANS. This yielded SANS data where only the membrane proteins were visible. The practical workflow is outlined in Fig. 1.

Initially, the membrane protein is purified in hydrogenated buffer and hydrogenated detergent. This step ensures that as many impurities as possible are removed before the more expensive deuterated buffers are introduced. Afterwards, a size exclusion purification in the equivalent deuterated buffer and match-out deuterated detergent is performed. This step ensures that the hydrogenated detergent and buffer are fully exchanged to the deuterated counterparts (see SI 3 for details). Furthermore, any aggregated protein that might be present in the sample arising from, for example, freezing and transport, is also removed to ensure optimal data quality. The peak fraction from the buffer and detergent exchange is collected and, without further manipulation, measured by SANS. Scattering data obtained this way can be reduced and

subsequently analysed by a number of available programs for small-angle scattering (such as ATSAS [30], WillItFit [31], SASSIE [32], IRENA [33], FoXS [34] etc.) due to the simplified contrast situation achieved through the use of *de facto* invisible detergents.

## Results

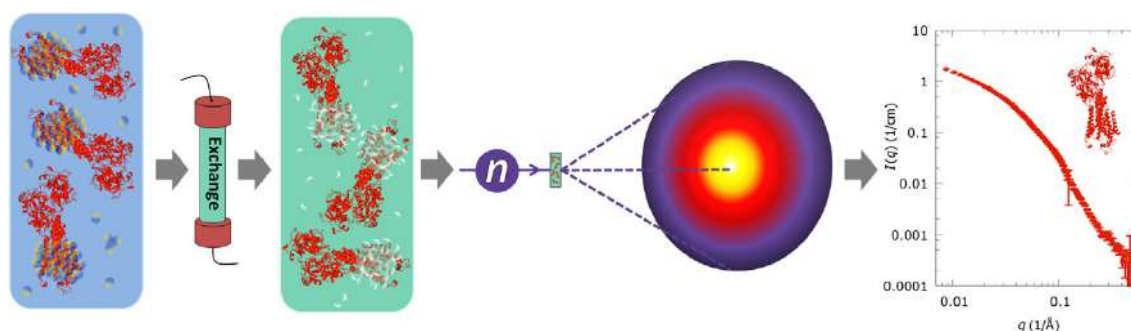
### Detergent deuteration

As described above, the chosen deuterated detergents were two commonly used sugar-based detergents, namely OG and DDM. The partial specific molecular volumes of the hydrophilic head and hydrophobic tail groups were determined by densitometry (see Supporting Information 1 for details) such that the deuteration level for complete match-out at 100% D<sub>2</sub>O at 20 °C could be calculated (see Table 1). Custom syntheses were then developed to produce the desired partially deuterated versions of both detergents.

### Synthesis of detergents with controlled deuteration levels in hydrophilic head groups and hydrophobic tails

Each of OG and DDM is made of two parts; the C8 or C12 alkyl chain tail and the corresponding sugar head of glucose or maltose respectively. In the case of OG, the required deuteration levels in the octyl chain and the glucose sugar head (seven nonexchangeable positions) groups were 94% D and 52% D respectively. In DDM, the required deuteration levels in the dodecyl chain and the maltose sugar head (14 nonexchangeable positions) groups were 89% D and 57% D respectively. Deuterating the detergent molecules

directly using hydrothermal reactions was not possible as the harsh conditions of high temperature and pressure were incompatible with the delicate nature of the molecules. The deuteration of methylene units in alkyl hydrocarbons can only be achieved using forced conditions of a hydrothermal catalytic exchange reaction [35], while the deuteration of the sugar groups can be achieved using milder reaction conditions [36]. An alternative approach of deuterating the corresponding sugar molecule and the alkyl chains separately before attaching them together would not be feasible since the sugar groups, glucose and maltose, are reducing sugars, which means they would quickly decompose in the presence of reducing reagents due to the hydrolysis of their hemiacetal moieties. Reducing sugars must be preceded by conversion into glycoside to prevent reduction at the anomeric carbon when exposed to Raney Nickel [36]. Therefore, to achieve specific deuteration levels in the two different parts of any of the two detergent molecules, it was necessary to follow a multiple-step synthesis approach and two sequential deuteration steps. This involved (illustrated in Fig. 2 for OG): **A**) deuterating the alkyl chain at the required deuteration level starting from the corresponding fatty acid and using hydrothermal Pt/C catalysed H/D exchange reactions at 220 °C in the appropriate molar ratio of deuterium and hydrogen atoms in the mixture, reducing the fatty acid molecule to the corresponding alcohol; **B**) attaching the deuterated alcohol to the corresponding acetylated bromo-sugar head group (i.e. 2,3,4,6-tetra-O-acetyl- $\alpha$ -D-glucopyranosyl bromide and 2,3,6,2',3',4',6'-hepta-O-acetyl- $\alpha$ -D-maltosyl bromide) according to standard procedures deacetylation of the sugar head group; **C**) deuterating the sugar head to achieve the required deuteration level by using mild



**Fig. 1.** Process outline of the experiment. Initially, the membrane protein is in H<sub>2</sub>O-based buffer and hydrogenated detergents. This sample is then applied to a size exclusion column equilibrated in D<sub>2</sub>O-based buffer and the match-out deuterated detergents, allowing the hydrogenated buffer and detergent to be exchanged to their deuterated counterparts. The obtained sample is used directly for SANS measurements to obtain a single contrast dataset yielding only the scattering from the membrane protein. Data from the experiment can then be analysed by generally available software developed for determining the low-resolution structure of proteins in solution.

**Table 1.** Deuteration levels needed and obtained for *n*-octyl  $\beta$ -D-glucopyranoside (OG) and *n*-dodecyl- $\beta$ -D-maltopyranoside (DDM).

	OG		DDM	
	Head group	Tail group	Head group	Tail group
Chemical composition of the detergent component	C <sub>2</sub> H <sub>11</sub> O <sub>6</sub>	C <sub>8</sub> H <sub>17</sub>	C <sub>12</sub> H <sub>21</sub> O <sub>11</sub>	C <sub>12</sub> H <sub>25</sub>
Exchangeable hydrogens	4	0	7	0
Theoretical level of deuteration needed for match-out at 100% D <sub>2</sub> O	C <sub>6</sub> D <sub>7.6</sub> H <sub>3.4</sub> O <sub>6</sub>	C <sub>8</sub> D <sub>15.9</sub> H <sub>1.1</sub>	C <sub>12</sub> D <sub>15.2</sub> H <sub>5.8</sub> O <sub>11</sub>	C <sub>12</sub> D <sub>22.4</sub> H <sub>2.6</sub>
Experimentally obtained level of deuteration in 100% D <sub>2</sub> O	C <sub>6</sub> D <sub>7.64</sub> H <sub>3.36</sub> O <sub>6</sub>	C <sub>8</sub> D <sub>15.98</sub> H <sub>1.12</sub>	C <sub>12</sub> D <sub>14.98</sub> H <sub>6.02</sub> O <sub>11</sub>	C <sub>12</sub> D <sub>22.25</sub> H <sub>2.75</sub>

conditions of Raney Nickel as a catalyst in D<sub>2</sub>O/H<sub>2</sub>O mixture at 80 °C for 18 h. The latter step allows the incorporation of deuterium atoms on carbons adjacent to free hydroxyl groups ( $\alpha$  positions) in the sugar head group with retention of configuration, but it does not affect any H/D back-exchange at the more inert alkyl chain sites [36–38] (see Supporting Information 2 for details).

### SANS contrast variation on empty micelles

The deuteration levels obtained from the custom synthesis were close to the desired values (Table 1), implying that the scattering length density from the detergents should theoretically be close to that of 100% D<sub>2</sub>O. To experimentally verify this, micelles of the match-out deuterated detergents were measured by SANS in a set of buffers with D<sub>2</sub>O content ranging from 60% to 100%. Experiments were performed at ANSTO and FRMII (See SI 4). Background subtracted data are plotted in Fig. 3.

The SANS intensity from both detergents is reduced by more than two orders of magnitude when changing the percentage of D<sub>2</sub>O in the buffer from 60% to 100%. Indeed, at 100% D<sub>2</sub>O, the signal from the match-out deuterated detergents is effectively at the noise-level over the measured  $q$ -range and exhibiting no significant  $q$ -dependence.  $I(0)$  values (Fig. 3C) were determined by indirect Fourier transformation [39]. The expected parabolic behaviour of the  $I(0)$  as a function of D<sub>2</sub>O contents was confirmed (Fig. 3C) and the match points were found to be at 102% D<sub>2</sub>O for DDM and 103% D<sub>2</sub>O for OG.

In Figures 3D and E, the obtained contrast-matched data are compared to previously obtained and published SANS data from the commercially available tail-deuterated counterparts to OG and DDM: d17-OG and d25-DDM [28]. These data, (obtained and plotted with permission from the authors of ref. [28]), are measured at the D22 instrument at ILL, France and plotted at their respective match points which is 90% D<sub>2</sub>O for d17-OG and 85% for d-25-DDM [28].

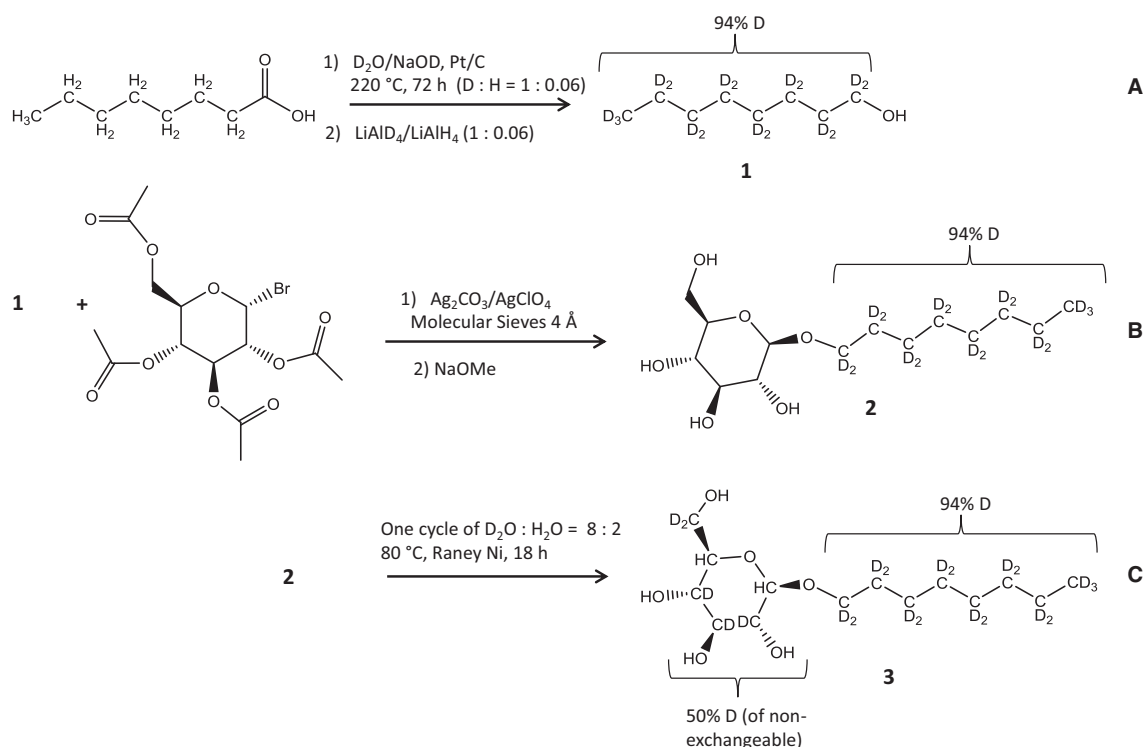
The data from the tail-deuterated detergent d17-OG exhibit the expected  $q$ -dependence with an oscillation having a maximum at around  $0.15 \text{ \AA}^{-1}$ , but overall a very low intensity of the  $I(q)$  at the match-point. Note however, that the d17-OG data had a small experimental discrepancy between the background levels of the measured buffer and sample. Due to the low signal to background ratio, this had a relatively large impact on the obtained scattering intensity as illustrated in Fig 3D which plots both the d17-OG data with background subtracted using the theoretical normalization (light grey curve) and the same data with a small renormalization of the background (dark grey curve). As it is seen, this causes a relatively large shift in the overall scattering intensity, which unfortunately makes a direct quantitative comparison to our data difficult (Fig. 3D). In any case, the density fluctuations expected in the d17-OG are systematic and clearly visible in the  $I(q)$ -data regardless of the buffer level, while there is no indication of this feature in the data from the match-out deuterated OG.

In the comparison between the d25-DDM at 85% D<sub>2</sub>O and the match-out d-DDM at 100% D<sub>2</sub>O (Fig. 3E) the above-mentioned experimental discrepancy between the buffer and sample background levels is not present to the same extent. Furthermore, a much stronger difference between the custom synthesized and commercial variants is observed which clearly demonstrates the added value of the custom synthesis of these match-out detergents. Indeed, the scattering intensity of the d25-DDM is 5-6 times as strong around  $0.1 \text{ \AA}^{-1}$  as compared to match-out d-DDM and its  $q$ -dependence is, as expected, very significant due to the internal core-shell contrast.

Figure 3F shows the simulated scattering intensity of bR reconstituted in, respectively, a match-out deuterated d-DDM micelle in 100% D<sub>2</sub>O and a d25-DDM micelle at 85% D<sub>2</sub>O. Note that, for d25-DDM, the overall scattering intensity is lower than for d-DDM. This is a trivial consequence of the different D<sub>2</sub>O concentrations. More importantly, the  $q$ -dependence differs significantly in the two systems.

S. R. Midtgaard *et al.*

Invisible deuterated detergents for membrane proteins



**Fig. 2.** Reaction scheme for the synthesis of specifically deuterated levels of OG (the corresponding scheme for DDM can be found in Fig. S1). Step (A) The fatty acid is deuterated by Pt/C catalysed H/D exchange reactions at  $220^\circ\text{C}$  to produce the correctly deuterated version and then it is reduced to the corresponding alcohol (**1**) using the specified D:H ratio. Step (B) The deuterated alcohol (**1**) is coupled with the acetylated bromo-sugar head group, producing the tail-deuterated version of the detergent, which is then deacetylated to produce (**2**). Step (C) The sugar head group is deuterated via Raney Nickel catalysis to produce the final partially deuterated detergent (**3**) with different levels of deuteration in the head and the tail groups.

The oscillation around  $0.1 \text{ \AA}^{-1}$  arising from the core-shell contrast of the d25-DDM is still visible in the bR-micelle complex, but already from  $q > 0.02 \text{ \AA}^{-1}$  the two  $I(q)$ -functions start differing (see dashed line in Fig. 3F) as a result of the internal contrast of the d25-DDM micelles. This is already visible in the Guinier range and show that not even a reliable value for the  $R_g$  of bR can be obtained with the d25-DDM as reconstitution system. Similar effects, although less pronounced are to be expected if using d17-OG as the reconstitution system.

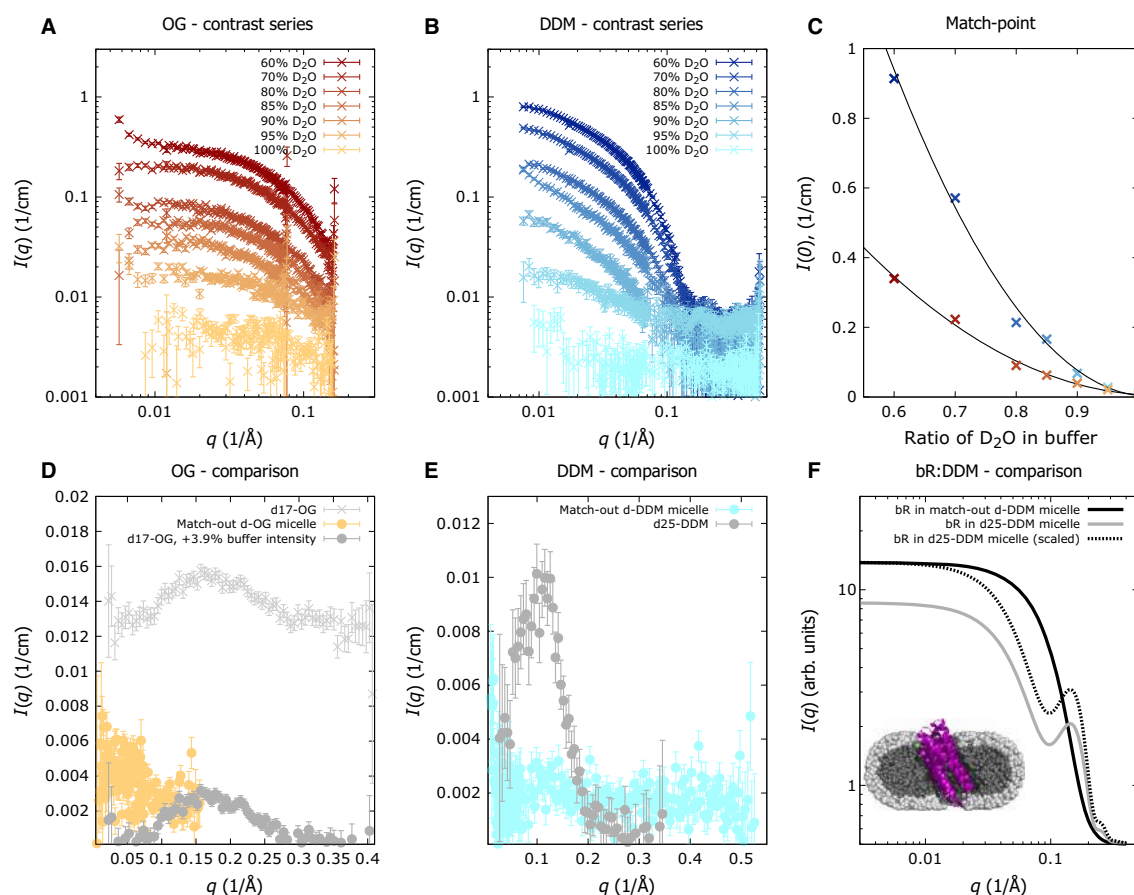
### Exchange of protonated to deuterated detergents

The extent of exchange from hydrogenated to deuterated detergents in samples of membrane proteins outlined in Fig. 1, was evaluated by monitoring the background SANS signal at high  $q$  as a function of SEC flowrate, as the hydrogen from any nonexchanged

buffer or detergent would generate an increase in incoherent background scattering. In initial experiments, the background scattering was found to be very high due to incomplete exchange of the detergent. The Agilent Bio SEC-3-300 column, used in combination with a high flowrate ( $0.7 \text{ mL}\cdot\text{min}^{-1}$ ) was found to be the cause and by lowering the flowrate to  $0.4 \text{ mL}\cdot\text{min}^{-1}$  the problem was solved. Hence, for this method to perform as intended, it is of paramount importance to use a sufficiently low flowrate in the SEC to enable an enhanced H/D exchange of detergent and solvent (see further details in Supporting Information 3).

### Membrane protein data and analysis

To evaluate this method, five structurally different membrane proteins were investigated; bacteriorhodopsin (bR), maltoporin (LamB), photosystem I (PSI), the ionotropic glutamate receptor A2 (GluA2) and the sarco/endoplasmatic reticulum  $\text{Ca}^{2+}$  ATPase



**Fig. 3.** (A) Background subtracted SANS data from match-out deuterated OG in increasing percentage of D<sub>2</sub>O in the buffer. (B) SANS data from match-out deuterated DDM in increasing percentage of D<sub>2</sub>O in the buffer. (C) Experimental  $I(0)$  values plotted (points) and the corresponding quadratic power law fits (lines). (D) Comparison of  $I(q)$  from match-out d-OG (yellow) and d17-OG (light and dark grey), both measured at their match point. Light grey curve plots d17-OG with the usual background subtraction. In the dark grey curve, the d17-OG background has been empirically increased by a factor of 1.039. (E) Comparison of match-out d-DDM (turquoise) and d25-DDM (grey) measured at its match point. Data from d17-OG and d25-DDM are reproduced from [28]. (A)–(E) Error bars are the standard deviation (SD), obtained from counting statistics and standard error propagation. (F) Simulation of a Bacteriorhodopsin monomer measured by SANS, respectively, in match-out d-DDM (black) and in d25-DDM (grey and dashed) micelles.

(SERCA). The first three proteins, bR, LamB and PSI, are all examples of membrane proteins that have either no or small domains outside the cell membrane. These three proteins were also chosen as they are relatively stable and at the same time representative of the wide

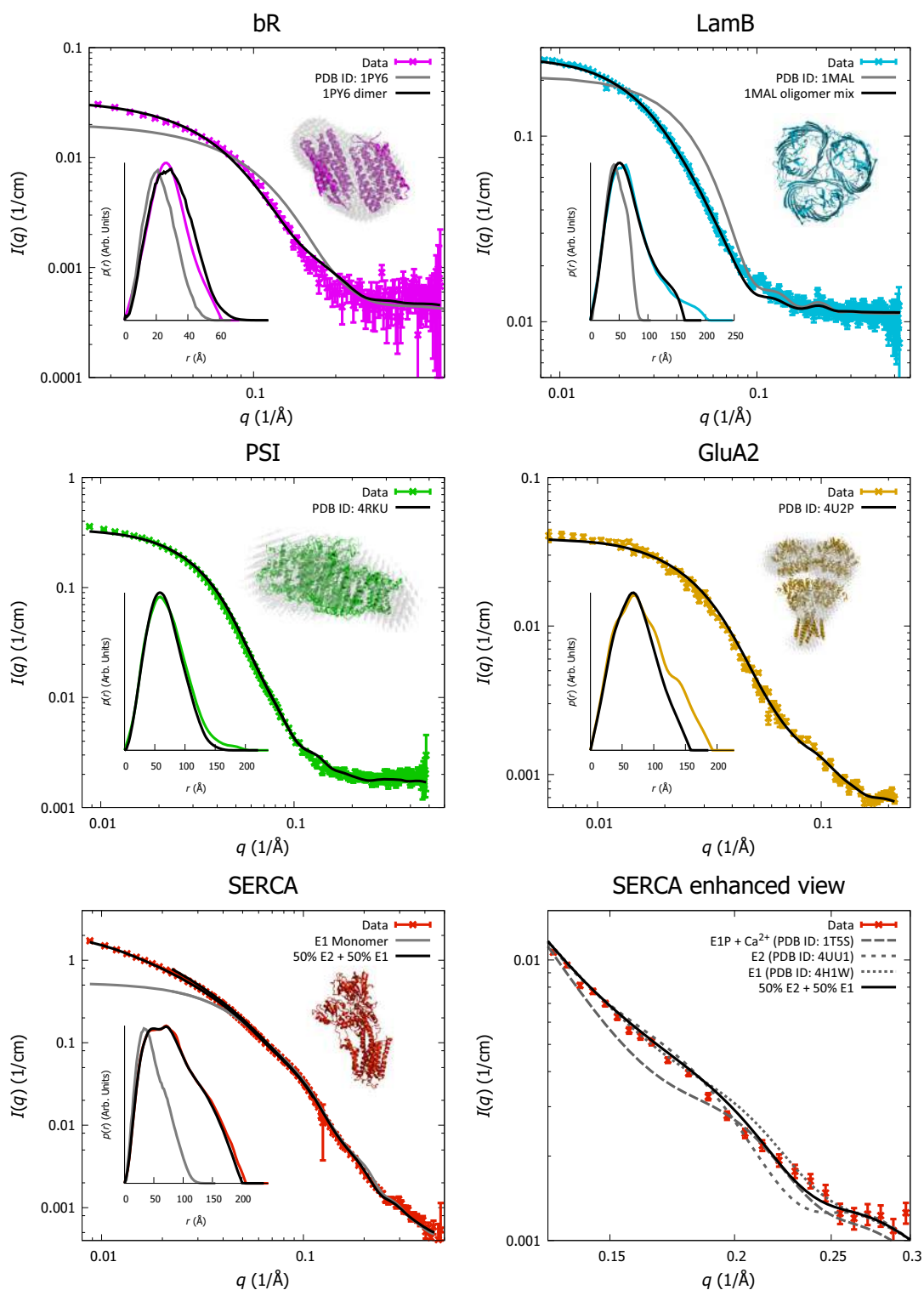
range of molecular weights found in membrane proteins viz. ~ 27 kDa (bR), ~ 47 kDa (LamB) and ~ 650 kDa (PSI). Two additional membrane proteins, GluA2 (~ 385 kDa) and SERCA (~ 110 kDa), were selected as examples of more unstable membrane

**Fig. 4.** Experimental SANS data (coloured points) and the resulting model fits (black full and dashed lines) for the five investigated membrane proteins. Error bars are the standard deviation (SD), obtained from counting statistics and standard error propagation. Note that instrumental resolution effects are included in the models, which then exhibit small discontinuities at medium  $q$ -values. Inserts: The corresponding pair-distance distribution functions as determined by Indirect Fourier transform [39] with identical colour scheme. Figures are the known structures from crystallography together with the *ab initio* bead models where appropriate. Note that SERCA (data, model and  $p(r)$ ) contain a fraction of aggregated protein.



S. R. Midtgaard *et al.*

Invisible deuterated detergents for membrane proteins



proteins with complex structures and large protein domains outside the membrane spanning region that may be captured in distinct functional states. SANS data on the LamB protein were obtained on the QUOKKA instrument at ANSTO, GluA2 was measured on KWS1 at FRM II and data from bR, PSI and SERCA were obtained on the D22 instrument at ILL. The SANS data and results obtained using the invisible detergents in 100% D<sub>2</sub>O are presented in Fig. 4 (See Supporting Information 4 for further details).

Crystal structures were available for all proteins, and these were used as a basis for modelling the theoretical scattering patterns to be compared to the experimental data (Fig. 4). bR was found to agree with a dimer structure, LamB, with a combination of homotrimers and higher oligomers of these, PSI data with a monomer structure, GluA2 with a homotetramer and SERCA with a combination of two structural states combined with a minor aggregated fraction of SERCA using a previously developed approach [40] (see SI 4 for more details). Hydration layers were added to all water-exposed surfaces. Instrumental resolution effects on the SANS data were taken into account in the analysis [41] and the required values for the  $\Delta q(q)$  for the resolution function calculations for the different instrumental settings were provided by the beamlines. The resolution effect is most clearly visible for the SERCA data where an (expected) discontinuity of the fitted model is observed around  $q = 0.03 \text{ 1/\AA}$  for all data. A constant background and a scale factor were fitted to obtain the best agreement between the data and model, shown in Fig. 4. The figure also plots the theoretical pair distance distribution functions  $p(r)$  calculated from the fitted models together with the ones obtained from the data by indirect Fourier transformation [39]. The values of radius of gyration,  $R_g$ , and maximum internal distances,  $D_{\max}$ , obtained from the  $p(r)$  functions, are listed in Table 2 together with molecular masses estimated from the estimated Porod volume and the partial specific molecular volume found by Mylonas *et al.* [42]. Table 2 compares these

values to the theoretical  $R_g$ ,  $D_{\max}$  and  $MW$  values based on the available crystal structures. Note that the experimental estimates for the Porod volume are not expected to have a precision better than  $\pm 20\%$  [30]. Finally, *Ab initio* bead modelling was performed for bR, PSI and GluA2 and compared with the known structures (See Fig. 4 and SI 4).

Bacteriorhodopsin with a monomeric molecular weight of  $\sim 27 \text{ kDa}$ , is a relatively small membrane protein with seven transmembrane helices. Due to its small size, it scatters weakly and any interfering signal from the detergent should be visible in the data. As it is evident from Fig. 4, the experimental data correspond well with a dimer of bR (PDB ID: 1PY6). While bR is traditionally found as a monomer or trimer, a dimeric form is well documented [43]. A closer look into the structural parameters in Table 2, show that the experimentally obtained  $R_g$ ,  $D_{\max}$  and  $MW$  values are in most cases slightly larger than what would be expected from the crystal structures of the dimer. This larger size is not only indicated in the obtained *ab initio* structure (Fig. 4) and can both be indications of slightly more loose or open dimer structures in solution than in the crystals but may also indicate the presence of small populations of higher order oligomers. These questions could be pursued further in a future work. However, the ability to discern the oligomeric state shows that the method can probe the structure of relatively small membrane proteins in solution without any significant contribution from the match-out deuterated detergent. bR has the same size and general fold as many G protein-coupled receptors (GPCRs) [44]. The data show that this interesting class of membrane proteins, typically too small for cryo-EM fall well within the range of this SANS-based method.

The LamB protein is a homotrimeric beta barrel porin structure with a monomer weight of  $\sim 142 \text{ kDa}$  [45], hence a membrane protein with intermediate molecular weight. The SANS data presented in Fig. 4 were found to agree with an oligomeric mixture of monomeric, trimeric and heptameric states of the

**Table 2.** Structural parameters deduced from the experimental data and the calculated theoretical values from the fitted models as obtained through Indirect Fourier Transform.  $R_g$ : Radius of gyration.  $D_{\max}$ : Maximum distance present in the scattering particle. MW: Molecular weight. Experimental values obtained from SANS data, theoretical values calculated from the available crystal structures.

	$R_g \text{ exp. [\AA]}$	$R_g \text{ theo. [\AA]}$	$D_{\max} \text{ exp. [\AA]}$	$D_{\max} \text{ theo. [\AA]}$	$MW \text{ exp. [kDa]}$	$MW \text{ calc. [kDa]}$
bR	$21.6 \pm 0.04$	17.4	$60.2 \pm 0.9$	53	44.6	27*2
LamB	$57.9 \pm 0.2$	33.4	$201 \pm 1$	87	210	47*3
PSI	$55.2 \pm 0.2$	51.3	$199 \pm 5$	159	473	650
GluA2	$60.0 \pm 0.4$	56.2	$166 \pm 1$	173	278	385
SERCA E1 + E2	$56.4 \pm 0.1$	38.4	$204.5 \pm 0.6$	123	460	110

underlying LamB homotrimer (PDB ID: 1MAL, See SI 4 for further details), making an *ab initio* structure irrelevant. While it is clear from the data and the values for  $R_g$ ,  $D_{\max}$  and  $MW$  listed in Table 2 that the data must arise from, on average, larger particles than the basic LamB homotrimeric structure, another base set of structures derived from the homo-trimer could likely also have provided a good representation of the measured data. A small deviation between the model and data is visible at  $0.11 \text{ \AA}^{-1}$ , this hints towards the possibility of investigating finer structural details of the LamB oligomerization with this method. However, a larger set of data would be required for this purpose.

Plant PSI is a large protein complex with a diameter of around 20 nm that comprises multiple, mostly alpha-helical subunits [46,47]. It is evident from the SANS data in Fig. 4 that the overall structure of PSI in solution is highly similar to the one recently found by X-ray crystallography (PDB ID: 4RKU). Interestingly, a small discrepancy is found around  $0.1 \text{ \AA}^{-1}$  and the experimental values for  $R_g$  and  $D_{\max}$  are slightly higher than the comparable value for the crystal structure. The slightly larger size observed by SANS is also reflected in the plotted *ab initio* structure (Fig. 4). The observed differences likely stem from the fact that  $\sim 10\%$  of the amino acid residues are not found in the crystal structure. Other contributing factors could be that the solution structure, due to increased flexibility, is slightly different from the one confined in the crystal or that the crystal structure is from pea PSI and the SANS data were obtained from barley PSI. In either case, it is interesting that this method is able to demonstrate minor discrepancies between the known model and the obtained SANS data for this otherwise highly studied membrane protein complex.

Ionotropic glutamate receptor A2 is a neurotransmitter-activated receptor with a structure consisting of a large, extended homotetramer [48]. The scattering data correspond very well with the known structure of GluA2 (PDB ID: 4U2P), which is also seen from the resemblance between the crystal structure and the *ab initio* model (see Fig. 4) and from the good agreement (within the experimental accuracy) between the experimental and expected values for the  $R_g$ ,  $D_{\max}$  and  $MW$  listed in Table 2. The values found in Table 2 also indicate a good agreement with the expected structure within the expected accuracy. This demonstrates that the method can readily probe the structure of large and complicated membrane proteins and, in this case, confirm that the crystal structure is representative of the solution structure.

Sarco/endoplasmic reticulum  $\text{Ca}^{2+}$  ATPase is a calcium-transporting ATPase and more difficult to stabilize in solution than the other proteins in this study. It has a complex fold with three major cytoplasmic domains outside the membrane spanning region that rearrange with respect to each other during the calcium transport cycle [49]. Several of the structural states relating to the pumping cycle of SERCA have been revealed by crystallography via stabilization with cofactors and inhibitors [50]. However, some states and dynamic transitions are still debated. Here, the structure in the absence of calcium and stabilized by the nonhydrolysable ATP analogue AMPPCP has been investigated, representing the so-called E1 and E2 states. High-quality SANS data have been collected as evidenced by the relatively small error bars and the presence of significant features in the high- $q$  region. The data correspond to monomeric SERCA together with a minor fraction of oligomeric SERCA ( $\sim 2\%$  of the molecules, see SI 4 for details). The data clearly delineate calcium-free SERCA structures, as is also evident from the enhanced view in Fig. 4 where the structure with calcium (PDB ID: 1T5S) completely fails to describe the finer details observed in the experimental data. The structural state probed in this study is expected to be dynamic and potentially switch between the E2 (PDB ID: 4UU1) and E1 (PDB ID: 4H1W) calcium-free states [50]. Fitting the individual E2 and E1 states and a superposition of the two structures to the data indeed reveals that while both the calcium-free states represent the features in the data better than the calcium-bound form, the best description of the data is obtained by a linear combination of the E2 and E1 states with a distribution of 50% E2 and 50% E1 (see SI 4 for details). Not only does this confirm what has been previously suggested in the literature but it also provides a result that could not have been obtained using a traditional crystallographic approach. Additionally, this result demonstrates that the workflow and method presented here is compatible with probing the finer structure of this challenging and unstable type of membrane proteins. There was aggregation in a fraction of the SERCA sample, as clearly seen from the experimental data in Fig. 4, and reflected in the  $R_g$ ,  $D_{\max}$  and  $MW$  listed in Table 2. This effect was included in the modelling as a structure factor for fractal aggregates and using a previously developed approach [40]. However, since the two structures of interest, E1 and E2, had approximately the same size, and hence the same scattering intensity at low- $q$ , the aggregation did not affect the assessment of the equilibrium between these states. A detailed description of the data analysis, the structure

factor as well as a more thorough discussion of this matter can be found in the supplemental information (SI 4).

## Discussion

The feasibility of using custom-synthesized match-out deuterated detergents for elucidating the structure of membrane proteins by solution SANS has been demonstrated. Firstly, the synthesis of partly deuterated versions of DDM and OG were developed and the desired match-out deuteration confirmed by SANS. Secondly, the detergents were successfully used for stabilizing membrane proteins while performing SANS experiments. Thirdly, by analysis of the obtained SANS data, it was demonstrated that this method does indeed perform as well as expected and hence provides a novel and easy-to-use tool for low-resolution structural studies of membrane proteins.

For the past decade, performing a SAXS experiment to complement high-resolution structural work by crystallography has become a standard approach when investigating soluble proteins [51]. This is not least due to the easy analysis made possible by the ATSAS suite [30] and development of stable and high-throughput SAXS beamlines [52]. This approach is now directly transferable to membrane proteins by using the approach of deuterated detergents and SANS.

The method of using match-out deuterated detergents to obtain solution structures of membrane proteins fills an important gap in the toolbox for their detailed investigation. While traditional protein crystallography has undeniably produced fantastic new insights into biological processes over several decades, obtaining crystals of membrane proteins diffracting to high resolution remains a bottleneck [53]. Indeed, larger flexible membrane proteins are under-represented in the protein data bank due to this obstacle. The method presented here allows for probing the structure in detail early in the process when a robust expression and purification protocol has been established. Recent advances in both the methodology and hardware have matured Cryo-EM to a technique that has become a major game changer in structural biology [8]. However, a major drawback of Cryo-EM is the requirement of proteins with a molecular mass above ~100–150 kDa to obtain sufficiently high signal-to-noise ratios to allow for the alignment and averaging of the individual frames required for obtaining the desired high-resolution data [8]. This means that many important membrane proteins are too small for Cryo-EM (although use of the Volta Phase Plate shows great promise for cryo-EM studies of even smaller

proteins [54]). NMR spectroscopy is also a popular method for determining the high-resolution structure of membrane proteins [55], but due to the overcrowding of the spectra, only relatively small proteins (~20–30 kDa) may be solved. Hence, the combined SANS and invisible detergent method described in this paper provides an easily applicable alternative that allows for analysing protein structures with a resolution down to about 10 Å, and that importantly contributes to bridging the gap between existing methods.

Several other strategies have been proposed in the literature which utilize some of the aspects of the method presented here. The use of small-angle X-ray scattering (SAXS) for elucidating the structure of membrane proteins in detergent micelles has been thoroughly tested [56–59] but thus far mainly revealed information on the detergent rim around the membrane proteins. Furthermore, using so-called nanodiscs coupled with small-angle scattering has also been shown to be possible but the sample preparation remains challenging [11,60]. This is contrasted with the approach presented here, where the sample preparation is uncomplicated and the signal almost exclusively comes from the membrane protein, significantly simplifying the data analysis and increasing the information that can be extracted about the structure of the investigated membrane protein. Importantly, this method allows for probing the structure of uncrystallizable dynamic structural states of proteins. Recent technical developments in SEC-SANS sample environments will furthermore allow for performing the whole workflow outlined in Fig. 1 *in situ* at the neutron instrument, together with removing any oligomeric species caused by unstable proteins [61].

Five different membrane proteins and two detergents were investigated in this study. The data demonstrate the general applicability of the approach and provide promising perspectives for future use. The basic idea of match-out deuteration may easily be generalized to other detergent types to accommodate special needs for particular membrane proteins or in combination with selective deuteration of different subunits.

## Methods

### General materials

Hydrogenated Octyl-β-D-glucopyranoside (OG) was purchased from AppliChem while hydrogenated n-dodecyl-β-D-maltopyranoside (DDM) was purchased from Sigma Aldrich. Match-out deuterated detergents were custom

synthesized as described briefly below and in detail in the supplemental information. The membrane proteins used in the present study were expressed and purified according to standard protocols as described in the supplemental information. All pH values have been measured using a standard glass electrode and all reported values are from the direct measurement [62].

### Densitometry

Densitometry was performed using a DMA 5000 density meter (Anton Paar). Octyl- $\beta$ -D-glucopyranoside (OG) (Sigma-Aldrich) was solubilized in 20 mM Tris/HCl, pH 7.5 and 100 mM NaCl in concentrations of 25 and 100 mM. *n*-dodecyl- $\beta$ -D-maltopyranoside (DDM) (Sigma-Aldrich) was solubilized in identical buffer at concentrations of 15 and 60 mM. All samples were prepared in triplicates and measured at 4 °C intervals between 4 °C and 28 °C. See further details in the supplemental information.

### General detergent synthesis

Chemicals and reagents of the highest grade were purchased from Sigma-Aldrich and Carbosynth Limited (Berkshire, UK) and were used without further purification. Solvents were purchased from Sigma-Aldrich and Merck and were purified by established methods [63]. NMR solvents were purchased from Cambridge Isotope Laboratories Inc. and Sigma-Aldrich and were used without further purification. D<sub>2</sub>O (99.8%) was supplied by Sigma-Aldrich. Thin layer chromatography (TLC) was performed on Fluka Analytical silica gel aluminium sheets (25 F254). Davisil<sup>®</sup> silica gel (LC60Å 40–63 micron) was used for bench-top column chromatography.

Deuteration of octanoic acid and dodecanoic acid was performed using hydrothermal H/D exchange reactions in D<sub>2</sub>O at 220 °C by mixing the appropriate amount of fatty acid with NaOD and Pt/C (10% w/w) in a Mini Benchtop 4560 Parr Reactor (600 mL vessel capacity, 3000 psi maximum pressure, 350 °C maximum temperature). This was followed by filtering the catalyst, acidifying the solution and then extracting the aqueous phase with ethyl acetate. Thin layer chromatography was used (referenced with the protonated compound) to estimate the purity and to develop separation protocols. <sup>1</sup>H NMR (400 MHz), <sup>13</sup>C NMR (100.6 MHz) and <sup>2</sup>H NMR (61.4 MHz) spectra were recorded on a Bruker 400 MHz spectrometer at 298 K. Chemical shifts, in p.p.m., were referenced to the residual signal of the corresponding NMR solvent. Deuterium NMR was performed using the probe's lock channel for direct observation. Electrospray ionization mass spectra (ESI-MS) were recorded on a 4000 QTrap AB Sciex spectrometer. The overall percentage deuteration of the molecules was calculated by MS using the isotope distribution analysis of the different isotopologues. This was calculated

taking into consideration the <sup>13</sup>C natural abundance, whose contribution was subtracted from the peak area of each *M* + 1 isotopologue to allow for accurate estimation of the percentage deuteration of each isotopologue.

### Transfer of membrane proteins into deuterated detergents

The transfer was performed using an Agilent Bio SEC-3-300 column with a flowrate of 0.4 mL·min<sup>-1</sup>. The flow through from the column was collected, measured and used for background subtraction.

Buffers used for PSI, GluA2 and bR contained 0.5 mM match-out deuterated DDM, 20 mM Tris/DCl pH 7.5 and 100 mM NaCl in D<sub>2</sub>O. For LamB, 40 mM match-out deuterated OG was used instead of DDM. For SERCA, the buffer used consisted of 20 mM MOPS pH 6.8, 100 mM KCl and 0.5 mM DDM.

### SANS measurements

The SANS measurements of the deuterated OG were performed at KWS 1, Forschungs Neutronenquelle Heinz Maier-Leibnitz, Munich (FRM II). A neutron wavelength of 4.5 Å (±10% FWHM) was used at two sample-to-detector distances covering *q*-ranges of 0.012–0.16 Å<sup>-1</sup> and 0.0057–0.077 Å<sup>-1</sup>. Precalibrated plexiglass was used as a standard for absolute calibration of the scattered intensity *I* (*q*) in units of 1/cm, where  $q = (4\pi/\lambda)\sin(\theta)$  and  $2\theta$  is the scattering angle and  $\lambda$  is the wavelength. GluA2 was measured at the same instrument using three sample-to-detector distances: 1.5 m and 4.0 m with 4 m collimation and 8.0 m with 8 m collimation were used.

The SANS measurements of the LamB protein in deuterated OG and empty deuterated DDM micelles were performed at QUOKKA [64], Australian Nuclear Science and Technology Organization (ANSTO), Sydney. A neutron wavelength of 5.0 Å (±10% FWHM) was used at two sample-to-detector distances, covering *q*-ranges of 0.017–0.51 Å<sup>-1</sup> and 0.0076–0.098 Å<sup>-1</sup>. Data were absolutely calibrated by an attenuated direct beam transmission measurement of the scattered intensity.

The SANS measurements of the bR, SERCA and PSI in deuterated DDM were performed at D22, Institut Laue-Langevin (ILL), Grenoble. A neutron wavelength of 6.0 Å (±10% FWHM) was used at two sample-to-detector distances covering *q*-ranges of 0.022–0.48 Å<sup>-1</sup> and 0.0087–0.12 Å<sup>-1</sup>. H<sub>2</sub>O was used as a standard for absolute calibration of the scattered intensity.

### SANS data analysis

Pair distance distribution functions, *p*(*r*) and the associated values for the radius of gyration, *R<sub>g</sub>*, and maximum particle diameter, *D<sub>max</sub>*, were obtained from the scattering data using

a Bayesian indirect Fourier transform [39]. Porod volume-based values for the protein molar mass were estimated by the GNOM program of the ATSAS package [65]. *Ab initio* models of the proteins were obtained using the DAMMIF program of the ATSAS package [65]. Theoretical scattering patterns for the relevant protein structures were calculated using the program WillItFit [31]. Hydrogen and deuterium was added to the structures by the Phenix program Ready-Set! [66], exchanging all labile hydrogens for deuterium. A surface hydration layer was added to the hydrophilic areas of the structures using an in-house routine which identifies surface residues and places beads at a distance of 5 Å from the C $\alpha$  of the surface residues. This routine allows for not adding a water layer at the hydrophobic portion of the protein surface as identified by the Orientations of proteins in Membranes' database (OPM) [67].

### Acknowledgements

The authors acknowledge Dr. Robert Knott at ANSTO and Dr. Adam Round at ESRF for their assistance in obtaining supporting SAXS data. We thank ANSTO, ILL and FRM II for awarding us beamtime for this project at, respectively, the QUOKKA, D22 and KWS-1 instruments. RV, CK, TST, JJD and JSK thank Eric Gouaux for providing the GluA2cryst construct and protein purification protocol. Dr. Frank Gabel and Dr. Christine Ebel are thanked for kindly providing data on the tail-deuterated detergents previously published in [28]. PhD student Nicolai Tidemand Johansen is thanked for helping with the SANS measurements of GluA2. We thank the Danish Agency for Science, Technology and Innovation funding agency for making this study possible through a Sapere Aude grant to LA. RV, CK, and JSK acknowledge the support from the Lundbeck Foundation, The Danish Council for Independent Research – Medical Sciences and GluTarget. The authors also acknowledge financial support from the Center for Synthetic Biology (bioSYNergy) funded by the UCPH Excellence Programme for Interdisciplinary Research and to the Lundbeck foundation “BRAIN-STRUC” project. The National Deuteration Facility is partly supported by the National Collaborative Research Infrastructure Strategy – an initiative of the Australian Government.

### Author contributions

SRM, TD and LA conceived the project. TD developed methods and synthesized the deuterated detergents. SRM, AJZN, PEJ, MB, CO, PN, RV, TST, JJD, CK and JSK produced and provided proteins.

SRM, MCP, SARK, PH, and AHL performed the experiments. AHL, NSG, GVJ, MCP and SARK performed the data analysis. HF, EG and AM were responsible for the SANS instrument and participated in the SANS measurements. SRM, TD and LA cowrote the paper.

### Conflict of interest

The authors declare no competing financial interests.

### References

- 1 Loll PJ (2014) Membrane proteins, detergents and crystals: what is the state of the art? *Acta Crystallogr Sect F Struct Biol Commun* **70**, 1576–1583.
- 2 Diller A, Loudet C, Aussenac F, Raffard G, Fournier S, Laguerre M, Grélard A, Opella SJ, Marassi FM & Dufourc EJ (2009) Bicelles: a natural “molecular goniometer” for structural, dynamical and topological studies of molecules in membranes. *Biochimie* **91**, 744–751.
- 3 Poulos S, Morgan JLW, Zimmer J & Faham S (2015) Bicelles coming of age. In *Membrane Proteins Engineering, Purification and Crystallization* 1st edn. pp. 393–416. Elsevier Inc., Amsterdam.
- 4 Midtgaard SR, Pedersen MC, Kirkensgaard JJK, Sørensen KK, Mortensen K, Jensen KJ & Arleth L (2014) Self-assembling peptides form nanodiscs that stabilize membrane proteins. *Soft Matter* **10**, 738–752.
- 5 Knowles TJ, Finka R, Smith C, Lin Y-P, Dafforn T & Overduin M (2009) Membrane proteins solubilized intact in lipid containing nanoparticles bounded by styrene maleic acid copolymer. *J Am Chem Soc* **131**, 7484–7485.
- 6 Popot J-L, Althoff T, Bagnard D, Banères J-L, Bazzacco P, Billon-Denis E, Catoire LJ, Champeil P, Charvolin D, Cocco MJ *et al.* (2011) Amphipols from A to Z. *Annu Rev Biophys* **40**, 379–408.
- 7 Ritchie TK, Grinkova YV, Bayburt TH, Denisov IG, Zolnerciks JK, Atkins WM & Sligar SG (2009) Chapter 11 reconstitution of membrane proteins in phospholipid bilayer nanodiscs. *Methods Enzymol* **464**, 211–231.
- 8 Nogales E & Scheres SHW (2015) Cryo-EM: a unique tool for the visualization of macromolecular complexity. *Mol Cell* **58**, 677–689.
- 9 Murray DT, Das N & Cross TA (2013) Solid state NMR strategy for characterizing native membrane protein structures. *Acc Chem Res* **46**, 2172–2181.
- 10 Marie S, Skar-Gislinge N, Midtgaard S, Thygesen MB, Schiller J, Frielinghaus H, Moulin M, Haertlein M, Forsyth VT, Pomorski TG *et al.* (2014) Stealth carriers for low-resolution structure determination of membrane

- proteins in solution. *Acta Crystallogr D Biol Crystallogr* **70**, 317–328.
- 11 Kynde SAR, Skar-Gislinge N, Pedersen MC, Midtgaard SR, Simonsen JB, Schweins R, Mortensen K & Arleth L (2014) Small-angle scattering gives direct structural information about a membrane protein inside a lipid environment. *Acta Crystallogr D Biol Crystallogr* **70**, 371–383.
  - 12 Skar-Gislinge N, Kynde SAR, Denisov IG, Ye X, Lenov I, Sligar SG & Arleth L (2015) Small-angle scattering determination of the shape and localization of human cytochrome P450 embedded in a phospholipid nanodisc environment. *Acta Crystallogr Sect D Biol Crystallogr* **71**, 2412–2421.
  - 13 Nawroth T, Dose K & Conrad H (1989) Neutron small angle scattering of detergent solubilized and membrane bound ATP-synthase. *Phys B Condens Matter* **156–157**, 489–492.
  - 14 Sverzhinsky A, Qian S, Yang L, Allaire M, Moraes I, Ma D, Chung JW, Zoonens M, Popot JL & Coulton JW (2014) Amphipol-trapped ExbB-ExbD membrane protein complex from *Escherichia coli*: a biochemical and structural case study. *J Membr Biol* **247**, 1005–1018.
  - 15 Wise DS, Karlin A & Schoenborn BP (1979) An analysis by low-angle neutron scattering of the structure of the acetylcholine receptor from *Torpedo californica* in detergent solution. *Biophys J* **28**, 473–496.
  - 16 Le RK, Harris BJ, Iwuchukwu IJ, Bruce BD, Cheng X, Qian S, Heller WT, O'Neill H & Frymier PD (2014) Analysis of the solution structure of *Thermosynechococcus elongatus* photosystem I in  $n$ -dodecyl- $\beta$ -D-maltoside using small-angle neutron scattering and molecular dynamics simulation. *Arch Biochem Biophys* **550–551**, 50–57.
  - 17 Hunt JF, McCrea PD, Zaccari G & Engelman DM (1997) Assessment of the aggregation state of integral membrane proteins in reconstituted phospholipid vesicles using small angle neutron scattering. *J Mol Biol* **273**, 1004–1019.
  - 18 Wang ZY, Muraoka Y, Nagao M, Shibayama M, Kobayashi M & Nozawa T (2003) Determination of the B820 subunit size of a bacterial core light-harvesting complex by small-angle neutron scattering. *Biochemistry* **42**, 11555–11560.
  - 19 Gall A, Dellerue S, Lapouge K, Robert B & Le L (2001) Scattering measurements on the membrane protein subunit B777 in a detergent. *Biopolymers* **58**, 231–234.
  - 20 Cardoso MB, Smolensky D, Heller WT & O'Neill H (2009) Insight into the structure of light-harvesting complex II and its stabilization in detergent solution. *J Phys Chem B* **113**, 16377–16383.
  - 21 Clifton LA, Johnson CL, Solovyova AS, Callow P, Weiss KL, Ridley H, Le Brun AP, Kinane CJ, Webster JRP, Holt SA *et al.* (2012) Low resolution structure and dynamics of a colicin-receptor complex determined by neutron scattering. *J Biol Chem* **287**, 337–346.
  - 22 Pachence JM, Edelman IS & Schoenborn BP (1987) Low-angle neutron scattering analysis of Na/K-ATPase in detergent solution. *J Biol Chem* **262**, 702–709.
  - 23 Perkins SJ & Weiss H (1983) Low-resolution structural studies of mitochondrial ubiquinol:cytochrome c reductase in detergent solutions by neutron scattering. *J Mol Biol* **168**, 847–866.
  - 24 Gabel F, Lensink MF, Clantin B, Jacob-Dubuisson F, Villeret V & Ebel C (2014) Probing the conformation of FhaC with small-angle neutron scattering and molecular modeling. *Biophys J* **107**, 185–196.
  - 25 Block MR, Zaccari G, Lauquin GJ & Vignais PV (1982) Small angle neutron scattering of the mitochondrial ADP/ATP carrier protein in detergent. *Biochem Biophys Res Commun* **109**, 471–477.
  - 26 Osborne HB, Sardet C, Michel-Villaz M & Chabre M (1978) Structural study of rhodopsin in detergent micelles by small-angle neutron scattering. *J Mol Biol* **123**, 177–206.
  - 27 Nogales A, García C, Pérez J, Callow P, Ezquerro TA & González-Rodríguez J (2010) Three-dimensional model of human platelet integrin  $\alpha$ IIb  $\beta$ 3 in solution obtained by small angle neutron scattering. *J Biol Chem* **285**, 1023–1031.
  - 28 Breyton C, Gabel F, Lethier M, Flayhan A, Durand G, Jault J-M, Juillan-Binard C, Imbert L, Moulin M, Ravaud S *et al.* (2013) Small angle neutron scattering for the study of solubilised membrane proteins. *Eur Phys J E Soft Matter* **36**, 71.
  - 29 Breyton C, Flayhan A, Gabel F, Lethier M, Durand G, Boulanger P, Chami M & Ebel C (2013) Assessing the conformational changes of pb5, the receptor-binding protein of phage T5, upon binding to its *Escherichia coli* receptor FhuA. *J Biol Chem* **288**, 30763–30772.
  - 30 Petoukhov MV, Franke D, Shkumatov AV, Tria G, Kikhney AG, Gajda M, Gorba C, Mertens HDT, Konarev PV & Svergun DI (2012) New developments in the ATSAS program package for small-angle scattering data analysis. *J Appl Crystallogr* **45**, 342–350.
  - 31 Pedersen MC, Arleth L & Mortensen K (2013) WillItFit: a framework for fitting of constrained models to small-angle scattering data. *J Appl Crystallogr* **46**, 1894–1898.
  - 32 Curtis JE & Krueger S. SASSIE. NIST.
  - 33 Ilavsky J & Jemian PR (2009) Irena: tool suite for modeling and analysis of small-angle scattering. *J Appl Crystallogr* **42**, 347–353.
  - 34 Schneidman-Duhovny D, Hammel M & Sali A (2010) FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res* **38**, 540–544.

- 35 Darwish TA, Luks E, Moraes G, Yepuri NR, Holden PJ & James M (2013) Synthesis of deuterated [D 32] oleic acid and its phospholipid derivative [D 64]dioleoyl- sn -glycero-3-phosphocholine. *J Label Compd Radiopharm* **56**, 520–529.
- 36 Koch HJ & Stuart RS (1977) A novel method for specific labelling of carbohydrates with deuterium by catalytic exchange. *Carbohydr Res* **59**, C1–C6.
- 37 Hans J & Koch RSS (1978) The synthesis of per-C-deuterated D-glucose. *Carbohydr Res* **64**, 127–134.
- 38 Koch H & Stuart R (1978) The catalytic C-deuteration of some carbohydrate derivatives. *Carbohydr Res* **67**, 341–348.
- 39 Hansen S (2012) BayesApp : a web site for indirect transformation of small-angle scattering data. *J Appl Crystallogr* **45**, 566–567.
- 40 Malik L, Nygaard J, Høiberg-Nielsen R, Arleth L, Hoeg-Jensen T & Jensen KJ (2012) Perfluoroalkyl chains direct novel self-assembly of insulin. *Langmuir* **28**, 593–603.
- 41 Pedersen JS (1993) Resolution effects and analysis of small-angle neutron scattering data. *Le J Phys IV* **3**, C8-491–C8-498.
- 42 Mylonas E & Svergun DI (2007) Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering. *J Appl Crystallogr* **40**, s245–s249.
- 43 Michel H, Oesterhelt D & Henderson R (1980) Orthorhombic two-dimensional crystal form of purple membrane. *Proc Natl Acad Sci USA* **77**, 338–342.
- 44 Rasmussen SGF, Choi H-J, Rosenbaum DM, Kobilka TS, Thian FS, Edwards PC, Burghammer M, Ratnala VRP, Sanishvili R, Fischetti RF *et al.* (2007) Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. *Nature* **450**, 383–387.
- 45 Schirmer T, Keller TA, Wang YF & Rosenbusch JP (1995) Structural basis for sugar translocation through maltoporin channels at 3.1 Å resolution. *Science* **267**, 512–514.
- 46 Fromme P, Jordan P & Krauß N (2001) Structure of photosystem I. *Biochim Biophys Acta - Bioenerg* **1507**, 5–31.
- 47 Ben-Shem A, Frolow F & Nelson N (2003) Crystal structure of plant photosystem I. *Nature* **426**, 630–635.
- 48 Sobolevsky AI, Rosconi MP & Gouaux E (2009) X-ray structure, symmetry and mechanism of an AMPA-subtype glutamate receptor. *Nature* **462**, 745–756.
- 49 Olesen C, Picard M, Winther A-ML, Gyrup C, Morth JP, Oxvig C, Møller JV & Nissen P (2007) The structural basis of calcium transport by the calcium pump. *Nature* **450**, 1036–1042.
- 50 Bublitz M, Musgaard M, Poulsen H, Thøgersen L, Olesen C, Schiøtt B, Morth JP, Møller JV & Nissen P (2013) Ion pathways in the sarcoplasmic reticulum Ca<sup>2+</sup> + -ATPase. *J Biol Chem* **288**, 10759–10765.
- 51 Putnam CD, Hammel M, Hura GL & Tainer JA (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q Rev Biophys* **40**, 191–285.
- 52 Pernot P, Theveneau P, Giraud T, Fernandes RN, Nurizzo D, Spruce D, Surr J, McSweeney S, Round A, Felisaz F *et al.* (2010) New beamline dedicated to solution scattering from biological macromolecules at the ESRF. *J Phys Conf Ser* **247**, 12009.
- 53 Columbus L (2015) Post-expression strategies for structural investigations of membrane proteins. *Curr Opin Struct Biol* **32**, 131–138.
- 54 Khoshouei M, Radjainia M, Baumeister W, Danev R (2017) Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. *Nat Commun* **8**, 16099. <https://doi.org/10.1038/ncomms16099>
- 55 Maslennikov I & Choe S (2013) Advances in NMR structures of integral membrane proteins. *Curr Opin Struct Biol* **23**, 555–562.
- 56 Calcutta A, Jessen CM, Behrens MA, Oliveira CLP, Renart ML, González-Ros JM, Otzen DE, Pedersen JS, Malmendal A & Nielsen NC (2012) Mapping of unfolding states of integral helical membrane proteins by GPS-NMR and scattering techniques: TFE-induced unfolding of KcsA in DDM surfactant. *Biochim Biophys Acta* **1818**, 2290–2301.
- 57 Døvling Kaspersen J, Moestrup Jessen C, Stougaard Vad B, Skipper Sørensen E, Kleiner Andersen K, Glasius M, Pinto Oliveira CL, Otzen DE & Pedersen JS (2014) Low-resolution structures of OmpA-DDM protein-detergent complexes. *ChemBioChem* **15**, 2113–2124.
- 58 Berthaud A, Manzi J, Pérez J & Mangenot S (2012) Modeling detergent organization around aquaporin-0 using small-angle X-ray scattering. *J Am Chem Soc* **134**, 10080–10088.
- 59 Bu Z & Engelman DM (1999) A method for determining transmembrane helix association and orientation in detergent micelles using small angle x-ray scattering. *Biophys J* **77**, 1064–1073.
- 60 Skar-Gislinge N, Simonsen JB, Mortensen K, Feidenhans'l R, Sligar SG, Lindberg Møller B, Bjørnholm T & Arleth L (2010) Elliptical structure of phospholipid bilayer nanodiscs encapsulated by scaffold proteins: casting the roles of the lipids and the protein. *J Am Chem Soc* **132**, 13713–13722.
- 61 Jordan A, Jacques M, Merrick C, Devos J, Forsyth VT, Porcar L & Martel A (2016) SEC-SANS: size exclusion chromatography combined *in situ* with small-angle neutron scattering. *J Appl Crystallogr* **49**, 2015–2020.
- 62 Glasoe PK & Long FA (1960) Use of glass electrodes to measure acidities in deuterium oxide. *J Phys Chem* **64**, 188–190.



S. R. Midtgaard *et al.*

Invisible deuterated detergents for membrane proteins

- 63 Armarego WLF & Perrin DD (1996) *Purification of Laboratory Chemicals*, 4th edn. Elsevier.
- 64 Gilbert EP, Schulz JC & Noakes TJ (2006) “Quokka”-the small-angle neutron scattering instrument at OPAL. *Phys B Condens Matter* **385–386**, 1180–1182.
- 65 Franke D & Svergun DI (2009) DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *J Appl Crystallogr* **42**, 342–346.
- 66 Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr Sect D Biol Crystallogr* **66**, 213–221.
- 67 Lomize MA, Lomize AL, Pogozheva ID & Mosberg HI (2006) OPM: orientations of proteins in membranes database. *Bioinformatics* **22**, 623–625.

### Supporting information

Additional Supporting Information may be found online in the supporting information tab for this article.



## **Invisible detergents for structure determination of membrane proteins by small-angle neutron scattering**

Søren Roi Midtgaard, Tamim A. Darwish, Martin Cramer Pedersen, Pie Huda, Andreas Haahr Larsen, Grethe Vestergaard Jensen, Søren Andreas Røssell Kynde, Nicholas Skar-Gislinge, Agnieszka Janina Zygadlo Nielsen, Claus Olesen, Mickael Blaise, Jerzy Józef Dorosz, Thor Seneca Thorsen, Raminta Venskutonytė, Christian Krintel, Jesper V. Møller, Henrich Frielinghaus, Elliot Paul Gilbert, Anne Martel, Jette Sandholm Kastrup, Poul Erik Jensen, Poul Nissen and Lise Arleth

DOI: 10.1111/febs.14345

## Supplemental information

1. Densitometry experiments and determining deuteration levels .....	2
2. Detergent synthesis .....	4
Synthesis of octanoic acid- $d_{15(94\% \text{ D})}$ : .....	4
Synthesis of octanol- $d_{17(94\% \text{ D})}$ (1): .....	6
Synthesis of 2,3,4,6-tetra-O-acetyl-octyl- $d_{17(94\% \text{ D})}$ - $\beta$ -D-glucopyranoside: .....	8
Synthesis of octyl- $d_{17(94\% \text{ D})}$ - $\beta$ -D-glucopyranoside (2): .....	10
Synthesis of octyl- $d_{17(94\% \text{ D})}$ - $\beta$ -D-glucopyranoside- $d_{7(52\% \text{ D})}$ (3): .....	13
Synthesis of dodecanoic acid- $d_{23(89\% \text{ D})}$ : .....	16
Synthesis of dodecanol- $d_{23(89\% \text{ D})}$ (1): .....	18
Synthesis of 2,3,6,2',3',4',6'-hepta-O-acetyl-dodecyl- $d_{25(89\% \text{ D})}$ - $\beta$ -D-maltopyranoside: .....	19
Synthesis of <i>n</i> -dodecyl- $d_{25(89\% \text{ D})}$ - $\beta$ -D-maltopyranoside (5): .....	23
Synthesis of <i>n</i> -dodecyl- $d_{25(89\% \text{ D})}$ - $\beta$ -D-maltopyranoside- $d_{14(57\% \text{ D})}$ (6): .....	25
3. Protein production, purification and handling .....	28
Bacteriorhodopsin (bR) production and purification .....	28
Photosystem I (PSI) production and purification .....	28
Maltoporin (LamB) production and purification .....	28
Ionotropic Glutamate Receptor A2 (GluA2) production and purification .....	29
Sarco/endoplasmatic reticulum $\text{Ca}^{2+}$ ATPase (SERCA) production and purification .....	30
Transfer of membrane proteins into deuterated detergents .....	31
Protein concentration measurements .....	31
4. Small-angle neutron scattering .....	32
Small-angle neutron scattering at FRMII .....	32
Small-angle neutron scattering at ANSTO .....	32
Small-angle neutron scattering at ILL .....	32
Measurement on empty DDM and OG micelles .....	32
Data analysis .....	33
Maltoporin (LamB) .....	34
SERCA .....	35
5. References .....	39

## 4. Small-angle neutron scattering

### Small-angle neutron scattering at FRMII

SANS measurements of the deuterated OG were performed at KWS 1[12], Forschungs Neutronenquelle Heinz Maier-Leibnitz, Munich (FRM II). Measurements on empty micelles were performed using neutrons with a wavelength of 4.5 Å with a wavelength spread of 10% (FWHM) at two sample-to-detector distances: 3.77 m with 4 m collimation and 7.77 m with 8 m collimation. These settings covered  $q$ -ranges of 0.012–0.16 Å<sup>-1</sup> and 0.0057–0.077 Å<sup>-1</sup>, respectively. For measurements on GluA2 three sample-to-detector distances: 1.5 m and 4.0 m with 4 m collimation and 8.0 m with 8 m collimation were used. These settings covered the  $q$ -range 0.006–0.44 Å<sup>-1</sup> with good overlap. Data were reduced using available software at the beamline and a pre-calibrated plexiglass was used as a secondary standard for absolute calibration of the scattered intensity  $I(q)$  in units of 1/cm, where  $q = (4\pi/\lambda)\sin(\theta)$  and  $2\theta$  is the scattering angle and  $\lambda$  is the wavelength.

### Small-angle neutron scattering at ANSTO

SANS measurements of the LamB protein in deuterated OG detergent and empty deuterated DDM micelles were performed at QUOKKA[13], Australian Nuclear Science and Technology Organization (ANSTO), Sydney. Measurements were performed using neutrons with a wavelength of 5.0 Å with a wavelength spread of 10% (FWHM) at two sample-to-detector distances: 2.0 m with 12.0 m collimation and 8.0 m with 8.0 m collimation. These settings covered  $q$ -ranges of 0.017–0.51 Å<sup>-1</sup> and 0.0076–0.098 Å<sup>-1</sup>, respectively. Data were reduced using available software[14] at the beamline and absolutely calibrated by an attenuated direct beam transmission measurement of the scattered intensity.

### Small-angle neutron scattering at ILL

SANS measurements of the bR, SERCA, PSI and GluA2 in deuterated DDM were performed at D22<sup>1</sup>, Institut Laue-Langevin (ILL), Grenoble. Measurements were performed using neutrons with a wavelength of 6.0 Å with a wavelength spread of 10% (FWHM) at two sample-to-detector distances: 2.0 m with 5.6 m collimation and 5.6 m with 5.6 m collimation. These settings covered  $q$ -ranges of 0.022–0.48 Å<sup>-1</sup> and 0.0087–0.12 Å<sup>-1</sup>, respectively. Data were reduced using available software GRASP<sup>2</sup> at the beamline and H<sub>2</sub>O was used as a standard for absolute calibration of the scattered intensity.

### Measurement on empty DDM and OG micelles

DDM micelles were prepared in a final concentration of 15 mM in buffer containing 20 mM Tris/DCI pH/D 7.5 and 100 mM NaCl in increasing percentages of D<sub>2</sub>O. OG micelles were prepared in a final concentration of 100 mM in the same buffer.

---

<sup>1</sup> There are currently no published papers that describe the D22 instrument. The reader is referred to the instrument website for details: <https://www.ill.eu/instruments-support/instruments-groups/instruments/d22/more/documentation/>

<sup>2</sup> There are currently no published papers that describe the GRASP software. The information can be found on the software website: <https://www.ill.eu/instruments-support/instruments-groups/groups/lss/grasp/home/>

## Data analysis

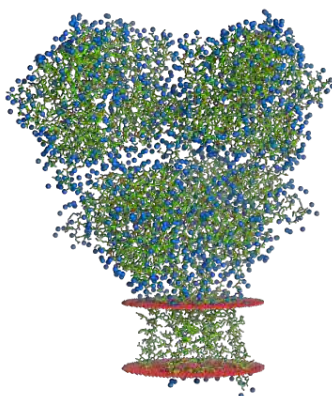
Pair distance distribution functions,  $p(r)$ , presented in Figure 4 (in the main text) and the associated values for the radius of gyration,  $R_g$ , and maximum particle distance,  $D_{max}$ , were obtained from the scattering data using a Bayesian indirect Fourier transform[15].

*Ab initio* models of the proteins were obtained using the DAMMIF program of the ATSAS package[16]. 20 Monte Carlo runs were performed, and the most typical structures were selected and averaged using the DAMAVER program of the ATSAS package[17]. The average bead structures are presented in figure 4 (main text), with the sizes of the beads representing the occupancy of their position.

The scattering data were compared to calculated scattering patterns from the relevant protein structures found in the protein data bank (PDB). Hydrogen and deuterium were added to the protein structures using the Phenix program ReadySet![18], exchanging all labile hydrogens for deuterium. For each of the protein structures, a surface hydration layer was added to the structure using an in-house routine which identifies surface residues and places beads at a distance of 5 Å from the  $C_\alpha$  of the surface residues. Each bead is associated with 4.1 water molecules, which is derived from the surface area per bead estimated by comparing the van der Waals surface and the amount of placed beads for a small number of structures. The hydrophobic portion of each protein surface expected to be covered by detergent molecules was identified using the Orientations of Proteins in Membranes' database (OPM)[19], and water beads were not added here. An example of a membrane protein structure with added water beads is given in figure S34.

In the model used to compute the scattering from the protein structures, the atomic coordinates in the PDB file were coarse-grained to represent individual amino acids[20] in an effort to speed up the calculations. The amplitudes of the amino acids were expanded on spherical harmonics and Bessel functions, which offers a convenient way of summing amplitudes for several components in the model similar to what is done in the Crysol algorithm [21].

The small-angle scattering from the pdb-files from the hydrogenated protein structures was finally calculated using the program WillItFit[22] which also allows for taking into account the resolution effects. A scale factor and a flat background were refined to obtain the best possible fit to the data. Resolution effects were taken into account by using the uncertainties in  $q$  generated from the SANS data reduction procedure(s). For bR, SERCA, and maltoporin, the models included protein oligomers, details for the latter two are given in the sections below.



**Figure S34: GluA2 in the apo state.** The structure from the protein data bank with ID 4U2P is shown, with added water beads (blue) and hydrophobic bilayer planes (red). The orientation of the membrane protein in the bilayer as well as the bilayer thickness were determined as described in Lomize *et al*[19].

#### Maltoporin (LamB)

The initial data processing of the scattering data from LamB in DDM indicated that the protein was not in the state as one might infer from the published crystal structures, as the MW,  $R_g$  and  $D_{max}$  (see table 2) were considerably greater than the values for the trimeric structure, (PDB ID: 1MAL) from the protein databank, and from which the analysis is based. The sample production and scattering experiment for this sample were repeated in order to verify this.

Based on the values in Table 2, an initial informed view would suggest the trimers self-assemble into a larger structure. Due to the manner in which LamB embeds in a bilayer, and due to the geometry and symmetry of the protein itself, an educated guess would be that three or seven trimeric subunits would self-assemble into a homo-heptamer or a homo-21-mer. Thus, we developed a model representing these structures and compared the resulting predicted scattering from them with the recorded scattering data. The data and fits are shown in figure S35. Apart from the scaling factor and background we also refined another parameter describing the rotation of the individual trimer substructures in the oligomers. Unsurprisingly, this parameter was not determined well by the data, but we have decided to report it for completeness.

The model based on the homo-heptameric structure (in red in figure S35) seems to produce the best fit to the data using a single structure. However, based on non-ideal fitting at the lower values of  $q$  and around  $0.09 \text{ \AA}^{-1}$ , we find it plausible that a model with several different oligomers might be able to rectify this. The fit shown in figure 4 in the main article of the 3 structures shown in figure S35 revealed a distribution of the monomer, trimer and heptamer of 53.7%:35.7%:10.6%. .

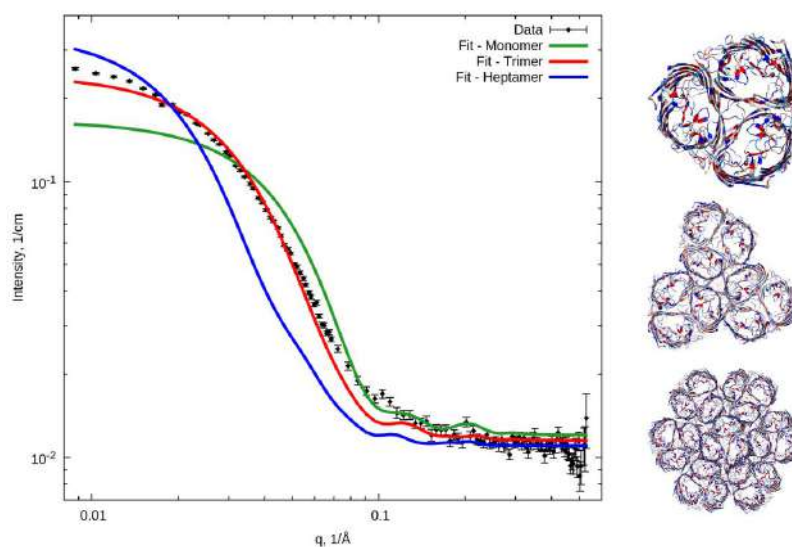
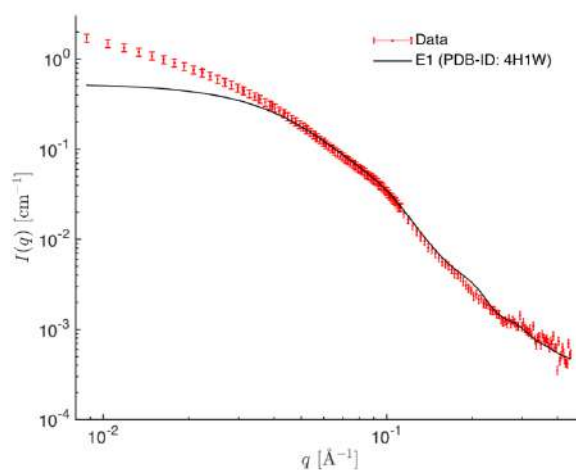


Figure S35: Data and fits for maltoporin. To the right, the structure from PDB ID: 1MAL in the protein data bank is shown on top, and a fit from a model based on this structure is plotted in dark green to the left. A trimer of these trimers is shown in the middle, and the scattering from this structure is plotted in red. Below, a heptamer constructed from the initial trimeric structure is shown, and the scattering computed from this structure is plotted in blue.

### SERCA

The data showed presence of oligomers in the sample, as seen in figure S36, where the monomeric E1 structure of SERCA clearly failed to match data for  $q < 0.06 \text{ \AA}^{-1}$ .



**Figure S36: Data from SERCA sample in red, and a fitted model of monomeric SERCA E1 (PDB-ID: 4H1W) in black.**

Therefore, it was hypothesized that the sample contained a minor fraction of protein aggregates. This fraction was modelled as a product of a fractal structure factor  $S'(q)$ , and a monomeric form factor  $P(q)$ , following the methodology of previous studies[23]

$$I(q)_{fractal} \propto S'(q)_{fractal} \cdot P(q)_{monomer}$$

The monomeric subunits of the fractal aggregate were assumed to be randomly oriented with respect to each other, which was taken into account by the decoupling approximation[24]

$$S'(q)_{fractal} = 1 + \beta(q) \cdot [S(q)_{fractal} - 1]$$

where  $\beta(q)$  is given in terms of the form factor amplitude  $F(q)$ . The brackets  $\langle \dots \rangle$  denote orientational averaging

$$\beta(q) = \frac{\langle F(q) \rangle^2}{\langle F(q)^2 \rangle} = \frac{(A_{m=0}^{l=0})^2}{P(q)}$$

The nominator is the square of the 0<sup>th</sup> order spherical harmonics expansion of  $F(q)$  [21,25], and the denominator is the form factor of a single SERCA protein.

The structure factor of a mass fractal was derived by Teixeira[26]

$$S(q)_{fractal} = 1 + \frac{1}{qr^D} \cdot \frac{D \cdot \Gamma(D-1)}{[1 + 1/(q\xi)^2]^{(D-1)/2}} \cdot \sin[(D-1) \cdot \text{atan}(q\xi)]$$

where  $\Gamma(\dots)$  is the gamma function,  $D$  is the dimensionality of the fractal ( $1 < D < 3$ ),  $r$  is the distance between SERCA molecules in the fractal aggregate, and  $\xi$  is the correlation length of the fractal, which is related to the radius of gyration  $R_g$  of the aggregates by

$$R_g^2 = \frac{D(D+1)}{2} \cdot \xi^2 \quad (1)$$

The intensity for the  $i^{th}$  structural state is thus a linear combination of monomers and fractal aggregates, with  $\gamma$  denoting the fraction of SERCA molecules in aggregated form, and  $N$  is the number of SERCA molecules per aggregate.  $B$  is a constant background contribution

$$I(q)_i = C \cdot [\gamma \cdot N \cdot S'(q)_{fractal} + (1 - \gamma)] \cdot P(q) + B \quad (2)$$

The prefactor  $C = K \cdot n_{SERCA} \cdot \Delta\rho_{SERCA}^2 \cdot V_{SERCA}^2$  is given in terms of  $n_{SERCA}$ , the molar concentration of SERCA molecules,  $\Delta\rho_{SERCA}$ , the excess scattering length density inside the protein,  $V_{SERCA}$ , the volume of a SERCA molecule, and  $K$ , a correction factor for the concentration measurement.  $K$  was fitted and should be close to



unity, whereas the other factors were calculated from the measured concentration and experimentally determined volumes of atoms in proteins[21,27].

To investigate whether the sample was in an equilibrium between two structural states, the parameter  $\alpha$  was introduced to denote the fraction of the intensity coming from the first state

$$I(q)_{equilibrium} = \alpha \cdot I(q)_1 + (1 - \alpha) \cdot I(q)_2$$

Hence, the model had a total of 8 parameters, as listed in Table ST4. The mean inter-molecular distance  $r$  of the fractal aggregate was however fixed, to equal the radius of a sphere fulfilling  $V_{sphere} = V_{SERCA}$ . So the total number of free parameters was 6 for the model assuming a single state, and 7 for the model assuming an equilibrium of two structural states. The parameters  $\gamma$ ,  $N$  were strongly correlated as seen from equation (2), and  $D$  and  $\xi$  were likewise correlated, as seen from equation (1), meaning that these parameters were not well-determined, as reflected in the large uncertainties. Hence, the parameters describing the fractal, as listed in Table ST4 represent one possible solution.

	Fractals + 1T5S	Fractals + 4UU1	Fractals + 4H1W	Fractals + 4UU1 4H1W
$K$	<b>0.84</b>	<b>0.81 ± 5</b>	<b>0.85 ± 4</b>	<b>0.83 ± 6</b>
$\gamma$	<b>0.02*</b>	<b>0.02*</b>	<b>0.02*</b>	<b>0.02*</b>
$N$	<b>8*</b>	<b>10*</b>	<b>9*</b>	<b>10*</b>
$B$ [ $10^{-4}\text{cm}^{-1}$ ]	<b>4.5 ± 0.1</b>	<b>4.3 ± 0.1</b>	<b>4.1 ± 0.1</b>	<b>4.2 ± 0.1</b>
$r$ [Å]	<b>24.1 (fixed)</b>			
$\xi$ [Å]	<b>50 ± 60</b>	<b>101 ± 53</b>	<b>90 ± 53</b>	<b>93 ± 45</b>
$D$	<b>2.7±0.3</b>	<b>1.85 ± 0.3</b>	<b>2.0 ±0.3</b>	<b>2.0 ±0.3</b>
$\alpha$	<b>Only 1 structure</b>			<b>0.50 ± 0.13</b>
<b>Derived parameters</b>				
$R_g$ [Å]	<b>112 ± 105</b>	<b>164 ± 89</b>	<b>155 ± 81</b>	<b>161 ± 81</b>
<b>Goodness of fit</b>				
$\chi^2_{reduced}$	<b>1.31</b>	<b>1.02</b>	<b>1.02</b>	<b>0.94</b>

**Table ST4 Parameters obtained from the fitting of SERCA. The errors are determined by profile likelihood. \* $N$  and  $\gamma$  were ill-defined due to high correlation, and errors could therefore not be determined.**

The structural state of the sample was probed in the high- $q$  part of the scattering curve, whereas the contribution from the fractal aggregates dominated the low- $q$  part of the curve. Therefore, instead of using the experimental errors as weights in the  $\chi^2$ -minimization, we used  $\sigma=10\%/l(q)$ , to give more weight to the high- $q$  part.

The aggregation had negligible effect of the refined equilibrium between the E1 and the E2 state (50% E1 and 50% E2). The fractal aggregates were built up of SERCA molecules, with the same fraction between the E1 and the E2 as for the single proteins, as seen from equation (3). That is, 50% of the aggregates consisted of SERCA E1 and 50% of SERCA E2. Therefore, the scattering from the aggregates at high- $q$  was practically identical to the corresponding scattering from the single proteins and the degree of aggregation did thus not effect the

refined equilibrium. To test the independency from the aggregates, a truncated dataset ( $q > 0.06 \text{ \AA}^{-1}$ ) was fitted with a model without the fractal aggregates, resulting in  $50 \pm 13\%$  E2 and  $50 \pm 13\%$  E1, thus confirming the result. If one of the two states had been substantially larger than the other, e.g. an equilibrium between a monomer and a dimer, then the result would be more effected by the degree of aggregation. However, this is not the case here.

Similarly, the equilibrium between the E1 and the E2 state did virtually not affect the parameters associated with the fractal aggregate ( $\gamma, N, \xi, \text{ and } D$ ), since the low- $q$  scattering from the E1 and E2 state were virtually identical. The form factors were calculated from atomic crystal structures by first calculating the pair distance distribution function  $p(r)$ , taking the corresponding Fourier transform and normalizing the result. This approach was utilized to gain calculation speed without decreasing the resolution of the model by coarse-graining.

## 5. References

- 1 Israelachvili JN (2011) *Intermolecular and Surface Forces*, 3rd editio Ademic Press - Elsevier.
- 2 Sears VF (1992) Neutron scattering lengths and cross sections. *Neutron News* **3**, 26–37.
- 3 Midtgaard SR, Pedersen MC, Kirkensgaard JJK, Sørensen KK, Mortensen K, Jensen KJ & Arleth L (2014) Self-assembling peptides form nanodiscs that stabilize membrane proteins. *Soft Matter* **10**, 738–52.
- 4 Oesterhelt D & Stoeckenius W (1974) Isolation of the cell membrane of *Halobacterium halobium* and its fractionation into red and purple membrane. *Methods Enzymol.* **31**, 667–78.
- 5 Dencher N (1982) Preparation and properties of monomeric bacteriorhodopsin. *Methods Enzymol.* **88**, 5–10.
- 6 Jensen PE, Rosgaard L, Knoetzel J & Scheller HV (2002) Photosystem I activity is increased in the absence of the PSI-G subunit. *J. Biol. Chem.* **277**, 2798–2803.
- 7 Lichtenthaler HK (1987) [34] Chlorophylls and carotenoids: Pigments of photosynthetic biomembranes. *Methods Enzymol.* **148**, 350–382.
- 8 Blaise M & Thirup S (2011) Crystallization of *Escherichia coli* maltoporin in the trigonal space group R3. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **67**, 114–116.
- 9 Sobolevsky AI, Rosconi MP & Gouaux E (2009) X-ray structure, symmetry and mechanism of an AMPA-subtype glutamate receptor. *Nature* **462**, 745–756.
- 10 Goehring A, Lee C, Wang KH, Michel JC, Claxton DP, Bacongus I, Althoff T, Fischer S, Garcia KC & Gouaux E (2014) Screening and large-scale expression of membrane proteins in mammalian cells for structural studies. *Nat. Protoc.* **9**, 2574–2585.
- 11 Sørensen TL-M, Olesen C, Jensen A-ML, Møller JV & Nissen P (2006) Crystals of sarcoplasmic reticulum Ca(2+)-ATPase. *J. Biotechnol.* **124**, 704–716.
- 12 Feoktystov A V., Frielinghaus H, Di Z, Jaksch S, Pipich V, Appavou MS, Babcock E, Hanslik R, Engels R, Kemmerling G, Kleines H, Ioffe A, Richter D & Brückel T (2015) KWS-1 high-resolution small-angle neutron scattering instrument at JCNS: Current state. *J. Appl. Crystallogr.* **48**, 61–70.
- 13 Gilbert EP, Schulz JC & Noakes TJ (2006) “Quokka”-the small-angle neutron scattering instrument at OPAL. *Phys. B Condens. Matter* **385–386**, 1180–1182.
- 14 Kline SR (2006) Reduction and analysis of SANS and USANS data using IGOR Pro. *J. Appl. Crystallogr.* **39**, 895–900.
- 15 Hansen S (2012) BayesApp : a web site for indirect transformation of small-angle scattering data. *J. Appl. Crystallogr.* **45**, 566–567.
- 16 Franke D & Svergun DI (2009) DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *J. Appl. Crystallogr.* **42**, 342–346.
- 17 Volkov V V. & Svergun DI (2003) Uniqueness of ab initio shape determination in small-angle scattering. *J. Appl. Crystallogr.* **36**, 860–864.
- 18 Adams PD, Afonine P V., Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC & Zwart PH (2010) PHENIX: A comprehensive Python-based system for macromolecular structure solution.

- Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 213–221.
- 19 Lomize MA, Lomize AL, Pogozheva ID & Mosberg HI (2006) OPM: Orientations of proteins in membranes database. *Bioinformatics* **22**, 623–625.
- 20 Kynde SAR, Skar-Gislinge N, Pedersen MC, Midtgaard SR, Simonsen JB, Schweins R, Mortensen K & Arleth L (2014) Small-angle scattering gives direct structural information about a membrane protein inside a lipid environment. *Acta Crystallogr. D. Biol. Crystallogr.* **70**, 371–83.
- 21 Svergun D, Barberato C & Koch MHJ (1995) CRY SOL – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J. Appl. Crystallogr.* **28**, 768–773.
- 22 Pedersen MC, Arleth L & Mortensen K (2013) WillItFit: A framework for fitting of constrained models to small-angle scattering data. *J. Appl. Crystallogr.* **46**, 1894–1898.
- 23 Malik L, Nygaard J, Hoiberg-nielsen R, Arleth L, Hoeg-jensen T & Jensen KJ (2012) Perfluoroalkyl Chains Direct Novel Self-Assembly of Insulin. , 593–603.
- 24 Kotlarchyk M & Chen SH (1983) Analysis of small angle neutron scattering spectra from polydisperse interacting colloids. *J. Chem. Phys.* **79**, 2461.
- 25 Høiberg-Nielsen R, Westh P, Skov LK & Arleth L (2009) Interrelationship of steric stabilization and self-crowding of a glycosylated protein. *Biophys. J.* **97**, 1445–1453.
- 26 Teixeira J (1988) Small-angle scattering by fractal systems. *J. Appl. Crystallogr.* **21**, 781–785.
- 27 Fraser RDB, MacRae TP & Suzuki E (1978) An improved method for calculating the contribution of solvent to the X-ray diffraction pattern of biological molecules. *J. Appl. Crystallogr.* **11**, 693–694.

## 9.4 Paper IV: Small-angle neutron scattering studies on the AMPA receptor GluA2 in the resting, AMPA and GYKI-53655 bound states

**List of Authors** Andreas Haahr Larsen, Jerzy Dorosz, Thor Seneca Thorsen, Nicolai Tidemand Johansen, Tamim Darwish, Søren Roi Midtgaard, Lise Arleth and Jette Sandholm Kastrup

**Status** Accepted in iUCrJ, 2018, ??, ??-??. Latest version attached with final changes highlighted.

**Abstract** The AMPA receptor GluA2 belongs to the family of ionotropic glutamate receptors, which are responsible for most of the fast excitatory neuronal signaling in the central nervous system. These receptors are important for memory and learning, but have also been associated with brain diseases such as Alzheimer's disease and epilepsy. Today, one drug is on the market for treatment of epilepsy targeting AMPA receptors, i.e. a negative allosteric modulator of these receptors. Recently, crystal structures and cryo-electron microscopy (cryo-EM) structures of full-length GluA2 in the resting (apo), activated and desensitized states were reported. Here, solution structures of full-length GluA2 is reported, using small-angle neutron scattering (SANS) and a novel, fully matched out detergent. The GluA2 solution structure was investigated in the resting state as well as in the presence of AMPA and the negative allosteric modulator GYKI-53655. In solution and at neutral pH, the SANS data clearly indicate that GluA2 in the resting state is in a compact form. The solution structure resembles the crystal structure of GluA2 in the resting state, with estimated maximum distance ( $D_{max}$ ) of  $179 \pm 11$  Å and radius of gyration ( $R_g$ ) of  $61.9 \pm 0.4$  Å. An ab initio model of GluA2 in solution generated using the program DAMMIF clearly showed the individual domains, i.e. the extracellular N-terminal domains and ligand-binding domains as well as the transmembrane domain. In the presence of AMPA and GYKI-53655, respectively, the solution structures remain in a compact form, also when it is under conditions with no restraints on the dynamics of the protein. Only at acidic pH, GluA2 in the presence of AMPA adopts a more open conformation of the extracellular part (estimated  $D_{max}$  of  $189 \pm 5$  Å and  $R_g$  of  $65.2 \pm 0.5$  Å), resembling the most open, desensitized cryo-electron microscopy structure of GluA2 in the presence of quisqualate (class 3).

**Contributions from AHL** AHL did the SANS measurements together with TST, NTJ and SRM. AHL did the SANS analysis. JSK wrote the paper with contributions from LA and AHL.

**Supporting Information** Attached

## **Small-angle neutron scattering studies on the AMPA receptor GluA2 in the resting, AMPA and GYKI-53655 bound states**

Andreas Haahr Larsen,<sup>a</sup> Jerzy Dorosz,<sup>b</sup> Thor Seneca Thorsen,<sup>b</sup> Nicolai Tidemand Johansen,<sup>a</sup> Tamim Darwish,<sup>c</sup> Søren Roi Midtgaard,<sup>a</sup> Lise Arleth<sup>a,\*</sup> and Jette Sandholm Kastrup<sup>b,\*</sup>

<sup>a</sup> Structural Biophysics, X-ray and Neutron Science, The Niels Bohr Institute, University of Copenhagen, Denmark

<sup>b</sup> Biostructural Research, Department of Drug Design and Pharmacology, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

<sup>c</sup> National Deuteration Facility, Australian Nuclear Science and Technology Organization, Australia

\* Correspondent author: Jette Sandholm Kastrup: email: [jsk@sund.ku.dk](mailto:jsk@sund.ku.dk), phone: +45 35336486.  
Correspondent author on SANS: Lise Arleth: email: [arleth@nbi.ku.dk](mailto:arleth@nbi.ku.dk)

### **Synopsis**

In this study, the behavior of the detergent solubilized tetrameric, full-length ionotropic glutamate receptor GluA2 in solution is investigated using small-angle neutron scattering. **We find that** the GluA2 solution structure in the resting state as well as in the presence of AMPA and the negative allosteric modulator GYKI-53655 is preferentially in a compact form.

## Abstract

The AMPA receptor GluA2 belongs to the family of ionotropic glutamate receptors, which are responsible for most of the fast excitatory neuronal signaling in the central nervous system. These receptors are important for memory and learning, but have also been associated with brain diseases such as Alzheimer's disease and epilepsy. Today, one drug is on the market for treatment of epilepsy targeting AMPA receptors, *i.e.* a negative allosteric modulator of these receptors. Recently, crystal structures and cryo-electron microscopy (cryo-EM) structures of full-length GluA2 in the resting (apo), activated and desensitized states were reported. Here, solution structures of full-length GluA2 is reported, using small-angle neutron scattering (SANS) and **a novel**, fully matched out detergent. The GluA2 solution structure was investigated in the resting state as well as in the presence of AMPA and the negative allosteric modulator GYKI-53655. In solution and at neutral pH, the SANS data clearly indicate that GluA2 in the resting state is in a compact form. The solution structure resembles the crystal structure of GluA2 in the resting state, with estimated maximum distance ( $D_{max}$ ) of  $179 \pm 11$  Å and radius of gyration ( $R_g$ ) of  $61.9 \pm 0.4$  Å. An *ab initio* model of GluA2 in solution generated using the program DAMMIF clearly showed the individual domains, *i.e.* the extracellular *N*-terminal domains and ligand-binding domains as well as the transmembrane domain. In the presence of AMPA and GYKI-53655, respectively, the solution structures **remain** in a compact form, **also when it is under conditions with no restraints on the dynamics of the protein**. **Only at** acidic pH, GluA2 in the presence of AMPA adopts a more open conformation of the extracellular part (estimated  $D_{max}$  of  $189 \pm 5$  Å and  $R_g$  of  $65.2 \pm 0.5$  Å), resembling the most open, desensitized cryo-electron microscopy structure of GluA2 in the presence of quisqualate (class 3).

**Keywords:** ionotropic glutamate receptor; small-angle neutron scattering; agonist; negative allosteric modulator; resting state

## 1. Introduction

Glutamate is the major excitatory neurotransmitter in the central nervous system (CNS) and mediates its function through interaction with metabotropic G protein coupled receptors (mGluRs) and ionotropic glutamate receptors (iGluRs). Located in the cell membrane at the synapse, the iGluRs mediate fast synaptic transmission in the CNS and have an important role in memory and learning (Sachser *et al.*, 2017). However, these receptors have also been associated with brain diseases or disorders, *e.g.* epilepsy, Parkinson's disease, Alzheimer's disease, depression and stroke (Lee *et al.*, 2016). Therefore, the iGluRs are considered important targets for **intervention** by medicines. For example, the drug Memantine used for treatment of Alzheimer's disease and Perampanel used for treatment of epilepsy both target the iGluRs.

The members of the iGluR family have been divided into four classes: the  $\alpha$ -amino-3-hydroxy-5-methylisoxazole-4-propionate (AMPA) receptors, the kainate receptors, the *N*-methyl-D-aspartate (NMDA) receptors and the delta receptors (Traynelis *et al.*, 2010). iGluRs form tetrameric ion channels composed of either identical subunits (homomeric receptors) or different subunits (heteromeric receptors). The AMPA receptors consist of subunits GluA1-4, of which the GluA2 subunit is the most studied. The GluA2 subunit is composed of four domains (**Fig. 1A**): the extracellular *N*-terminal domain (NTD; also abbreviated the ATD) followed by the ligand-binding domain (LBD), the transmembrane domain (TMD) forming the ion channel and the cytosolic *C*-terminal domain (CTD; **not included in structures**).

The iGluRs have been shown to adopt various conformational states upon activation and inactivation (Fig. 1A). The first X-ray structure of a full-length homomeric GluA2 was published in 2009 (Sobolevsky *et al.*, 2009). This structure was of GluA2 with a competitive antagonist bound. Then followed structures of GluA2 in different states, *e.g.* with agonists and positive allosteric modulators (see *e.g.* Dürr *et al.*, 2014), with negative allosteric modulators (*e.g.* perampanel and GYKI-53655) (Yelshanskaya *et al.*, 2016) as well as GluA2 in the resting (apo) state (Dürr *et al.*, 2014; Yelshanskaya *et al.*, 2016). Recently, the first structures of GluA2 in the activated state and in the desensitized state (an inactive form of the receptor where glutamate is still bound) were reported using cryo-electron microscopy (cryo-EM) (Twomey *et al.*, 2017). To date, approximately 40 full-length GluA2 structures have been deposited in the Protein Data Bank (PDB; [www.rcsb.org](http://www.rcsb.org)), of which half were determined by X-ray crystallography (Sobolevsky *et al.*, 2009; Chen *et*



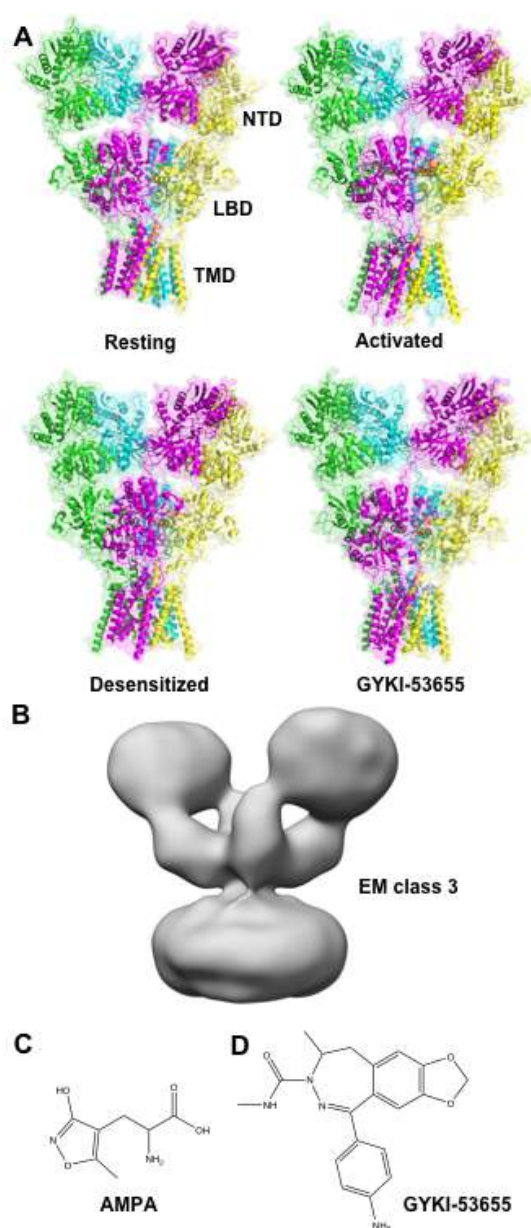


Figure 1. Structures. (A) GluA2 in the resting (apo) state (pdb-code 4u2p; Dürr *et al.*, 2014), the activated state (pdb-code 5weo; Twomey *et al.*, 2017), the desensitized state (pdb-code 5vhz; Twomey *et al.*, 2017) and with negative allosteric modulator bound (pdb-code 511h; Yelshanskaya *et al.*, 2016). The *N*-terminal domain (NTD), ligand-binding domain (LBD) and transmembrane domain (TMD) are indicated on the figure. Parts of the TMD have not been modelled in the structures. (B) The EM class 3 structure (EMDataBank EMD-2688; Meyerson *et al.*, 2014). (C) The agonist AMPA. (D) The negative allosteric modulator GYKI-53655.

*al.*, 2014; Dürr *et al.*, 2014; Yelshanskaya *et al.*, 2016) and the other half by electron microscopy (Meyerson *et al.*, 2014; Herguedas *et al.*, 2016; Meyerson *et al.*, 2016; Twomey *et al.*, 2016 and 2017; Zhao *et al.* 2016; Chen *et al.* 2017). As seen from Fig. 1A, the X-ray structures all represent compact and also rather similar structures, whereas structures with more open extracellular domains have been reported using cryo-EM, *e.g.* GluA2 in complex with the agonist quisqualate, class 3 (Fig 1B). The structure of GluA2 with quisqualate was considered to represent a desensitized state of the receptor (Meyerson *et al.*, 2014).

Here, we report small-angle neutron scattering (SANS) data on detergent solubilized full-length GluA2 (GluA2cryst with deletion of CTD; Sobolevsky *et al.*, 2009), using a novel match-out deuterated DDM (Midtgaard *et al.*, 2018). The match-out DDM ensures that only the GluA2 contributed to the measured SANS signal. GluA2 was investigated in its resting state, in the presence of the agonist AMPA (Fig. 1C) and with the negative allosteric modulator GYKI-53655 (Fig. 1D). We show that GluA2 in solution at neutral pH primarily adopts a tetrameric compact structure that resembles the X-ray and cryo-EM structures, whereas acidic pH leads to a more open structure. To our knowledge this is the first time that the structure of the full-length GluA2 is studied as a detergent solubilized protein in solution. Also it is the first time that the structural effect of AMPA binding at full-length GluA2 is studied.

## 2. Methods

### 2.1 Expression and purification of GluA2

The construct GluA2cryst was kindly provided by E. Gouaux (Sobolevsky *et al.*, 2009). The receptor was expressed in the HEK293F cell line and purified as previously described (Midtgaard *et al.*, 2018). In brief, the membranes were isolated by resuspending the cell pellet in buffer containing 20 mM Tris pH 8.0, 150 mM NaCl, 45 mM n-dodecyl- $\beta$ -D-maltopyranoside (DDM; Anatrace) and protease inhibitors (Roche). The supernatant was supplemented with 50 mM imidazole, mixed with TALON metal affinity resin (Clontech) and rotated overnight. The receptor was eluted using the same buffer but containing 250 mM imidazole and 1 mM DDM. GFP and His-tag were removed by adding thrombin (Sigma T4393) over night at 4 °C. Finally, the receptor was purified using size exclusion chromatography on a Superose 6 column (GE Healthcare) and the peak fractions flash frozen in liquid nitrogen.

## 2.2 Solvent and detergent exchange

Deuterated DDM was synthesized to match the scattering length density of D<sub>2</sub>O in both the detergent head and tail, as described by Midtgaard *et al.* (Midtgaard *et al.*, 2018). Purified GluA2 in H<sub>2</sub>O-based buffer (20 mM Tris pH 8.0, 150 mM NaCl, 1 mM DDM Anagrade (Anatrace) was applied to a Superose 6 10/300 GL (GE Helthcare) column equilibrated in D<sub>2</sub>O-based buffer (20 mM Tris/DCl pH 7.5, 100 mM NaCl) with 0.5 mM deuterated DDM to exchange solvent and detergent and obtain match-out conditions, where only the protein is visible. The exchange was performed at 5 °C with a flow rate of 0.25 ml/min to ensure full exchange (Midtgaard *et al.*, 2018).

## 2.3 Addition of ligands

The agonist AMPA was dissolved in the above mentioned D<sub>2</sub>O-based buffer to a 100 mM stock solution with a pH of 4.2. Stock solution was added to GluA2 in D<sub>2</sub>O-based buffer by gentle mixing in the SANS cuvettes to obtain one sample with 1 mM AMPA (pH 7.6) and one with 10 mM AMPA (pH 5.5). GYKI-53655 was dissolved in hydrogenated DMSO to a 100 mM stock solution, and added to GluA2 in the D<sub>2</sub>O-based buffer to get a sample of GluA2 with 1 mM GYKI-53655. The protein concentrations were measured to 0.20 mg/ml (0.54  $\mu$ M) for the apo sample, 0.31 mg/ml (0.84  $\mu$ M) for the 1 mM AMPA and the 1 mM GYKI-53655 samples, and 0.17 mg/ml (0.46  $\mu$ M) for the 10 mM AMPA sample. The concentrations were determined with UV absorption at 280 nm using a NanoDrop 1000 (Thermo Fisher Scientific) for the 1 mM AMPA and the 1 mM GYKI-53655 sample, using extinction coefficient of 519100 M<sup>-1</sup> cm<sup>-1</sup> as calculated from the construct sequence with ExPASy ProtParam. The concentrations of the apo sample and the 10 mM AMPA sample were determined with QuantiPro BCA assay (Sigma).

## 2.4. SANS data collection

SANS data were collected at the KWS1 SANS instrument at FRM-II, MLZ, Munich, using a neutron wavelength  $\lambda$  of 5.0 Å and a wavelength spread  $\Delta\lambda/\lambda$  of 10% (FWHM). Three instrumental settings were used, with collimation/sample detector distances of 4 m/1.5 m, 4 m/4 m and 8 m/8 m to cover a nominal  $q$ -range of 0.006-0.3 Å<sup>-1</sup>, where  $q = 4\pi \sin(\theta)/\lambda$  and  $\theta$  is half the scattering angle. Samples were measured in 2 mm Hellma quartz cuvettes at 10 °C. The data were reduced according to the standard procedures of the

beamline (Feoktystov *et al.*, 2015), *i.e.* azimuthally averaged, absolute calibrated with plexiglass as standard and background subtracted using the QtiKWS software to yield the reduced scattering intensity,  $I(q)$  in units of 1/cm. Due to parasitic scattering at high- $q$ , the data sets were truncated for  $q > 0.2 \text{ \AA}^{-1}$ . The overlap between the three settings was optimized by multiplying the data from the 4 m/1.5 m setting and the 8 m/8 m setting with factors close to unity. This optimization was performed with an implementation of the Indirect Fourier Transformation (IFT) method (Glatter, 1977) which allows for varying the multiplication factors to obtain the best fit to the data. The program was obtained from Jan Skov Pedersen, Aarhus University, Denmark.

## 2.5 Model for SANS data analysis

An initial inspection of the obtained data suggested the presence of small fractions of higher order oligomers of GluA2 in some of the samples, besides from the expected GluA2 tetramers. Models were therefore developed that allowed for including this effect and each data set was evaluated through fitting by the following four different models: Model 1 compared data directly with the theoretical scattering from relevant GluA2 atomic structures from the Protein Data Base, *i.e.* GluA2 in the resting state (pdb-code 4u2p; X-ray; Dürr *et al.*, 2014), the activated state (pdb-code 5weo; cryo-EM; Twomey *et al.* 2017), the desensitized state (pdb-code 5vzh; cryo-EM; Twomey *et al.* 2017), and in the negative allosteric modulator inhibited state (pdb-code 5l1h; X-ray; Yelshanskaya *et al.*, 2016) (Fig. 1A). In addition, data were compared to the GluA2 desensitized class 3 cryo-EM structure with quisqualate, determined at 22.9 Å resolution by Meyerson *et al.* (2014) and deposited in the EMDDataBank (EMD-2688) (Fig. 1B). In order to calculate the scattering from the EM class 3 structure, an approximate atomic model had to be generated. This was done by manually fitting an atomic structure of GluA2 in the desensitized state (pdb-code 5vzh) into the EM density map using Chimera (Pettersen *et al.*, 2004) to generate an EM class 3 atomic structure. The model contains no detergents and is thus directly comparable with the obtained SANS data. This approximate atomic structure was then used for calculating the theoretical scattering from the EM class 3 state. Model 2 was a linear combination of single tetrameric GluA2 in one of the aforementioned states and random oligomers of tetrameric GluA2 in the same state. For the description of the oligomers, the GluA2 tetramers were assumed to be randomly oriented with respect to each other and could thus be modelled as mass fractals (as described below). Model 3 was a linear combination of scattering from one of the atomic structures from model 1 and

the GluA2 desensitized class 3 EM structure. Model 4 was a linear combination of the scattering from the atomic structures, of fractal oligomers and of the generated class 3 EM atomic structure, *i.e.* a combination of models 2 and 3.

The models were assessed using the F-test, based on the reduced  $\chi^2$  values, with a significance criteria of  $P < 5\%$ . The reduced  $\chi^2$  is defined as  $\chi_r^2 = \chi^2/f$ , where  $f$  is the number of degrees of freedom, given in terms of the number of data points  $N$  and the number of model parameters  $K$  as  $f = N - K$ .  $\chi^2$  is defined as

$$\chi^2 = \sum_{i=1}^N \frac{(I_i^{fit} - I_i^{exp})^2}{\sigma_i^2}$$

where  $I_i^{exp}$  and  $\sigma_i$  are the  $i$ 'th experimental intensity and error respectively, and  $I_i^{fit}$  is the  $i$ 'th intensity from the model fit. Residual plots are also shown to ease the visual comparison of data and fit with  $(\Delta I/\sigma)_i = (I_i^{fit} - I_i^{exp})/\sigma_i$ .

**Model 1:** The scattering intensity was given in terms of the prefactor  $C$ , the background  $B$  and the form factor  $P(q)$ :

$$I(q)_{M1} = C \cdot P(q) + B \quad (1)$$

The form factor  $P(q)$  was calculated directly from the relevant atomic structures in PDB-format using the in-house developed software CaPP which is adapted for membrane proteins (see further details under model implementation). The prefactor was given as  $C = K \cdot n \cdot \Delta b^2$ , where  $K$  is a correction factor for the protein concentration measurement,  $n$  is the molar concentration,  $\Delta b$  is the excess scattering length of the protein.  $K$  and  $B$  were fitted,  $n$  was measured by UV absorption, and the  $\Delta b$  was found by summing up atomic scattering lengths (provided *e.g.* by NIST, <https://www.ncnr.nist.gov/resources/n-lengths/>) and subtracting the total scattering length of the corresponding excluded water volume.

**Model 2:** The random oligomers were described as mass fractals of GluA2 using a previously developed approach (Malik *et al.*, 2011), which applies the structure factor for fractals of spherical subunits derived by Teixeira (Teixeira, 1988):

$$S(q) = 1 + \frac{1}{(qr)^D} \cdot \frac{D \cdot \Gamma(D-1)}{\left[1 + \frac{1}{(q\xi)^2}\right]^{\frac{D-1}{2}}} \cdot \sin[(D-1) \cdot \text{atan}(q\xi)]$$

where  $\Gamma$  is the gamma function,  $D$  is the dimensionality of the fractal ( $1 < D < 3$ ),  $r$  is the mean distance between the fractal subunits, and  $\xi$  is the correlation length of the fractal oligomers, which is directly related to the radius of gyration  $R_g$  of the oligomers (Teixeira, 1988):

$$R_g^2 = \frac{D(D+1)}{2} \cdot \xi^2$$

The GluA2 tetrameric subunits were assumed to be randomly oriented with respect to each other in the fractal oligomer. This was taken into account by the decoupling approximation (Kotlarchyk & Chen, 1983):

$$S'(q) = 1 + \beta(q) \cdot [S(q) - 1]$$

with  $\beta(q) = \langle \psi(q) \rangle^2 / \langle \psi(q)^2 \rangle$ , where  $\psi(q)$  is the form factor amplitude and the brackets  $\langle \dots \rangle$  denote the orientational averaging. As discussed in (Høiberg-Nielsen *et al.*, 2009) this can be rewritten into:

$$\beta(q) = \frac{(A_{m=0}^{l=0})^2}{P(q)}$$

where  $A_0^0$  is the amplitude corresponding to the 0th order spherical harmonics expansion of  $\psi(q)$  (Svergun *et al.*, 1995) which was calculated from the atomic structures. The intensity from one fractal oligomer could then be expressed as a product of  $P(q)$  and  $S'(q)$ .

$$I(q)_{agg} \propto P(q) \cdot S'(q)$$

The scattering intensity of model 2 was a linear combination of the scattering from single proteins and fractal oligomers, with  $\gamma$  denoting the fraction of GluA2 molecules in oligomeric form, and  $N$  denoting the number of GluA2 molecules per oligomer. The intensity could then be written as

$$I(q)_{M2} = C \cdot [\gamma \cdot N \cdot S'(q) + (1 - \gamma)] \cdot P(q) + B = C \cdot S(q)_{eff} \cdot P(q) + B \quad (2)$$

where  $S(q)_{eff}$  is an effective structure factor for the linear combination.  $N$  is related to  $R_g$ ,  $D$  and  $r$  by the fractal scaling relationship (Sorensen & Roberts, 1996).

$$N = k \left( \frac{R_g}{r} \right)^D$$

where  $k$  is the structural coefficient. In this study, we fixed  $D$  to 2, since the information in the data about the fractals was limited. According to Sorensen and Roberts (Sorensen & Roberts, 1996),  $k \approx 1$  when  $D \approx 2$ , so  $k$  was furthermore fixed to unity. The mean inter-molecular distance  $r$  of the fractal oligomer was also fixed to equal the radius of a sphere fulfilling  $V_{sph} = V_{GluA2}$ , where  $V_{GluA2}$  was calculated from the protein sequence as a sum of the atomic van der Waals volumes. Hence, the model had a total of four free fitting parameters,  $K$ ,  $B$ ,  $R_g$  and  $\gamma$ , where the last two were directly related to the fractal oligomer. Each data set had 3-5 so-called “good parameters” (Vestergaard & Hansen, 2006), making it possible to determine well all four model parameters.

**Model 3:** It was investigated whether the samples were in an equilibrium between two structural states, namely the atomic structures deposited in the PDB and the class 3 EM structure deposited in the EMDataBank. Denoting by  $\alpha$  the fraction of the intensity coming from the class 3 EM state, the intensity could be expressed as:

$$I(q)_{M3} = C \cdot [\alpha \cdot P(q)_{EM} + (1 - \alpha) \cdot P(q)_{atm}] + B \quad (3)$$

where  $P(q)_{EM}$  and  $P(q)_{atm}$  are the form factors for the class 3 EM structure and one of the atomic structures, respectively. Model 3 had the three parameters  $K$ ,  $B$  and  $\alpha$ .

**Model 4:** The fourth model had four contributions to the scattering: GluA2 in one of the atomic structures, GluA2 in the class 3 EM state and fractal oligomers of, respectively, one of the atomic structures and the class 3 EM structure. The intensity was given as:

$$I(q)_{M4} = C \cdot [\alpha \cdot S(q)_{eff,EM} \cdot P(q)_{EM} + (1 - \alpha) \cdot S(q)_{eff,atm} \cdot P(q)_{atm}] + B \quad (4)$$

with  $S(q)_{eff}$  given as in equation (2). Model 4 had five free parameters, namely  $K, B, R_g, \gamma$  and  $\alpha$ .

**Model implementation:** The models were implemented in WillItFit (Pedersen *et al.*, 2013). Resolution effects were included in the modelling using the resolution function,  $\sigma_q(q)$  provided by the beamline in the fourth column of the SANS data. The GluA2 form factors were calculated using the in-house developed C/Python software program CaPP (source code freely available at <https://github.com/Niels-Bohr-Institute-XNS-StructBiophys/CaPP>). CaPP is adapted for membrane proteins and allows for including a hydration layer to only the water exposed part of the membrane protein surface and, in the case of SANS studies, exchanges the scattering length of exchangeable hydrogens to the average H/D scattering length relevant for the sample. The hydration layer is included as a single layer of water molecules with a density 10% higher than that of bulk water, in accordance with (Svergun *et al.*, 1995). The layer is represented by dummy beads, each corresponding to 4.13 water molecules and added to the surface of the protein, except in the region embedded by the core of the DDM micelle. In CaPP, the thickness and orientation of the water depleted layer is determined either using the database Orientation of Proteins in Membranes ([opm.phar.umich.edu](http://opm.phar.umich.edu); Lomize *et al.*, 2006) or defined manually by the user.

## 2.6 Experimental pair distance distribution functions

The experimental  $p(r)$  functions were calculated by Bayesian indirect Fourier transformations (BIFT) as implemented in BayesApp ([www.bayesapp.org](http://www.bayesapp.org); Hansen, 2012). Backgrounds were fitted, and for some data sets and fits, the regularization parameter and  $D_{max}$  values were varied manually in the proximity of the automatically determined values, to obtain  $p(r)$  functions with a smooth decay to zero at  $D_{max}$  and a sensible smoothness. For the data sets with GluA2 in the presence of AMPA at pH 7.6 and GYKI-53655, respectively, slight aggregation was seen. In these cases, a  $p(r)$  function was also calculated for a low- $q$  truncated data set in order to limit the effect of aggregation on the refined  $p(r)$  function. The data set with GluA2 in the presence of AMPA at pH 7.5 was truncated after the first four points ( $q_{min} = 0.011 \text{ \AA}^{-1}$ ), the data set with GluA2 and AMPA at pH 5.5 was truncated after 14 points ( $q_{min} = 0.019 \text{ \AA}^{-1}$ ), and with GYKI-53655 after five points ( $q_{min} = 0.012 \text{ \AA}^{-1}$ ). These were the minimum number of data points necessary to get rid of the “tail” in the  $p(r)$  function for large  $r$  values. The given experimental values for the radius of gyration  $R_g$ , maximal distances in the particle  $D_{max}$  and forward scattering  $I(0)$



(Supplementary Table S1) were estimated from the truncated data sets, since the non-truncated values were influenced by aggregations and thus gave large, and irrelevant, values for  $R_g$  and  $D_{max}$ , not comparable with those based on atomic GluA2 structures deposited in the PDB.  $R_g$  and  $I(0)$  were also determined using Guinier analysis (Supplementary Fig. S1). The dataset for GluA2 AMPA bound state at pH 5.5, had no valid Guinier region fulfilling  $qR_g < 1.3$  (Supplementary Fig. S1 C), due to the oligomeric contribution. This led to overestimation of  $R_g$  and  $I(0)$  from the Guinier analysis. For the three other samples, the  $R_g$  and  $I(0)$  values from the Guinier analysis were consistent with those from the  $p(r)$  (Supplementary Table S1). In the manuscript we refer to the  $R_g$  values from the  $p(r)$ . Likewise, the  $R_g$  and  $I(0)$  values from the  $p(r)$  were used for  $M_w$  determination.

## 2.7 Theoretical pair distance distribution functions

The theoretical  $p(r)$  functions were calculated directly from the atomic structures deposited in the PDB, using CaPP. First, a hydration shell was added to the structure as described in Section 2.5, then the  $p(r)$  functions were calculated using the positions and scattering length of the atoms and water beads. The calculated  $p(r)$  functions had a slowly decreasing asymptotic behavior for large pair-distances,  $r$ , because every single atom was included in the calculation. This resulted in a  $D_{max}$  much larger than the experimental, where the far most distances were not detectable. Therefore, in order to obtain a  $D_{max}$  that could be compared directly with experiments, the theoretical  $D_{max}$  values were calculated with a 1% threshold, *i.e.* the  $D_{max}$  was defined as the first  $r$  where  $p(r)$  had decreased to 1% of its maximal value.

## 2.8 *Ab initio* modelling

Since the scattering from DDM was eliminated by deuteration, data analysis tools usually only applicable for soluble proteins (without detergents) could be applied. *Ab initio* modelling was performed using DAMMIF (Franke & Svergun, 2009). The only input was a pair distance distribution function,  $p(r)$ , which was calculated with DATGNOM (Petoukhov *et al.*, 2007) to obtain the right input data format for DAMMIF. No symmetry was assumed and DAMMIF was run 10 times. Alignment, clustering, selection, averaging and filtering of the 10 runs were performed using the automatic algorithm provided in the ATSAS online framework (Franke *et al.*, 2017).

## 2.8 $M_W$ determination for assessment of oligomeric state

The oligomeric state was assessed by comparing the  $M_W$  found from SANS with that of the construct (GluA2cryst; 368 kDa). The  $M_W$  was found from the  $I(0)$  and the concentration ( $c$ ), as well as the average excess scattering length density ( $\Delta\rho$ ) calculated from the sequence and the average protein density ( $\rho_p = 1.37 \text{ g/cm}^3$ ), by  $M_W = (I(0)/c) \cdot (N_A \rho_p^2 / \Delta\rho^2)$ , where  $N_A$  is Avogadro's number.  $M_W$  was determined to be close to the expected (368 kDa) for GluA2 in the AMPA bound state at pH 7.5 (347 kDa) and GluA2 in the GYKI bound state (347 kDa). The determined values of  $M_W$  were, however, unrealistically low for GluA2 in the resting state (220 kDa) and GluA2 in the AMPA bound state at pH 5.5 (240 kDa). The discrepancies from the expected value of 368 kDa may reflect the uncertainty in the concentration measurement of these proteins. As an alternative approach, that is independent of protein concentration measurements, the  $M_W$  was determined from the scattering invariant  $Q$  (Porod, 1982) using the method described by Fischer *et al.* (Fischer *et al.*, 2010; Supplementary Tables S1 and S2) and the method described by Petoukhov *et al.* (Petoukhov *et al.*, 2012; Supplementary Table S2). Constant backgrounds were determined using Porod plots (Supplementary Fig. S2) when calculating the value for  $Q$ . As the Fischer method takes the size of the protein into account, it is more precise for the large GluA2 tetrameric protein (368 kDa) than the Petoukhov method (Fischer *et al.*, 2010). Therefore, the  $M_W$  values obtained with the Fischer method are used throughout the main text and reported in Table 1. All values can however be seen in Supplementary Tables S1 and S2.

## 3. Results

The AMPA receptor GluA2 from rat with deletion of the disordered intracellular C-terminal domain (GluA2cryst; Sobolevsky *et al.*, 2009) was used for investigating receptor conformations in solution in the resting state (apo), in the presence of the agonist AMPA and in the presence of the negative allosteric modulator GYKI-53655, by use of SANS and fully matched out detergent.

### 3.1 GluA2 in the resting state

SANS data for GluA2 in conditions corresponding to the resting state were obtained in the  $q$ -range from  $0.006 \text{ \AA}^{-1}$  to  $0.2 \text{ \AA}^{-1}$ . This data set was also shown in a recent publication about the contrast optimized

detergents (Midtgaard *et al.*, 2018) and we showed that the solution structure of GluA2 resembles the X-ray crystal structure. Here, we analyze the data in detail. A flat low- $q$  region was observed as well as no indications of scattering from the detergent molecules around the transmembrane part of the receptor or of free detergent micelles in the data (Midtgaard *et al.*, 2018). The Fischer analysis yielded a  $M_W$  of  $396 \pm 52$  kDa, which should be compared with the expected  $M_W$  of the construct (GluA2cryst; 368 kDa) (Table 1). This indicated that the protein was in the expected tetrameric state, in line with other studies (Dürr *et al.*, 2014; Yelshanskaya *et al.*, 2016). Kratky plot of the data shows that the protein is partially or fully folded (Supplementary Fig. S3).

The GluA2 SANS data were then compared with the 3.2 Å resolution X-ray crystal structure of GluA2 in the resting state (pdb-code 4u2p; Dürr *et al.*, 2014) by fitting of model 1. This crystal structure of GluA2 in the resting state represents a compact structure (Fig. 1A). The structure fitted well with experimental data (with goodness of fit  $\chi_r^2 = 4.7$ ; Fig. 2A), confirming that the crystal structure was essentially maintained in solution. The desensitized structure of GluA2 with quisqualate (class 3; EMD-2688; Meyerson *et al.*, 2014) was also incorporated in model 1 and fitted to the experimental data as it is representing a structure with the extracellular domains in a more open form (Fig. 1B). This structure did clearly not fit as well ( $\chi_r^2 = 12.9$ ) as the crystal structure of GluA2 in the resting state (Fig. 2A). This was also confirmed using an F-test, showing that the difference in the goodness of fit between the resting state and the class 3 EM structure was significant, as the P-value was below the significance level ( $P = 0.0001\%$ , significance level 5%).

Table 1. Maximal distance ( $D_{max}$ ) and radius of gyration ( $R_g$ ) as determined from the experimental  $p(r)$  functions ( $D_{max,SANS}$  and  $R_{g,SANS}$ ), and from the theoretical  $p(r)$  functions ( $D_{max,THE}$  and  $R_{g,THE}$ ) from the structures. Molecular weight ( $M_W$ ) based on solution SANS data ( $M_{W,SANS}$ ) determined with Fischer analysis (Fischer *et al.*, 2010) and  $M_W$  calculated from the sequence of the construct ( $M_{W,CON}$ ), and from the sequence of the crystal and cryo-EM structures ( $M_{W,STR}$ ).

SANS data	$D_{max,SANS}$ [Å]	$R_{g,SANS}$ [Å]	$M_{W,SANS}$ [kDa]	$M_{W,CON}$ [kDa]
SANS, Resting	$179 \pm 11$	$61.9 \pm 0.4$	396	368
SANS, AMPA bound (pH 7.5)	$184 \pm 11$	$61.0 \pm 0.6$	379	368

SANS, AMPA bound (pH 5.5)	$189 \pm 5$	$65.2 \pm 0.5$	442	368
SANS, GYKI bound	$186 \pm 5$	$62.1 \pm 0.3$	373	368
<b>Structures</b>	$D_{max,TH}$ [Å]	$R_{g,TH}$ [Å]		$M_{W,STR}$ [kDa]
X-ray, resting (pdb-code 4u2p)	171.0	56.1	---	369
EM, active (pdb-code 5weo)	175.0	58.7	---	366
EM, desensitized (pdb-code 5vzh)	167.0	55.8	---	366
EM, class 3 (EMDB-2688)	179.0	64.1	---	372
X-ray, GYKI bound (pdb-code 5l1h)	169.5	57.3	---	359

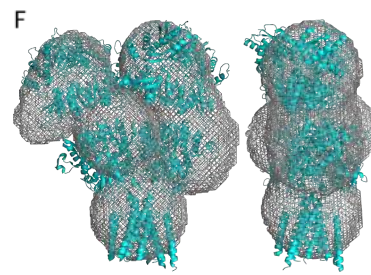
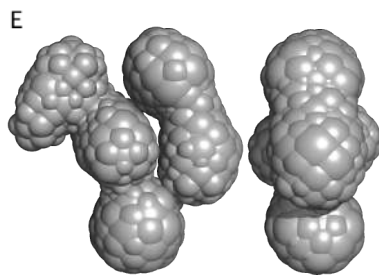
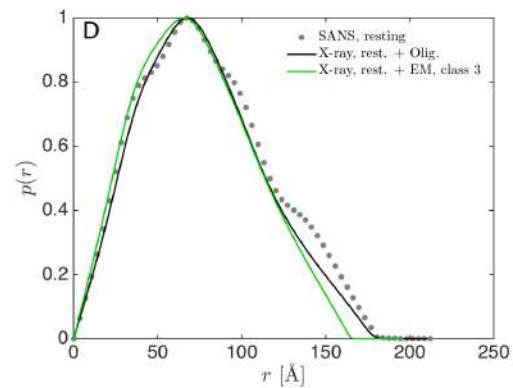
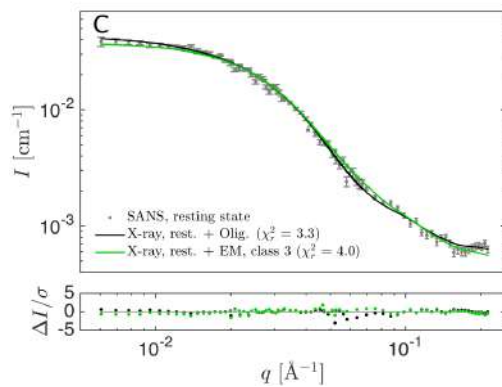
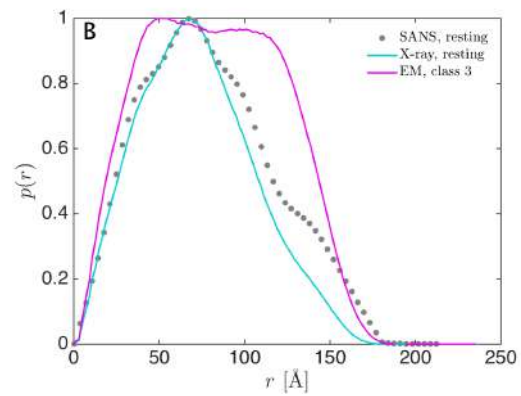
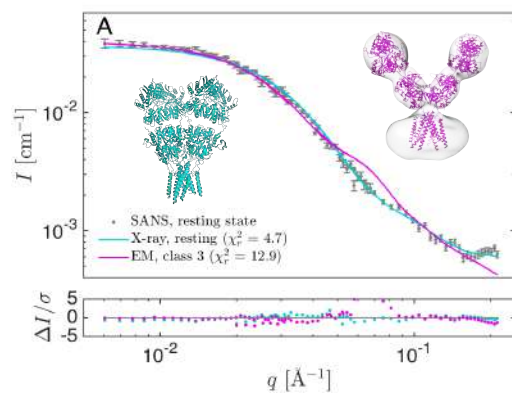


Figure 2. SANS data of GluA2 in the resting state (apo). (A) Experimental data (grey points), the resulting fits and residual plots. Fitted with the X-ray crystal structure of GluA2 in the resting state (pdb-code 4u2p; cyan; Dürr *et al.*, 2014) and the EM structure of GluA2 with quisqualate (class 3; EMD-2688; magenta; Meyerson *et al.*, 2014). A cartoon representation of the crystal structure of GluA2 is shown in cyan and an atomic model fitted into the class 3 EM structure in magenta. (B) Pair distance distribution function ( $p(r)$ ) for the experimental GluA2 data in solution (grey points) and theoretical  $p(r)$  functions for the GluA2 crystal structure in the resting state (cyan) and for the GluA2 class 3 EM structure (magenta). (C) Experimental SANS data and the fit of a linear combination of the crystal structure of GluA2 in the resting state and fractal oligomers (black), and a linear combination of the crystal structure and the class 3 EM structure (green). (D) Pair distance distribution functions for the fits. (E) *Ab initio* model generated with DAMMIF. The size of the beads was weighted by occupancy. (F) The GluA2 crystal structure in the resting state (pdb-code 4u2p; cyan) was manually overlaid with the DAMMIF envelope (grey).

The  $p(r)$  function of the GluA2 SANS data was compared to the theoretical  $p(r)$  functions calculated from the compact crystal structure of GluA2 in the resting state as well as from the more open class 3 EM structure of GluA2. A plot of experimental data and theoretical curves is displayed in Fig. 2B. The  $p(r)$  function for the solution GluA2 data had no tail, *i.e.* no indication of oligomerization or aggregation. However, the experimental  $p(r)$  function differed from the crystal structure for the resting state by lying above the theoretical function for all distances above  $r \sim 100$  Å. It also had a slightly larger maximal distance ( $D_{max}$ ) of  $179 \pm 11$  Å compared to the  $D_{max}$  of 171 Å based on the crystal structure (Table 1). Also, the radius of gyration ( $R_g$ ) was larger for the solution structure ( $61.9 \pm 0.4$  Å) compared to the crystal structure (56.1 Å). However, it should be noted that the  $D_{max}$  and  $R_g$  for the crystal structures are presumably underestimated due to parts in the TMD not being modelled (Supplementary Fig. S4).

Next, a linear combination of the crystal structure and fractal oligomers, model 2, was fitted to the experimental data, resulting in an even better fit to the data with  $\chi_r^2 = 3.3$  (linear combination) compared to  $\chi_r^2 = 4.7$  (atomic crystal structure alone). The improvement is minor, but significant (F-test:  $P = 4.7\%$ , significance level 5%). The fractal oligomers amount to  $0.9 \pm 6.5\%$ . Note that the amount of fractal oligomers is strongly correlated with the  $R_g$  of the oligomers (Supplementary Table S3), and thus poorly determined (large uncertainty). However, this correlation did not affect the refined values of the remaining

model parameters, nor the goodness of fit. As expected, the inclusion of fractal oligomers improved the fit in the low- $q$  region, for  $q < 0.02 \text{ \AA}^{-1}$  (Fig. 2C). A linear combination of the resting state and the class 3 EM structure, (model 3, Fig. 2C;  $\chi_r^2 = 4.0$ ) did not fit significantly better than the crystal structure alone (F-test:  $P = 22\%$ ). Model 4 in which both the crystal structure, fractal oligomers and the class 3 EM structure were included, resulted in the best goodness of fit ( $\chi_r^2 = 2.5$ ), but was not statistically better than model 2 (F-test:  $P = 10.0\%$ ). This suggests that besides the compact structure of GluA2, species with larger dimensions than the X-ray structure of GluA2 in the resting state are present in solution (Fig. 2C). On the other hand, there is no significant evidence for the presence of a more open conformation like the EM class 3 structure.

An *ab initio* structure was generated using DAMMIF (Fig. 2E). The *ab initio* bead model clearly showed the transmembrane domain and indicated a dimeric arrangement of the ligand-binding domains and of the *N*-terminal domains. The *ab initio* model is similar to the X-ray structure of GluA2 in the resting state but clearly more asymmetric (Fig. 2F), especially at the NTD level.

### 3.2 GluA2 in the presence of AMPA

SANS data were also collected on GluA2 in the presence of AMPA (Fig. 3A). As no X-ray crystal or EM structure is available of GluA2 with AMPA, we investigated the data by fitting three different structures: the crystal structure of GluA2 in the resting state, the recently reported structure of GluA2 in the activated state (cryo-EM structure of GluA2 as a complex bound to glutamate, cyclothiazide and stargazin in digitonin; pdb-code 5weo; Twomey *et al.* 2017) and the cryo-EM structure of GluA2 in the desensitized state (bound to L-quisqualate and germ cell-specific gene 1-like protein; pdb-code 5vhz; Twomey *et al.* 2017), see Fig. 1A.

The experimental  $M_w$  based on the SANS data was found to be  $379 \pm 49 \text{ kDa}$  as obtained from Fischer analysis. This is close to the expected  $M_w$  of  $368 \text{ kDa}$  for the construct (Table 1), and consistent with the protein being in a tetrameric state. As for the resting state, the Kratky plot of the data shows that the protein was folded or partially folded (Supplementary Fig. S3). The best fit was obtained with the structure of GluA2 in the activated state ( $\chi_r^2 = 13.6$ ) (Fig. 3A,B). The resting state gave a similar fit, although with slightly worse goodness of fit ( $\chi_r^2 = 16.9$ ). However, the goodness of fit was not significantly different

between the two structures ( $P = 12.4\%$ ). We also fitted structures of GluA2 in the desensitized state (pdb-code 5vhz and class 3 EM), which resulted in **even worse** fits with  $\chi_r^2 = 29.1$  and  $\chi_r^2 = 52.8$ , respectively.

Next, a linear combination of the four structures and fractal oligomers was fitted to the experimental data (model 2; Fig. 3C,D). Inclusion of fractal oligomers resulted in a marked improvement of the goodness of fit for the activated state (pdb-code 5weo;  $P = 0.0001\%$ ), desensitized state (pdb-code 5vhz;  $P = 0.000005\%$ ) and resting state (pdb-code 4u2p;  $P = 0.7\%$ ) with  $\chi_r^2$  of 5.0, 5.8 and 5.9, respectively. However, it was not possible to distinguish among these fits, in agreement with that the structures of GluA2 in the resting, activated and desensitized states are similar with  $D_{max}$  in the range 167-175 and  $R_g$  of 55.8-58.7 (Table 1).

**Linear combinations of, respectively, the resting, activated and desensitized (pdb-code 5vhz) state, with the open class 3 EM structure (model 3;  $\chi_r^2$  of 13.0, 12.3 and 21.7 respectively; Fig. 3E-F) did not fit as well as linear combination with fractal oligomers (model 2; Fig. 3C-D). Model 4 was also tested (linear combination of a compact structure, the loose EM class 3 structure and fractal oligomers). However, using model 4 did not improve the fit significantly as compared to model 2 (compact structure and fractal oligomers).** These observations support a compact form in solution, combined with a small amount of oligomers of tetrameric GluA2 **(approximately 1-2%, see Supplementary Table S3).**



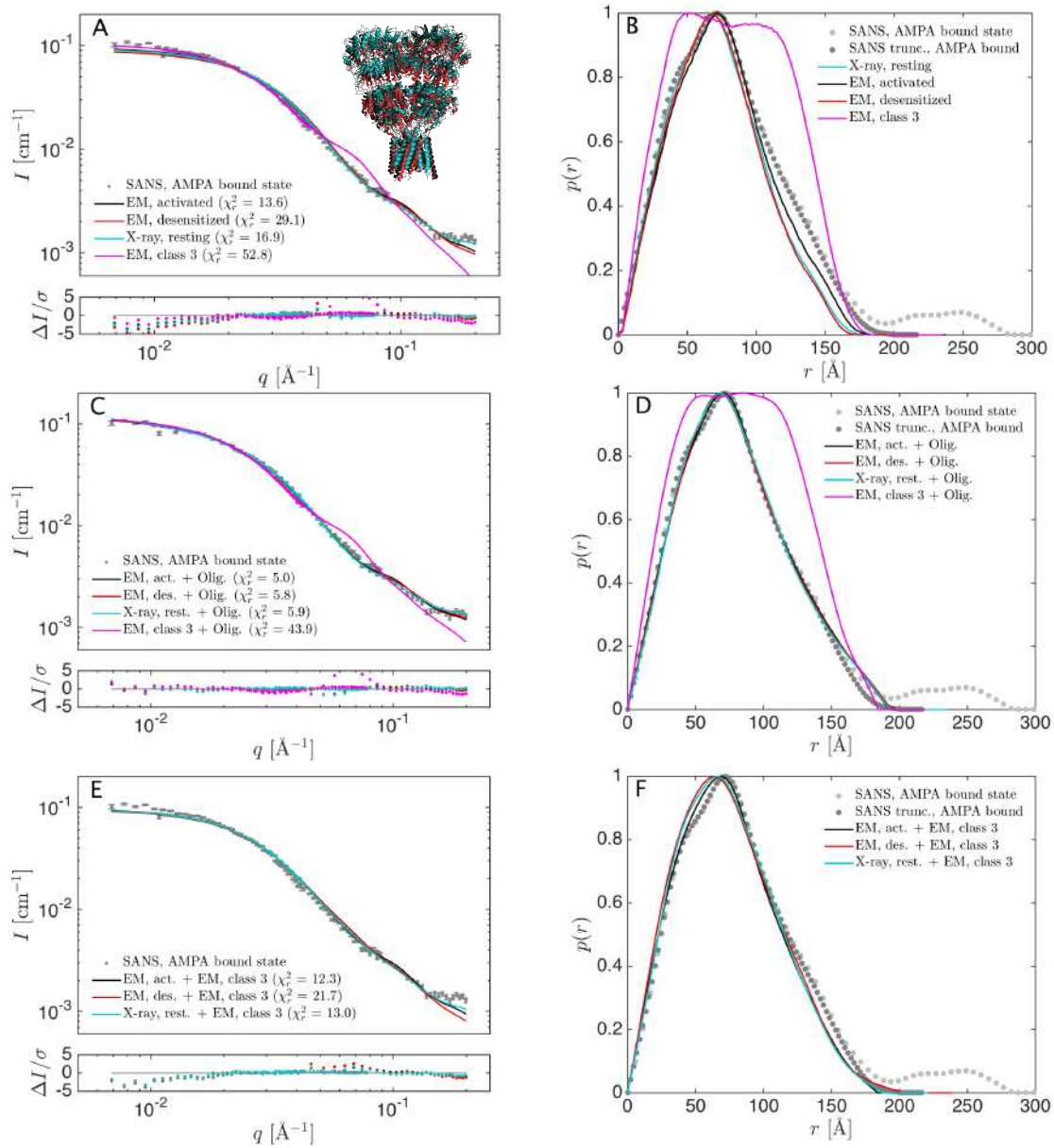


Figure 3. SANS data of GluA2 in the presence of AMPA at pH 7.5. (A) Experimental SANS data (grey points) and the resulting structure fits **and residual plots** for the crystal structure of GluA2 in the resting state (pdb-code 4u2p; cyan; Dürr *et al.*, 2014), GluA2 in the activated state (pdb-code 5weo; black; Twomey *et al.* 2017), GluA2 in the desensitized state (pdb-code 5vhz; red; Twomey *et al.* 2017) and GluA2 in the class 3 EM structure (EMD-2688; magenta; Meyerson *et al.*, 2014). A cartoon representation of the three structures overlaid are shown in respective colors. (B)  $p(r)$  functions for the SANS data (light grey points) and for a truncated data set ( $q \geq 0.011 \text{ \AA}^{-1}$ ; dark grey points) with the  $p(r)$  for four structures. (C-D) Resulting fits and  $p(r)$  functions for linear combinations of the atomic structures and fractal oligomers. (E-F) Resulting fits and  $p(r)$  functions for the combinations of the atomic structures and the class 3 EM structure.



When adding ~10 mM AMPA resulting in acidic pH of 5.5, we observed a significant structural change (Fig. 4 and Supplementary Fig. S5). The difference in the SANS data with 10 mM AMPA (Fig. 4) compared to the data with 1 mM AMPA (Fig. 3) is primarily seen in the low- $q$  region and in the  $q$ -range 0.02 to 0.06  $\text{\AA}^{-1}$ . The calculated  $M_w$ , estimated with Fischer analysis to be  $442 \pm 57$  kDa, is larger than that of the construct (368 kDa) (Table 1), but still in fair agreement with the expected tetrameric state. The Kratky plot showed that the protein was still in a folded or partially folded state (Supplementary Fig. S3). Interestingly, the SANS data at low pH are fitted relatively well by the more open GluA2 class 3 EM structure (Fig. 1B). The fit of the GluA2 class 3 EM structure resulted in  $\chi_r^2 = 10.1$ , whereas the goodness of fit was worse for the structures of GluA2 in the resting, activated and desensitized states (pdb-code 5vzh; Twomey *et al.* 2017) (17.0, 15.5 and 20.5, respectively; Fig. 4A). The values of  $R_g$  and  $D_{max}$  from the  $p(r)$  of the experimental SANS ( $65.2 \pm 0.5$   $\text{\AA}$  and  $189 \pm 5$   $\text{\AA}$ ) are larger than for GluA2 in the resting state or activated state (Table 1). On the other hand, these values are in accordance with the theoretical values calculated for the GluA2 class 3 EM structure with a hydration layer (64.1  $\text{\AA}$  and 179  $\text{\AA}$ ). When including fractal oligomers in the fit (EM class 3 and fractal oligomers; model 2;  $1.0 \pm 3.0\%$  oligomers), the goodness of fit was improved significantly ( $\chi_r^2 = 1.9$ ), now taking species of larger dimensions into account (Fig. 4C-D). The data were also fitted with combinations of GluA2 in the resting ( $\chi_r^2 = 5.2$ ), activated ( $\chi_r^2 = 4.6$ ) and desensitized ( $\chi_r^2 = 4.8$ ) state, respectively; all combined with fractal oligomers (model 2, Supplementary Fig. S6) to check if a combination of a compact structure and fractal oligomers could explain the data. The obtained  $\chi_r^2$ -values are significantly larger than the  $\chi_r^2$  of 1.9 obtained for the combination of the loose EM class 3 structure and fractal oligomers. From this, we conclude that the data show the best agreement with the EM class 3 structure, indicating a transition from a compact form to a loose form at low pH and in the presence of AMPA.

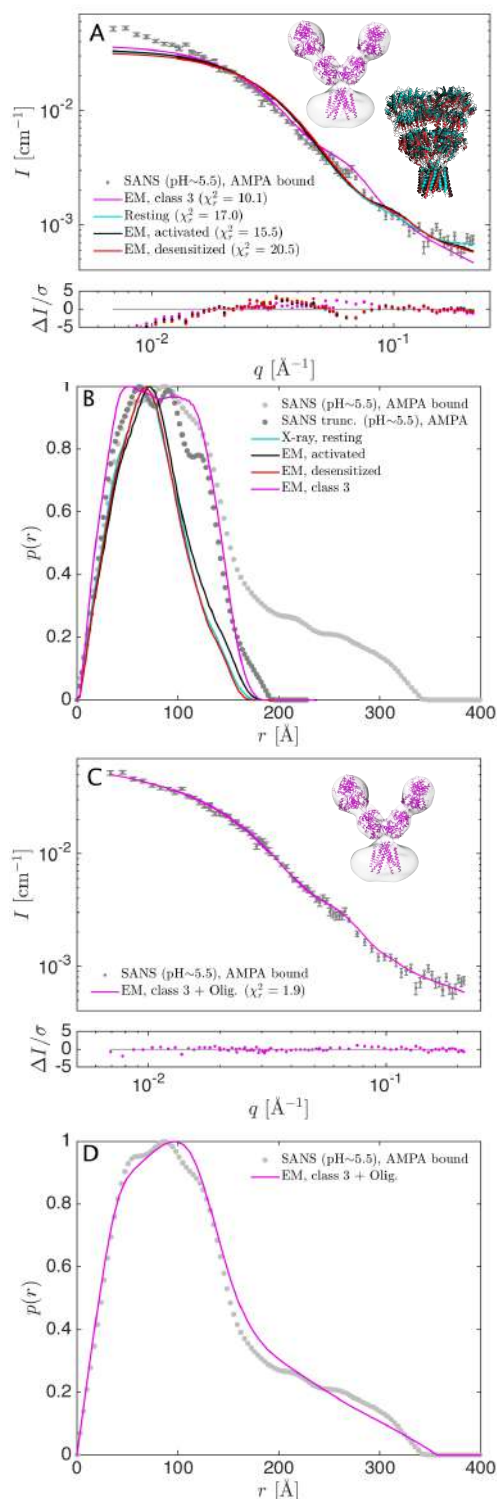


Figure 4. SANS data of GluA2 in the presence of AMPA at pH 5.5. (A) Experimental SANS data of GluA2 in the presence of AMPA at pH 5.5 (grey points) and resulting fits and residual plots of the GluA2 EM structure in the activated state (pdb-code 5weo; black; Twomey *et al.* 2017), the EM structure in the desensitized state (pdb-code 5vzh; red; Twomey *et al.* 2017), the X-ray crystal structure in the resting state (pdb-code 4u2p; cyan; Dürr *et al.*, 2014) and the class 3 EM structure (EMD-2688; magenta; Meyerson *et al.*, 2014). A cartoon representation of each of the three atomic resolution structures were aligned and shown in the respective colors. A cartoon representation of the atomic structure fitted to the EM class 3 density map is also shown. (B)  $p(r)$  functions for the SANS data (light grey points) and for a truncated data set ( $q \geq 0.019 \text{ \AA}^{-1}$ ; dark grey points) together with the theoretical  $p(r)$  functions for the four structures. (C-D) Experimental data (non-truncated; grey points), resulting fit and  $p(r)$  function for a linear combination of the EM class 3 structure and fractal oligomers (magenta).

### 3.3 GluA2 in the presence of GYKI-53655

We also looked into the solution structure of GluA2 in the presence of the negative allosteric modulator (non-competitive antagonist), GYKI-53655. Assessment of the molecular weight with Fischer analysis suggests that the protein was in the tetrameric state, with a calculated  $M_W$  of  $373 \pm 48$  kDa close to the  $M_W$  of the construct (368 kDa; Table 1). The Kratky plot implied a folded or partially folded structure (Supplementary Fig. S3).

The SANS data of GluA2 in the presence of GYKI-53655 was fitted by the crystal structure with the same ligand (pdb-code 5l1h; Yelshanskaya *et al.*, 2016) as well as the resting state (Fig. 5A). The goodness of fit was not optimal, neither to the GluA2 structure with GYKI-53655 ( $\chi_r^2 = 19.1$ ) nor GluA2 in the resting state ( $\chi_r^2 = 18.8$ ), especially in the low- $q$  region. An even worse fit was observed with the class 3 EM structure ( $\chi_r^2 = 45.5$ ). Again, including a small amount of fractal oligomers in the fitting procedure (model 2) improved the goodness of fit significantly. For example, when fitting the SANS data using the compact crystal structure of GluA2 with GYKI-53655, the fit was improved from  $\chi_r^2 = 19.1$  to  $\chi_r^2 = 7.5$  by inclusion of  $0.2 \pm 0.6\%$  fractal oligomers.

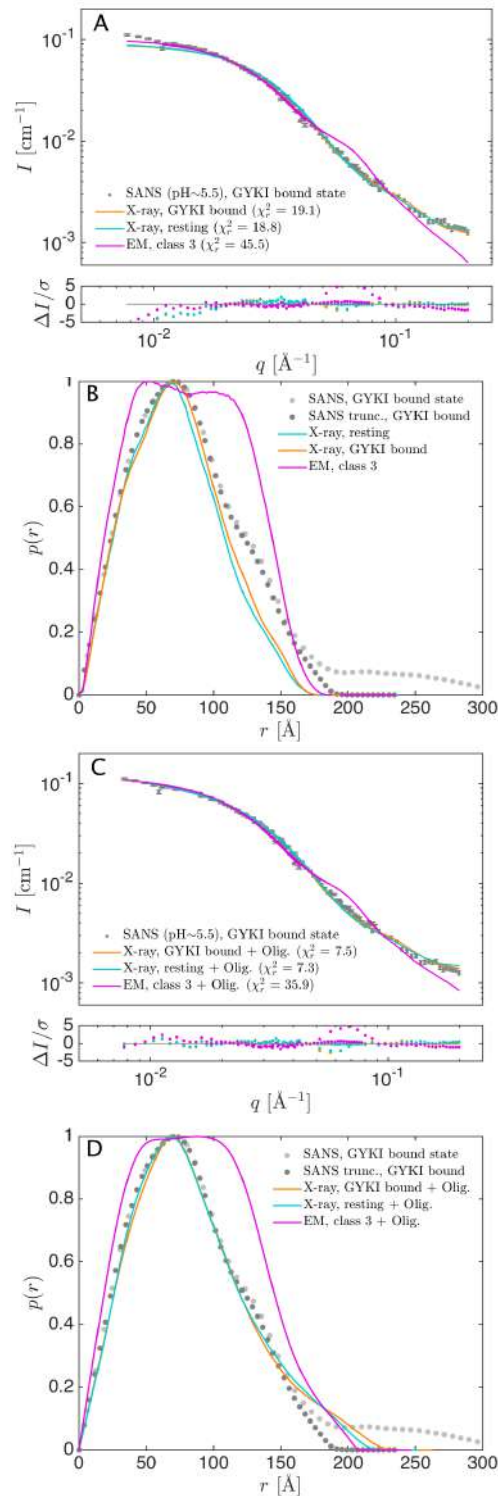


Figure 5. SANS data of GluA2 in the presence of GYKI-53655. (A) Experimental SANS data (grey points) and the resulting structure fits and residual plots for the X-ray crystal structures of GluA2 with GYKI-53655 bound (pdb-code 5l1h; orange; Yelshanskaya *et al.*, 2016), GluA2 in the resting state (pdb-code 4u2p; cyan; Dürr *et al.*, 2014) and GluA2 in the class 3 EM structure (EMD-2688; magenta; Meyerson *et al.*, 2014). (B)  $p(r)$  functions for the SANS data (light grey) and for a truncated data set ( $q \geq 0.012 \text{ \AA}^{-1}$ ; dark grey) and the theoretical  $p(r)$  functions for the two X-ray structures and the class 3 EM structure. (C) SANS data (grey points) and fits with a linear combination of the three structures and fractal oligomers. (D)  $p(r)$  functions for the SANS data (non-truncated and truncated) and for the linear combinations of the three structures and fractal oligomers.

#### 4. Discussion

Methods to study the structure of ionotropic glutamate receptors are essential in order to understand how these receptors function as well as for understanding of their role in diseases and as target for medicines. Within recent years, the AMPA receptor GluA2 has been thoroughly characterized in the resting, activated and desensitized states using X-ray crystallography and cryo-electron microscopy (see *e.g.* Dürr *et al.*, 2014; Twomey *et al.*, 2017). In this study, we investigated GluA2 in solution at 10 °C using SANS. This was made possible due to the very recent development of deuterated detergents with separate hydrogen/deuterium balances of the head- and tail groups that eliminated all signal from the detergent micelles solubilizing the membrane proteins in deuterated water-based buffer (Midtgaard *et al.*, 2018). Thus, such detergents allow for directly measuring the solution structure of the receptor **without seeing the surrounding micelles.**

Fractal oligomers were included in the fit of atomic structures to the experimental SANS data (model 2 and 4). Ideally, oligomerization in the sample should be avoided, *e.g.* by running a SEC-SANS experiment where SANS data are collected *in situ* as the sample leaves the purification column (Jordan *et al.*, 2016). **However, SEC-SANS is still an emerging technique only demonstrated at the D22 instrument at ILL, and is generally not a feasible technique for studies with many ligands due to a large sample consumption of protein, deuterated detergent and ligands.** Therefore, the data were instead “filtered” for the scattering contribution from large oligomers by inclusion of fractal oligomers in the model. **The information in the data about the detailed structure of the fractal oligomers is limited, which is reflected in the poorly determined values of  $\gamma$  and  $R_g$  (Supplementary Table S3). These are, however, not the parameters of interest, as the fractal oligomer model merely serves as a mean to minimize misinterpretations due to the effects of oligomerization.** Such models, as well as the associated molecular constraints, constitute a useful tool for future experiments. Combined with the recently developed match-out detergents (Midtgaard *et al.*, 2018), it enables the retrieval of information from samples that are not fully monodisperse.

In solution at 10 °C, **we find that** GluA2 primarily adopts a compact tetrameric structure both in the resting state as well as in the presence of 1 mM AMPA and 1 mM GYKI-53655, resembling the compact X-ray and cryo-EM structures determined at cryogenic temperature (Fig. 2, 3 and 5). **Therefore, this study adds support to the observation that GluA2 preferably adopts a compact conformation, also under conditions with no**

restraints on the dynamics of the protein. This was surprising as the cryo-EM study by Meyerson *et al.* (Meyerson *et al.*, 2014) showed that GluA2 was more dynamical in the presence of the agonist quisqualate, adopting a range of conformations of which three were modelled. Furthermore, GluA2 was also previously shown by cryo-EM to be conformationally heterogeneous in the presence of the partial agonist fluorowillardiine under desensitizing conditions, suggesting that GluA2 assumed a variety of different conformations (Dürr *et al.*, 2014). As the X-ray and cryo-EM structures of the resting, activated and desensitized states are very similar (Fig. 1A and Supplementary Fig. S7), it was statistically not possible to distinguish between these structures when fitted to the SANS data. In all cases, the fits of the compact structures were improved when fitting a linear combination of the atomic structure and small amounts of fractal oligomers, corresponding to a few percent. The F-test turned out to be a very useful tool for comparing hypothesized models to the SANS data. The P-values showed that despite minor improvements of the goodness of fit ( $\chi^2_r$ ) when fitting different compact structures, or using more complex models, it was not always statistically significant.

It is a characteristic of the X-ray and cryo-EM structures of the resting, activated and desensitized states that they all lack several amino-acid residues in the TMD. Also, the amino-acid sequences are not exactly the same as the sequence used in this study (Supplementary Fig. S4). This might affect the goodness of fit, and especially  $R_g$  and  $D_{max}$ . To address this issue, we introduced the missing amino-acid residues into the X-ray structure of GluA2 in the resting state with the program Modeller (Fiser *et al.*, 2000), using the “missing residue” procedure and assuming loop structure (Supplementary Fig. S8). This structural model led to an  $R_g$  (60.7 Å) more similar to the experimental value but at the same time had a larger  $D_{max}$  (201.0 Å) than the experimental data, while the deposited structures underestimated  $R_g$  and  $D_{max}$  due to missing residues in the structures (Table 1). The model with inserted loops thus partially explains this discrepancy between the theoretical and experimental scattering, but we do not consider the model to be accurate. Therefore, it was decided to use the deposited structures in the comparisons with the experimental SANS data. It should be noted that including/excluding the missing residues does not change the conclusion that GluA2 forms a compact structure in solution.

An *ab initio* model was generated based on SANS data of GluA2 in the resting state, clearly showing the individual domains: the TMD as well as the extracellular LBD and NTD layers. This *ab initio* model resembles the atomic structures of GluA2, but seems to be more asymmetric than the X-ray and cryo-EM structures. The discrepancy between the *ab initio* model and the crystal structure may, however, be caused by the scattering contribution from the fractal oligomers, since the sample was assumed to be solely in the tetrameric form in the *ab initio* modelling.

It has previously been reported using negative stain electron microscopy that GluA2 in the resting state adopted 60% compact structure, whereas addition of 3 mM glutamate led to only 3% compact structure (Nakagawa *et al.*, 2005). This distribution differs from what we observe for GluA2 in solution, where primarily compact structures of GluA2 are seen. We therefore speculated whether the dramatic shift towards more open GluA2 conformations in the negative stain electron microscopy studies could to some extent be due to a pH effect as the use of uranyl acetate typically results in pH below 5. Interestingly, when measuring on GluA2 in the presence of 10 mM AMPA, resulting in pH of 5.5, we observed an increase in the calculated average molecular weight ( $M_w$ ) to 442 kDa (Table 1), which indicated the presence of oligomers in the sample. Differences in scattering signals (Figs. 3 and 4) from addition of 10 mM AMPA or 1 mM AMPA could, however, not be explained by oligomerization alone. Whereas GluA2 in the presence of 1 mM AMPA adopts a compact structure, the SANS data for GluA2 in the presence of 10 mM AMPA could be fitted significantly better by a structure with a more open conformation of the extracellular part of GluA2, resembling the class 3 EM structure (Fig. 1B). Therefore, it is important to consider the impact of ligand concentration and/or pH on the GluA2 structure. As AMPA is present in large excess compared to GluA2 in this study (~1000-fold with 1 mM AMPA and ~10,000-fold with 10 mM;  $K_d$  of 16.8 nM (Coquelle *et al.*, 2000)), we suggest that the structural change in GluA2 observed in solution in the presence of 10 mM AMPA is primarily due to a pH effect. Protein stability is well-known to be affected by pH, and the structural change could very well be partial unfolding or aggregation. However, given the structural resemblance to the class 3 EM structure, it could be speculated that some ionization dependent interaction in the extracellular domain was destabilized at this low pH. Interestingly, two histidine residues (His229 in the NTD of chains B and D, respectively; numbering with signal peptide) are located in close proximity on the relatively small interaction surface between the NTDs (417 Å<sup>2</sup> for GluA2 in the resting state, pdb-code 4u2p;

'Protein interfaces, surfaces and assemblies' service PISA at the European Bioinformatics Institute. [http://www.ebi.ac.uk/pdbe/prot\\_int/pistart.html](http://www.ebi.ac.uk/pdbe/prot_int/pistart.html); Krissinel & Henrick, 2007). The pKa of histidine is most often in the interval 6-7 (Edgcomb & Murphy, 2002). Therefore, a pH decrease from 7.6 to 5.5 would effectively change the ionization of histidine from neutral to positive causing repulsion as well as unfavorable interactions to hydrophobic amino-acid residues. Whether this apparent widely open conformation of GluA2 observed at acidic pH is physiologically relevant is unclear and will require additional studies.

## 5. Conclusion

In this study we, to our knowledge, for the first time report data on full-length GluA2 (GluA2cryst with deletion of CTD) in solution as detergent solubilized protein. This was made possible by the recently developed fully matched out detergent described by us (Midtgaard *et al.*, 2018). We show that GluA2 primarily adopts a compact structure in solution at neutral pH, both in the resting state as well as in the presence of AMPA or GYKI-53655. Therefore, the solution structures of GluA2 are in accordance with most structures determined by X-ray crystallography and cryo-electron microscopy, but not with the more open class 3 EM structure. This study therefore adds support to the observation that GluA2 preferably adopts a compact conformation, also under conditions with no restraints on the dynamics of the protein. Moreover, we observed an altered and more open state at acidic pH in the presence of AMPA, resembling the class 3 EM structure. This observation should stimulate future structural studies. In conclusion, this study can serve as an example for future SANS studies on membrane proteins due to its methodological focus.

## SASBDB accession codes

The SANS data and the best fits have been uploaded to the small-angle scattering biological data bank (SASBDB; [www.sasbdb.org](http://www.sasbdb.org); Valentini *et al.*, 2018) with the following accession codes: SASDDY5 (GluA2 in the resting state), SASDDZ5 (GluA2 in the AMPA bound state, neutral pH), SASDD26 (GluA2 in the AMPA bound state, acidic pH), and SASDD36 (GluA2 in the GYKI-53655 bound state).

## Supporting Information



Supplementary figures and tables: Table S1: Information about the samples, the SANS measurements and the software used for the data analysis. Table S2: Results from Fischer/Petoukhov analysis. Table S3: Radius of gyration of oligomers and %oligomer fraction for each model. Figure S1: Guinier plots and residual plots. Figure S2: Porod plots. Figure S3: Kratky plots. Figure S4: Sequence alignment of GluA2 structures used in present study. Figure S5: SANS data of GluA2 in the presence of 1 mM AMPA at pH 7.5 and 10 mM AMPA at pH 5.5. Figure S6: Additional fits to SANS data of GluA2 in the AMPA bound state at pH 5.5. Figure S7: Theoretical SANS scattering for all investigated structures. Figure S8: Generated structure of GluA2 in the resting state. DOI: ????

### Acknowledgments

We would like to thank Eric Gouaux for providing the GluA2cryst construct. We thank FRM2 for awarding beamtime at KWS1, and Henrich Frielinghaus for support with data collection.

### Funding information

We thank the Danish Council for Independent Research – Medical Sciences (J.D., T.S.T., J.S.K.), CoNeXT (conext.ku.dk) (A.H.L., L.A., J.S.K.), BioSynergy (synbio.ku.dk) (N.T.J. and S.R.M.) and the Lundbeck Foundation Brainstruc (S.R.M. and L.A.) for financial support. The National Deuteration Facility is partly supported by the National Collaborative Research Infrastructure Strategy – an initiative of the Australian Government (T.D.). Beamtime travels were partly supported by the Danscatt program from the Danish Agency for Science, Technology and Innovation.

## References

- Chen, L., Dürr, K. L. & Gouaux, E. (2014). *Science* **345**, 1021-1026.
- Chen, S., Zhao, Y., Wang, Y., Shekhar, M., Tajkhorshid, E. & Gouaux E. (2017). *Cell* **170**, 1234-1246.
- Coquelle, T., Christensen, J. K., Banke, T. G., Madsen, U., Schousboe A. & Pickering, D. S. (2000). *Neuroreport* **11**, 2643–2648.
- Dürr, K. L., Chen, L., Stein, R. A., De Zorzi, R., Folea, I. M., Walz, T., Mchaourab, H. S. & Gouaux, E. (2014). *Cell* **158**, 778-792.
- Edgcomb, S. P. & Murphy, K. P. (2002). *Proteins* **49**, 1-6.
- Fiser, A., Do, R. K. G. & Šali, A. (2000). *Protein Sci.* **9**, 1753-1773.
- Fischer, H., de Oliveira Neto, M., Napolitano, H. B., Polikarpov, I. & Craievich, A. F. (2010). *J. Appl. Cryst.* **43**, 101–109.
- Feoktystov, A. V., Friehlinghaus, H., Di, Z., Jaksch, S., Pipich, V., Appavou, M.-S., Babcock, E., Hanslik, R., Engels, R., Kemmerling, G., Kleines, H., Ioffe, A., Richter, D. & Brückel, T. (2015). *J. Appl. Crystallogr.* **48**, 61-70.
- Franke, D. & Svergun, D. I. (2009). *J. Appl. Crystallogr.* **42**, 342-346.
- Franke, D., Petoukhov, M. V., Konarev, P. V., Panjkovich, A., Tuukkanen, A., Mertens, H. D. T., Kikhney, A. G., Hajizadeh, N. R., Franklin, J.M., Jeffries, C. M. & Svergun, D.I. (2017). *J. Appl. Crystallogr.* **50**, 1212-1225.
- Gekko, K. & Noguchi, H. (1979). *J. Phys. Chem.* **83**, 2706-2714.
- Glatter, O. (1977). *J. Appl. Crystallogr.* **10**, 415-421.
- Hansen, S. (2012). *J. Appl. Crystallogr.* **45**, 566-567.
- Herguedas, B., García-Nafria, J., Cais, O., Fernández-Leiro, R., Krieger, J., Ho, H. & Greger, I. H. (2016). *Science* **352**(6285):aad3873.
- Hoiberg-Nielsen, R., Westh, P., Skov, L. K. & Arleth, L. (2009). *Biophys. J.* **97**, 1445-1453.

Jordan, A., Jacques, M., Merrick, C., Devos, J., Forsyth, V. T., Porcar, L. & Martel, A. (2016). *J. Appl. Crystallogr.* **49**, 2015-2020.

Konarev P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J. & Svergun, D. I. (2003). *J. Appl. Crystallogr.* **36**, 1277-1282.

Kotlarchyk, M. & Chen, S.-H. (1983). *J. Chem. Phys.* **79**, 2461-2469.

Krissinel, E. & Henrick, K. (2007). *J. Mol. Biol.* **372**, 774-797.

Lee, K., Goodman, L., Fourie, C., Schenk, S., Leitch, B. & Montgomery, J. M. (2016). *Adv. Protein Chem. Struct. Biol.* **103**, 203-261.

Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. (2006). *Bioinformatics* **22**, 623-625.

Malik, L., Nygaard, J., Hoiberg-Nielsen, R., Arleth, L., Hoeg-Jensen, T. & Jensen, K. J. (2011). *Langmuir* **28**, 593-603.

Meyerson, J. R., Kumar, J., Chittori, S., Rao, P., Pierson, J., Bartesaghi, A., Mayer, M. L. & Subramaniam, S. (2014). *Nature* **514**, 328-334.

Meyerson, J. R., Chittori, S., Merk, A., Rao, P., Han, T. H., Serpe, M., Mayer, M. L. & Subramaniam, S. (2016). *Nature* **537**, 567-571.

Midtgaard, S. R., Darwish, T. A., Pedersen, M. C., Huda, P., Larsen, A. H., Jensen, G. V., Kynde, S. A. R., Skar-Gislinge, N., Nielsen, A. J. Z., Olesen, C., Blaise, M., Dorosz, J. J., Thorsen, T. S., Venskutonytė, R., Krintel, C., Møller, J. V., Friehlinghaus, H., Gilbert, E. P., Martel, A., Kastrup, J. S., Jensen, P. E., Nissen, P. & Arleth, L. (2018). *FEBS J.* **285**, 357-371.

Nakagawa, T., Cheng, Y., Ramm, E., Sheng, M. & Walz, T. (2005). *Nature* **433**, 545-549.

Pedersen, M. C., Arleth, L. & Mortensen, K. (2013). *J. Appl. Crystallogr.* **46**, 1894-1898.

Petouknov, M. V., Konarev, P. V., Kikhney, A. G. & Svergun, D. I. (2007). *J. Appl. Crystallogr.* **40**, 223-228.

Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., Gorba, C., Mertens, H. D.,

Konarev, P. V. & Svergun, D. I. (2012). *J. Appl. Crystallogr.* **45**, 342-350.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). *J. Comput. Chem.* **25**, 1605-1612.

Porod, G. (1982). *Small Angle X-ray Scattering* (Glatter, O. & Kratky, O., ed.), New York: Academic Press, Chapter 2, 17-52.

Sachser, R. M., Haubrich, J., Lunardi, P. S. & de Oliveira Alvares, L (2017). *Neuropharmacology* **112**(Pt A), 94-103.

Sobolevsky, A. I., Rosconi, M. P. & Gouaux, E. (2009). *Nature* **462**, 745-756.

Sorensen, C. M. & Roberts, G. C. (1997). *J. Colloid. Interface Sci.* **186**, 447-452.

Squire, P. G. & Himmel, M. E. (1979). *Arch. Biochem. Biophys.* **196**, 165-177.

Svergun, D. I., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Crystallogr.* **28**, 768-773.

Svergun, D. I., Richard, S., Koch, M. H. J., Sayers, Z., Kuprin, S. & Zaccai, G. (1998). *Proc. Natl. Acad. Sci. USA* **95**, 2267-2272,

Teixeira, J. (1988). *J. Appl. Crystallogr.* **21**, 781-785.

Traynelis, S. F., Wollmuth, L. P., McBain, C. J., Menniti, F. S., Vance, K. M., Ogden, K. K., Hansen, K. B., Yuan, H., Myers, S. J. & Dingledine R. (2010). *Pharmacol. Rev.* **62**, 405-496.

Trehwella, J., Duff, A.P., Durand, D., Gabel, F., Guss, J. M., Hendrickson, W. A., Hura, G. L., Jacques, D. A., Kirby, N. M., Kwan, A. H., Pérez, J., Pollack, L., Ryan, T. M., Sali, A., Schneidman-Duhovny, D., Schwede, T., Svergun, D. I., Sugiyama, M., Tainer, J. A., Vachette, P., Westbrook, J. & Whitten A. E. (2017). *Acta Crystallogr. D Struct. Biol.* **73**, 710-728.

Twomey, E. C., Yelshanskaya, M. V., Grassucci, R. A., Frank, J. & Sobolevsky, A. I. (2016). *Science* **353**, 83-86.

Twomey, E. C., Yelshanskaya, M. V, Grassucci, R. A., Frank. J. & Sobolevsky, A. I. (2017). *Nature* **549**, 60-65.

Valentini E., Kikhney, A. G., Previtali G., Jeffries C. M. & Svergun D. I. (2015). *Nucleic Acids Res.* **43**, D357-363.

Vestergaard, B. & Hansen, S. (2006). *J. Appl. Crystallogr.* **39**, 797-804.

Yelshanskaya, M. V., Singh, A.K., Sampson, J. M., Narangoda, C., Kurnikova, M. & Sobolevsky, A. I. (2016). *Neuron* **91**, 1305-1315.

Zhao, Y., Chen, S., Yoshioka, C, Bacongus, I. & Gouaux, E. (2016). *Nature* **536**, 108-111.

Webb, B. & Sali, A. (2014). *Current Protocols in Bioinformatics*, John Wiley & Sons Inc., 5.6.1-5.6.32.

## Supporting Information

### **Small-angle neutron scattering studies on the AMPA receptor GluA2 in the resting, AMPA and GYKI-53655 bound states**

Andreas Haahr Larsen,<sup>a</sup> Jerzy Dorosz,<sup>b</sup> Thor Seneca Thorsen,<sup>b</sup> Nicolai Tidemand Johansen,<sup>a</sup> Tamim Darwish,<sup>c</sup> Søren Roi Midtgaard,<sup>a</sup> Lise Arleth<sup>a,\*</sup> and Jette Sandholm Kastrup<sup>b,\*</sup>

<sup>a</sup> Structural Biophysics, X-ray and Neutron Science, The Niels Bohr Institute, University of Copenhagen, Denmark

<sup>b</sup> Biostructural Research, Department of Drug Design and Pharmacology, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

<sup>c</sup> National Deuteration Facility, Australian Nuclear Science and Technology Organization, Australia

\* Correspondent author: Jette Sandholm Kastrup: email: [jsk@sund.ku.dk](mailto:jsk@sund.ku.dk), phone: +45 35336486.

Correspondent author on SANS: Lise Arleth: email: [arleth@nbi.ku.dk](mailto:arleth@nbi.ku.dk)

## Table of Content

- Table S1: Information about the samples, the SANS measurements and the software used for the data analysis.
- Table S2: Fischer and Petoukhov  $M_w$  determination.
- Table S3:  $R_g$  of fractal oligomers and the amount of oligomers in the fitted models.
- Figure S1: Guinier plots and residual plots.
- Figure S2: Porod plots.
- Figure S3: Kratky plots.
- Figure S4: Sequence alignment of GluA2 structures used in present study.
- Figure S5: SANS data of GluA2 in the presence of 1 mM AMPA at pH 7.5 and 10 mM AMPA at pH 5.5.
- Figure S6: Additional fits to SANS data of GluA2 in the AMPA bound state at pH 5.5.
- Figure S7: Theoretical SANS scattering for all investigated structures.
- Figure S8: Generated structure of GluA2 in the resting state.
- References.

Table S1. Information about the samples, the SANS measurements and the software used for the data analysis.

	GluA2, Apo, in dDDM.	GluA2, AMPA-bound state, in dDDM, neutral pH.	GluA2, AMPA-bound state, in dDDM, acidic pH.	GluA2, GYKI-53655 bound state, in dDDM.
Sample details				
Uniprot ID	P19491 (GRIA2_RAT)			
Organism	Rattus Norvegicus			
Ligands	None	1 mM AMPA	10 mM AMPA	1 mM GYKI-53655
Buffer	20 mM Tris/DCl, 100 mM NaCl, 0.5 mM dDDM, pH 7.5	20 mM Tris/DCl, 100 mM NaCl, 0.5 mM dDDM, pH 7.5	20 mM Tris/DCl, 100 mM NaCl, 0.5 mM dDDM, pH 5.5	20 mM Tris/DCl, 100 mM NaCl, 0.5 mM dDDM, pH 7.5
Extinction coefficient* <sup>1</sup> [M <sup>-1</sup> cm <sup>-1</sup> ]	519100			
Density* <sup>2</sup> [g/ml]	1.37			
Molecular weight* <sup>1</sup> [kDa]	367.7			
Mean scattering length density* <sup>2</sup> of protein in D <sub>2</sub> O [10 <sup>-6</sup> Å <sup>-2</sup> ]	3.0			
Mean scattering length density* <sup>2</sup> of DDM tail groups in D <sub>2</sub> O [10 <sup>-6</sup> Å <sup>-2</sup> ]	6.4			
Mean scattering length density* <sup>2</sup> of DDM head groups in D <sub>2</sub> O [10 <sup>-6</sup> Å <sup>-2</sup> ]	6.4			
Mean scattering length density* <sup>2</sup> of solvent (D <sub>2</sub> O) [10 <sup>-6</sup> Å <sup>-2</sup> ]	6.4			
Protein concentration* <sup>3</sup>	0.20 mg/ml 0.54 μM	0.31 mg/ml 0.84 μM	0.17 mg/ml 0.46 μM	0.31 mg/ml 0.84 μM
SANS data collection details				
Instrument	KWS1@FRM2 ( <a href="https://www.mlz-garching.de/kws-1">https://www.mlz-garching.de/kws-1</a> )			
Date for data collection	19/09 2017	8/12 2016	19/09 2017	8/12 2016
Wavelength (λ <sub>mean</sub> , Δλ/λ)	5.0 Å, 10 % (FWHM)			
Beam dimensions	Rectangular beam, 6x10 mm <sup>2</sup> (at sample), 30x30 mm <sup>2</sup> (first pinhole)			
Resolution effects	Width of the resolution function Δq(q) was calculated by the beamline software and given in the 4 <sup>th</sup> column in data, which was used in WillItFit.			
Settings (Sample-detector/Collimation)	1.5m/4.0m, 4.0m/4.0m, 8.0m/8.0m			
Measured q-range	0.006-0.3 Å <sup>-1</sup>			
Absolute calibration	By plexiglass			
Exposure time (total for all 3 settings)	~ 2.5 hours	~ 4.5 hours	~ 4.5 hours	~ 4.0 hours
Temperature	10 °C			
Software employed				
Indirect Fourier transformations to obtain p(r)	BayesApp <sup>R1,R2</sup> ( <a href="http://www.bayesapp.org">www.bayesapp.org</a> )			
Calculation of theoretical p(r) Addition of water layer to protein	CaPP ( <a href="https://github.com/Niels-Bohr-Institute-XNS-StructBiophys/CaPP">https://github.com/Niels-Bohr-Institute-XNS-StructBiophys/CaPP</a> )			
Fitting of data with combined analytical and atomistic models	WillItFit <sup>R3</sup> ( <a href="https://sourceforge.net/projects/willitfit">https://sourceforge.net/projects/willitfit</a> )			



Fischer/Petoukhov $M_w$ determination	Own implementation in MATLAB (Table S3)			
Missing sequence modelling	MODELLER <sup>R4</sup> ( <a href="https://salilab.org?modeller">https://salilab.org?modeller</a> )			
Graphic model visualization	PyMOL			
Guinier analysis	Own implementation in MATLAB (Fig. S6)			
<i>Ab initio</i> dummy bead modelling	DAMMIF <sup>R5</sup> ( <a href="https://www.embl-hamburg.de/biosaxs/dammif.html">https://www.embl-hamburg.de/biosaxs/dammif.html</a> ) Note: DATGNOM <sup>R6,R7</sup> was used to generate the $p(r)$ function needed as input to DAMMIF			
Structural parameters				
<i>Guinier analysis</i>				
$I(0)$ [cm <sup>-1</sup> ]	0.043 ± 0.002	0.108 ± 0.004	0.052 ± 0.07 <sup>*4</sup>	0.109 ± 0.004
Mw from $I(0)$ [kDa]	220	347	240	347
(ratio to expected)	(0.60)	(0.94)	(0.66)	(0.94)
$R_g$ [Å]	60.8 ± 4.3	62.2 ± 3.0	79.1 ± 11.5 <sup>*4</sup>	62.8 ± 3.3
Minimum $q$ used [Å <sup>-1</sup> ]	0.0069	0.0078	0.0104	0.0103
Maximum $q \cdot R_g$	1.26	1.24	1.29	1.30
<i>p(r) analysis</i>				
$I(0)$ [cm <sup>-1</sup> ]	0.0444 ± 0.0002	0.106 ± 0.001	0.0418 ± 0.0004	0.109 ± 0.001
$R_g$ [Å]	61.9 ± 0.4	61.0 ± 0.6	65.2 ± 0.5	62.1 ± 0.3
$D_{max}$ [Å]	179 ± 11	184 ± 11	189 ± 5	186 ± 5
Used $q$ -range	[0.011,0.20]	[0.011,0.20]	[0.019,0.21]	[0.012,0.20]
Fitted constant background [cm <sup>-1</sup> ]	0.00052	0.00089	0.00048	0.00090
Reduced $\chi^2$	2.15	6.84	1.39	6.60
Number of good parameters	5.2	4.2	4.1	5.5
Number of Shannon channels	11.8	11.0	11.7	11.0
Number of error calculations	260	759	95	264
Regularization parameter $\log(\alpha)$	14.3	14.7	14.0	14.3
<i>Fischer <math>M_w</math> determination</i> <sup>*5</sup>				
Molecular weight [kDa]	396 ± 52	379 ± 49	442 ± 57	373 ± 48
(ratio to expected)	(1.08)	(1.03)	(1.20)	(1.01)
Model fitting parameters				
<i>Combined analytical and atomistic model</i>				
$q$ -range for fitting [Å <sup>-1</sup> ]	[0.006,0.3]	[0.006,0.3]	[0.006,0.3]	[0.006,0.3]
Reduced $\chi^2$ (best fit)	3.3	5.0	1.9	7.5
<i>Ab initio dummy bead modelling</i>				
Number of calculations	10	-- <sup>*6</sup>	-- <sup>*6</sup>	-- <sup>*6</sup>
Symmetry	P1, none			
NSD	1.5 ± 0.1			
Resolution (from SASRES <sup>R8</sup> ) [Å]	56 ± 4			
Filtered volume [nm <sup>3</sup> ]	380			
Mw from filtered volume	238 kDa			
(ratio to expected)	(0.92)			
SASBDB IDs for data and models				
SASBDB ID	SASDDY5	SASDDZ5	SASDD26	SASDD36
Footnotes and references				
<sup>*1</sup> Calculated with ExPASy Protparam ( <a href="https://web.expasy.org/protparam/">https://web.expasy.org/protparam/</a> ).				
<sup>*2</sup> Calculated with Biomolecular Scattering Length Density Calculator ( <a href="http://pslde.isis.rl.ac.uk/Pslde">http://pslde.isis.rl.ac.uk/Pslde</a> ).				
<sup>*3</sup> Protein concentration determined by UV280 absorption for GluA2 AMPA-bound at pH 7.5 and GluA2 GYKI-bound. Determined with BCA assay for GluA in the resting state and GluA2 AMPA-bound at pH 5.5				

\*<sup>4</sup> It was not possible to obtain fully linear region at  $qR_G < 1.3$  (Fig. S6C), so the the values may be incorrect.

\*<sup>5</sup>  $M_W$  determined with the Fischer method (Fischer *et al.*, 2011) with parameters given in Table S3.

\*<sup>6</sup> The dummy atom model for GluA2 apo is approximate, since aggregation was not taken into account. For the same reason, a dummy atom model was only generated for the GluA2 apo sample, where the aggregation scattering contribution was very minor.

<sup>R1</sup> Hansen, S. (2000). *J. Appl. Cryst.* **33**, 1415-1421.

<sup>R2</sup> Hansen, S. (2014). *J. Appl. Cryst.* **47**, 1469-1471.

<sup>R3</sup> Pedersen, M. C., Arleth, L. & Mortensen, K. (2013). *J. Appl. Cryst.* **46**, 1894-1898.

<sup>R4</sup> Fiser, A., Do, R.K. & Sali, A. (2000). *Protein Sci.* **9**, 1753-1773.

<sup>R5</sup> Franke, D. & Svergun, D. I. (2009). *J. Appl. Cryst.* **42**, 342-346.

<sup>R6</sup> Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J. & Svergun, D. I. (2003). *J. Appl. Cryst.* **36**, 1277-1282.

<sup>R7</sup> Petoukhov, M. V., Konarev, P. V., Kikhney, A. G. & Svergun, D. I. (2007). *J. Appl. Cryst.* **40**, 223-228.

<sup>R8</sup> Tuukkanen, A. T., Kleywegt, G. J. & Svergun, D. I. (2016). *IUCrJ* **3**, 440-447.

Table S2. Fischer and Petoukhov  $M_W$  determination (Petoukhov *et al.*, 2012; Fisher *et al.*, 2010), where  $M_W$  is determined via. the scattering “invariant”  $Q$  (Porod, 1982). The upper integration limit  $q_m$  used to determine  $Q$  was  $8/R_g$  (Petoukhov *et al.*, 2012).  $V_{app}$  is the apparent volume, and is the same for the two methods. In the Fischer method, linear coefficients  $A$  and  $B$  given in the table are used to convert  $V_{app}$  to the Porod volume  $V_p$ , and the weight-to-volume conversion constant of  $0.83 \text{ kDa/nm}^3$  to obtain  $M_W^F$ . In the Petoukhov method,  $M_W^P$  is determined directly from the  $V_{app}$  using the conversion constant  $0.625 \text{ kDa/nm}^3$ . The constant subtracted backgrounds  $K$  were used to assure a constant plateau in the Porod plots (Fig. S2) and the data sets were extrapolated to  $q = 0$  by simple linear extrapolation. An implementation in MATLAB of the methods was used. The value for  $M_W$  obtained with the Fisher method is given in Table S1 and used in the paper, since this method takes the size of the particle into account, which adds an important correction for large proteins such as GluA2. Values of  $R_g$  and  $I(0)$  from the  $p(r)$  analysis were used (Table S1).

Fischer/Petoukhov $M_W$ determination	Resting	AMPA pH 7.5 <sup>*1</sup>	AMPA pH 5.5	GYKI-53655
$V_{app} [\text{nm}^3]$	871.7	833.4	970.1	820.4
$M_W^F$ <sup>*1</sup> [kDa] (ratio to expected) $\Delta M_W/\sigma$ <sup>*2</sup>	$396 \pm 52$ (1.08) 0.56	$379 \pm 49$ (1.03) 0.22	$442 \pm 57$ (1.20) 1.3	$373 \pm 48$ (1.01) 0.1
$M_W^P$ <sup>*</sup> [kDa] (ratio to expected) $\Delta M_W/\sigma$	$545 \pm 109$ (1.48) 1.5	$521 \pm 104$ (1.42) 1.5	$606 \pm 121$ (2.65) 2.8	$513 \pm 103$ (1.39) 1.4
$K [10^{-3} \text{cm}^{-1}]$	0.65	1.30	0.61	1.30
$q_m = 8/R_g [\text{\AA}^{-1}]$	0.133	0.131	0.127	0.131
$A [\text{\AA}^3]$ $B$	-10500 0.56	-10500 0.56	-10500 0.56	-10500 0.56

<sup>\*1</sup> Assuming a 13% uncertainty on  $M_W^F$  (Fischer *et al.*, 2010, p. 106), and a 20% uncertainty on  $M_W^P$  (Petoukhov *et al.*, 2012, p. 344).

<sup>\*2</sup>  $\Delta M_W/\sigma$  is the normalized residual molecular weight, i.e. the difference between the experimentally determined value and the expected molecular weight in units of the experimental error. If  $M_W/\sigma < 2$  then the null-hypothesis (tetrameric state) cannot be rejected, given a significance level of 5%.

Table S3.  $R_g$  of fractal oligomers and the amount of oligomers in the fitted models.

Data	Apo	AMPA pH 7.5			AMPA pH 5.5	GYKI-53655
Model	X-ray, rest. + frac. olig.	X-ray, rest + frac. olig.	EM, act. + frac. olig.	EM, des. + frac. olig.	EM, class3 + frac. olig.	EM, GYKI + frac. olig.
$R_g$ for fractal oligomers [ $\text{\AA}$ ]	$126 \pm 250$	$190 \pm 177$	$145 \pm 135$	$116 \pm 62$	$175 \pm 166$	$240 \pm 254$
Fraction in oligomeric form, $\gamma$ [%]	$0.9 \pm 6.5$	$0.4 \pm 1.0$	$0.7 \pm 2.1$	$2.7 \pm 5.3$	$1.0 \pm 3.0$	$0.2 \pm 0.6$

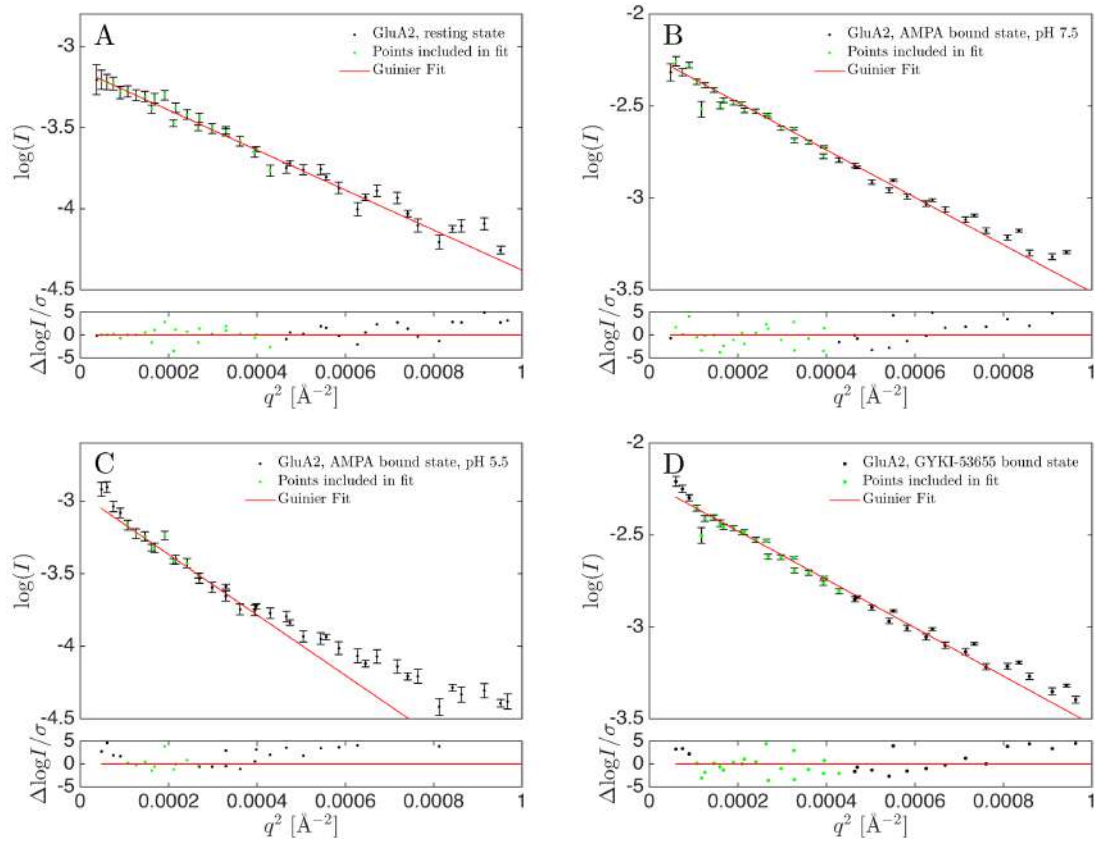


Figure S1. Guinier plots and residual plots for GluA2 in the resting state (A), in the AMPA bound state at pH 7.5 (B), in the AMPA bound state at pH 5.5 (C) and in the GYKI-53655 bound state (D). Residuals show the difference between  $\log(I)$  and the fit, weighted with the errors on  $\log(I)$ . Resulting values for  $I(0)$  and  $R_g$  are given in Table S1. The AMPA bound state at pH 5.5 (panel C) does not have a fully linear Guinier region at  $qR_g < 1.3$ , meaning that the values for  $I(0)$  and  $R_g$  may be wrong. The values of  $I(0)$  and  $R_g$  from the  $p(r)$  function was therefore used for  $M_w$  determination.

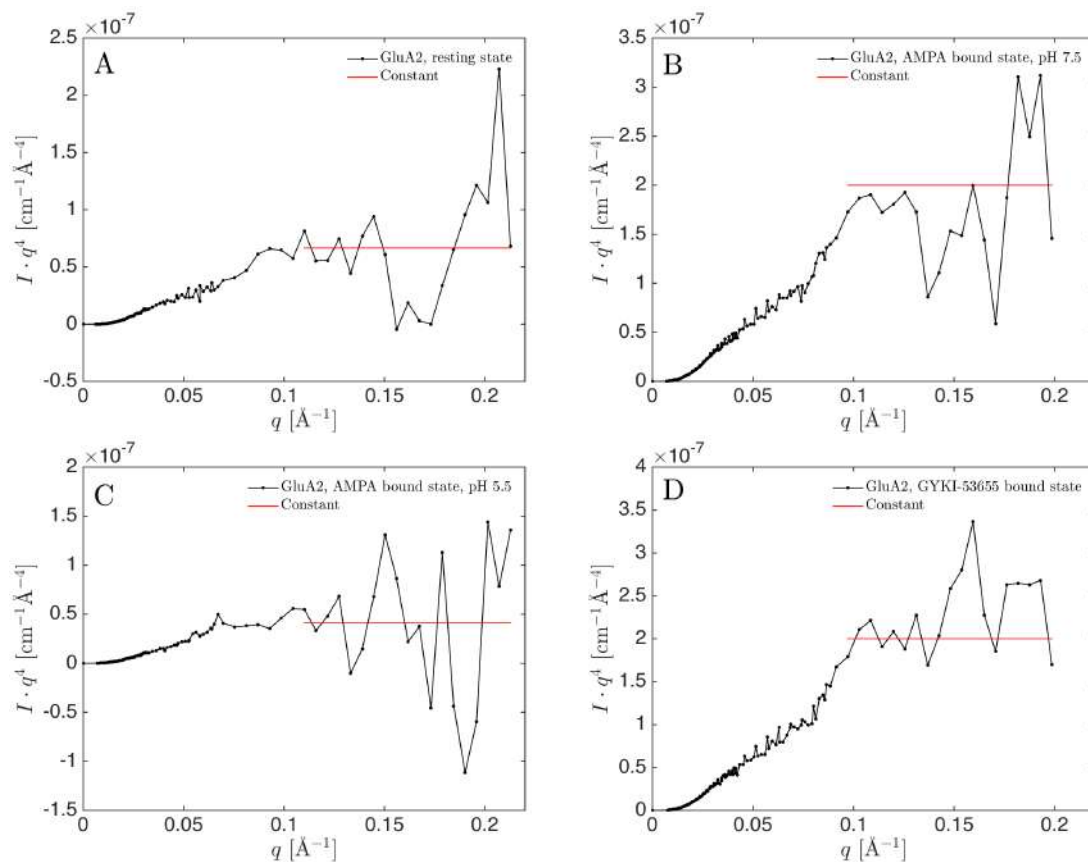


Figure S2. Porod plots (black) for GluA2 in the resting state (A), in the AMPA bound state at pH 7.5 (B), in the AMPA bound state at pH 5.5 (C) and in the GYKI-53655 bound state (D). Additional constant backgrounds were subtracted to give a constant behavior at high- $q$  (red). The constants are listed in Table S2.

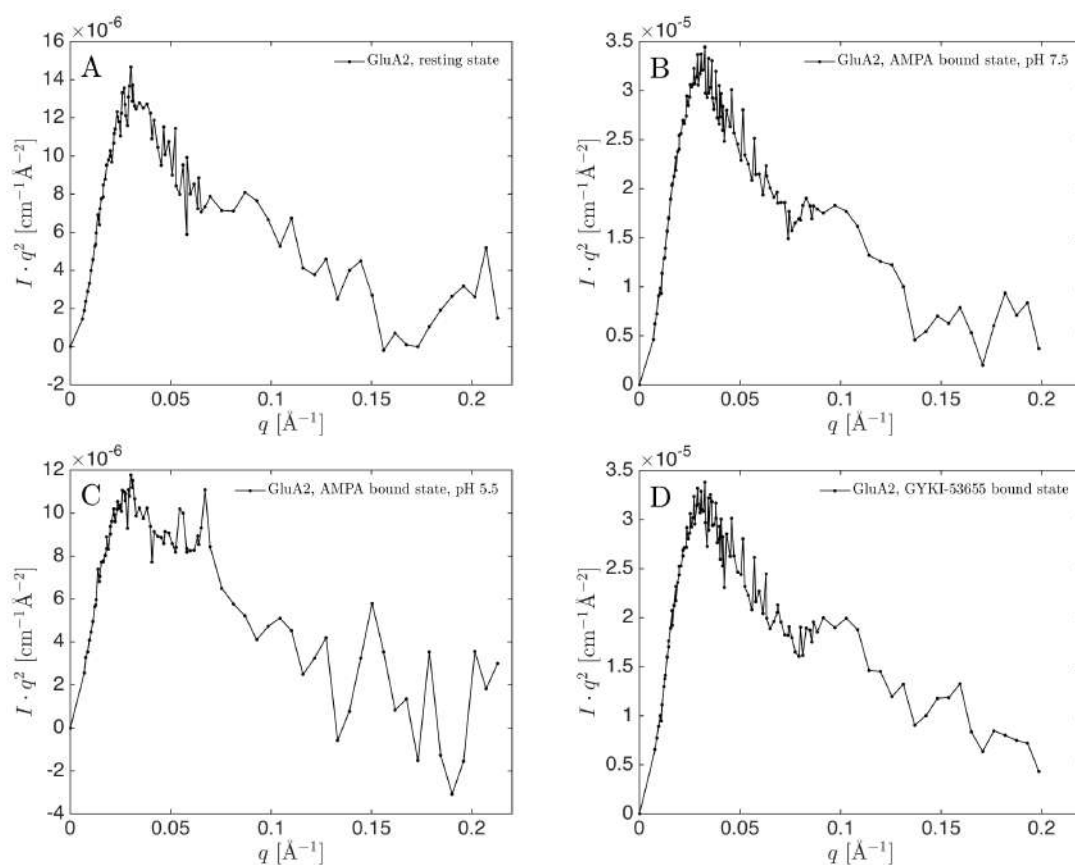


Figure S3. Kratky plots for GluA2 in the resting state (A), in the AMPA bound state at pH 7.5 (B), in the AMPA bound state at pH 5.5 (C) and in the GYKI-53655 bound state (D). Constant backgrounds were subtracted, and listed in Table S2.

3KG2:	NSIQIGGLFPRGADQEYSAFRVGMVQFSTSEFRLTPHIDNLEVANSFAVTNAFCSQFSRG	60
4U2P:	NSIQIGGLFPRGADQEYSAFRVGMVQFSTSEFRLTPHIDNLEVANSFAVTNAFCSQFSRG	60
5WEO:	NSIQIGGLFPRGADQEYSAFRVGMVQFSTSEFRLTPHIDNLEVANSFAVTNAFCSQFSRG	60
5VHZ:	NSIQIGGLFPRGADQEYSAFRVGMVQFSTSEFRLTPHIDNLEVANSFAVTNAFCSQFSRG	60
5L1H:	NSIQIGGLFPRGADQEYSAFRVGMVQFSTSEFRLTPHIDNLEVANSFAVTNAFCSQFSRG	60
3KG2:	VYAIFGFYDKKSVNTITSFCGTLHVSFITPSFPTDGTHPFVIQMRPDLKGALLSLIEYYQ	120
4U2P:	VYAIFGFYDKKSVNTITSFCGTLHVSFITPSFPTDGTHPFVIQMRPDLKGALLSLIEYYQ	120
5WEO:	VYAIFGFYDKKSVNTITSFCGTLHVSFITPSFPTDGTHPFVIQMRPDLKGALLSLIEYYQ	120
5VHZ:	VYAIFGFYDKKSVNTITSFCGTLHVSFITPSFPTDGTHPFVIQMRPDLKGALLSLIEYYQ	120
5L1H:	VYAIFGFYDKKSVNTITSFCGTLHVSFITPSFPTDGTHPFVIQMRPDLKGALLSLIEYYQ	120
3KG2:	WDKFAYLYDSRGLSTLQAVLDSAAEKKWQVTAINVGNINNDKKDETYRSLFQDLELKKE	180
4U2P:	WDKFAYLYDSRGLSTLQAVLDSAAEKKWQVTAINVGNINNDKKDETYRSLFQDLELKKE	180
5WEO:	WDKFAYLYDSRGLSTLQAVLDSAAEKKWQVTAINVGNINNDKKDETYRSLFQDLELKKE	180
5VHZ:	WDKFAYLYDSRGLSTLQAVLDSAAEKKWQVTAINVGNINNDKKDETYRSLFQDLELKKE	180
5L1H:	WDKFAYLYDSRGLSTLQAVLDSAAEKKWQVTAINVGNINNDKKDETYRSLFQDLELKKE	180
3KG2:	RRVILDCERDKVNDIVDQVITIGKHVKGYHYIIANLGFTDGDLLKIQFGGAEVSGFQIVD	240
4U2P:	RRVILDCERDKVNDIVDQVITIGKHVKGYHYIIANLGFTDGDLLKIQFGGAEVSGFQIVD	240
5WEO:	RRVILDCERDKVNDIVDQVITIGKHVKGYHYIIANLGFTDGDLLKIQFGGAEVSGFQIVD	240
5VHZ:	RRVILDCERDKVNDIVDQVITIGKHVKGYHYIIANLGFTDGDLLKIQFGGAEVSGFQIVD	240
5L1H:	RRVILDCERDKVNDIVDQVITIGKHVKGYHYIIANLGFTDGDLLKIQFGGAEVSGFQIVD	240
3KG2:	YDDSLVSKFIERWSTLEEKEYPGAHTATIKYTSALTYDAVQVMTEAFRNLRKQRIEISRR	300
4U2P:	YDDSLVSKFIERWSTLEEKEYPGAHTATIKYTSALTYDAVQVMTEAFRNLRKQRIEISRR	300
5WEO:	YDDSLVSKFIERWSTLEEKEYPGAHTATIKYTSALTYDAVQVMTEAFRNLRKQRIEISRR	300
5VHZ:	YDDSLVSKFIERWSTLEEKEYPGAHTATIKYTSALTYDAVQVMTEAFRNLRKQRIEISRR	300
5L1H:	YDDSLVSKFIERWSTLEEKEYPGAHTATIKYTSALTYDAVQVMTEAFRNLRKQRIEISRR	300
3KG2:	GNAGDCLANPAVPWGQGVIEIERALKQVQVEGLSGNIKFDQNGKRINYTIMELKTNNGPR	360
4U2P:	GNAGDCLANPAVPWGQGVIEIERALKQVQVEGLSGNIKFDQNGKRINYTIMELKTNNGPR	360
5WEO:	GNAGDCLANPAVPWGQGVIEIERALKQVQVEGLSGNIKFDQNGKRINYTIMELKTNNGPR	360
5VHZ:	GNAGDCLANPAVPWGQGVIEIERALKQVQVEGLSGNIKFDQNGKRINYTIMELKTNNGPR	360
5L1H:	GNAGDCLANPAVPWGQGVIEIERALKQVQVEGLSGNIKFDQNGKRINYTIMELKTNNGPR	360
3KG2:	KIGYWSEVDKMV--LTEDDTSGLEQKTVVVTTILESPPYMMKANHAALAGNEREYEGYCDV	418
4U2P:	KIGYWSEVDKMV--LTEDDTSGLEQKTVVVTTILESPPYMMKANHEMLEGNERYEGYCDV	420
5WEO:	KIGYWSEVDKMV--LTEDDTSGLEQKTVVVTTILESPPYMMKANHEMLEGNERYEGYCDV	418
5VHZ:	KIGYWSEVDKMV--LTEDDTSGLEQKTVVVTTILESPPYMMKANHEMLEGNERYEGYCDV	418
5L1H:	KIGYWSEVDKMV--LTEDDTSGLEQKTVVVTTILESPPYMMKANHEMLEGNERYEGYCDV	418
3KG2:	LAAEIAKHCGFKYKLTIVGDGKYGARDADTKIWNMGVGLVYGKADIAIAPLTITLVREE	478
4U2P:	LAAEIAKHCGFKYKLTIVGDGKYGARDADTKIWNMGVGLVYGKADIAIAPLTITLVREE	480
5WEO:	LAAEIAKHCGFKYKLTIVGDGKYGARDADTKIWNMGVGLVYGKADIAIAPLTITLVREE	478
5VHZ:	LAAEIAKHCGFKYKLTIVGDGKYGARDADTKIWNMGVGLVYGKADIAIAPLTITLVREE	478
5L1H:	LAAEIAKHCGFKYKLTIVGDGKYGARDADTKIWNMGVGLVYGKADIAIAPLTITLVREE	478
3KG2:	VIDFSKPFMSLGISIMIKKPQKSKPGVFSFLDPLAYEIWMCIWFAYIGVSVVFLVSRFS	538
4U2P:	VIDFSKPFMSLGISIMIKKPQKSKPGVFSFLDPLAYEIWMCIWFAYIGVSVVFLVSRFS	540
5WEO:	VIDFSKPFMSLGISIMIKKPQKSKPGVFSFLDPLAYEIWMCIWFAYIGVSVVFLVSRFS	538
5VHZ:	VIDFSKPFMSLGISIMIKKPQKSKPGVFSFLDPLAYEIWMCIWFAYIGVSVVFLVSRFS	538
5L1H:	VIDFSKPFMSLGISIMIKKPQKSKPGVFSFLDPLAYEIWMCIWFAYIGVSVVFLVSD	536
3KG2:	PYEWHTTEEFEDGRETQSSESTNEFGIFNSLWFSLGAFMQQGADISPRSLSGRIVGGVWWF	598
4U2P:	PYEWHTTEEFEDGRETQSSESTNEFGIFNSLWFSLGAFMQQGADISPRSLSGRIVGGVWWF	600
5WEO:	PYEWHTTEEFEDGRETQSSESTNEFGIFNSLWFSLGAFMQQGADISPRSLSGRIVGGVWWF	598
5VHZ:	PYEWHTTEEFEDGRETQSSESTNEFGIFNSLWFSLGAFMQQGADISPRSLSGRIVGGVWWF	598
5L1H:	-----TDSTNEFGIFNSLWFSLGAFMQQGADISPRSLSGRIVGGVWWF	579
3KG2:	FTLIISSYTANLAAFLTVERMVSPIESAEDLSKQTEIAYGTLDGSGSTKEFFRRSKIAVF	658



4U2P:	FTLIIISSYTANLAAFLTVERMVSPIESAEDLSKQTEIAYGTLDSGSTKEFFRRSKI	660
5WEO:	FTLIIISSYTANLAAFLTVERMVSPIESAEDLSKQTEIAYGTLDSGSTKEFFRRSKI	658
5VHZ:	FTLIIISSYTANLAAFLTVERMVSPIESAEDLSKQTEIAYGTLDSGSTKEFFRRSKI	658
5L1H:	FTLIIISSYTANLAAFLTVERMVSPIESAEDLSKQTEIAYGTLDSGSTKEFFRRSKI	639
3KG2:	DKMWTYMRSAEPSVFVRTTAEGVARVRKSKGKYAYLLESTMNEYIEQRKPCDTMKVGGNL	718
4U2P:	DKMWTYMRSAEPSVFVRTTAEGVARVRKSKGKYAYLLESTMNEYIEQRKPCDTMKVGGNL	720
5WEO:	DKMWTYMRSAEPSVFVRTTAEGVARVRKSKGKYAYLLESTMNEYIEQRKPCDTMKVGGNL	718
5VHZ:	DKMWTYMRSAEPSVFVRTTAEGVARVRKSKGKYAYLLESTMNEYIEQRKPCDTMKVGGNL	718
5L1H:	DKMWTYMRSAEPSVFVRTTAEGVARVRKSKGKYAYLLESTMNEYIEQRKPCDTMKVGGNL	699
3KG2:	DSKGYGIATPKGSSSLGTPVNLAVLKLSEQGLLDKLNKWWYDKGECGAKDSGSKEKTSAL	778
4U2P:	DSKGYGIATPKGSSSLGTPVNLAVLKLSEQGLLDKLNKWWYDKGECGAKDSGSKEKTSAL	780
5WEO:	DSKGYGIATPKGSSSLGTPVNLAVLKLSEQGLLDKLNKWWYDKGECGAKDSGSKEKTSAL	778
5VHZ:	DSKGYGIATPKGSSSLGTPVNLAVLKLSEQGLLDKLNKWWYDKGECGAKDSGSKEKTSAL	778
5L1H:	DSKGYGIATPKGSSSLGTPVNLAVLKLSEQGLLDKLNKWWYDKGECGAKDSGSKEKTSAL	759
3KG2:	SLSNVAGVFYILVGGLGLAMLVALIEFCYKSRAEAKRMKGLVPRG	823
4U2P:	SLSNVAGVFYILVGGLGLAMLVALIEFCYKSRAEAKRMKGLVPR-	824
5WEO:	SLSNVAGVFYILVGGLGLAMLVALIEFCYKSRAEAKRMK-----	817
5VHZ:	SLSNVAGVFYILVGGLGLAMLVALIEFCYKSRAEAKRMK-----	817
5L1H:	SLSNVAGVFYILVGGLGLAMLVALIEFCYKSRAEAKRMKGLVPR-	803

Figure S4. Sequence alignment of GluA2 structures used in present study. The alignment was made using Clustal Omega (Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J. & Lopez, R. A new bioinformatics analysis tools framework at EMBL-EBI (2010) *Nucl. Acids Res.* W695-699). Residues in green are differing from the target sequence (3kg2). The residues marked in italics were not seen in the structures (for chain A, similar for the other chains).

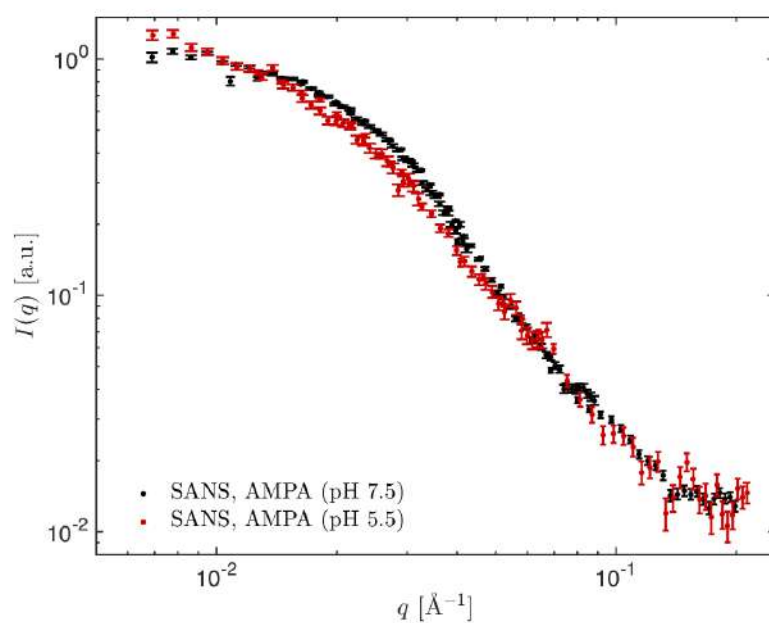


Figure S5: SANS data of GluA2 in the presence of 1 mM AMPA at pH 7.5 (black) and 10 mM AMPA at pH 5.5 (red)

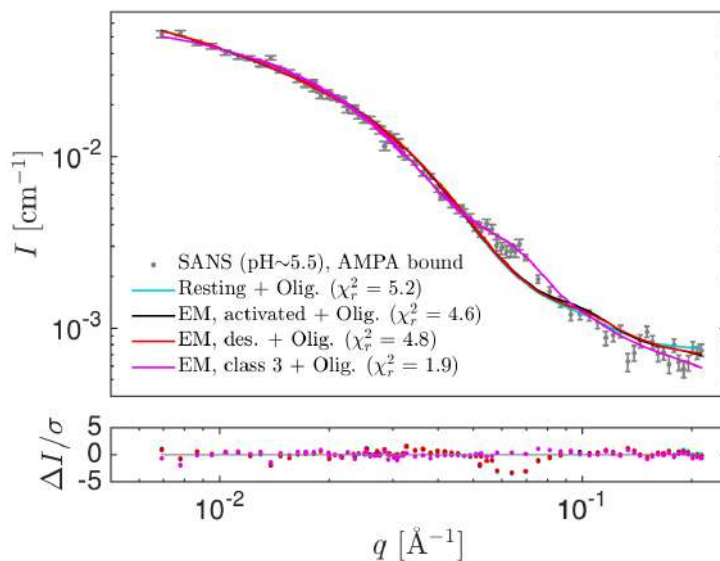


Figure S6. Additional fits to SANS data of GluA2 in the AMPA bound state at pH 5.5 (grey). The data were fitted with models of tetrameric GluA2 in combination with fractal oligomers. Models included GluA2 in the resting state (cyan, pdb-code 4u2p,  $\chi_r^2 = 5.2$ ), GluA2 in the activated state (black; pdb-code 5weo;  $\chi_r^2 = 4.6$ ), GluA2 in the desensitized state (red; pdb-code 5lhv;  $\chi_r^2 = 4.8$ ) and the class 3 EM structure (magenta; EMD-2688;  $\chi_r^2 = 1.9$ ).

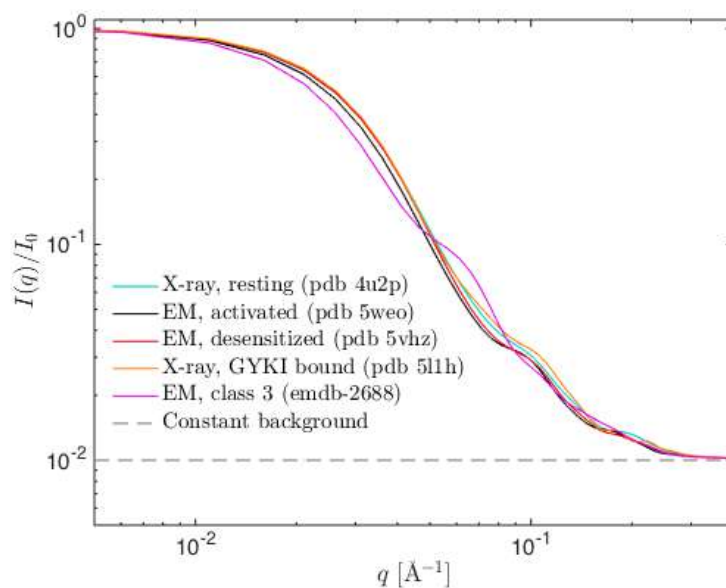


Figure S7. Theoretical SANS scattering for all investigated structures. GluA2 in the resting state (X-ray; cyan; pdb-code 4u2p), in the activated state (EM; black; pdb-code 5weo), in the desensitized state (EM; pdb-code 5vhz), in the GYKI-53655 bound state (X-ray; orange; pdb-code 5l1h) and GluA2 in the class 3 state (EM; magenta; EMDB-2688). Data are normalized and a constant background of  $0.01 \cdot I(0)$  is subtracted (grey dashed line). The compact forms are similar, whereas the scattering curve for the more open EM class 3 structure is clearly distinguishable by eye. The compact structures differs only at high  $q$ -values, where the signal to noise ratio is low.

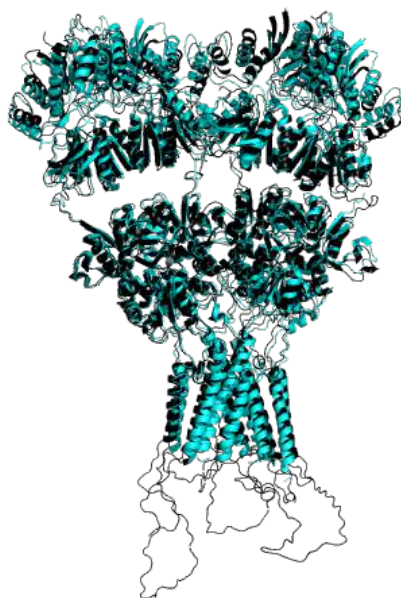


Figure S8. Generated structure of GluA2 in the resting state. Due to missing residues in the X-ray structure of GluA2 in the resting state (cyan; pdb-code 4u2p), a model structure was generated of GluA2 (black) using Modeller (Fiser et al., 2000), with the missing residues inserted as loops.

### References

- Fischer, H., de Oliveira Neto, M., Napolitano, H. B., Polikarpov, I. & Craievich, A. F. (2010). *J. Appl. Cryst.* **43**, 101-109.
- Fiser, A., Do, R.K.G. & Šali, A. (2000). *Protein Sci.* **9**, 1753-1773.
- Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., Gorba, C., Mertens, H. D., Konarev, P.V. & Svergun, D. I. (2012). *J. Appl. Cryst.* **45**, 342-350.
- Porod, G. (1982). *Small Angle X-ray Scattering* (Glatter, O. & Kratky, O., ed.), New York: Academic Press, Chapter 2, 17-52.

## 9.5 Paper V: Structure, dynamics and function of a lipid pool at the centre of the bacterial holo-translocon

**List of Authors** Remy Martin\*, Andreas Haahr Larsen\*, Robin A. Corey, Søren Roi Midtgaard, Nathan Zaccai, Marie-Sousai Appavou, Christiane Schaffitzel, Lise Arleth and Ian Collinson

\*These authors contributed equally to the work.

**Status** Not submitted yet.

**Abstract** The bacterial Sec translocon, SecYEG, associates with accessory proteins YidC and the SecDF-YajC subcomplex to form a bacterial holo-translocon (HTL). The HTL is a dynamic and flexible protein transport machine capable of coordinating protein secretion across the membrane, and efficient lateral insertion of membrane proteins. Previously, it has been hypothesized a central lipid core facilitates the controlled passage of membrane proteins into the bilayer, ensuring efficient formation of their native state. By performing small angle neutron scattering (SANS) on protein solubilized in deuterated detergent, we have been able to interrogate a “naked” HTL with the scattering contribution of detergent rendered invisible. Such an approach has allowed the confirmation of a lipid core at the heart of the complex, accommodating approximately  $22 \pm 2$  lipids. Large-scale coarse-grained MD simulations of the HTL corroborate this value, demonstrating a dynamic, central pool of lipids. An opening between YidC and the SecY lateral gate may provide an exit gateway for newly synthesized, correctly oriented membrane protein helices to emerge from the HTL.

**Contributions by AHL** AHL did the SANS measurements together with SRM and RM. AHL took part in defining the core questions to be answered from the SANS data with LA, SRM, IC and RM. AHL did the SANS analysis. RM and AHL co-wrote the paper with contributions from IC and LA.

**Structure, dynamics and mechanism of a lipid pool at the centre of the bacterial  
holo-translocon**

Remy Martin<sup>1†</sup>, Andreas Haahr Larsen<sup>2†</sup>, Robin Adam Corey<sup>1£</sup>, Søren Roi  
Midtgaard<sup>2</sup>, Nathan Zaccai<sup>1§</sup>, Marie-Sousai Appavou<sup>4</sup>, Christiane Schaffitzel<sup>1</sup>, Lise  
Arleth<sup>2\*</sup> and Ian Collinson<sup>1\*</sup>

<sup>1</sup>, *School of Biochemistry, University of Bristol, Bristol, BS8 1TD, United Kingdom*

<sup>2</sup>, *Niels Bohr Institute, University of Copenhagen, Universitetsparken 5, 2100  
Copenhagen, Denmark.*

<sup>3</sup>, *Institut Laue Langevin, 71 Avenue des Martyrs, F-38042 Grenoble, France*

<sup>4</sup>, *Jülich Centre for Neutron Science JCNS at Heinz Maier-Leibnitz Zentrum (MLZ),  
85748 Garching, Germany*

<sup>†</sup>, these authors contributed equally to this work

<sup>§</sup>, present address: Cambridge Institute for Medical Research, University of  
Cambridge, Cambridge CB2 0XY, United Kingdom

<sup>£</sup>, present address: Department of Biochemistry, University of Oxford, OX1 3QU

<sup>\*</sup>, corresponding authors; [ian.collinson@bristol.ac.uk](mailto:ian.collinson@bristol.ac.uk) and [arleth@nbi.ku.dk](mailto:arleth@nbi.ku.dk)



**Abstract**

The bacterial *Sec* translocon, SecYEG, associates with accessory proteins YidC and the SecDF-YajC subcomplex to form a bacterial holo-translocon (HTL). The HTL is a dynamic and flexible protein transport machine capable of coordinating protein secretion across the membrane, and efficient lateral insertion of nascent membrane proteins. It has been hypothesized that a central lipid core facilitates the controlled passage of membrane proteins into the bilayer, ensuring efficient formation of their native state. By performing small angle neutron scattering (SANS) on protein solubilized in deuterated detergent, we have been able to interrogate a “naked” HTL complex, with the scattering contribution of the detergent micelle rendered invisible. Such an approach has allowed the confirmation of a lipid core within the HTL, which accommodating  $27 \pm 5$  lipids. Coarse-grained molecular dynamics simulations of the HTL corroborate this value, demonstrating a dynamic, central pool of lipids. An opening between YidC and the SecY lateral gate may provide an exit gateway for newly synthesized, correctly oriented, membrane protein helices to emerge from the HTL.

## Introduction

The general process of protein secretion and membrane protein insertion is achieved by the conserved secretory, or *Sec*, machinery at the plasma membrane of bacteria and archaea, and the endoplasmic reticulum (ER) of eukaryotes. The protein-conducting channel is formed by a core hetero-trimeric assembly – the SecY (bacteria/archaea) / Sec61 complex (eukaryotes) (Brundage et al. 1990; Görlich & Rapoport 1993). Secretory and membrane proteins are driven through the complex, passing across or inserting into the membrane. This process occurs either during protein synthesis, involving the direct binding of co-translating ribosomes to Sec, or post-translationally, powered by associated energy transducing factors. Additional components combine with the core complex to facilitate the lateral passage of trans-membrane  $\alpha$ -helices into the bilayer, or for the implementation of specific post-translational modification, such as glycosylation. Bacterial SecYEG associates with additional membrane protein factors to facilitate and modulate successful translocation and insertion of substrates. Using immunoprecipitation, SecYEG was co-purified with SecDF, YajC (Duong & Wickner 1997) and an unknown protein of approximately 60kDa, identified as YidC (Scotti et al. 2000). This 7 protein super-complex of SecYEG, SecDF-YajC and YidC is known as the holo-translocon (HTL).

The resultant holo-translocon ensures efficient translocation, modification, folding and assembly of secretory and membrane proteins. The structure of the eukaryotic holo-translocon engaged with the ribosome illustrates how the core complex and accessory factors could streamline the efficient translocation and processing of proteins at the ER membrane (Pfeffer et al. 2015).

HTL can now be produced in sufficient quantities for structural and functional analyses (Bieniossek et al. 2009; Schulze et al. 2014; Komar et al. 2016). The material allowed the joint cryo electron microscopy (cryo-EM) and small-angle neutron (SANS) structural analyses of the complex composed of the core complex SecYEG, the membrane protein insertase YidC and the accessory sub-complex SecDF-YajC (Botte et al. 2016) (Figure 1). Interestingly, the proteins are arranged around a central cavity, most likely constituted of lipids, which we suppose forms a protected environment for the insertion of trans-membrane  $\alpha$ -helical bundles. The encapsulation of nascent unfolded membrane proteins would prevent catastrophic proteolysis or aggregation and thus promote protein folding, much in the same way that GroEL facilitates the folding of globular proteins within a secluded hydrophilic chamber (Xu et al. 1997).

High-resolution structures of the individual components of the HTL are known (Van Den Berg et al. 2004; Kumazaki et al. 2014; Tsukazaki et al. 2011). These structures could be fitted into the low-resolution cryo-EM structure to create a preliminary atomic model of the HTL, supported also by biochemical data (Botte et al. 2016). In this model the lateral gate of SecY, from which nascent trans-membrane helices enter the membrane (Van Den Berg et al. 2004), faces the central lipid cavity. YidC is located on the opposite side of the cavity, with its putative binding site for inserting trans-membrane helices (Kumazaki et al. 2014) also facing the lipid pool. The juxtaposition of these regions at the lipid core of the HTL provides a compelling case for their concerted action in membrane protein insertion.

To explore further the structure and arrangement of the central lipid core we conducted an analysis of the HTL, combining SANS and coarse-grained (CG) molecular dynamics (MD) simulations. HTL was solubilised in the deuterated

detergent n-dodecyl- $\beta$ -D-maltoside (DDM). This d-DDM was deuterated separately in the head and tail group to fully match out – i.e. appear invisible – in SANS in a D<sub>2</sub>O-based buffer (Midtgaard et al. 2018). This allowed us to distinguish and unambiguously describe the lipid component of the translocon. The CG MD simulations support the notion of a stable and persistent lipid-filled cavity within the centre of the HTL.

## **Materials and Methods**

### **HTL preparation and d-DDM exchange**

HTL was purified as described (Schulze et al. 2014). Purified HTL in hydrogenated DDM was exchanged into a 100% D<sub>2</sub>O buffer containing deuterated DDM. HTL was purified as described previously (Schulze et al. 2014). Detergent exchange was performed on a Superose 6 (10/300) column equilibrated in a simple TS buffer (20mM Tris pH 7.5, 100mM NaCl<sub>2</sub>), made with 100% D<sub>2</sub>O, and 0.02% deuterated DDM.

### **SANS data collection for deuterated detergent**

Samples were prepared and measured in 2 mm quartz cuvettes (Hellma), temperature controlled at 10 °C. Data were collected on KWS-1 at FRMII (Garching), at a wavelength of  $\lambda = 5 \text{ \AA} \pm 10\%$  (FWHM). Detector/collimation distances of 1.5m/4m and 8m/8m (sample-detector/collimation distance) were used, to obtain a  $q$ -range of 0.006 to 0.44  $\text{\AA}^{-1}$ , with a good overlap between the settings. The wave vector is defined as  $q = 4\pi \sin(\theta) / \lambda$ , where  $2\theta$  is the scattering angle. Data were calibrated and placed on an absolute scale using plexiglass as a calibrant.

Correction and averaging was performed using QtiKWS (v. 10; [www.qtikws.de](http://www.qtikws.de)), and the buffer measurement was subtracted subsequently. The sample was measured for ~4 hours to obtain sufficient signal over background. 15 minute measurement windows were used to monitor change in scattering over time. No change was observed, meaning that the sample was stable during the measurements.

### SANS data analysis

Home-written software CaPP (v. 3.8; [github.com/Niels-Bohr-Institute-XNS-StructBiophys/CaPP](https://github.com/Niels-Bohr-Institute-XNS-StructBiophys/CaPP)) was used to fit the data. A water layer with 6% higher scattering length than bulk D<sub>2</sub>O was added (Persson et al. 2018), but was excluded from the transmembrane region. A hydrophobic bilayer thickness of 30.6 Å was assumed in accordance with the orientations of proteins in the OPM membrane database (Lomize et al. 2012). Resolution effects were included using the 4<sup>th</sup> column in data (uncertainty in  $q$ ), as provided by the beamline. CaPP was also used to calculate the theoretical pair distance distribution,  $p(r)$ , functions for the full-atom structures. Experimental  $I(r)$  functions were calculated using BayesApp (Hansen 2012) including a constant background in the fit and truncation of data at  $q = 0.3 \text{ Å}^{-1}$ . The fit to obtain the  $p(r)$  had a  $\chi_r^2$  of 2.7. As the model is generic and thus true for this dataset as well, this value is surprisingly large. There were 112 points in the fitted range, and the degrees of freedom of the model was estimated as  $N_g = 8.3$  (Vestergaard & Hansen 2006). The probability of obtaining a  $\chi_r^2$  of 2.7 given  $N = 112$  and  $N_g = 8.3$  and the a true model is  $\sim 10^{-16}$ . The errorbars were therefore underestimated and were renormalized by  $\sigma_{new} = \beta \sigma_{old}$ , where  $\beta = \sqrt{2.7}$ .

Forward scattering, as determined by Guinier analysis, was used to calculate a model-free estimation of the number of lipids in the lipid core (Fig. S1). The protein concentration of the sample was calculated from a measurement of the UV280 absorption of  $0.65 \text{ cm}^{-1}$ , and an extinction coefficient of  $234600 \text{ cm}^{-1} \text{ M}^{-1}$ , as calculated from the protein sequence, using ExPASy ProtParam ([web.expasy.org/protparam](http://web.expasy.org/protparam)). The forward scattering from the protein-lipid complex (HTL + lipid core) is given as:

$$I(0) = c |\Delta\rho_{HTL}V_{HTL} + \Delta\rho_{LIP}V_{LIP}|^2,$$

where  $c$  is the concentration (number of complexes per  $\text{cm}^3$ ),  $\Delta\rho_{HTL}$  and  $\Delta\rho_{LIP}$  are the excess scattering length densities (scattering contrasts) of the protein and lipid respectively, and  $V_{HTL}$  and  $V_{LIP}$  are the corresponding volumes. The sample was purified with an *E. coli* lipid extract ([avantilipids.com/product/100600](http://avantilipids.com/product/100600)), with known lipid composition. Thus  $\Delta\rho_{LIP}$  could be estimated, and the only unknown was  $V_{LIP}$ , the volume of the lipid core:

$$V_{LIP} = \frac{\sqrt{I(0)/c} - |\Delta\rho_{HTL}| \cdot V_{HTL}}{|\Delta\rho_{LIP}|}.$$

$V_{LIP}$  is calculated by subtraction of two numbers,  $\sqrt{I(0)/c}$  and  $|\Delta\rho_{HTL}| \cdot V_{HTL}$ , equal in magnitude, and each with an uncertainty, which result in a relatively large error on the calculated result. The major contributions to the uncertainty stems from the UV280 absorption measurement used to estimate the molar concentration. We assumed a 15% uncertainty on the concentration measurement, 10% on the estimation of  $I(0)$  and 2% on the estimated volumes of HTL and the lipids. The number of lipids could then be found by dividing  $V_{LIP}$  by the mean volume of the *E. coli* lipids of 1216

$\text{\AA}^3$ , as calculated by the composition (avantlipids.com/product/100600) using known volume for the different lipid components (Armen et al. 1998).

A fit was made where the fitting algorithm was allowed to mix HTL-2 and HTL-3 (Figure 1) to obtain the optimal fit. The intensity of the mix was given as:

$$I(q) = I(0) \cdot [A \cdot P_{HTL-2}(q) + (1 - A) \cdot P_{HTL-3}(q)] + B,$$

where  $A$  is the fraction of the sample in the HTL-2 form.

The goodness of the fits was evaluated using the reduced  $\chi^2$ , given as  $\chi_r^2 = \chi^2 / (N - K)$ , where  $N$  is the number of datapoints and  $K$  the number of fitting parameters. The  $\chi^2$  is defined in terms of the measured experimental intensities  $I_i^{exp}$  and corresponding uncertainties  $\sigma_i$  and the fitted theoretical intensities  $I_i^{fit}$

$$\chi^2 = \sum_{i=1}^N \frac{(I_i^{exp} - I_i^{fit})^2}{\sigma_i^2}.$$

There was a minor contribution of aggregates in the sample, as seen from the upturn in the Guinier plot (SI, Fig. S1). The presence of aggregations were also clear from the “tail” of the  $p(r)$  with a large  $D_{\max}$  of  $\sim 200 \text{ \AA}$  (Fig. 4A). These were taken into account in the final fits (Fig. 4B) by including a fractal structure factor  $S(q)$  to the model, as previously described in (Larsen et al. 2018 – ADD REF TO LIST SOON). Shortly, the a fractal aggregate description was used (Teixeira 1988– ADD REF TO LIST) in combined with the decoupling approximation (Kotlarchyk & Chen 1983 – ADD REF TO LIST) and the form factor of the complex,  $P(q)$ :

$$I_{frac}(q) = I(0) \cdot [A \cdot P_{HTL-2}(q)S_{frac}(q) + (1 - A) \cdot P_{HTL-3}(q)S_{frac}(q)] + B,$$

where  $S(q)$  is the effective form factor after the decoupling approximation was applied. A mean radius of  $R = 42.1 \text{ \AA}$  was used for HTL, corresponding to radius of a sphere with volume equal to the sum of Van der Waals volumes of the atoms in the protein (Svergun et al. 1999 – ADD REF TO LIST). The models were implemented in WillItFig (Pedersen et al. 2013 – ADD REF TO LIST).

### **Molecular Dynamics**

PDB structure files (PDB: 5MG3) were converted to a coarse-grained (CG) description using the Martinize script (Monticelli et al., 2008), with secondary structure defined using DSSP (Define Secondary Structure of Proteins, (Kabsch & Sander 1983)). Additional harmonic bonds were applied between protein beads, with a force constant of  $500 \text{ kJ mol}^{-1} \text{ nm}^{-2}$  and an upper bond length cut-off of 0.9 nm. Topologies for the POPE, POPG and cardiolipin lipids were obtained from Lipid Book (Domański et al. 2010). The systems were solvated in MARTINI water and neutralized with ions to 0.15 M.

Simulations were performed using GROMACS 5.0.4 (Berendsen et al. 1995). Non-bonded interactions were treated with a switch function from 0-1.2 nm and 0.9-1.2 nm for the Coulombic and Lennard-Jones interactions respectively. Temperature and semi-isotropic pressure coupling were achieved with the Bussi-Donadio-Parrinello thermostat (Bussi et al. 2007) and the Parrinello-Rahman barostat (Parrinello & Rahman 1981).

The systems were minimized using steepest descents for 2500 steps of 20 fs, before equilibration of 100 ns at 310 K with 2 fs time steps. Finally, simulations were run for



3  $\mu$ s at 310 K using 20 fs time steps. Simulations were run on Phase 3 of BlueCrystal, the University of Bristol's High Performance Computer (HPC) and with 20 fs integration steps. Images of proteins were made in PyMOL and VMD, and data were plotted with gnuplot.

## Results

### SANS confirms a central lipid core within the HTL

Previous studies show that the purified HTL complex is composed of its constituent protein subunits and significant proportions of lipid and detergent (Botte et al. 2016). The majority of this lipid and detergent is localised at the centre of the complex with the protein at the periphery. Due to the relatively close contrast match points of DDM (21.7% D<sub>2</sub>O) and *E. coli* lipids (13.1% D<sub>2</sub>O) it is difficult to distinguish and separate the scattering contributions from the lipid and the solubilizing detergents. In order to address this, SANS experiments were performed on the HTL using partially deuterated DDM (d-DDM) to quench the detergent signal. The DDM sugar head group and tail moieties were chemically deuterated independently, with differing levels of deuteration, such that the contrast match point of both head and tail group of the d-DDM is at 100% D<sub>2</sub>O.

Purified HTL was detergent exchanged into d-DDM buffer by gel filtration (see Materials & Methods). The d-DDM buffer was made up with 100% D<sub>2</sub>O, so the recorded measurements would be conducted at the d-DDM contrast match point and with minimum incoherent scattering background from the buffer. Thus, only scattering contributions of the protein and lipid components were measured and the detergent rendered effectively invisible.

Guinier analysis of the collected data indicates a radius of gyration ( $R_g$ ) for the HTL in apparent absence of DDM as  $41.2 \pm 0.3$  Å, slightly higher than the calculated theoretical  $R_g$  of 37.0 Å (Supplementary Table 1). The forward scattering,  $I(0)$ , determined by Guinier analysis (Guinier & Fournet 1955) can be used to calculate a model-free estimation of the lipid volume of the HTL (see Materials & Methods). From an  $I(0)$  value of  $0.23 \text{ cm}^{-1}$ , the volume of lipids ( $V_{LIP}$ ) can be estimated to  $12000 \text{ Å}^3$ . Assuming a mean volume of an *E. coli* lipid as  $1200 \text{ Å}^3$  (calculated from (Armen et al. 1998)), the model-free estimation indicates a lipid volume of  $\sim 10$  lipids, supporting a significant lipid-based scattering contribution. However, due to the nature of the determination of  $I(0)$  and the cumulative uncertainties involved in calculating  $V_{LIP}$ , the estimated error on the result was  $\pm 20$ . A more precise estimation could be obtained by model based analysis.

As the forward scattering estimated 10 lipids with a volume of approximately  $12000 \text{ Å}^3$ , a model lipid core of this volume was created, with the height corresponding to a typical lipid bilayer (50 Å). The lipid core was positioned in the central cavity of the EM-fitted HTL structure (PDB: 5MG3; Botte et al. 2016). The structure without lipids was termed HTL-1 and the structure with lipids HTL-2 (Figure 1). Theoretical scattering curves for both HTL-1 and HTL-2 were calculated and fitted to the experimental data, showing that inclusion of a lipid core significantly improved the fit to experimental data (Figure 2A).

### **Probing the Flexibility of SecDF Periplasmic domains by SANS**

In the SecDF-YajC subcomplex, the SecD periplasmic domain 1 (P1), has been observed in two distinct orientations, the F and the I form, with an approximate  $100^\circ$

rotation of P1 between the two structures (Tsukazaki et al. 2011; Furukawa et al. 2017). To further refine the lipid containing HTL-2, and assess the effect of domain flexibility on model fit, a model was created, taking the lipid containing HTL-2 structure and replacing SecDF in the F form with SecDF in the I form (PDB: 5XAM). This model was termed HTL-3 (Figure 1).

An improved fit to the observed data was achieved with a combination of HTL-2 and HTL-3. With the number of lipids fixed to 10, a mixture of 43 $\pm$ 8% HTL-2 and 57 $\pm$ 8% HTL-3 fitted to the data better than either structure separately indicating that the SecD periplasmic domain is flexible in the HTL in solution (Figure 2B). As described below, when the number of lipids was varied, this balance shifted.

### **Refining the number of lipid molecules in the core**

The forward scattering calculation estimated ~10 lipids in the core, but with a high uncertainty (roughly  $\pm$ 20). Theoretical scattering was therefore calculated for a series of HTL-2 and HTL-3 structures containing a range of lipid molecules in the core (0-40), which could be fitted to the experimental data (Figure 3A). The model fits better to the experimental data as the number of lipids increases up to a count of 17 lipids, as assessed by the calculated  $\chi^2_r$  values (Materials & Methods), and rapidly worsens at numbers above this value (Figure 3B).

As previously, the fitting algorithm also took into account the ratio of HTL-2 to HTL-3. Assuming a complete absence of central lipids, the entire sample was calculated to be in the HTL-3 form. As lipid numbers were increased, the calculated proportion of HTL-2 increases. At lipid numbers higher than 25, 100% of the sample was predicted

to be HTL-2 (Figure 3C). The number of lipids and the structural conformation of HTL are thus highly correlated parameters.

The best model fit for the SANS experimental data includes 17 lipids in the core of the complex, with 53% of the protein in the HTL-2 and 47% in the HTL-3, and 1% protein in aggregated form (Figure 4B). Using the calculated  $\chi_r^2$  values, an uncertainty of the lipid content can be calculated at 17 lipids  $\pm 5$  (see e.g. Pedersen et al. 2013). The overall structure and arrangement of the HTL including the lipid core is believed to be well described by the model (Figure 4B). That is, the SANS analysis conjectures flexibility in the complex.

### **Coarse-grained simulation supports the existence of a lipid core**

In order to assess the stability of the HTL complex, and begin characterisation of a central lipid core to the complex as indicated by SANS (Botte et al. 2016) a coarse-grained (CG) molecular dynamics (MD) study was performed. An atomic model of HTL was constructed using *E. coli* YidC (Kumazaki et al. 2015), SecYEG (Tanaka et al. 2015), and *E. coli* homology models from *T. thermophilus* SecDF (Botte et al. 2016; Eswar et al. 2007). These structures were arranged to fit the experimental cryo-EM density of the HTL (PDB: 5MG3). The atomic structures were converted to coarse-grained models using the Martini forcefield (Marrink et al. 2007; Monticelli et al. 2008). The CG HTL was inserted into a simulation box filled with randomly oriented CG lipids (65% POPE, 30% POPG, 5% cardiolipin), and solvated with CG water and ions. After energy minimisation and position restrained equilibration, the system was allowed to self-assemble, forming a clear lipid bilayer around the HTL. Following bilayer formation, and an initial settling period the radius of gyration of the

protein complex settles at approximately 38 Å, corresponding with the scattering data (Figure 5). The HTL was simulated for a total duration of 3 μs (Figure 5) and remained stable. As a check for distortion and stability, and also as a comparison to previous all-atom MD studies on SecY (Allen et al. 2016) the width of the lateral gate of SecY was measured over the course of the simulation. Measured across 3 residue pairs (SecY 119:276, 122:279, 125:283), the distance appeared stable over the duration of the measurement at 1.25 nm (Figure 5), corresponding to previous measurements of the closed SecY lateral gate (Corey et al. 2016).

CG modelling of the HTL complex shows the presence of a stable lipid pool between the trans-membrane domains of all of the subunits of HTL, within the centre of the complex (Figure 5). This contiguous pool of lipids is predominantly located at the interface of the 3 major HTL subcomponents SecYEG, SecDF and YidC, and is stable during the course of a 3 μs simulation. The number of lipids within this island remains between 7 and 13 for the duration of the simulation (Figure 6B). The average number of lipids remaining in the centre of the HTL complex is 9.4 +/- 0.8 lipids, for the final 2 μs of a 3 μs simulation. Lipids are seen to diffuse in and out of the pool, predominantly through the gap between the SecY lateral gate and YidC, which may act as an opening point of the complex. Due to lipid diffusion, the lipid pool fluctuates in shape throughout the simulation, but remains between approximately 20 and 40 Å in diameter depending on the number of lipids present.

The model CGMD model was converted to a full-atom model with 7 POPE and 2 POPG in the core. This model had the overall structure similar to HTL2. A model of HTL-2 with the lipids from the simulations was also generated. The scattering from these models were calculated and compared with data (Figure 7). Intriguingly, HTL-2 with the core of CG MD lipids only fitted slightly better to data than the lipid-free

HTL-1 structure ( $\chi^2_r = 160$  for HTL-2 as compared to 206 for HTL-1). The HTL-3 structure with the 9 CG MD lipids on the other hand fitted the data significantly better ( $\chi^2_r = 31$ ). Including aggregates in the model did not improve the fit. When comparing SANS data with the simulations, it should be noted that in the simulations, the HTL complex is in a lipid bilayer, allowing for dynamic exchange between the bilayer and the lipid core. In the SANS experiment, the HTL complexes are isolated and solubilized in DDM, so the lipid core is “captured” in the center.

### **Discussion**

The results presented show the HTL to be a dynamic complex, unequivocally demonstrating that the individual subunits are arranged around a central lipid core. The SANS data supports a model of the HTL containing  $17 \pm 5$  lipids (Figure 4) at its centre. The fit is improved further by accounting for flexibility of SecDF indicating that with 17 lipids in the core, approximately 47% of the HTL was calculated to have a rotated SecD periplasmic domain. The correlation between the number of lipids in the core of the complex and the conformational variation of certain aspects of HTL (Figure 3) is intriguing, and may point to a level of interplay between structural arrangement of the protein subunits and the volume of lipid content.

The lipid pool at the centre of the HTL complex was observed to be stable for during the CG MD simulations. This correlates well with both the SANS data in this study, and previous structural studies of the HTL, indicating protein is located towards the periphery of the particle in solution with lipid and/or detergent material located

towards the centre (Botte et al. 2016) The number of lipids observed in the central core during the simulation remained between 7 and 13, slightly lower than the calculated 17 from SANS. However, the lipids were observed to diffuse into and out of the core during the simulations, suggesting that there is natural fluctuation of the lipid core volume in bilayer conditions. It may pe

The SecD P1 periplasmic domain remained in close proximity to the to the other parts of the complex may also prevent formation of a periplasmic pool of lipids. SecD interacts with unfolded proteins (Tsukazaki et al. 2011), suggesting it may play a facilitative role in the HTL of binding and facilitating passage of periplasmic domains of membrane proteins as they are inserted by HTL. Insertion of a multi-spanning membrane protein by HTL may sequentially open up the complex as additional trans-membrane helices insert, allowing lipids to fill the periplasmic side of the core. It is proposed that the lipid core may provide substrates with a protected folding environment, in a manner analogous to cytoplasmic chaperones such as GroEL (Xu et al. 1997).

The MD simulation points to a potential gateway in the cytoplasmic membrane between SecY and YidC through which lipids are able to diffuse in and out of the lipid pore (Figure 6). During insertion, YidC is known to function in concert with SecY performing chaperone activities which facilitate correct folding of trans-membrane helices as they sequentially exit the lateral gate of SecY (Nagamori et al. 2004; Zhu et al. 2013). The proposed flexibility of the HTL, could potentially allow YidC to initially facilitate correct membrane protein folding in the central lipid pool, before flexing away from SecY and opening what could be considered a ‘membrane airlock’ to the membrane proper. In this way, the HTL could provide a protected lipidic microenvironment in which proteins can achieve the correct topology as they

exit the SecY lateral gate (facilitated by YidC) before the whole complex opens to allow release of the proteins (or portions of proteins) into the adjacent bilayer.

Soluble chaperone proteins such as GroEL provided protected micro-environments for protein folding. It is therefore not unreasonable to propose a similar system for membrane proteins. YidC has been shown previously to provide important chaperone functions in conjunction with SecY, to help polytopic membrane proteins obtain correct conformations (Zhu et al. 2013). YidC's proximity to both the SecY lateral gate and the lipid core during simulation (Figure 5) suggest how it might perform this function. Cross-linking experiments on insertion substrates have shown that upon an initial insertion event, polytopic membrane proteins cross-link to both SecY and lipids, before being transferred to a YidC/lipid environment upon elongation and insertion of subsequent helices (Urbanus et al. 2001).

It is proposed that the lipid island identified within the HTL may form a lipidic microenvironment, which would facilitate transfer of substrates between SecYEG and YidC during insertion. YidC has been shown to interact with and 'occupy' all 4 helices of the SecYEG lateral gate (trans-membrane helices 2b, 3, and 8) (Sachelaru et al. 2013) before being sequentially displaced by insertion substrates. During the MD simulations (Figure 5), the face of YidC and the SecY lateral gate were adjacent to a lipid 'gateway' connecting the central lipid pool to the external bilayer. A tighter interaction between the face of YidC and the lateral gate of SecY, as has been observed during insertion events (Sachelaru et al. 2013), could act as a convenient way of opening and closing the HTL, allowing sequential release of correctly folded substrate trans-membrane helices into the surrounding bilayer.





## Figures

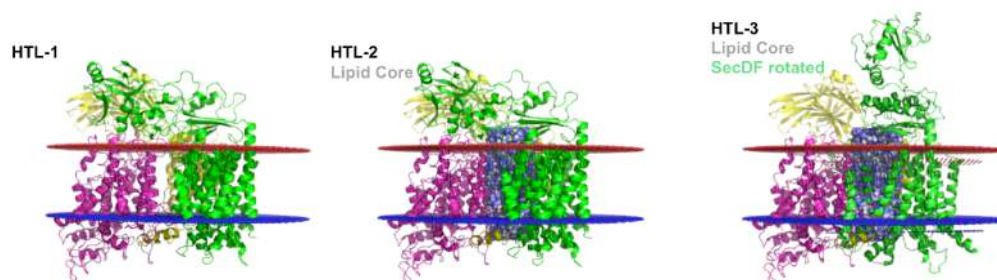


Figure 1) 3 HTL structures used to fit the experimental SANS data. HTL-1 is the starting structure, and is based on the EM fitted structure from Botte et al. 2016. SecYEG is shown in magenta, with SecDF in green, and YidC in yellow. HTL2 is the same protein arrangement with the addition of a lipid core. HTL-3 is the same structure as HTL-2 (i.e. contains lipids) but has had the SecDF P1 domain rotated through 60°. Lipid bilayer planes are marked in red and blue.

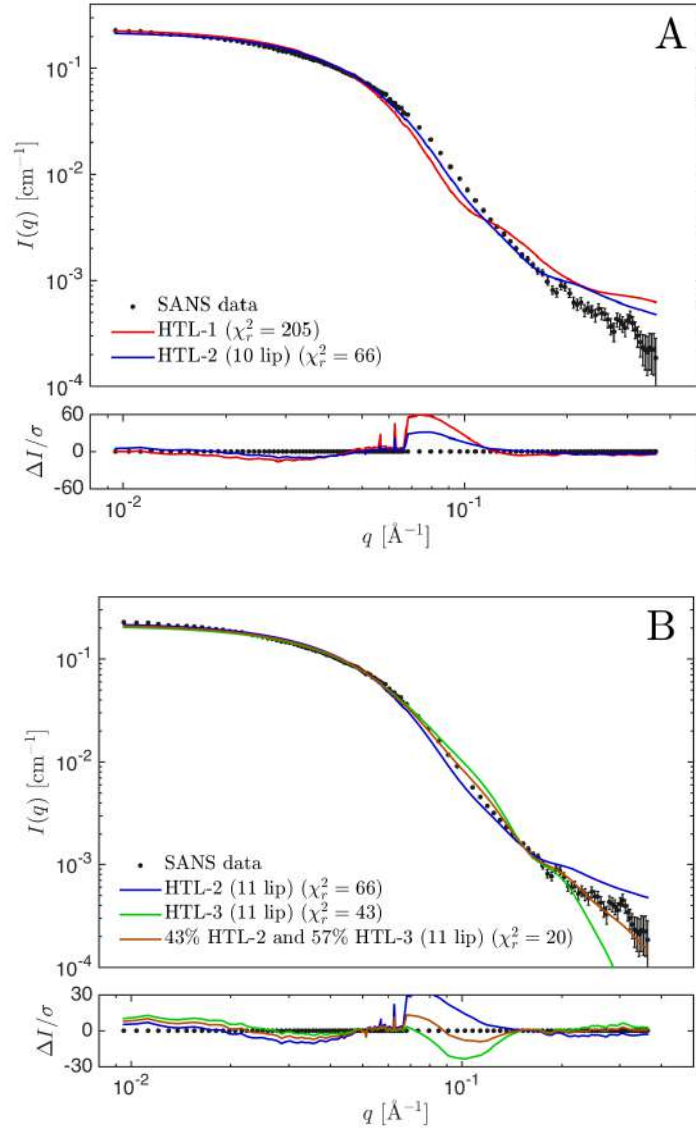


Figure 2) A) Experimental HTL SANS data (black dots) fitted with the HTL-1 structure without lipids (red), and the HTL-2 structure with 10 lipids, as per initial forward scattering calculations (blue). B) HTL SANS data (black dots) with the HTL-2 structure with 10 lipids (blue), the HTL-3 (SecD P1 rotated) structure with 10 lipids (green) and a linear combination of the two structures (brown).

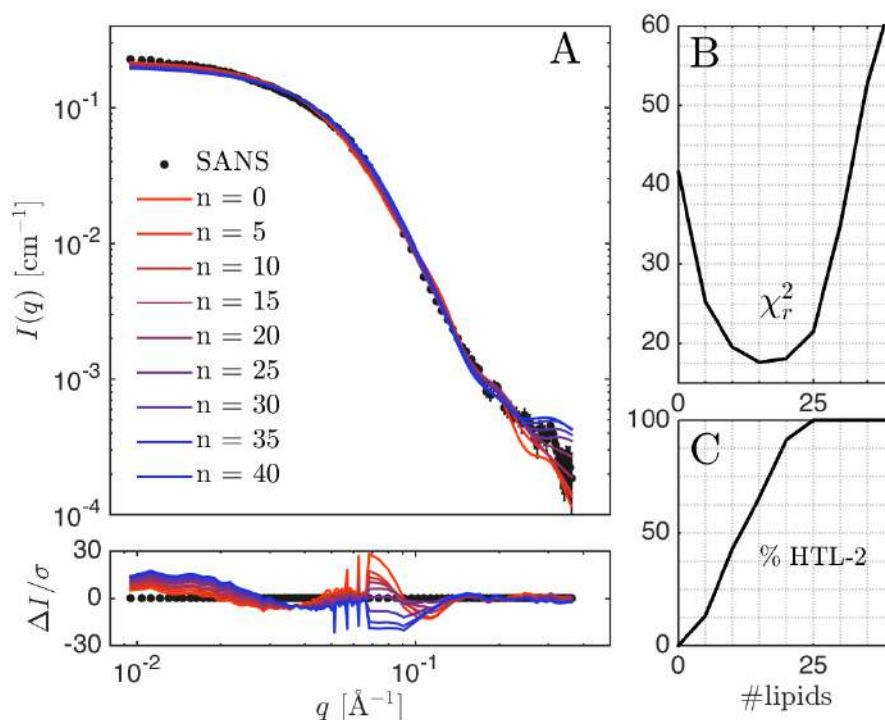


Figure 3) A) Theoretical scattering of a linear combination of model HTL-2 and HTL-3 in which the number of central lipids is varied from 0-40 (red-blue), plotted against experimental HTL data (black dots). B) The  $\chi_r^2$  plot showing the best fit for the number of lipids at  $17 \pm 5$ . C) Plot showing the percentage of HTL-2 in the linear combination of HTL-2 and HTL-3, as a function of the number of lipids.

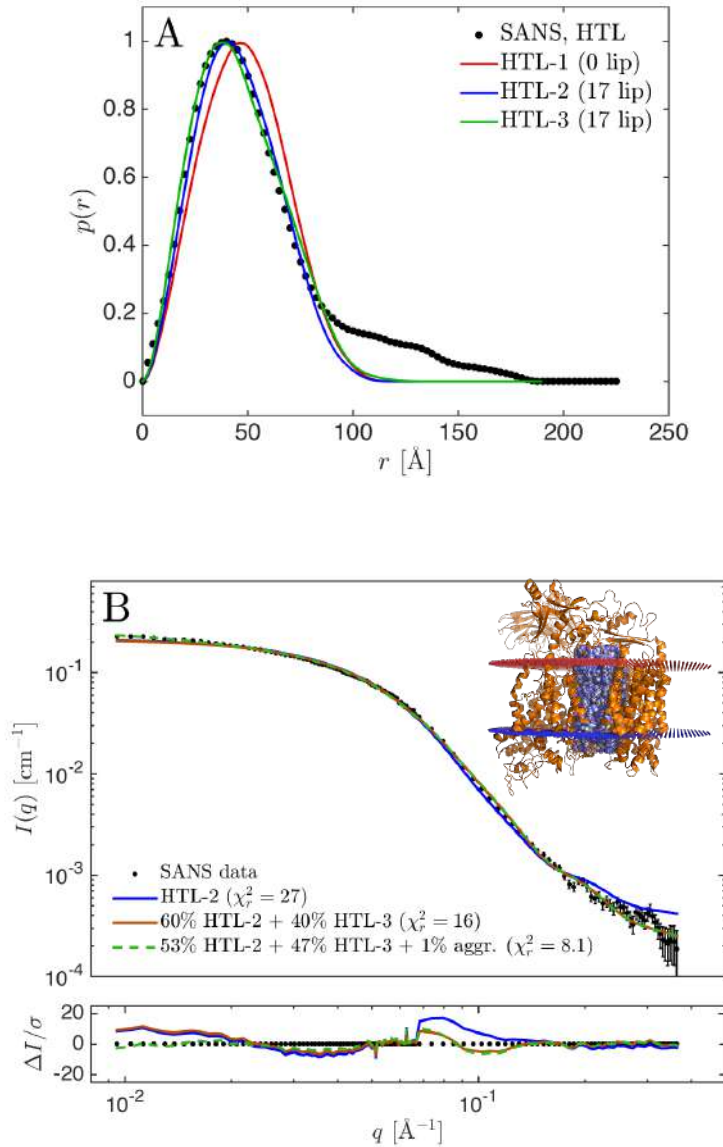


Figure 4) A)  $p(r)$  plot of HTL SANS data (black), HTL-1 (red), HTL-2 (17 lipids, blue) and HTL-3 (17 lipids, green). B) Experimental HTL SANS data (black dots) fitted with the HTL-2 (blue) and a combination of 53% HTL-2 (17 lip) and 3% HTL-3 (17 lip) (brown), and with a combination of 53% HTL-2 (17 lip), HTL-3 (17 lip) and 1% unspecific aggregates. (green dashed line) The inset image shows the HTL2 in cartoon representation (orange) with a lipid core representative of 17 lipids (blue and white). Lipid bilayer planes are marked in red and blue.

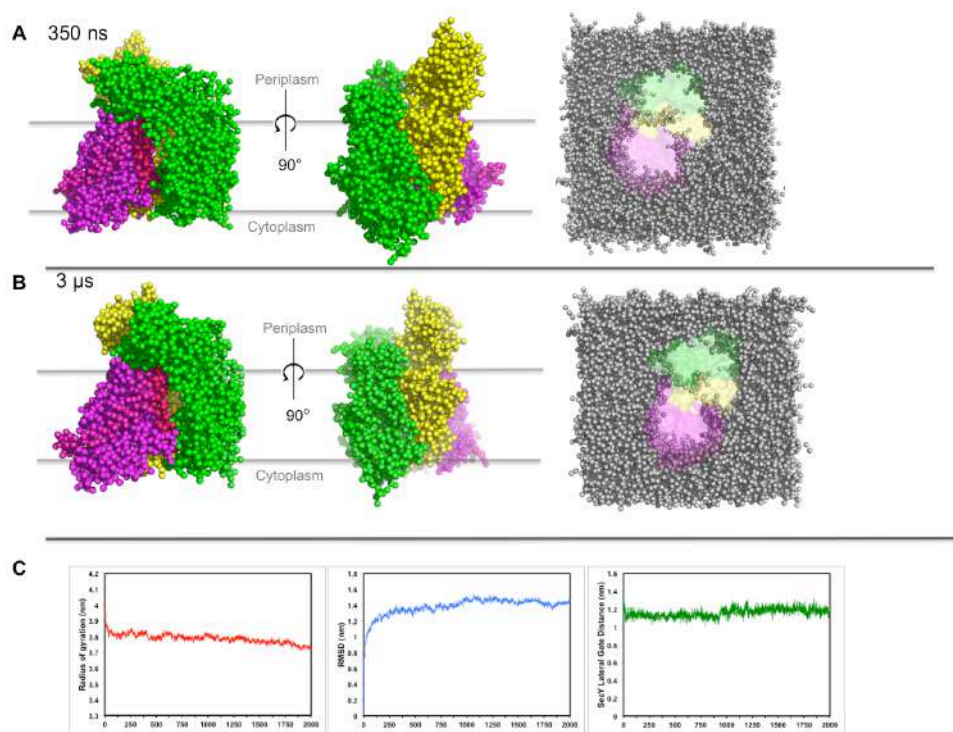


Figure 5) Coarse-grained HTL after 350 ns and 3 μs simulation. A) HTL shown after 350 ns simulation, viewed transversely through the membrane from two orientations, and from cytoplasmic face, showing the lipid arrangement within the complex. SecYEG is shown in Magenta, with SecDF in Green, and YidC in yellow. B) As previous, but after 3 μs of simulation. C) Graphs showing the stability of the structure over 3 ms, from l-r: Radius of gyration, RMSD, SecY lateral gate distance. SecY LG distance measured across the average of 3 residue pairs 119:276, 122:279, 125:283.

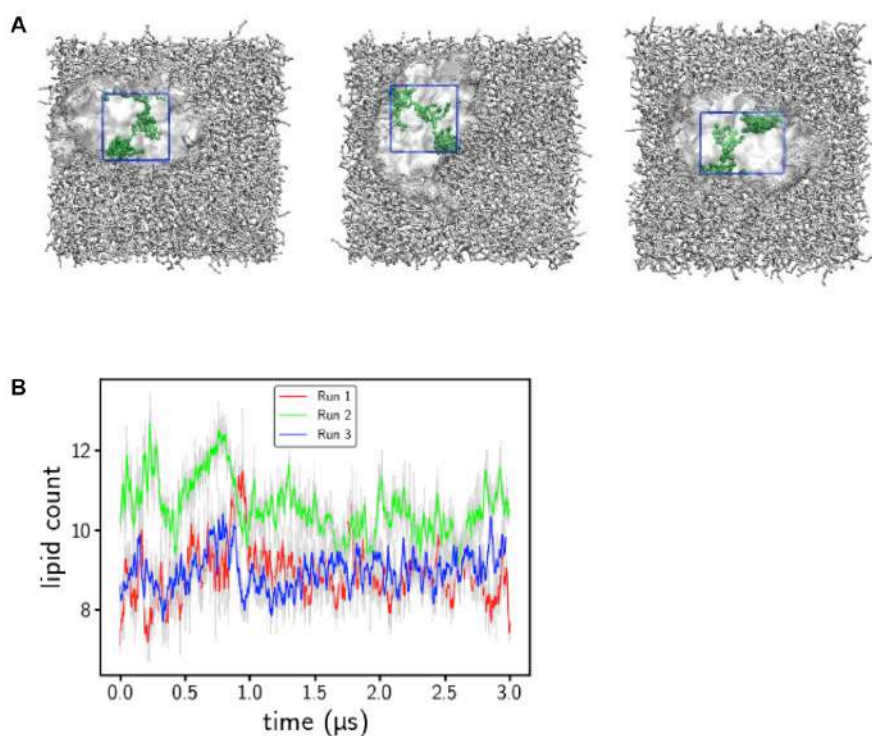


Figure 6) The localisation of the lipids within the HTL during simulations. A) Shots of 3 independent coarse-grained simulations of HTL in a mixed lipid bilayer. In each image, the lipids present in the centre after 3 $\mu$ s simulation are highlighted green. A boundary box was created for each simulation, and lipid presence within the area quantified. B) Graph showing the number of lipids within the core of the HTL, as defined by the boundary box, over the course of the simulation time.

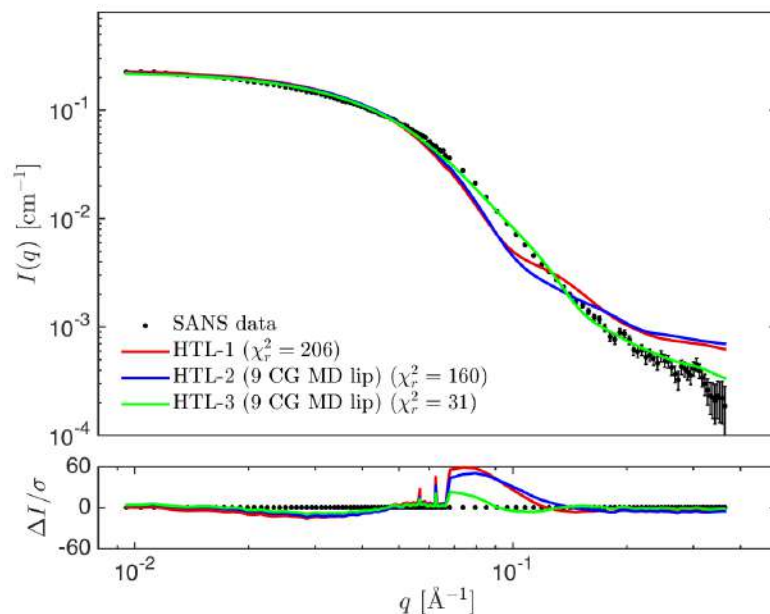


Figure 7) A) Experimental HTL SANS data (black dots) fitted with the HTL-1 structure without lipids (red), and the HTL-2 structure with the 9 lipids from the CG MD simulation (blue) and with the HTL-3 structure with the 9 CGMD lipids (green).

### Acknowledgements

R.M. received a University of Bristol Postgraduate Scholarship. Additional support was gratefully received by I.C. from the BBSRC (BB/M003604/1 and BB/I008675/1). L.A. and A.H.L. thank CoNeXT for co-funding the project. We are grateful to the Institut Laue Langevin for the use of their facilities and assistance of Anne Martel. Additionally, the authors gratefully acknowledge the financial support provided by JCNS to perform the neutron scattering measurements at the Heinz



Maier-Leibnitz Zentrum (MLZ), Garching, Germany. Part of this work is based upon experiments performed at the KWS-1 instrument and the authors would like to thank Henrich Friehlinghaus for support during the experiments. The deuterated DDM (d-DDM) was synthesised by Dr. Tamim Darwish (ANSTO, NSW, Australia).

**References:**

- Allen, W.J. et al., 2016. Two-way communication between SecY and SecA suggests a Brownian ratchet mechanism for protein translocation. *Proceedings of the National Academy of Sciences of the United States of America*, 5, pp.6545–6549.
- Armen, R.S., Uitto, O.D. & Feller, S.E., 1998. Phospholipid component volumes: Determination and application to bilayer structure calculations. *Biophysical Journal*.
- Berendsen, H.J.C., van der Spoel, D. & van Drunen, R., 1995. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1–3), pp.43–56.
- Van Den Berg, B. et al., 2004. X-ray structure of a protein-conducting channel. *Nature*, 427(6969), pp.36–44. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/14661030>.
- Bieniossek, C. et al., 2009. Automated unrestricted multigene recombineering for multiprotein complex production. *Nature methods*, 6(6), pp.447–450. Available at: <http://dx.doi.org/10.1038/nmeth.1326>.
- Botte, M. et al., 2016. A central cavity within the holo-translocon suggests a mechanism for membrane protein insertion. *Scientific Reports*, 6, p.38399. Available at: <http://www.nature.com/articles/srep38399>.
- Brundage, L. et al., 1990. The purified E. coli integral membrane protein SecY/E is sufficient for reconstitution of SecA-dependent precursor protein translocation. *Cell*, 62(4), pp.649–657.

- Bussi, G., Donadio, D. & Parrinello, M., 2007. Canonical sampling through velocity rescaling. *Journal of Chemical Physics*.
- Corey, R.A. et al., 2016. Unlocking the Bacterial SecY Translocon. *Structure*.  
Available at: <http://linkinghub.elsevier.com/retrieve/pii/S0969212616000460>  
[Accessed March 18, 2016].
- Domański, J. et al., 2010. Lipidbook: A public repository for force-field parameters used in membrane simulations. *Journal of Membrane Biology*, 236(3), pp.255–258.
- Duong, F. & Wickner, W., 1997. The SecDFyajC domain of preprotein translocase controls preprotein movement by regulating SecA membrane cycling. *EMBO Journal*, 16(16), pp.4871–4879.
- Eswar, N. et al., 2007. Comparative Protein Structure Modeling Using MODELLER. In *Current Protocols in Protein Science*. Hoboken, NJ, USA: John Wiley & Sons, Inc., p. 2.9.1-2.9.31. Available at:  
<http://www.ncbi.nlm.nih.gov/pubmed/18429317> [Accessed December 28, 2016].
- Furukawa, A. et al., 2017. Tunnel Formation Inferred from the I-Form Structures of the Proton-Driven Protein Secretion Motor SecDF. *Cell Reports*.
- Gold, V.A. et al., 2010. The action of cardiolipin on the bacterial translocon. *Proceedings of the National Academy of Sciences of the United States of America*, 107(22), pp.10044–10049. Available at:  
<http://www.ncbi.nlm.nih.gov/pubmed/20479269>  
<http://www.pnas.org/content/107/22/10044.full.pdf>.

- Görllich, D. & Rapoport, T.A., 1993. Protein translocation into proteoliposomes reconstituted from purified components of the endoplasmic reticulum membrane. *Cell*, 75(4), pp.615–630.
- Guinier, A. & Fournet, G., 1955. *Small-angle scattering of X-rays*, Wiley.
- Hansen, S., 2012. BayesApp: A web site for indirect transformation of small-angle scattering data. *Journal of Applied Crystallography*, 45(3), pp.566–567.
- Hendrick, J.P. & Wickner, W., 1991. SecA protein needs both acidic phospholipids and SecY/E protein for functional high-affinity binding to the Escherichia coli plasma membrane. *Journal of Biological Chemistry*, 266(36), pp.24596–24600.
- Kabsch, W. & Sander, C., 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. *Biopolymers*, 22, pp.2577–2637.
- Komar, J. et al., 2016. Membrane protein insertion and assembly by the bacterial holo-translocon SecYEG-SecDF-YajC-YidC. *Biochemical Journal*, 0(2016), pp.1–35. Available at: <http://biochemj.org/cgi/doi/10.1042/BCJ20160545>.
- Kumazaki, K. et al., 2015. Crystal structure of Escherichia coli YidC, a membrane protein chaperone and insertase. *Scientific reports*, 4, p.7299. Available at: <http://www.nature.com/srep/2014/141203/srep07299/full/srep07299.html>.
- Kumazaki, K. et al., 2014. Structural basis of Sec-independent membrane protein insertion by YidC. *Nature*, 509(7501), pp.516–20. Available at: <http://dx.doi.org/10.1038/nature13167>.
- Lill, R., Dowhan, W. & Wickner, W., 1990. The ATPase activity of secA is regulated

- by acidic phospholipids, secY, and the leader and mature domains of precursor proteins. *Cell*, 60(2), pp.271–280.
- Lomize, M.A. et al., 2012. OPM database and PPM web server: Resources for positioning of proteins in membranes. *Nucleic Acids Research*, 40(D1), pp.370–376.
- Marrink, S.J. et al., 2007. The MARTINI force field: Coarse grained model for biomolecular simulations. *Journal of Physical Chemistry B*, 111(27), pp.7812–7824.
- Midtgaard, S.R. et al., 2018. Invisible detergents for structure determination of membrane proteins by small-angle neutron scattering. *FEBS Journal*, 285(2), pp.357–371.
- Monticelli, L. et al., 2008. The MARTINI coarse-grained force field: Extension to proteins. *Journal of Chemical Theory and Computation*, 4(5), pp.819–834.
- Nagamori, S., Smirnova, I.N. & Kaback, H.R., 2004. Role of YidC in folding of polytopic membrane proteins. *Journal of Cell Biology*, 165(1), pp.53–62.
- Parrinello, M. & Rahman, A., 1981. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*.
- Pedersen, M.C., Arleth, L. & Mortensen, K., 2013. WillItFit: A framework for fitting of constrained models to small-angle scattering data. *Journal of Applied Crystallography*.
- Persson, F., Söderhjelm, P. & Halle, Bertil, 2018. The geometry of protein hydration. *Journal of Chemical Physics*. 148, 215101.

- Pfeffer, S. et al., 2015. Structure of the native Sec61 protein-conducting channel. *Nature Communications*, 6, p.8403. Available at:  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4598622&tool=pmcentrez&rendertype=abstract>.
- Sachelar, I. et al., 2013. YidC occupies the lateral gate of the SecYEG translocon and is sequentially displaced by a nascent membrane protein. *Journal of Biological Chemistry*, 288(23), pp.16295–16307.
- Schulze, R.J. et al., 2014. Membrane protein insertion and proton-motive-force-dependent secretion through the bacterial holo-translocon SecYEG-SecDF-YajC-YidC. *Proceedings of the National Academy of Sciences of the United States of America*, 111(13), pp.4844–9. Available at:  
<http://www.pnas.org/content/111/13/4844.short>.
- Scotti, P.A. et al., 2000. YidC, the Escherichia coli homologue of mitochondrial Oxa1p, is a component of the Sec translocase. *The EMBO journal*, 19(4), pp.542–9. Available at:  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=305592&tool=pmcentrez&rendertype=abstract>.
- Tanaka, Y. et al., 2015. Crystal Structures of SecYEG in Lipidic Cubic Phase Elucidate a Precise Resting and a Peptide-Bound State. *Cell Reports*, 13(8), pp.1561–1568. Available at: <http://dx.doi.org/10.1016/j.celrep.2015.10.025>.
- Tsukazaki, T. et al., 2011. Structure and function of a membrane component SecDF that enhances protein export. *Nature*, 474(7350), pp.235–8. Available at:  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3697915&tool=pmcentrez&rendertype=abstract>.

ntrez&rendertype=abstract.

Urbanus, M.L. et al., 2001. Sec-dependent membrane protein insertion: sequential interaction of nascent FtsQ with SecY and YidC. *EMBO reports*, 2(6), pp.524–529.

Xu, Z., Horwich, a L. & Sigler, P.B., 1997. The crystal structure of the asymmetric GroEL-GroES-(ADP)<sub>7</sub> chaperonin complex. *Nature*, 388(6644), pp.741–750.

Zhu, L., Kaback, H.R. & Dalbey, R.E., 2013. YidC protein, a molecular chaperone for LacY protein folding via the SecYEG protein machinery. *Journal of Biological Chemistry*, 288(39), pp.28180–28194.

## Supplemental Information

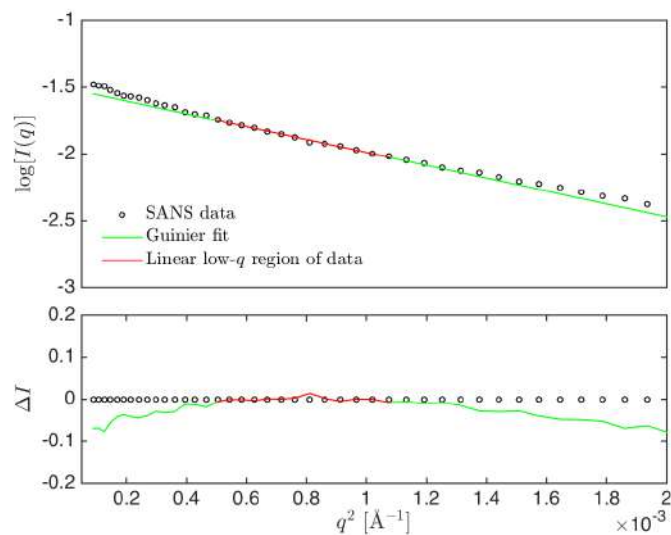


Figure S1) Guinier analysis of the SANS data of HTL-1.  $R_g$  was determined to be  $41.2 \pm 0.3 \text{ \AA}$ , with  $q_{\max} R_g = 1.21$ .  $I(0)$  was determined to be  $0.234 \pm 0.001 \text{ cm}^{-1}$ .

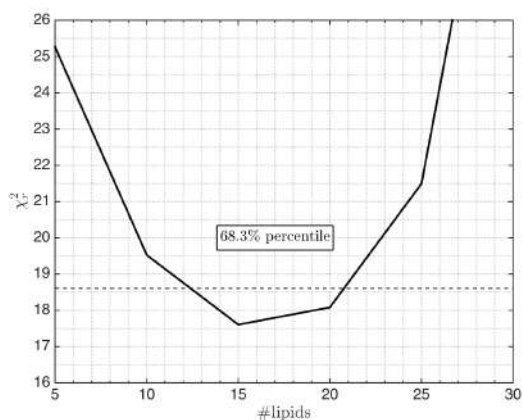


Figure S2) The 68.5 percentile for  $\chi^2_r$  for the number of lipids in the core of HTL. The percentile was used to estimate the uncertainty for the optimal number of lipids,  $17 \pm 5$ .



