New Methods for the Analysis of Small-Angle Solution X-ray Scattering Data from Biomacromolecules

Submitted in accordance with the requirements for the degree of

Doktor der Naturwissenschaften

Yunyun Gao

Universität Hamburg
Fachbereich Chemie
Max-Planck-Institut für Struktur und Dynamik der Materie

April 2020

Evaluators of the dissertation:

Prof. Dr Andrew Torda (Vorsitz, Universität Hamburg)

Prof. Dr Arwen Pearson (Universität Hamburg)

Prof. Dr Henning Tidow (Universität Hamburg)

Prof. Dr Dorota Koziej (Universität Hamburg)

Dr. Christian Löw (EMBL, Hamburg)

Date of the oral defense: 18th September 2020

Declaration/Erklärung

I hereby declare that this doctoral dissertation is my own work and that I have not used any sources other than those listed.

Hiermit erkläre ich an Eides statt, die vorliegende Dissertation selbst verfasst und keine anderen als die angegebenen Hilfsmittel benutzt zu haben.

Würzburg 21.04.2020 Yunyun Gao

Place, Date, Signature: Yunyun Gao

Acknowledgements

I would like to extend my thanks and appreciation to Dr Arwen Pearson for giving me the opportunity to perform this project and making the process so enjoyable. Her supervision, help and motivation have been invaluable in leading me into the biological world and developing the algorithms in this thesis. I have enjoyed her company during the course of my PhD and feel terribly sad that I have to leave to work elsewhere at the completion of this project.

My collaborators have also proved invaluable. I couldn't have completed any of my accomplishments without the support from them. I thank Dr Diana Monteiro and Dr Inokentijs Josts for giving me the opportunity to work with them on their brilliant time-resolved experiments. I thank Dr Mohammad Vakili for bringing me on board to his project and keeping coming to my deserted office instead asking me to come to his. I thank Dr Chris Orr and Hayden Fisher for introducing me to the wonderful scenario of antibodies. I thank Dr Anna Polyakova for allowing me access to her experimental data on Mhp1 solution structures. I thank Diogo Melo for providing me with enormous assistance before and during each beamtime. I thank Dr Maria Kokkinidou for her company and useful suggestions during every beamtime. I thank Dr Haydyn Mertens at P12 and Dr Martha Brennich at BM29 for their patience and those midnight discussions on the beamlines. I would also like to give my special thanks to my summer student, Samuel Holst. His hard work during the summer made Chapter 1 possible.

I had the great privilege to work with Dr Henning Tidow, Dr Martin Trebbin, Dr Godfrey Beddard, Dr Ivo Tews, Dr Edward Snell and Dr Thomas Grant. The time spent with them in both physical and virtual meetings was always an eye opening experience. All their individual skills and expertise have been required to make this project possible.

This work was made possible by the funding provided by the Max Planck Institute for the Structure and Dynamics of Matter as well as the International Max Planck Research School for Ultrafast Imaging and Structural Dynamics (IMPRS-UFAST). I thank these bodies deeply for funding, Dr Sonia Utermann for selecting my application for funding and Dr Neda Lotfiomran and Dr Julia Quante for the support provided.

It is hard to finish a PhD project let alone study abroad alone. Therefore, I would like to thank those who gave me emotional support during my PhD. Jo, he joined the group at the same time as I did and she also finished about the same time as I would. The journey together with

her means a lot. Diogo, he and his girlfriend helped me avoid being homeless. Henry, he is always nice and perhaps the nicest person I have been working with. Marta, she is the western life mentor of mine. Elena, she kept me away from lonely Christmas and new years and treated me as one of her family members. Susanne, the combination of humour and toughness of her made me enjoy the countable lab days. Heike, every bureaucracy I experienced in Elmshorn couldn’t be solved without her.

Last but definitely not least, I would like to thank my family, without them and their support this would not mean anything.

Abstract

Small-angle X-ray scattering (SAXS) of macromolecules in solution is subject to ambiguity. At each stage of SAXS data processing, subjective decisions are involved to a greater or lesser degree. This is due either to a lack of appropriate statistical tools or because the processing procedure is unstable (ill-posed). This thesis aimed to develop new tools for solution SAXS from (bio-)macromolecules to improve the objectivity of data analysis. Based on the data collected from experimental configurations at modern synchrotron SAXS beamlines (i.e. time-resolved SAXS, size-exclusion-chromatography SAXS (SEC-SAXS)), the following questions were asked: How can one assess the data quality objectively? How can the data interpretation be conducted in a model-free way? How can the fidelity of SAXS-oriented modelling be improved?

Chapter 2 of the thesis describes an algorithm to determine the correctness-state score (CSS), with which the quality of post-processed data from SEC-SAXS can be objectively evaluated. Chapter 3 explores the deconvolution of time-resolved SAXS observations with few to no assumptions. Chapter 4 introduces an algorithm to model membrane proteins and a workflow to model proteins with varying flexibility, both designed to reject any potentially non-physiologically-likely conclusions. Finally a user-friendly graphical user interface was developed for the methods and figurative demonstrations developed and used in this thesis.

Kurzfassung

Die Röntgenkleinwinkelstreuung (SAXS) von Makromolekülen in Lösung ist mehrdeutig. In jeder Phase der SAXS-Datenverarbeitung sind subjektive Entscheidungen bis zu einem gewissen Grad involviert. Dies ist entweder auf das Fehlen statistischer Werkzeuge zurückzuführen, oder der Verarbeitungsprozess ist instabil (d.h. inkorrekt gestellt). Diese Arbeit zielt darauf ab, neue Werkzeuge zur Lösung von SAXS-Daten aus (Bio-)Makromolekülen zu entwickeln, um die Objektivität der Datenanalyse zu verbessern. Basierend auf den Daten, die aus experimentellen Konfigurationen an den modernen Synchrotron-SAXS-Strahlführungen (z.B. zeitaufgelöstes SAXS, size-exclusion -chromatography SAXS (SEC-SAXS)) gesammelt wurden, werden folgende Fragen gestellt: Wie kann man die Datenqualität objektiv beurteilen? Wie kann man die Dateninterpretation modellfrei durchführen? Wie kann die Treue der SAXS-orientierten Modellierung verbessert werden?

Kapitel 2 beschreibt einen Algorithmus, den correctness-state score (CSS), mit dem die Qualität bearbeiteter Daten aus SEC-SAXS objektiv bewertet werden kann. Kapitel 3 untersucht die Möglichkeit, zeitaufgelöste SAXS-Beobachtungen mit wenigen bis gar keinen Annahmen zu dekonvolieren. Kapitel 4 stellt einen Algorithmus zur Modellierung von Membranproteinen und den Arbeitsablauf bei der Modellierung einer Reihe von Proteinen mit unterschiedlicher Flexibilität vor. Beide sind so konzipiert, dass sie eine mögliche nicht-physiologische Schlussfolgerung zurückweisen. Es wurden weitere Anstrengungen unternommen, um eine benutzerfreundliche Schnittstelle für die in dieser Arbeit verwendeten Methoden und bildlichen Darstellungen zu entwickeln.

Glossary

ATP - adenosine triphosphate

NPE-ATP - 1-(2-nitrophenyl)-caged ATP

BSA - bovine serum albumin

CG - conjugate gradient method

CHROMS - chromatographic signal

CSS - correctness-state score

CSV-CORMAP - cumulative first-ranked singular-values correlation map

DDM - *n*-dodecyl-*β*-D-maltopyranoside

DESY - Deutsches Elektronen-Synchrotron, Hamburg, Germany

DTW - dynamic time warping

EMG - exponential modified Gaussian

ESRF - European Synchrotron Radiation Facility, Grenoble, France

FWHM - full width at half-maximum

GMG - half-Gaussian modified Gaussian

IMP - integral membrane protein

L-BFGS-B - limited-memory Broyden–Fletcher–Goldfarb–Shanno for bound-constrained optimization

LMTR - Levenberg–Marquardt least square with trust region reflective optimization

MD - molecular dynamics

Mhp1 - Na^+-hydantoin membrane transport protein

MX - macromolecular crystallography

NBD - nucleotide-binding domains of a bacterial lipid flippase, MsbA

NITRO - nonlinear interior point trust region optimizer

NLL - negative log likelihood

NM (as compound) - *n*-nonyl-*β*-D-maltopyranoside

NM (as minimization algorithm) - Nelder-Mead method

PDB - protein data bank

PDC - protein-detergent complex

PDMAm - poly(N,N-dimethylacrylamide)

PMEA - poly(2-methoxyethyl acrylate)

RSS - residual sum of squares

SAXS - small-angle X-ray scattering

SEC - size exclusion chromatography

SIMD - single instruction multiple data

SNR - signal-to-noise ratio

SVD - singular value decomposition

SSRL - Stanford Synchrotron Radiation Light Source, Stanford, USA

THF - Tetrahydrofuran

TNC - truncated Newton method

XFEL - X-ray free electron laser

XXS - X-ray scattering signal

Table of Contents

Chapter 1

Introduction

Small angle X-ray scattering (SAXS) in solution plays a unique role among methods for the structural determination of proteins. 1) The solution-based approach of SAXS is convenient as well as efficient in terms of both sample preparation and data collection. 2) X-ray crystallography does not reveal the structure of proteins (especially for membrane proteins) in their native biological environment but rather the structure of the protein packed into a crystal lattice (Stansfeld et al. 2015). Solution-based SAXS provides an alternative situation where conformational distortions due to crystal packing forces are largely eliminated. 3) When combined with intense X-ray sources, SAXS has the potential to monitor large-scale changes (Levantino et al. 2015), such as molecular folding, protein aggregation and tertiary structure rearrangement, in a time-resolved manner.

The aim of this thesis was to improve solution SAXS data analysis by employing algorithmic methods.

1.1 Solution Small-angle X-ray Scattering of Biomacromolecules: A Historical Perspective

Two years before the first high-resolution structures of myoglobin and haemoglobin were determined by M. Perutz and J. Kendrew in 1957, methods for extracting structural information of non-crystalline samples from scattering patterns with almost no use of Bragg's law had already made much progress (Guinier et al. 1956). A. Guinier, G. Fournet and other pioneers in the field realised that the angular region that SAXS covers was suitable for particles with a radius of gyration (R_g) of the order of 10 to 50 Å. The corresponding molecular weight of such particles is from 5,000 to 250,000 g/mol, which is the order of magnitude of the molecular weights of biological macromolecules, for example, proteins.

Guinier compared the works of Fournet on hemoglobin in solution (Fournet 1951) and Perutz and coworkers on crystallized hemoglobin (Perutz 1949) and drew the conclusion that the two methods represented completely different levels of challenge, in terms of carrying out the experiment. The crystallographic method requires a considerable amount of work both in the preparation of a usable crystal and in the interpretation of the resulting diffraction data, whereas solution-SAXS was already considered a routine operation 60 years ago. However, it was quite evident that the information obtained from a SAXS experiment concerned only the exterior form and not the structural details of the molecule, while much more information on the structure was contained in the diffraction pattern of crystallized proteins. In the following decades, the details provided by X-ray crystallography have become superordinate: macromolecular crystallography (MX) has been supplying unparalleled excitement to the entire biological community, whereas solution SAXS is considered a supplementary method (Forstner 2000; Grant et al. 2011; Kikhney & Svergun 2015; Pearson et al. 2015). Despite it not being as "sexy" as crystallography, considerable endeavours have been made to improve the details that can be obtained from solution SAXS. Since Guinier's inspiring work, methods to extract ever more structural information about proteins from SAXS data , such as shape, physicochemical parameters, secondary, tertiary and quaternary structure, as well as conformational changes, have been developed (Glatter & Kratky 1982). The theoretical improvements contributed by O. Glatter and G. Porod, the experimental techniques by O. Kratky and K. Holmes, along with the many other important scientific results obtained by SAXS have helped biologists understand, beyond the structure, the solution behavior of proteins. Rigorous numerical recipes were given by S. Doniach (Doniach 1985), P. Cachon (Chacón et al. 1998), P. Moore (Moore 1980) L. Feigin and D. Svergun (Feigin & Svergun

1987), showing the enormous potential of solution SAXS for analyzing the spatial structure of biomacromolecules. Applied mathematical instruments, such as numerical approximation, orthonormal expansion and regularization of ill-posed problems, introduced by them and other researchers, make it possible to evaluate SAXS data, extract structural information and optimize real-space modelling using modern computers.

Advances in instruments have also played a huge role in the development of SAXS research. Due to the use of diluted solution, the signal for solution SAXS is relatively weak and thus long exposures are required. Intense and stable X-ray sources are therefore major requirements of SAXS experiments. The slit-collimator invented by O. Krakty, and named after him (compact Kratky camera), made the collection of SAXS data down to very low scattering angles possible, provided there is enough flux. It is still one of the major designs for modern commercial SAXS instruments. However, the data must be numerically desmeared to yield the true scattering pattern. Since the 1970s, the gradually increasing flux of synchrotron radiation, the improvement of X-ray optics (e.g. point collimator systems) and the modernization of data management have facilitated high-throughput SAXS experiments (Hura et al. 2009). On modern synchrotron beamline as well as the X-ray free electron lasers (XFEL), SAXS data can be collected, processed and analysed in a few minutes or even seconds (Svergun et al. 2013). This dramatically broadens the use of SAXS for biomacromolecules. For example, as the proteins being studied become increasingly complex, SAXS can be conveniently used for rapid screening of samples in various buffer conditions to identify and optimize crystallization conditions.

Nowadays, *in situ* techniques and advanced protocols at synchrotron beamlines and XFEL end stations have enormously expanded the scale-limits measureable by SAXS in both space and time. The widely applied in-line SEC-SAXS (size exclusion chromatography coupled SAXS) has enabled the monodispersity requirement to be sufficiently fulfilled for most biological systems (Hopkins et al. 2017; Jeffries et al. 2016; Malaby et al. 2015; Ryan et al. 2018). Automated pipelines provide the possibility of remote management (De Maria Antolinos et al. 2015). Many novel applications have rapidly emerged: protein dynamics validation (Cammarata et al. 2008; Chen & Hub 2014; Vestergaard 2016), time-resolved experiments (Cammarata et al. 2008; Josts et al. 2018a; Levantino et al. 2015), ensemble selection (Antonov et al. 2016; Cheng et al. 2017; Tria et al. 2015), and the use of SAXS data as a refinement constraint for serial crystallography (Ayyer et al. 2016). The prospect of a solution SAXS has gone beyond a merely complementary method to X-ray crystallography.

1.2 From Data to Models: the Current Pipeline of Solution SAXS Practice

1.2.1 Theory

The fundamental formula of SAXS has not changed since it was first expressed by P. Debye in his famous article "*Zerstreuung von Röntgenstrahlen*" (Debye 1915):

$$I(\boldsymbol{q}) = \Sigma_{\mathrm{i}}\Sigma_{\mathrm{j}}\, \boldsymbol{\psi}_i\boldsymbol{\psi}^*{}_j = \Sigma_{\mathrm{i}}\Sigma_{\mathrm{j}} f_i f_j \exp[-i\boldsymbol{q}\cdot(\boldsymbol{r}_i - \boldsymbol{r}_j)] = \Sigma_{\mathrm{i}}\Sigma_{\mathrm{j}} f_i f_j \sin(qr_{ij})/qr_{ij} \qquad (1.1)$$

where I is the scattering intensity; $\boldsymbol{q}$ is the scattering vector (or momentum transfer, $|\boldsymbol{q}|$ = $4\pi\sin(\theta)/\lambda$, where θ is the diffraction half-angle and λ is the wavelength); $\boldsymbol{\psi}$ is the scattering amplitude; r_{ij} is the euclidean norm of $\boldsymbol{r}_i - \boldsymbol{r}_j$. f_i is the measure of the scattering factor of the i-th atom.

The essential theoretical simplifications of classical SAXS inherently exist in this formula: the system is statistically isotropic (in terms of both space and time) so that the phase factor (the exponential term) can be spherically averaged over all directions of $\boldsymbol{r}$.

The Debye function also assumes that the scattering intensities are added to give the total diffraction pattern. However, for a dynamic system (i.e. non-ideal solution), the scattering is better described as a function of the microscopic state of the system. In this case, the whole system can be treated as one disordered object consisting of N identical particles. Separating the terms with $i = j$ and averaging over the ensemble, from (1.1) the scattering intensity can be written as the sum of the self-correlative term and the interference term

$$I(q) = Nf^2 + \Sigma\Sigma_{i\neq j} f_i f_j \sin(qr_{ij})/qr_{ij} \qquad (1.2)$$

According to Zernike and Prins (Zernike & Prins 1927), a radial distribution function $P(r)$ can be introduced to express the interference term. Statistically, each particle has the same surroundings which are also isotropic. Considering a single particle, the mean value of the probability that another particle will be found in the volume element dV at a distance r away is (N/V)dV. Any deviation from this may be accounted for by a factor $P(r)$. Assume the impenetrable distance of an individual particle is D. In the range of $r < D$, P(r) = 0 due to the impenetrability. At long range, where $r >> D$, $P(r)$ = 1. If there is no long range order, the interference term will be exactly canceled. Only the difference 1 − P(r) is relevant for the scattering. Equation (1.2) then takes the form

$$I(q) = NF(q)\{1\text{-}N/V\int_{0}^{\infty} 4\pi r^{2}[1 - P(r)]\mathrm{d}r \sin(qr)/qr\} \tag{1.3}$$

where $F(q)$ represents the average scattering intensity of a particle and the latter term is the statistical function governing the particle's arrangement.

Equation (1.3) is one of the most important analytical forms for SAXS, as the function is expressed as the product of the form factor of an atom and the term contains all interparticle interference. Similar equations can be derived for non isotropic systems and nonidentical particles (Feigin & Svergun 1987).

1.2.2 Sample Preparation

SAXS requires only a suitable contrast between the particle of interest and background, whereas for most biological solution SAXS experiments and the corresponding structural interpretation, stable samples in terms of composition and intermolecular interaction are also essential. The stability of biological samples is known to be an issue compared to inorganics, nanoparticles or polymers in solution (Chi et al. 2003). Therefore, it is critical to prepare buffers with appropriate pH, cosolute/salt type and concentration, and surfactants where appropriate. Except for samples with intrinsic polydispersity (i.g. multi-domain proteins) or large second virial coefficients (i.e. native oligomer), weak-monodispersity and non-interacting protein solutions are required. To fulfil these requirements, the concentration of either proteins or additives should be low enough that the interparticle interactions are negligible (Weyerich et al. 1999). A typical SAXS measurement (in so-called "batch" mode) consists of measuring the scattering from the biomacromolecular solution and the scattering from a second solution with exactly the same composition as the one in which the protein is dissolved. For a solution containing mixed states, it is hard to either pre-separate the complex or prepare a matching buffer. For example, solutions of membrane proteins are particularly challenging as when a membrane protein is solubilized with a detergent the solution always contains free detergent micelles. In practice, it is not easy to ensure that an equilibrium between protein-detergent complex (PDC), micelles and free detergent molecules has been reached. Therefore, it has become common practice to measure such samples (e.g. native mixture, membrane protein) using SEC-SAXS (Jeffries et al. 2016; Ryan et al. 2018). A protocol for preparing monodisperse macromolecular samples for successful synchrotron SAXS experiments is summarized in (Jeffries et al. 2016).

1.2.3 Data Collection and Data Preprocessing

Synchrotron SAXS

The modern synchrotron SAXS beamlines, including SIBYLS (ALS), B21 (DIAMOND), EMBL P12 (DESY), BM29 (ESRF), BL4-2 (SSRL), BioCAT (APS), SWING (Soleil), SAXS beam line at ANSTO, BL23A1 (NSRRC) etc. have hugely reduced the difficulty of conducting SAXS experiments (Acerbo et al. 2015; Hopkins et al. 2017; Pernot et al. 2013). In the main, these SAXS beamlines have optics allowing energy tunability from 7 to 15 keV, fluxes ranging from 10^{13} to 10^{15} photons/s and beam focusing systems which deliver improved X-ray beam characteristics, resulting in reduced parasitic scattering, shorter exposure times and extend the detectable q-range. In batch mode, samples are usually placed in a temperature-controlled automatic sample changer. Once ready, samples of 20 - 50 µL are injected into an on-axis quartz capillary and exposed to X-rays. In the SEC-SAXS mode, a HPLC system is integrated and delivers samples through a SEC column, enabling SAXS measurement of the progressively eluted fractions and providing a good estimation of matched buffer background if sufficient column equilibrium has been performed.

Pattern, Profile and Background Reduction

In a SAXS experiment, the scattered photons are recorded as counts on the detector. This two dimensional array of photon counts is a measurement of the scattering pattern. For solution SAXS, particles are assumed to be randomly oriented. The scattering is thus isotropic and has no azimuthal dependence. The only dependence of the scattering intensity is on the radial position corresponding to the scattering vector (or the scattering angle). The two dimensional array of photon counts can thus be azimuthally averaged about the direct beam. This so-called radial integration converts the photon counts to a one dimensional profile, and can be further corrected to take into account the instrumental parameters (Kieffer & Wright 2013). If the buffer composition of solution and solvent are exactly matched, then the scattering of the dissolved macromolecule $I_{\mathrm{sample}}(q)$ is simply ,

$$I_{\text{sample}}(q) = I_{\text{solution}}(q) - I_{\text{solvent}}(q) \tag{1.4}$$

where $I_{\text{solution}}(q)$ is the scattering of the sample solution; $I_{\text{solvent}}(q)$ is the scattering of the matched buffer (Svergun et al. 2013).

Absolute Scaling

In order to perform structural analysis, the scattering intensity must be brought onto an absolute scale. This is normally done by comparison to the X-ray scattering from pure water. Since the forward scattering intensity of water $I_{\text{water}}(0) = \rho^2 k_{\text{B}} T \chi_{\text{T}}$ is only temperature-dependent (Guinier et al. 1956), one can calculate the absolute scattering $I_{\text{abs}}(q)$ using (Orthaber et al. 2000)

$$I_{\text{abs}}(q) = I_{\text{sample}}(q) / \boldsymbol{C}\,[\rho^2 k_{\text{B}} T \chi_{\text{T}} / I_{\text{water}}(q)] \tag{1.5}$$

where $\boldsymbol{C}$ is a constant proportional to the product of flux, detector sensitivity and exposure volume; ρ is the scattering length density; k_{B} is Boltzmann's constant; T is the temperature; χ_{T} is the isothermal compressibility of water; $I_{\text{water}}(q)$ is the experimentally measured water scattering.

The value of ρ can be found at https://sld-calculator.appspot.com/. The values of $I_{\text{water}}(0)$ and χ_{T} can be found at http://PhysChem.kfunigraz.ac.at/.

Experimental Errors

Like any result of physical measurements, SAXS data include experimental errors. Since SAXS patterns are measured by photon-counting detectors, the number of photon counts per pixel follows the Poisson distribution (Bevington et al. 1993). The true values of counting errors are always unknown and have to be estimated from the experimental data assuming Poisson statistics (Franke et al. 2015). This error propagates throughout the radial integration, the absolution scaling and background subtraction steps that eventually yield the scattering profile $I(q)$ with its associated standard deviation $\sigma(q)$. Due to the fact that the uncertainty propagates quadratically, a mismatched buffer for background determination will hugely degrade the signal-to-noise ratio (SNR), especially when the sample concentration is low. Another source of error are systematic errors. These include, for example, gaps between detector modules, hot pixels, non-uniform response of the detector and variation in the incident beam. All these factors may lead to additional errors during the data processing, and the standard deviation estimated purely from statistical error propagation may underestimate the real uncertainty of the experiments (Svergun et al. 2013). Another often ignored point which makes the Gaussian-like error model in the subtraction inappropriate at high q is that scattered intensity in this region is usually very low, and errors are dominated by counting statistics i.e. strictly in the Poisson regime (R. Rambo, private discussion).

Radiation Damage

For synchrotron SAXS, another unavoidable problem exists during the data collection, X-ray induced radiation damage. X-ray induced radiation damage can cause macromolecule aggregation, fragmentation and unfolding, and SAXS is extremely sensitive to such large-scale alterations (Hopkins & Thorne 2016). Unlike MX, cryogenics cannot be employed in SAXS measurements to minimize radiation damage. Radiation damage is one of the major obstacles to successful SAXS data interpretation (Brookes et al. 2016; Franke et al. 2015; Jeffries et al. 2016). To reduce radiation damage methods including constant/oscillating flow, scavengers, exposure period slicing and intensity attenuation, have been applied, although these are not without cost in terms of higher sample consumption, worse SNR, requirements for more data storage space and less incident flux, respectively. In addition, as the effects of radiation damage are different for different proteins and different experimental setups (Hopkins & Thorne 2016), it is also very hard to rescue data contaminated by radiation damage after the experiment.

1.2.4 Data Interpretation

The common parameters derived from SAXS analysis for a monodisperse, homogeneous, interference-free system have been the subject of an excellent review (Putnam et al. 2007). A brief summary of the calculations of these parameters is given here.

Pair Distribution Function

For a homogeneous particle, the Debye formula can be expressed in a continuous form as

$$I(q) = 4\pi \int_0^{D_{\max}} p(r)\frac{\sin(qr)}{qr}\mathrm{d}r \tag{1.6}$$

where $p(r)$ is the pair distribution function. $p(r)$ can be considered as the counts of the number of distances within the interval r and $r + \mathrm{d}r$. $D_{\max}$ is the largest distance between any two points inside the particle. To be specific, $D_{\max}$ is weighted by the product of the excess scattering length densities at these two points. It is defined that $p(0)$ and $p(r \geq D_{\max}) = 0$.

Since the Debye formula is exactly a Fourier sine transform, one can write the inverse Fourier transform as

$$p(r) = \frac{1}{2\pi^2}\int_0^{\infty} I(q)qr\sin(qr)\mathrm{d}q \tag{1.7}$$

However, equation (1.7) is not usually used directly because $I(q)$ must be measured over the full q interval $[0, \infty]$ which is obviously non-trivial. Practically, therefore $p(r)$ is approximated by the indirect Fourier transform method (Brunner-Popela & Glatter 1997; Svergun 1992).

Guinier Approximation

The Guinier approximation is deduced from the Taylor expansion of (1.6) and the expansion of a negative exponential. The resulting formula is the famous Guinier’s law,

$$I(q) \approx I(0)\exp(-\frac{q^2R_{\mathrm{g}}^2}{3}) \quad (0 < q < 1/R_{\mathrm{g}}) \tag{1.8}$$

where R_{g}, the second moment of $p(r)$, is the radius of gyration and $I(0)$ is the scattering intensity at zero-angle (forward scattering).

A Guinier plot is a linear regression of $\ln(I(q)) \sim q^2$ on the dataset at low-q region. Hence, $\ln(I(0))$ can be read from the zero intercept of the Guinier plot, and R_{g} can be deduced from the slope of the Guinier plot. For SAXS, R_{g} can be considered as the electronic radius of gyration of the particle about its electronic center of mass. R_{g} for most macromolecules can be regarded as a measure on the hydrodynamic shape of the particle.

Additional important information that can be deduced from the Guinier plot is the effective molecular weight of the protein (M), which is the molecular weight “visible” by X-rays. If absolute scaling has been conducted, $I(0)$ is the value of the differential scattering cross-section in the forward direction relating to the total excess scattering length density. The relationship between $I(0)$ and M can be expressed as

$$I(0) = \frac{cr_{\mathrm{e}}^2 N_{\mathrm{A}}}{M}(\Delta\rho\overline{v})^2 \tag{1.9}$$

where c is the protein concentration; r_{e} is the electron radius (scattering length of an electron); N_{A} is Avogadro's number; $\overline{v}$ is the partial specific volume of a protein; $\Delta\rho$ is the electron density contrast of protein and detergent relative to the solvent.

Porod Plot and Kratky Plot

According to Porod’s Law, a quasi two-phase system with a diffusion phase boundary (i.e. hydration layer) and with constant electron density contrast has a decay of scattering proportional to q^{-4} at high angles. Hence in a plot of $I(q)q^4$ *versus* q^4 (Porod plot), a plateau (Porod plateau) at higher-q suggests the protein has a compacted conformation (Rambo and Tainer 2013).

Minor excesses of Porod plots above the Porod plateau suggest the protein has adopted an intermediate state between globular and random coiled. A Kratky plot is a plot of $q^2I(q)$ *versus* q. Kratky plots can qualitatively assess the flexibility and/or degree of unfolding in samples. Unfolded (highly flexible) proteins have a significant increase in the Kratky plot at high q, while compact, globular proteins will only have a Gaussian-like peak. A partially unfolded (flexible) protein may have a combination of the Gaussian-like peak and the increase at high-q. A dimensionless Kratky plot is a plot of $(qR_g)^2I(q)/I(0)$ v*ersus* qR_{g}. The peak position of the first feature in the dimensionless Kratky plot is a semi-quantitative indicator of flexibility. For a compacted globular protein, the peak position is around $qR_{\mathrm{g}} \sim \sqrt{3}$. The Porod and Kratky plots reflect the thermodynamic state of the protein. The changes in these plots semi-quantitatively implicate changes in the thermodynamic state of the given protein.

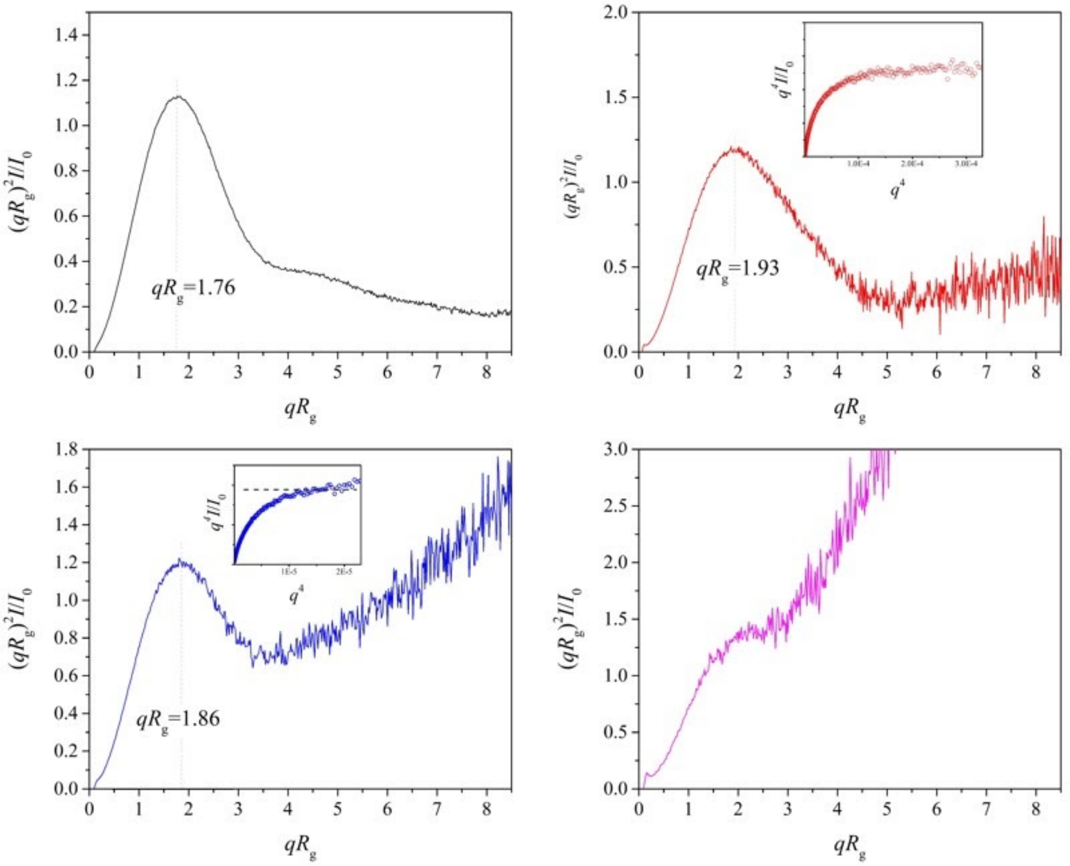


Figure 1.1 Examples of Porod (inset) and Kratky plots. (Top left) The dimensionless Kratky plot of a compacted globular protein. The main feature is a Gaussian-like peak where the peak position is around $\sqrt{3}$ (qR_g = 1.76). (Top right) The dimensionless Kratky plot of a loosely compacted globular protein. The Gaussian-like peak largely remains but an increasing feature at high-q appears. The inset figure shows a Porod plot for the same sample. The positive deviation from the Porod plateau indicates the protein is not tightly compacted. (Bottom left) The dimensionless Kratky plot of a two-domain protein with flexible linker. A clearly increasing feature is presentent at high-q. The peak position of the first feature is 1.86 $>\sqrt{3}$, suggesting the protein has relatively high flexibility. A significant positive deviation from the Porod plateau (dashed line) can be observed in the Porod plot (inset figure). (Bottom right) The dimensionless Kratky plot of an unfolded protein. The Gaussian feature is barely observable whereas there is a significant increase at high-q.

Calculating Theoretical SAXS Profiles

Numerical methods, approximating theoretical scattering, have become one of the most effective ways to analyze SAXS data (Grudinin et al. 2017; Putnam et al. 2007). A number of computational tools have been developed dedicated to calculating the theoretical solution SAXS profiles. The most prominent method is the multipole expansions of scattering intensity (Svergun et al. 1995). The general idea of this method is to express the SAXS profile using a sum of a series of spherical harmonics (multipole expansion process). An efficient implementation of this method is found in the popular program CRYSOL (Petoukhov et al. 2012). Similarly, the package SASTBX adapts the multipole expansion implementation but uses Zernike polynomials for the real-space representation of the electron density (Schneidman-Duhovny et al. 2013). Another method is the direct approximation using the Debye formula with modified scattering form factors, which is implemented in the program FoXS (Schneidman-Duhovny et al. 2013). Other notable efforts include SASSIM (Merzel & Smith 2002), AquaSAXS (Poitevin et al. 2011) and pepsi-SAXS (Grudinin et al. 2017). If the experimental data is available, a hypothesis test can be performed. The most common statistical estimator is the reduced chi-squared (χ^2) (Andrae et al. 2010; Svergun et al. 1995)

$$\chi^2 = \frac{1}{N-1}\sum_i [\frac{I_{\text{exp}}(q_i) - I_{\text{calc}}(q_i)}{\delta(q_i)}]^2 \tag{1.10}$$

where $I_{\text{exp}}(q_i)$ and $I_{\text{calc}}(q_i)$ are the experimental scattering intensity and calculated scattering intensity, respectively, at the resolution bin *i*; $\delta(q_i)$ is the corresponding experimental error.

1.2.5 Modelling

To determine a high-quality model is arguably the ultimate goal for most biologists who use structural biology as a tool. There are two main types of SAXS-oriented modelling methods: *ab initio* modelling and modelling with prior structural information.

Ab initio modelling

Reconstruction of a three-dimensional low resolution envelope from SAXS data has been pursued since 1960. The early attempts tried to map the experimental scattering patterns to the patterns from simple geometrical bodies (Feigin & Svergun 1987). P. Chacon's pioneer work with a genetic algorithm carved out a way of modern *ab initio* modelling (Chacón et al. 1998). With improved computational ability, finite element based bead-modelling (DAMMIN, DAMMIF) (Franke & Svergun 2009; Svergun 1999) and dummy-residue modelling (GASBOR) (Svergun et al. 2001) significantly improved the resolution of models derived from scattering data and made SAXS modelling more reliable for low resolution structural characterization of proteins. Recently, an *ab initio* electron density determination method (DENSS) has been suggested (Grant 2018). This method enables the reconstruction of non-uniform density and hence has the ability to model the structure within which domains with different scattering length density exist, for example, detergent-solubilized membrane proteins and protein complexes with a core-shell structure. In principle, *ab initio* modelling can only be performed on data from monodisperse samples. The modelling based on SAXS data from more complex mixtures requires the inclusion of prior information.

Modelling with prior structural information

The ability to rapidly calculate the theoretical scattering profile of a hydrated particle plays a central role in modeling molecular complexes. Because of this one can use an all-atom structure, derived for example from X-ray crystallography, as prior information. Solution and crystallization conformations of a protein are usually different, as the MX structure represents a biased sampling of protein's solution behaviors. With the MX structure as a starting point, structural rearrangements of the high resolution static structure can be applied for the interpretation of experimental SAXS data. This leads to the question of how to explore the conformational space of a solubilized macromolecule. Depending on the scale of conformational changes, there exist different strategies.

On the small-scale, conformational changes are induced by internal motion. The most commonly used strategy is through molecular dynamics (MD) simulation (Karplus & McCammon 2002). MD simulation can capture the local fluctuations of proteins rather precisely. Each frame of a MD trajectory can therefore be used to calculate the theoretical SAXS profile and this can be fit to the experimental data. Another strategy is normal mode analysis (NMA). NMA is based on an elastic network model and uses a well-established coarse-grained method (Tirion 1996) to study protein conformational transitions (Gorba et al. 2008; Mahajan & Sanejouand 2015; Zheng & Tekpinar 2011). In a recent implementation, SREFLEX (Panjkovich & Svergun 2016), NMA is used in combination with iterative optimization and chemical restraints to find the best-fitted refined model.

For large-scale, e.g. allosteric, motions, rigid-body modelling approaches can be applied. In rigid-body modelling, the individual domains or protomers are considered as uninterrupted rigid bodies. The conformational landscape is explored by the kinematical motion of these individual rigid bodies. SAXS-based rigid-body modelling has been used to study a variety of structural problems, including: validation of crystallographic oligomers in solution (Korasick & Tanner 2018), identification of distinctly different quaternary structures (Petoukhov & Svergun 2005; Petoukhov et al. 2012) and organisation of full complexes with known sub-unit structures (Jiménez-García et al. 2015),

Hybrid modelling

Pure rigid-body modelling is possible only when the full-length structure or reliable homology models of all subunits are available. In the situation where the MX structure is lacking electron density for certain parts, such as flexible loops and disordered subunit, the approximate configurations can be reconstructed using the experimental SAXS data as a guide. During such modelling, a combination of rigid-body modelling and *ab initio* modelling can be used. Since the precise conformations of the missing residues or subunits are not required to adequately compute the scattering intensity from the entire molecule (Svergun et al. 2013), coarse-grained pseudo-fragments represented by beads or chains of dummy-residues can be employed to link the rigid-domains. By iteratively optimizing the positions and orientations, together with the probable conformers and the pseudo-fragments, the best-fitted model is found by minimizing the reduced χ^2. The current state-of-the-art programming tools for hybrid modelling are BUNCH, SASREF, FOXSDOCK and CORAL (Petoukhov et al. 2007, 2012; Schneidman-Duhovny et al. 2016).

1.3 Testing Membrane Protein: Sodium-Hydantoin Transporter Mhp1

The bacterial sodium-hydantoin transporter, Mhp1, is extensively used in Chapter 3 and Chapter 4 as a test sample for methods development. A brief introduction to its structure and function is therefore given here.

Mhp1 is a secondary active transporter of the nucleobase-cation-symporter family and a member of the widespread 5-helix inverted repeat superfamily of transporters (LeuT family). Mhp1 couples the transport of sodium ions down their concentration gradients with the symport of 5-substituted hydantoin compounds into the cytoplasm of bacterial cells. Inside the cells, the hydantoin compounds are hydrolysed and converted into optically pure amino acids. The structure of Mph1 was previously solved in three different conformations: outward-facing, inward-facing and outward occluded (Faham et al. 2008; Krishnamurthy & Gouaux 2012; Krishnamurthy et al. 2009). The outward-facing open and the outward occluded states show very high similarity, but neither superimposes well with the inward-facing open state, raising the question of whether or not there is only one occluded state. Molecular dynamics simulations suggest the occurrence of an additional inward-facing occluded state (Polyakova 2015). The significant conformational changes derived from all known structures appear to be an overall rigid-body rotation of the hash motif (helices 3, 4, 8, 9) relative to the bundle motif (helices 1, 2, 6, 7).

Previous work suggests that the crystallisation conditions used, as well as the resulting crystal packing, may be the limiting factors determining which parts of the conformational landscape of the transport mechanism of Mhp1 can be explored crystallographically. It appears that the outward open or outward occluded crystal form is highly favoured for Mhp1 crystallization as the inward open state has only been observed in a single crystal. This could be a result of the conformational equilibrium of Mhp1 in solution, where the inward-open form may be a short-lived intermediate that does not exist for long enough to form protein crystal nuclei. Alternatively, the crystal lattice could conformationally select only the outward open or occluded states of Mhp1. Given these challenges and the limitations that MX has encountered with this membrane protein, SAXS was explored as a route to additional structural information.

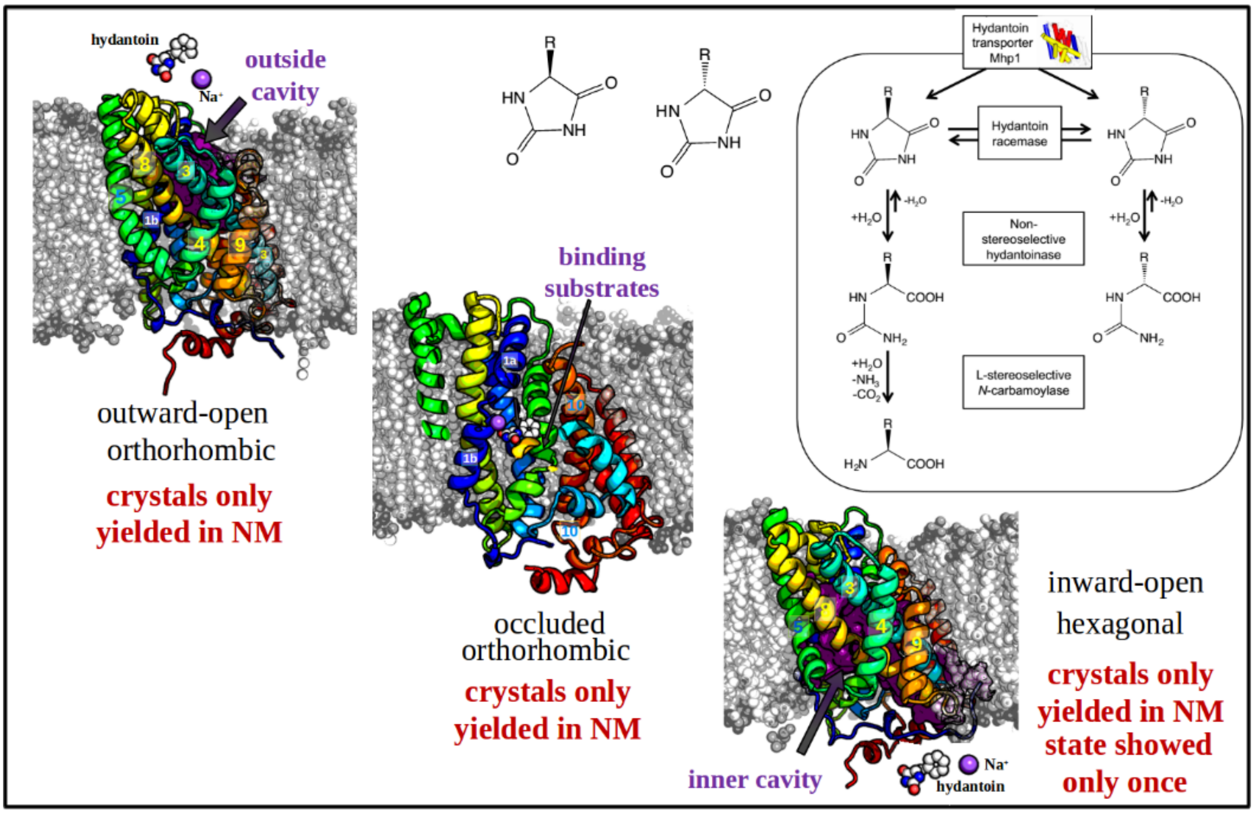

hydantoin
Na+
outside cavity
8
3
5
1b
4
9
3
outward-open orthorhombic
crystals only yielded in NM
R
HN
O
NH
O
R
HN
O
NH
O
binding substrates
1a
1b
10
10
occluded orthorhombic
crystals only yielded in NM
Hydantoin transporter Mhp1
Hydantoin racemase
+H2O
-H2O
Non-stereoselective hydantoinase
COOH
NH2
+H2O
-NH3
-CO2
L-stereoselective N-carbamoylase
H2N
COOH
inward-open hexagonal
crystals only yielded in NM state showed only once
3
5
4
9
inner cavity
Na+
hydantoin

Figure 1.2 The proposed reaction cycle of Mhp. In accordance with the alternating access model, the protein alternates between an outward-open state, where it accepts sodium and hydantoin compounds from the extracellular side of the membrane and an inward-open state, which is adopted after substrates have been released into the inside of the cell. The transition between these two states is thought to go via an occluded state, where the bound substrates are shielded from both sides of the membrane but with either the outward or inward gate remaining partially open. Upon ligand binding, during the transition between the outward-open to occluded states, helice 10 shifts to close the ligand binding site as a lid (arrow). Upon the subsequent transition to the inward-open state, the hash-motif undergoes a rotation with respect to the bundle-motif. Mhp1 transports hydantoin compounds into the cytoplasm of bacterial cells, where they are broken down into *L*-amino acids by other members of the hydantoin utilization gene cluster. Other than these three previously observed conformations, no additional conformational state has been structurally characterised for the Mhp1 protein.

1.4 Ambiguity and Subjectivity: Remaining Challenges

The solution SAXS pattern of a molecule is inherently ambiguous (Zhao & Shukla 2018) due to its isotropic nature. Generally speaking, recovering three-dimensional parameters from one-dimensional experimental data is difficult as the variety of solutions is practically infinite. Although several general parameters can be obtained through rigorous analytical interpretation of SAXS data, the correctness of the interpretation is, however, largely affected by the basic data characteristics, such as perfectness of the background subtraction, evaluation of the structure factor (i.e polydispersity and intermolecular inferences), assessment of the meaningful data-range, and estimation of the appearance of radiation damage. Due to the fact that the SAXS profile is lacking correlated measured outputs, it is hard to statistically give an objective assessment on these crucial data qualities. A common situation for the SAXS-user is that data interpretation can only be performed in an ill-posed way, which means the solution is not unique or the solution procedure is unstable, and the assessment of data quality is in fact a measurement of solution stability (Petoukhov et al. 2007). Unfortunately, the criteria of stability itself also includes subjective judgement, since there is no objective meaning for *many solutions*, and there is no sharp dividing line between “many" and “too many”. The ambiguity becomes even larger when one tries to reconstruct $p(r)$ or $\Delta\rho(\mathrm{r})$ from $I(q)$, given that it is an inverse problem of reconstructing the shape of a body from its spherically-averaged projection. As for the reduced χ^2 test, it is possible to reject a hypothesis in such a way that the violation against the experimental observation is not tolerable (i.e $\chi^2 >> 1$). However, the subjective sense of “no obvious violation” (i.e $0.7 < \chi^2 < 1.3$) cannot be converted into an objective criteria that indicates the hypothesis should be accepted as the physically meaningful conclusion.

As a SAXS-user myself, I often have to strike a balance between the goodness-of-fit and an uncertain parameter when encountering questions that need subjective judgement. Browsing the popular SAXS forum, SAXIER, the majority of questions being discussed are related to the problems brought by ambiguity and/or subjectivity.

It has been shown that the solution of an ill-posed problem can be stabilized by imposing *a priori* information (restraints) as well as by correlating with an independent observation of a closely related effect (constraint) (Frick 1995; Grant 2018; Kabanikhin 2008; Svergun 1992). The objective of this thesis is to explore the possibility of using both these approaches to alleviate the ambiguity and/or subjectivity currently included in post-processing, interpretation and modelling of experimental SAXS data.

Chapter 2
Towards Better Data Quality

Data are one of the key parts of every scientific experiment. The quality of data that are collected, and of their processing and interpretation ultimately determines the conclusions we can draw from an experiment.

There are many definitions of data quality (Bohm & Zech 2010). From a statistical point of view, the quality of data is dominated by two factors: precision and accuracy. Precision refers to how consistent results are when measurements are repeated and accuracy to how close a measurement is to the true value. The ongoing efforts in experiment optimisation and data processing for both synchrotron and XFEL experiments have been of huge benefit in recording better raw diffraction/scattering data. Besides the improvements at the data collection stage, the precision of data is also emphasized. In MX, all the commonly used statistical indicators report on data precision (Karplus & Diederichs 2015). In contrast to MX, SAXS data require only limited processing post data acquisition, i.e. subtraction of the buffer signal and potentially some correction for radiation damage. This latter often takes the form of simply discarding data frames where damage is observed.

In this chapter the challenge of determining the accuracy of SAXS data is addressed.

2.1 Introduction

A little appreciated aspect of SAXS data analysis is that if substantial systematic errors are present, the currently applied statistical indicators can fail to accurately reflect the data quality. This problem arises partly because of the limited availability of statistical tools for assessing SAXS data (Putnam et al. 2007; Toby 2006). In addition, a set of data is sometimes regarded as being of low quality if it is poorly fit by a defined model. However, model fitting of scattering data is an inverse problem. Any measure of the accuracy of SAXS data should therefore include consideration of whether the data correctly represent the physical phenomena. This measure should be determined before, instead of after, any interpretation of the data is made. In practice, it is currently nontrivial to perform SAXS data processing, where substantial corrections of the data are needed, guided by an accurate systematic error estimation. This means that with today's high throughput beamlines at brilliant modern light sources we risk collecting more data but having that data contaminated with poorly determined propagated errors (Franke et al. 2015).

The situation is even worse for SAXS of biological macromolecules in solution, due to the lack of Bragg amplification as well as the use of aqueous sample conditions. Both compromise the signal-to-noise ratio (SNR) and bring significant difficulties in data post-processing. Many optimisations of experimental configuration have been pursued to reduce data ambiguity and these have yielded impressive results (Acerbo et al. 2015; Bizien et al. 2016; Pernot et al. 2013; Ryan et al. 2018). New statistical methods have also been developed to improve solution SAXS data quality. Here a short introduction is given to two innovative model-free statistical methods that utilize the oversampling possible in a modern SAXS experiment.

The first applies information theory to SAXS (Rambo & Tainer 2013). According to the Nyquist-Shannon sampling theorem, the minimum sampling rate required to completely reconstruct the SAXS profile is determined by the largest distance between any two points inside the particle ($D_{\max}$). This relationship determines the minimum number of equally spaced sampling points, *ns* (Shannon points).

$$ns = q_{\max}\, D_{\max} / \pi \tag{2.1}$$

In the modern SAXS experiment, *ns* is much smaller than the typical angular sampling rate. In principle, the oversampled points should provide zero additional information. However, due to the uncertainties in both q and $I(q)$, the oversampled dataset can be used as a constraint on the Shannon points determination. The inferred SAXS profile from a set of Shannon points can be derived from the Shannon-Whittaker interpolation formula,

$$I(q) \ = \sum_{i=1}^{n} (\frac{ns \cdot \pi}{D_{max}} (\frac{sinc(qD_{max} - ns \cdot \pi)}{qD_{max} - ns \cdot \pi}) \tag{2.2}$$

where $n_{max} = q_{max} D_{max} \pi^{-1}$. The interpolation keeps $I(q)$ at all Shannon points but expressess $I(q)$ at any other q as a linear combination of all the Shannon points. Although the *SNR* affects the interpolation, the Shannon-Hartley theorem suggests that, as long as the sampling rate Δq is smaller than the maximum rate of information, C

$$C = \frac{2\pi}{D_{max} log2(1+SNR)} \tag{2.3}$$

, the interpolation is able to reconstruct the scattering profile. For the modern beamline, the sampling frequency and *SNR* that is routinely achieved guarantees the success of the reconstruction.

The second approach takes a step forward from the traditional cross-correlation approach (Franke et al. 2015). In synchrotron SAXS, scattering data are usually recorded in multiple short frames. Given the matrix of the Pearson correlation factor r_{kl} of any two frames, where

$$r_{kl} = \frac{\mathrm{cov}(I(q_k), I(q_l))}{\delta(I(q_k))/\delta(I(q_l))} \tag{2.4}$$

a “heatmap” of r_{kl} plotted according to the sign of its numeric values (−1 if r_{kl} is negative or +1 if positive) displays a random pattern without any obvious significance if no systematic differences exist between the two frames. Nonrandom and contiguous areas called patches occur in the presence of differences.

The quantification of whether or not there are systematic differences is achieved by evaluation of the largest observed patch. The maximum edge length of the patches (an integer number), L, follows a Binomial distribution with $p = 0.5$. The probability of obtaining a patch with an edge longer than L, $P(R_n > L)$, is given by

$$P(R_n > L) = 1 - \frac{A_n(L)}{2^n} \tag{2.5}$$

$A_n(L)$ is defined by Schilling's equation

$$A_n(L) = \begin{cases} \displaystyle\sum_{j=0}^{L} A_{1-n-j}(L) & \text{for } n > L \\ 2^n & \text{for } n \leq L \end{cases} \tag{2.6}$$

A predetermined significance level alpha, often 0.01 or less, is used to determine statistical significance. $A_n(L)$ that leads to a *$P(R_n > L)$* smaller than 0.01 means the differences between the two distributions or datasets are significant.

This idea can be extended to multiple repeated tests, as in a frame series from a SAXS experiment. The *P* values are now adjusted by

$$P_{adj} = mp \tag{2.7}$$

where $p = P(R_n > L)$ and *m* is the number of tests. The P_{adj} is then compared to the predefined significance level alpha. Any frame with a P_{adj} smaller than alpha should be rejected in order to stabilize the final scattering profile.

Despite the success of these methods, they are mostly useful for "rejecting" significantly different data points or data frames, based on the assumption that the rest of the data are close to the "truth". If we do not have objective evidence showing this assumption indeed holds, we end up improving the data precision instead of the accuracy. This can lead to a situation where mediocre data are recorded that are not correctable or simply cannot be assessed before any interpretation is done. In my opinion, due to the current paucity of tools for assessment of true data quality, a huge confusion exists in the post-processing of SAXS data, especially when considering how best to perform both background corrections and data averaging.

In this chapter I propose an objective metric, based on the dynamic time warping algorithm, that can be used to both verify the data quality and identify the optimal data correction procedure for post-processing. This approach takes advantage of the independent information that can be determined for the sample under study in, for example, a SEC-SAXS experiment.

2.2 Correctness-State Score: an Objective Metric

2.2.1 Background

One approach often used to perform SAXS data quality assessment is to estimate how well the data fulfills the intended purpose. For example, a scattering profile with inconsistent linearity in the Guinier region is regarded as low quality because of the subsequent difficulty in determining the parameters needed for the Guinier approximation (Petoukhov et al. 2007). Another example is that implemented in the popular program GNOM (Svergun 1992). Here, when an indirect Fourier transform is performed to obtain the real-space pair distribution, the data are considered as good if the resulting pair distribution satisfies certain criteria, such as oscillation, discrepancy, stability, positivity, etc. This latter approach requires that one strikes a balance among many factors and is entirely subjective. Moreover, GNOM uses post-processed data as input, and the post-processing itself may be flawed. Establishing an objective approach to assess the quality of both SAXS data and its processing must therefore start with the background correction.

In modern synchrotron bio-SAXS experiments in-line chromatography has been introduced to separate the often complex mixtures that occur in these samples, for example, the mixture of protein-detergent complexes, oligomers and empty micelles/vesicles that occur even in well behaved membrane protein samples. By separating any potential contaminants and different components of mixtures (conformational or compositional) using a chromatographic column, chromatography-SAXS facilitates ideality and weak mono-dispersion of the biological particles under study (Brennich et al. 2016; Chaudhuri et al. 2017; Kikhney & Svergun 2015). It is important to realise that the chromatographic signal (*CHROMS*) and the X-ray scattering signal (*XSS*) are essentially two time series. The *CHROMS* is normally a measurement of a certain molecular property, such as refractive index, light scattering or absorption at 280 nm. The *XSS* is a measurement of the total X-ray scattering of the sample and its surrounding environment (buffer, sample capillary). *CHROMS* and *XSS* recorded in the same experiment both respond to the sample concentration and are hence correlated. As long as the weak monodispersity and non-interaction assumptions hold, the resulting background corrected time series of the X-ray scattering intensity at each resolution bin of the scattering vector q, $I_q(t)$, is semantically similar to the *CHROMS*. Therefore, an objectively good background correction should in principle lead to a good similarity between the *CHROMS* and $I_q(t)$. However, the configuration of the chromatography-SAXS experiment creates a number of issues that reduce the agreement between the two signals. Problems such as flow gradient, sample dilution, X-ray induced changes and pump cycles can reduce the stability of the data. More often than not, as extra generic in-line detection modules (i.e UV-detector, multi-angle light scattering) are added, the situation worsens due to peak broadening and signal asynchronization. Therefore, traditional correlation statistics fail to predict the similarity between the two signals as the *CHROMS* and *XSS* are not necessarily aligned in either time or signal amplitude.

It is clear that in order to exploit the possibility of using the correlation between those two independent signals as an indicator for assessment of the data quality, the similarity metric used must be sensitive yet robust to nonlinear relationships. To address this challenge a Dynamic Time Warping (DTW)-based scoring method has been developed. The DTW algorithm is one of the most robust methods to compute the optimal alignment between two time series by taking into account the nonlinearities, such as local warping, phase shifts and scaling distortions (Keogh & Pazzani 2001; Keogh & Ratanamahatana 2005; Sakoe & Chiba 1978; Salvador & Chan 2007). It has been successively applied in the fields such as pattern mining (Hong et al. 2017; Lu et al. 2014), word recognition (Filatova et al. 2012; Myers et al. 1980) and protein sequence alignment (Loose et al. 2016; Lyons et al. 2014).

As a metric of "similarity", a Correctness-State Score (CSS) was formulated by utilising several mathematical concepts originating from the DTW. CSS can be considered as a numerical likelihood of the fit with a scale of 0 to 1. With the guidance of this objective score, CSS, it is possible to quantitatively assess the "goodness" or appropriateness of a background correction for chromatography-SAXS data.

2.2.2 the Algorithm

2.2.2.1 Derivative Warping Distance and Warping Cost

A time series is a sequence

$$\boldsymbol{Q} = [q_1, q_2, ...q_n],\ q_i \in \mathbb{R} \text{ for all } i \in [1, n] \tag{2.8}$$

Let $q_i^{'}$ be the estimated derivatives of q_i

$$q_i^{'} = \frac{1}{2}(q_i - q_{i-1} + \frac{q_{i+1} - q_{i-1}}{2}) \tag{2.9}$$

We denote ***Q***' as the derivative time series,

$$\boldsymbol{Q}^{'} = [q_1^{'}, q_2^{'}, ...q_n^{'}] \tag{2.10}$$

Given two time series,

$$\boldsymbol{Q} = [q_1, q_2, ...q_n], \ \ \boldsymbol{C} = [c_1, c_2, ...c_n] \tag{2.11}$$

the set of all possible warping paths is denoted by $\mathbb{P}$.

Suppose a sequence ***P***

$$\boldsymbol{P} = [p_1, p_2, ...p_l] \tag{2.12}$$

where p is a indices-pair of ***Q'*** and ***C'*** with

$$p_k = (i_k, j_k) \quad \text{for all } k \in [1, l] \tag{2.13}$$

The mapping between indices follows following rules:

1. The first index from ***Q'*** must be matched with the first index from ***C'***;
2. The last index from ***Q'*** must be matched with the last index from ***C'***;
3. The mapping of the indices from ***Q'*** to indices from ***C'*** must be monotonically increasing, and vice versa.

The derivative warping path $\boldsymbol{P}_{\text{warp}}$ is determined by

$$\arg\min_{\boldsymbol{P}} \sum_{(i,j)\in\boldsymbol{P}} (q_i^{'} - c_j^{'})^2 \ : \ \boldsymbol{P} \in \mathbb{P} \tag{2.14}$$

Define,

$$cost(\boldsymbol{Q}^{'}, \boldsymbol{C}^{'}) := \sum_{(i,j)\in\boldsymbol{P}_{\text{warp}}} (q_i^{'} - c_j^{'})^2 \tag{2.15}$$

The derivative warping distance, D, of ***Q*** and ***C*** is of the form

$$D = \sqrt{cost(\boldsymbol{Q}^{'}, \boldsymbol{C}^{'})} \tag{2.16}$$

The warping cost, W, is defined by,

$$W = \sum_{(i,j)\in \boldsymbol{P}_{\text{warp}}} ||\boldsymbol{q}_i - \boldsymbol{c}_j|| \tag{2.17}$$

where ||·|| denotes the Euclidean norm.

Fig. 2.1 and Fig. 2.2 show a figurative explanation of the above terms.

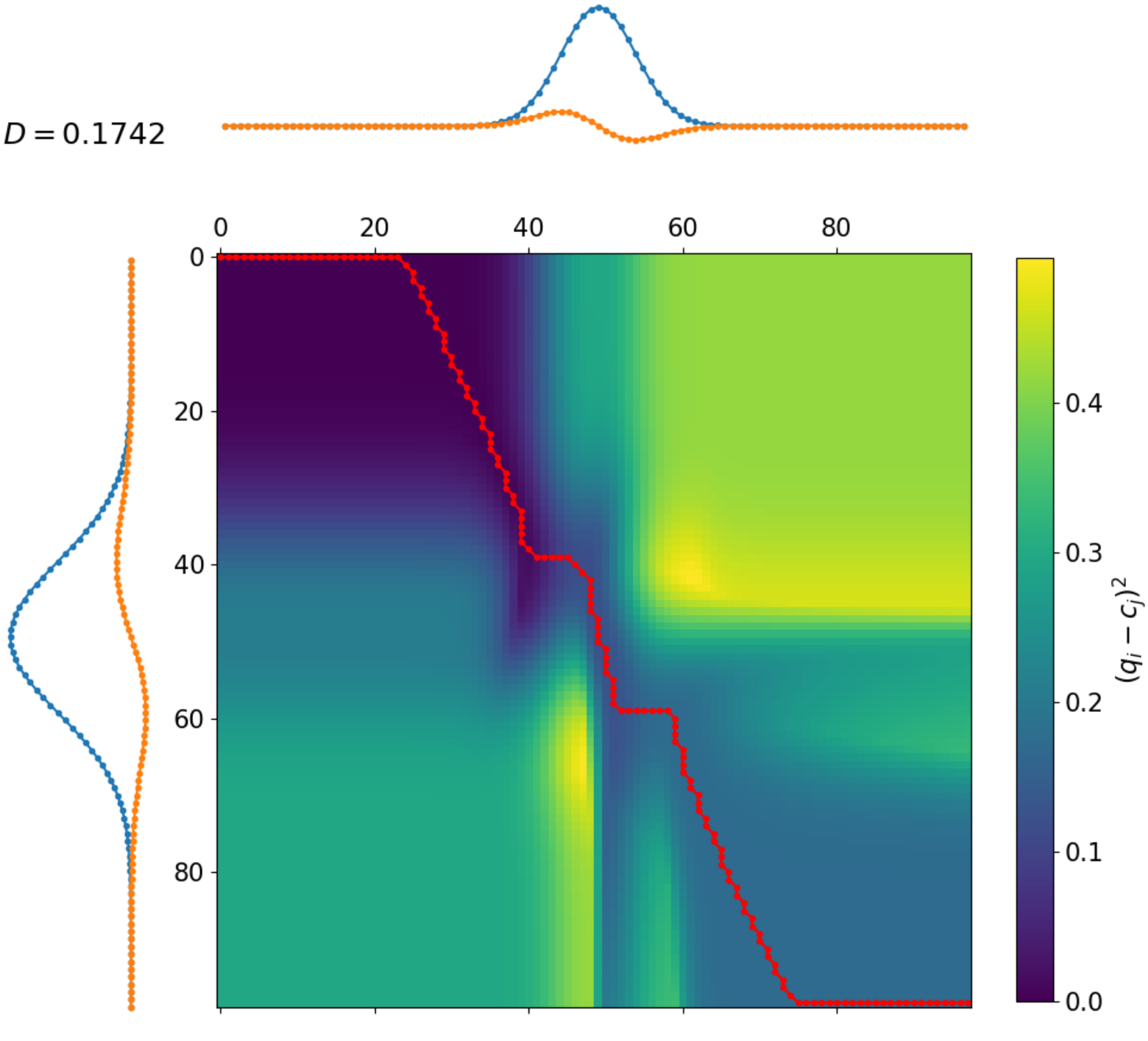


Figure 2.1 Figurative explanation of DTW (part 1). y-axis: normal distribution, $\mathcal{N}(50,\ 10^2)$; x-axis: normal distribution, $\mathcal{N}(50,\ 5^2)$. Blue dotted line: time series in the form of a normal distribution; orange dotted line: the derivative of corresponding time series. Red dotted line: the derivative warping path of the two time series which takes the lowest possible squared

Euclidean norm (blue-yellow colour ramp) between every pair of two derivative time points in a non-backtracking way. The derivative warping distance is 0.1742.

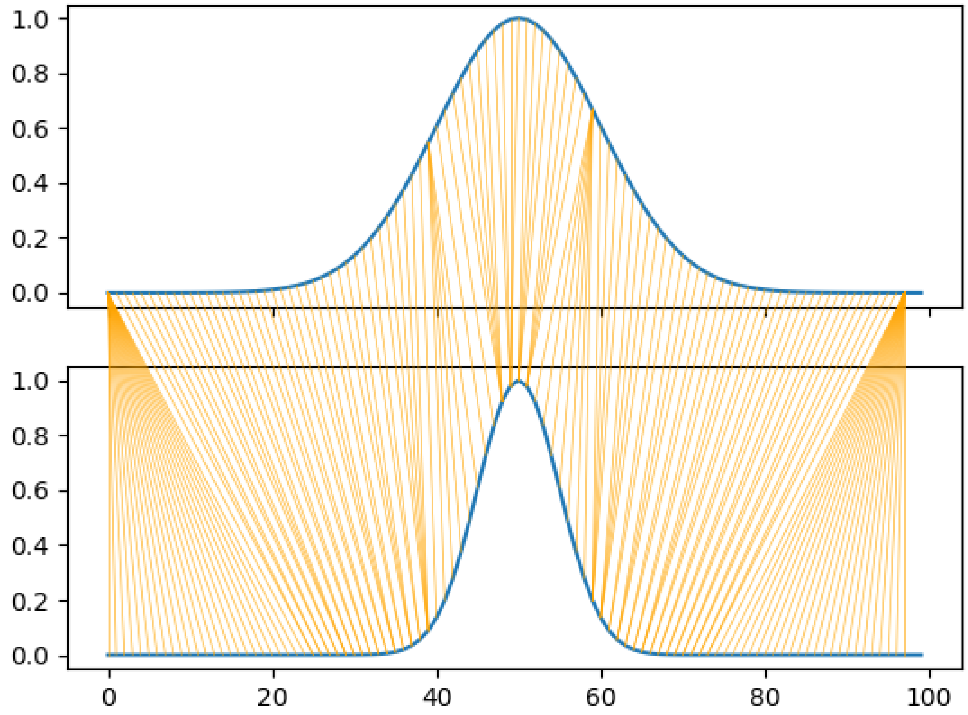


Figure 2.2 Figurative explanation of DTW (part 2). Top panel: normal distribution, $\mathcal{N}$(50, 10^2); bottom panel: normal distribution, $\mathcal{N}$(50, 5^2). Distributions shown as blue lines. Orange solid lines show the warping path between the two time series plotted in a way that illustrates the optimal point-to-point alignment.

2.2.2.2 Notation and Definitions for *CHROMS-XSS*

Now we extend DTW to *CHROMS-XSS* data. The *CHROMS* is a series of observations with one variable.

We can write vector ***u***,

$$\boldsymbol{u} = [u_1, u_2, ...u_l] \in R^l,\ u_i \in R\,(i \in [1, l]) \tag{2.18}$$

where u_i is the detector observation at time point indexed by i and l is the length of the vector *CHROMS*.

The small-angle X-ray scattering signal (*XSS*) is a series of observations with a number of variables equaling the detector resolution.

We can write matrix $\boldsymbol{X}_{2\mathrm{D}}$,

$$\boldsymbol{X}_{2\mathrm{D}} = [\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_k] \in R^{k\times(h\times w)} \tag{2.19}$$

where $\boldsymbol{X}_j \in R^{h\times w}$ $(j \in [1, k]])$ is a photon detector reading at time point indexed by j, $h\times w$ is the number of pixels, k is the total number of collected frames.

After proper radial integration and post-process, we denote matrix $\boldsymbol{X}$,

$$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k] \in R^{k\times m} \tag{2.20}$$

where $\boldsymbol{x}_j \in R^m$ $(j \in [1, k]])$ is the scattering intensity of the scattering vector and m is the sampling size along the scattering vector space (number of resolution bins).

If $l \neq k$, the *CHROMS* is interpolated so that $l = k$.

We call the indices of the boundary elements of the subgroup

$$[\boldsymbol{X}]_{TFR} \rightarrow \boldsymbol{X} \tag{2.21}$$

as total frame/time range *TFR* (*TFR* ∈ $[1, k]^2$).

The same is applied to U so that

$$[\boldsymbol{u}]_{TFR} \rightarrow \boldsymbol{u} \tag{2.22}$$

Let D_n then be the derivative warping distance of $[\bar{\boldsymbol{u}}]_{TFR}$ and $[\bar{\boldsymbol{x}}_n]_{TFR}$ at the resolution bin of n (n ∈ $[1,m]$). Using (2.16) we have

$$D_n = \sqrt{cost([\bar{\boldsymbol{u}}']_{TFR}, [\bar{\boldsymbol{x}}_n']_{TFR})} \tag{2.25}$$

where $^{-}$ denotes the global max-min scaling of the signal amplitude.

Let $\boldsymbol{P}_{\mathrm{warp}}$ be the derivative warping path of $[\bar{\boldsymbol{u}}']_{TFR}$ and $[\hat{\boldsymbol{x}}_n']_{TFR}$, where ^ denotes dimension -wise max-min scaling.

We define peak signal range *PSR* (*PSR* ∈ $[1, l]^2$) as a boundary condition. Accordingly the subgroup $[\boldsymbol{P}_{\mathrm{warp}}]_{PSR}$ is generated.

$$[\boldsymbol{P}_{\mathrm{warp}}]_{PSR} \rightarrow \boldsymbol{P}_{\mathrm{warp}} \tag{2.24}$$

We can then obtain W_n, the warping cost of $[\bar{\boldsymbol{u}}]_{TFR}$ and $[\hat{\boldsymbol{x}}_n]_{TFR}$, within *PSR*.

$$W_n = \sum_{(i,j)\in[\boldsymbol{P}_{\text{warp}}]_{PSR}} ||[\bar{u}_i]_{TFR} - [\hat{\boldsymbol{x}}_n(j)]_{TFR}|| \tag{2.25}$$

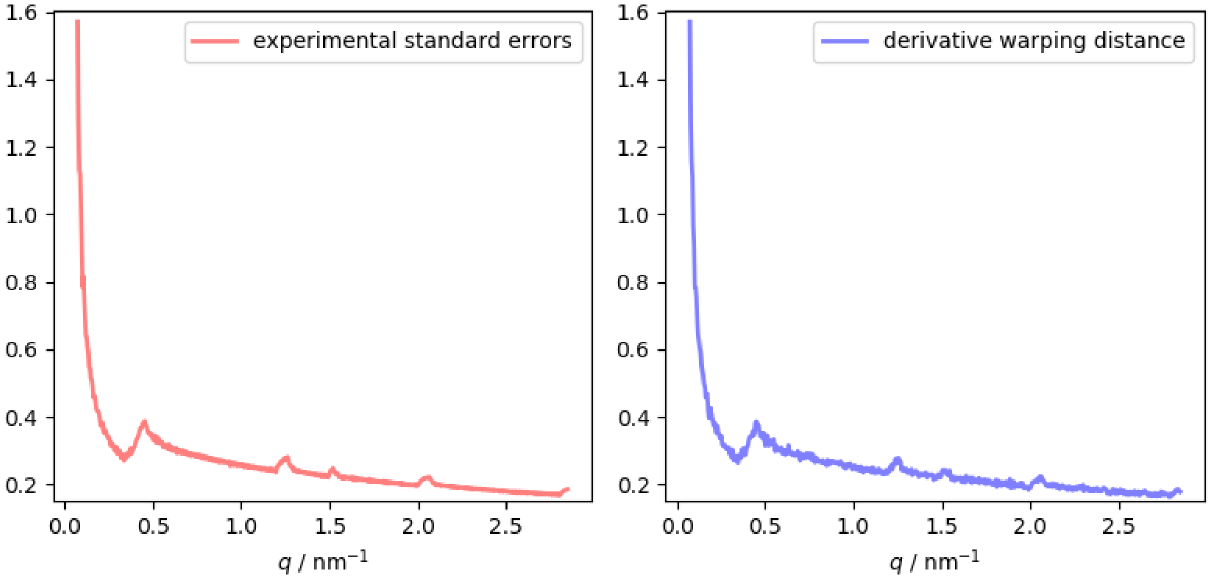


Figure 2.3 Physical implication of derivative warping distance D_n. Left: the experimental standard error estimates of 3,000 frames of buffer background collected at BioSAXS BM29, ESRF (Grenoble, France) using a Pilatus 1M detector. The spikes in the various resolution bins are caused by the gaps in the detector modules. Right: the derivative warping distance maps the experimental errors.

Interestingly, D_n is a measure of noise level. For a group of repeated measurements in the form of time series $\boldsymbol{Q}$, $q_i^{'}$ is directly proportional to the experimental variance. Assuming the variance of the reference time series $\boldsymbol{C}$ is small (which is usually the case for *CHROMS* signals), the calculation of derivative warping distance, *D*, has the same form to the uncertainty of a measured value. Physically, the experimental standard error of a group of repeated measurements gives the variance of the mean value. Statistically, these two values should be on the same scale. Note that the value of D_n follows the same pattern of the estimated experimental errors (Fig. 2.3).

The warping cost is a measure of the similarity between $[\bar{\boldsymbol{u}}]_{TFR}$ and $[\hat{\boldsymbol{x}}_n]_{TFR}$. A pair of identical time series will give a warping cost of 0.

2.2.2.3 Formulation of the Correction-State Score

We can construct the correction-state score (CSS), as a function of q, as a likelihood estimation. However, in practice, the data are binned, so that it is more useful to describe the likelihood estimation per resolution bin, n:

$$\xi_n = e^{-D_n\left(\frac{Wn}{An} + \frac{||An - Bn||}{Bn}\right)} \tag{2.26}$$

where D_n is the derivative warping distance of $[\bar{\boldsymbol{u}}]_{TFR}$ and $[\bar{\boldsymbol{x}}_n]_{TFR}$, W_n is the warping cost of $[\bar{\boldsymbol{u}}]_{TFR}$ and $[\hat{\boldsymbol{x}}_n]_{TFR}$.

$$A_n = \sum_j |[\hat{\boldsymbol{x}}_n(j)]_{TFR}| \ , \ B_n = \sum_i |[\bar{u}_i]_{TFR}| \ , \text{ given } (i,j) \in [\boldsymbol{P}_{\text{warp}}]_{PSR} \tag{2.27}$$

ξ_n ranges from 0 to 1 (For simplicity, ξ_n at a certain q is denoted by *CSS*). The formulation of ξ_n has a clear physical implication. ξ_n is accounted for two parts. The first part comprises the likelihood of semantic likeness, where W_n is a measure of similarity and the "cost" of a certain specific correction state. The smaller the W_n is the better the correction will be, given a certain support signal. The second part is the penalty of baseline drifts, which downscales the score once the signals have inconsistent background. Denominator A_n and B_n refer to the scaled feature of the two signals. The use of D_n is to compensate for the fact that *XSS* signals may have different noise levels and those noise should be weighted off. It is worth mentioning that severe oversubstraction results in that A_n becomes larger than the length of x_n, in which case ξ_n equals 0.

Finally, we have,

$$\boldsymbol{CSS}(q) = [\xi_1, \xi_2, \xi_3, \xi_4, \ldots, \xi_m] \tag{2.28}$$

2.2.2.4 Current Optimal Protocol for Calculating *CSS*(*q*)

The current protocol for calculating ***CSS***(q) for SEC-SAXS data where a *CHROMS* signal is available is as follows:

1. Radial integration of scattering patterns (normally automated on-line in the beamline software)
2. Construct an *N*-by-*M* *XSS* matrix using all 1D scattering profiles. *N* is the total number of frames. *M* is the total number of resolution bins
3. Background subtraction to yield a "corrected" *XSS*(*t*, *q*) (denominated as *I*(*t*, *q*))
4. Construct a 1-by-*K* *CHROMS* matrix, where *K* is the total number of recorded data points
5. Convert the sampling rate of *XSS* and *CHROMS* into the same units (e.g second)
6. Interpolate the *CHROMS* into a 1-by-*M* array
7. Choose the TFR
8. Choose the PSR
9. Calculate a denoised *XSS* using SVD
10. Calculate $[\bar{\boldsymbol{u}}]_{TFR}$, $[\bar{\boldsymbol{x}}_n]_{TFR}$, and $[\hat{\boldsymbol{x}}_n]_{TFR}$.
11. Calculate CSS metadata including, W_n, D_n, A_n, B_n
12. Calculate ***CSS***(q) for each resolution bin of the *I*(*t*, *q*)

2.2.3 Results and Discussion

2.2.3.1 CSS on Synthetic Data

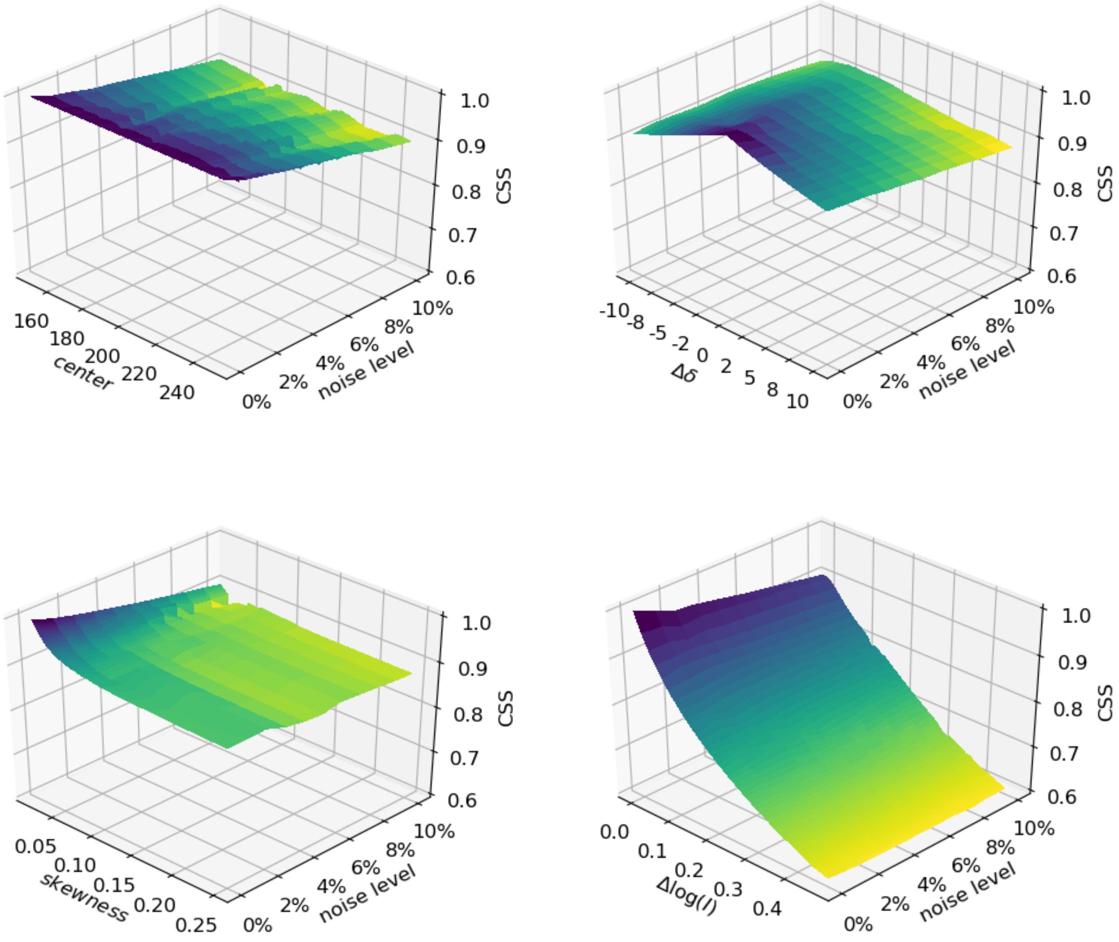


Figure 2.4 (Top left) The co-effect of peak center and noise level. (Top right) The co-effect of peak width and noise level. $\Delta\delta$ is the difference of the variance to $\mathcal{N}(200, 40)$. (Bottom left) The co-effect of peak skewness and noise level. (Bottom right) The co-effect of the accumulative background and noise level. $\Delta\log(I)$ is the logarithmic difference of the value of $\mathcal{N}(200, 40)$ and $\mathcal{N}(200, 40) + a(x - 150)$ ($x \in [150, 400]$) at the position $x = 200$, where a ranges from 0 to 0.02.

Synthetic data were used to demonstrate the implication of the value of *CSS*. A normal distribution, $\mathcal{N}(200, 40)$ in the range of [0, 400], was used as the target data. The source data were alternated accordingly based on $\mathcal{N}(200, 40)$. *TFR* and *PSR* were both set as [0, 400]. Random Gaussian noise was added to test the robustness of the score function. Fig. 2.4 shows the effects of peak center, peak width, peak skewness and accumulative background on *CSS* respectively.

2.2.3.2 Applications of Correctness-State Score

Use of CSS to define the optimal background correction for SEC-SAXS data

Here, I demonstrate the application of CSS to SEC-SAXS data deposited for Bovine Serum Albumin (SASDF99) in the small-angle scattering biological database (SASBDB) (Valentini et al. 2015). The *XSS* of these data show clear traces of radiation damage after the monomeric BSA elution peak (Fig. 2.5). Three *CHROMS* are available, the excess Rayleigh ratio (*Rex*), the refractive index contrast (*dRI*), and absorption at 280 nm (*UV*) and can be used to assess the appropriateness of different corrections to the *XSS* signal using CSS.

Radiation damage disproportionately increases scattering at both low and high resolution due to the progressive accumulation of aggregates which persist in the X-ray exposed region of the capillary through which the sample is flowing. Three different background corrections *FANCY*, *BEFORE* and *AFTER* were applied as benchmarks to test the CSS against correlated data from Rex, dRI and UV (Fig. 2.6 top). *BEFORE* and *AFTER* are the standard ways of conducting the background correction, where the “background” signal from a “flat” region of the mean *XSS* either before or after the BSA peak appears is obtained by determining the average *I*(*q*) over a series of frames which can then be subtracted from the *I*(*q*) of the BSA peak. The *FANCY* correction uses an approach I have developed called self-adaptive background correction which assumes background changes resulting from irradiation are cumulative yet inconsistent with respect to q (this is discussed in detail in §2.3.3). *BEFORE* and *AFTER* are inevitably flawed. *BEFORE* is incapable of correcting for the radiation damaged component contribution underlying the BSA peak, whereas *AFTER* results in an overestimation of the background signal in the regions of the time-trace where damage has not yet occured. A summary of the background correction methods and parameters used to calculate the ***CSS***(*q*) is given in Table 2.1

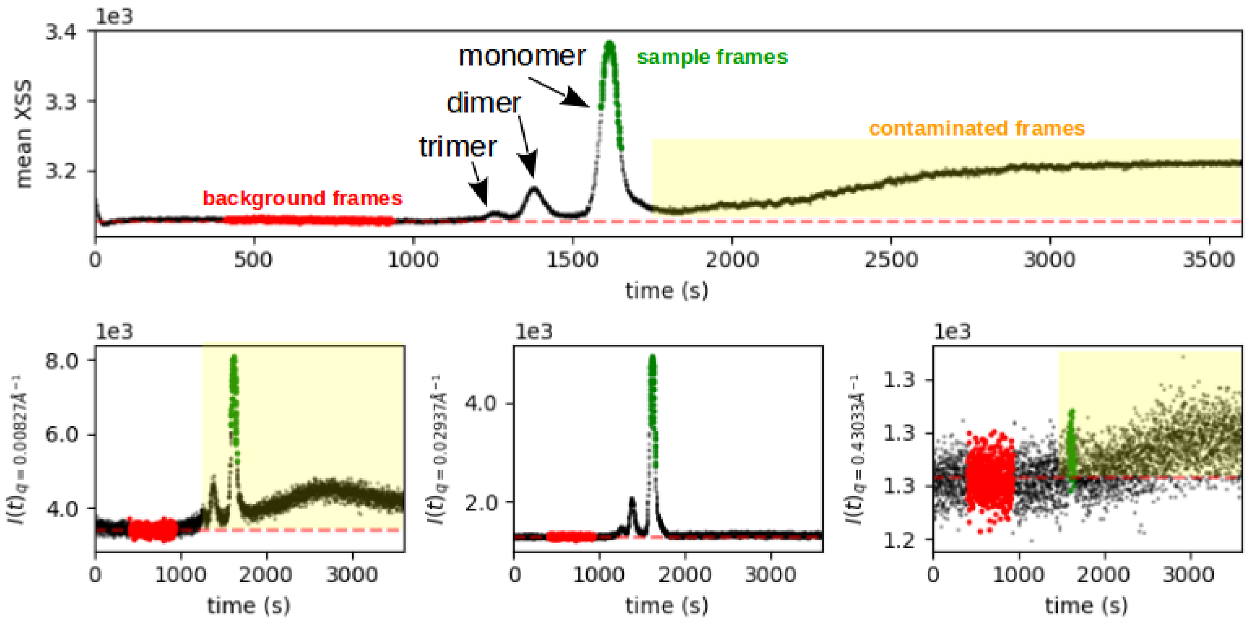


Figure 2.5 Top panel: the mean *XSS* against time/frames plot for SASDF99, showing the sequential elution of the BSA trimer, dimer and monomer species, as well as a progressively increasing contribution from radiation induced changes that manifests as an increasing *XSS* with time. Red and green points indicate the background and sample frames used by SASBDB to generate the final background corrected average scattering profile for SASDF99. Bottom panel (left to right): $I_q(t)$ at q-value of 0.00827 Å^{-1}, 0.02937 Å^{-1} and 0.43033 Å^{-1}, respectively, to illustrate that patterns of radiation damage vary with q. Red dashed lines indicate the $I(q)$ derived from the background frames used by SASBDB that are subtracted from the *XSS* to yield the background corrected signal. Yellow shading indicates the radiation-contaminated frames where simple visual examination indicates that the background correction is not appropriate. At low resolution (left), radiation damage components become evident immediately after the trimer peak. At medium resolution (middle), no apparent radiation damage can be seen. At high resolution (right), although the data are noisy, radiation damage is again evident. In addition, several of the sample points used by SASBDB are below the background $I(q)$, demonstrating how poor the default correction applied by SASBDB is in this q-range.

Table 2.1 A summary of the background correction methods and parameters used to calculate the ***CSS*** for SASDF99

post-processing method	***Frames used for background correction***	***TFR***	***PSR***
BEFORE	{500, 501, …, 900}	{500, 501, …., 3090}	{1550, 1551, …, 1680}
AFTER	{2100, 2101,..., 2500}	{500, 501, …., 3090}	{1550, 1551, …, 1680}
FANCY	self-adaptive background†	{500, 501, …., 3090}	{1550, 1551, …, 1680}

† See §2.3.3

To illustrate the impact of poor background correction on the data and the utility of the CSS metric for assessing background correction quality, we can consider three different Shannon points in the SASDF99 *XSS* data (Fig. 2.6 bottom). At the second Shannon point (q = 0.0315 Å^{-1}), where the contribution from radiation damage is negligible, all three background correction approaches give a *CSS* score close to 1, indicating a nearly perfect correction. On the other hand, at the 36th Shannon point (q = 0.79752 Å^{-1}), where the signal is highly affected by radiation damage, the BEFORE background correction leads to a large reduction in the *CSS* value, indicating it is a poor correction approach for this resolution range. Similarly, *AFTER* performs even worse, with a severe over-subtraction leading to a *CSS* value of zero. A more sophisticated background correction for these data is thus clearly needed and *FANCY* yields a well-corrected signal, reflected in a high *CSS* value. It is important to note that the requirement for a sophisticated background correction is not obvious if the data are only considered where the sample scattering is not heavily contaminated with the radiation damaged component (illustrated for the 2nd and 7th Shannon points in Fig. 2.6) and therefore it is vital to assess the background correction quality across the whole resolution range of the data and over the entire time-series. This is not routinely presented to a user in the currently widely used SAXS data analysis packages.

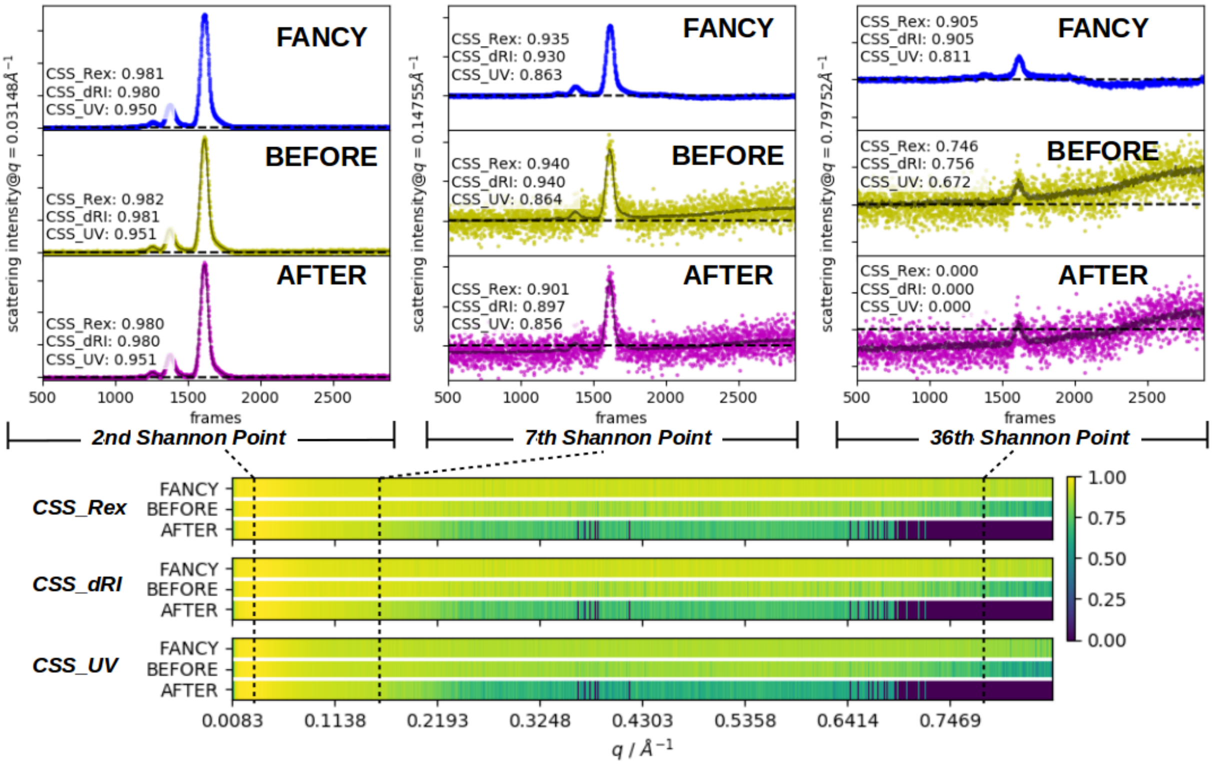


Figure 2.6 The utility of CSS for assessing choice of background correction algorithms. The bottom shows the ***CSS*** values for the three different X-ray signal background corrections, as a function of resolution, using each of the three *CHROMS* signals deposited in the SASDF99 dataset (*Rex*, *dRI* and *UV*). The top panels show the background corrected *I*(*t*, *q*) at the 2nd (left), the 7th (middle) and the 36th (right) Shannon points. *FANCY* (blue), *BEFORE* (gold) and AFTER (purple). The solid lines in BEFORE and AFTER are derived from the summation of the top 5 SVD components. *CSS_Rex*, *CSS_dRI* and *CSS_UV* are correction state scores using Rex, dRI and UV, respectively, as supports.

While CSS is calculated for each q, the assessment of SAXS data quality requires the full profile at every resolution bin. A statistical analysis of ***CSS***(*q*) is helpful to decide the overall data quality before any interpretation. Given a perfect support, the best correction is a set of scores with an expectation of one and a standard deviation of zero in the ***CSS***(*q*) plot. Therefore, the goodness of correction can be indicated by the values of the mean and the slope of ***CSS***(*q*) from its linear regression.

Use of CSS to define the optimal sample frames for averaging

Since a single frame of *XSS*(t) is very noisy, normally multiple *XSS*(t) frames containing signals from the molecule of interest are averaged. This process, however, will introduce systematic errors if the frames included in the average have contributions from non-sample components, such as radiation-damage-induced aggregation. CSS provides an objective criteria for deciding the ideal set of *XSS*(t) frames that can be averaged to yield the final background corrected sample SAXS curve.

To demonstrate this, the mean and the slope of ***CSS***(q) were calculated using a sliding Peak Signal Range (*PSR*) window of 50 frames (Fig. 2.7). First we can immediately see that there is a series of *PSR* windows where the values of ***CSS***(q) *mean* and ***CSS***(q) *slope* indicate that the data are reliable and can be averaged. Second, the ***CSS***(q) statistics for *FANCY*, *BEFORE* and *AFTER* background corrections around the peak position are consistent. This means the perceptual criteria for selecting the ideal *PSR* frames are valid even if the background correction itself is non-ideal. However, it is also clear from Figure 2.7 that choice of an optimal background correction, *FANCY* which has the largest ***CSS***(q) *mean* and the least steep ***CSS***(q) *slope*, also results in the maximum number of frames that can be averaged together, enhancing the sample signal. In conclusion, for this dataset, given the consistency and goodness of the ***CSS***(q) statistics, an average over the whole monomeric peak range (1050 to 1180) after the *FANCY* background correction should be applied in order to obtain the best signal-to-noise ratio as well information completeness.

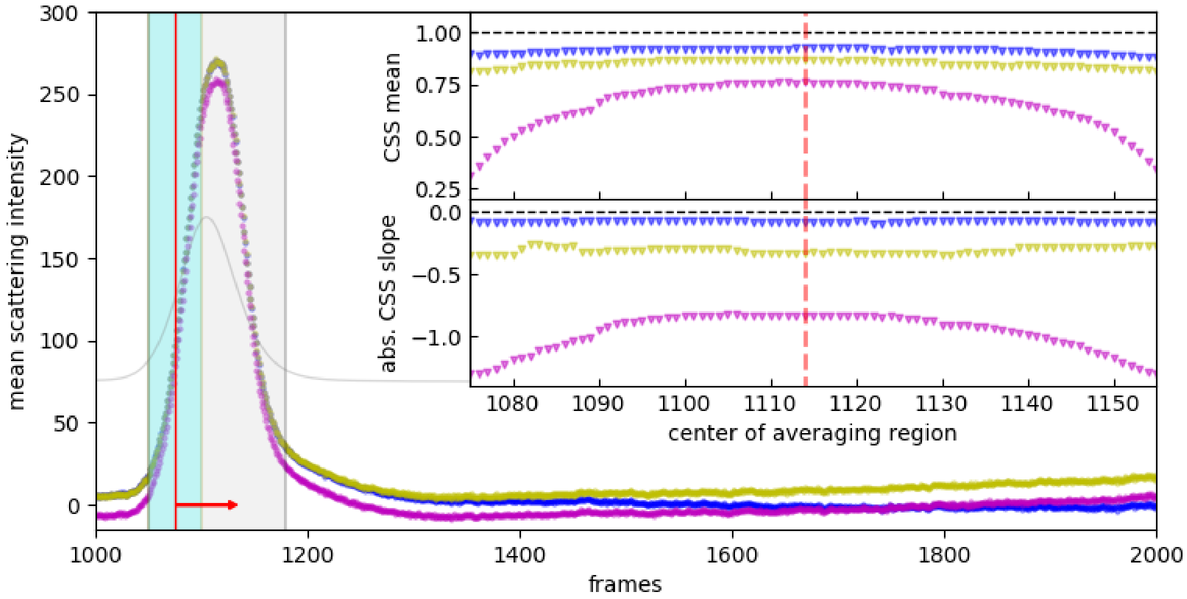


Figure 2.7 CSS used to select the ideal averaging region. The main plot shows the mean *XSS*(t) for *FANCY* (blue dots), *BEFORE* (gold dots) and *AFTER* (purple dots) background corrected data. The excess Raleigh ratio (*Rex*) trace is shown as a grey line. The first *PSR* window ranging from 1050 to 1100 is highlighted in cyan. Inset: Evolution of the ***CSS***(q)

mean and ***CSS***(*q*) *slope* as the *PSR* window is slid over the sample peak. The center of the monomeric peak in *XSS*(*t*) is marked by the dashed red line.

Use of CSS to identify the optimal region for calculation of the radius of gyration

One of the most important pieces of protein geometric information that can be derived from SAXS data is the radius of gyration (R_g), which can be estimated using the Guinier approximation (Svergun et al. 2013). When using the Guinier approximation, the choice of the *q*-range to be used for fitting usually requires some manual estimation. The resulting small residuals are not ideal to really determine whether the *q*-range chosen is the best option. ***CSS***(*q*) can be used as a direct criteria to evaluate the data quality of the Guinier region (Fig. 2.8). By examining ***CSS***(*q*), unreliable data points can be readily identified and excluded as their low *CSS* suggests non-standard divergence from the support. The region with the best linearity and the largest mean in ***CSS***(*q*) should be chosen to conduct the Guinier approximation. Again, here, ***CSS***(*q*) are independent of the choice of background correction.

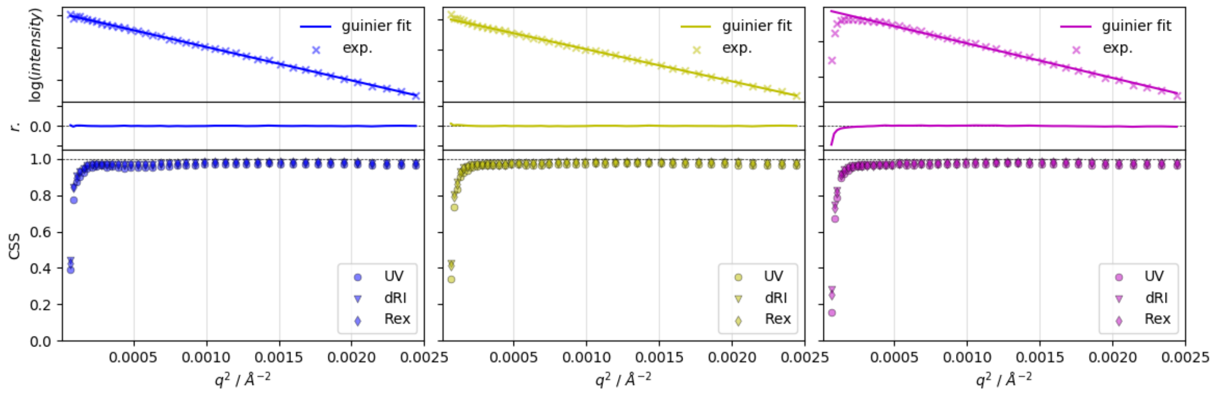


Figure 2.8 CSS used to identify the optimal *q*-range for Guinier approximation. The relationship between the ***CSS*** and the quality of the Guinier approximation is shown for the same data using the *FANCY* (blue), *BEFORE* (gold) and *AFTER* (purple) background corrections. Top: log(*I*) against q^2. The data are shown as x and the Guinier fit as a solid line. Middle: the corresponding fitting residuals. Bottom: the *CSS versus* q^2 using the *CHROMS* singles (*UV* (circles), *Rex* (triangles) and *dRI* (diamonds)) as support.

In summary, proper use of CSS leads to a unique and well-behaved data model for background correction, choice of sample frames to be averaged and choice of *q*-range for the calculation of R_g. Further details are discussed in §2.3.

2.2.3.3 Robustness Testing

The utility of CSS was further validated using publicly available *CHROMS-XSS* datasets in the small-angle scattering biological database (Table 2.2). For each dataset, the database includes the raw and background corrected *XSS* data for each frame, as well as the *CHROMS* signals. In the following analysis I have reproduced the background corrections defined in SASBDB, and have calculated the corresponding ***CSS***(*q*).

Table 2.2 Details of benchmarking datasets from SASDBD

code/sample	buffer	experiment†	*CHROMS*	background
SASDFP8 carbonic anhydrase 2 (Fig. 2.9-2.10)	50 mM HEPES 150 mM NaCl 2% *v/v* glycerol pH 7	flow rate: 0.5 mL/min concentration: 11.9 mg/mL injection volume: 100.00 uL	*UV* *dRI* *Rex*	1288-1354
SASDFS8 alcohol dehydrogenase 1 tetramer (Fig. 2.11-2.12)	50 mM HEPES 150 mM NaCl 2% *v/v* glycerol pH 7	flow rate: 0.5 mL/min, concentration: 9.2 mg/mL injection volume: 100.00 uL	*UV* *dRI* *Rex*	1274-1408, 2226-2267
SASDFR8 BSA dimer (Fig. 2.13-2.14)	50 mM HEPES 150 mM NaCl 2% *v/v* glycerol pH 7	flow rate: 0.5 mL/min, concentration: 5.0 mg/mL injection volume: 100.00 uL	*UV* *dRI* *Rex*	1-1049, 1953-2129
SASDFN8 Apoferritin light chain 24-mer (Fig. 2.15-2.16)	50 mM HEPES, 150 mM NaCl, 2% *v/v* glycerol, pH 7	flow rate: 0.5 mL/min, concentration: 11.0 mg/mL injection volume: 50.00 uL	*UV* *dRI* *Rex*	519-1009, 1992-2823

† Data were collected on EMBL P12, PETRA III (DESY, Germany). A GE superdex 200 increase 10/300 column was used for the SEC system. The hutch temperature was 20 ℃.

SASDFP8 - Carbonic anhydrase 2

The X-ray data (*XSS*) for this data set show clear evidence of radiation damage during the data acquisition. This manifests as an increasing background at low resolution (Fig. 2.9 left). For this dataset three *CHROMS* signals are available for use in ***CSS***(*q*) calculation. All give a similar result, showing the first few data points are strongly interfered with by the radiation damage components. In addition, some clear outlier data points at high *q* can be identified. In Fig. 2.10 shortcomings in the background subtraction can also be identified. At very low *q*, the later frames used for averaging are clearly under corrected, whereas at high *q* there are an increasing number of outlier data points where the centre of the sample elution peak is over corrected and the overall signal is within the noise. Even at mid-*q*-ranges, where the background subtraction is reasonable, a slightly compromised score is observed due to a combination of the background modulation and peak broadening caused by an unstable flow.

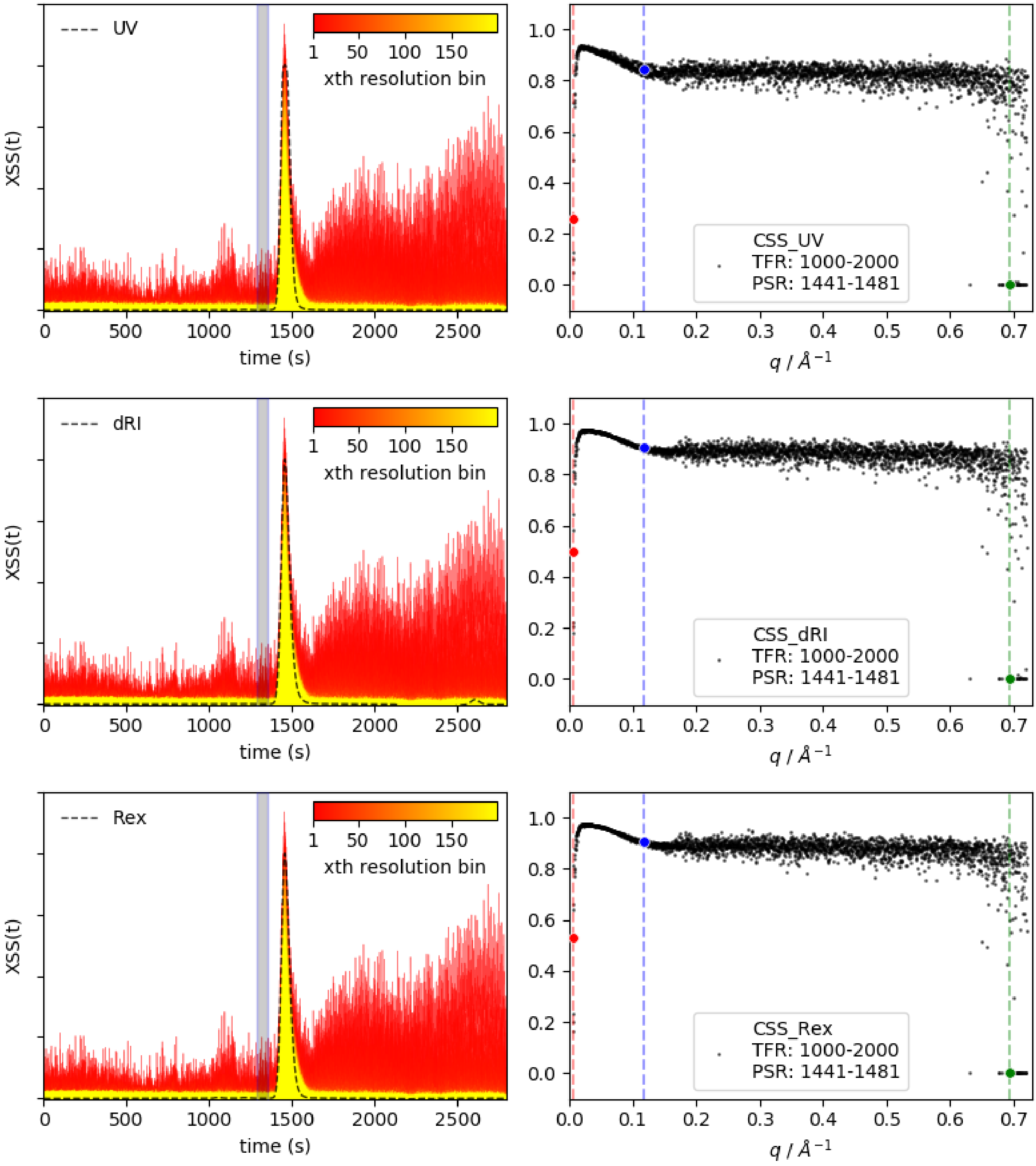


Figure 2.9 CSS analysis of post-processed *XSS* data for SASDFP8. Left panel: a comparison between the *XSS* of SASDFP8 (coloured lines) and the three different *CHROMS* signals (dotted line). The background frames are highlighted by the grey bar. Right panel: the resulting ***CSS***(q) given *Rex*, *dRI* and *UV* as supports, respectively. *TFR* is set from 1000 to 2000 and *PSR* is set from 1441 to 1481. The dashed vertical lines with circle markers mark the $I_q(t)$ at the corresponding q positions in Fig. 2.10.

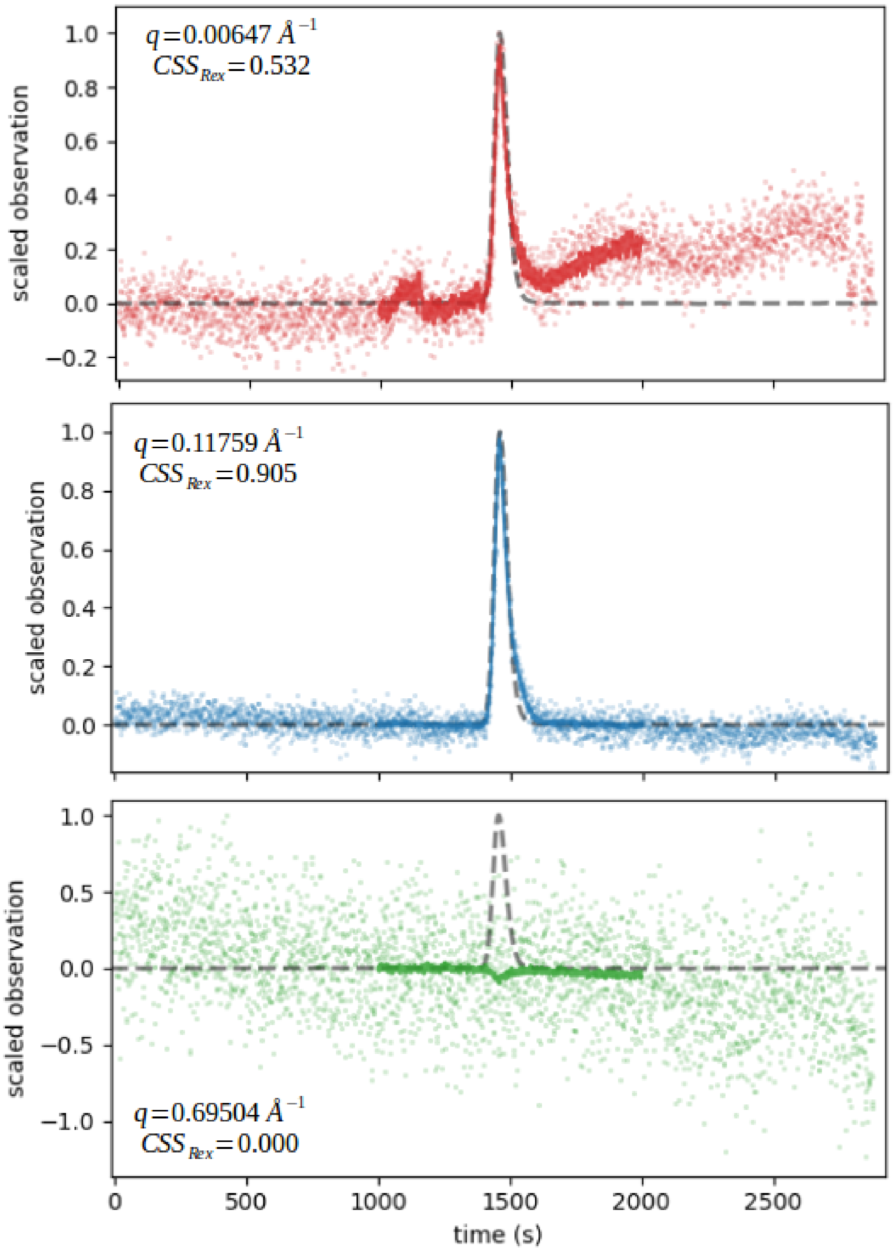


Figure 2.10 *XSS* signal at different *q* for post-processed SASDFP8. *CHROMS* signal (grey dotted line), *XSS* data (coloured dots), denoised *XSS* across the elution peak (solid line). Top panel: at low *q* the first few data points are strongly compromised by the radiation damage components, reflected in the low *CSS*. Middle panel: in the mid-range *q*, a background modulation and peak broadening caused by unstable flow result in a slightly compromised *CSS*. Bottom panel: in the high *q*-region for this outlier data point the signal is indistinguishable from the noise and thus has a *CSS* of zero.

SASDFS8 - Alcohol dehydrogenase 1 tetramer

The X-ray data (*XSS*) for this dataset indicate the difficulties in making a background correction, particularly at low resolution (Fig.2.11 left, and Fig. 2.12 top), as a result of radiation damage. For this dataset three *CHROMS* signals are available for use in ***CSS***(*q*) calculation. All give similar results. At low q, despite the rising baseline after background correction, the elution peak itself is relatively well treated, reflected in the reasonable *CSS*. At mid and very high *q* some clear outlier data points can be identified, where the *CSS* is very low, but overall the ***CSS***(*q*) suggests that the background correction is reasonable.

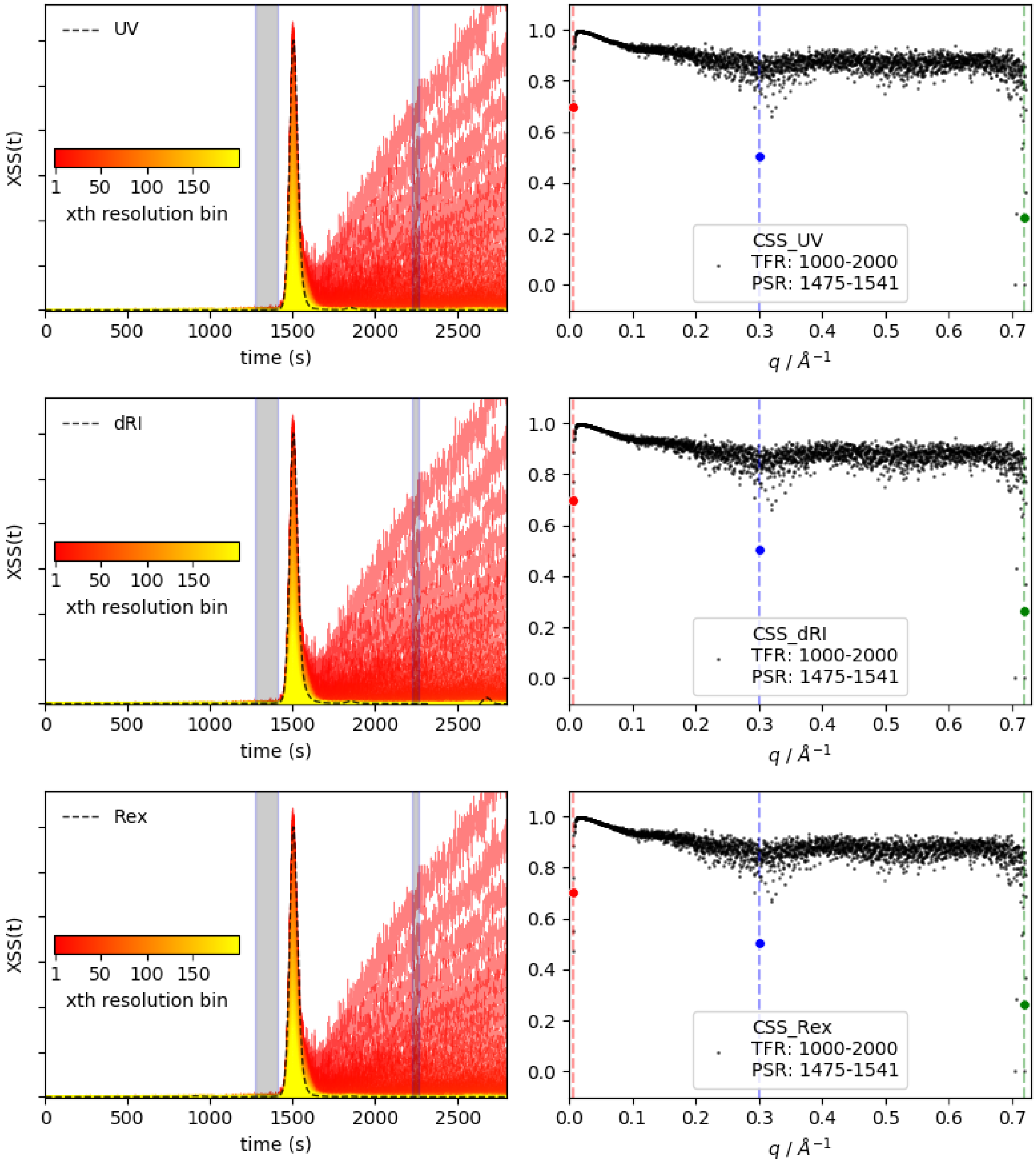


Figure 2.11 CSS analysis of post-processed *XSS* data forSASDFS8. , Left panel: a comparison between the *XSS* of SASDFS8 (coloured lines) and the three different *CHROMS* signals (dotted line). Frames used for the background correction are highlighted with a grey bar. Right panel: the resulting ***CSS***(q) given *Rex*, *dRI* and *UV* as supports, respectively. *TFR* is set as from 1000 to 2000 and *PSR* is set as from 1475 to 1541. The dashed vertical lines with circle markers mark the $I_q(t)$ at the corresponding q positions in Fig. 2.12.

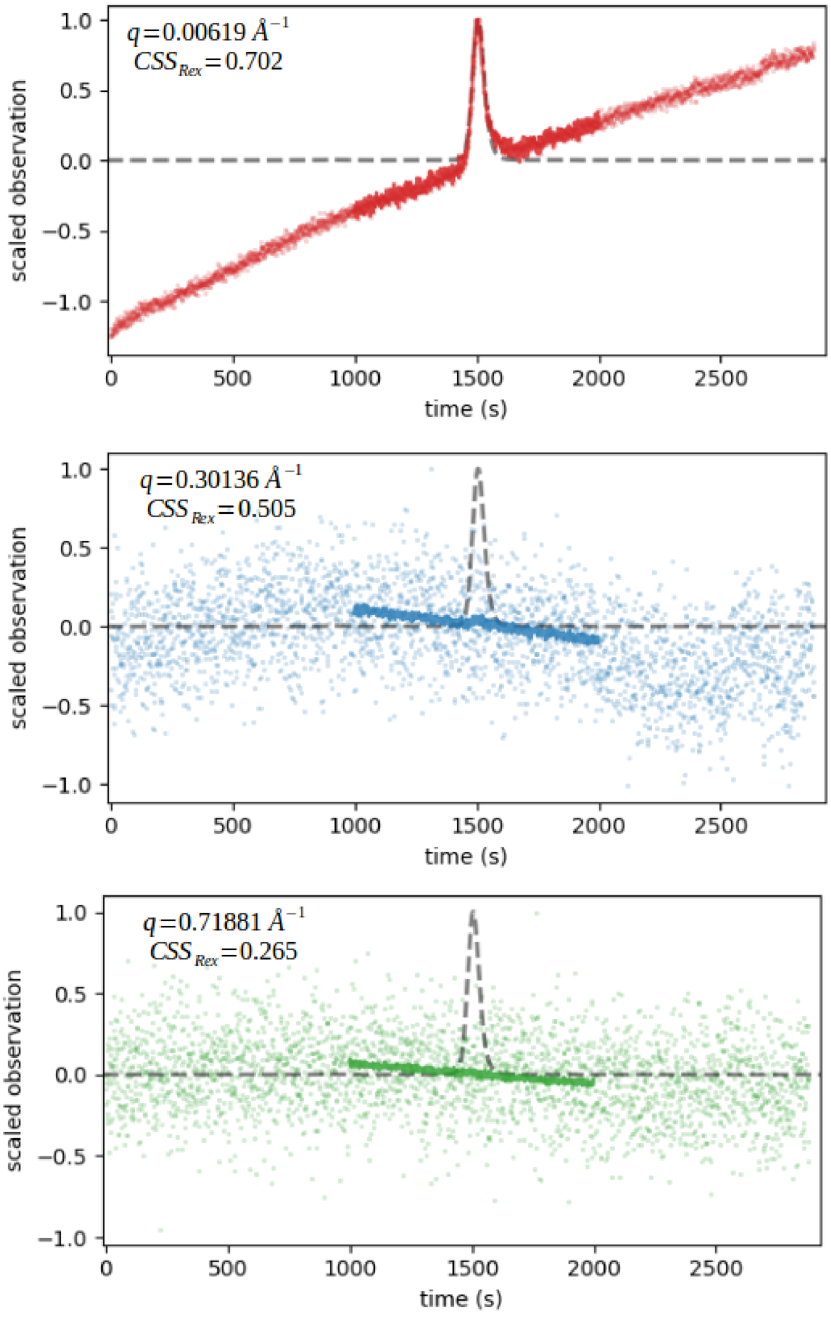


Figure 2.12 *XSS* signal at different q for post-processed SASDFS8. *CHROMS* signal (grey dotted line), *XSS* data (coloured dots), mean *XSS* across the elution peak (solid line). Top panel: at low q, despite the rising baseline after background correction, the *CSS* remains reasonable. Middle panel: in the mid q-range certain clear outlier data points are observed due to the background modulation. Bottom panel: similarly, outliers appear at high q.

SASDFR8 - BSA dimer

The X-ray data (*XSS*) for this data set indicate some contributions from radiation damage particularly at low resolution (Fig. 2.13 left). For this dataset three *CHROMS* signals are available for use in ***CSS***(*q*) calculation. All give similar results. The first few data points at low *q* are not strongly interfered with by the radiation damage but are undermined by an over-subtraction during the background correction. This suggests the background frames are not correctly chosen. From a *q*-range of 0.2 onwards, there are a large number of low-scored data points that strongly comprise the quality of the dataset, potentially due to the radiation damage. However, in the mid-high *q* range, a good signal is obtained and this is confirmed by a reasonable *CSS* except for some outliers with very low *CSS*. This means that the background averaged from two parts of elution indeed helps to stabilize the correction in the mid-high range.

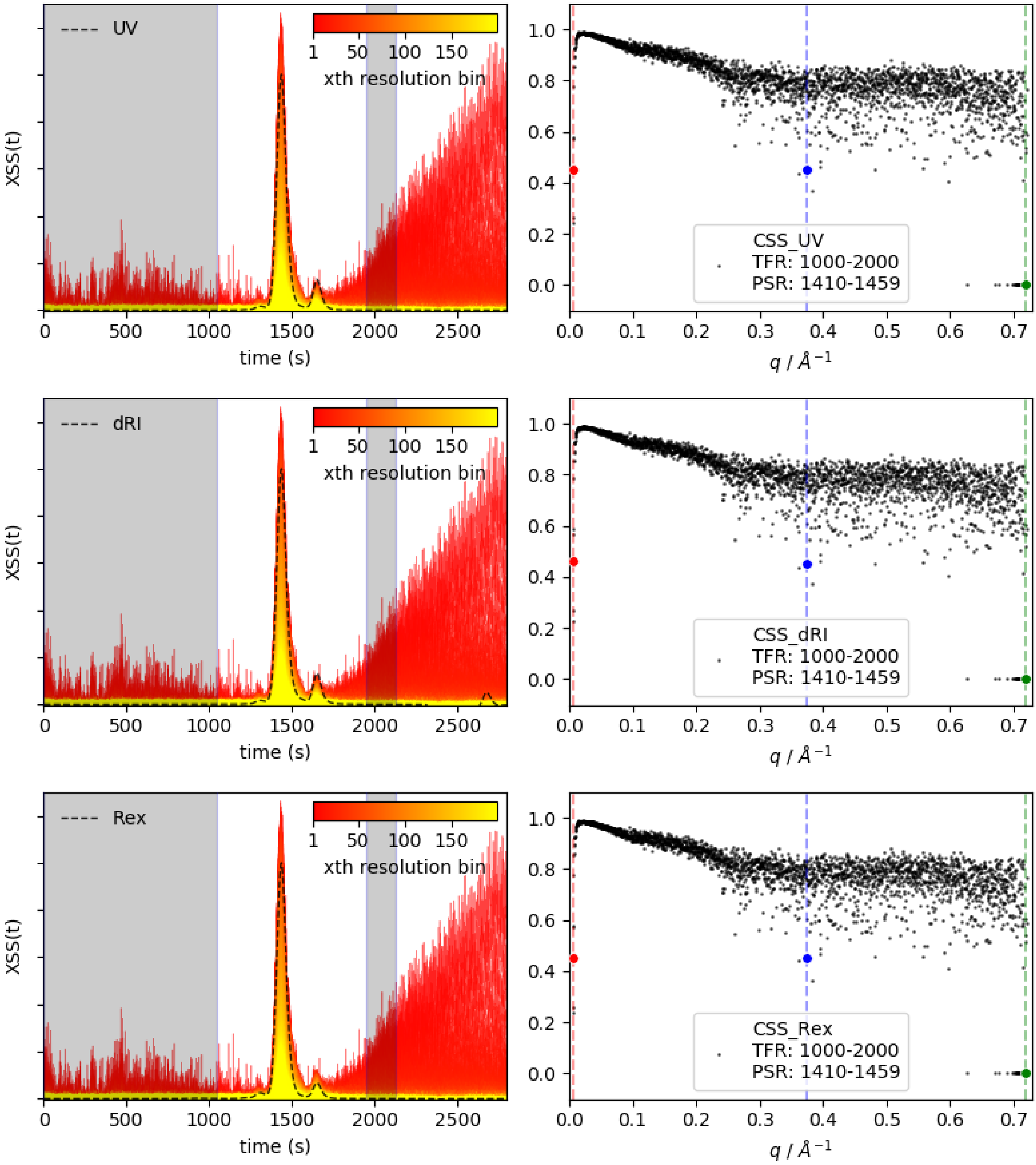


Figure 2.13 CSS analysis of post-processed *XSS* data for SASDFR8 Left panel: a comparison between the *XSS* of SASDFR8 (coloured lines) and the three different *CHROMS* signals (dotted line). Right panel: the resulting ***CSS***(q) given *Rex*, *dRI* and *UV* as supports, respectively. The *TFR* is from 1000 to 2000 and the *PSR* is from 1410 to 1459. The dashed vertical lines with circle markers mark the $I_q(t)$ at the corresponding q positions in Fig. 2.14.

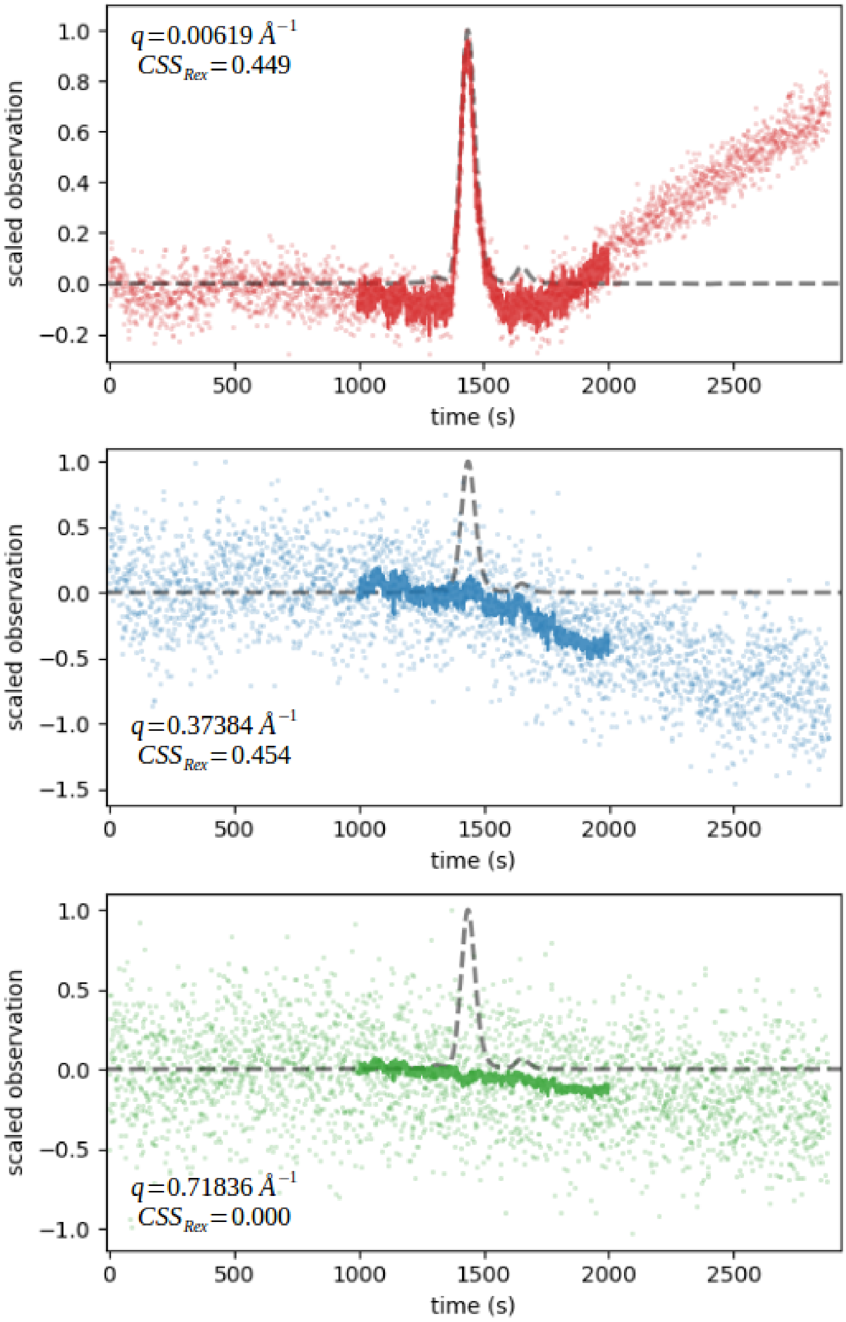


Figure 2.14 *XSS* signal at different q for post-processed SASDFR8. Top panel: at low q the data points are not strongly interfered with by the radiation damage but are undermined by the over-subtraction. Middle panel: In the mid q-range, a detected outlier. The instability of the signal leads to an unacceptable *CSS*. Bottom panel: For the outlier data points at higher q over-subtraction leads to a *CSS* of zero.

SASDFN8 Apoferritin light chain 24-mer

For this dataset a complex background correction was applied by the depositors which included frames from both before and after the sample elution peak (519-1009, 1992-282). Examination of the post processed *XSS* shows that this results in a reasonable background correction. For this dataset three *CHROMS* signals are available for use in ***CSS***(*q*) calculation. All give similar results. At low *q*, the background correction is not perfect, but this has little negative effect on the post processed *XSS* signal. The ***CSS***(*q*) is reminiscent of the sink function often observed for spherical particles such as apo-ferritin. Indeed, the period drops in ***CSS***(*q*) occur at exactly the *q*-ranges where the intensity of the *XSS* signal also drops. This is not unexpected. As the intensity drops by orders of magnitude at the valleys of the sink function compared to the peaks, the signal quality also drops, reflected in the decreased *CSS*. Again, from a *q* of 0.2 onwards, the number of outlier data points with low *CSS* begins to grow, although the overall quality of the post-processed data remains high.

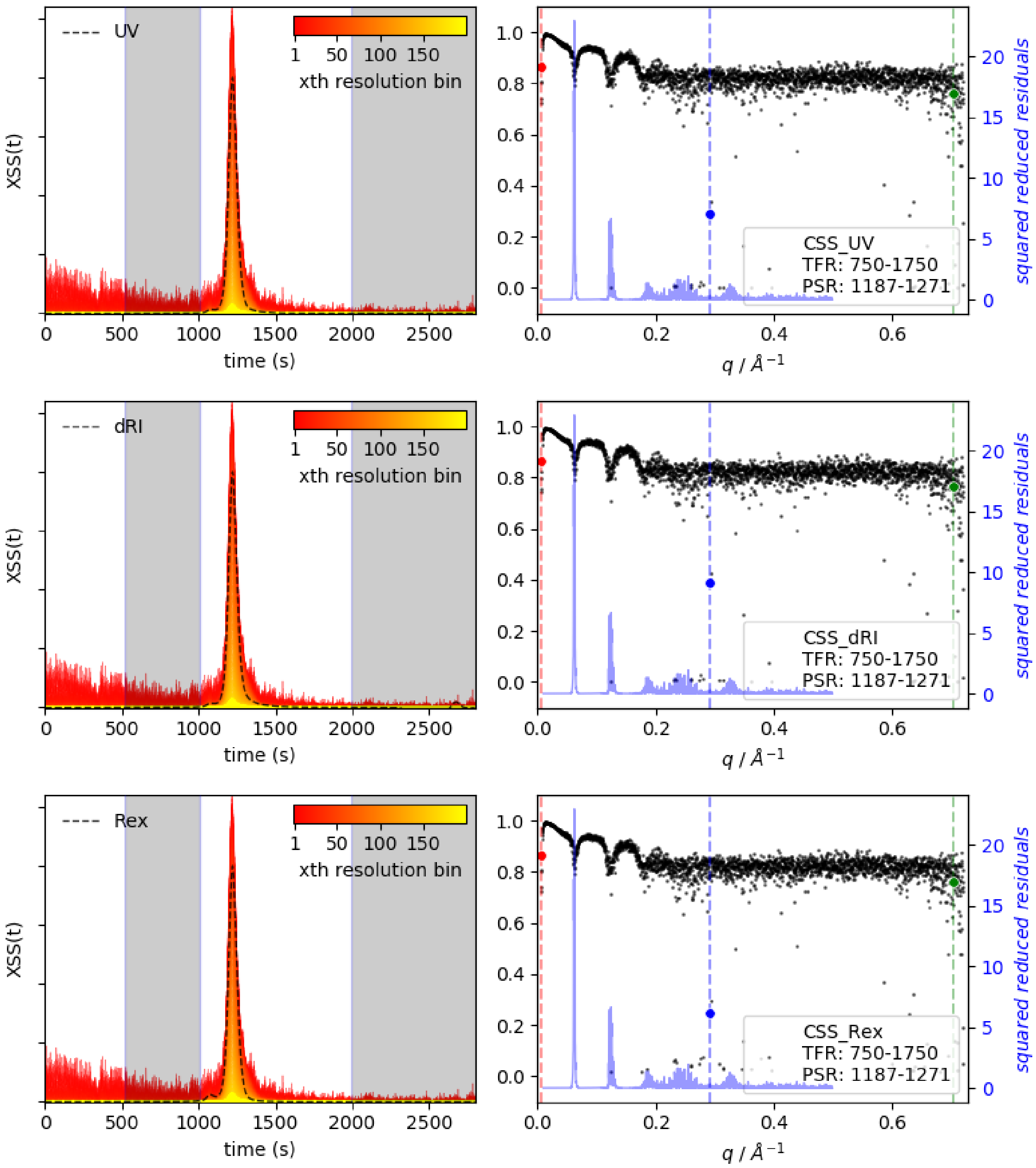


Figure 2.15 CSS analysis of post-processed *XSS* data for SASDFN8. Left panel: a comparison between the *XSS* of SASDFN8 (coloured lines) and the three different *CHROMS* signals (dotted line). Right panel: the resulting *CSS* given *Rex*, *dRI* and *UV* as supports, respectively. The *TFR* is from 750 to 1750 and *PSR* is from 1189 to 1271. The squared normalized residuals between the theoretical scattering of the crystallographic model (1IER) and the post-processed scattering profile are shown as blue lines. Interestingly, the positions where large discrepancies of fitting show are in accordance with the sudden drops of ***CSS***(q). This suggests the large fitting residuals are a consequence of imperfect post-processing instead of an improper model.

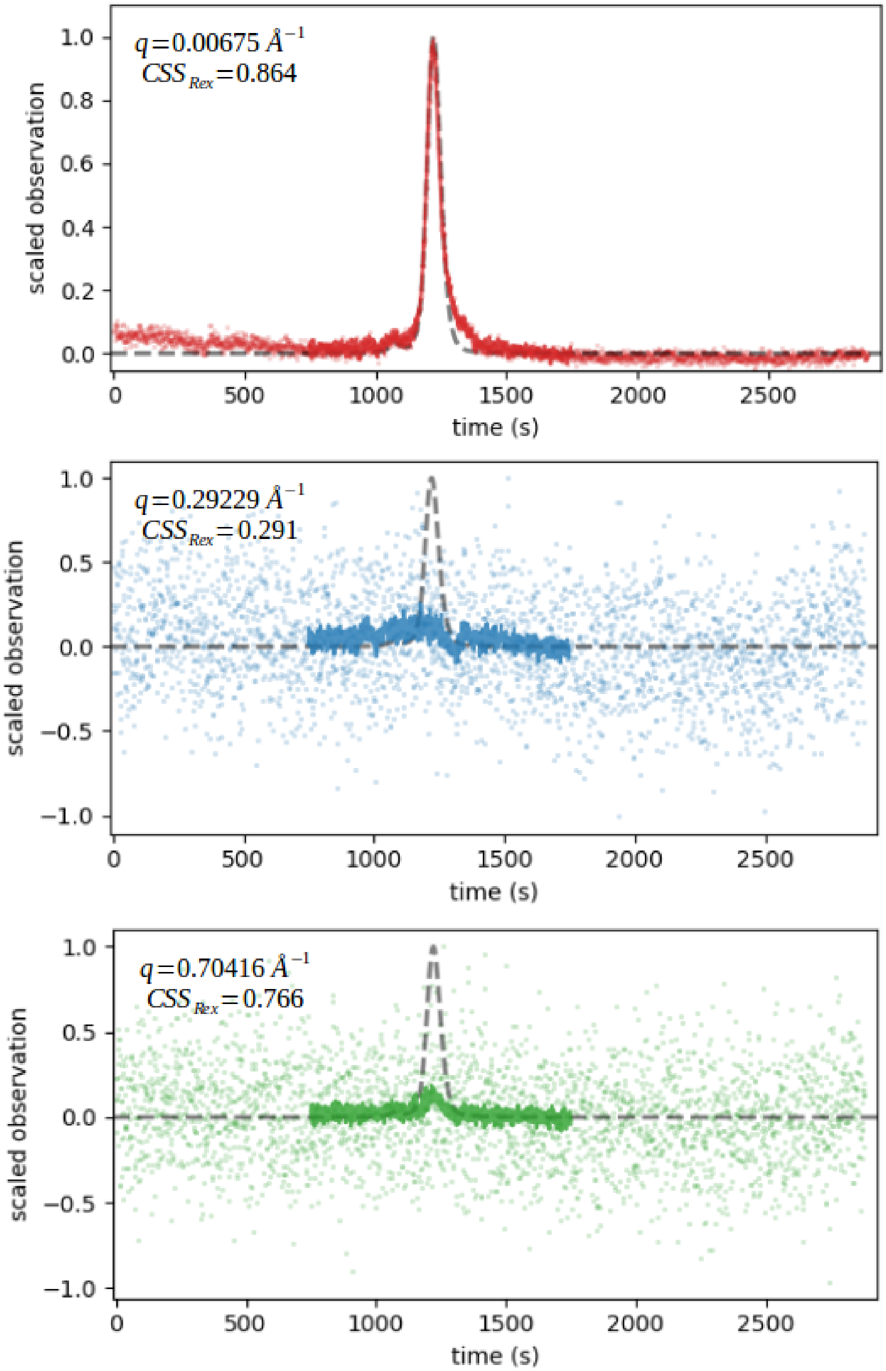


Figure 2.16 *XSS* signal at different q for post-processed SASDFN8. Top panel: the construction of the background is not perfect yet has little negative effect on the correct intensity at the first few bins. Middle panel: An outlier in the mid-q range. Modulation of background leads to an unacceptable *CSS*. Bottom panel: At the high-q range, although the *SNR* is extremely low, the *CSS* is still reasonable. This is an uncertain situation that would normally have had to rely on a purely subjective impression but can now be quantitatively assessed by *CSS*.

2.3.3.3 Implications of CSS for the Effective Structure Factor

Under certain conditions (see §1.2), the total scattering intensity can be represented by the product of the form factor *P*(*q*) and the structure factor *S*(*q*) (Brunner-Popela & Glatter 1997):

$$I(\mathrm{q}) \propto NP(q)S(q) \qquad (2.29)$$

where N is the number of protein molecules per unit volume in the solution; *P*(*q*) is the form factor of a given protein (the scattering from a single protein molecule after orientational averaging); *S*(*q*) is the structure factor, which contains information on the inter-particle interaction, particle geometry and solution (in-)homogeneity.

In real space the pair distance distribution function *p*(*r*) can be approximated by the linear combination of a finite number of functions $\varphi_{\mathrm{i}}(\mathrm{r})$ with weighting factors of c_{i}:

$$p(r) = \sum c_{\mathrm{i}}\varphi_{\mathrm{i}}(\mathrm{r}) \qquad (2.30)$$

The Fourier transform of *p*(*r*), taking account of the linearity property, is

$$P(q) = \mathrm{FT}[\sum c_{\mathrm{i}}p_{\mathrm{i}}(\mathrm{r})] = \sum c_{\mathrm{i}}\mathrm{FT}[p_{\mathrm{i}}(\mathrm{r})] = \sum c_{\mathrm{i}}\psi_{\mathrm{i}}(\mathrm{r}) \qquad (2.31)$$

For a system that fulfills weak-monodispersity, *P*(*q*) is a function that only relates to the averaged geometrical volume of the particle. If the non-intermolecular interaction assumption is also satisfied, the *S*(*q*) is a constant equal to 1. As mentioned in §2.2.1, at any given *q* this leads to a resolved peak in the scattering elution trace, *I*(*t*, *q*), as a function of *N* (i.e. particle concentration). It is obvious that under the same assumptions UV, Rex and dRI also all respond only to *N*. The deviation from 1 of the ***CSS***(*q*) between the *XSS* and *CHROMS* signals is therefore due almost entirely to an uncorrected background mismatch.

However, from the practical point of view, a system without any particle interference and polydispersity is hardly ever observed, and concentration extrapolation is not routinely explored or utilised in SEC-SAXS. The intermolecular interaction, ionic strength and polydispersity will generate oscillations in *S*(*q*) (Brunner-Popela & Glatter 1997). With the increases of the intermolecular interaction (as a consequence of increased concentration, Fig. 2.17 left), the oscillation becomes more pronounced and the maxima are slightly shifted to higher-*q*. The value of *S*(0) drops significantly due to the decrease in compressibility. If the polydispersity increases (Fig. 2.17 right), the oscillation becomes smoother. At lower-*q*, the effect of polydispersity is not significant. At the higher-*q*, however, The average effect of polydispersity flattens *S*(*q*) towards a value smaller than 1.0 and approaches unity at very high *q*. Because of the oscillation in *S*(*q*), the experimental scattering profile also shows oscillations and therefore deviates from that expected for an ideal situation where there is no interparticle interference and the system is monodisperse (Fig. 2.17). Since *S*(*q*) is *q*-dependent, this effect is observable in the ***CSS***(*q*).

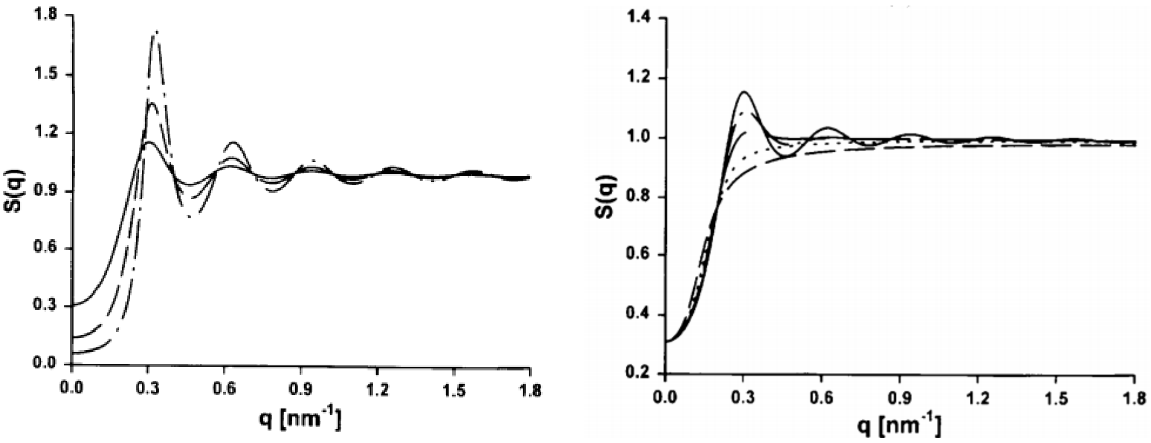


Figure 2.17 Oscillation of the structure factor S(q) due to (left) increasing concentrations and (right) increasing polydispersity. (Left) The volume fractions are 0.15 (solid line), 0.25 (dashed line) and 0.35 (dashed-dotted line). The polydispersity coefficient is fixed at 0.0. (Right) The polydispersity parameters are 0.0 (solid line), 0.15 (dashed-dotted line), 0.25 (dashed line), 0.40 (dotted line) and 0.60 (long-dashed line). The volume fraction is fixed at 0.15. The analytical expression of the structure factor is referred to as the Percus-Yevick approximation (Baxter 1970). Polydispersity parameters are defined as the width of the normal distribution divided by the maximum particle radius. Figure adapted from (Brunner-Popela & Glatter 1997).

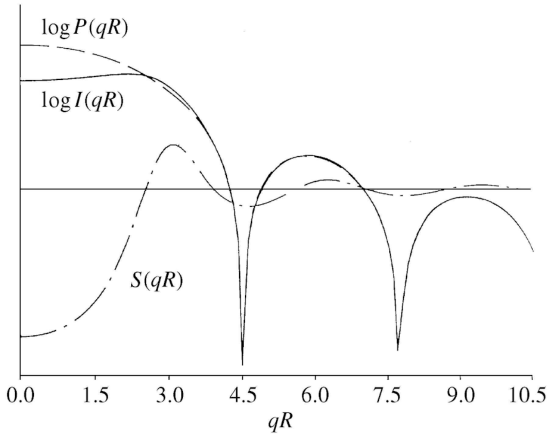


Figure 2.18 The effect of particle interference on the scattering profile. The scattering intensity *I*(*q*) (dashed line) of the ideal particle deviates from the form factor *P*(*q*) (dashed line) because of the presence of oscillating *S*(*q*) (dashed-dotted line). This effect is dimensionless (independent of the particle size). Figure adapted from (Weyerich et al. 1999).

In this study, three *CHROMS* signals were used, *UV*, *dRI* and *Rex*. The integrated *UV* absorbance (*A*) can be expressed by the Lambert-Beer law (Mayerhöfer et al. 2019), $A = \varepsilon \ell N/N_a$, where *N* is the number of molecules per unit volume in the solution, N_a is Avogadro's constant, ε is the molecular extinction coefficient. For any solution of particles, the refractive index contrast, *dRI*, depends linearly on the concentration of the particles (Tan & Huang 2015). The excess Rayleigh scattering ($R(q_L)$) is a measurement of the contribution of the particles in solution to the scattered light intensity. Since the theory of small-angle scattering is uniform, the expression of $R(q_L)$ is similar to the expression of X-ray scattering intensity *I*(*q*) and can be written as $R(q_L) \propto NP(q_L)S(q_L)$ (Schmitz 1990), where the light scattering vector $|\boldsymbol{q}_\mathrm{L}| = 4\pi n_\mathrm{s}\sin(\theta)/\lambda$ (n_s is the refractive index of the sample). The difference between $R(q_L)$ and *I*(*q*) is that the scattering length density difference for light scattering is the excess ability of a particle to reflect light, compared to the electron density contrast for SAXS. Light scattering instruments report the excess Rayleigh scattering at zero angle (R_ex).

For these three *CHROMS* signals it is obvious that, when *S*(*q*) is close to 1.0 for the whole *q*-range, background mismatch is the major source compromising ***CSS***(*q*) because both *I*(*t*, *q*) and *CHROMS* signal are functions of *N*. However, if the effect of *S*(*q*) cannot be ignored, ***CSS***(*q*) will reflect the deviation from *S*(*q*) and manifest a similar oscillation pattern. I hypothesise that, if perfect background subtraction is achieved, ***CSS***(*q*) will in fact represent the effective structure factor (although the absolute value will not be correct).

Table 2.3 A summary on the results of CSS-based resampling

code		**# of outliers (q_{min}-0.55 Å^{-1})**	q_{min} **(Å^{-1})**	**CRYSOL χ^2 †**	**GNOM estimate †**
SASDFP8	initial	0	0.00730	2.208	0.9159/13.17
	resampled	0	0.01753	2.115	0.9287/10.32
SASDFS8	initial	0	0.00730	1.496	0.7074/46.59
	resampled	0	0.01006	1.130	0.8392/6.036
SASDFR8	initial	8	0.01338	4.609	0.8683/7.002
	resampled	1	0.01310	3.487	0.9361/15.07
SASDFN8	initial	34	0.00923	8.433	0.7575/2.325
	resampled	0	0.01089	8.332	0.8351//11.77

† Data are fitted against the theoretical profile from the all-atomic model included in the corresponding deposition.

† GNOM estimate is reported as "total quality estimate/ALPHA" ((Brunner-Popela & Glatter 1997; Svergun 1992)). Total quality estimate is an assessment on the solution quality of indirect Fourier transform. It has a scale from 0.0 (worst) to 1.0 (best). ALPHA values were automatically searched by maximizing the total quantity estimate. D_{max} was set as the same as the value in the corresponding deposition.

2.3.3.4 Resampling by CSS

Since SAXS profiles are oversampled, it is possible to design a resampling method according to ***CSS***(q) so that individual low-scoring data points can be filtered out. Two principles are applied: (1) For the Guinier region (very low q region), the points with *CSS* below the consistent value are discarded due to structure factor related deviations. (2) After ***CSS***(q) has reached a consistent value, for every five q points, the point with maximum *CSS* is kept whereas the other four are discarded. This is a safe filter for particles with a size between 10 to 50 Å (i.e. proteins) and most automated radical integration protocols on synchrotron beamlines to retain enough information according to Nyquist-Shannon sampling theorem (§3.1)

2.3 CSS-guided Self-Adaptive Background Correction

2.3.1 Background

Optimization of data quality relies strongly on the ability to test different hypotheses about the data model used for post processing (choice of background model, outlier rejection, choice of data range *etc*.). Two recent studies have taken different approaches to both describing the data model and optimising the post-processing (Table 2.4). In these cases, the final post processing is guided by a model-based hypothesis test based on the value of the estimator parameter.

In my opinion, suggesting a strong assumption that forms the basis of the data model is not problematic. However, the fundamental issue of such model-based hypothesis tests is that they are forced to solve an inverse problem and the consequential solutions are inevitably ill-posed (Adler & Öktem 2017). The estimators described above essentially perform reliability tests. This means that measurement errors, even if small, will interfere with the solution algorithm and potentially dramastically alter the resulting interpretation. A much better approach is to find an estimator that can evaluate data accuracy (López Cárdenas 2016). The ability of CSS to provide an objective assessment of the accuracy of post-processed *I*(*t,q*) data collected using chromatography-SAXS, and where an independent measure of the sample elution behaviour is available, has been demonstrated earlier in this Chapter. In this section, I will now describe the iterative method guided by a CSS-based estimator that I have developed to maximise data quality.

As a -proof-of-principle, I demonstrate my approach by targeting data contaminated by radiation damage, as this is the major issue leading to poor post-processing of chromatography-SAXS datasets deposited in SASBDB.

Table 2.4 Summary of two methods for optimising the post-processing

	Integral baseline (Brookes et al. 2016)	**SVD with Guinier-optimized linear combination (SVD-LC) (Malaby et al. 2015)**
Usage	background subtraction for SEC-SAXS data with capillary fouling deposits	extract individual sample component from two-components SEC-SAXS data
Assumption(s)	a) the capillary fouling is proportional to the sample's	a) the scattering profile can be reconstructed accurately by the linear

	scattering intensity while exposed to the beam. b) the proportionality coefficient is species independent.	combination of first (U0) and second-ranked (U1) significant bases in the U matrix calculated from SVD. b) the Guinier region must be perfectly linear.
Model	linear recurrence sequences	single parameter tuned linear combination of U0 and U1
Estimator	absolute difference between the coefficient of the previous and the present iterations	coefficient of determination (R^2) for the linear regression of selected Guinier region
Iterative parameter inference	minimising the estimator	global maximum of the estimator

2.3.2 the Algorithm

2.3.2.1 Formulation of the Self-Adaptive Scattering Background

Let $[\boldsymbol{X}]_r \to \boldsymbol{X}$ be a subset of $\boldsymbol{X}$ with frames within the range r. The number of frames in r is N. The standard approach to background correction constructs the scattering background ($B(q)$) as the numerical average of the frames in r:

$$B(q) = \frac{1}{N}\sum[\boldsymbol{X}]_r \tag{2.32}$$

It is clear in the examples presented in the previous section that the scattering background is almost never constant over the whole elution period of the chromatography-SAXS measurement. Instead, it varies but adapts certain patterns related to the flux of the incident X-ray beam, the flow pattern over the column and through the sample chamber, the appearance of radiation damage, etc.

For a particular dataset $\boldsymbol{X}$, we begin with the assumption that the solution scattering background (includes three components: buffer ($\beta(t$,q)), a growing radiation-damage component ($\varphi(t,q)$), and a decreasing radiation-damage component ($\theta(t,q)$). The total background $B(t,q)$ can then be written as:

$$B(t,q) = \beta(t,q) + \varphi(t,q) + \theta(t,q) \tag{2.32}$$

The buffer background $\beta(t,\mathrm{q})$ is then predicted using a linear model of M features derived from the svd-denoised $[\boldsymbol{X}]_s$. M is the total number of resolution bins. Range s is a stable non-sample region.

The growing radiation-damage component can be expressed as a cumulative distribution function (CDF) of a normal distribution.

$$\varphi(t,q) = 0.5C(q)[1 + \mathrm{erf}(\frac{(t-t_i)}{\sqrt{2}\sigma})] \tag{2.33}$$

where $\mathrm{erf}(\mathrm{x})$ denotes an error function. t_i is the time point at which the rate of formation of the radiation-damage component reaches its maximum. σ is the standard deviation of the normal distribution. $C(q)$ is a scaling constant and can be estimated from

$$C(q) = \frac{1}{n_F}\sum_{t\in F}(\boldsymbol{X} - \beta(t,q)) \tag{2.34}$$

Range F is a user-defined region where the total amount of the radiation-damage component is believed to reach a maximum, n_F is the size of range F.

The decreasing radiation-damage component is expressed as a modulated CDF of a normal distribution:

$$\varphi(t,q) = 0.5\mathfrak{M}(t)[1 + \mathrm{erf}(\frac{(t-t_i^{'})}{\sqrt{2}\sigma'})] \tag{2.35}$$

where $\mathfrak{M}(t)$ is predicted using the linear model of M features derived from $[\boldsymbol{X} - \beta(t,q) - \beta(t,q))]_D$. Range D is of the same length as F.

2.3.2.2 Accuracy Estimator

Let ξ be a set of ***CSS***(q) calculated from a particular scattering background $B(t,q)$ and a corresponding CHROMS. Suppose that $b(\xi)$ is the regression coefficient of q on ξ and $\bar{\xi}$ is the arithmetic mean value of the ***CSS***(q). A statistically perfect correction would be indicated by

$b(\xi) = 0$ and $\bar{\xi} = 1$. The $b(\xi) = 0$ indicates that the correction is consistent for every resolution bin while $\bar{\xi} = 1$ means maximum similarity between the post-processed XSS and the chosen *CHROMS* has been reached.

The estimators ($\bar{\xi}$ and $b(\xi)$) can be affected by the resolution cutoff and the choice of *CHROMS*. Therefore, to generalize the estimation, scaling is required. A reasonable solution is to measure the relative improvement of the self-adaptive background correction with respect to the standard background subtraction. The target function is thus formulated as:

$$0.7\exp(-\frac{\bar{\xi}-1}{\bar{\xi}_{\mathrm{ref}}-1}) + 0.3\exp(-\frac{b(\xi)}{b(\xi_{\mathrm{ref}})}) \qquad (2.36)$$

where $b(\xi_{\mathrm{ref}})$ and $\bar{\xi}_{\mathrm{ref}}$ refer to the same values derived from the standard background subtraction. A correction that is better than the reference gives a $\boldsymbol{T}(\xi)$ larger than e^{-1}. E.q. 2.36 is empirical and assumes the goodness of a correction weighs higher than its consistency over q.

2.3.2.3 Iterative Optimization

To search for the optimal $B(t, q)$, a simple Monte Carlo optimization involving following steps is used:

1. select range r (successive or non-successive) and compute the corresponding standard background $B(q)$
2. standard background subtraction $\boldsymbol{X}_{\mathrm{standard}} = \boldsymbol{X} - B(q)$
3. select *TFR* and *PSR* and calculate the ***CSS*** for $\boldsymbol{X}_{\mathrm{standard}}$
4. calculate $b(\xi_{\mathrm{standard}})$ and $\bar{\xi}_{\mathrm{standard}}$
   iteration starts, set loop index $i = 0$ and maximum iterations
5. initialize s_0 , F_0, $t_{i,0}$, $t'_{i,0}$, σ_0, σ'_0 and the searching ranges
6. randomly select s_n , F_n, $t_{i,n}$, $t'_{i,n}$, σ_n and σ'_n
7. compute $B_i(t, q)$ and background corrected $\boldsymbol{X}_i = \boldsymbol{X} - B_i(t, q)$
8. using the same *TFR* and *PSR* as in step 3 calculate the ***CSS*** for $\boldsymbol{X}_i$
9. calculate $b(\xi)$ and $\bar{\xi}$
10. the quality of the correction is evaluated by $\boldsymbol{T}(\xi)$
11. stop if $\boldsymbol{T}(\xi)$ >= threshold. Otherwise add i by 1 and go back to step 6
12. stop if maximum iterations reaches

2.3.3 Results and Discussion

2.3.3.1 Checking the Validity of Background Modelling

For any optimization, the choice of initialization can be critical for the success of the algorithm. Creating a background correction method specific to radiation damage is equivalent to building an attractor. An attractor is a term used in the study of dynamical systems and is used to describe a manifold in Euclidean space which draws the iterates towards it (Boeing 2016). If the iterate enters the wrong attractor, it will either never converge or will require an extremely large number of iterations to escape (Lo 2011). To check whether the background correction approach described above is a proper attractor, a comparison was made between standard background corrections and the self-adaptive background correction of SASDF99 where the data are strongly contaminated by radiation damage. If a rational initialization of the latter case leads to a significant improvement of *CSS*, the self-adaptive background correction can be considered as a good attractor.

Table 2.5 Defining the initial parameters for the self-adaptive background correction for SASDF99

	criteria	initial parameters	Wiener + SGD	SVD + SGD
s_0	$\forall\, t \leq 1000$, $\lvert I'(t)_{\mathrm{mean}}\rvert < 0.1$	{500, 501, …, 900}	{156, 157, …,920 }	{147, 148, …, 905}
$t_{i,\,0}$	$t \geq 2000 \wedge$ $\max\{I'(t)_{\mathrm{mean}}\}$	2350	2364	2342
F_0	$\forall\; t \geq 2000$, $\lvert I'(t)_{\mathrm{mean}}\rvert < 0.1$	{3000, 3001, …, 3090}	{3015, 3016, ... , 3092}	{2938, 2939, …, 2988}
σ_0	approx. FDHM*	350	805	844
$t'_{i,\,0}$	$t \geq 2000 \wedge$ $\min\,\{I'(t)_{\mathrm{avrg}}\} \wedge$ $I'(t)_{\mathrm{mean}} <= -0.1$	∞	N/A	N/A
σ'_0	--	∞	--	--
D	--	{}	--	--

* FDHM stands for full duration of half maximum

The initial guess is made by neutralizing two sets of parameters acquired through signal processing. Wiener filter and SVD denoising are applied to the mean $I(t, q)$ ($I(t)_{\text{mean}}$) and the first derivative ($I'(t)_{\text{mean}}$) of the mean XSS is deduced by Savitzky-Golay Differentiation (SGD). The criteria for parameter searching and the results are listed in Table 2.5

It should be noted that, in this dataset, the decomposition of the radiation-damage component cannot be detected near the time range where the monomeric BSA elutes (*PSR*).

The sigma is calculated according to

$$\sigma = \frac{\text{FDHM}}{2\sqrt{2\ln(2)}} \tag{2.37}$$

The self-adaptive background correction derived from the initial parameters in Table 2.5 is the aforementioned *FANCY* background correction. A comparison of ***CSS***(q) can be made between *BEFORE*, *AFTER* and *FANCY* (Fig. 2.19) using *BEFORE* as a reference. A significant improvement of ***T***(ξ) is evident for all the *CHROMS* for the *FANCY* background correction. This confirms the background model is a suitable attractor for optimization of data contaminated by radiation damage.

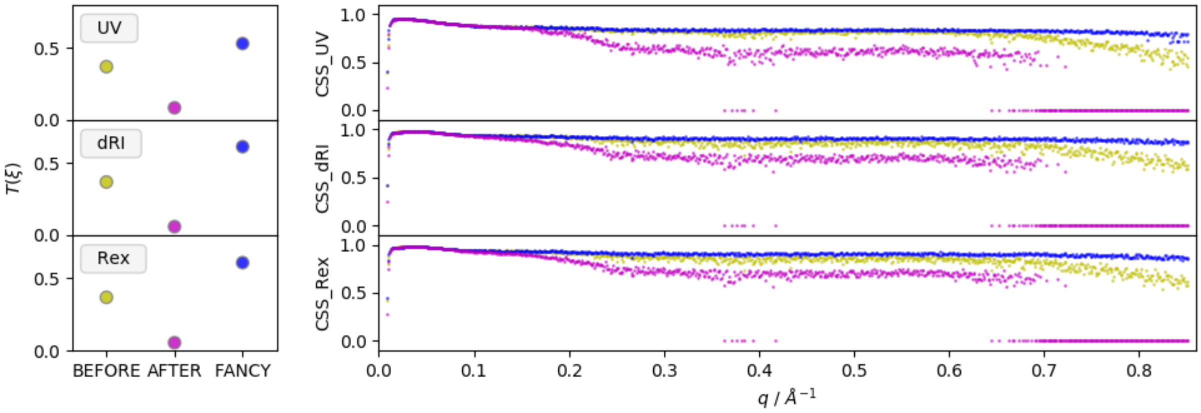


Figure 2.19 Left panel: $T(\xi)$ of *BEFORE*, *AFTER*, *FANCY*, using *BEFORE* as a reference, for three different *CHROMS*. Right panel: ***CSS***(q) for the *BEFORE* (gold), *AFTER* (purple) and *FANCY* (blue) background corrections with respect to q for each of the *CHROMS* signals.

2.3.3.2 Estimator Diagnostics

While the pursuit of perfectly corrected data through this iterative process is beyond the scope of this section, attention should be paid to how the value of the accuracy estimator evolves during the iteration. An interesting case (mBSA) in which $\boldsymbol{T}(\xi)$ hardly converges is found for a dataset collected by our collaborators (Dr Edward Snell & Tim Stachowski, Hauptman-Woodward Medical Research Institute, Buffalo, NY, USA) (Fig. 2.20). The experimental details are listed in Table 2.6.

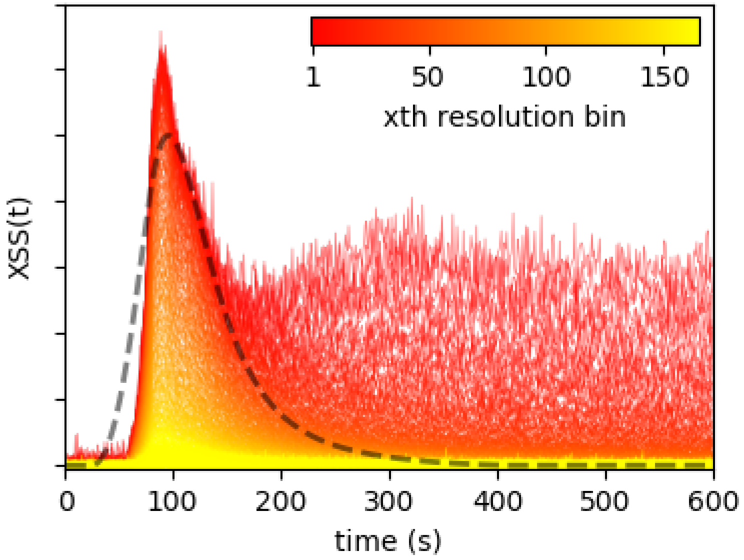


Figure 2.20 Standard background corrected $I(t, q)$ of mBSA before optimization using the CSS-guided self adaptive background correction approach. XSS data are shown as coloured lines. The dashed grey line shows the dRI signal collected with an in-line MALS system (DAWN HELOS-II, Wyatt Technology).

Table 2.6 The experimental details of the SEC-SAXS run

sample	**buffer**	**experiment**	***CHROMS***	***r***
BSA, monomer (mBSA)	137 mM NaCl 2.7 mM KCl 1.8 mM PBS pH 7.4	G1, CHESS column bypass mode temperature:20 °C flow rate: 0.18 mL/min concentration: 9.3 mg/mL injection volume:100.00 uL	*dRI*	{1, 2,..., 50}

During the search for an optimal background correction using the CSS-guided self-adaptive model, a rather stable ξ around q = 0.06 Å^{-1} is observed (Fig 2.21). At the same time, a negatively correlated trend in $T(\boldsymbol{\xi})$ before and after q = 0.06 Å^{-1} is observed. This is in agreement with the observations made during a sophisticated SAXS radiation damage quantification by Hopkins and Thorne (Hopkins & Thorne 2016).

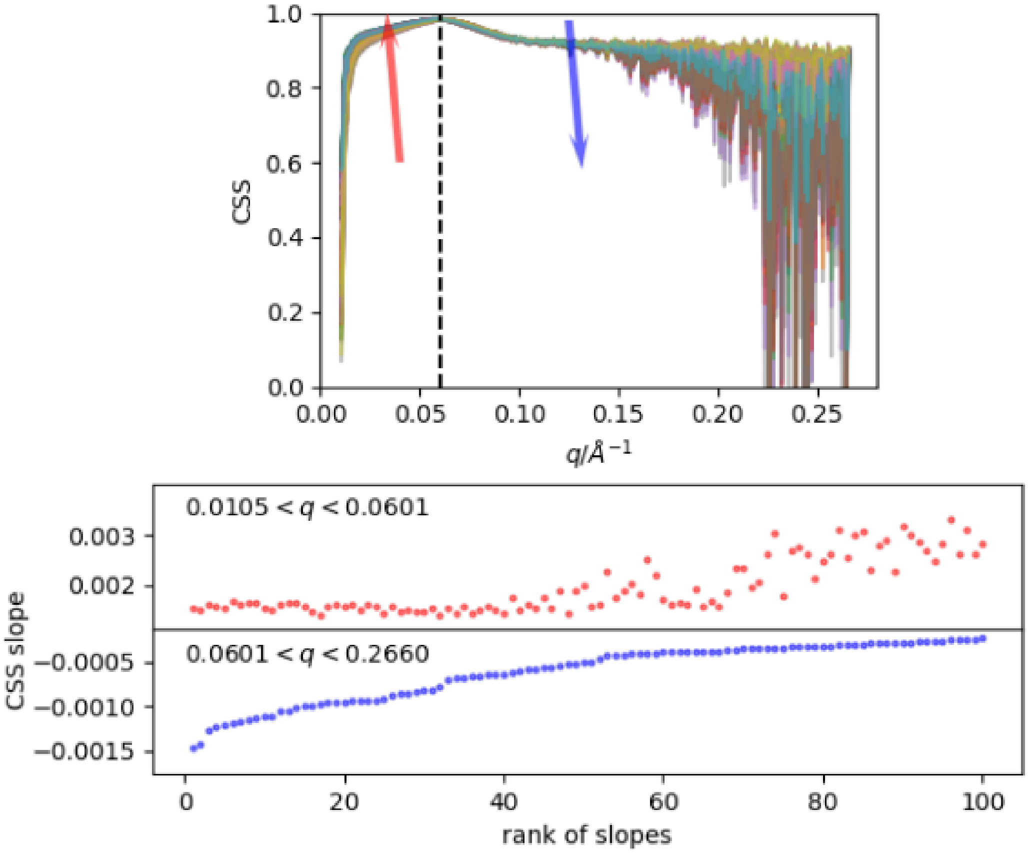


Figure 2.21 Top panel: The search (100 loops) reveals two contradictory trends before and after q = 0.06 Å^{-1}. Improving the correction at lower q undermines the quality of the correction at higher q and *vice versa*. Bottom panel: Regression coefficients for the lower q region and the higher q region ranked by the results determined by the higher q region, showing the two values are negatively correlated.

It is clear that the self-adaptive background model can identify the response of biomolecules to radiation, but the assumption that this effect is uniform across the whole q range is incorrect. Therefore, I adjusted the background model to allow the parameters of the lower q region and the higher q region to be decided separately. This requires a search for:

$$\mathbf{max}\ T(\boldsymbol{\omega},\boldsymbol{\upsilon}),\quad \begin{aligned} T(\boldsymbol{\omega},\boldsymbol{\upsilon}) = &0.35\exp(-(\bar{\boldsymbol{\omega}}-1)/(\bar{\boldsymbol{\omega}}'-1))+\\ &0.35\exp(-(\bar{\boldsymbol{\upsilon}}-1)/(\bar{\boldsymbol{\upsilon}}'-1))+\\ &0.15\exp(-b(\boldsymbol{\omega})/b(\boldsymbol{\omega}'))+\\ &0.15\exp(-b(\boldsymbol{\upsilon})/b(\boldsymbol{\upsilon}')) \end{aligned} \tag{2.38}$$

where, $\boldsymbol{\omega}$ denotes the subset of $\boldsymbol{CSS}(q)$ before q of 0.06 ($\{\boldsymbol{CSS}(q) \mid q <= 0.06\}$), $\boldsymbol{\upsilon}$ denotes the subset of $\boldsymbol{CSS}(q)$ after q of 0.06 ($\{\boldsymbol{CSS}(q) \mid q > 0.06\}$), and $\boldsymbol{\omega}$' and $\boldsymbol{\upsilon}$' are the values for the same regions derived from the standard background subtraction. Eq. 2.38 is again empirical and assumes two subsets have an equal contribution.

Monte Carlo searching with the maximum steps of 3,000 was then re-run with the variable vector drawn from the uniformly distributed searching space. The searching space and the most optimal set of parameters identified are listed in Table. 2.7.

The resulting optimised correction (Fig 2.23) is able to successfully account for the radiation-induced changes. The $\boldsymbol{CSS}(q)$ of the optimal correction shows a huge improvement against the standard background subtraction. The post processed SAXS profile obtained using this optimised correction (Fig 2.24) shows significant differences compared to the one from the standard background subtraction and clearly has better stability.

Table. 2.7 The searching space and the most optimal set of parameters of the self-adaptive search.

	searching space	**optimal parameters**
s	$\{1, 2, \ldots, 50\}$	--
$t_{\mathrm{i},\,\boldsymbol{\omega}}$	$120\leq t_{\mathrm{i},\,\boldsymbol{\omega}}\leq 220$	177
$t'_{i,\,\boldsymbol{\omega}}$	$120\leq t'_{i,\,\boldsymbol{\omega}}\leq 300$	292
$\sigma_{\boldsymbol{\omega}}$	$20\leq \sigma_{\boldsymbol{\omega}}\leq 80$	41
$\sigma'_{\boldsymbol{\omega}}$	$20\leq \sigma'_{\boldsymbol{\omega}}\leq 80$	68
$F_{\boldsymbol{\omega}}$	$\{x\text{-}5, x\text{-}4, \ldots, x, \ldots, x+4, x+5 \mid 220\leq x \leq 400\}$	$\{222, 223, \ldots, 232\}$
$D_{\boldsymbol{\omega}}$	$\{x\text{-}50, x\text{-}49, \ldots, x, \ldots, x+49, x+50 \mid 490\leq x \leq 540\}$	503
$t_{\mathrm{i},\,\boldsymbol{\upsilon}}$	$120\leq t_{\mathrm{i},\,\boldsymbol{\nu}}\leq 220$	135
$t'_{i,\,\boldsymbol{\upsilon}}$	$120\leq t'_{i,\,\boldsymbol{\nu}}\leq 300$	269
$\sigma_{\boldsymbol{\upsilon}}$	$20\leq \sigma_{\boldsymbol{\nu}}\leq 80$	31
$\sigma'_{\boldsymbol{\upsilon}}$	$20\leq \sigma'_{\boldsymbol{\nu}}\leq 80$	21
$F_{\boldsymbol{\upsilon}}$	$\{x\text{-}5, x\text{-}4, \ldots, x, \ldots, x+4, x+5 \mid 220\leq x \leq 400\}$	$\{355, 356, \ldots, 366\}$
$D_{\boldsymbol{\upsilon}}$	$\{x\text{-}50, x\text{-}49, \ldots, x, \ldots, x+49, x+50 \mid 490\leq x \leq 540\}$	504

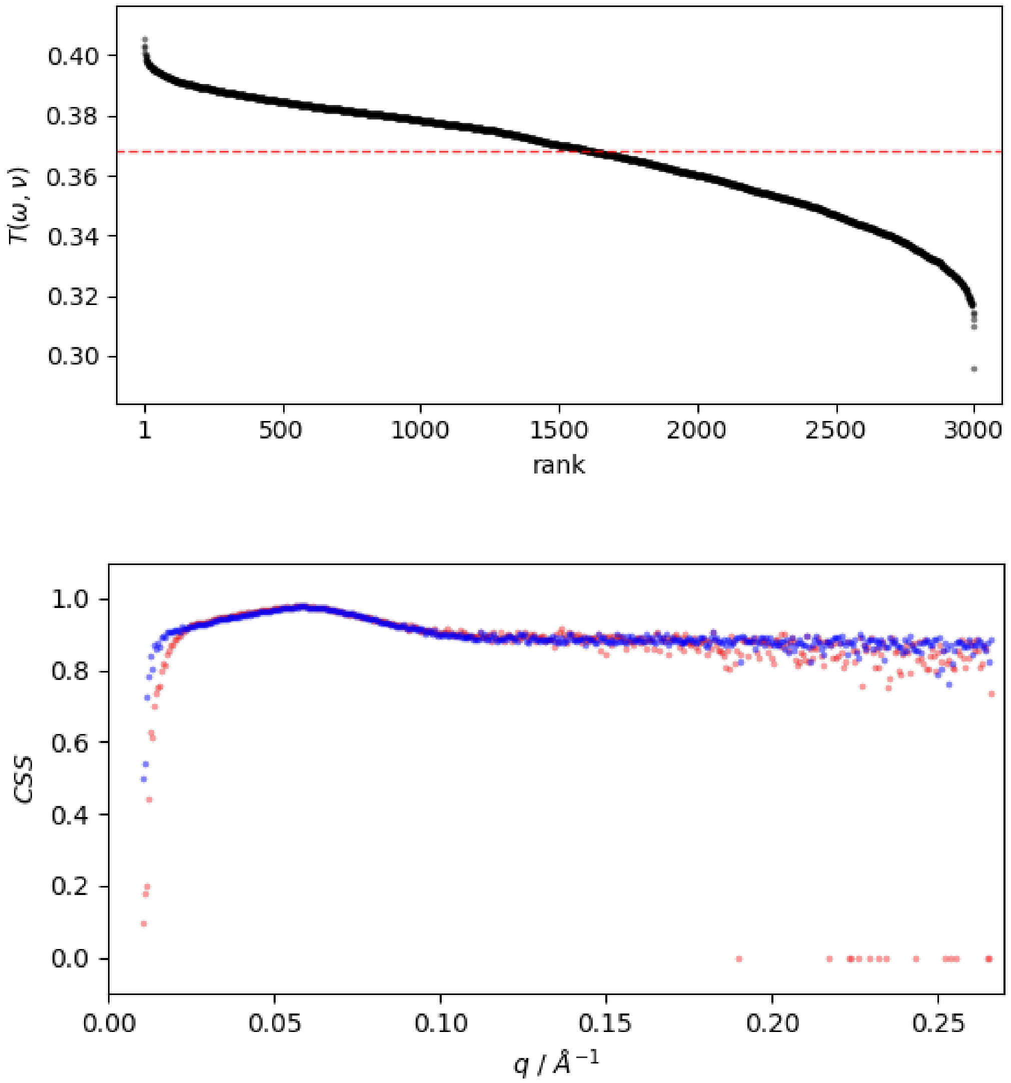


Figure 2.22 (Top) The result of the Monte Carlo search ranked by the value of $T(\boldsymbol{\omega}, \boldsymbol{\upsilon})$. The dashed red line highlights the value of $T(\boldsymbol{\omega}^{'}, \boldsymbol{\upsilon}^{'})$. (Bottom) A comparison between $\{\boldsymbol{\omega}_{\text{optimal}} \cup \boldsymbol{\upsilon}_{\text{optimal}}\}$ (blue dots) and $\{\boldsymbol{\omega}^{'} \cup \boldsymbol{\upsilon}^{'}\}$ (red dots).

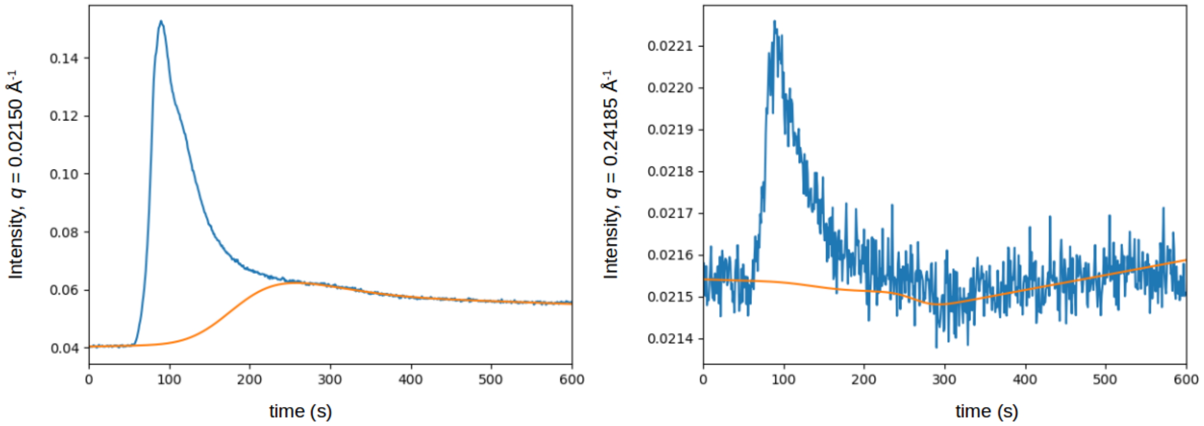


Figure 2.23 A comparison of optimal corrections for lower q region (left panel) and higher q region (right panel). Blue lines are the $I(t, q)$ of q = 0.02150 Å^{-1} (left) and of q = 0.24185 Å^{-1} (right). Orange lines are the optimal $B(t$, 0.02150) and $B(t$, 0.24185), respectively. This highlights how the responses to radiation for the two q-regions are different.

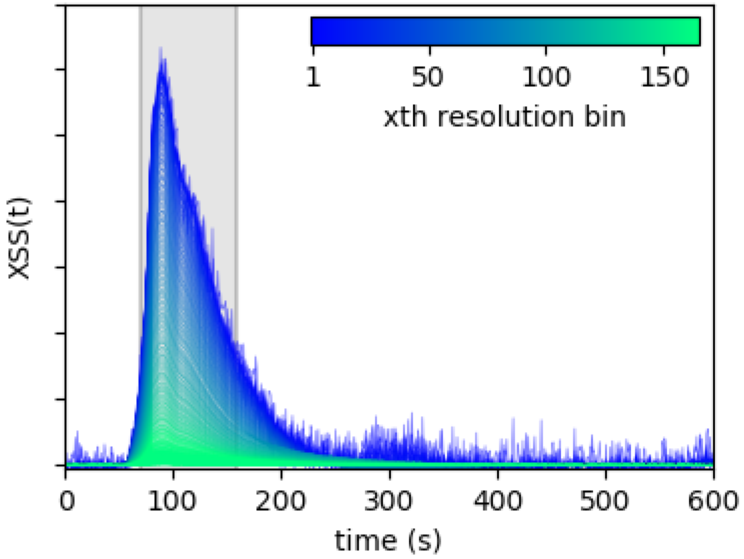


Figure 2.24 The optimally corrected *XSS* against time of mBSA. The grey shade indicates *PSR* used for calculating ***CSS***(q). TFR is from 0 to 600.

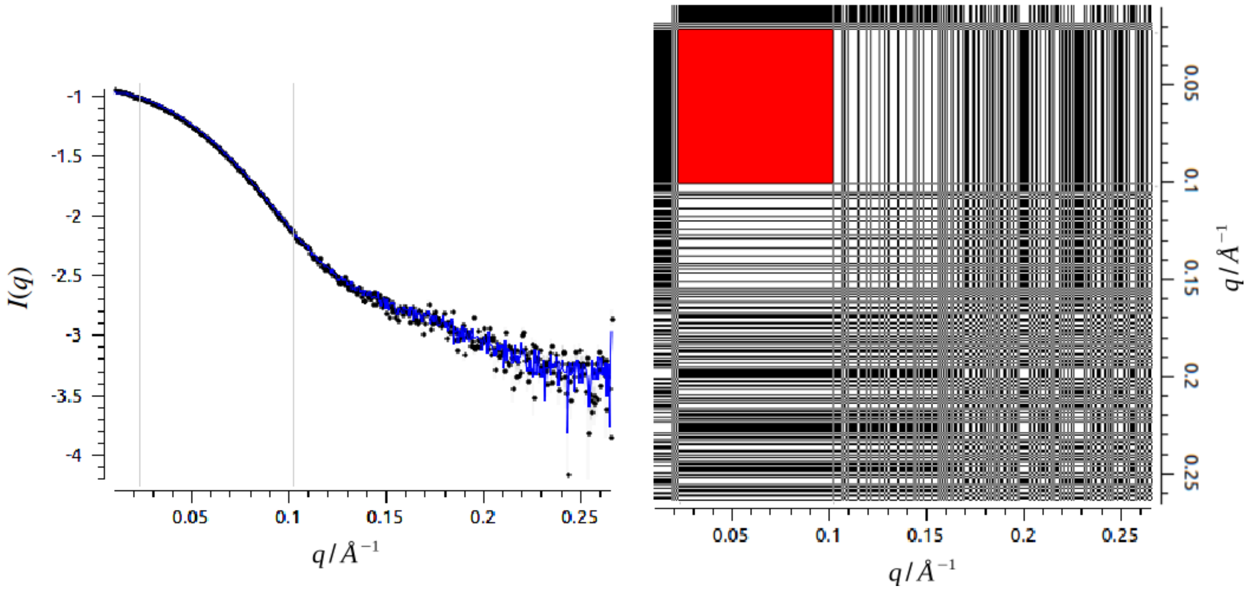


Figure 2.25 Left panel: a comparison between the SAXS profiles manifested from the standard correction (black dot) and the optimal correction (blue line). The region between two vertical lines refers to the location of the largest patch in CORMAP. Right panel: CORMAP of two corrections. The largest patch is highlighted in red.

2.3.3.3 Relationship between CSS and Reduced χ^2

The reduced χ^2 (§1.2.4) is a quantity widely used in SAXS (Franke et al. 2015). Its main purpose is to allow easy comparison of models, particularly real-space atomic models. The best model for any particular dataset is chosen as the one whose value of χ^2 is closest to one. However, the value of χ^2 is subject to uncertainty. Therefore a reduced χ^2 test is used as this better tolerates data with higher noise or experimental errors (for example arising from over-fitting). Practically in SAXS, this leads to a situation where residuals with large absolute discrepancy at lower q are acceptable whereas the ones with large relative discrepancy at higher q are ignored. The most severe impact of this is probably that individual models with the same reduced χ^2 may not be unique. In my opinion, as high-quality data are arguably decisive to generate a good model, good ***CSS***(q) statistics must lead to a successful reduced χ^2 test but the reverse is not necessarily true. The case of mBSA in the last section is a good example to test this argument as the ***CSS***(q) shows contradicting trends for the low q and high q regions.

Using CRYSOL (Svergun et al. 1995), the SAXS profiles obtained from each step of the Monte Carlo optimization were fitted by the model from SASDF99 (SASDF99_fit3_model1.pdb). This shows that $T(\boldsymbol{\omega}, \boldsymbol{\upsilon})$ is a multivalued mapping to the reduced χ^2 (Fig 2.26). Corrections with a low $T(\boldsymbol{\omega}, \boldsymbol{\upsilon})$ can still pass the reduced χ^2 test. However, as $T(\boldsymbol{\omega}, \boldsymbol{\upsilon})$ improves, the ambiguity of the reduced χ^2 decreases. At the highest values of $T(\boldsymbol{\omega}, \boldsymbol{\upsilon})$ the reduced χ^2 is also single-valued. This suggests that just relying on the reduced χ^2 hypothesis test may yield incorrect results, whereas a CSS-guided approach delivers high-quality corrected data suitable for subsequent interpretation steps.

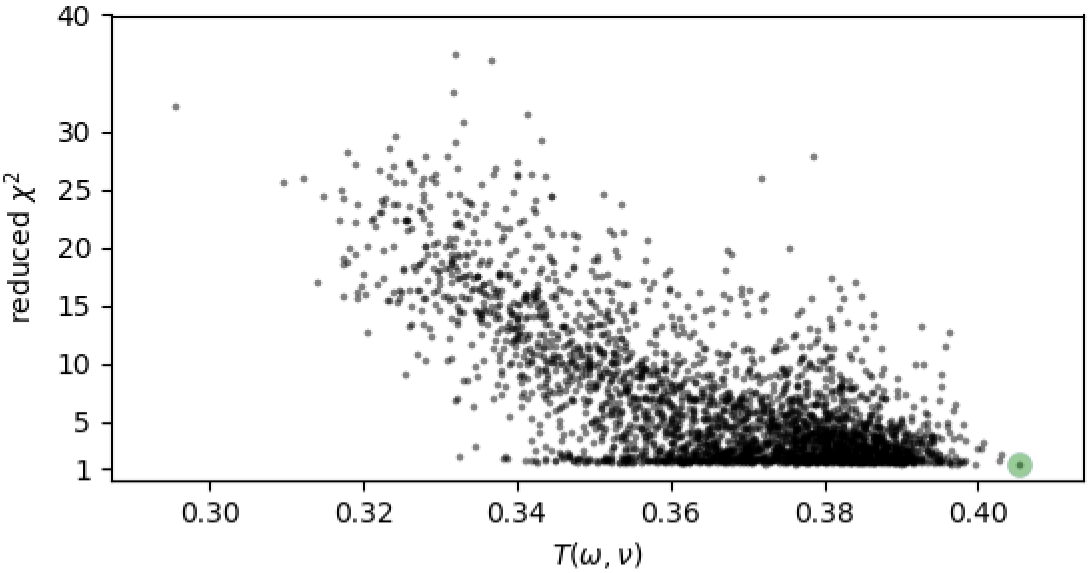


Figure 2.26 The reduced χ^2 of 3,000 non-repeating optimization steps plotted against $T(\boldsymbol{\omega}, \boldsymbol{\upsilon})$. The green dot highlights the optimal result with a $T(\boldsymbol{\omega}, \boldsymbol{\upsilon}) = 0.405$ and a reduced $\chi^2 = 1.40$.

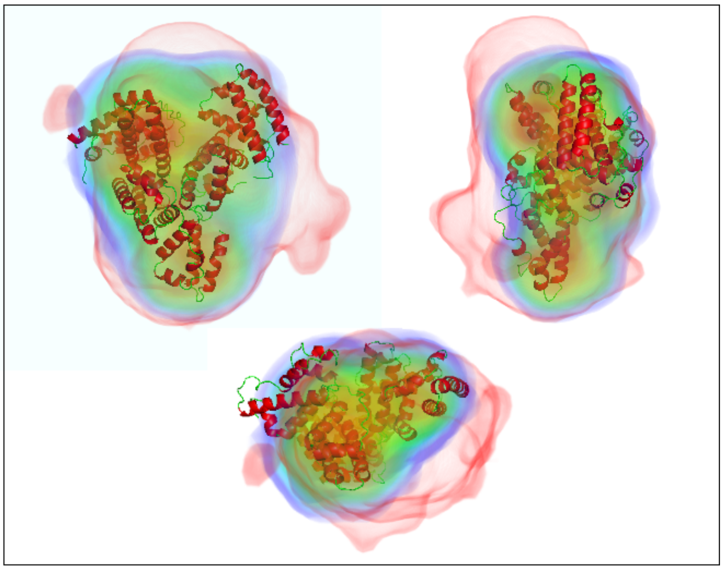


Figure 2.27 The comparison between the models derived from the standard background corrected mBSA data and optimally background corrected data using the CSS-guided self-adaptive approach. The atomic-model was refined using normal mode analysis (SASFLEX) (Panjkovich & Svergun 2016) of PDB code 4F5S against the SAXS profile obtained from the CSS-guided self-adaptive background corrected mBSA. The electron density map shown is calculated by DENSS (Grant 2018). The red envelope is from the standard corrected data. The blue-colored envelope uses the CSS-guided self-adaptive background corrected data.

2.4 Summary

Although all the operations involved in calculating ***CSS***(q) have unique solutions, it should be noted that the nature of calculating ***CSS***(q) with respect to a certain *CHROMS* support is a form of Bayesian inference. ***CSS***(q) can even be, to some extent, derived from supporting *CHROMS* which are imperfectly correlated with the *XSS*. However, any incompleteness of the supporting information will hamper the objectivity of ***CSS***(q). For example, the sample proteins may have intermolecular interactions that manifest in the *XSS* signal, but to which the *CHROMS* are blind. In principle, this problem can be solved by adding additional *CHROMS* detection modules to the experiment that are sensitive to different molecular properties. Such extra information may eventually give us a route to overcoming the resolution limits of chromatographic columns, for example when studying samples with varied conformers that are similar in hydrodynamic size.

I wish to also emphasize that CSS is not a validation method for all bio-SAXS data. This is because the so-called “batch mode”, which is the most commonly used method for synchrotron SAXS due to its speed, has no supporting *CHROMS* information. Interestingly, a recent report from the Australian Synchrotron described a modified SEC-SAXS setup that used small columns with low resolving power and a fibre optic coupled UV-spectrometer to allow rapid SEC-SAXS data acquisition with only a few minutes required per sample instead of hours (as is the case in most current SEC-SAXS experiments) (Ryan et al. 2018). Such an approach could allow the routine collection of both CHROMS and XSS signals for all bio-SAXS experiments, and enable us to use ***CSS***(q) as a general evaluation system for solution SAXS data quality.

As diverse as the observed response of biomolecules to radiation is, the perfect background models clearly vary on a case-by-case basis. Situations such as changes in incident beam flux and modulation of the HPLC pumping cycles can constantly interrupt the stability of the background. A further refinement of the self-adaptive background model to take this into account may be achievable by dynamic programming. With the recent improvements in machine learning, as well the possibility of recording *CHROMS* signals for all bio-SAXS data, CSS, as a data quality metric, has huge potential. This should allow us to not only provide better background corrected data, but also to better deconvolute complex signals, such as those from membrane proteins which so far can only be experimentally addressed by methods such as contrast mapping.

Chapter 3

Towards Better Interpretation of Data

3.1 Introduction

To interpret data is the same as asking the data questions. Consider, for example, a simple electron density map in macromolecular crystallography (F_{obs}, φ_{calc}) as an approximation of the true structure. The data are neutral. We cannot immediately make scientific sense of the electron density map unless questions are asked. By plotting a difference map ($F_{\text{obs}} - F_{\text{calc}}$, φ_{calc}), we ask the data "is there any difference between the true and the currently modeled structures". For this question, "*the parts existing in the structure, but not included in the model, should show up in the positive map contours, whereas the parts wrongly introduced into the model and absent in the true structure will be visible in negative contours*" (Wlodawer et al. 2008). Now the data has turned into information. A simple mathematical operation has allowed us to ask the data a question. In the first chapter of this thesis, I have already reviewed the information, such as the overall parameters of particles, that can be deduced from the scattering profile in a SAXS experiment. In this chapter, the focus is providing "ways of asking questions" of the SAXS data, for both kinetics and static structural studies.

3.2 Contributions to this Chapter

The time-resolved experiment and photocaging strategy in §3.3.3.1 were designed by Dr. Diana Monteiro (Universität Hamburg). Biological samples were prepared by Dr. Inokentijs Josts (Universität Hamburg). The experiment was conducted on ID09 at ESRF (Grenoble, France). The time-resolved experiment in §3.3.3.2 wes designed by Dr. Mohammad Vakili (Universität Hamburg). He also synthesized the experimental samples and performed the flow simulation. The experiment was conducted on ID02 at ESRF (Grenoble, France). The Mhp1 samples used in §3.4 were purified and solubilized in corresponding detergent solutions by Dr. Maria Koukkinidou and Diogo Melo (Universität Hamburg). The data were collected on P12 at PETRA III (DESY, Germany).

The results have been reported in the following manuscripts:

Inokentijs Josts, Stephan Niebling, Yunyun Gao, Matteo Levantino, Henning Tidow and Diana Monteiroa. Photocage-initiated time-resolved solution X-ray scattering investigation of protein dimerization. *IUCrJ*, 2018, 5, 6, 667-672

Inokentijs Josts, Yunyun Gao, Diana Monteiro, Stephan Niebling, Julius Nitsche, Katharina Veith, Tobias Gräwert, Clement Blanchet, Martin Schroer, Nils Huse, Arwen Pearson, Dmitri Svergun and Henning Tidow. Structural Kinetics of MsbA Investigated by Stopped-Flow Time-Resolved Small-Angle X-Ray Scattering. *Structure*. 2020, 28, 3, 348-354

Mohammad Vakili, Stefan Merkens, Yunyun Gao, Paul Gwozdz, Ramakrishna Vasireddi, Lewis Sharpnack, Andreas Meyer, Robert Blick and Martin Trebbin. 3D Micromachined Polyimide Mixing Devices for in situ X-ray Imaging of Solution-Based Block Copolymer Phase Transitions. *Langmuir.* 2019, 35, 32, 10435-10445

3.3. Extracting Crucial Transitions in Time-Resolved SAXS

3.3.1 Background

Understanding data can be very different depending on whether there is prior information available. Most advanced SAXS data evaluation is based on the inclusion of prior knowledge. The most valuable inclusion is probably the structural information from MX experiments. Once obtained, the theoretical solution scattering profiles can be calculated from the experimental all-atom or coarse-grained structure. The dynamics in the solution state, such as large conformational changes (Panjkovich & Svergun 2016; Sweeny et al. 2013), multi-domain proteins with extensive linkers (Bernadó et al. 2010; Tian et al. 2015), mixtures of transient or flexible complexes (Blobel et al. 2009; Møller et al. 2013), can be captured by appropriate simulations using the prior structural information.

The recent development of brilliant light sources has enabled the pump-probe solution scattering experiment to become a new tool to investigate the relationship between the dynamics of a biomolecular structure and its function (Josts et al. 2018a; Levantino et al. 2015; Thompson et al. 2019; Zaitsev-Doyle et al. 2019). Due to the relatively low SNR of time-resolved experiments, the experiment usually relates either to larger conformational changes (Chen & Hub 2014; Ramachandran et al. 2011) or to distinguishing allosteric changes where there are static structures of the metastable intermediate states one can refer to (Cammarata et al. 2008). However, determining the subtle transitions of intermediates with unknown structures is non-trivial as the system has to be perturbed from its non-equilibrium state and followed in real time. Moreover, in practice, “on-the-fly” decisions must be made during the experiments considering that beam time is not unlimited. For example, deciding the necessity of testing an extra time point in order to validate/invalidate the identification of a transient state. These limits are in conflict with the fact that more sophisticated analysis is required to unveil a plausible kinetic model.

To alleviate this problem, I have proposed a semi-quantitative model-free method. With this, it is possible to rapidly locate the potential transition points and the time scales of kinetic changes.

3.3.2 the Algorithm

3.3.2.1 Singular Value Decomposition (SVD)

In the field of real numbers (ℝ), the singular value decomposition of a matrix ***A*** is the factorization of ***A*** into the product of three matrices (Press et al. 2007)

$$\boldsymbol{A} = \boldsymbol{USV}^{\mathrm{T}} \tag{3.1}$$

where the columns of ***U*** and ***V*** are orthonormal and the matrix ***S*** is diagonal with positive real entries. The superscript T indicates the matrix transpose.

A non-negative real number σ is a singular value for ***A*** if and only if there exist unit-length vectors ***u*** in R^m and ***v*** in R^n such that

$$\boldsymbol{Av} = \sigma\boldsymbol{v} \text{ and } \boldsymbol{A}^{\mathrm{T}}\boldsymbol{u} = \sigma\boldsymbol{u} \tag{3.2}$$

The vector ***u*** and ***v*** are the left-singular and right-singular vectors for σ respectively. The diagonal entries of ***S*** are equal to the singular values of ***A***. The first $k = \min(m, n)$ columns of ***U*** and ***V*** are the left- and right-singular vectors for the corresponding singular values.

Properties of SVD matrices:

(a) An m-by-n matrix ***A*** has at most k distinct singular values.

(b) It is always possible to find a unitary basis ***U*** for R^m with a subset of basis vectors spanning the left-singular vectors of each singular value of ***A***.

(c) It is always possible to find a unitary basis ***V*** for R^{n} with a subset of basis vectors spanning the right-singular vectors of each singular value of ***A***.

(d) Non-degenerate singular values always have unique left- and right-singular vectors, up to multiplication by a sign. Consequently, the singular value decomposition is unique, if all singular values of ***A*** are non-degenerate and non-zero, up to arbitrary unitary transformations on vectors of ***U*** and ***V***.

3.3.2.2 Cumulative First-Ranked Singular-Values Correlation Map (CSV-CORMAP)

Let the m-by-n matrix $\boldsymbol{X}_{m\times n}$ be the absolute scattering intensities of all the time points.

$$\boldsymbol{X}_{m\times n} = [I_{\mathrm{m}}(q_{\mathrm{n}})]\ n \in [1, N], m \in [1, M] \tag{3.3}$$

where N is the number of resolution bins and M is the number of experimental points.

Let us construct a M-by-M squared matrix $\mathbb{C}$ where for each element in $\mathbb{C}$, we have

$$c_{i\rightarrow j} = \frac{\sigma_{1,i\rightarrow j}}{\sum_{r} \sigma_{r,i\rightarrow j}} \tag{3.4}$$

where $\sigma_{r,i\rightarrow j}$ is the r-ranked singular value of the mean centering matrix of the scattering profiles (I_i, …, I_j) at the consecutive time delays.

According to (a), $r \in [1, k]$.

According to (b), each $c_{i,j}$ is uniquely decided if $\sigma_{r,i\rightarrow j} \neq 0$ for all $r \in [1, k]$.

A figurative demonstration of $\mathbb{C}$ is a CSV-CORMAP. $c_{i\rightarrow j}$ is the weight of the first-ranked singular value. It can be considered as a scale of how significant the (dis-) similarity is between all the time points from i to j.

Denotations:

(a) =: means that the left term is defined as the right term.

(b) $\boldsymbol{J}_M$ is an all-*ones* vector with a size of M. M is the total number of time delays. N is the size of observations for each time delay.

(c) $\boldsymbol{X}[n]$ means taking the n-th element from vector $\boldsymbol{X}$. $\boldsymbol{X}[n{:}m]$ means taking the n-th to m-th elements from vector $\boldsymbol{X}$.

```
algorithm csv_cormap{
    # constructing $X_{m\times n}$
    $X_{m\times n}$ :: empty matrix with a size of $M \times N$
    for $t$ in (range 1 to $M$){
        I_t = read(absolute scattering intensity of the time point t)
        vector $X[t]$ =: I_t
    }
    # end of constructing $X_{m\times n}$
    # constructing $\mathbb{C}$
    $centered\_X_{m\times n} = X_{m\times n} - J_M * (\sum_{i}^{M} X_{i,j}/M)^{T}$
    $\mathbb{C}$ :: empty matrix with a size of $M \times M$
    for i in (range 1 to $M$){
        for $j$ in (range i to $M$){
            vector $\sigma\_ij$ =: singular_values($centered\_X_{m\times n}$[i:j])
            if (all $\sigma$_ij ≠ 0){
                c_ij = $\sigma$_ij[1] / sum($\sigma$_ij)
                $\mathbb{C}$[i, j] =: c_ij
                $\mathbb{C}$[j, i] =: c_ij
            }
        }
    }
    # end of constructing $\mathbb{C}$
    # plotting $\mathbb{C}$
    normalized_colormap =: colormap(min=min($\mathbb{C}$), max=max($\mathbb{C}$))
    display 2d_image(2d_array=, colormap=normalized_colormap)
    # end of plotting $\mathbb{C}$
}
```

3.3.3 Case Studies

3.3.3.1 Photocage-Initiated Protein Dimerization

Photocage-initiated time-resolved X-ray scattering is a new experimental approach that allows tracking of time-resolved structural transitions initiated by small-molecule binding using X-ray solution scattering. The ligand is rapidly released into the protein solution by photocleavage and scattering profiles are collected at appropriate time delays after decaging. Photodecaging of small-molecule ligands offers a path to the observation of single turnover events of proteins that are not naturally photoactivatable, opening up many possibilities for studying reaction kinetics and structural dynamics in a much wider range of biological systems.

Adenosine triphosphate (ATP) hydrolysis drives numerous enzymatic reactions and biological processes from the translocation of substances across cell membranes, signaling cascades, protein folding and chaperoning to protein degradation. These enzymes, termed ATPases, utilize the binding and breakdown of ATP as a source of energy to undergo changes in their tertiary or quaternary structure during their functional cycle. Caged ATP has previously been used to study the mechanism of protein function (Clapp & Gurney 1992).

In this study, laser-flash photolysis of 1-(2-nitrophenyl)-caged ATP (NPE-ATP) and time-resolved X-ray-scattering were combined to investigate the ATP-dependent dimerization of soluble nucleotide-binding domains (NBDs) from a bacterial lipid flippase, MsbA. These domains are key drivers for the “power stroke” mechanism of substrate translocation in the ATP-binding cassette-transporter family of proteins. The binding of ATP to each of the NBDs induces conformational changes in each domain with the subsequent formation of a closed dimer, while further allosteric changes in the transmembrane domain of the transporter results in the translocation of solutes across the membrane through the alternating access mechanism (Ward et al. 2007) (Fig. 3.1). Time-resolved Fourier transform infrared spectroscopy was previously employed to indirectly investigate the dimerization kinetics of the soluble NBDs from MsbA by ATP-decaging (Syberg et al. 2012).

Here, we used the strategy of ATP-decaging in combination with X-ray scattering to gain time-resolved structural insight into the dimerization of NBDs, using the experimental setup outlined in Fig. 3.1. To follow the dimerization of NBDs using X-ray scattering by decaging NPE-ATP, we used a 5 ns 355 nm laser pulse which overlapped with the X-ray pulses in the sample capillary. The achievable time-resolution of the experiment was limited by the decaging time of NPE-ATP (~10 ms) following flash photolysis. A slow-flowing liquid-delivery system was employed to provide automatic refreshment of the sample between each pump–probe pulse, while keeping the laser-illuminated volume in the X-ray path.

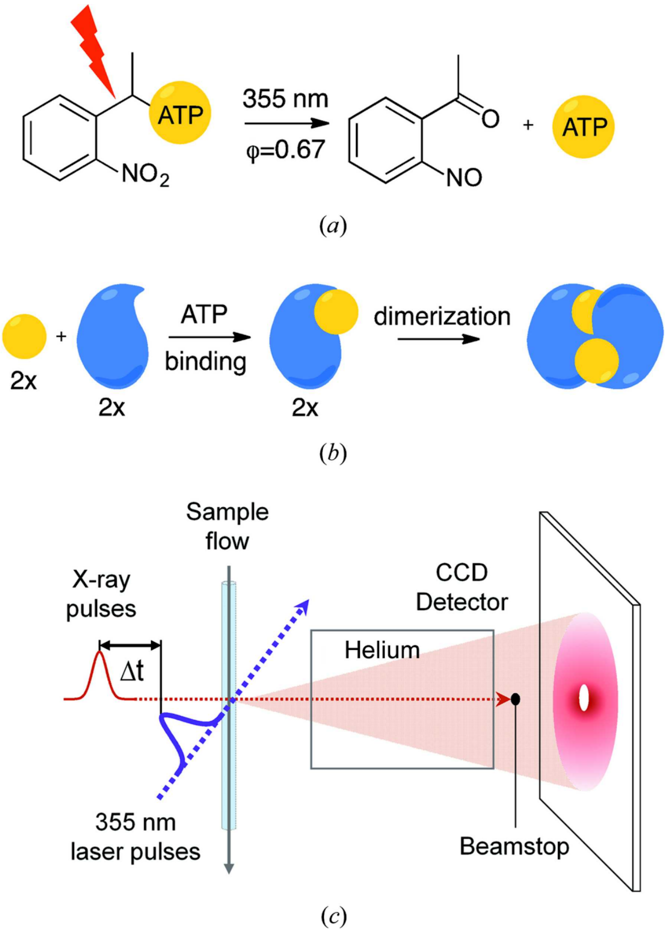


Figure 3.1 Overview of the experimental setup and protein structural transitions.
(a) Photoexcited decaging of the NPE group from NPE-ATP using laser irradiation at 355 nm releases free ATP into solution. The decaging reaction happens on a 10 ms timescale. The quantum yield of the reaction is 0.67.
(b) Mechanism of ATP-dependent (yellow circle) dimerization of NBDs (blue).
(c) Schematic diagram of the experimental set up at the beamline.

Figure adapted from (Josts et al. 2018b).

The experiment was carried out at beamline ID09 at the European Synchrotron Radiation Facility (ESRF, Grenoble, France). ID09 provides polychromatic X-rays centered at 15 keV with a 3% bandpass in a 100 × 60 µm (H × W) full width at half-maximum (FWHM) focused beam at the sample position. A Rayonix MX170-HS detector was used at a distance of 300 mm to the sample. A helium cone with a 1.6 mm diameter beamstop was placed between the sample and the detector. The sample-to-beamstop distance was 210 mm. For this experiment, 15 µs X-ray pulses were used. A 0.5 mM protein solution with 1.5 mM NPE-caged ATP {adenosine, 5′-(tetrahydrogen triphosphate), P″-[1-(2-nitrophenyl)ethyl] ester disodium salt, Jena Bioscience} was flowed continuously through a 1 mm diameter quartz capillary using a Miniplus 3 peristaltic pump (Gilson). The sample was photolyzed using a 355 nm ns laser (Vibrant from Opotek, Carlsbad, CA, USA, sold by Quantel, Les Ulis, France) with 4 mJ per 5 ns pulse and a 1.7 × 0.2 mm (H × W) FWHM. The relative timing of the laser and X-ray pulse was controlled electronically (laser–X-ray jitter < 5 ps). At this wavelength, NPE-ATP has an extinction coefficient of 430 M^{-1} cm^{-1}. The low extinction coefficient allows for uniform penetration of the laser through the sample. The laser power was calculated to deliver ∼4 photons per caged ATP. NPE-ATP has a quantum yield of 0.6 at a photolysis wavelength of 360 nm, yielding a final concentration of ∼0.95 mM ATP post-photolysis.

At t_0, the laser pulse for photoexcitation of the sample was delivered and after the desired time delay, the photoexcited sample volume was probed by the X-ray pulse. The maximum possible time delay was calculated from the flow velocity and the area of overlap between the laser and the X-ray pulses, to guarantee that only photoexcited samples were measured. After each X-ray probe pulse, the sample flowed continuously to ensure full refreshment of the laser-interaction volume. Depending on the flow-rate, this step determined the maximum repetition rate of the measurement. The laser-pump, X-ray-probe and sample refresh cycle was repeated multiple times and multiple X-ray pulses (20–100) were integrated onto the detector before readout to increase the signal-to-noise ratio. After the detector readout, the time delay was changed and the cycle repeated. Multiple images were collected for each time delay and merged during data processing to further increase the signal-to-noise ratio. Non-photoexcited data were collected with a negative time delay between the pump-laser and the X-ray probe pulse (-100 µs). Photoexcited and non-photoexcited images were interleaved to keep the experimental conditions between dark and light data sets as similar as possible. The detailed experimental parameters are presented in Table 3.1 and Figure 3.2.

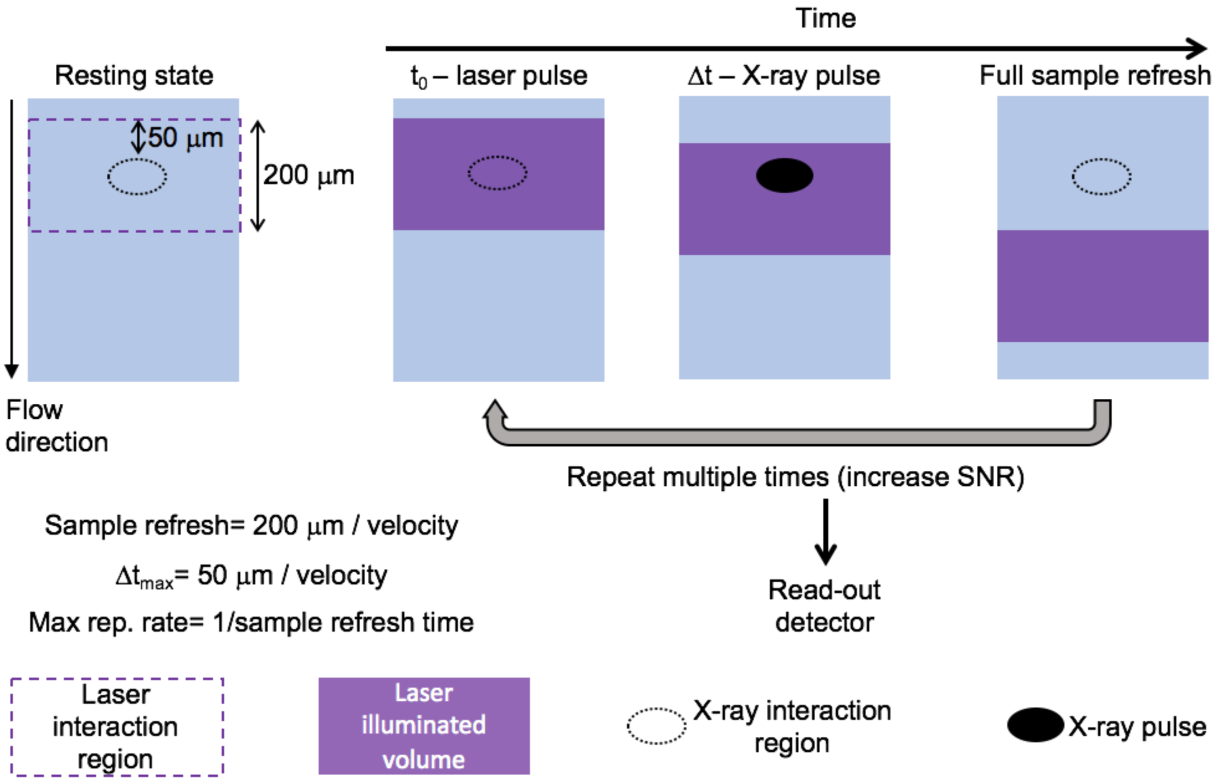


Figure 3.2 Flow scheme for illumination, probing and refreshing of the sample. The X-ray and laser interaction regions are centered and aligned to the capillary. The resting state shows the overlap of the pump and the probe. The X-ray spot is 0.10 x 0.06 mm^2 and the laser 1.7 x 0.2 mm^2 (HxV, FWHM). At t0, the laser pulse illuminates the sample. After a specified time-delay, the X-ray pulse hits the activated sample volume. The sample is allowed to flow for a full 200 micrometers before the measurement is repeated to ensure full sample refresh. The pump-probe-refresh cycle is repeated multiple times, accumulating several X-ray pulses on the detector to increase the signal- to-noise ratio before the detector is read out. After the image is read out, a new time delay is set and the cycle repeated. The sample was allowed to flow a maximum of 50 micrometers before the X-ray probe pulse arrived to ensure the X-rays probe a laser-illuminated sample volume. Figure adapted from (Josts et al. 2018b, SI).

The normalization factor for each exposure was calculated using the sector integration of the total counts within the radial range of 74.0–83.4 pixels. Appropriate normalized buffer profiles were subtracted. The background-corrected images were then azimuthally averaged. The resulting profiles for each time delay were averaged to acquire a sufficient signal-to-noise ratio.

Table 3.1 The experimental details of illumination, probing and image collection of sample

dataset	**time delay (Δt, ms)**	**flow rate (mm/s)**	**pump-probe pulses**	**merged images**	**data acquisition rate (Hz)**
1	-0.1, 600, 800, 1000	0.02	20	23	0.1
2	-0.1, 50, 100, 200	0.20	100	25	1.0
3	-0.1, 300, 400, 500	0.10	80	20	0.5
4	-0.1, 250	0.10	40	20	0.5
5	-0.1, 900, 1200, 1400	0.02	20	8-9	0.1

A CSV-CORMAP was plotted using the absolute scattering profiles from the dark state and the 13 time delays (Fig. 3.3 center). Without any assumptions, CSV-CORMAP shows a continuous transition, and a similar trend can be seen in the intensity difference (Fig. 3.3 a). Intensity difference is calculated by subtracting the scattering of the dark state from all following time delays. Although a clear increase in the forward scattering intensity is easily noticeable, it is hard to quantitatively assign any subtle transitions. From CSV-CORMAP, a fast transition in the first 100 ms is revealed as $c_{0\text{->}50}$ and $c_{0\text{->}100}$ both drop by half compared to $c_{0\text{->}0}$ and there is decay over $c_{50\text{->}100}$. After 200 ms, a steady evolution is visible between 200 and 1000 ms time delays. Within this region, two significant time points are observed: a global minimum at $c_{150\text{->}800}$ and neighboring minimum at $c_{300\text{->}400}$. The global minimum is the time point where the major conformational components change most, and corresponds to the appearance or disappearance of certain conformational species. The neighboring minimum implies the time point where the most rapid changes happen. The final stage of structural change is noticeable from 1200 ms to 1400 ms by the dissimilarity of $c_{1200\text{->}1400}$ to the former time points. This might indicate the conformational evolution has significantly slowed down by this point.

Further model-/hypothesis-based studies strongly correlate with the conclusions drawn from CSV-CORMAP (Fig 2.3 b-f). The initial fast transition can be assigned to conformational changes introduced by the nucleotide binding event. It takes roughly 100-200 ms to yield a dimerization-capable state, based on a non-decay pattern for the first component and the second component in the factor analysis. This is also consistent with the unchanged R_g. The neighboring minimum in CSV-CORMAP coincides with the crosspoint of the evolution trends of two major factors, as well as the steepest ascent in the second-order kinetic fitting.

The global minimum matches the first time point where the contribution of the second factor becomes negative. The *ab initio* modelling of the data at 1400 ms, the factor analysis, and the kinetic fitting all confirm that by this point dimerization has reached its maximum.

The ability to track the dimerization of the NBDs in real-time, by combining the photoexcited decaging of ATP and time resolved solution scattering, offers new possibilities for the investigation of structural dynamics for numerous ATP-dependent reactions, and potentially in any photo-decageable system. This ability to study non-equilibrium and unidirectional protein function coupled to structure opens up new possibilities to obtain invaluable insight into molecular structure–function relationships. Several parameters must be taken into consideration for a successful time resolved solution scattering experiment in order to overcome the technical difficulties associated with conducting these experiments. Most importantly, the signal-to-noise ratio and the magnitude of the difference signal in the *q* range expected for the structural change under investigation has to be detectable in order for a clear structural and kinetic model to be extracted. For example, for large proteins and protein complexes undergoing local structural perturbations, the difference in signal would be small compared with the overall scattering power of the protein. CSV-CORMAP paves the way for extracting crucial transitions within a dataset containing subtle changes without any assumptions. This method allows a general interpretation based on first-hand data and facilitates a quick on-site interaction with the information obtained in the experiment.

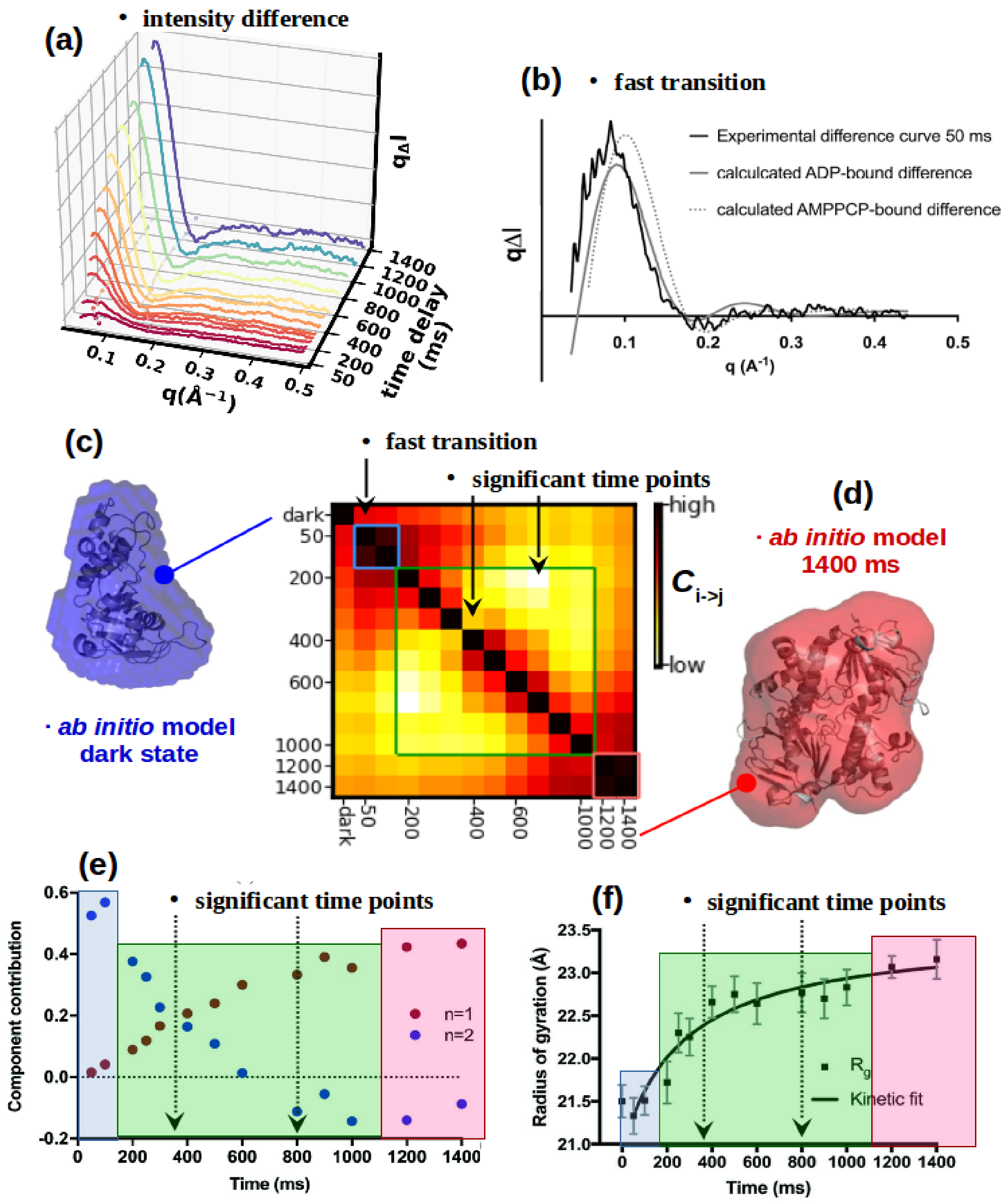


Figure 3.3 (Center) CSV-CORMAP based on the absolute scattering data. Three zones relating to different stages of conformational changes of NBD are distinguished. (a) Plot of the scattering difference curves ($q\Delta I$ versus q) calculated by subtracting the protein scattered with a negative laser offset (-100 µs) from all the subsequent time point measurements. (b) Intensity difference at the fast transition point marked in the CSV-CORMAP. A comparison is made between experimental difference curves and theoretical difference curves. Theoretical scattering difference from crystal structures of the NBD monomer bound to ADP (PDB: 5DGX), non-hydrolysable ATP analogue (AMPPCP, PDB: 3B60) has a good agreement with the 50 ms time-delay difference curve. (c) Ab initio models of the dark state superimposed with monomeric (PDB: 5IDV). (d) Ab initio models of the analysis of the TR-XSS data reveals the presence of two components. (e) Factor analysis assuming only two major factors exists. It shows the relative amplitudes of the eigenvectors (component 1 and 2 in red and

blue, respectively) versus time. The blue, green, red shades highlight the three zones distinguished in CSV-CORMAP. Two significant time points marked by dashed arrows correspond to the neighboring minimum and the global minimum, respectively, in CSV-CORMAP. (f) Radius of gyration representing NBD dimerization and kinetic fit with an increase in R_g over time (black squares with corresponding standard deviations). The time points after laser excitation were fitted with a second-order kinetic function (black line). Shades and dashed arrows represent the same concepts as ones in (e).

3.3.3.2 Flow-Mixing Induced Biocompatible Block Copolymer Phase Transitions

Time-resolved solution SAXS, combined with a novel 3D flow-focusing microfluidic platform, was used to investigate the amphiphilic diblock copolymer phase transitions from relaxed dissolved polymer chains to spherical micellar assemblies. Such block polymer nanostructures belong to the most widely studied systems in soft matter sciences (Narayanan et al. 2017) and offer great possibilities in pharmaceutical applications due to their drug encapsulation properties (Mura et al. 2013). Therefore, knowledge of their phase transitions is essential for tailoring efficient drug encapsulation vehicles.

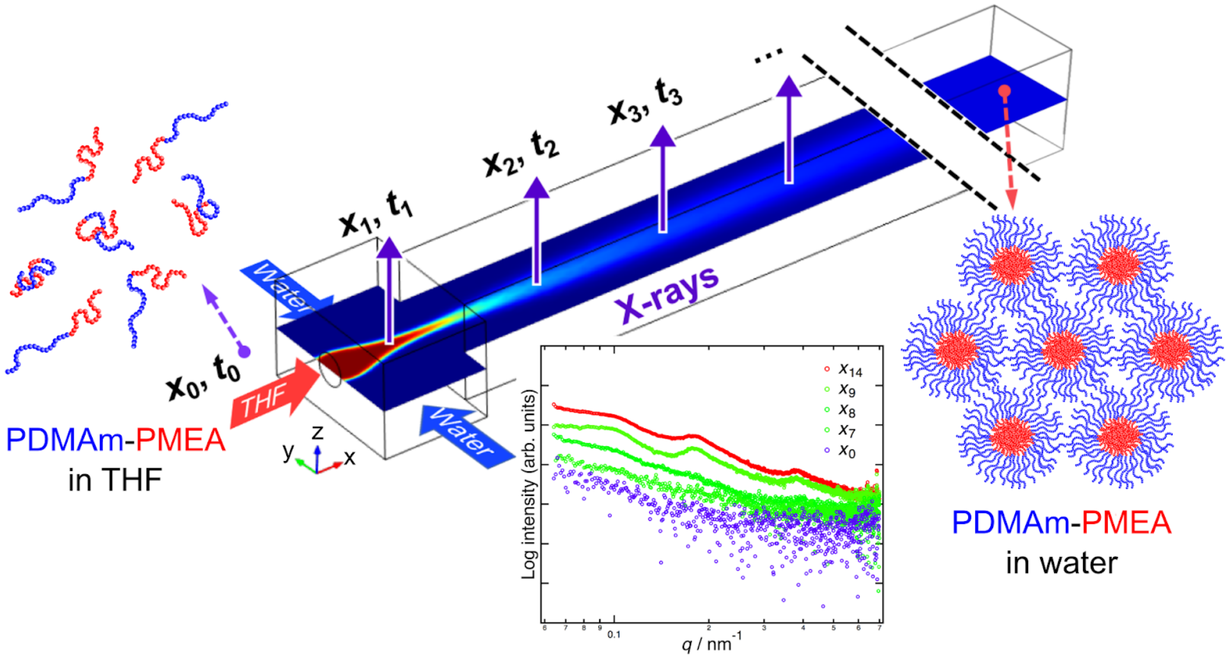


Figure 3.4 Time resolved SAXS, combined with a novel 3D flow-focused X-ray compatible microfluidic device, was used to study rapid mixing-induced self-assembly of diblock copolymers PDMAm-PMEA. The incoming focused stream has initially no wall contact, efficiently preventing radiation induced wall agglomeration and enabling a more uniform mixing. The time resolution is defined by the positional interval (*x*0, *x*1...) within the flow channel. Figure adapted from (Vakili et al. 2019).

Here, as a self-assembly system, biocompatible poly(N,N-dimethylacrylamide)−poly(2-metho-xyethyl acrylate) (PDMAm−PMEA) diblock copolymers were investigated. The block copolymers can self-assemble into self-stabilized nano-objects such as spherical micelles (Charleux et al. 2012). The polymerization proceeds from a monomer-in-water dispersion where the monomer (MEA) is initially water-soluble but becomes water-insoluble upon chain growth. The block length of the hydrophilic block (PDMAm) and the hydrophobic block (PMEA) were targeted to be of the same magnitude to facilitate the formation of spherical micelles in water.

In order to study the self-assembly dynamics, the finished and purified diblock copolymers were dissolved in THF and flow focused with water (a nonsolvent for poly(MEA)) to trigger the formation of micelles. For the first time, with the use of a full 3D flow-focusing microfluidic device and *in situ* X-ray probes, we were able to monitor the entire phase transition process in a true focusing flow (Fig 3.4).

The diblock polymer has repeating units of 48 and 61 for PDMAm and PMEA, respectively. The molecular weight (Mn) is 17 kg mol^{-1} with a dispersity (Mw/Mn) of 1.4. Before injection, the polymerized mixture was freeze-dried from its native aqueous medium and subsequently dissolved in THF to give a 10% *w/w* solution. The SAXS patterns were collected at beamline ID02 at ESRF42 (Grenoble, France) using monochromatic X-ray radiation with a wavelength of $\lambda = 0.10$ nm (12.46 keV) and q ranging from 0.01 to 0.75 nm^{-1} , where $q = 4\pi/\lambda \sin(\theta)$ is the length of the scattering vector and θ is the half-scattering angle. The beam diameter was adjusted to 72.4 μm in the horizontal (x) direction and 42.3 μm in the vertical (y) direction (FWHM at the sample). Assuming a Gaussian distribution, the portion of the beam that is hitting outside the channel can be estimated. When the channel is centered, this is about 0.3%, but closer to the edge there is more beam overlap with the sample cell. The beamstop diameter was 2 mm. A 2D Rayonix MX-170HS detector with a pixel size of 44 × 44 μm^2 was used, which was housed in an evacuated flight tube at a sample-to-detector distance of 10 m.

The microfluidic channel geometry and the beam size of the microfocused X-rays determined the limiting time scales of the flow experiment which are the time resolution (t_{res}) and the maximum residence time (t_{max}). t_{res} is given by the X-ray beam size in horizontal direction, $\Delta x_{\mathrm{beamsize}} = 72.4$ μm, and the average flow velocity of the fluid stream v by

$$\mathrm{t}_{\mathrm{res}} = \Delta x_{\mathrm{beamsize}}/v \tag{3.5}$$

With a measured channel width w_{c} = 120 μm, a channel height h_{c} = 125 μm, and the sum of the flow rates Q = 1100 μL/h, one obtains an average flow velocity of v

$$v = Q/(w_{\mathrm{c}} \times h_{\mathrm{c}}) = 20.4 \text{ mm s}^{-1} \tag{3.6}$$

and therefore a t_{res} of 3.5 ms. The maximum residence time $t_{\max}$ is given by the channel length l_{c} (47.6 mm) divided by the flow velocity,

$$t_{\max} = l_{\mathrm{c}}/v = 2.3 \text{ s} \tag{3.7}$$

The downstream flow direction is referred to as x, while the channel width ranges over y (perpendicular to the flow direction).

SAXS patterns of the sample and background were collected in the microchannel under continuous-flow conditions with an exposure time of 1 s. They were normalized by incident flux and transmission before the sample image was subtracted with its corresponding background pattern. The background-corrected SAXS patterns were recorded at 15 downstream positions x_0–x_{14} (Fig 3.5 and Table 3.2). All 1D profiles were background corrected and azimuthally-averaged using a custom pyFAI code.

CSV-CORMAP was plotted using the invariant-normalized 1D profiles of the 15 downstream positions (Fig. 3.6 center). The structural evolution shows three distinct regions. The indicated green region is dominated by the colors in the upper range of the color scale (black to dark red), which implies the major structural components undergo no significant changes either in composition or in the amount of each component. The global minimum is clearly located at $c_{X8\text{->}X9}$, where the neighboring minimum also occurs. This sudden drop strongly suggests the incidence of a fast structural transition. This structural evolution lasts until x_{12} as the remarkable restoration of similarity is evident in the whole column of $c_{x\mathrm{n}\text{->}x13}$. Although the neighboring similarity increases in the last region, it is hard to conclude whether the structure of particles has stabilized, due to the fact that moderate dissimilarity can be found for every set of three time points.

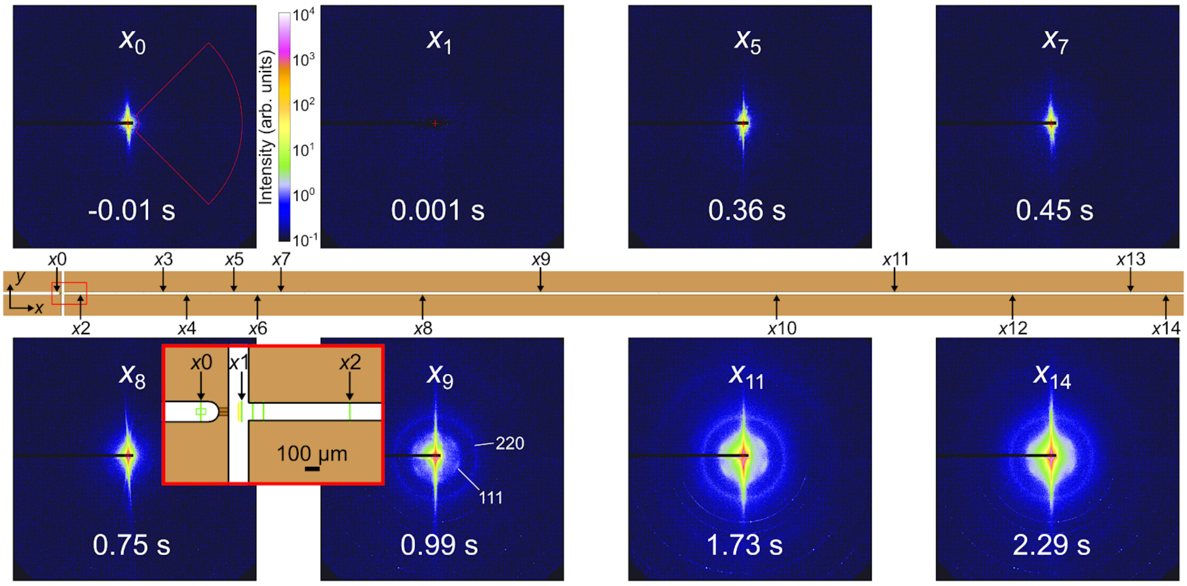


Figure 3.5 Overview of the scan positions along the microchannel with indication of the corresponding SAXS patterns. The patterns show the structural evolution along the *x* direction (flow direction) at the center of the channel. Position x_0 is located before the main channel inlet pore, and the image shows the 10% *w/w* polymer in THF solution. The subsequent images with increasing position number are recorded at increasing distance from the mixing cross, showing the transition from disorder to order. The distance of the detection points (x_1–x_{14}) from the starting point x_0 is given in Table 3.2. The dimensions of the X-ray beam are indicated by the small green rectangle in the enlarged depiction of the mixing cross. The patterns suffer from a symmetrical streak (resulting from grazing-incidence reflection at the channel walls in combination with beamstop scattering), which has been masked out for further data processing by integrating a custom region of interest, as indicated by the red area in the top left detector image (x_0). Figure adapted from (Vakili et al. 2019).

Table 3.2 Overview over the scan positions *xn* with the distance *x* (mm) from the middle of the mixing cross ($x = 0$, $y = 0$) and the corresponding average flow time *t* (s).

xn	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
Δ*x* (mm)	-0.25	0.02	0.75	4.25	5.25	7.25	8.25	9.25	15.25	20.25	30.25	35.25	40.25	45.25	46.25
Δ*t* (s)	-0.01	0.001	0.04	0.21	0.26	0.36	0.40	0.45	0.75	0.99	1.45	1.73	1.98	2.22	2.29
cTHF†	100	97.7	12.0	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1

† wt%, simulated using the finite-element method in COMSOL Multiphysics

With the help of CSV-CORMAP, we are able to discuss the data before and after x_8/x_9 from different theoretical aspects (Fig 3.6 and Table 3.3). For the early dilution stage, the particles are treated as non-interactive flexible random coils. After the transition point, the particles are considered as semi-crystalline micelles. The scattering profiles of the early stage have no long-term order yet are also not fully consistent with the expected scattering of Gaussian-chain-like polymers. The *R*g value of the flowing diblock copolymer at x_0 is 18.4 nm which is significantly larger than 4.95 nm, the *R*g of the flexible random coil in unperturbed THF. This, as well as the less steep Porod region, indicates a flow-induced anisotropic chain conformation (Svergun et al. 2013). The solvent dilution only induces the onset of phase separation while no significant number of micelles have formed yet. From x_9 onward, pronounced Bragg reflections and Debye-Scherrer rings appear, which indicates the solution-mediated chain interaction is strong enough to form long-term order. The Bragg reflection has a 6-fold rotational symmetry. This is in agreement with the known shear orientation of FCC lattices that orient in a way that the [110] direction, i.e., the line of highest micellar density, is parallel to the flow direction (Förster et al. 2007). Further downstream, a sharpening of the peaks with increased scattering intensity occurs, indicating shear-induced alignment and the formation of a lyotropic liquid-crystalline micellar phase. In the last stage, the size of the micelles becomes gradually smaller along the flow, possibly indicating a deswelling of the micelles, as evidenced by the fitting results.We speculate that a competition between micelle formation and its deswelling might already start around x_{11}/x_{12} but the signal is eventually dominated by the latter process.

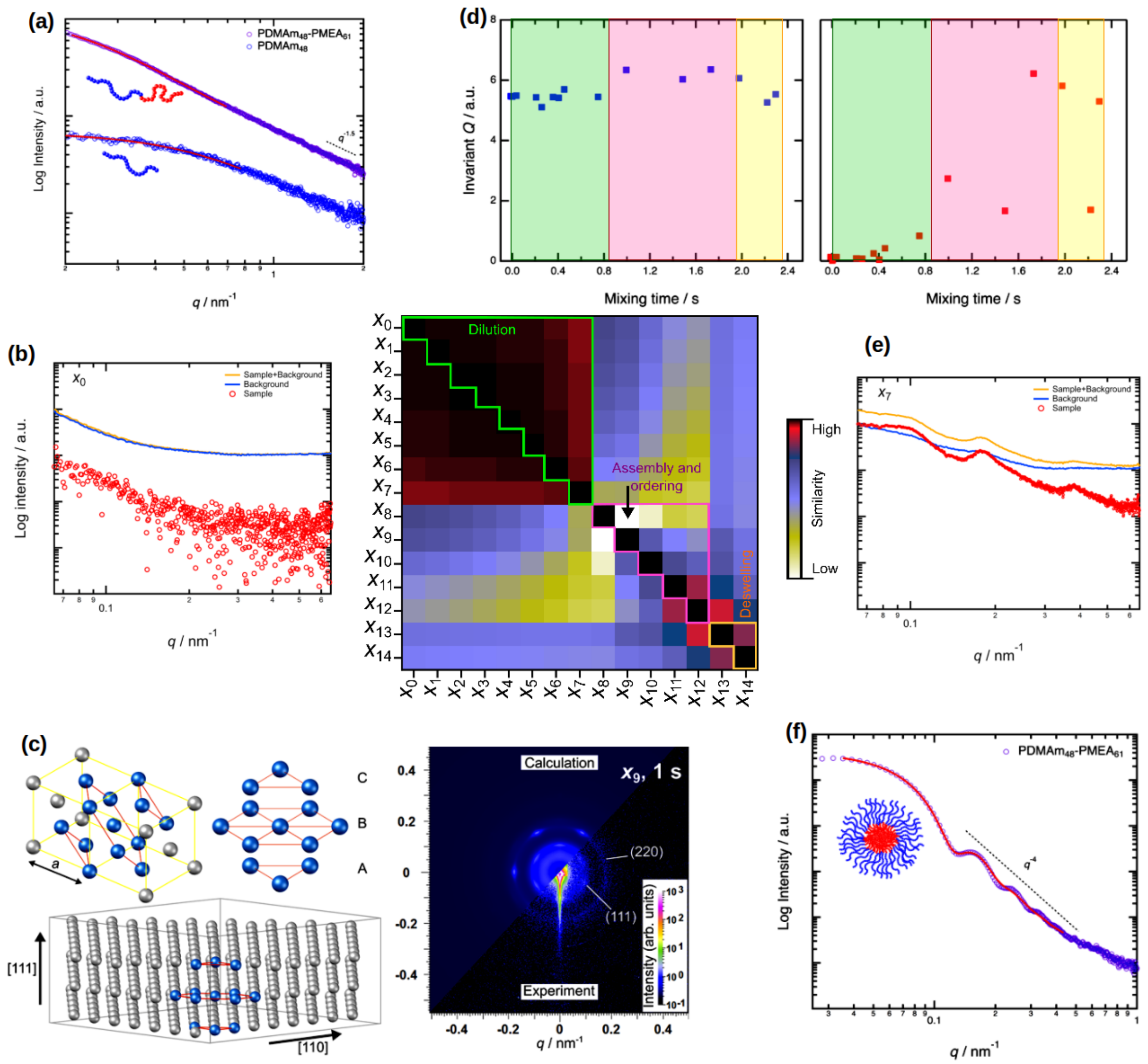


Figure 3.6 (Center) CSV-CORMAP demonstrating structure-related transitions during and after mixing. The onset due to ordered micelles is indicated by the significant change of metrics between x8 and x9 (black arrow). The map suggests three mixing regions which can be assigned to 1: solvent dilution (green); 2: assembly (pink); and 3: deswelling (orange). (a) SAXS profiles for the PDMAm48 macro-CTA in water (blue) and the PDMAm-PMEA block copolymer in THF (purple), each measured at 2% *w/w*. A Porod exponent of *s* ~ 1.5 points to swollen chains in good solvents. (b) SAXS profile at the channel position *x*0. Obvious elongation can be observed. It has a Porod exponent of *s* ~ 2.5 indicating a flow-induced anisotropy. (c) The experimental scattering pattern is in good agreement with a calculated 2D pattern showing the characteristic 6-fold symmetry expected when the X-ray beam is parallel to the [111] direction, thus placing the (111) plane perpendicularly. The fact that scattering from the (220) plane does not show distinctive reflections at the detector but rather an isotropic ring is strong evidence that the shear-induced ordering of (220) plane is experiencing continuous relaxation due to the insufficient shear stress at the center of the channel. (d) Plots of the Porod invariant *Q* as a function of the mixing time for the pure water (*left*) and for the background-corrected sample (*right*). The increase of the invariant for the sample at x_8 (after 0.75 s of mixing) is due to the increment of the volume fraction of the particles and therefore demonstrates the structure formation of spherical micelles upon

mixing in the microchannel. The onset was observed in the CSV-CORMAP with the color change between x_8 and x_9. (e) SAXS profiles at the downstream position x_7. It can be fitted with a model for the compacted FCC lattice of core-shell-spheres. (f) The red curves indicate obtained fits to the data using a sphere model (core-shell-sphere with a core radius of R = 35±5 nm, shell thickness of 3.4 nm)

Table 3.3 The fitted parameters from the *in situ* SAXS profiles

	x_0	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
R_g/ nm	18.2	--	--	--	--	--	--
R/nm	--	35.4	35.3	35.2	35.1	35.0	34.6
σ_R/nm	--	3.89	4.70	3.73	4.49	4.73	4.33
α/nm	--	100.0	99.9	99.6	99.4	99.0	98.0
σ_α/nm	--	10	13	14	11	12	17
D/nm	--	140	135	130	130	130	125

* R_g of the dissolved diblock polymer was obtained by fitting the data against the Debye function for the Gaussian chain. The micellar core radius (R), the unit cell length (α), the standard deviation of R and α (σ_R and σ_α, respectively), the ordered domain size (D) was obtained by fitting the model for a compacted FCC lattice of core-shell-spheres against the data using the package SCATTER (ver. 2.4) (Förster et al. 2011). The shell thickness was fixed at 3.4 nm.

The time-resolved structural evolution of the dispersed nanoparticles during mixing could be mapped onto different positions along the microchannel to realize millisecond time resolution. According to CSV-CORMAP, an important kinetics-related transition point occurs 200 ms after the end of solvent exchange, when the assembly of dissolved diblock copolymer chains to ordered spherical micelles is triggered. We also propose the existence of three dynamic states: solvent dilution, assembly, and deswelling. During the second phase, the gradual formation of ordered micellar phase and a shear-induced arrangement into the ordered lattice can be observed. Upon further and more complete mixing, a decrease in the spheres' size is observed. CSV-CORMAP also indicates that a dynamic balance has not been reached. Longer time monitoring would require a longer microchannel at the same flow rate. This is yet to be fabricated.

3.3.3 Conclusion

With the above cases, I have shown that CSV-CORMAP enables the semi-quantitative identification of crucial transitions in time-resolved experiments using solely SAXS profiles with no further assumptions. As a first hand analysis step, CSV-CORMAP excels in simplicity, since all one needs to do is input the absolute scattering data and compare the similarity by visualising. I note that CSV-CORMAP is neither a single-reference comparison such as the intensity difference plot, which creates the contrast by subtracting from the data from all the time delays the data at time 0, nor a simple pairwise comparison. Statistically, CSV-CORMAP can instead be considered as a test of separability (DeAngelis et al. 1995). $c_{i\text{->}j}$ indicates that, from time point *i* to time point *j*, the probability of the data can be represented by a single model. If the data can be better described by separate models, the value of $c_{i\text{->}j}$ would be less. CSV-CORMAP is thus very sensitive since it reports the degree of relative changes. One should bear in mind that once the SNR is lower than the actual changes, CSV-CORMAP might raise a false-positive. Nevertheless, when conducting time-resolved SAXS experiments, this method provides us a way to ask “is there any difference in my dataset?” and “whether this difference can be quantified?” The answer to such questions provides actionable information and gives immediate feedback to the users of high-throughout light sources. This actionable information is necessary for them to deal with the situation at hand. It helps, hopefully, both in making use of precious beamtime and in guiding the further analysis.

3.4 Deconvolution of SEC-SAXS Data from Membrane Proteins

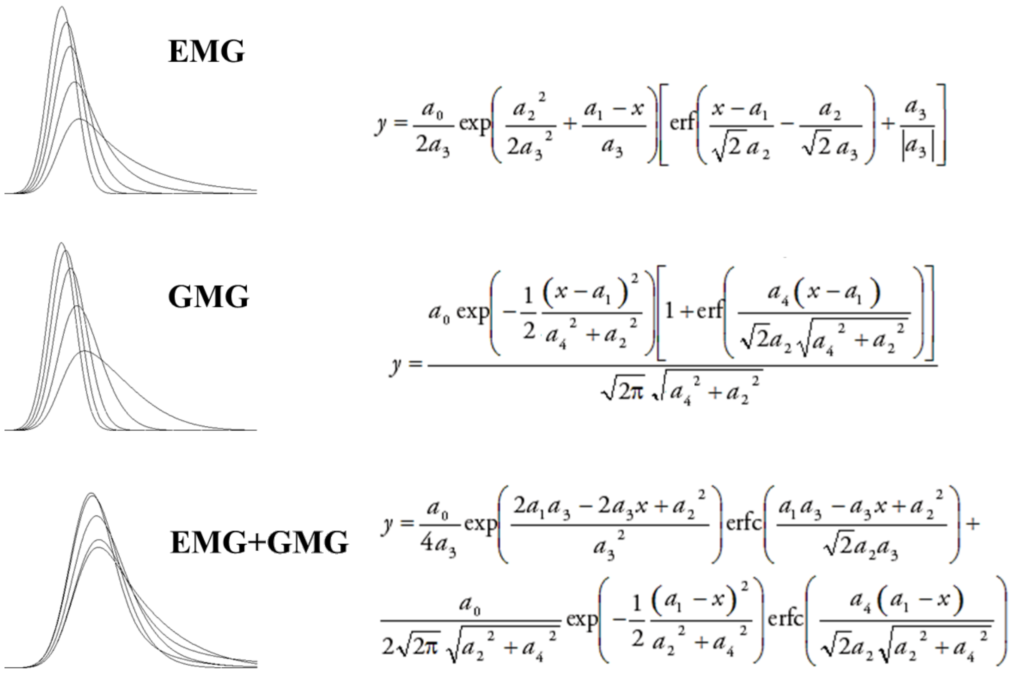
3.4.1 Background

Integral membrane proteins (IMPs), once extracted from cell membranes, are extremely challenging to study in solution due to their insolubility in aqueous media. In order to solubilize them, lipophilic agents such as detergents, lipids or nanodiscs (Chen & Hub 2015; Loll 2014; Ujwal & Bowie 2011) (Jeffries et al. 2016; Josts et al. 2018b) have to be used. Among these lipophilic agents, the most feasible and widely-used are detergents (Loll 2014). However, an IMPs/detergent solution contains free detergent micelles, due to the fact that the stoichiometric equivalence is not easy to find for different combinations of IMPs and detergents. For solution SAXS, the existence of an unspecified amount of detergent micelles hugely compromises the post processing. This problem can in principle be alleviated by using

SEC-SAXS, the data collected from the buffer eluting in the immediate vicinity of the protein peak provides a satisfactory estimation of the background (Berthaud et al. 2012). In such a way, the SAXS profile of the separated protein-detergent complex (PDC) can be determined. In practice, the situation is more complicated. Even though the background matching can be achieved, the signals from each elution fraction, more often than not, are not well resolved. Issues can include, for example, e elution peak tailing, excess detergent micelles, continuous elution of multiple oligomers and a conflict between the optimal flow rate and radiation damage. Thus, major efforts have been devoted to deconvolute the SEC-SAXS elution peaks so that the SAXS profile of the fraction of interest can be isolated.

One approach is utilizing the idea of chromatographic peak fitting of the SEC-SAXS dataset (Brookes et al. 2016), since $I(t, q)$ at each q has similar chromatographic features. It is assumed that the scattering elution peaks have the shape of modified Gaussians (Fig. 3.7): EMG+GMG, a linear combination of an exponential modified Gaussian (EMG) and a half-Gaussian modified Gaussian (GMG) with each factor contributing a half.

The initial parameters are estimated using one $I_q(t)$ selected from a low-q region. The subsequent global fit by q by applying the initial guess.



$$y = \frac{a_0}{2a_3}\exp\left(\frac{a_2^{\,2}}{2a_3^{\,2}} + \frac{a_1 - x}{a_3}\right)\left[\operatorname{erf}\left(\frac{x-a_1}{\sqrt{2}\,a_2} - \frac{a_2}{\sqrt{2}\,a_3}\right) + \frac{a_3}{|a_3|}\right]$$

$$y = \frac{a_0\exp\left(-\frac{1}{2}\frac{(x-a_1)^2}{a_4^{\,2}+a_2^{\,2}}\right)\left[1+\operatorname{erf}\left(\frac{a_4(x-a_1)}{\sqrt{2}a_2\sqrt{a_4^{\,2}+a_2^{\,2}}}\right)\right]}{\sqrt{2\pi}\,\sqrt{a_4^{\,2}+a_2^{\,2}}}$$

$$y = \frac{a_0}{4a_3}\exp\left(\frac{2a_1a_3 - 2a_3x + a_2^{\,2}}{a_3^{\,2}}\right)\operatorname{erfc}\left(\frac{a_1a_3 - a_3x + a_2^{\,2}}{\sqrt{2}a_2a_3}\right) + \frac{a_0}{2\sqrt{2\pi}\,\sqrt{a_2^{\,2}+a_4^{\,2}}}\exp\left(-\frac{1}{2}\frac{(a_1-x)^2}{a_2^{\,2}+a_4^{\,2}}\right)\operatorname{erfc}\left(\frac{a_4(a_1-x)}{\sqrt{2}a_2\sqrt{a_2^{\,2}+a_4^{\,2}}}\right)$$

Figure 3.7 Informative definitions of EMG (E.q. 3.8), GMG (E.q. 3.9) and EMG+GMG (E.q. 3.10). *a*1 and *a*2 are the center and width, respectively, of the classical normal distribution. a0 is the peak area. *a*3 and *a*4 are the distortions for EMG and GMG, respectively.

Another novel method is based on SVD of SEC-SAXS datasets and the so-called adaptation of evolving factor analysis (EFA) (Meisburger et al. 2016). EFA starts with the background subtracted SEC-SAXS datasets, the $I(q, t)$ matrix (the transpose of $I(t, q)$). The authors note that the columns of the right-singular vectors of $I(q, t)$ have the shapes that are reminiscent of elution peaks. Unfortunately, these vectors cannot represent physical elution peaks due to the sign changes within the vectors (Fig. 3.8). EFA manipulates the significant right-singular vectors (C_{jk}) with an operation called basis rotation so that the non-physical negative contribution is either inverted or set as 0 (Hopkins et al. 2017). The rotation stops when the estimator δ, which is the absolute change in the numerical sum of the C_{jk}, converges to a user-defined value. The remapped columns of C_{jk} are considered as the physical elution peaks and the multiplication of $I(q, t)$ and C_{jk} as the SAXS profiles of the pure components (Fig 3.9). This method provides a model-free validation of the deconvolved scattering profiles.

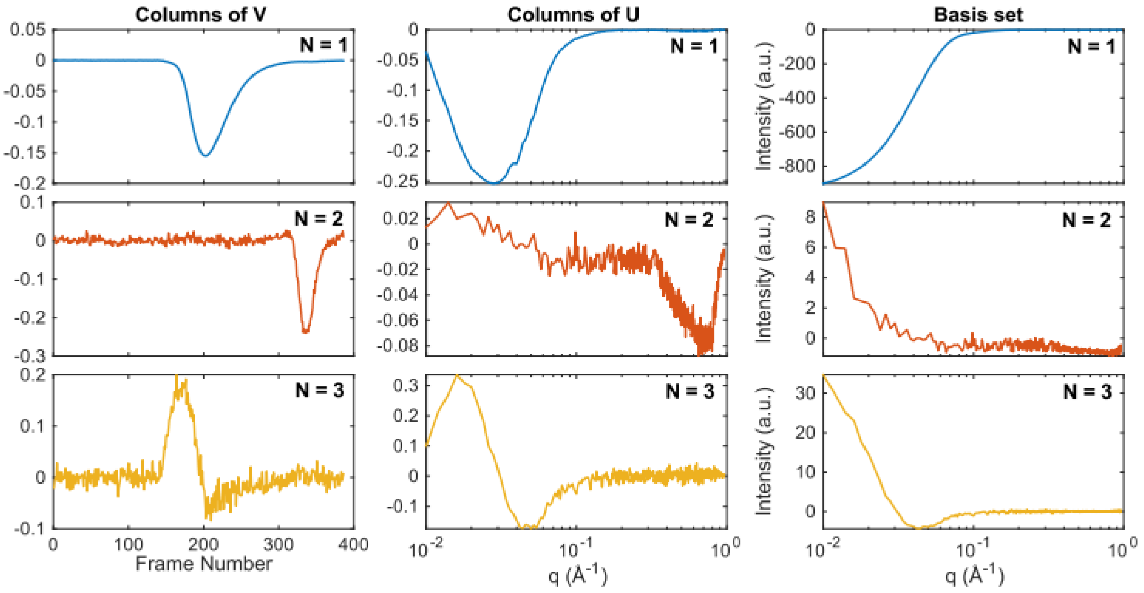


Figure 3.8 SVD analysis of SEC-SAXS data adapted from Figure S10 in (Meisburger et al. 2016), showing the three significant singular vectors. The right-singular vectors (columns of ***V***) have shapes that are reminiscent of elution peaks of each potential fraction, but have non-physical negative values. The basis set recovered by multiplying the column of ***U*** with the experimental error is reminiscent of the actual scattering profiles, although the sign can also be negative.

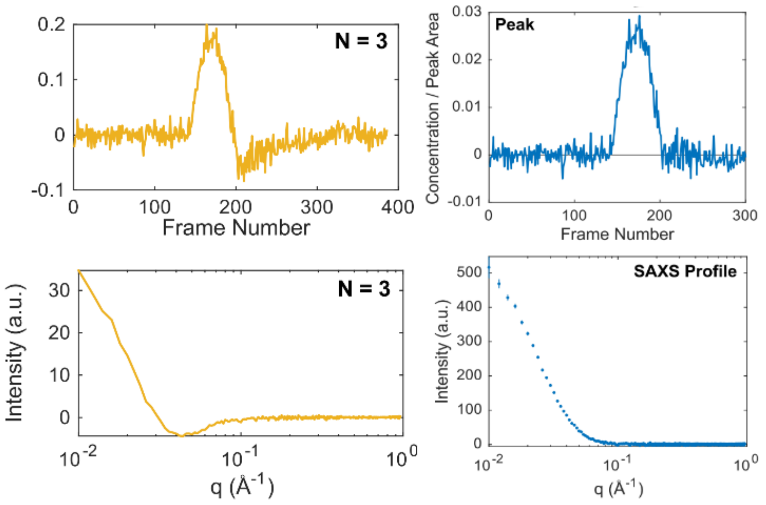


Figure 3.9 A comparison between the original (left) and the EFA remapped (right) 3rd left-singular vector and the respective scattering basis set. The remapped vectors are considered as the physical elution peak and the actual SAXS profile. Figure adapted from (Meisburger et al. 2016)

However, for EFA, the physical meaning of the basis rotation is not clear. Moreover, in my view, for SEC-SAXS of membrane protein EFA is a poor predictor of the scattering profiles. This in turn leads to doubts about the physical implication behind the remapped significant right-singular vectors.

Deconvolution using the peak fitting approach is rather straightforward. The deconvoluted peaks indeed have clear physical significance. However, despite its advantages, this method also has two major issues that are worth discussing: the choice of peak model and of the fitting algorithm. First, complex peak models such as EMG+GMG suffer from ill-conditioning problems. EMG is widely used in chromatography for peak analysis (Golubev 2017; Grushka 1972; Purushothaman et al. 2017). It is the convolution of a normal distribution and a negative exponential response function. It aims to describe the asymmetry in peak shape that occurs as a result of band broadening. Physically, this can be related to the first order response from a detector but can also be used empirically to aggregate the column effects. Both GMG and EMG+GMG were originally developed in a commercial package, Peakfit (Systat Software, Inc. USA). GMG is the convolution of a normal distribution and a half-Gaussian response function. It is used to take the intra column effects such as axial diffusion, dispersive effects, and mass transfer resistances into consideration. EMG+GMG, as mentioned in the documentation of Peakfit, is a "furnished empirical function" to include the features of both EMG and GMG. In convolution science, the parameters of a Gaussian in any form cannot be assumed to have any significance without a solid ground. The nature of the peak deconvolution requires a single set of parameters, except for $a1$, to be shared across all the resolution bins. However, it is practically not trivial to randomly pick a $I_q(t)$ and then to fix the initial parameters inferred by the first prediction. A certain degree of flexibility should be allowed to better deal with the trend of structure factor and the decrease of SNR. Under these circumstances, one is in effect fitting an empirical model multiple times. As an empirical model, all three forms of modified Gaussians are able to describe a wide variety of fronted and tailed peaks. As the famous Occam's razor principle suggests, when presented with competing hypotheses that make the same predictions, one should select the solution with the fewest assumptions. That means EMG should be efficient enough to fit the SEC-SAXS data when, if at all, peak fitting is the way to deconvolute the elution fractions.

The question is therefore whether one should blindly trust the fitting result considered just as a curve fitting problem which may have multiple minima (ill-posed). Fitting and finding local non-global minima is a very common, if not universal, problem. It is not specifically mentioned in (Brookes et al. 2016) which solver has been implemented. Nonetheless, for a non-linear curve fitting problem such as the one we are facing, the convergence to the global minimum should be carefully diagnosed. There are many optimization algorithms designed to find the global minimum of a non-linear fitting. Before any fine tuning (e.g. return starting at final point, change tolerances, rescale coordinates), it is necessary to develop a proof-of-concept study on the quality of the popular well-implemented optimization algorithms. The suitable algorithm should be able to not only achieve the mathematical approximation of the global minimum but also return a physically comprehensive output.

In this section, I have designed an experiment to test the performance of different optimization algorithms using the SEC-SAXS data collected for the Mhp1 (Na^+-Hydantoin Membrane Transport Protein)/detergent system. The results from different minimization algorithms are evaluated by the residual sum of squares by *q*, the ***CSS***(*q*) and the parameter stability by *q*.

3.4.2 Experiments

3.4.2.1 SEC-SAXS data collection

SEC-SAXS data were collected on P12 at PETRA III (DESY, Germany) and BM29 at ESRF (France). In order to investigate the applicability of SEC-SAXS to Mhp1, it was prepared in detergents with varying alkyl chain lengths. His_6-tagged Mhp1 was initially purified in *n*-dodecyl-*β*-D-maltopyranoside (DDM). The protein (10-20 mg/mL) was stored in 10-20 μL aliquots at −80 ℃ after flash-freezing in liquid nitrogen. Mhp1-DDM complex solutions were prepared by diluting the protein in a storage buffer (10 mM Tris-HCl pH 7.6, 2.5% *v/v* glycerol, 0.02% *w/v* DDM, 140/500/1000 mM NaCl) to a final concentration of 10~15 mg/ml. Mhp1-NM complex solutions were prepared by detergent exchange with an additional washing step using a Ni^{2+}-NTA column. The first washing step involved washing the resin with a buffer containing DDM, in order to initially wash off impurities and unbound protein. The second washing step was performed using a buffer containing *n*-decyl-*β*-D-maltopyranoside (NM) (10mM TrisHCl pH8, 2,5% *v/v* Glycerol, 0.7% *w/v* NM). The bound protein sample was then eluted with an elution buffer (10mM TrisHCl pH 8, 200mM imidazole pH 8, 2,5% *v/v* glycerol, 0.7% *w/v* NM). The eluted protein was concentrated using a Vivaspin 20 MWCO 100,000 spin concentrator at 3000 *g* to a volume of approximately 3 ml. The concentrated protein solution was applied onto an Econo-pac® 10DG desalting column and exchanged into a storage buffer (10mM TrisHCl pH8, 2,5% *v/v* Glycerol, 0.5% *w/v* NM, l140 mM NaCl) with a final concentration of 10~15 mg/mL. A Superdex™ 200 Increase 3.2/300 column was pre-equilibrated with the same storage buffer at a flow rate of 0.1 mL/min for at least two column volumes. 25 µL of sample solutions were injected onto the column system at a flow rate of 0.15 mL/min. Continuous exposure was applied with an exposure period of 1 s and 0.05 s read-out. An in-line UV detector was mounted and the absorption at 280 nm was recorded simultaneously. Absolute scattering intensity was calibrated according to E.q. (1.5).

3.4.2.2 Peak Model

The original formulation of EMG (Grushka 1972), instead of the chromatographically informative formulation (Kalambet et al. 2011), is used as a computational-friendly analytical model.

$$f(t;A,K) = \frac{A}{2K}\text{exp}(\frac{1}{K^2})\text{exp}(-\frac{t}{K})\text{erfc}(-\frac{t-1/K}{\sqrt{2}}) \tag{3.11}$$

This parameterization is a rewritten alternative to the definition in E.q. 3.8 where *K* equals to *a*3/*a*2 and *t* to (*x*-*a*1)/*a*2. erfc(*x*) is the complementary error function defined as erfc(*x*) = 1 − erf(*x*). The vector of parameters ***p*** = {*A*, *a*1, *a*2, *a*3}.

3.4.2.3 Minimization Algorithms

The minimization algorithm (minimizer) plays a central role when dealing with model fitting problems. Minimization algorithms adjust the function parameters so that the model fits the data as closely as possible. The objective function defines the concept of how close a fit is to the data. There are no efficient mathematical methods to solve this problem in general. To compare the performance of different minimization algorithms in deconvoluting SEC-SAXS dataset, eight common and well-implemented minimization methods were used (Table 3.4). All the Newton's methods (L-BFGS-B, TNC) used here have a certain inner iteration for the numerical approximation of Jacabian and Hessian, as the analytical Jacabian and Hessian are not trivial to calculate, and are therefore called quasi-Newton methods. The conjugate gradient method (CG) can be considered as a moderate case between the steepest descent method and Newton's method (S. Djordjevic 2019): a positive deflector is always used to orthogonally direct the searching step, thus Hessian is not required. The downhill simplex method is a direct search method which explores the conformational space, instead of being guided by the gradient, by heuristically evaluating the points in a n-dimensional simplex, where n is the dimension of the conformational space. The trust region method is similar to the quasi-Newton method in the sense of iterative approximation but is different to others as the step size is decided before the step direction. Nonlinear least-squares are the default methods for the most optimization solvers. However, nonlinear least-squares is only robust in finding the local minimum and can only be used to minimize a sum of squared function values for which the Hessian is not required. The detailed description of each method can be found in the relative references.

Besides the commonly used residual sum of squares (*RSS*), negative log likelihood (*NLL*) has also been tested as the objective function because it presumably gives greater suppression of outliers.

Table 3.4 The summary of tested local minimization algorithms

minimizer	type	objective function
limited-memory BFGS for bound-constrained optimization (L-BFGS-B) (Byrd et al. 1995)	bound-constrained quasi-Newton method	*RSS*/*NLL*
conjugate gradient with Polak-Ribiere updating factor (CG) (Shewchuk 1994)	nonlinear conjugate gradient method	*RSS*/*NLL*

truncated Newton (TNC) (Nash 2000)	bound-constrained quasi-Newton method	*RSS*/*NLL*
adaptive Nelder-Mead simplex optimization (NM) (Gao & Han 2012)	downhill simplex heuristic global minimization	*RSS*/*NLL*
nonlinear interior point trust region optimizer (NITRO) (Byrd et al. 1999)	inequality bound-constrained trust-region method	*RSS*/*NLL*
Levenberg–Marquardt least square with trust region reflective (LMTR) (Branch et al. 1999)	bound-constrained nonlinear least squares	*RSS*

* all the methods used above have been implemented in the scipy.optimize module (Virtanen et al. 2019).

* residual sum of squares (RSS) is defined as $RSS = \sum(Res/\delta)^2$, where $Res = (I_q(t) - f(t; A, K))$, δ is the corresponding estimated experimental errors. If δ is composed of only Gaussian noise, RSS equates to reduced χ^2.

* negative log likelihood (NLL) of a truncated Cauchy (Nadarajah 2011) is defined as

$$NLL = \frac{1}{N}\sum\left\{-\log[\frac{1}{\arctan(1)}(1+(\frac{Res}{\delta})^2)^{-1}]\right\} \tag{3.12}$$

δ is the corresponding estimated experimental error, N is the number of samples.

If needed, the Jacobian (J_{ij}) is numerical approximate using central differences,

$$J_{ij} = \frac{\partial I_q(t_i)}{\partial p_j} = \frac{I_q(t_i; p+\delta p_j) - I_q(t_i; p-\delta p_j)}{2\left\|\delta p_j\right\|} \tag{3.13}$$

3.4.2.4 Evaluation of Fitting Results

Evaluation based on the convergence of the objective function is the rule of thumb for any minimization problem. It is worth mentioning that the fittings at different q regions are of different levels of difficulty. There are four regions for this minimization problem (Fig. 3.10). For the very low q part (VLQ), the intensity is interfered with by the beam stop shadow. The peak models may not be adequate to approximate the data and the fitting can thus be badly converged. For the low q part (LQ), the difficulty is relatively low for the minimizers because in this region the SNR is high and the intensity ratio between fractions is rather stable. Moreover, the initial point is deduced from one $I_q(t)$ from this region which provides enormous advantages. The local minimum part (MIN) has the highest difficulty. In this region, the noise is often on the same level as the signal. The minimizer can easily get stuck in a local minimum. For the part where the detergent feature (DF) dominates, the problem has medium difficulty. Here the difficulty mainly comes from the potentially larger distance between the starting point and the local minimum. In general, producing better results at a certain *q* region does not imply that the minimizer can better fit the data of another q region. The definition of regions is listed in Table 3.5.

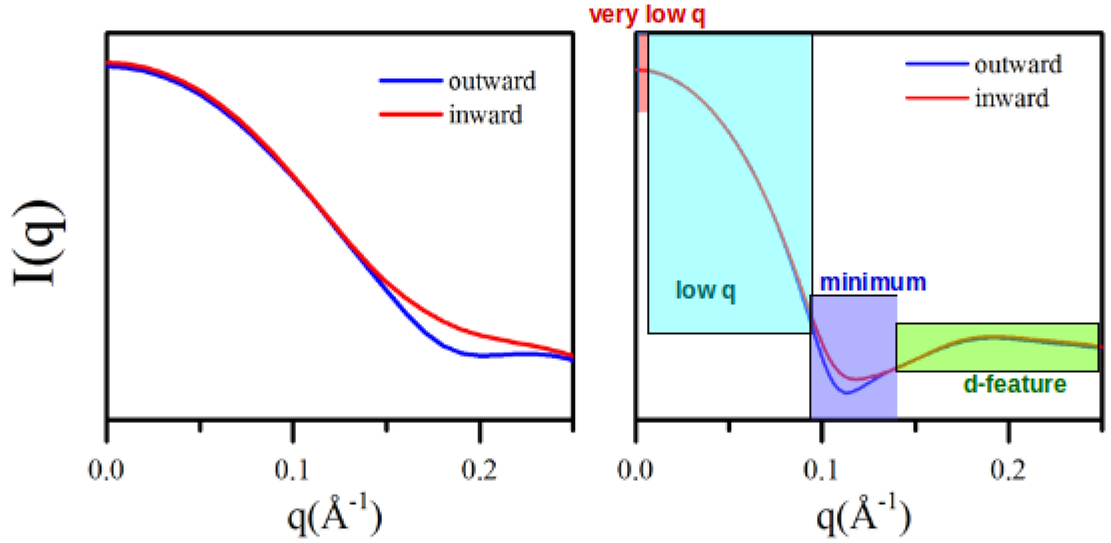


Figure 3.10 The theoretical scattering profiles of MHP1 with different conformational states. (Left) Scattering of protein alone without detergent. (Right) Scattering of the PDC. Four regions are highlighted. These are the very low-*q* region (red), the low-*q* region (cyan), the local minimum region (blue) and the detergent-feature region (green), respectively.

Table 3.5 Definition of the regions based on the difficulty level

	VLQ	LQ	MIN	DF
q range (10^{-2} Å^{-1})	0.88974-1.5550	1.5827-6.4332	6.4609-9.2405	9.2326-22.398
difficulty	high	low	high	medium

However, due to the fact that there is no certified solution available, the evaluation of convergence can only rely on the final optimal solution, which is not necessarily the global minimum. Nevertheless, in Chapter 2, I have shown that CSS is capable of objectively assessing the truthfulness of a background correction. This idea can also be applied to SEC-SAXS data of membrane proteins. The scattering intensity of the monomeric PDC is proportional to the protein concentration, provided the configuration of the detergent corona is stable. The UV signal of the monomeric PDC can be considered semantically similar to its SAXS signal. A globally converged or physically meaningful deconvolution should give an overall good ***CSS***(*q*) between the SEC-SAXS elution peak and the UV elution peak of the monomeric PDC. Here, the SEC-SAXS elution peak of monomeric PDC identified from the global fitting is used as $I(t, q)$.

3.4.3 Results and Discussion

3.4.3.1 Starting Point

A good practice for comparing optimization algorithms should NOT allow different algorithms to use different starting points. However, an obviously bad starting point will bias the comparison result. For example, if the starting point is at the boundary of a constraint set, an inner iteration method such as TNC, L-BFGS-B will be severely disadvantaged (Beiranvand et al. 2017). Because full automation of a global peak fitting by q is practically not trivial, here I manually deduce the first set of parameters using a single low-q $I_{\mathrm{q}}(t)$ or the q-averaged $I(\mathrm{t})$. This initial successful prediction in principle places the parameter set at a position where the objective function is not unacceptably distant from the global minimum. Since the efficiency of the algorithms is not within the scope of this proof-of-concept study, the parameter set inferred from the first prediction is hence considered as a “fair” starting point.

Fig. 3.11 shows the deconvolution result using EMG of the UV elution profile and the SAXS elution profile. For both of them, the oligomer (peak1) and the major fraction (peak2) are identified. The third peak in the SAXS elution profile, which cannot be detected in the UV elution profile, relates to the empty DDM micelles.

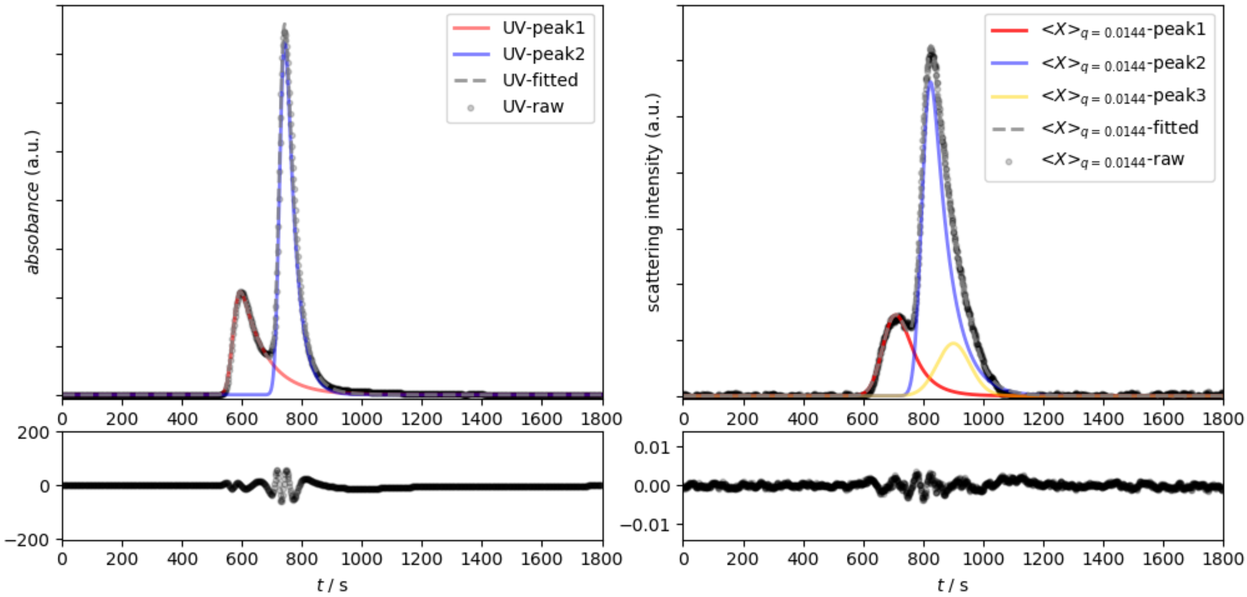


Figure 3.11 (left) The UV elution profile recorded at 280 nm. The R^2 of fitting is 0.9986. Two fractions (blue and red solid lines) are identified. (right) SAXS elution profile represented using $I_q(t)$ at q = 0.0144 Å^{-1} .The R^2 of fitting is 0.9988. Three fractions (blue, red and yellow solid lines) are identified. The bottom panel shows the residuals of the fittings for the UV elution profile (left) and the SAXS elution profile (right), respectively. The plot scale is set to be one tenth of the actual intensity range for the raw data.

For SEC-SAXS, the model used for further procedures is written as,

$$F(t) = f_1(t_1; \boldsymbol{p}_1) + f_2(t_2; \boldsymbol{p}_2) + f_3(t_3; \boldsymbol{p}_3) \tag{3.14}$$

$F(t)$ is the composition of three individual EMGs indexed in the order of elution time.

The starting point inferred from the fitting of $I_{q=0.0144}(t)$ is listed in Table 3.6

Table 3.6 Initial parameters for global fitting.

t_1	A_1	K_1	t_2	A_2	K_2	t_3	A_3	K_3
$\frac{t-676.62}{34.50}$	3.86	1.67	$\frac{t-800.41}{22.32}$	12.04	2.85	$\frac{t-912.68}{44.25}$	2.58	0.300

3.4.3.2 Performance of Minimization Methods

Convengency of Objective Function

The *RSS* (Fig. 3.12) of all local minimizers clearly shows the fitting results are *q*-region dependent. For VLQ, none of the minimizers can converge to a reasonable value. This strongly suggests the peak model is not capable of describing the SEC-SAXS elution profiles at VLQ. A comparison is made for the remaining three regions (Fig 3.13). NM has a slight lead in LQ as *RSS* = 1.0 implies a good fit. Although CG and NITRO are comparable here, both of them have larger discrepancies suggesting the minimizer is less stable. According to the reduced χ^2 statistics, a fit (in our case with an effective sample size ~ 800) within the interval $0.8 < \chi^2 < 1.2$ can be considered converged (Andrae et al. 2010). A χ^2 larger than 1.2 is considered a bad fit whereas smaller than 0.8 an overfit. By this standard, MIN and DF, especially the former, experience overfitting. However, the expectation value of reduced χ^2 distribution only equals one under the conditions that the data only have Gaussian noise and the true model with the true parameter values are applied, which are both not necessarily true in our case.

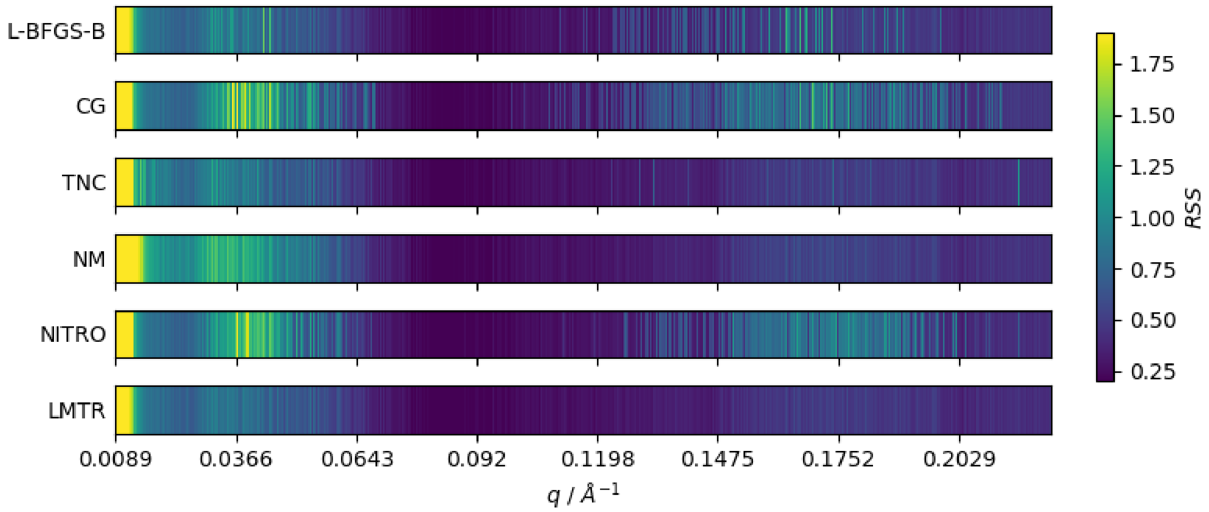


Figure 3.12 The 1D colormap showing *RSS* of non-linear fitting results by *q*. A perceptually uniform color palette "viridis" is used to map the scale of 0.2 to 1.9.

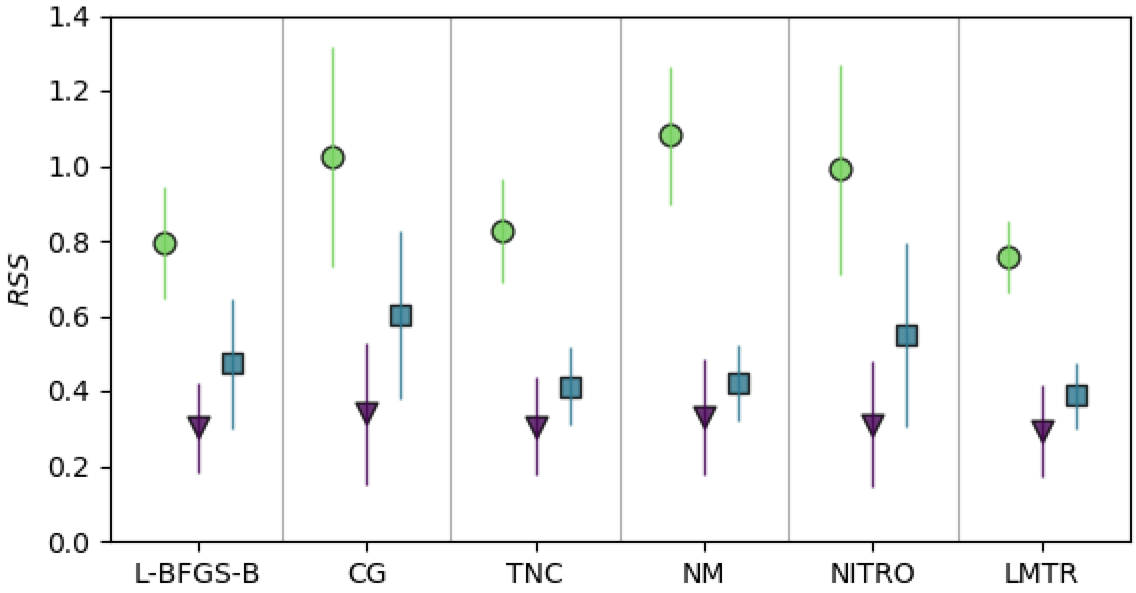


Figure 3.13 A comparison between the fitting results of different regions. The green circles represent the mean value of *RSS* in LQ. The purple triangles represent the mean value of *RSS* in MIN. The blue squares represent the mean value of *RSS* in DF. The error bars represent the corresponding variance.

By using *NLL* as the objective function, the minimization is equivalent to maximising the likelihood function. The very high values in VLQ again confirm the model is not able to predict the dataset in this region. Among the four applicable methods, NM shows clear advantages in goodness-of-fit as well as stability, especially for fitting the data in the DF (Fig. 3.14 and Fig. 3.15). Unlike reduced χ^2 statistics, *NLL* is less informative in regards to the quality of prediction. Nevertheless, it is clear that NITRO results are set with significantly larger deviations, which implies the occurrence of overfitting.

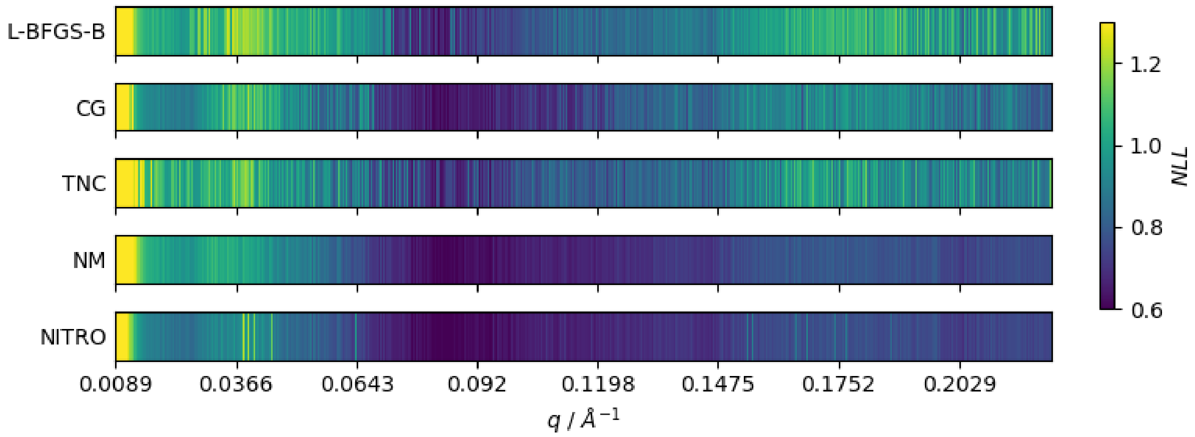


Figure 3.14 The 1D colormap showing *NLL* of non-linear fitting results by *q*. A perceptually uniform color palette “viridis” is used to map the scale of 0.6 to 1.3.

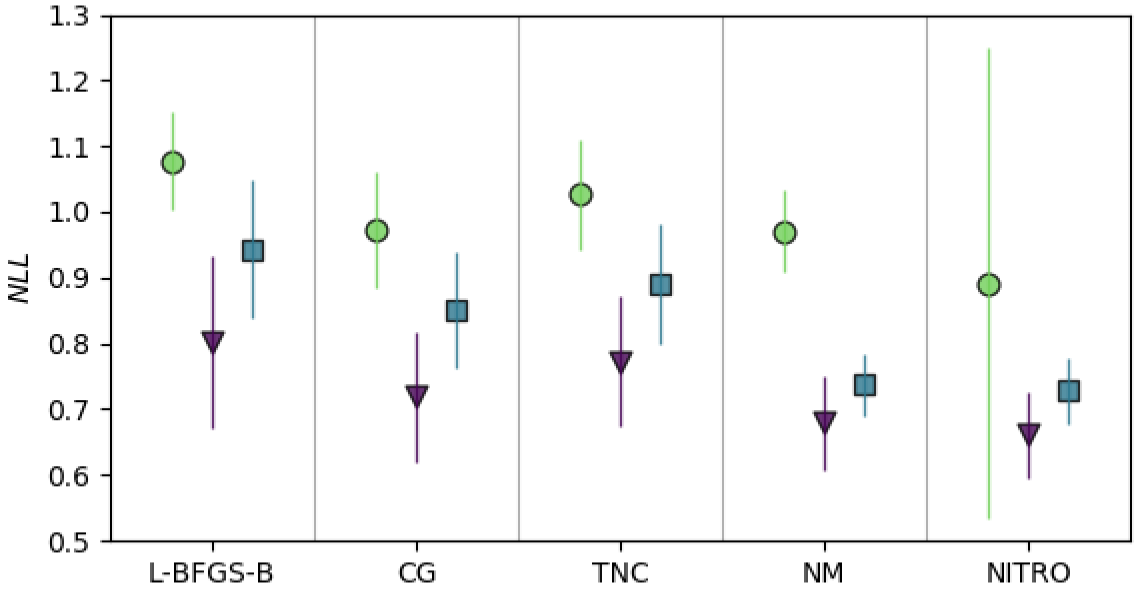


Figure 3.15 A comparison between the fitting results of different regions. The green circles represent the mean value of *NLL* in LQ. The purple triangles represent the mean value of *NLL* in MIN. The blue squares represent the mean value of *NLL* in DF. The error bars represent the corresponding variance.

Parameter Stability

If the inferred parameters are stable over a wide range of q, the minimizer is considered as a "broad-spectrum" minimizer. The trajectory of parameters along *q* is plotted to visualise the stability of each minimization method using either *RSS* (Fig. 3.16-3.21) or *NLL* (Fig. 3.22-3.26).

For each individual EMG (from left to right: f_1, f_2 and f_3) and its parameter vector $\boldsymbol{p}(q) = \{A(q), a1(q), a2(q), a3(q)\}$: the logarithmic amplitude ($\ln(\frac{A(q)}{a(q)})$), the peak location of its Gaussian component ($a1(q)$), the mean normalized scale ($\frac{a2(q)\ - mean(a2)}{max(a2) - min(a2)}$) and the mean normalized skewness ($\frac{a3(q)\ - mean(a3)}{max(a3) - min(a3)}$) are presented.

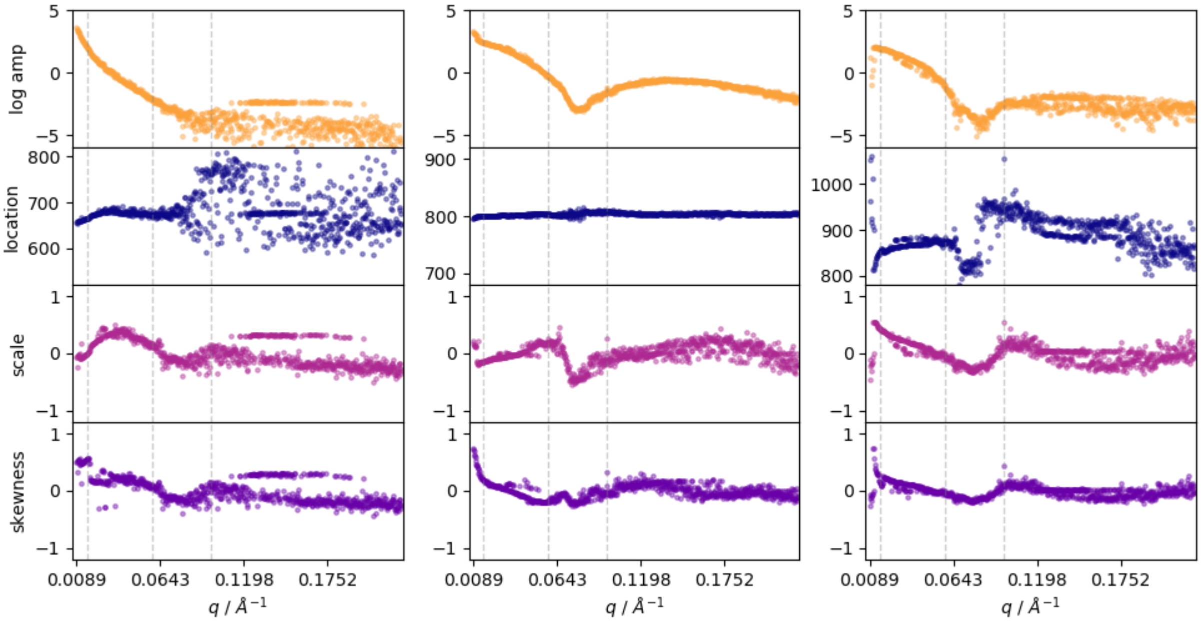


Figure 3.16 The trajectory of parameters along the q for L-BFGS-B using *RSS* as the objective function. All parameters, except for $a1(q)$ of f_2 , are q-dependent.

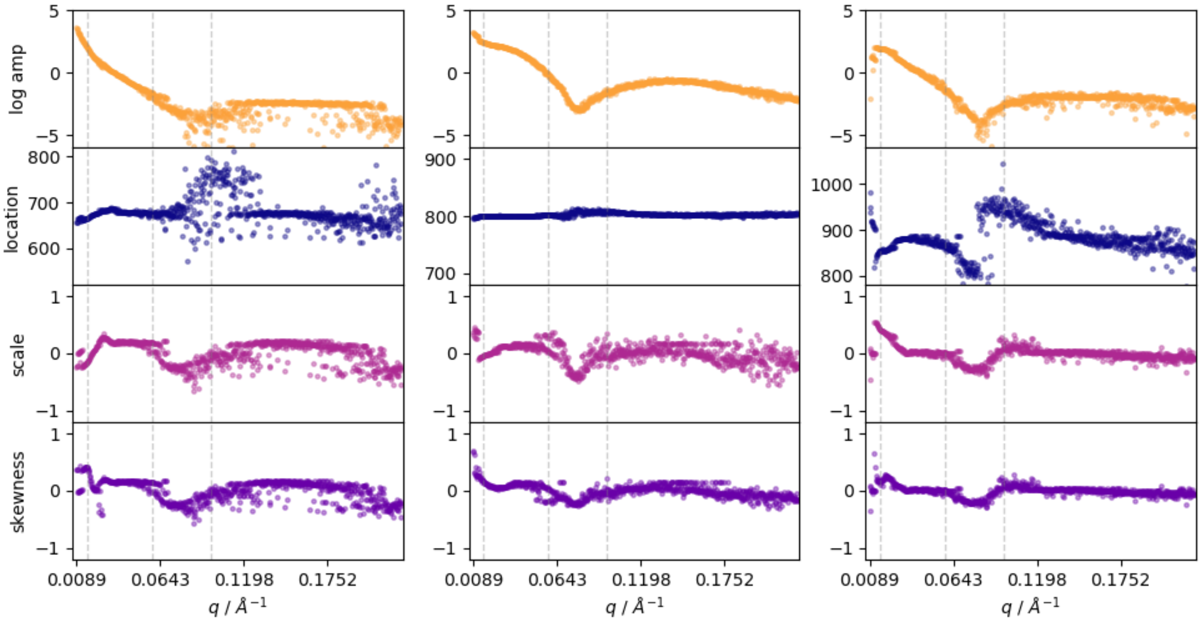


Figure 3.17 The trajectory of parameters along the q for CG using *RSS* as the objective function. All parameters, except for $a1(q)$ of f_2, are q-dependent. $a3(q)$ of f_2 and $a3(q)$ of f_3 show a reasonable stability in DF.

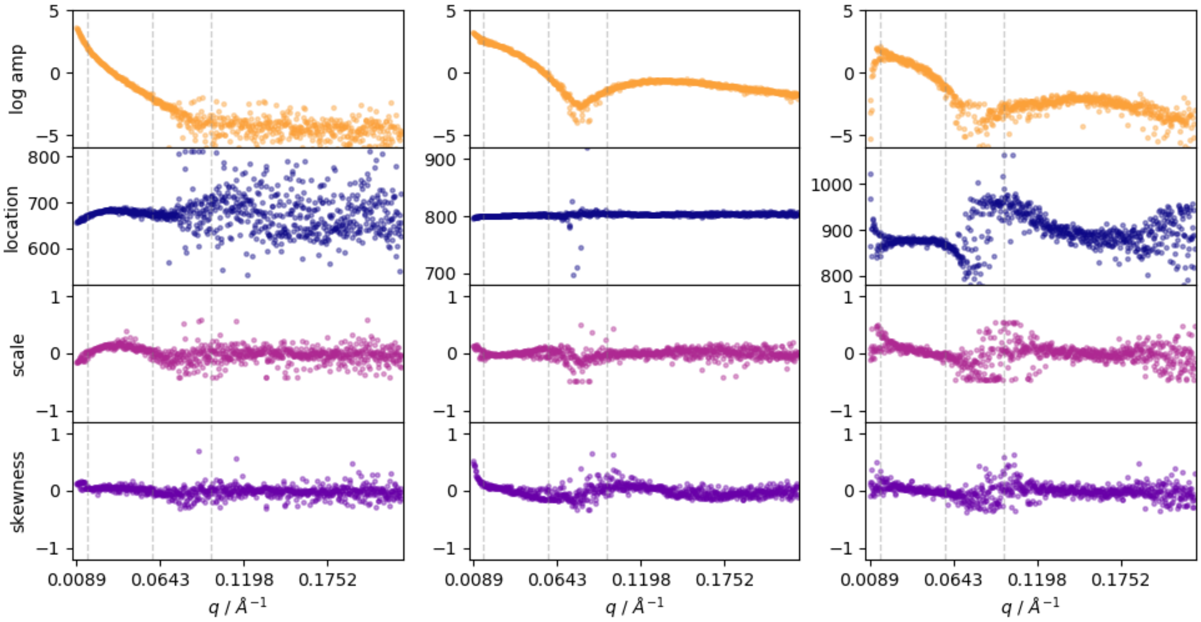


Figure 3.18 The trajectory of parameters along the q for TNC using *RSS* as the objective function. All parameters, except for $a1(q)$ of f_2, are q-dependent. The parameters of f_2 show a reasonable stability in LQ and DF.

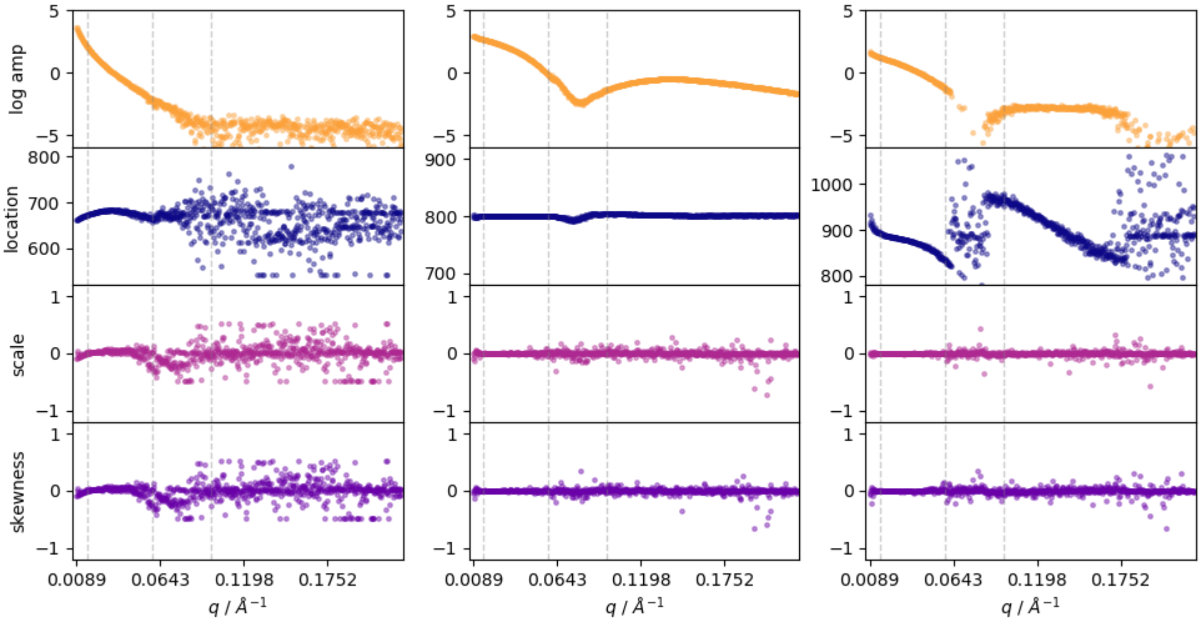


Figure 3.19 The trajectory of parameters along the q for NM using *RSS* as the objective function. Only $a1(q)$ of f_3 has clear q-dependence. All the parameters of f_2, $a2(q)$ of f_3 and $a3(q)$ of f_3 show a reasonable stability globally.

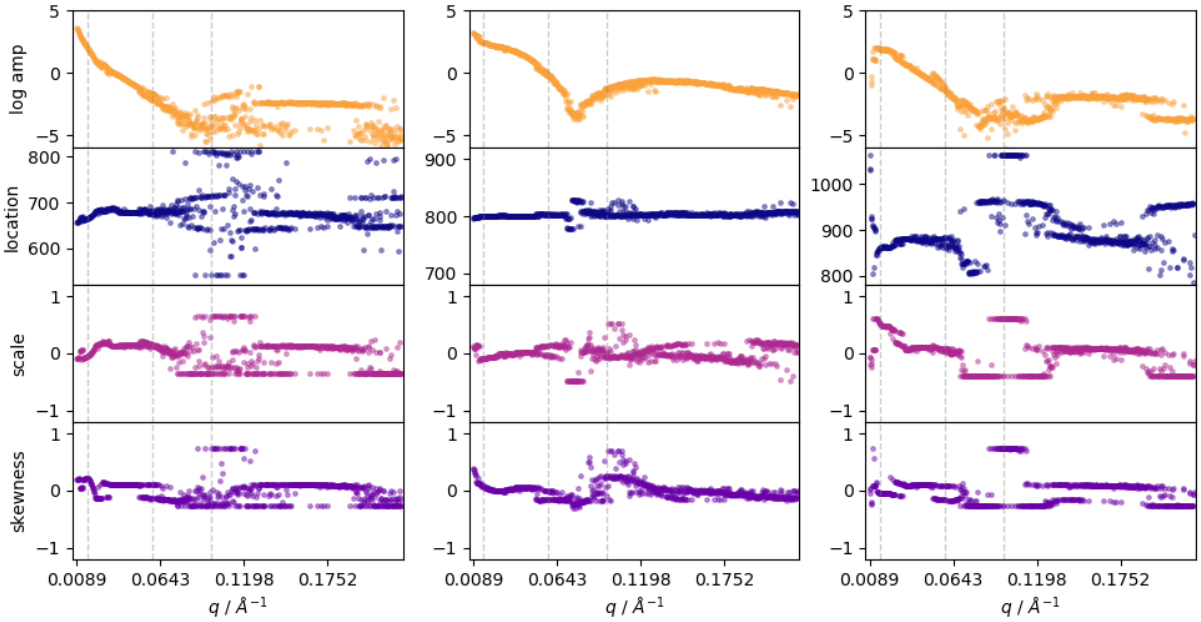


Figure 3.20 The trajectory of parameters along the q for NITRO using *RSS* as the objective function. All parameters, except for $a21$, are considered stochastic. Many outliers can be found for the parameters of f_2 and f_3.

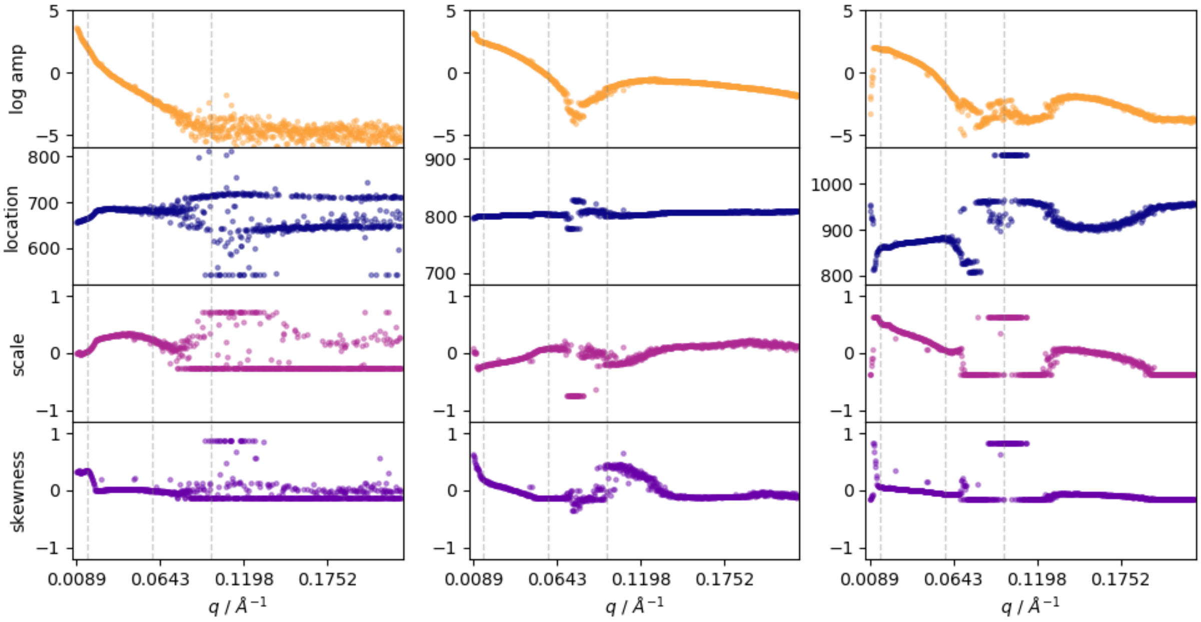


Figure 3.21 The trajectory of parameters along the q for LMTR using *RSS* as the objective function. All parameters are q-dependent. Parameter stability is very low in MIN and the low resolution part of DF.

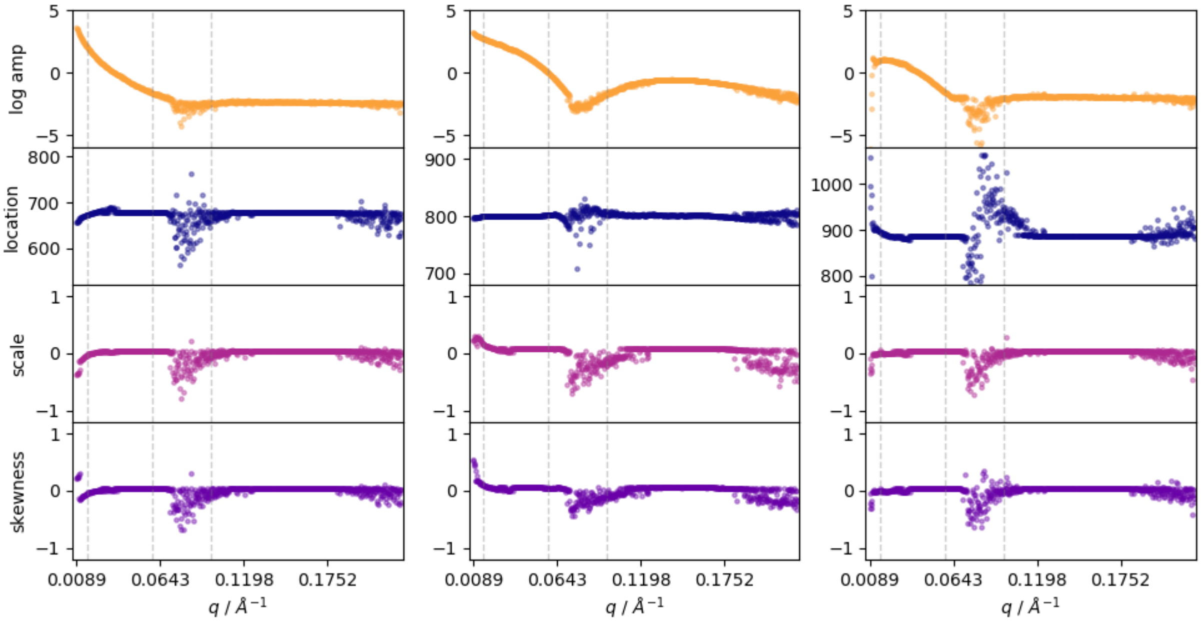


Figure 3.22 The trajectory of parameters along the *q* for L-BFGS-B using *NLL* as the objective function. All parameters show reasonable stability in LQ and the middle part of DF. The *q*-dependency of parameters has rather different patterns as the one using *RSS* as an objective function.

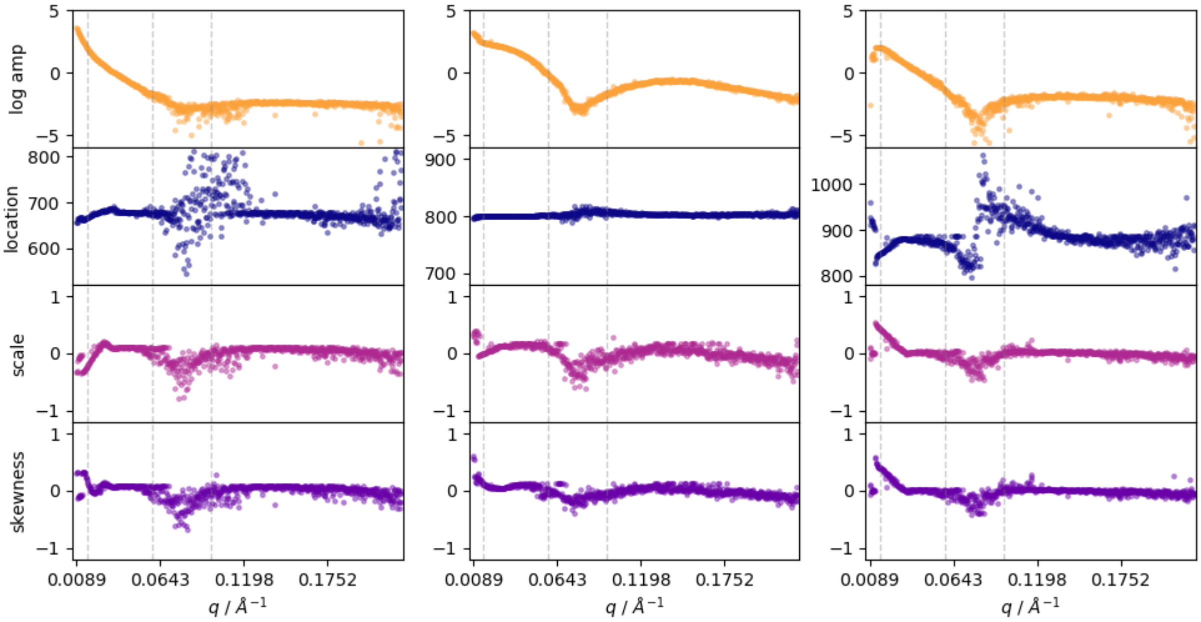


Figure 3.23 The trajectory of parameters along the *q* for CG using *NLL* as the objective function. All parameters, except for $a1(q)$ of f_2, are *q*-dependent. $a2(q)$ of f_3 and $a3(q)$ of f_3 show a reasonable stability globally.

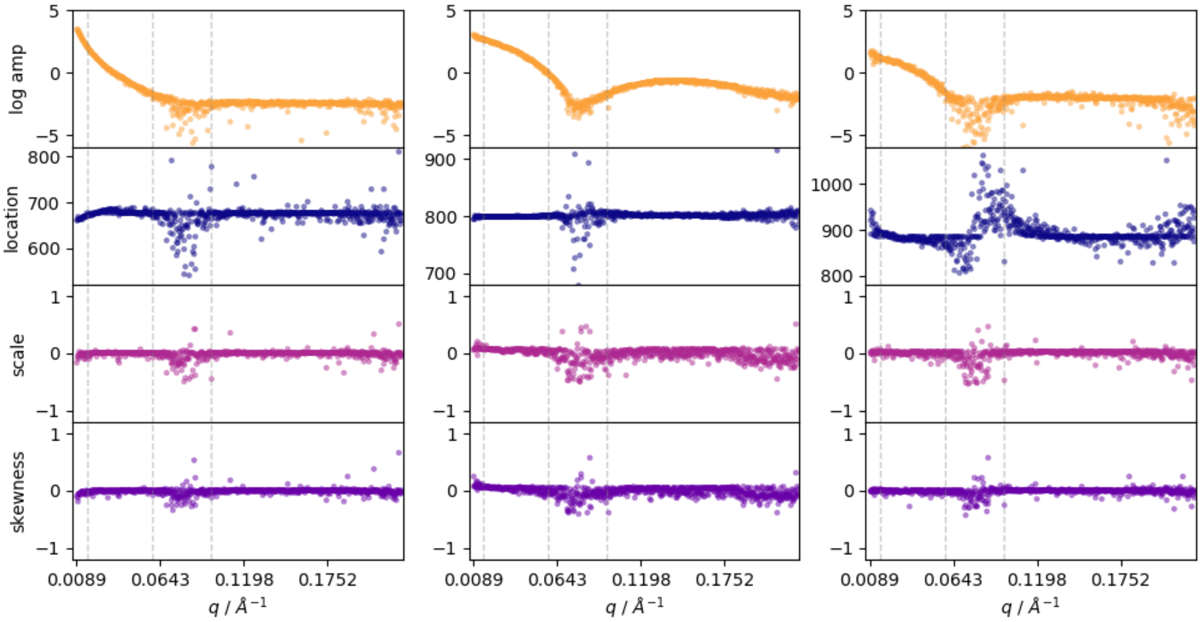


Figure 3.24 The trajectory of parameters along the q for TNC using *NLL* as the objective function. All parameters show reasonable stability in LQ and DF. The q-dependency of parameters has rather different patterns as the one using *RSS* as an objective function.

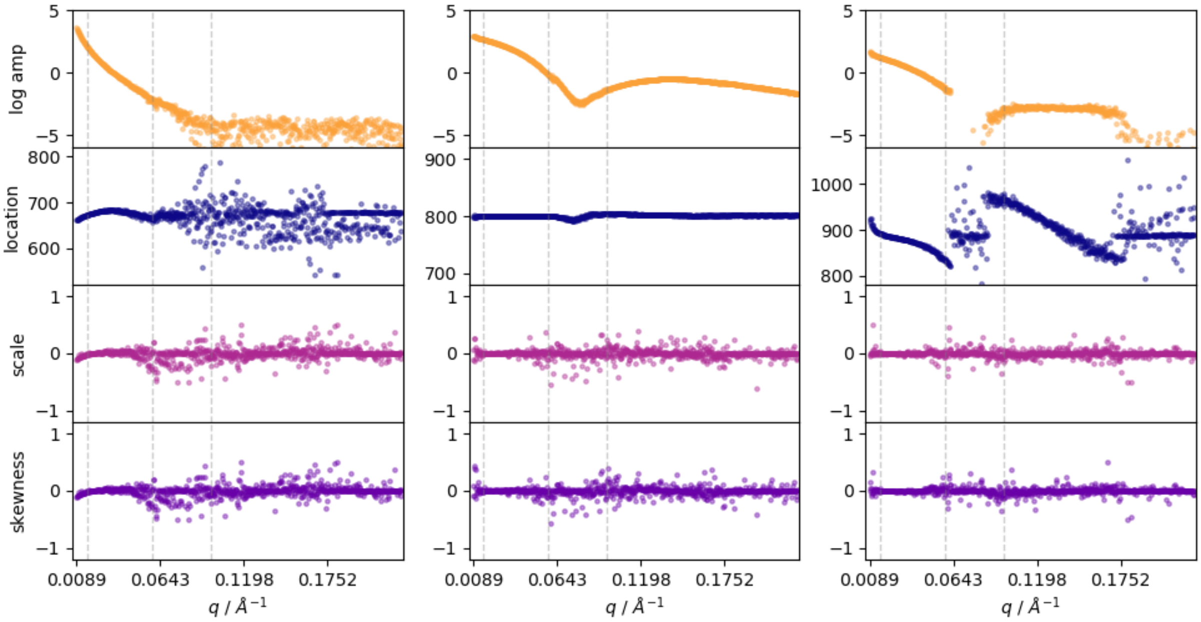


Figure 3.25 The trajectory of parameters along the q for NM using *NLL* as the objective function. Only $a1(q)$ of f_3 has clear q-dependence. All the parameters of f_2, $a2(q)$ of f_3 and $a3(q)$ of $f_3$show a reasonable stability globally.

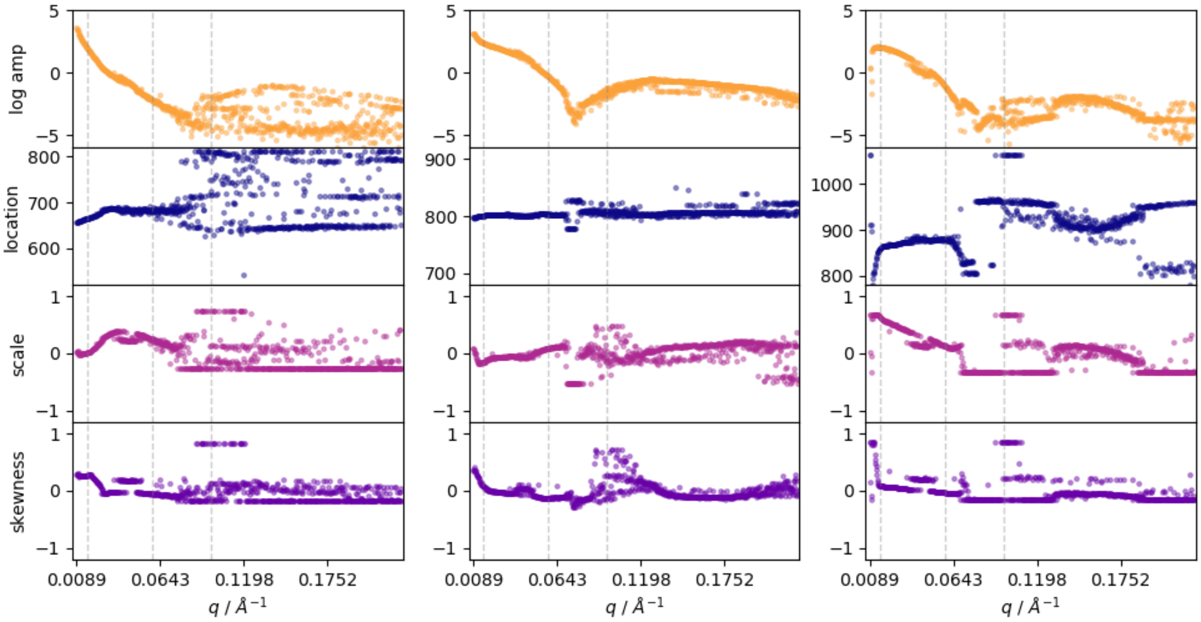


Figure 3.26 The trajectory of parameters along the q for NITRO using *NLL* as the objective function. All parameters show relatively large instability along q.

Inspection of all the trajectories shows that SNR has a considerable effect on parameter stability, especially for the gradient-based methods. In principle for a certain problem, if the situation that the true model has the true parameters has been approximated, parameters tend to remain sufficiently stable. I must emphasize, however, the parameter non-constancy does not necessarily imply that the solutions are over-fitted or under-fitted. Parameters that have a certain pattern may be a symptom of misinterpretation, which could involve the violation of certain assumptions of the applied model and/or the misspecification of other possible mechanisms (Coutts et al. 1997). Interestingly, other than CG and NM, the use of different objective functions results in non-unique patterns of q-dependent parameter variation. Notwithstanding the above caveats , this phenomenon is seemingly an implication that Newton's method based minimizers experience over-fitting problems in this study case.

CSS Assessment

The SAXS profiles of PDC manifested by different minimizers using *RSS* as the objective function yield subjectively comparable results (Fig. 3.27). This might be because the merging only includes a few frames from the separated f_2, which lessens the negative effects of parameter non-consistency. The ***CSS***(q) assessment suggests that the minimization results in VLQ are not faithful. This is in accord with the convergence analysis of the objective functions. For the given peak model and starting points, all minimizers, except for NITRO, are able to at least return supposedly acceptable deconvolution results for the certain resolution bins in LQ and DF. However, subtle differences for both the scattering profiles and the ***CSS***(q) can be found. The most noticeable one exists in MIN. Clear outliers in MIN are observed for TNC, NITRO and LMTR, which indicates these methods have poorer prediction if the SNR is low.Therefore, it is hard to conclude the merit rank of these minimizers based only on the values of objective function and the parameter stability. L-BFGS-B, CG and NM give few significant outliers in MIN, within which the NM result is most likely to be trustable. The CSS statistics suggest LMTR results in an overall better deconvolution for a higher ***CSS***(q) mean and a smaller regression coefficient in LQ as well DF (Fig. 3.28). A less obvious but rather important difference is that LMTR is slightly better than other minimizers for the problem in DF while NM is similarly competitive, especially in the high-q region, with smaller variances. Interestingly, although most outliers can be identified by CSS, some are not visible in the NITRO case. Combining this with the observed parameter instability, it appears that NITRO misassigns $a1(q)$ of f_1(the location of Gaussian component of f_1) and $A(q)$ of f_2 (the area of f_2) occasionally. Due to the deductive method generating the $I(t, q)$ for calculating ***CSS***(q), this error propagates and produces false similarity between $I(t, q)$ and *UV* within the selected *PSR*. Certainly, this operation has already involved subjectivity which CSS is blind to. This admissible failure is an example showing that the information incompleteness of *CHROMS* or/and *XSS* hampers the use of CSS.

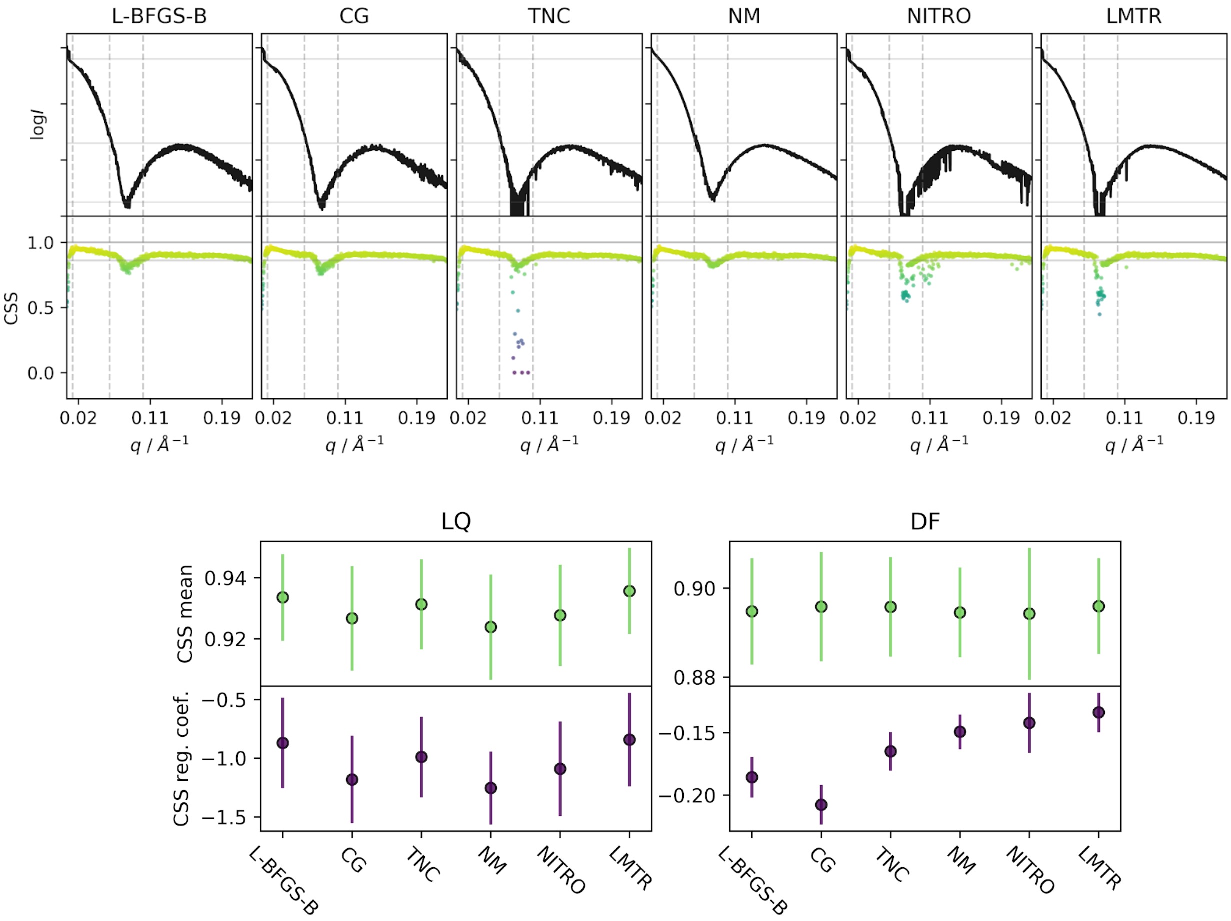


Figure 3.27 (Top) A comparison of the manifested SAXS profiles of the monomeric PDC from the different minimizer and the corresponding ***CSS***(*q*). (Bottom) The resulting ***CSS***(*q*) statistics of the LQ region and the DF. The objective function is *RSS*.

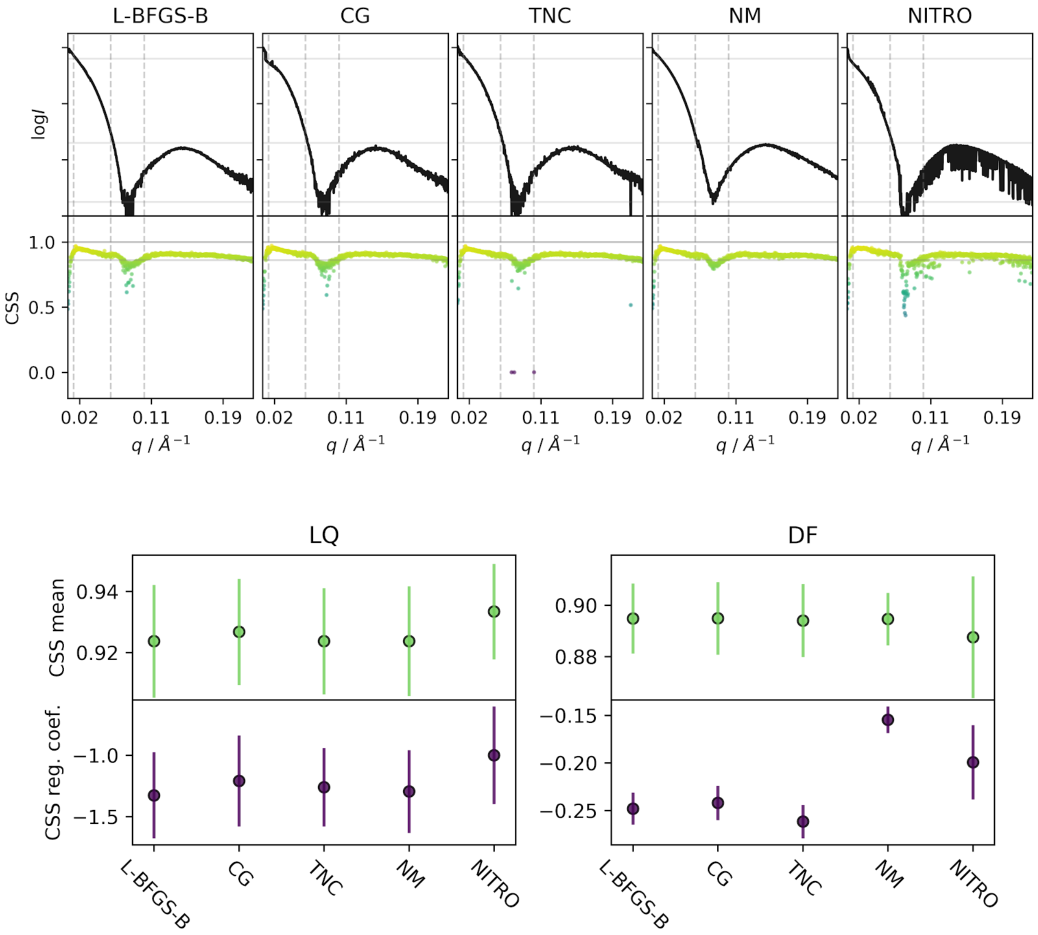


Figure 3.28 (Top) A comparison of the manifested SAXS profiles of the monomeric PDC from the different minimizer and the corresponding ***CSS***(q). (Bottom) The resulting ***CSS***(q) statistics of the LQ region and the DF. The objective function is *NLL*.

Compared to *RSS*, no significant improvement over ***CSS***(q) is observed when using *NLL* (Fig. 3.29 and Fig. 3.30). NM returns almost identical results from the two objective functions. It is expected that the nature of direct search method makes NM robust against the small perturbation of objective functions (McKinnon 1998). The trajectories of $a1(q)$ of f_3 (the location of Gaussian component of f_3) for NM however, show a pattern which is normally considered as over-fitting. This is an interesting observation since the parameter trajectories of other minimizers using *NLL* are more stable than the ones using *RSS* but generate worse ***CSS***(q). On one hand, this suggests that, for this problem, *NLL* may have a rugged-like surface which traps the gradient-based methods into a local minimum. On the other hand, it is intriguing to consider the possibility that the parameter non-consistency indeed reflects the physical truth which is misspecified by the model. A plausible explanation is that the empty detergent micelles have a rather broader size distribution. The larger sized micelles elute earlier and have greater contribution in the LQ region whereas the smaller micelles elute late and the peak position of DF shifts to larger q (Oliver et al. 2013) (Fig. 3.31). At a certain q, the larger the discrepancy between the SAXS curves of the larger and the smaller micelles the greater $a1(q)$ of f_3 will be. This is of course also affected by the relative concentration of each subspecies. Since the distortion of elution peaks is mainly affected by internal column factors (Purushothaman et al. 2017), it is reasonable to hypothesise that this effect of size polydispersity appears as a positional shift of the Gaussian components.

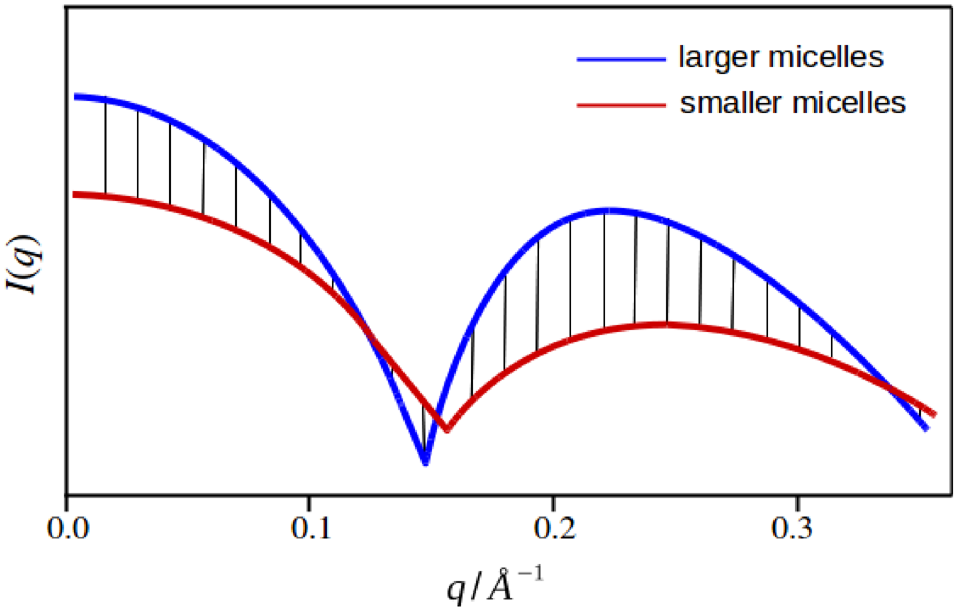


Figure 3.31 Schematics of the SAXS profiles of larger micelles (blue line) and smaller micelles (red line). The discrepancy (vertical solid lines) between two curves is q-dependent.

In summary, it appears that the NM minimizer has a balanced behavior. It returns the most robust results with regards to the different difficulties of questions. Most parameters are stable over the q-range, which to a certain extent signals that there is a less potential to introduce artifacts in the recovered deconvoluted curves. Moreover, although the speed performance has not been fully investigated, it must be emphasized that the speed of calculation for NM is comparably faster (the second fastest after LMTR). This advantage also makes further detailed optimization possible (e.g. multi-starting global optimization using Basin-hopping (Wales & Doye 1997)).

3.4.3.3 Validation by Studying the Effects of Ionic Strength

To further validate the above conclusion, the same procedure was used to deconvolute the SEC-SAXS data from the Mhp1-DDM system with different salt concentrations. It is known that for non-ionic detergents such as DDM, addition of monovalent salt has less effects on the aggregation number compared to ionic detergents (Seddon et al. 2004). However when ionic strength is sufficiently high, the effect becomes significant (Neale et al. 2013). This interesting phenomenon can be used to validate the robustness of the deconvolution method.

The SEC-SAXS data used are of three different NaCl concentrations, 140 mM, 500 mM and 1000 mM. The initial parameters were predicted using the SAXS elution profiles $I_{\mathrm{q}}(t)$ at q = 0.0144 Å^{-1}. The NM minimizer was used to derive the optimal parameters of the model adopted with the EMG peak models with *RSS* as the objective function. The SAXS profiles (Fig. 3.32 left) show that the size of the monomeric PDC increases with the increase of the salt concentrations. The characteristic length scale of the detergent layers can be derived from the peak position (q_{max}) at the mid-q range given $2\pi/q_{max}$. The resulting length scale of 42.7 nm for the PDC at the NaCl concentration of 140 mM is comparable with the value of 40 mm from solution SAXS experiments examining DDM micelles in the presence of 150 mM NaCl (Lipfert et al. 2007). From 140 mM to 500 mM, the overall size of PDC only increases slightly. The aggregation number of the detergent layers also increases to a certain extent as evidenced by the higher intensity in DF and the shifts of the peak positions towards smaller q. When the salt concentration is increased to 1000 mM, the deconvolution results suggest that the size of the PDC increases significantly. The inspection of the DF region shows that at a relatively high salt concentration DDM tends to exhibit larger aggregation numbers and forms a more pronounced ellipsoidal corona. A comparison $a1(q)$ of f_3 trajectories (Fig. 3.32 right) also indicates that the sizes of empty micelles are notably larger at the higher salt concentration. Notably, similar trends along q are observed, which implies the parameter instability may be introduced by the deficiency of a single EMG component to describe the rather broadly distributed empty micelles.

The outcome of this study is in accordance with the experimental observations and leads to the recommendation that for SEC-SAXS datasets the use of the EMG model, NM minimizer and objective function *RSS* is effective to deconvolute the SAXS profiles of monomeric PDCs.

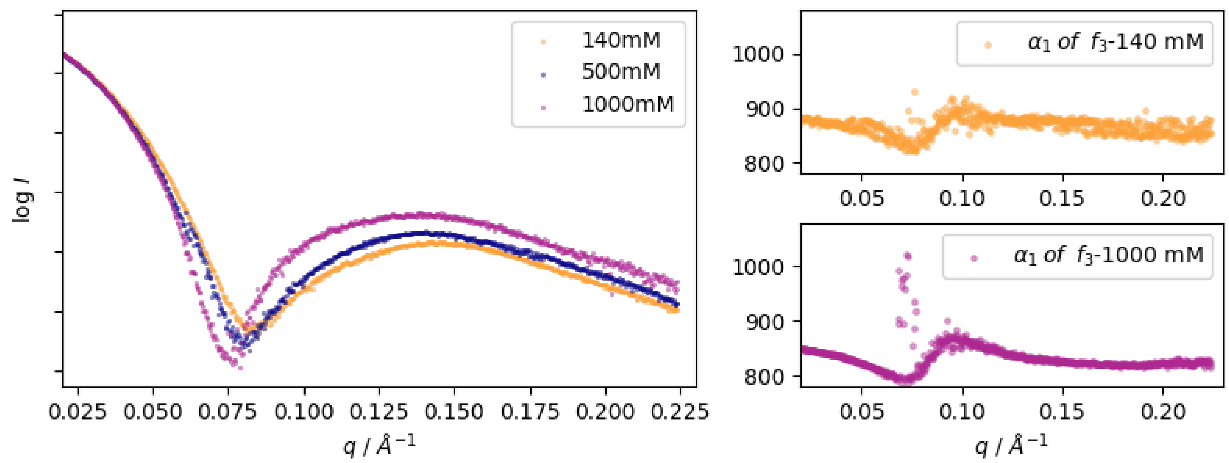


Figure 3.32 (left) The deconvoluted SAXS profiles of the monomeric PDC at the NaCl concentrations of 140mM (yellow), 500mM (blue) and 1000mM (pink). The sizes of PDC increase with the increase of salt concentration. The R_g are 48.7 Å, 49.3 Å and 50.1 Å for the NaCl concentration of 140 mM, 500 mM and 500 mM, respectively. The characteristic length scale of the detergent layers are 42.7 Å, 44.6 Å and 46.5 Å, respectively. (right) A comparison between the parameter trajectories of $a1(q)$ of f_3 for the deconvolution results at the NaCl concentrations of 140 mM and 1000 mM. The size increase of the empty micelles is supported by the shifts of $a1(q)$ of f_3 towards the earlier elution time.

3.4.4 Conclusion

The aim of this section is to investigate the effects of minimizer and objective function selection and to arrive, if possible, at some recommendations for this selection which will reduce the apparent subjectivity involved. The main conclusion of the study is a demonstration of the inadequacy of minimizers to serve as a complete tool for optimization of deconvolution of SEC-SAXS data from PDC at any resolution. All the minimizers fail to complete the task in the VLQ region. However, it appears universally robust results are obtained by using the EMG model, NM minimizer and *RSS* as the objective function. The results obtained are limited by the type of data and models included in the study but it is expected that for SEC-SAXS datasets similar studies using other membrane proteins or other peak models will yield similar conclusions. With this study as a starting point, further improvements allowing a search for the global optimum can be done by fine tuning the parameters of the minimizer (Beiranvand et al. 2017) or introducing Monte-Carlo and Tabu-based algorithms (Lasdon et al. 2010; Wales & Doye 1997; Xiang & Gong 2000).

3.5 Summary

This chapter, I have described two methods targeting the problems existing in the kinetic and the static SAXS studies. The interpretation of the data from the above two types of investigations, time-resolved SAXS and the SEC-SAXS of membrane proteins, is a non-trivial task. This is mostly because the information contained is convoluted. Often than not, the procedures of deconvolution involve subjectivity. In §3.2, by asking the question "at which time point do the data profiles change most", a model-free method based on first-hand data is introduced to provide an objective impression of the crucial transitions in time-resolved SAXS experiments. In §3.3, by asking the question, "which method can make the consequential result the most comprehensible", a recommendation of approaches is proposed to reduce the subjectivity involved in choosing an optimization method for deconvolution of SEC-SAXS datasets of membrane proteins. The procedures developed contribute to setting up guidelines for researchers to better interpret their SAXS experimental data. However, it must be emphasized that by no means the above conclusions are the universal solution for all cases. The user should carefully report how the input data are generated and validate the suggested outputs by alternative means.

Chapter 4

Towards Better Modelling

4.1 Introduction

To build a convincing model is arguably one of the most important goals for structural biologists. Structural modelling based on solution SAXS data generates three-dimensional descriptions of hydration envelopes for macromolecules (Vestergaard 2016). Recently, SAXS modelling for bio-macromolecules has been widely employed in many novel structure-related applications such as quaternary structure analysis (Jiménez-García et al. 2015; Schindler et al. 2016), determination of equilibrium mixtures (Blobel et al. 2009; Cheng et al. 2017), supporting crystallographic phasing (Ayyer et al. 2016) and flexibility analysis (Bernadó et al. 2010; Tian et al. 2015).

Due to the isotropic and diluted nature of the particles in the solution environment, *ab initio* SAXS modelling is subject to large ambiguity. To perform SAXS-based modelling to an advanced level one must include prior knowledge. One very important example of such prior knowledge is the atomic structural information from an MX experiment. The atomic resolution details can either be used as a rigid-body entity to conjugate intricate multi-domain proteins or be exploited to help understand the dynamics of the solution state. For such analyses, the keys are 1) to calculate the theoretical scattering profiles based on the high-resolution structure, 2) to generate a pool of reasonably manipulated high-resolution structure and from that, 3) to select the one(s) that fit the data best. While the first part has many implementation (e.g. CRYSOL (Svergun et al. 1995), FoXS (Schneidman-Duhovny et al. 2016), SASTBX (Liu et al. 2012), Pepsi-SAXS (Grudinin et al. 2017)), the methods for generating the plausible pool and further selection need to be developed on a case-by-case basis. In this chapter, I describe two methods enabling model generation based on the pre-available MX structures with corresponding case studies to show that, by properly conducting SAXS-oriented modelling, one is able to validate or elaborate on structural hypotheses.

4.2 Contributors to this Chapter

Dr. Anna Polyakova (University of Leeds, UK) provided the SEC-SAXS dataset in §4.3. She conducted the SEC-SAXS experiments on BL4-2 at SSRL (Carlifornia, USA). Dr. Chris Orr (University of Southampton, UK) provided the X-ray crystallographic models in §4.4. Data were collected on I04-1 at DLS (Oxford, UK). The molecular dynamics simulation was run by Hayden Fisher (University of Southampton, UK). Dr. Ivo Tews (University of Southampton, UK) coordinated the Fab project presented in §4.4.

The results have been reported in the following manuscripts:

Mark Cragg, Patrick Duriez, Hayden Fisher, Yunyun Gao, Chris Orr, Arwen Pearson, Ann White. A covalent activity switch mechanism in IgG2 antibodies. *Manuscript in preparation for submission*

4.3 Modelling Protein-Detergent Complexes

4.3.1 Background

SAXS can be used to determine the overall hydrated protein shape. The reconstructed low-resolution shapes are usually described as a rigid-body with a near-uniform electron density. This assumption is true for soluble proteins. However, integral membrane proteins (IMP) were believed for a long time to be difficult to study using SAXS, as the agents used to solubilise the membrane protein disturb the basic homogeneity assumption of solution scattering. The complex buffer conditions needed for stabilising membrane proteins leads to non-trivial background matching. The protein-detergent complex (PDC) has layered electron density contrasts which make *ab initio* reconstruction difficult. Nevertheless, even if it is possible to recover the shape information of a PDC, molecular biologists would probably still ask why one should care about the structure of the PDC. The short answer is that the commonly used schematic cartoon of the PDC is incorrect. This is partially revealed through the structural investigation of detergent micelles by SAXS (Ivanović et al. 2019; Lipfert et al. 2007). Instead of the typical sphere representation often used, the experimental observation of ellipsoid micelles may indicate a bilayer can form in the central part of a detergent micelle. To date a detailed investigation of PDC structure has yet to be done. Molecular dynamics simulations give results that do not match well to the experimental data, especially at high salt concentration (Neale et al. 2013). A second reason why knowledge of the PDC structure is important for crystallographers is that the structural constitution of PDC plays an important role in getting a crystal. The majority of structures currently being determined utilize direct crystallization of PDC and so detergents will be directly packed into the lattice. The stacking pattern of the PDC may facilitate PDCs assembling into a specific crystal system (Loll 2014). The lack of information on detergent-mediated oligomerization means the crystallization of membrane proteins relies heavily on a process of trial and error.

MEMPROT is the most recent approach targeting PDC reconstruction (Pérez & Koutsioubas 2015). The software builds a space-filling model of a coarse-grained detergent corona around an all-atom model of the IMP. The search is based on the evaluation of goodness-of-fitting between the theoretical SAXS curves and the experimental SAXS profiles. Using a coarse-grained method to model the detergent corona is an excellent idea because the detergent layer has only long-range order (i.e the characteristic length of the head-to-head group). However, the implemented algorithm in MEMPROT is only really applicable to proteins with geometrically isometric transmembrane regions such as ß-barrels (Fig. 4.1). Another drawback is that the searching can only be done in a brute-force way.

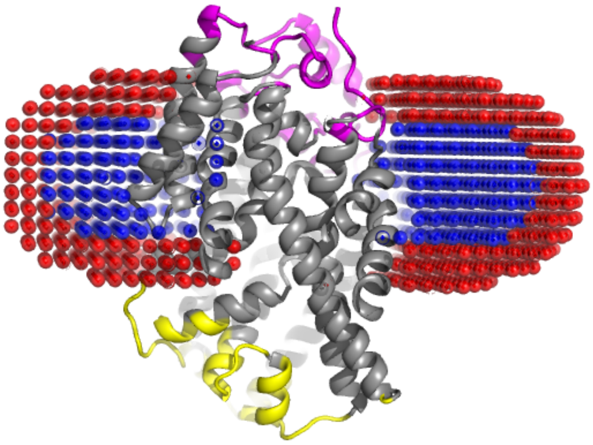


Figure 4.1 A PDC model of an anisometric IMP suggested by the MEMPROT algorithm. It returns a clear false-positive. Physically, the protein placement is not entirely correct as the hydrophobic region (grey) is not properly wrapped by the detergent corona.

The considerations to improve the modelling effectiveness of membrane protein include three aspects:

1) implement an envelope based detergent corona placing algorithm. This facilitates the modelling of asymmetrical IMPs and will not shade the actual cavity in certain conformations.
2) Enable the sampling of the orientational states of proteins and the shape of the detergent corona at the same time.
3) apply a proper parameter searching algorithm.

4.3.2 Algorithm

4.3.2.1 Convex Hull

In geometry, the convex hull is the smallest convex set that contains a finite set of points. Mathematically, it is a set of points P in n-dimensions that is the intersection of all convex sets containing P. For N points $p_1,\dots, p_N$, the convex hull $Conv(P)$ is defined as,

$$Conv(P) \equiv \{\sum_{j=1}^{N} \lambda_j p_j : \lambda_j \geq 0 \; for\; all\; j\; and\; \sum_{j=1}^{N} \lambda_j = 1\} \tag{4.1}$$

Convex hull is a fundamental construction for computational geometry and has applications in many fields, such as file searching, cluster analysis, collision detection, image processing and crystallography (Aurenhammer 1991).

A robust three-dimensional implementation of computing the convex hull was originally given by O'Rourke (O'Rourke 1998), which has complexity $O(N^2)$. An algorithm named Quickhull reduces the complexity to $O(N \log r)$ for three-dimensional problems, where r is the number of output points (Barber et al. 1996). Up till now, Quickhull has been one of the most cited algorithms with regards to the computation of convex hulls (*Convex hull algorithms - Wikipedia*). Quickhull represents a convex hull with a set of facets and a set of adjacency lists recording the neighbors and vertices for each facet. Specifically in R^3, the facets are triangles and the boundary elements of a facet are edges. Quickhull has two geometric operations: oriented planes through three points, and signed distance to the plane. It represents a plane by its outward-pointing unit normal and its offset from the origin. The signed distance of a point to a plane is the inner product of the normal point and normal plus the offset. The plane defines a halfspace of points that have negative distance from the plane. If the distance is positive, the point is called above the plane.

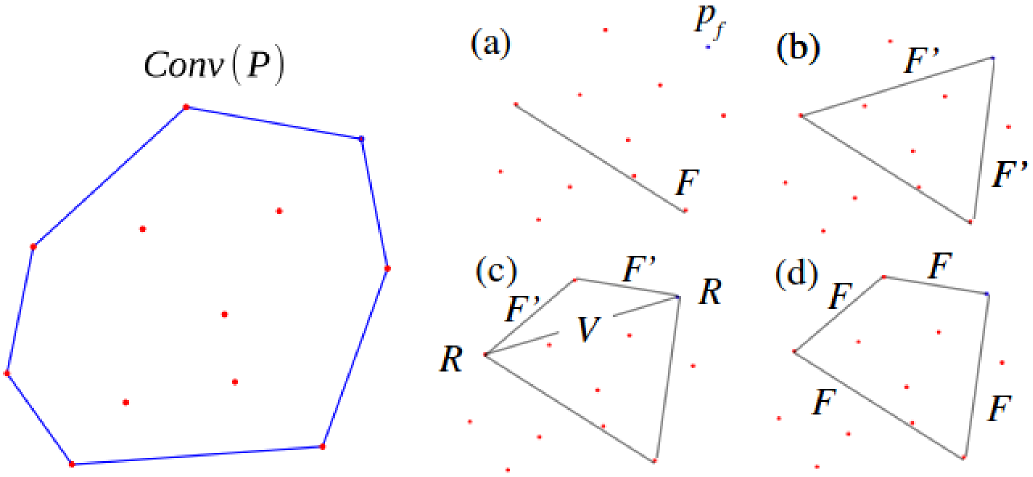


Figure 4.2 A two-dimensional demonstration of the Quickhull algorithm. (Left) the convex hull (blue edges) of the given points set P (red dots). (Right) The basic steps of Quickhull algorithm. (a) initialization, (b) partition and construction of new facets, (c) determination of visible set, (d) deletion of visible set and repartition.

In three-dimensional space (R^3), the convex hull is returned as a set of signed triangular vertices with the information of neighboring facets. Quickhull also repairs any faults appearing in the resulting convex hull: more than two facets meeting at a edge, a facet contained in another facet, a facet with fewer than three neighbors, a facet with flipped orientation, a newly processed point that is coplanar with an horizon facet, coplanar facets and redundant vertices. With this thorough trouble-shooting process, the appearances of cracks or very thin facets is unlikely.

4.3.2.2 Inclusion Test

An inclusion test in R^3 is a test on whether a certain point lies inside of a given polyhedron. The inclusion test is one of the basic primitive tests in computational geometries and has many applications in areas such as computer graphics, topological information and real-time collision detection (Ericson 2004). There are many algorithms for solving inclusion problems. Some well-known cases are the ray-tracing method, winding number method, solid BSP tree algorithm and GJK method (Ericson 2004; Foley et al. 1995; Gombosˇi & Žalik 2005; Heckbert 1994). If a polyhedron is given as a surface mesh or as an intersection of halfspace, the ray-tracing method can be applied. This method is based on the theorem that when casting a ray from the tested point in any direction, if an odd number of facets is intersected, the point lies inside the polyhedron; otherwise the point lies outside it (Ericson 2004). The number of intersections is called the crossing number (*cn*). The convex hull calculated with the Quickhull algorithm is a robust triangulated surface. It is thus possible to use a ray-tracing algorithm for the inclusion test upon a set of points and the derived convex hull.

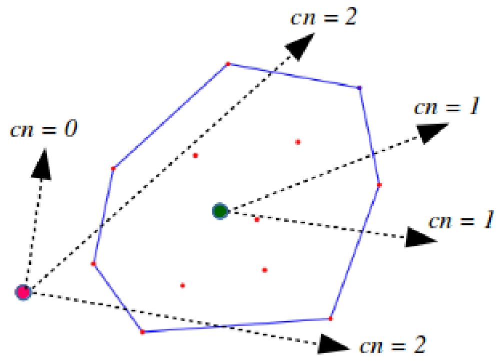


Figure 4.3 A two-dimensional demonstration of the ray-tracing method. A polygon boundary separates the points in the plane to its outside and inside. Shoot a ray from the tested point in an arbitrary direction. For each ray there is a crossing number (*cn*) of intersections. The outside points (red dot) have a *cn* of even numbers; the inside points have a *cn* of odd numbers (green dot).

4.3.2.3 Ray-Triangle Intersection

Obviously, the intersection test is the core of the inclusion test. A Plücker test is an intersection test that takes advantage of the algebraic properties of Plücker coordinates (Ericson 2004; Shevtsov et al. 2007). Instead of using barycentric coordinates, the Plücker test relies on testing the relation between a ray and the triangle edges, with which the intersection test is able to be quickly performed.

Plücker coordinates, introduced by Julius Plücker in the 19th century, are an alternative way to assign sex homogeneous coordinates to each line in R^3. Each directed line in three-dimensions can be rewritten in a six-dimensional Plücker space. Given two three-dimensional points a and b, the corresponding vector $\boldsymbol{L}$ in Plücker space is defined as,

$$\boldsymbol{L} = (a - b, a \times b) \tag{4.2}$$

where × refers to a cross product.

A ray $\boldsymbol{R}$ with origin o and unit direction $\boldsymbol{d}$ can be defined as,

$$\boldsymbol{R} = (\boldsymbol{d}, \boldsymbol{d} \times o) \tag{4.3}$$

For two given vectors $\boldsymbol{L0} = (\boldsymbol{u0}, \boldsymbol{v0})$ and $\boldsymbol{L1} = (\boldsymbol{u1}, \boldsymbol{v1})$, define the inner-product operation as,

$$\boldsymbol{L0} * \boldsymbol{L1} = \boldsymbol{u0} * \boldsymbol{u1} + \boldsymbol{v0} * \boldsymbol{v1} \tag{4.4}$$

where * refers to a dot product.

If the result of the inner-product operation equals 0, the vectors intersect, while a negative/positive result indicates ***L1*** passes ***L0*** from the far/near neighbor.

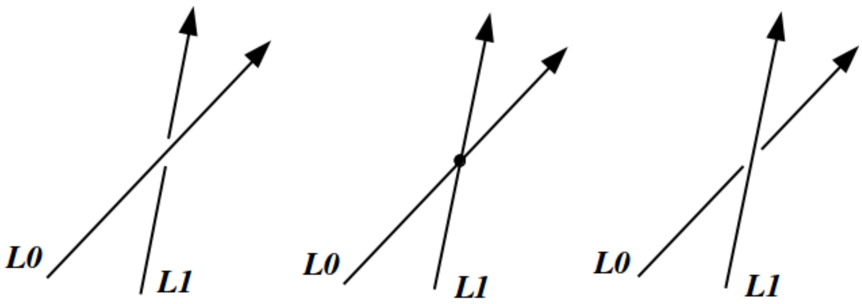


Figure 4.4 An inner-product operation in Plücker space has three different results depending on their signs. (Left) If ***L1*** passes ***L0*** from the far neighbor the inner-product operation

returns a positive scalar number. (Middle) If ***L1*** intersects with ***L0*** the inner-product operation returns 0. (Right) If ***L1*** passes ***L0*** from the near neighbor the inner-product operation returns a negative scalar number.

Given a triangle ***T*** with three edges ***E0, E1*** *and* ***E2*** and a ray ***R*** the criteria of a ray-triangle intersection (denoted as $\boldsymbol{R \bigcap T}$) Plücker test is defined as, a ray intersects the triangle if the inner-products between the ray and all edges of the triangle have the same sign.

$$\text{for } t0 = \boldsymbol{E0} * \boldsymbol{R},\ t1 = \boldsymbol{E1} * \boldsymbol{R},\ t2 = \boldsymbol{E2} * \boldsymbol{R}$$
$$\boldsymbol{R \bigcap T} \ \text{ iff. t0, t1, t2} > 0 \text{ or t0, t1, t2} < 0 \tag{4.5}$$

It is worth mentioning that the inner-product operation uses only multiplications and additions, which means single-precision floating-point arithmetic is sufficient for performing the Plücker test.

4.3.2.4 the Single Instruction Multiple Data implement of the Plücker-Based Inclusion Test

Although the basic idea of ray-tracing is simple, the computation of ray-triangle intersections may not be efficient if every facet is tested for each point separately. It is obvious that the Plücker test can be performed in a way that multiple floating points are operated in parallel since it uses only multiplications and additions. The Single Instruction Multiple Data (SIMD) architecture of modern processors allows a common operation such as adding and multiplying the same value to a large number of data points to be applied to all the data points at the same time. Inspired by the work of Shevtsov et al. (Shevtsov et al. 2007), an SIMD implication of the Plücker test was designed.

Denotations

(1) $A[i,...,j]$: take the $(i,\ ...,\ j)$th row from an arbitrary matrix A

(2) **sort**(*vec*): sort a vector *vec* by the magnitudes of normal projections in the ascending order

(3) **indicessort**(*vec*): record the positional indices of *vec* in the sorted order

(4) (*vec1*, *vec2*): compose *vec1* and *vec2* into a matrix

(5) **det**(*A*): determination of a squared matrix *A*

(6) <*vec1*, *vec2*>: the inner product operation of two vectors *vec1* and *vec2*

(7) $A1 \odot A2$: Hadamard product of two matrices *A1* and *A2*

(8) *A1* 🚫 *A2*: Hadamard division of two matrices *A1* and *A2*

(9) $\Sigma(A1 \odot A2)$: the inner product of each rows of two matrices *A1* and *A2*

(10) A>=0: return a Boolean matrix recording the truth value of each element in *A* larger than 0

(11) *A1* **&** *A2*: element-wise logical AND of two matrices *A1* and *A2*

```
algorithm SIMD_pluecker_inclusion{
INPUT Qhull, a set of signed triangular vertices returned by the Quickhull algorithm
        $N$ x (3 x 3) matrix $Tri$ := $Qhull$ , $N$ is the number of facets

        ## rewrite the triangle from the three vertices to a point plus two vectors
        $Tri$ :=  ($a$, $vec1 = a - b$, $vec2 = a - c$) for all elements ($a$, $b$, $c$) in $Tri$

        ## rewrite the triangle in Plücker space
        # calculate the facet normal
        $N$ x 3 matrix $norm\_unsorted$ :=
                $vec\_un = (vec2 \times vec1)$ for all elements in $Tri$

        # sort each element in norm_unsorted and record the positional indices of norm_unsorted
        $N$ x 3 matrix $norm\_sorted$ := $vec\_sort$ = sort($vec\_un$) for all $vec\_un$ in $norm\_unsorted$
        $N$ x 3 matrix $norm\_sorted\_arg$ :=
                $vec\_indsort$ = indicessort($vec\_un$) for all $vec\_un$ in $norm\_unsorted$

        # normalise each element in norm_sorted by its largest normal projection
        # and reduce the dimension of each element
        $N$ x 2 matrix $norm$ :=
                $vec\_n = (vec\_sort\,[1, 2]/vec\_n[3])$

        # order the components of each element in Tri in the order of norm_sorted_arg
        # and reduce the dimension of the components
        $N$ x 2 matrix $facets\_a$ :=
                $a\_sorted = a[vec\_indsort][1, 2]$ for all $a$ in $Tri$
        $N$ x 2 matrix $facets\_vec1$ :=
                $vec1\_sorted = vec1[vec\_indsort][1, 2]$ for all $vec1$ in $Tri$
        $N$ x 2 matrix $facets\_vec2$ :=
                $vec2\_sorted = vec2[vec\_indsort][1, 2]$ for all $vec2$ in $Tri$
        # calculate the determination of facets_vec1 and facets_vec2
        $N$ x 1 vector $facets\_det$ :=
                $det$ = det($vec1\_sorted$, $vec2\_sorted$)
                for all ($vec1\_sorted$ and $vec2\_sorted$)  in ($facets\_vec1$ and $facets\_vec2$)
```

```
    # calculate the projection of point a on the normal for each facet
    # and normalize it by the largest projection of that normal
     N x 1 vector facets_proj :=
            a_proj_vec_n = <a_sorted, vec_n>
            for all (a_sorted, vec_n) in (facets_a, norm)

    # rewrite facets_vec1 and facets_vec2 according to Cramer's rule
    N x 2 vector facets_vec1_det :=
            vec1_det = $(vec1\_sorted\ \ *[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}])/det$
            for all (vec1_sorted , det) in (facets_vec1, facets_det)
    N x 2 vector facets_vec2_det :=
            vec2_det = $(vec1\_sorted\ \ *[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}])/det$
            for all (vec2_sorted, det) in (facets_vec2, facets_det)

    ## end of rewriting the triangle in Plücker space

INPUT a M x 3 matrix P (a set of three-dimensional coordinates of points being tested)
    ## intersection test
    # sort all the points for each facet and reduce dimension
    M x  N x 2 matrix Ori :=
            for all p in P {
                    N x 2 matrix origin = p[vec_indsort][1, 2]
                    for all vec_indsort in norm_sorted_arg
                    }

    # cast a random ray direction d (i.g. point to the +x axis d=(1, 0, 0))
    # then sort and reduce dimension
    N x 2 matrix D :=
            dir = d[vec_indsor][1, 2] for all vec_indsort in norm_sorted_arg

    # calculate the scaled projection of the rays on each facet normal
    M x N matrix T_mat  :=
            for all origin in Ori {
                    N x 1 vector t =  (facets_proj − Σ(origin ⊙ norm)) 🚫 Σ(D ⊙ norm)
            }

    # calculate the scaled intersection points of the rays on each projection plane of facet
    M x N x 2 matrix C :=
            for all origin in Ori {
                    N x 2 matrix c = origin + $[\begin{smallmatrix} t \\ t \end{smallmatrix}]$ ⊙ D −  facets_a

            }

    # perform the inner-product operations of the Plücker test
     M x N matrix U_mat :=
            for all c in C {
                    N x 1 vector u = Σ((c  *$[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}]$⊙ facets_vec1_det)
            }
    $M$ x $N$ matrix $V\_mat$ :=
            for all $c$ in $C$ {
                    $N$ x 1 vector $v = \Sigma(c \odot facets\_vec2\_det)$
            }

    # intersection cretia. Every intersection will be assigned as 1 (true)
    $M$ x $N$ boolean matrix $Intersection$ :=
            ($U\_mat$ >= 0) & ($V\_ma\ t$>= 0) & (($U\_mat$+$V\_mat$) <= 1) & ($t\_mat$ >= 0)

    ## end of intersection test

    ## inclusion test
            $M$ x 1 vector $number\_of\_Intersection$ :=
                    summation over the $N$ column of $Intersection$
            $M$ x 1 Boolean vector : $bool\_points\_inside$ :=
                    $number\_of\_Intersection$ % 2 == 0
            $M$ x 1 Boolean vector : $bool\_points\_outside$ :=
                    $number\_of\_Intersection$ % 2 == 1

    ## end of inclusion test
}
```

It is worth mentioning that the operation "**for all … in …**" maps to the SIMD vectorization by many popular scientific computing languages such as Fortran, Matlab, Julia, Wolfram language and Python if a proper data structure is designed.

4.3.2.5 Constructing Convex Hulls for IMP

Parsing the MX structural information. The Protein Data Bank (PDB) format provides a standard representation for macromolecular structure data derived from MX and NMR studies. The structural information can be processed in a hierarchical configuration. In this study, the biopython.structure package is adapted (Cock et al. 2009). The top level class is *structure*, followed by *chain* and *residue* and the bottom level is the *atom*. A principal component analysis (PCA) is conducted on the *structures* level in order to get the three principal components of atom coordinates. The major principal (the component with the largest Euclidean norm) is set to point towards the inner side of the membrane. Three principal components are arranged to fulfil the right-handed coordinates.

Hydrophobic boundaries Assignment. The PDB of an IMP is sent to the Protein Positioning in Membranes (PPM) server to calculate the hydrophobically embedded residues using the Orientation of Proteins in Membranes (OPM) algorithm (Lomize et al. 2012). The output contains the indices of the embedded residues. With this list of residues, the outer, the transmembrane and the inner regions are then assigned to the relative *residues* correspondingly. Useful information such as the tilt angle and hydrophobic depth are also predicted from the calculation. The *structure* is rotated according to this tilt angle so that the hydrophobic core boundaries are parallel to the *xy*-plane.

Reconstruction of protein envelope. The atomic coordinates of the inner membrane region, the transmembrane region and the outer membrane regions are taken as three entities of point clouds. The convex hulls of the three regions are calculated separately according to the method described in §4.2.2.1.

4.3.2.6 Constructing Coarse-Grained Detergent Corona

The detergent layer is modelled as a three-dimensional FCC lattice of pseudo atoms with various geometries. The lattice parameter, *a*, and the type of atoms are initially set according to the electron density of the detergent (Lipfert et al. 2007; Mo et al. 2008). Reference values for the common detergents are listed in Table 4.1.

Computationally, the detergent lattice is first generated in a bounding volume with its thickest direction aligned with the z-axis. A mask is then applied according to the analytical description of the expected geometrical shapes to exclude the points sitting outside of the shapes. Then the detergent lattice is placed at the position where the mass centers of the detergent atoms and the IMP coincide. The initial thickness of the corona is set to the value of the characterized length scale of the detergent corona (*d,* defined in §3.3.3). The initial double length of the shell layer is derived by *d* minus the hydrophobic depth calculated by OPM. For other geometrical parameters one can refer to the following works (Berthaud et al. 2012; Chen & Hub 2015; Kunji et al. 2008; Lipfert et al. 2007; Mo et al. 2008; Neale et al. 2013; Oliver et al. 2013; Yang et al. 2014). The atoms colliding with the protein envelope (pass the inclusion test mentioned in §4.2.2.1) are deleted at the end. In the output PDB file, the shell layer is assigned to the chain identifier “S” and the core layer to “C”.

Table 4.1 The suggested initial value of *a* by the types of detergents: *n*-decylphosphocholine (FC-10), *n*-dodecylphosphocholine (FC-12), *n*-nonyl-*β*-*D*-maltoside (NM), *n*-dodecyl-*β*-*D*-maltoside (DDM), *n*-octyl-*β*-D-glucoside (OG), *n*-nonyl-*β*-*D*-glucoside (NG), 1,2-dihexanoyl-*sn*-glycerophosphocholine (DHPC) and 1-palmitoyl-2-hydroxy-*sn*-glycero-3-[phospho-*rac*-(1-glycerol)] (LLPG)

	***ρ*-core**	**atom type**	***a* (Å)**	***ρ*-shell**	**atom type**	***a* (Å)**
	(e/Å^3)			**(e/Å^3)**		
FC-10	0.273	CD2/LEU	3.2	0.490	NZ/LYS	3.0
FC-12	0.277	CD2/LEU	3.1	0.490	NZ/LYS	3.0
NM	0.272	CD2/LEU	3.2	0.520	NZ/LYS	2.8
DDM	0.277	CD2/LEU	3.1	0.520	NZ/LYS	2.8
OG	0.268	CD2/LEU	3.2	0.45-0.54	NZ/LYS	3.1-2.9
NG	0.271	CD2/LEU	3.2	0.50-0.53	NZ/LYS	2.7-2.8
DHPC	0.253	CD2/LEU	3.2	0.464	NZ/LYS	3.1
LPPG	0.281	CD2/LEU	2.8	0.46-0.48	NZ/LYS	3.1

4.3.2.7 Construction of *SO*(3) grid

A theoretical scattering prediction for the initial PDC model will likely fail to fit the experimental data. This is not only because the detergent corona varies in geometrical shape (Pérez & Koutsioubas 2015) but also because the theoretical orientation may deviate from the truth. Hence, the candidate pool should be generated by adjusting the shape of the detergent corona as well as the orientation of the IMP in the detergent layer. The latter can be achieved by creating a *SO*(3) transform grid, a space of 3D rotation. *SO*(3) is diffeomorphic to the 3-*sphere* S^3, embedded in R^4. Euler angles are often used to represent rotations. Each rotation is a vector (θ, φ, ψ) with the limits [-π, π]. The topology of the resulting space is S^1xS^1xS^1. Hence Euler angles do not correctly capture the structure of *SO*(3). Sampling *SO*(3) using Euler angles may lead to failure in producing deterministic rotations (Yershova & LaValle 2004). Because of the topological relationship between the 3-*sphere* and *SO*(3), hyperspherical coordinates can instead be used for *SO*(3) (Yershova et al. 2010). Consider a point (α, β, γ) in S^3. For each γ, the full ranges of α and β define a 2-s*phere*. Let $\gamma \in [0, \pi]$ be the angle parametrizing the circle, S^1,and (α, β), $\alpha \in [0, \pi]$, $\beta \in [0, 2\pi]$ be the spherical coordinates parametrizing the sphere, S^2. It is possible to construct the *SO*(3) into a grid in the space S^2xS^1 using the quaternion $x = (x0, x1, x2, x3)$ corresponding to (α, β, γ) based on the formula:

$$x0 = \cos\gamma,\ x1 = \sin\gamma,\ x2 = \sin\gamma\sin\alpha\cos\beta,\ x3 = \sin\gamma\sin\alpha\sin\beta \qquad (4.6)$$

This quaternion can be used to derive the deterministic rotation matrix.

A quasi-uniform angular grid on the 2-*sphere* can be generated using the following algorithm (Svergun 1994),

$$\begin{gathered} \alpha_i = \arccos[1 - 2(i - 1)/f_k] \\ \beta_i = 2\pi\{[(i - 1) + f_{k-1} \% f_k]\}/f_k \\ i = 1,\dots, f_k + 1 \end{gathered} \qquad (4.7)$$

where f_k is the k-th Fibonacci number defined as $f_1 = f_2 = 1$. The number of points (directions) for (α, β) on the grid over the 2-*sphere* are hence the corresponding k-th Fibonacci number .

The next problem is the resolution of γ. According to Yershova et al. (Yershova et al. 2010), the relationship between the number of points of the grids between the circle and the 2-*sphere* is given by

$$\frac{\pi}{\#\_on\_circle} = \sqrt{\frac{\pi}{\#\_on\_sphere}} \qquad (4.8)$$

For example, by choosing f_{11} , 145 points are generated over the 2-*sphere* of (α, β) and therefore lead to a discretization of size 21 over ψ. In total, 21 x 145 = 3045 unique rotations will be created.

4.3.2.8 Parameter Inference

The parameter space is configured by the geometrical parameters of the detergent corona, the lattice parameters for the shell and core layers, and the orientation of the protein. The parameter optimization is carried out using the Powell minimization implemented in the scipy package (Ivanović et al. 2019; Virtanen et al. 2020). The theoretical scattering profile is calculated using CRYSOL (Svergun et al. 1995). The objective function is the reduced χ^2 which can be directly read from the CRYSOL output.

4.3.3 Results and Discussion

4.3.3.1 Construction of the IMP Envelope

The outward-facing conformation (PDB code: 2jln) was used for demonstrating the envelope reconstruction. The PDB file was fetched from the Protein Data Bank in Europe (PDBe) (Velankar et al. 2011) and sent to the PPM server for OPM calculation. The embedded subunits were assigned to following residues:

Embedded 27-49, 52, 58-81, 89-95, 97-99, 101-137, 139-184, 194,198, 208-231, 249-271, 283, 296-327, 329-330, 333-356, 359-380, 382-383, 404, 406-445, 447-448.

The periplasmic and the cytoplasmic regions were calculated according to the hydrophobic boundaries and assigned to the following residues, respectively:

Out 50-53, 54-55, 57, 138, 185-193, 195-197, 199-207, 272-282, 284-295, 357-358

In 8-26, 82-88, 96, 100, 232-248, 328, 331-332, 381, 384-403, 405, 446, 449-471

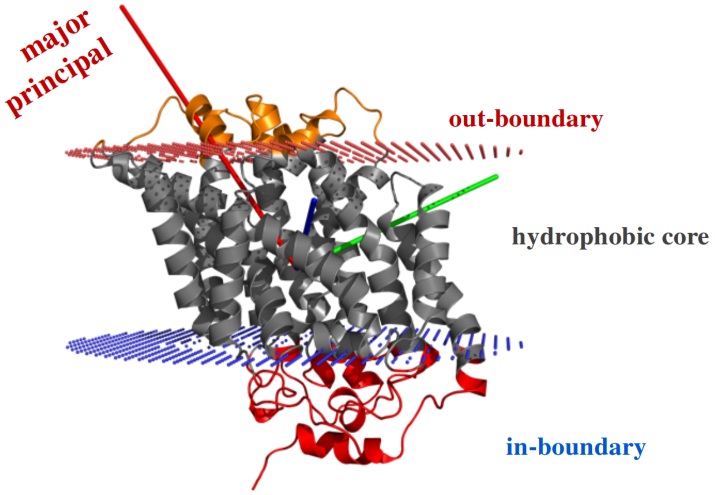


Figure 4.5 The hydrophobic boundaries assigned according to the OPM result for 2jln. The red dots form a layer marking the outer membrane boundary (out-boundary); whereas the blue dots mark the inner membrane boundary (in-boundary). The hydrophobic region of protein is colored in grey. The outer and inner regions are highlighted in yellow and red, respectively. The red axis shows the major principal axis of the protein. To fulfil the right-handed rule, the green axis and the blue axis (behind the protein) are taken as the first and the second principals, respectively.

The visualisation of the envelope is shown in Fig. 4.6.

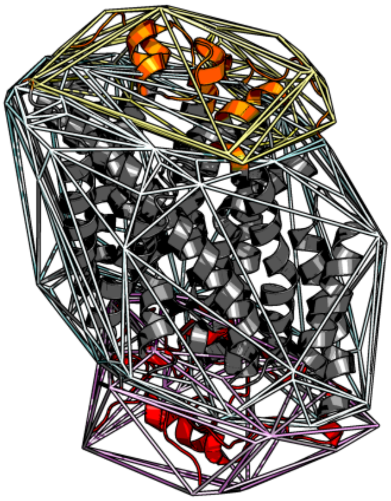


Figure 4.6 The protein envelope constructed based on the convex hulls of the three regions. Overall the convex hulls form a concave shape.

4.3.3.2 Reconstruction of the PDC

The construction of the detergent corona follows the descriptions in §4.2.2.6. The two examples shown here (Fig. 4.7) adopt the geometrical shapes of oblate ellipsoid and oblate spheroid, respectively. It is worth mentioning that, on an Intel Core-i7 machine, the construction with 20946 pseudo-atoms and 162 facets takes $8\text{x}10^8$ clock cycles for the SIMD-implementation whereas for the standard implementation the construction takes $3\text{x}10^{10}$ clock cycles.

The parameter inference is done following the descriptions in §4.2.2.7. The best-fitted model suggests a PDC formed by outward-facing Mhp1 and DDM at a NaCl concentration of 140 mM has an elliptic capsule-like (*Capsule -- from Wolfram MathWorld*) detergent corona (Fig. 4.8). The long-axis is 29.8 Å; the short-axis is 25.2 Å; the half height of the cylinder is 15.2 Å; the length of the shell layer is 6.2 Å. The lattice parameters for the core layer and the shell layer are 3.1 and 2.8, respectively. The χ^2 between the theoretical scattering curve and the experimental data is 1.37.

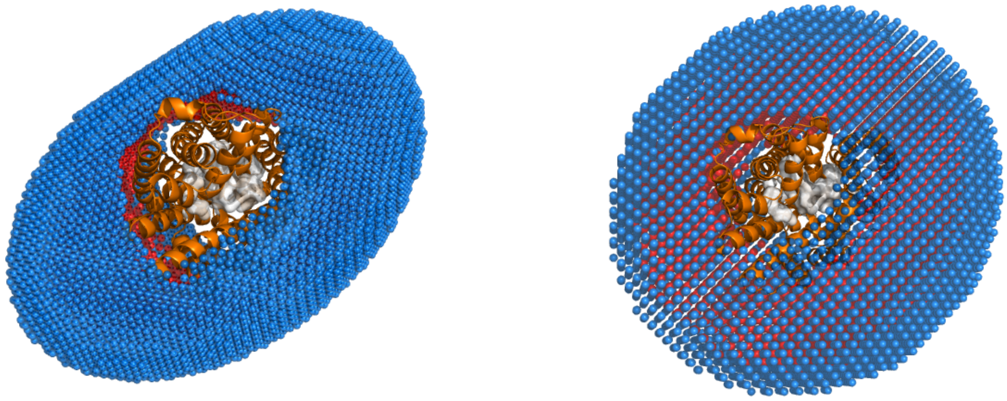


Figure 4.7 Examples of different detergent corona geometries (left) oblate ellipsoid, (right) oblate spheroid. In both cases the cavity of the binding site (grey blobs) remains unobscured.

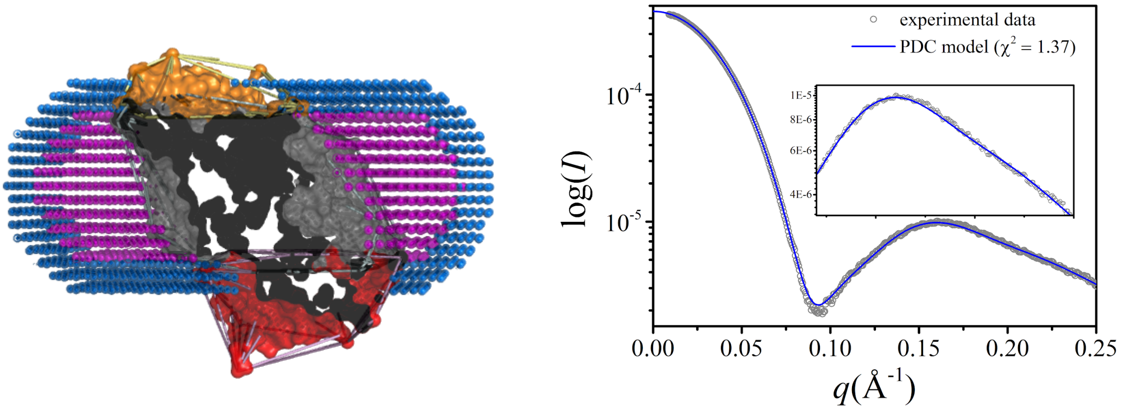


Figure 4.8 (Left) the best-fitted PDC model for the outward-facing Mhp1 (PDB code: 2jln) and a DDM corona with an elliptic capsule-like shape. The shell layer and core layer are highlighted in blue and purple, respectively. A physiologically reasonable result is obtained. (Right) the fitting result of the best-fitted PDC model against the experimental data.

4.3.3.3 The Relationship between Conformational Flexibility and Ligand Binding Affinity

A fluorescence quenching assay can be used to monitor the functionality of Mhp1 as the hydantoin ligand-binding event involves distinct changes in the environment of tryptophan residues (Weyand et al. 2008). A systematic investigation of the ligand-binding affinity of wild-type Mhp1 in the presence of two different detergents (DDM and NM) suggests that the binding affinity decreases with the shorter aliphatic detergent chain (Fig. 4.9) (Polyakova 2015). The reason for this phenomenon is proposed to be due to the fact that *the shorter* *detergents may make the protein more rigid and restrict the conformational changes required to bind ligands efficiently*. Lee et al., on the other hand, conclude that the short chain nonionic detergents destabilize the IMP because the highly mobile detergent molecules form small micelles around the protein which inevitably results in loss of packing interactions (Lee et al. 2016). Orientational flexibility is a way to mimic the thermal motion of an IMP in a PDC. An analysis of major conformers and their orientation in different detergents based on the above model fitting method can be used to give an insight on the effect of detergent size on the functionality of the IMP.

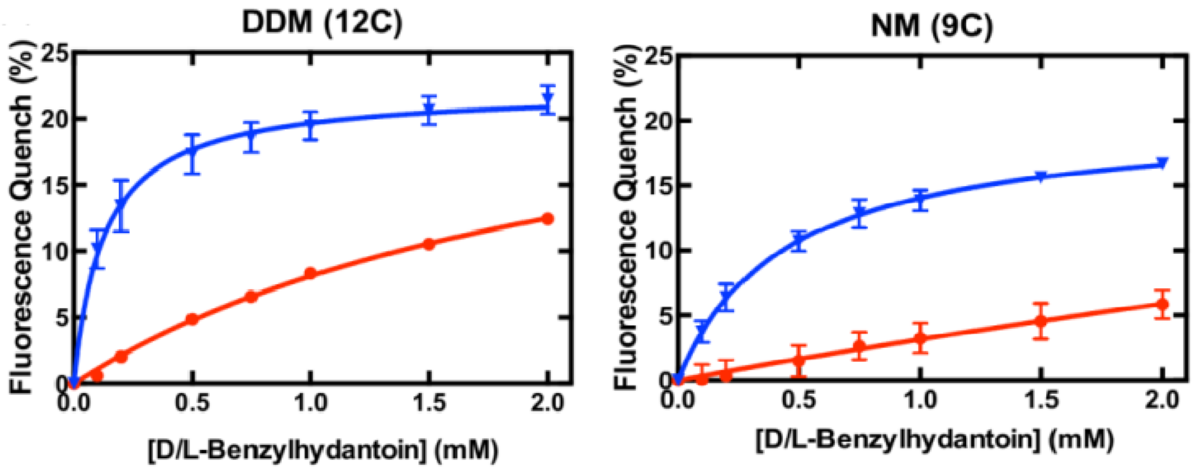


Figure 4.9 The fluorescence quench during titrations of L-BH (0-2 mM) with wild-type Mhp1 solubilised in DDM (left) and NM detergents (right). The titrations were performed in the presence of 0 mM (red) and 15 mM (blue) added NaCl. Mhp1 samples (in 10 mM Tris-HCl pH 7.6, 2.5% *v/v* glycerol, 10 mM DTT, 0.5% *w/v* NM) were diluted to 140 µg/ml using buffers containing 50 mM Tris-HCl pH 7.6, 2% *v/v* DMSO. The figures are courtesy of Dr. Anna Polyakova, University of Leeds.

A set of Euler angles can be converted from the uniform quaternion number. Let the order for Euler angle rotation be intrinsic ZXZ. Then the *tilt angle* (θ), the *spin angle* and the *rotation angle* are defined as shown in Fig. 4.10.

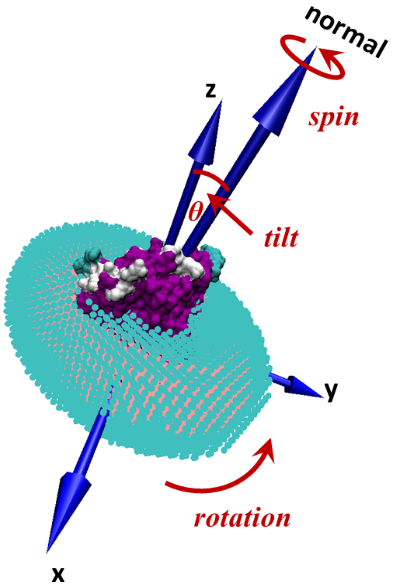


Figure 4.10 The definition of Euler angles. The rotation convention is intrinsic ZXZ. The spin is performed first, which is a rotation around the normal of the hydrophobic boundary by a *spin angle*. The tilt is followed by a rotation around the first principal (showing as vector *x*) of the IMP by a *tilt angle* (*θ*). The final step is represented by a rotation around the tilted major principal (showing as vector *y*) by a *rotation angle*.

Given an experimental SAXS profile, the parameter space of orientations is explored without an explicit goal of attempting to refine the solution. Each orientation is converted to a set of Euler angles as defined above. The distribution of objective function can be plotted as a function of the three Euler angles. This distribution can be understood as the orientational states Mhp1 can explore, where certain conformations are forbidden because the models with these orientations fail in describing the experimental data. The conformational space where better fitting the experimental data occurs is interpreted as representing the preferred orientations. The outward-open conformer in the DDM (Fig. 4.11) has clear preferred orientations. Upon binding of the hydantoin ligand, the occluded conformer shows less flexibility than the ligand-free outward-open form. It seems the hydantoin binding affinity of the Mhp1 solubilized in DDM is not hugely affected, suggesting that the long-chain detergent is a reasonable mimic of the native bilayer environment (Loll 2014). On the other hand, the outward-open conformer in the NM corona (Fig. 4.12) has a far broader distribution with respect to the orientational states. This might be due to the high mobility of the NM molecules (Lee et al. 2016). This is a clear hindrance for the conformer to access a stable native state as the corona is flexible and the PDC is energetically stable even when the protein samples other orientational states. Notebly, in NM, the orientational flexibility of the occluded form is even higher than the outward-open form. This increases the risks of hydrophobic mismatch (Srivastava et al. 2018; van Duyl et al. 2002) may explain the reduced binding efficiency of the protein in this detergent.

In summary, the ligand-binding affinity is affected by the chain length of the non-ionic detergents used to solubilise the IMP. This case study has explored the conformational space Mhp1 can adapt in the NM and the DDM detergent coronas. The results suggest that, instead of making the protein more rigid, shorter-chain detergents promote protein orientational flexibility within the detergent corona and that this adversely affects the binding affinity of Mhp1. Nevertheless, the statement that shorter *detergents restrict the conformational changes* is plausible. However, the reason is not that the NM corona is too compacted but that the NM is a bad mimic of the bilayer environment and therefore cannot stabilize Mhp1.

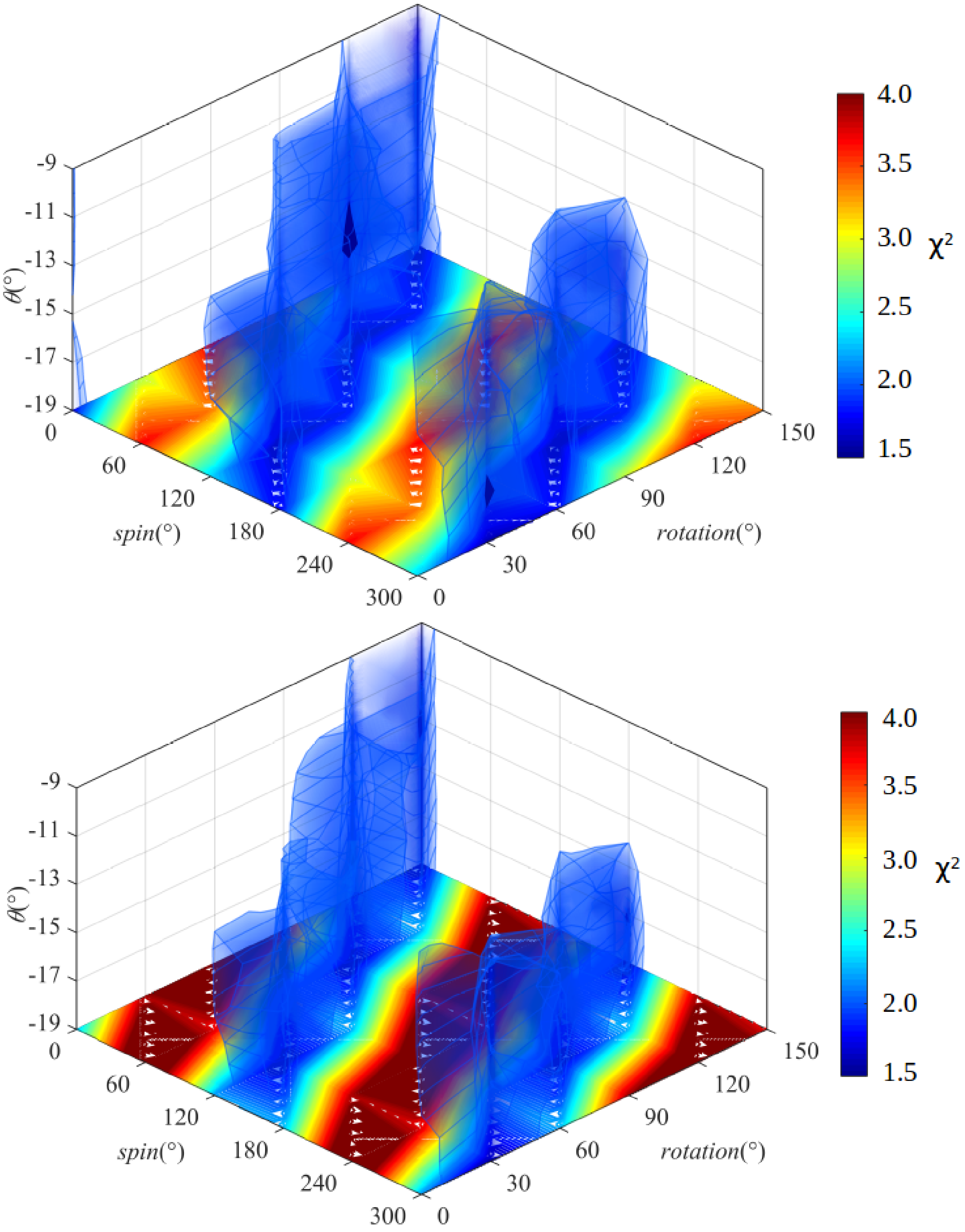


Figure 4.11 χ^2 as a function of three Euler angles for the Mhp1-DDM PDC of (top) outward-open conformer (PDB code: 2jln) and (bottom) occluded conformer (PDB code: 4d1b). The data (SEC-SAXS) were collected at BM29, ESRF, France and P12, Petra III, Germany. Wild-type Mhp1 was solubilized in 10 mM Tris-NaOH pH 8, 2.5% *v/v* Glycerol, 0.05% *w/v* DDM, 140 mM NaCl in the presence of 0 mM (top) and 2 mM (bottom) added L-BH.

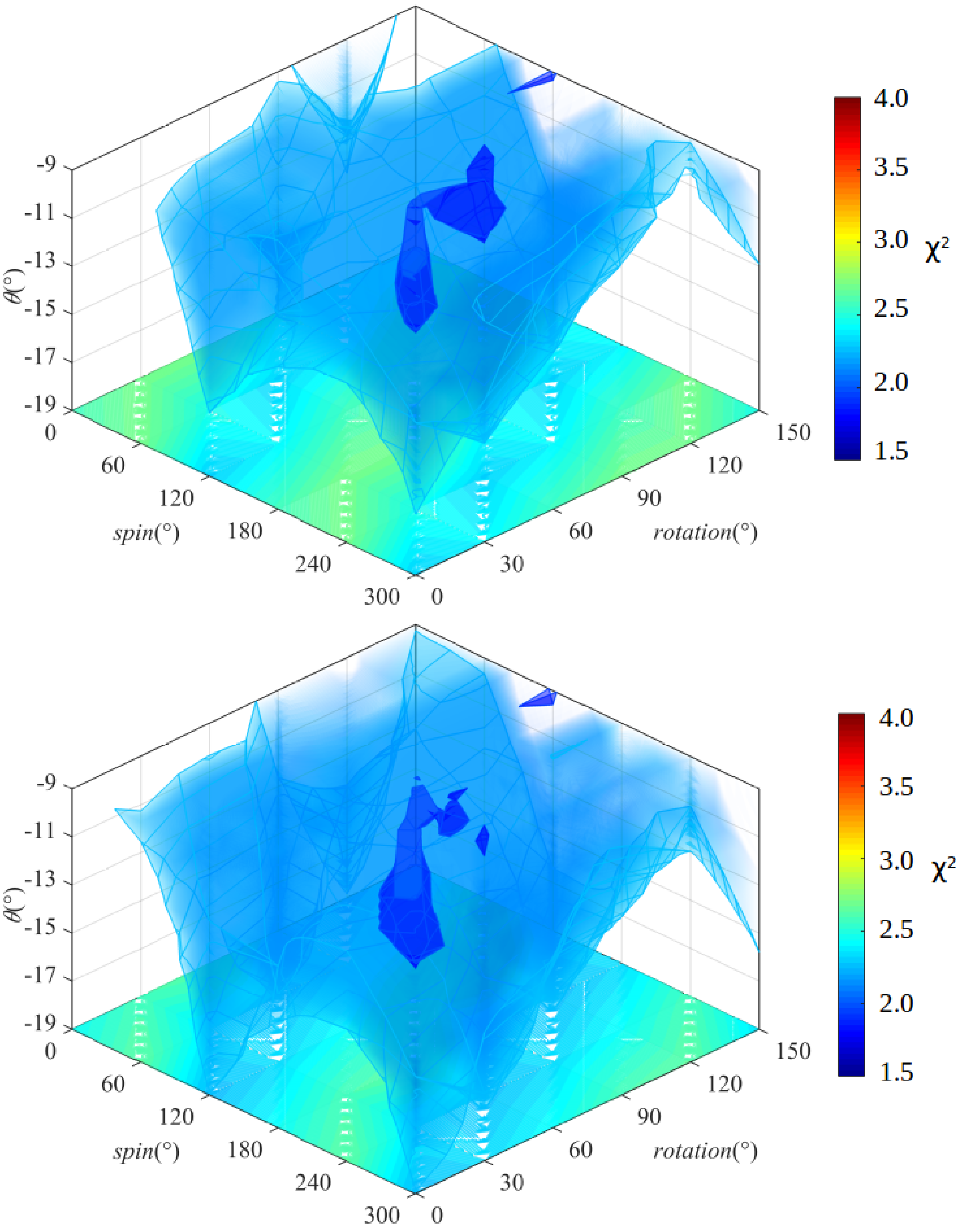


Figure 4.12 χ^2 as a function of three Euler angles for the Mhp1-NM PDC of (top) outward-open conformer (PDB code: 2jln) and (bottom) occluded conformer (PDB code: 4d1b). The data (SEC-SAXS) were collected at BM29, ESRF, France and P12, Petra III, Germany. Wild-type Mhp1 was solubilized in 10mM Tris-NaOH pH8 2.5% *v/v* Glycerol, 0.5% *w/v* NM, 140 mM NaCl in the presence of 0 mM (top) and 2 mM (bottom) added L-BH.

4.3.4 Conclusion

The method described in this chapter improves the SAXS modelling of the PDC for integral membrane proteins with known MX structures. The Pluecker based SIMD implementation of detergent corona placement shifts the speed-limiting step during parameter inference from the model building to the calculation of the theoretical scattering curve. Here I have demonstrated the potential of this method to generate physiologically meaningful models. New insights into the stability of PDCs can be drawn using this method. Furthermore, with the proper formulation, the modelling of other lipid bilayer mimics, such as nanodiscs can be readily achieved (due to the presumably stable structure of nanodiscs, this is in fact less troublesome than for detergents). Since little effort has been devoted to deciphering the basic physical chemistry principles underlying the crystallization process of membrane proteins (Loll 2014), prior knowledge of the geometrical conformation of the lipid-solubilized IMP obtained from SAXS modelling opens a door for more rational approaches to crystallization. For example, a better starting point is available for experiment-oriented simulations on the stability of IMPs in different detergents or other lipid-like environments. Pre-crystallization screens can also be carried out to investigate the orientational homogeneity of PDC as a guide to choice of optimal detergent and buffer conditions.

4.4 SAXS-Driven Ensemble Optimization of Antigen Binding Fragment of human IgG2

4.4.1 Background

Recent discoveries of antibodies that bind the same or very similar epitopes but display a large variation of biological activity indicate the presence of additional regulatory layers mediating the antibody response (Dillon et al. 2008; Hubbard et al. 2013; White et al. 2015). Antibodies directed against the key immune receptor CD40 are a case in point. CD40 is a member of the Tumor Necrosis Factor Receptor Superfamily and is essential for the initiation and regulation of adaptive immunity (Aspeslagh et al. 2016; Bartkowiak et al. 2015). Antibodies directed against CD40, anti-CD40 monoclonal antibodies (mAb), display a wide range of activities, from antagonism to strong agonism, with several in clinical testing (Ascierto et al. 2013; Bruhns et al. 2009; Vonderheide & Glennie 2013; Yamniuk et al. 2016). Recent studies have shown that both the antibody isotype and epitope targeted by these reagents is central to their activity (White et al. 2015) . In particular, the human IgG2 isotype affords stronger agonism than other isotypes for multiple anti-CD40 antibodies, even though the epitope binding regions and binding affinity are identical.

A key feature of IgG2 antibodies is their ability to undergo disulfide-switching in the hinge region. The hinge connects the Fc domain (Fc = fragment crystallisable) with the F(ab)$_2$

domain (Fab = antigen binding fragment). Disulfide switching is a natural redox process occurring over time as antibodies circulate in the blood, although the evolutionary basis and functional consequence of this change is unclear.

Due to the disulfide switching, IgG2 is present in the body in two distinct isoforms, termed A and B. It is produced in the A form and converted by redox conditions in the blood to a 1:1 ratio of A:B. The A form of IgG2 has been shown to have little to no activity as an anti-CD40 mAb (Liu et al. 2013). The B form, however, has been shown to have strong agonistic activity, both as a full-length IgG and in F(ab)$_2$ form. It is predicted that the causative factor between A and B forms is the shuffling of disulfide bonds in the hinge region (Dillon et al. 2008; Liu et al. 2008; White et al. 2015; Zhang et al. 2010). Importantly, the agonistic CD40 activity of the IgG2A/B antibodies was retained in F(ab)$_2$ fragments where the Fc portion had been cleaved off, supporting the proposal that biological activity is conferred through the hinge, a hypothesis confirmed by hinge swapping experiments between hinge-IgG1 and hinge-IgG2 (White et al. 2015). This conformational switching and the Fc independent activity significantly depart from the classical epitope recognition model of activity. The structural basis behind this striking difference, mediated solely by the hinge disulfides, with no change to F(ab)$_2$ binding or affinity to CD40, remains obscure.

Solution SAXS is a technique that is sensitive to the conformational variability of multi-domain proteins and large-scale protein fluctuations. The quantitative analysis of SAXS data is a way to exploit the biological insight behind the structure-activity relationship of F(ab)$_2$-IgG2. A series of anti-CD40 antibodies were designed using the same F(ab)$_2$ region specificity and were prevented from disulfide shuffling by site directed mutagenesis of cysteine residues, which yielded F(ab)$_2$s that were ‘locked’ into the postulated A-form or B-form configurations. Using F(ab)$_2$ fragments of these reagents we determined their biological activity and related this to resolved structural models for multiple locked A- and B-form variants. The crystallographic structures reveal a domain cross-over by cysteine bridges in the B-form that provides the basis for a model of conformational restriction, supported by in-solution studies using SAXS analysis. However, the modelling strategy required for capturing the conformational variability in F(ab)$_2$-IgG2 is challenging.

The relatively broad conformational space that can be sampled by highly flexible proteins or protein complexes with flexible linkers in solution make the single-model representation a poor description of proteins’ solution behaviors. For the SAXS measurement of F(ab)$_2$, the scattering intensity is a noisy observation of an ensemble average reflecting the solution behaviour of the F(ab)$_2$ fragments. Recently, SAXS-based ensemble modelling has become increasingly significant to characterize proteins with structural polydispersities (Cordeiro et al. 2017; Shevchuk & Hub 2017; Tria et al. 2015). The main idea of ensemble methods is that theoretically the experimental SAXS profile can be described by a linear combination of SAXS curves from a group of preferable conformations. There are two key aspects of SAXS-based ensemble modelling: 1) generation of a large ensemble pool that describes the conformational landscape of the protein; 2) use of optimization methods to select a sub-ensemble from the pool under the guidance of SAXS data. While many works focus mainly on the novel selection algorithm (Antonov et al. 2016; Bertini et al. 2010; Daughdrill et al. 2012; Pelikan et al. 2009; Różycki et al. 2011; Tria et al. 2015), only two strategies are readily available for the ensemble pool generation: molecular dynamics trajectory and kinematic rigid-body motion. Molecular dynamics (MD) simulation, when an appropriate force-field is provided, is suitable to approximate the conformational space of flexible proteins. However, the computational costs of sampling a landscape of large conformational changes hampers the exhaustive exploration of conformational space. On the other hand, a kinematic motion which treats the domains with well-defined secondary structure as undisrupted rigid-bodies and the linker or restriction sites as kinematic joints allows constrained motion of different domains. However, problems such as under-/over-represented motion and nonphysiologic conformers cannot be prevented.

To better understand the structure-activity relationship of the designed A-forms and B-forms of IgG2-F(ab)$_2$ I have developed a SAXS-guided ensemble modelling method that combines the principles of both molecular dynamics simulation and rigid-body motion. Here the focus is on providing statistical distributions of structural conformations that are sufficient to explain the observed disulfide-switching effects on antibody activity.

4.4.2 Methods

4.4.2.1 Samples of A-form and B-form

Specific IgG2 cys-ser mutants used in this study are listed in the order of increasing CD40 agonist activity: C22S/C214S (A-form), C225S (A-form), C224S (A-form), C225S/KC214S (B-form), C228S (B-form), C224K/KC214S (B-form). The plausible hinge patterns and the MX models for each sample are shown in Fig. 4.13

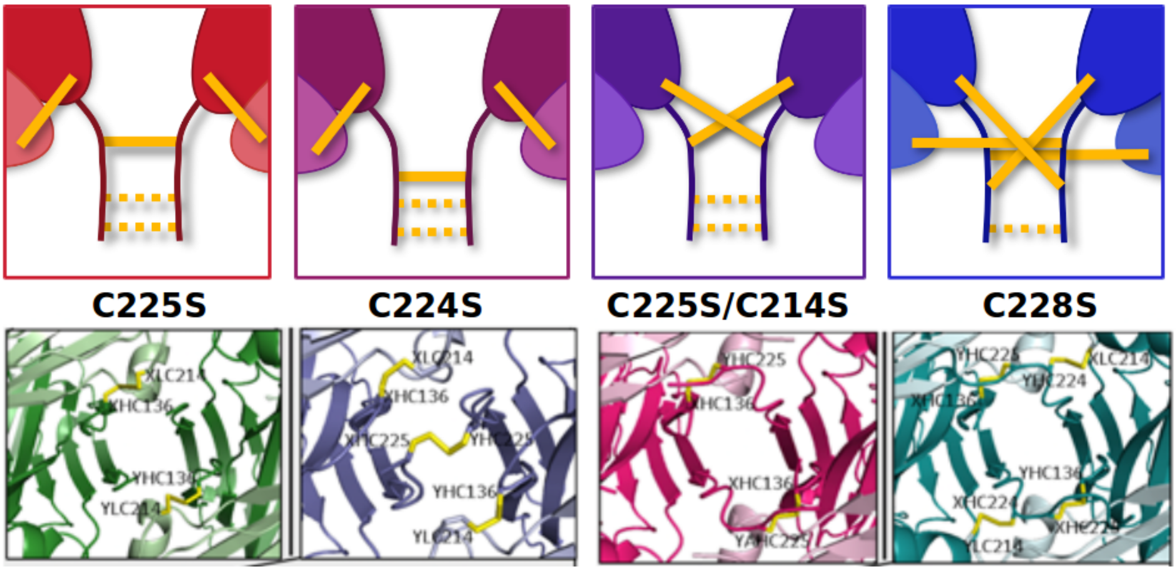


Figure 4.13 The hinge region patterns and MX structure of selected A-forms and B-forms. The PDB codes are 6tkb, 6tkc, 6tke, 6tkd for C225S, C224S, C225S/C214S and C228S, respectively. The disulfide bonds are highlighted as yellow sticks. The figure is courtesy of Dr. Chris Orr, University of Southampton.

4.4.2.2 SAXS data collection

Purified F(ab)$_2$ were concentrated to ~10 mg/ml using VivaSpin concentrators with a 10,000 Da MW cutoff. SAXS data were collected at the BM29 beamline, ESRF. Scattering images were collected using a Dectris Pilatus 1M detector with a sample to detector distance of 2.9 m. Data were collected at a wavelength of 0.99 Å. Samples were loaded using a SEC-SAXS setup, passing through a Superdex 200 10/300 column before entering a 1 mm quartz glass capillary to be exposed to X-rays at a temperature of 20 ℃.

4.4.2.3 Molecular Dynamics Simulation and Principal Component Analysis

Atomistic molecular dynamics (MD) simulations were performed using the Amber16 molecular dynamics package (Salomon-Ferrer et al. 2013). Crystal structures were used as the starting point with unresolved hinge residues built in using Modeller (Eswar et al. 2006). The protein was represented by the Amber ff14SB force field. Ions were represented by the parameters of Joung and Cheatham (Joung & Cheatham 2008). Water was represented by a TIP3P water model. Both the A-form and B-form were run for 1500 ns each, which formed three independent repeat runs of 500 ns with randomised ion starting positions and different randomised seeds for the Langevin thermostat. Frames for analysis were extracted at 1 ns intervals giving 1500 structures for analysis for both the A and B form. Principal component analysis (PCA) was performed using the Bio3d package (Grant et al. 2006). Translation and rotation between frames was removed by C_α alignment.

4.4.2.4 Generation of the Conformational State Pool

Based on the crystallographic structure of the $F(ab)_2$ fragment of IgG2, the dimer model was created using the symmetry mates, (x, y, z) and (x-y, $\bar{y}$, $\bar{z}$) (Fig. 4.14). The two $F(ab)_2$ arms were defined as two rigid bodies. According to the cross-over pattern of the hinge region, rotational centers and flexible linkers were assigned for individual A-from and B-form.

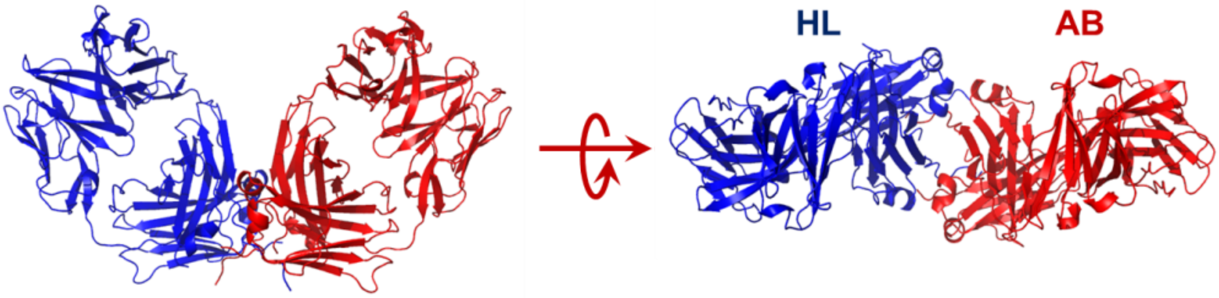


Figure 4.14. The $F(ab)_2$ dimer model generated from symmetry mates. The two heavy chains are indexed as H and A. The two light chains are indexed as L and B. Two rigid-body arms are HL and AB.

Each candidate in the ensemble pool is generated via the following steps :

(1) Set a rotational center at the mass center of the hinge region listed in Table 4.2. The whole entity constructed by the identified residues on chain H and chain A is taken as a spherical joint.

(2) In order to reduce the rigidity of the kinematic pair, flexible linkers are introduced. The selected residues (Table 4.2) in the sequences are built a C_α trace with Flory's model using backbone C_α-s with size limitations (Tria et al. 2015). This stage can be seen as a heuristic process.

(3) Shift the rotational centers for two $F(ab)_2$ arms to the H-Glu221 and A-Glu221.

(4) Randomly place the two $F(ab)_2$ arms according to the method described in §4.3.2.7

(5) Discard any rotation leading to two arms on opposite hemispheres (non-physiological candidates).

(6) Calculate the vectors between the center of each arm and the rotational center (vec_{rot}). Define wag plane by taking the sum of two vec_{rot} as the plane normal. Calculate the major principal vector of two arms. Define the arm vector (vec_{arm}) as the sum of vec_{rot} and the maximum projection of F(ab)$_2$ arm along the major principal.

Table 4.2 Definition of rotational centers and flexible linkers for each sample

		arm model	rotational center	flexible linker
C22S/C214S	A-form	6tkb	H, A - C225SVECP	H, A - E221RK
C225S	A-form	6tkb	H, A - C225SVECP	H, A - E221RK
C224S	A-form	6tkc	H, A - S225CVECP	H, A - E221RK
C225S/KC214S	B-form	6tke	H, A - A225CAECP	H, A - E221AG
C228S	B-form	6tkd	H, A - C225CVESP	H, A - E221RK
C224K/KC214S	B-form	6tkd	H, A - C225CVESP	H, A - E221RK

4.4.2.5 Ensemble Optimization

For each sample, the conformational state pool (10,000 randomly sampled candidates) was generated using the method described in §4.3.2.4. Selection of the best-fit sub-ensembles for each sample was performed using Genetic Algorithm Judging Optimisation of Ensembles (GAJOE) (Bernadó et al. 2007). 100 ensembles composed of randomly chosen individuals from the conformational state pool were then used as the initial generation. 150 cycles were carried out per run to obtain a statistically significant result. Each cycle comprised 1500 generations. The reduced χ^2 is reported using the fit of best-fit ensembles in the final generation of the most successful cycle to the experimental SAXS data.

The Information entropy of the resulting ensembles was calculated in order to quantitatively evaluate the overall flexibility of the F(ab)$_2$ structures. The entropy $H(i)$ was calculated as $—\Sigma P_i \log(P_i)$, where P_i is the population frequency of each ensemble. The scale is −1 to 0, where $H(i) = -1$ indicates maximum flexibility (a uniform distribution without selectivity).

t4.4.3 Results and Discussion

4.4.3.1 Conformational Space of F(ab)$_2$ can be Described by a Kinematic Model

Considering a composing transformation on two arbitrarily shaped rigid-bodies with a kinematic contact, let two stick arms be vector C_{AB} and vector C_{HL} and one of the arms be fixed as the frame of reference. The schema (Fig. 4.15) shows the simplified definition of the two characterizing variables, *bend motion* and *wag motion*. There is a third motion, *twist motion* , if the rigid-bodies, with the major principals of C_{AB} and C_{HL} , are anisometric.

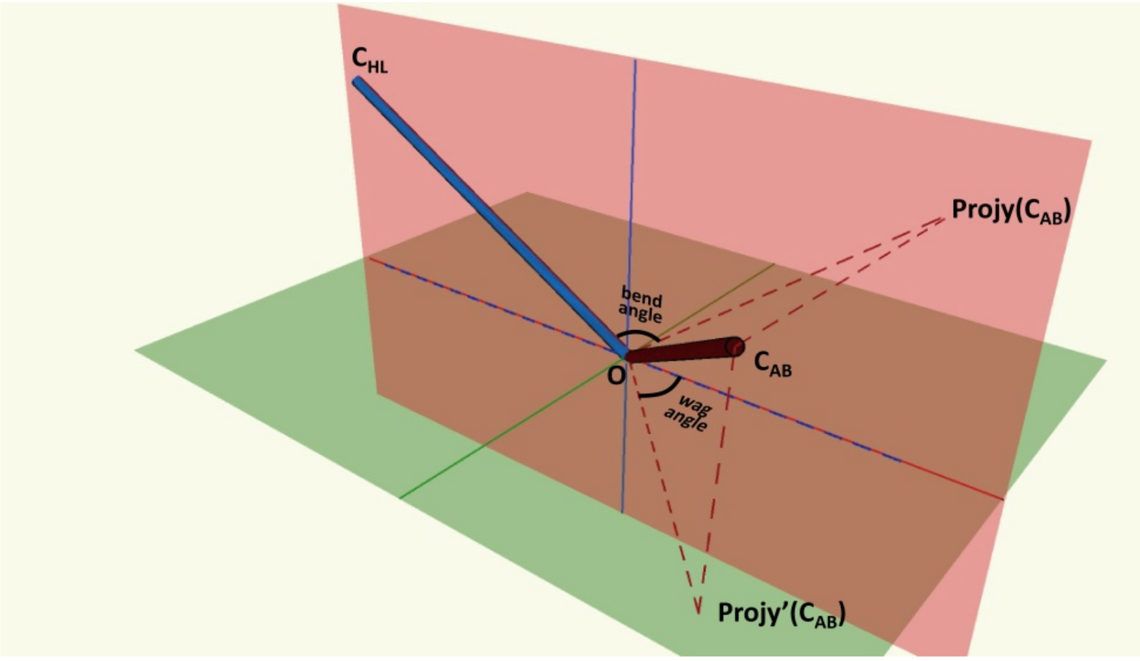


Figure 4.15 The simplified kinematic model. *O-C*$_{HL}$ and *O-C*$_{AB}$ together divide the *S*^{2} space by two perpendicular planes, namely plane *y* (red) and plane *y*' (green). Bend motion is the relative movement related to the angle between the link *O-C*$_{HL}$ and the projection component of the link O-C$_{AB}$ onto the plane *y*. *Wag motion* is the movement related to the angle between the two respective projection components of the *O-C*$_{HL}$ and *O-C*$_{AB}$ onto the plane *y*'.

PCA analysis was employed to examine the relationship between different conformations sampled during the MD trajectory. The examination of the contribution of each residue to the principal components suggests that the first, second and third principal motion (PC1, PC2 and PC3) can be referred to as the bend motion, wag motion and twist motion of the simplified kinematic model, respectively (Fig. 4.16). The percentage of the total mean square displacement (variance) of atom positional fluctuations captured in each principal component is characterized by their corresponding eigenvalue. The proportion of variance can then be plotted as the function of principal components (Fig. 4.17). For the A-form (simulated using C225S), PC1, PC2 and PC3 account for 45.24 %, 24.35 % and 7.8 % of variance, respectively, which gives a total 77.39% cumulative variance. For the B-form (simulated using C228S), PC1, PC2 and PC3 account for 35.13 %, 20.38% and 11.58 % of variance, respectively, which give them a total 67.10 % cumulative variance. The 4th and 5th principal motions are intra-F(ab) hinging motion and take the cumulative variances up to 81.33 % and 77.39 %. The conformational changes triggered by these two motions are small-scale because the disulfide bonds between heavy chain and light chain lock the F(ab) arm as a non-flexible entity (Wypych et al. 2008). Moreover, experience on PCA suggests that principal components capturing overing 70 % of the total variance are sufficient to provide a useful description while still retaining most of the variance in the original distribution and indeed a standard molecular dynamics trajectory (Grant et al. 2006). Therefore, it is expected that the first three components are sufficient and the motions involved in the rigid-body kinematic model can be indeed used to describe the conformational landscape of $F(ab)_2$.

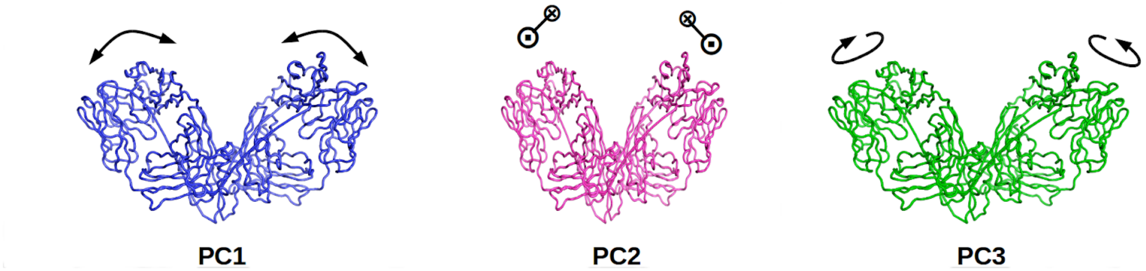


Figure 4.16 The demonstration of the first three decomposed principal motions. PC1, PC2 and PC3 can be mainly assigned to the $F(ab)_2$ hinging on the paper plane, the $F(ab)_2$ wagging perpendicular to the paper plane and the twist originating from the hinging. These three principal motions map the bend, wag and twist motion of the simplified kinematic model.

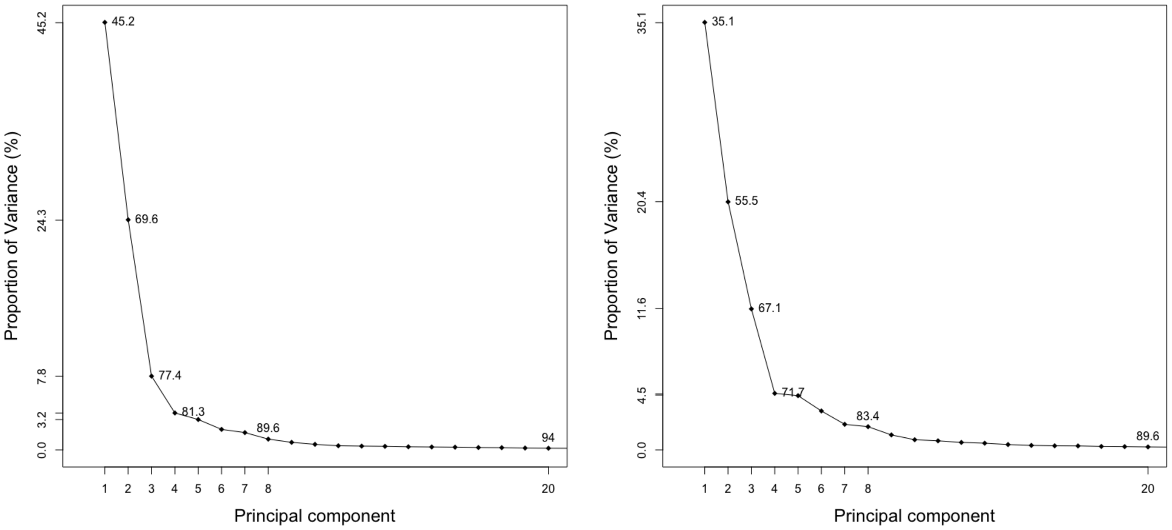


Figure 4.17 The proportion of variance for the PCA of MD trajectories of A-form (left) and B-form (right). The principal motions take the cumulative variance up to 70 % should be considered as relevant motions for $F(ab)_2$ function. The figure is courtesy of Hayden Fisher, University of Southampton.

4.4.3.2 MD Trajectory Cannot Sufficiently Explore the Conformational Space

During the time-scale of MD simulations, a local energy minimum as a function of the chosen force field cannot be reached (personal discussion with Prof. Ivo Tews), which suggests the conformational landscape has not been fully explored within the total time duration of the simulation. Except for the convergence consideration, the ensemble optimization based on 3,000 frames of MD trajectory shows that the ensemble pool generated with the MD trajectory cannot accurately describe the solution behavior of $F(ab)_2$ for either A-form or B-form (Fig. 4.18). The radius of gyration (R_g) and maximum length scale (D_{max}) of the candidates derived from the MD trajectory are both in a Gaussian-like distribution, with the MX model falling at the peak position. The most populated species obviously deviate from the confidence interval observed from the original pool. The reason that the selected ensembles bias towards the high positive *z*-score interval is that $F(ab)_2$ are much more extended in the solution environment than in the crystal (Ayyer et al. 2016). A MD simulation starting with the MX model with limited simulating time and standard force field may not be able to escape the energy barrier imposed by the crystal packing and therefore leads to a pool including more compacted conformation states than are actually present in the solution environment.

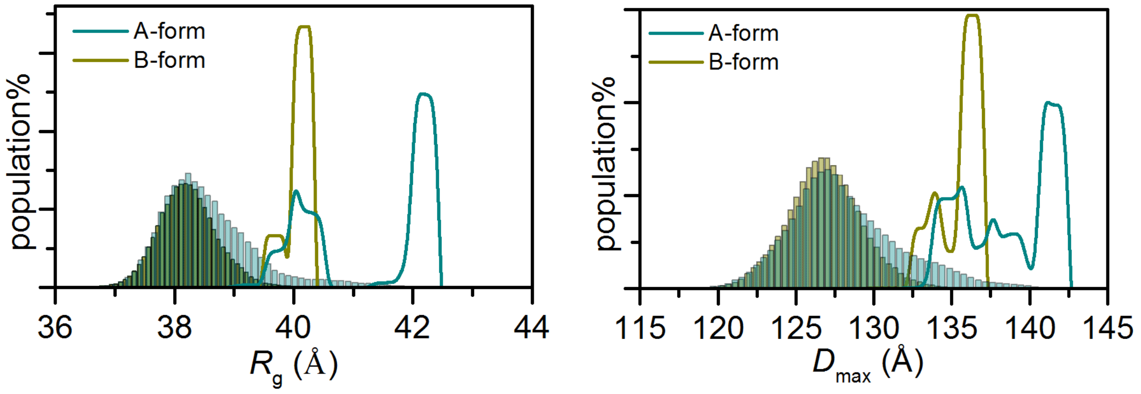


Figure 4.18 GAJOE results using the 3,000 frames of MD trajectory as an external conformational states pool. (Left) The R_g distribution of the ensembles in the pool (bars) and selected by GAJOE (lines) for A-form (dark cyan) and B-form (dark yellow). The reduced χ^2 is 2.506. (Right) The D_{max} distribution of the ensembles in the pool (bars) and selected by GAJOE (lines) for A-form (dark cyan) and B-form (dark yellow). The reduced χ^2 is 3.863. Although the result suggests B-form is more compacted than A-form, the selected ensembles are biased.

4.4.3.3 Reducing Algorithmic Bias by Heuristic Process

Apart from the problem of efficient sampling, another question should be addressed: will the selection be directed along an *a priori* direction or be biased by systematic sampling (Orellana 2019)? A direct rotation of $F(ab)_2$ arms can be applied by setting the rotation center at the geometrical center of the observed bonded cystines, which excludes any non-hinge flexibility and the possibility of any disulfide shuffling. A Monte-Carlo sampling by this schema generates a randomly-sampled yet constrained conformational states pool. For the two parameters, R_g and D_{max}, the GAJOE optimization on the constraint pool shows the evolution is explicitly affected by the distribution of the constraint pool (Fig. 4.19). The final result is a straight amplification of the most populated species in the initial pool. On the other hand, after flexible linkers being introduced (§4.4.2.4), a much broader landscape is explored. The linkers act as a heuristic interpolation of rigid-body motions, which corresponds to a physical approximation to any off-script transition. The final result shows little to no biasing with both R_g and D_{max}. However, the ensemble averages derived from constraint pool and heuristic pool fit the SAXS data with similar reduced χ^2 (2.946 for constraint pool and 2.780 for heuristic pool). Therefore, it is hard to conclude which approach is significantly more appropriate. In statistics, this is a typical bias-variance problem (Neal 2019): a less-biased optimization suffers high-variance (heuristic strategy) whereas a less-variant optimization is prone to have high-bias (constraint strategy). In this specific study, the balance leans towards the side of applying the heuristic process based on three factors: 1) bias decreases with increasing number of observations (Bohm & Zech 2010); 2) the optimization should be guided by the SAXS data instead of being oriented by the method used to generate the ensemble pool; 3) it allows the pool generation of all IgG2-F(ab)$_2$ samples adapt to a similar strategy. For the rest of this chapter, the strategy described in §4.4.2.4 is used to generate the conformational states pool.

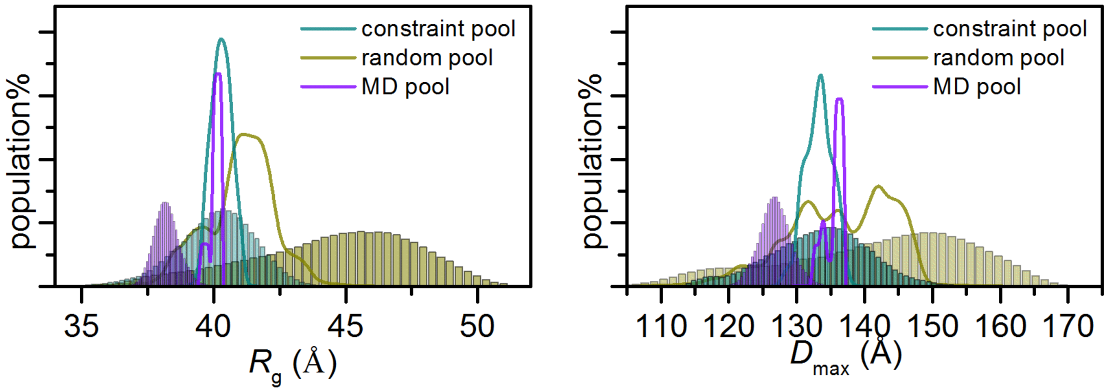


Figure 4.19 A comparison of GAJOE results using different pool generation strategies. (Left) The R_g distribution of the ensembles in the pool (bars) and selected by GAJOE (lines) using constraint pool (dark cyan) , heuristic pool (dark yellow) and MD pool (dark purple). (Right) the D_{max} distribution of the ensembles in the pool (bars) and selected by GAJOE (lines) using constraint pool (dark cyan) , heuristic pool (dark yellow) and MD pool (dark purple).

4.4.3.4 Changed Flexibility of A-form and B-form Validated by SAXS

To understand the reason behind the changes in flexibility, the crystal structures of the F(ab)$_2$ fragments of A-form and and B-form were determined by crystallographic analysis. However, despite the distinct hinge patterns being observed, on the larger-scale all F(ab)$_2$ fragments adopted a similar conformation in the crystal. This is due to the fact that the conformation is dominated by the crystal packing of the F(ab)$_2$ fragments (Panjkovich & Svergun 2016). The difference in the confromations or the distribution of conformers of two forms is thus hard to understand from these data. How the differences in the hinge structure as seen in the crystal structures of F(ab)$_2$ variants influence flexibility was therefore further investigated with solution SAXS experiments.

The SAXS-oriented ensemble analysis reveals that the conformer populations of IgG2-F(ab)$_2$ variants differ. B-form subtypes show an overall narrower D_{max} and R_g than A-form subtypes, and contain fewer ensembles that are larger in size than the A-from subtypes. The highly active B-forms (C228S and C224K/KC214S) also show a more clear unimodal size distribution when compared to other samples. In contrast, a greater variation in the ensemble is evident in the A-form variants. Interestingly, a distinct population of ensembles around a R_g of 41 Å and a D_{max} of 135 Å gradually increases with activity (Fig. 4.20).

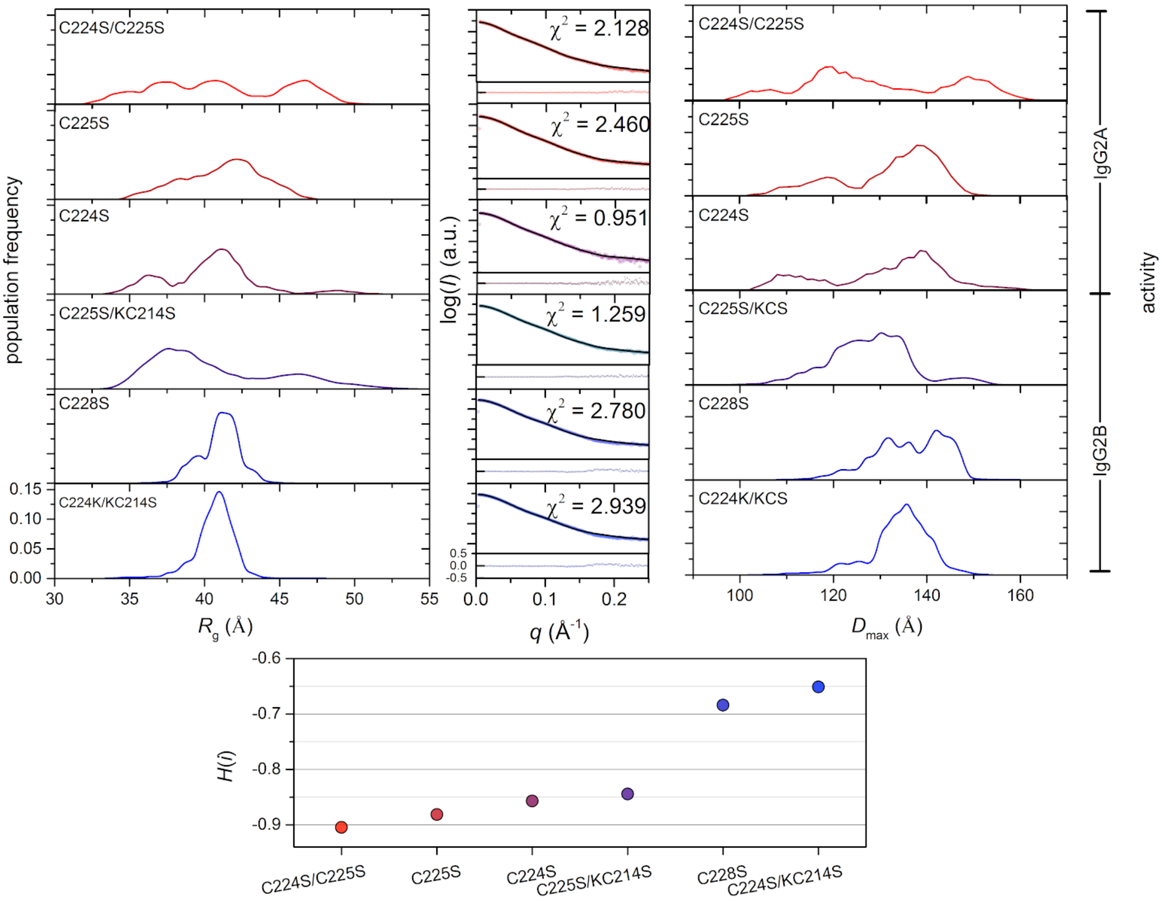


Figure 4.20 SAXS-oriented ensemble analysis over the F(ab)$_2$ variants. F(ab)2 variants are sorted from A-form (top) to B-form by increasing agonist biological activity (Orr 2019). The final conformer population distribution of the different F(ab)$_2$ variants are shown with two indicators, R_g (top left panel) and D_{max} (top right panel). (Top middle panel) χ^2 fitting result returned by GAJOE and resolution dependent relative residues of the fit. (Bottom) prediction of the information entropy $H(i)$ as a measure of flexibility. The entropy $H(i)$ of −1 indicates maximum flexibility (uniform distribution, no selectivity).

While the distribution of conformations changes between A-form and B-form, changes in the accessible conformational space for the molecules are also observable.The information entropy of the ensembles is quantified to illustrate this trend. The conformation space that can be sampled by the F(ab)$_2$ fragments is significantly confined (a totally random population would have an information entropy of −1). The less active "A-like" variants show a higher information entropy than the more "B-like" variants, indicating they explore a much wider region of conformational space.

This analysis allows us to propose a model of CD40 agonist antibody function based on conformational flexibility. We hypothesis that the disulphide bond patterns define the accessible conformational space for A- and B-forms. While the A-form readily accesses all conformations that the B-form does, the B-form does not access all conformations sampled by the A-form in the same way. The observation of the tighter conformational distribution of the highly active B-forms suggests certain conformational states are preferentially occupied. We propose that one or more of these preferred states constitutes an "on-state" form which is a semi-compact state with a R_g between 41 Å and a D_max between 110 and 140 Å. In the B-form, the accessible conformational space is restricted, resulting in a higher proportion of the molecules sampling the "on-state" and staying there long enough to form a productive interaction with CD40. In this model the CD40s are also diffusing around in the membrane – with random transient interactions that are the same as the "signaling" state, however the residence times and densities of these clusters are insufficient to signal productively (i.e. below threshold). In the B-form, the F(ab)$_2$ residence time in the "on-state" is long enough to catch two CD40s in the right spatial arrangement and the binding stabilizes both Fab and CD40s in a "signaling-on" state. If multiple F(ab)s bind the surface this can produce a network of optimally spaced CD40s to create an over threshold signal. In the A-form, the kinetics are mis-matched. Although the A-form also samples the on-state, it has fewer chances of catching two CD40s in the right relative spatial arrangement.

To illustrate this a spherical map is plotted, which represents the real-space orientation of the two F(ab)$_2$ two arms relative to the defined rotational center (Fig. 4.21). It is clear that the preferential conformations are very different for the A- and B-forms. The two vec_arm are plotted on a hemi-spherical surface that gives an intuitive picture of the conformational space sampled and the F(ab)$_2$ configuration the "on-state" corresponds to. While the A-form evenly samples throughout the conformational space, the B-form predominantly samples a more limited region of conformational space that clearly clusters in certain orientations. With this insight, one may be able to construct a more detailed biological mechanism model which correlates to the experimental observation of membrane CD40 clustering (Grassmé et al. 2002).

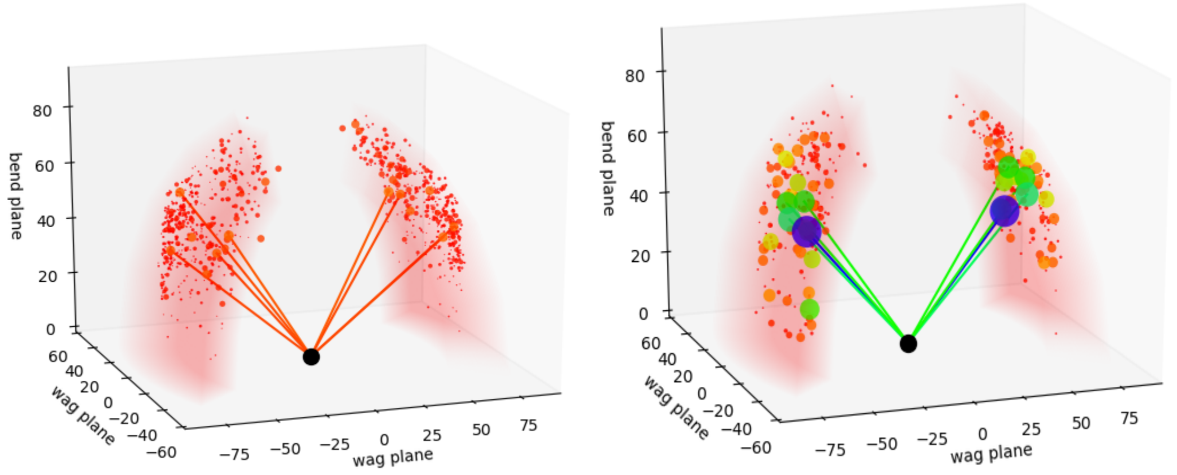


Figure 4.21 The F(ab)$_2$ arm orientation of the optimized ensembles for the A-from (left) and the B-form (right). Figures are plotted using spherical coordinates, where the z-axis goes up through the hinge region and the other two angles effectively describe the wag and bend. For each conformer there is a pair of "dots" on the hemisphere surface that indicates the end points of the *vec*$_{\mathrm{arm}}$. The "dots" represent the conformational space sampled after selection against the SAXS data. The colour (‘jet’ pallet, red-1, green-1750 , blue-3500) of the dot represents the population of selected conformers. The radius of each pair of "dots" is also related to the population. The red halo represents the orientational configuration space that can be explored by the candidates in the constraint pool. The lines highlight the five most important conformers selected in each sample.

4.4.4 Conclusion

In this section, SAXS-oriented ensemble modelling was used to provide detailed insights into the hinge disulfide mediated C40 agonist activity of IgG2-F(ab)$_2$. Although the workflow of ensemble modelling is similar, the strategies used in each algorithm vary. A key problem is how to reasonably expand the static snapshots to multi-state structural ensembles. First, coarse-grained sampling based on a kinematic model was shown to be sufficient to predict the conformational distribution of F(ab)$_2$ by PCA. Next, the insufficiency of MD sampling was demonstrated. The sampling was further refined by a heuristic process, where certain atomistic features were applied to alleviate the bias introduced by the kinematic model. The final heuristic sampling strategy was used to generate conformational states-pool for a series of $F(ab)_2$ isotopes with different activity, namingly A-form and B-form. The results of SAXS-oriented ensemble modelling suggests that A-forms have no preferential conformation whereas an “on-state” cluster exists in the B-form. Interestingly, the CD40s are found to form trimeric receptor complexes (Grassmé et al. 2002; Haswell et al. 2001; Wajant 2015). A topological comparison between the CD40 complex and the potential $F(ab)_2$ clusters may help further unveil the mechanism of modulated target receptor signaling.

4.4 Summary

Due to the isotropic nature of solution scattering, SAXS-oriented modelling suffers ambiguity. This ambiguity can lead to non-physiological conclusions. Unlike the problems presented in the previous two chapters, the subjectivity of modelling comes mainly from the algorithmic bias instead of personal impression. In this chapter, efforts are made to detect the algorithmic bias and thus improve the fidelity of the SAXS modelling. Since SAXS is a low resolution method, coarse-grained modelling is a good way to sufficiently explore the conformational landscape while retaining the structural envelope-level details. In the first section, a general algorithm was developed to construct a physically meaningful model for protein-detergent complexes. The bias was tackled by introducing *a priori* chemistry information. Rather than a general method, the second section is merely a case study. However, the core of the second section is still applicable for other practical applications of SAXS-oriented modelling, that is for an *in silico* analysis each step should be carefully inspected and validated efficiently by whether the extrapolation of a certain step violates the principle of other computational steps. I sincerely hope the approaches and scripts developed here can be used and tested in more cases to help the bio-SAXS user community gather more useful information from their data.

Chapter 5

Conclusions

The purpose of this thesis was to develop new methods for solution SAXS of biomacromolecules capable of alleviating the ambiguity and subjectivity involved in SAXS data processing. In chapter 1, the commonly-used data analysing methods as well as their drawbacks were discussed and the need of stabilizing solutions of SAXS data analysis by imposing *a priori* information (restrain) or/and by correlating with an independent observation on a closely related effect (constraint) was suggested. Chapter 2-4 detail how each stage of data analysis can be improved by specific strategies.

Chapter 1 discusses the range of analytical and numerical tools available for studying biological systems by solution SAXS. While there are many methods for extracting information from SAXS patterns, they are mainly ill-posed problems and are subject to ambiguity. The lack of objective criterion compromises the quality of interpretations of SAXS data.

Chapter 2 focuses on the post-processing of data. A new metric, correctness-state score (CSS) is introduced to assess the quality of a background correction for SEC-SAXS. Due to the inherent correlation, the spectroscopic data collected simultaneously along with the scattering pattern can be used as constraints. The algorithm is validated against publicly available datasets. The corresponding use of CSS on the correction of radiation-damaged dataset is demonstrated.

Chapter 3 attempts to interpret SAXS data in an assumption-free way. Specifically, chapter 3 suggests new approaches for analysing data from time-resolved SAXS and SEC-SAXS of membrane proteins (representing kinetic and static SAXS studies, respectively). For time-resolved SAXS, a quasi-quantitative figurative metric, cumulative first-ranked singular-values correlation Map (CSV-CORMAP), is proposed to make “on-the-fly” interpretations of time-resolved SAXS data without any assumptions. For SEC-SAXS of membrane proteins, the different approaches for deconvolution of SAXS elution profiles are evaluated. A combination of an exponential-modified Gaussian (EMG) model, Nelder-Mead (NM) minimizer and residual sum of squares (*RSS*) as the objective function is recommended for its universal robustness. Each part of Chapter 3 is concerned with validating the algorithm against the corresponding case studies and understanding the implication of the consequential results.

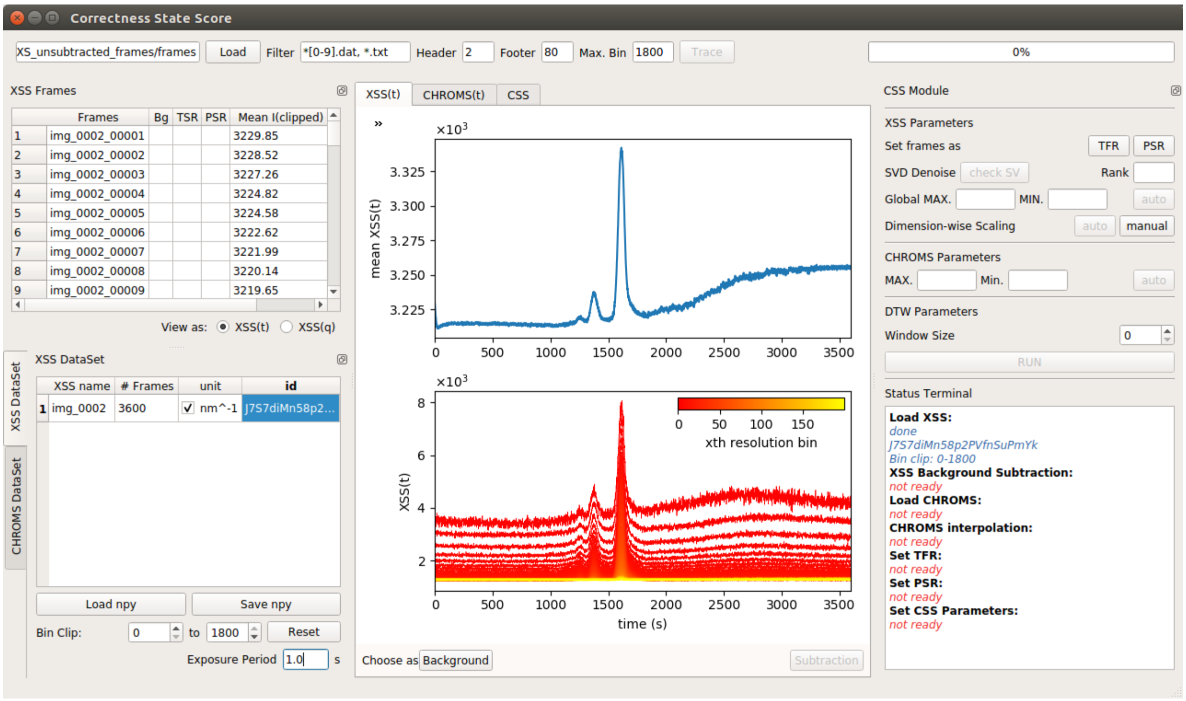
Chapter 4 targets SAXS modelling. The first part of Chapter 4 describes a novel coarse-grained method enabling the modelling of the protein-detergent complex (PDC) formed by integrated membrane proteins (IMP) and detergent molecules. By introducing chemical information and the Plücker test, it is able to stabilize the modelling result so that physiologically meaningful models are obtained. This method is used to gain insights into the stability of PDCs and to provide input into the design of subsequent crystallization trials. The second part of Chapter 4 discusses the ensemble approach of modelling a flexible dimeric system (i.e. a series $F(ab)_2$ isotopes of IgG2). A heuristic process is proposed to strike a balance between molecular dynamics sampling and rigid-body sampling and to overcome the drawbacks of both. The results of SAXS-oriented ensemble modelling reveals the relationship between the structural flexibility and the activity of $F(ab)_2$.

A cross-platform graphical user interface (GUI) has been designed using PyQt5, a Python binding for v5 of the Qt application framework, for CSS (Fig. 5.1-5.7). Command line interfaces have also been developed for CSV-CORMAP and coarse-grained PDC reconstruction. Such programs would help the bio-SAXS users to test and use the methods being discussed in this thesis. The python scripts for the methods in this thesis can be found via following links:

https://github.com/PearsonCUI/CSSGUI

https://github.com/PearsonCUI/DETPROT

https://github.com/PearsonCUI/SspGui



Correctness State Score

XS_unsubtracted_frames/frames | Filter *[0-9].dat, *.txt | Header 2 | Footer 80 | Max. Bin 1800 | 0%

XSS Frames

	Frames	Bg	TSR	PSR	Mean I(clipped)
1	img_0002_00001				3229.85
2	img_0002_00002				3228.52
3	img_0002_00003				3227.26
4	img_0002_00004				3224.82
5	img_0002_00005				3224.58
6	img_0002_00006				3222.62
7	img_0002_00007				3221.99
8	img_0002_00008				3220.14
9	img_0002_00009				3219.65

View as: ● XSS(t) ○ XSS(q)

XSS DataSet

	XSS name	# Frames	unit	id
1	img_0002	3600	☑ nm^-1	J7S7diMn58p2...

Bin Clip: 0 to 1800

Exposure Period 1.0 s

Choose as Background

CSS Module

XSS Parameters

Set frames as TFR PSR

SVD Denoise check SV Rank

Global MAX. MIN.

Dimension-wise Scaling auto manual

CHROMS Parameters

MAX. Min.

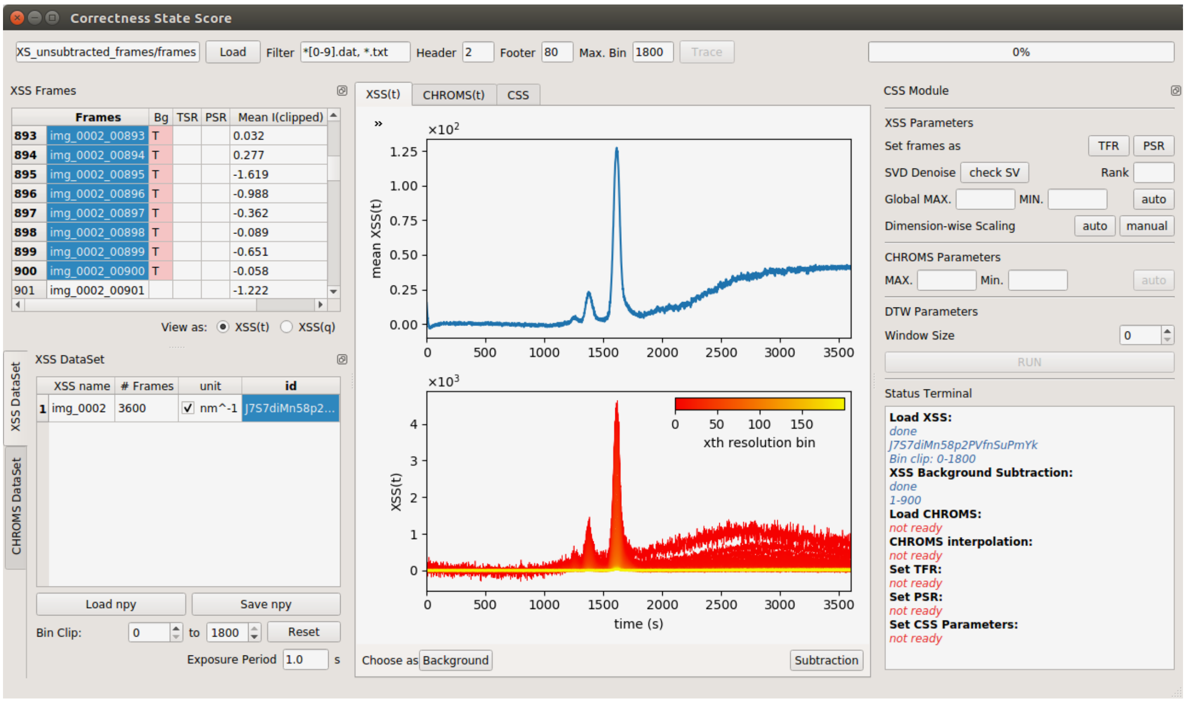
DTW Parameters

Window Size 0

Status Terminal

```
Load XSS:
done
J7S7diMn58p2PVfnSuPmYk
Bin clip: 0-1800
XSS Background Subtraction:
not ready
Load CHROMS:
not ready
CHROMS interpolation:
not ready
Set TFR:
not ready
Set PSR:
not ready
Set CSS Parameters:
not ready
```

Figure 5.1 The GUI of CSS. Load *XSS* dataset.



Correctness State Score

XS_unsubtracted_frames/frames | Filter *[0-9].dat, *.txt | Header 2 | Footer 80 | Max. Bin 1800 | 0%

XSS Frames

	Frames	Bg	TSR	PSR	Mean I(clipped)
893	img_0002_00893	T			0.032
894	img_0002_00894	T			0.277
895	img_0002_00895	T			-1.619
896	img_0002_00896	T			-0.988
897	img_0002_00897	T			-0.362
898	img_0002_00898	T			-0.089
899	img_0002_00899	T			-0.651
900	img_0002_00900	T			-0.058
901	img_0002_00901				-1.222

View as: ● XSS(t) ○ XSS(q)

XSS DataSet

	XSS name	# Frames	unit	id
1	img_0002	3600	☑ nm^-1	J7S7diMn58p2...

Bin Clip: 0 to 1800

Exposure Period 1.0 s

Choose as Background

CSS Module

XSS Parameters

Set frames as TFR PSR

SVD Denoise check SV Rank

Global MAX. MIN. auto

Dimension-wise Scaling auto manual

CHROMS Parameters

MAX. Min.

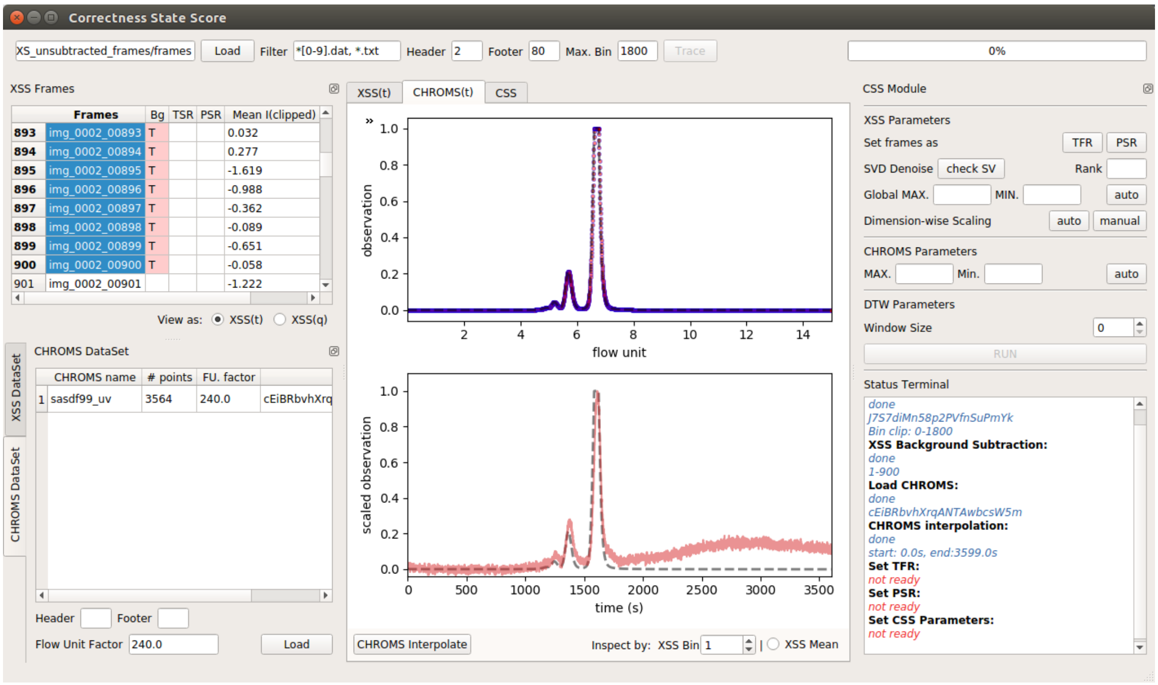
DTW Parameters

Window Size 0

Status Terminal

```
Load XSS:
done
J7S7diMn58p2PVfnSuPmYk
Bin clip: 0-1800
XSS Background Subtraction:
done
1-900
Load CHROMS:
not ready
CHROMS interpolation:
not ready
Set TFR:
not ready
Set PSR:
not ready
Set CSS Parameters:
not ready
```

Figure 5.2 The GUI of CSS. Background subtraction to yield a “corrected” $XSS(t, q)$.



XS_unsubtracted_frames/frames Filter *[0-9].dat, *.txt Header 2 Footer 80 Max. Bin 1800

	Frames	Bg	TSR	PSR	Mean I(clipped)
893	img_0002_00893	T			0.032
894	img_0002_00894	T			0.277
895	img_0002_00895	T			-1.619
896	img_0002_00896	T			-0.988
897	img_0002_00897	T			-0.362
898	img_0002_00898	T			-0.089
899	img_0002_00899	T			-0.651
900	img_0002_00900	T			-0.058
901	img_0002_00901				-1.222

	CHROMS name	# points	FU. factor	
1	sasdf99_uv	3564	240.0	cEiBRbvhXrq

Header Footer
Flow Unit Factor 240.0

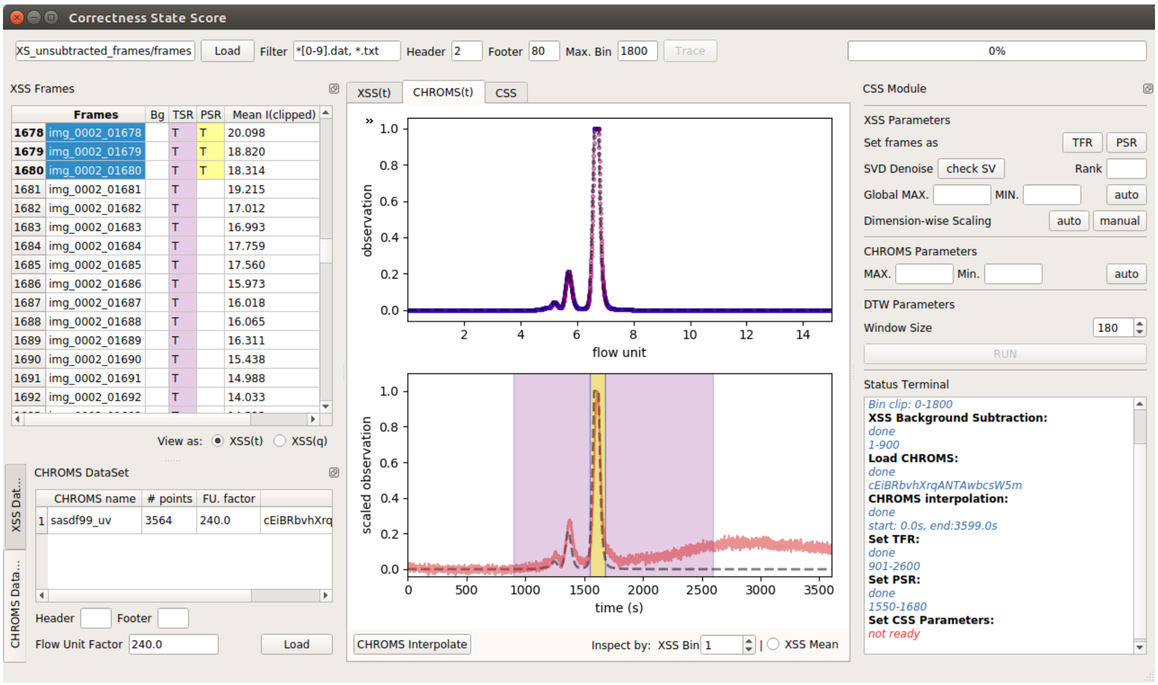
Inspect by: XSS Bin 1 | ○ XSS Mean

Set frames as
SVD Denoise Rank
Global MAX. MIN.
Dimension-wise Scaling
CHROMS Parameters
MAX. Min.
DTW Parameters
Window Size 0

Status Terminal

```
done
J7S7diMn58p2PVfnSuPmYk
Bin clip: 0-1800
XSS Background Subtraction:
done
1-900
Load CHROMS:
done
cEiBRbvhXrqANTAwbcsW5m
CHROMS interpolation:
done
start: 0.0s, end:3599.0s
Set TFR:
not ready
Set PSR:
not ready
Set CSS Parameters:
not ready
```

Figure 5.3 The GUI of CSS. Load *CHROMS* and convert the sampling rate of *XSS* and *CHROMS* into the same units



XS_unsubtracted_frames/frames Filter *[0-9].dat, *.txt Header 2 Footer 80 Max. Bin 1800

	Frames	Bg	TSR	PSR	Mean I(clipped)
1678	img_0002_01678		T	T	20.098
1679	img_0002_01679		T	T	18.820
1680	img_0002_01680		T	T	18.314
1681	img_0002_01681		T		19.215
1682	img_0002_01682		T		17.012
1683	img_0002_01683		T		16.993
1684	img_0002_01684		T		17.759
1685	img_0002_01685		T		17.560
1686	img_0002_01686		T		15.973
1687	img_0002_01687		T		16.018
1688	img_0002_01688		T		16.065
1689	img_0002_01689		T		16.311
1690	img_0002_01690		T		15.438
1691	img_0002_01691		T		14.988
1692	img_0002_01692		T		14.033

	CHROMS name	# points	FU. factor	
1	sasdf99_uv	3564	240.0	cEiBRbvhXrq

Header Footer
Flow Unit Factor 240.0

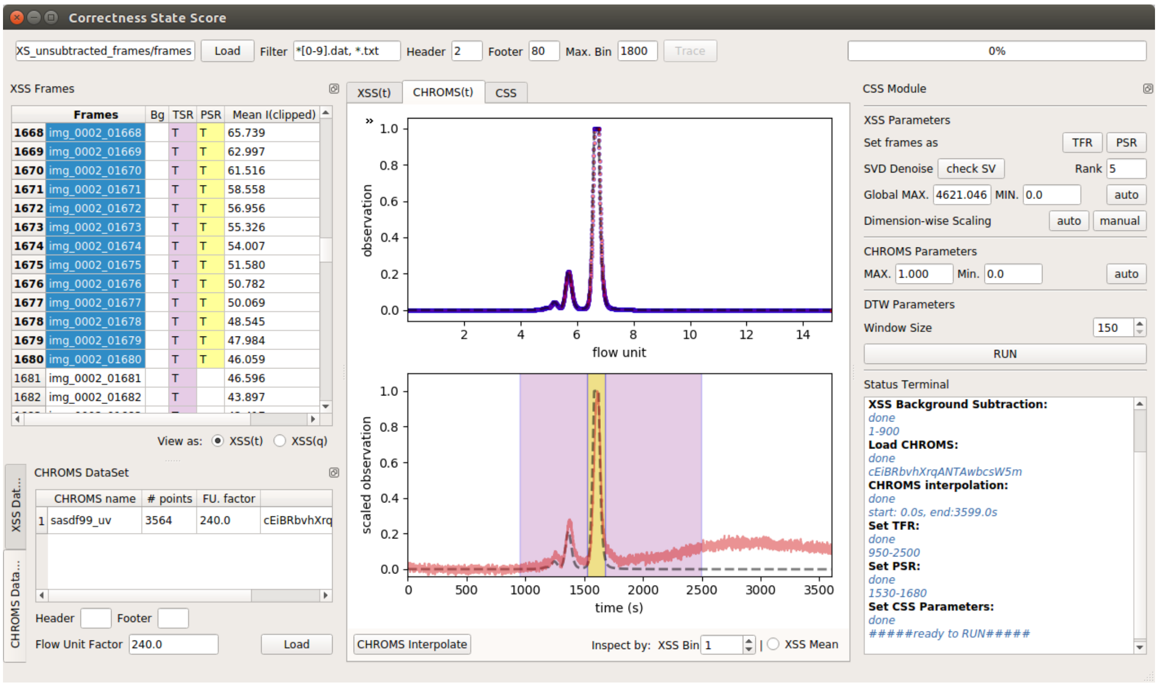
Inspect by: XSS Bin 1 | ○ XSS Mean

Set frames as
SVD Denoise Rank
Global MAX. MIN.
Dimension-wise Scaling
CHROMS Parameters
MAX. Min.
DTW Parameters
Window Size 180

Status Terminal

```
Bin clip: 0-1800
XSS Background Subtraction:
done
1-900
Load CHROMS:
done
cEiBRbvhXrqANTAwbcsW5m
CHROMS interpolation:
done
start: 0.0s, end:3599.0s
Set TFR:
done
901-2600
Set PSR:
done
1550-1680
Set CSS Parameters:
not ready
```

Figure 5.4 The GUI of CSS. Choose *TFR* (purple bar) and *PSR* (yellow bar)



XS_unsubtracted_frames/frames Filter *[0-9].dat, *.txt Header 2 Footer 80 Max. Bin 1800 0%

XSS Frames

	Frames	Bg	TSR	PSR	Mean I(clipped)
1668	img_0002_01668		T	T	65.739
1669	img_0002_01669		T	T	62.997
1670	img_0002_01670		T	T	61.516
1671	img_0002_01671		T	T	58.558
1672	img_0002_01672		T	T	56.956
1673	img_0002_01673		T	T	55.326
1674	img_0002_01674		T	T	54.007
1675	img_0002_01675		T	T	51.580
1676	img_0002_01676		T	T	50.782
1677	img_0002_01677		T	T	50.069
1678	img_0002_01678		T	T	48.545
1679	img_0002_01679		T	T	47.984
1680	img_0002_01680		T	T	46.059
1681	img_0002_01681		T		46.596
1682	img_0002_01682		T		43.897

View as: ● XSS(t) ○ XSS(q)

CHROMS DataSet

	CHROMS name	# points	FU. factor	
1	sasdf99_uv	3564	240.0	cEiBRbvhXrq

Header Footer

Flow Unit Factor 240.0

Inspect by: XSS Bin 1 | ○ XSS Mean

XSS Parameters

Set frames as TFR PSR

SVD Denoise check SV Rank 5

Global MAX. 4621.046 MIN. 0.0 auto

Dimension-wise Scaling auto manual

CHROMS Parameters

MAX. 1.000 Min. 0.0 auto

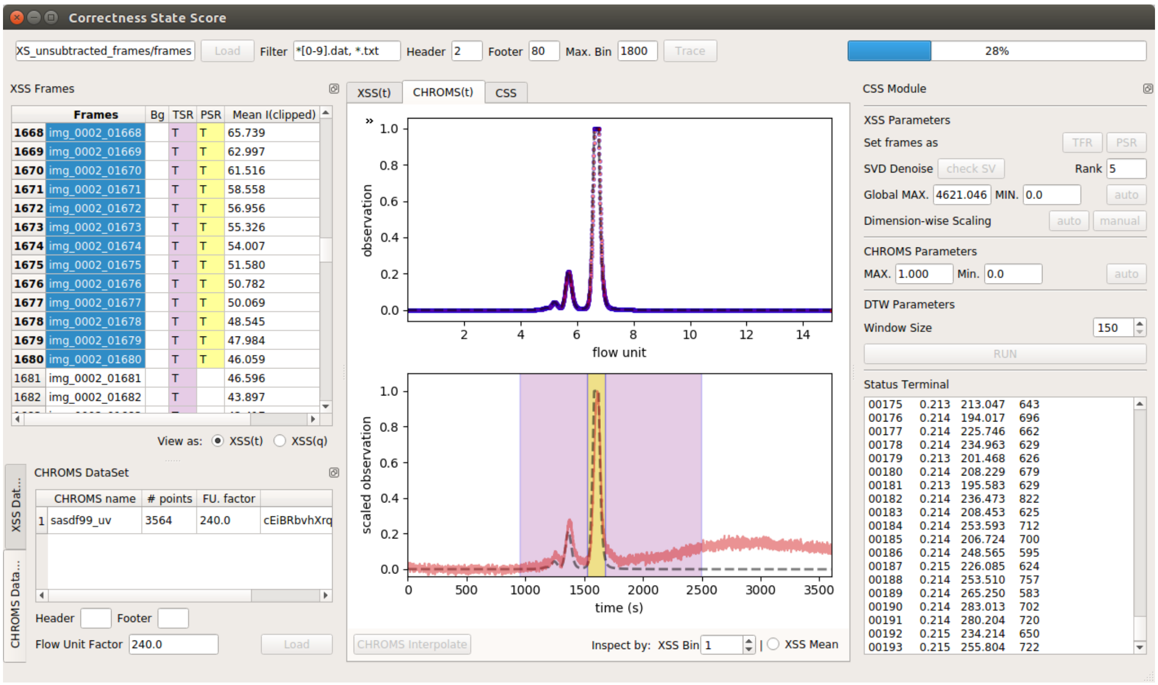
DTW Parameters

Window Size 150

Status Terminal

```
XSS Background Subtraction:
done
1-900
Load CHROMS:
done
cEiBRbvhXrqANTAwbcsW5m
CHROMS interpolation:
done
start: 0.0s, end:3599.0s
Set TFR:
done
950-2500
Set PSR:
done
1530-1680
Set CSS Parameters:
done
#####ready to RUN#####
```

Figure 5.5 The GUI of CSS. Calculate CSS parameters: $[\bar{\boldsymbol{u}}]_{TFR}$, $[\bar{\boldsymbol{x}}_n]_{TFR}$, and $[\hat{\boldsymbol{x}}_n]_{TFR}$.



XS_unsubtracted_frames/frames Filter *[0-9].dat, *.txt Header 2 Footer 80 Max. Bin 1800 28%

XSS Frames

	Frames	Bg	TSR	PSR	Mean I(clipped)
1668	img_0002_01668		T	T	65.739
1669	img_0002_01669		T	T	62.997
1670	img_0002_01670		T	T	61.516
1671	img_0002_01671		T	T	58.558
1672	img_0002_01672		T	T	56.956
1673	img_0002_01673		T	T	55.326
1674	img_0002_01674		T	T	54.007
1675	img_0002_01675		T	T	51.580
1676	img_0002_01676		T	T	50.782
1677	img_0002_01677		T	T	50.069
1678	img_0002_01678		T	T	48.545
1679	img_0002_01679		T	T	47.984
1680	img_0002_01680		T	T	46.059
1681	img_0002_01681		T		46.596
1682	img_0002_01682		T		43.897

View as: ● XSS(t) ○ XSS(q)

CHROMS DataSet

	CHROMS name	# points	FU. factor	
1	sasdf99_uv	3564	240.0	cEiBRbvhXrq

Header Footer

Flow Unit Factor 240.0

Inspect by: XSS Bin 1 | ○ XSS Mean

XSS Parameters

Set frames as TFR PSR

SVD Denoise check SV Rank 5

Global MAX. 4621.046 MIN. 0.0 auto

Dimension-wise Scaling auto manual

CHROMS Parameters

MAX. 1.000 Min. 0.0 auto

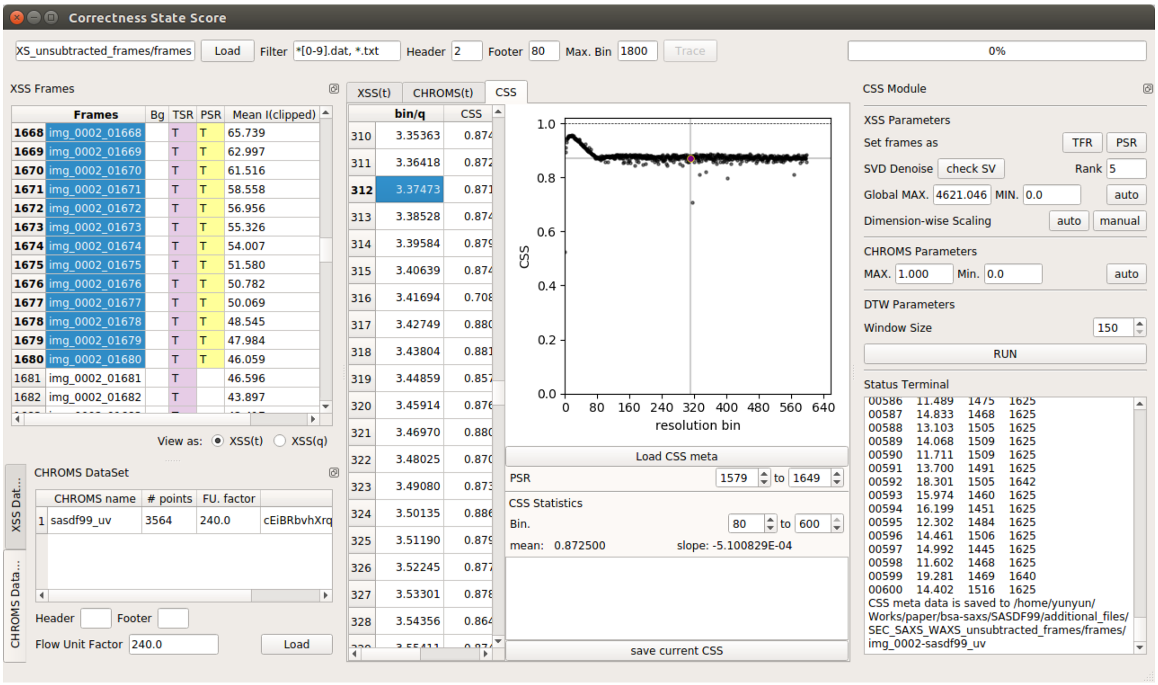
DTW Parameters

Window Size 150

Status Terminal

```
00175   0.213  213.047   643
00176   0.214  194.017   696
00177   0.214  225.746   662
00178   0.214  234.963   629
00179   0.213  201.468   626
00180   0.214  208.229   679
00181   0.213  195.583   629
00182   0.214  236.473   822
00183   0.214  208.453   625
00184   0.214  253.593   712
00185   0.214  206.724   700
00186   0.214  248.565   595
00187   0.215  226.085   624
00188   0.214  253.510   757
00189   0.214  265.250   583
00190   0.214  283.013   702
00191   0.214  280.204   720
00192   0.215  234.214   650
00193   0.215  255.804   722
```

Figure 5.6 The GUI of CSS. Calculating ***CSS***(q).



XS_unsubtracted_frames/frames | Filter *[0-9].dat, *.txt | Header 2 | Footer 80 | Max. Bin 1800

	Frames	Bg	TSR	PSR	Mean I(clipped)
1668	img_0002_01668		T	T	65.739
1669	img_0002_01669		T	T	62.997
1670	img_0002_01670		T	T	61.516
1671	img_0002_01671		T	T	58.558
1672	img_0002_01672		T	T	56.956
1673	img_0002_01673		T	T	55.326
1674	img_0002_01674		T	T	54.007
1675	img_0002_01675		T	T	51.580
1676	img_0002_01676		T	T	50.782
1677	img_0002_01677		T	T	50.069
1678	img_0002_01678		T	T	48.545
1679	img_0002_01679		T	T	47.984
1680	img_0002_01680		T	T	46.059
1681	img_0002_01681		T		46.596
1682	img_0002_01682		T		43.897

View as: ● XSS(t) ○ XSS(q)

	CHROMS name	# points	FU. factor	
1	sasdf99_uv	3564	240.0	cEiBRbvhXrq

Header □ Footer □

Flow Unit Factor 240.0

	bin/q	CSS
310	3.35363	0.874
311	3.36418	0.872
312	3.37473	0.871
313	3.38528	0.874
314	3.39584	0.879
315	3.40639	0.874
316	3.41694	0.708
317	3.42749	0.880
318	3.43804	0.881
319	3.44859	0.857
320	3.45914	0.876
321	3.46970	0.880
322	3.48025	0.870
323	3.49080	0.873
324	3.50135	0.886
325	3.51190	0.879
326	3.52245	0.877
327	3.53301	0.878
328	3.54356	0.864

PSR 1579 to 1649

CSS Statistics

Bin. 80 to 600

mean: 0.872500 slope: -5.100829E-04

Set frames as

SVD Denoise | Rank 5

Global MAX. 4621.046 MIN. 0.0

Dimension-wise Scaling

MAX. 1.000 Min. 0.0

Window Size 150

```
00586   11.489   1475   1625
00587   14.833   1468   1625
00588   13.103   1505   1625
00589   14.068   1509   1625
00590   11.711   1509   1625
00591   13.700   1491   1625
00592   18.301   1505   1642
00593   15.974   1460   1625
00594   16.199   1451   1625
00595   12.302   1484   1625
00596   14.461   1506   1625
00597   14.992   1445   1625
00598   11.602   1468   1625
00599   19.281   1469   1640
00600   14.402   1516   1625
CSS meta data is saved to /home/yunyun/
Works/paper/bsa-saxs/SASDF99/additional_files/
SEC_SAXS_WAXS_unsubtracted_frames/frames/
img_0002-sasdf99_uv
```

Figure 5.7 The GUI of CSS. An overview of $\boldsymbol{CSS}(q)$ and $\boldsymbol{CSS}(q)$ statistics

Bibliography

Acerbo AS, Cook MJ, Gillilan RE. 2015. Upgrade of MacCHESS facility for X-ray scattering of biological macromolecules in solution. *J. Synchrotron Radiat.* 22(1):180–86

Adler J, Öktem O. 2017. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Probl.* 33(12):124007

Andrae R, Schulze-Hartung T, Melchior P. 2010. Dos and don'ts of reduced chi-squared. *arXiv*

Antonov LD, Olsson S, Boomsma W, Hamelryck T. 2016. Bayesian inference of protein ensembles from SAXS data. *Phys. Chem. Chem. Phys.* 18(8):5832–38

Ascierto PA, Kalos M, Schaer DA, Callahan MK, Wolchok JD. 2013. Biomarkers for immunostimulatory monoclonal antibodies in combination strategies for melanoma and other tumor types. *Clin. Cancer Res.* 19(5):1009–20

Aspeslagh S, Postel-Vinay S, Rusakiewicz S, Soria J-C, Zitvogel L, Marabelle A. 2016. Rationale for anti-OX40 cancer immunotherapy. *Eur. J. Cancer*. 52:50–66

Aurenhammer F. 1991. Voronoi diagrams---a survey of a fundamental geometric data structure. *ACM Comput. Surv.* 23(3):345–405

Ayyer K, Yefanov OM, Oberthür D, Roy-Chowdhury S, Galli L, et al. 2016. Macromolecular diffractive imaging using imperfect crystals. *Nature*. 530(7589):202–6

Barber CB, Dobkin DP, Huhdanpaa H. 1996. The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.* 22(4):469–83

Bartkowiak T, Singh S, Yang G, Galvan G, Haria D, et al. 2015. Unique potential of 4-1BB agonist antibody to promote durable regression of HPV+ tumors when combined with an E6/E7 peptide vaccine. *Proc Natl Acad Sci USA*. 112(38):E5290-9

Baxter RJ. 1970. Ornstein–zernike relation and percus–yevick approximation for fluid mixtures. *J. Chem. Phys.* 52(9):4559–62

Beiranvand V, Hare W, Lucet Y. 2017. Best practices for comparing optimization algorithms. *Optim. Eng.* 18(4):815–48

Bernadó P, Modig K, Grela P, Svergun DI, Tchorzewski M, et al. 2010. Structure and dynamics of ribosomal protein L12: an ensemble model based on SAXS and NMR relaxation. *Biophys. J.* 98(10):2374–82

Bernadó P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI. 2007. Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem.*

Soc. 129(17):5656–64

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(24):10080–88

Bertini I, Giachetti A, Luchinat C, Parigi G, Petoukhov MV, et al. 2010. Conformational space of flexible biological macromolecules from average data. *J. Am. Chem. Soc.* 132(38):13553–58

Bevington PR, Robinson DK, Blair JM, Mallinckrodt AJ, McKay S. 1993. Data reduction and error analysis for the physical sciences. *Comput. Phys.* 7(4):415

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

Blobel J, Bernadó P, Svergun DI, Tauler R, Pons M. 2009. Low-resolution structures of transient protein-protein complexes using small-angle X-ray scattering. *J. Am. Chem. Soc.* 131(12):4378–86

Boeing G. 2016. Visual Analysis of Nonlinear Dynamical Systems: Chaos, Fractals, Self-Similarity and the Limits of Prediction. *Systems*. 4(4):37

Bohm G, Zech G. 2010. *Introduction to Statistics and Data Analysis for Physicists*

Branch MA, Coleman TF, Li Y. 1999. A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems. *SIAM J. Sci. Comput.* 21(1):1–23

Brennich ME, Kieffer J, Bonamis G, De Maria Antolinos A, Hutin S, et al. 2016. Online data analysis at the ESRF bioSAXS beamline, BM29. *J. Appl. Crystallogr.* 49(1):203–12

Brookes E, Vachette P, Rocco M, Pérez J. 2016. US-SOMO HPLC-SAXS module: dealing with capillary fouling and extraction of pure component patterns from poorly resolved SEC-SAXS data. *J. Appl. Crystallogr.* 49(Pt 5):1827–41

Bruhns P, Iannascoli B, England P, Mancardi DA, Fernandez N, et al. 2009. Specificity and affinity of human Fcgamma receptors and their polymorphic variants for human IgG subclasses. *Blood*. 113(16):3716–25

Brunner-Popela J, Glatter O. 1997. Small-Angle Scattering of Interacting Particles. I. Basic Principles of a Global Evaluation Technique. *J. Appl. Crystallogr.* 30(4):431–42

Byrd RH, Hribar ME, Nocedal J. 1999. An Interior Point Algorithm for Large-Scale Nonlinear Programming. *SIAM J. Optim.* 9(4):877–900

Byrd RH, Lu P, Nocedal J, Zhu C. 1995. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Sci. Comput.* 16(5):1190–1208

Cammarata M, Levantino M, Schotte F, Anfinrud PA, Ewald F, et al. 2008. Tracking the structural dynamics of proteins in solution using time-resolved wide-angle X-ray scattering. *Nat. Methods*. 5(10):881–86

Capsule -- from Wolfram MathWorld. http://mathworld.wolfram.com/Capsule.html

Charleux B, Delaittre G, Rieger J, D'Agosto F. 2012. Polymerization-Induced Self-Assembly: From Soluble Macromolecules to Block Copolymer Nano-Objects in One Step. *Macromolecules*. 45(17):6753–65

Chaudhuri B, Muñoz IG, Qian S, Urban VS, eds. 2017. *Biological Small Angle Scattering: Techniques, Strategies and Tips*, Vol. 1009. Singapore: Springer Singapore

Cheng P, Peng J, Zhang Z. 2017. SAXS-Oriented Ensemble Refinement of Flexible Biomolecules. *Biophys. J.* 112(7):1295–1301

Chen P-C, Hub JS. 2014. Validating solution ensembles from molecular dynamics simulation by wide-angle X-ray scattering data. *Biophys. J.* 107(2):435–47

Chen P-C, Hub JS. 2015. Structural Properties of Protein-Detergent Complexes from SAXS and MD Simulations. *J. Phys. Chem. Lett.* 6(24):5116–21

Chi EY, Krishnan S, Randolph TW, Carpenter JF. 2003. Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation. *Pharm. Res.* 20(9):1325–36

Clapp LH, Gurney AM. 1992. ATP-sensitive K+ channels regulate resting potential of pulmonary arterial smooth muscle cells. *Am. J. Physiol.* 262(3 Pt 2):H916-20

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 25(11):1422–23

Convex hull algorithms - Wikipedia. https://en.wikipedia.org/wiki/Convex_hull_algorithms

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

Coutts JA, Roberts J, Mills TC. 1997. Parameter stability in the market model: tests and time varying paramete r estimation with UK data. *J. Royal Statistical Soc. D*. 46(1):57–70

Daughdrill GW, Kashtanov S, Stancik A, Hill SE, Helms G, et al. 2012. Understanding the structural ensembles of a highly extended disordered protein. *Mol. Biosyst.* 8(1):308–19

DeAngelis GC, Ohzawa I, Freeman RD. 1995. Receptive-field dynamics in the central visual pathways. *Trends Neurosci.* 18(10):451–58

Debye P. 1915. Zerstreuung von Röntgenstrahlen. *Ann. Phys.* 351(6):809–23

De Maria Antolinos A, Pernot P, Brennich ME, Kieffer J, Bowler MW, et al. 2015. ISPyB for BioSAXS, the gateway to user autonomy in solution scattering experiments. *Acta Crystallogr. D Biol. Crystallogr.* 71(Pt 1):76–85

Dillon TM, Ricci MS, Vezina C, Flynn GC, Liu YD, et al. 2008. Structural and functional characterization of disulfide isoforms of the human IgG2 subclass. *J. Biol. Chem.* 283(23):16206–15

Ericson C. 2004. *Real-Time Collision Detection (the Morgan Kaufmann Series In Interactive 3-d Technology)*. Amsterdam: Crc Press

Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, et al. 2006. Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics*. Chapter 5:Unit 5.6

Faham S, Watanabe A, Besserer GM, Cascio D, Specht A, et al. 2008. The crystal structure of a sodium galactose transporter reveals mechanistic insights into Na+/sugar symport. *Science*. 321(5890):810–14

Feigin LA, Svergun DI. 1987. *Structure Analysis by Small-Angle X-Ray and Neutron Scattering*. Boston, MA: Springer US

Filatova OA, Deecke VB, Ford JKB, Matkin CO, Barrett-Lennard LG, et al. 2012. Call diversity in the North Pacific killer whale populations: implications for dialect evolution and population history. *Animal Behaviour*. 83(3):595–603

Foley JD, Van Dam A, Feiner SK, Hughes JF. 1995. *Computer Graphics: Principles And Practice In C (2nd Edition)*. Reading, Mass: Addison-wesley Professional. 2nd ed.

Förster S, Fischer S, Zielske K, Schellbach C, Sztucki M, et al. 2011. Calculation of scattering-patterns of ordered nano- and mesoscale materials. *Adv. Colloid Interface Sci.* 163(1):53–83

Förster S, Timmann A, Schellbach C, Frömsdorf A, Kornowski A, et al. 2007. Order causes secondary Bragg peaks in soft materials. *Nat. Mater.* 6(11):888–93

Forstner M. 2000. SAXS, SANS and X-ray crystallography as complementary methods in the study of biological form and function. *J. Appl. Crystallogr.* 33(3):519–23

Fournet G. 1951. Étude théorique et expérimentale de la diffusion des rayons X par les ensembles denses de particules. *bulmi*. 74(1):37–172

Franke D, Jeffries CM, Svergun DI. 2015. Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra. *Nat. Methods*. 12(5):419–22

Franke D, Svergun DI. 2009. DAMMIF , a program for rapidab-initio shape determination in small-angle scattering. *J. Appl. Crystallogr.* 42(2):342–46

Frick RW. 1995. Accepting the null hypothesis. *Mem. Cognit.* 23(1):132–38

Gao F, Han L. 2012. Implementing the Nelder-Mead simplex algorithm with adaptive parameters. *Comput. Optim. Appl.* 51(1):259–77

Glatter O, Kratky O, eds. 1982. *Small Angle X-Ray Scattering*. Academic Press. illustrated, reprint ed.

Golubev A. 2017. Exponentially modified peak functions in biomedical sciences and related disciplines. *Comput. Math. Methods Med.* 2017:7925106

Gombosˇi M, Žalik B. 2005. Point-in-polygon tests for geometric buffers. *Comput. Geosci.* 31(10):1201–12

Gorba C, Miyashita O, Tama F. 2008. Normal-mode flexible fitting of high-resolution structure of biological molecules toward one-dimensional low-resolution data. *Biophys. J.* 94(5):1589–99

Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. 2006. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*. 22(21):2695–96

Grant TD. 2018. Ab initio electron density determination directly from solution scattering data. *Nat. Methods*. 15(3):191–93

Grant TD, Luft JR, Wolfley JR, Tsuruta H, Martel A, et al. 2011. Small angle X-ray scattering as a complementary tool for high-throughput structural studies. *Biopolymers*. 95(8):517–30

Grassmé H, Jendrossek V, Bock J, Riehle A, Gulbins E. 2002. Ceramide-rich membrane rafts mediate CD40 clustering. *J. Immunol.* 168(1):298–307

Grudinin S, Garkavenko M, Kazennov A. 2017. Pepsi-SAXS: an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. *Acta Crystallogr. D Struct. Biol.* 73(Pt 5):449–64

Grushka E. 1972. Characterization of exponentially modified Gaussian peaks in chromatography. *Anal. Chem.* 44(11):1733–38

Guinier A, Fournet G, Walker CB, Vineyard GH. 1956. *small-angle scattering of x-rays*. *Phys. Today*. 9(8):38–39

Haswell LE, Glennie MJ, Al-Shamkhani A. 2001. Analysis of the oligomeric requirement for signaling by CD40 using soluble multimeric forms of its ligand, CD154. *Eur. J.*

Immunol. 31(10):3094–3100

Heckbert PS. 1994. *Graphics Gems*. Boston: AP Professional (Academic Press). 4th ed.

Hong L, Qu Y, Dhupia JS, Sheng S, Tan Y, Zhou Z. 2017. A novel vibration-based fault diagnostic algorithm for gearboxes under speed fluctuations without rotational speed measurement. *Mech. Syst. Signal Process.* 94:14–32

Hopkins JB, Gillilan RE, Skou S. 2017. BioXTAS RAW: improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. *J. Appl. Crystallogr.* 50(Pt 5):1545–53

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

Hubbard MA, Thorkildson P, Kozel TR, AuCoin DP. 2013. Constant domains influence binding of mouse-human chimeric antibodies to the capsular polypeptide of Bacillus anthracis. *Virulence*. 4(6):483–88

Ivanović MT, Hermann MR, Wójcik M, Pérez J, Hub JS. 2019. SAXS curves of detergent micelles: effects of asymmetry, shape fluctuations, disorder, and atomic details. *BioRxiv*

Jeffries CM, Graewert MA, Blanchet CE, Langley DB, Whitten AE, Svergun DI. 2016. Preparing monodisperse macromolecular samples for successful biological small-angle X-ray and neutron-scattering experiments. *Nat. Protoc.* 11(11):2122–53

Jiménez-García B, Pons C, Svergun DI, Bernadó P, Fernández-Recio J. 2015. pyDockSAXS: protein-protein complex structure by SAXS and computational docking. *Nucleic Acids Res.* 43(W1):W356-61

Josts I, Niebling S, Gao Y, Levantino M, Tidow H, Monteiro D. 2018a. Photocage-initiated time-resolved solution X-ray scattering investigation of protein dimerization. *IUCrJ.* 5(Pt 6):667–72

Josts I, Nitsche J, Maric S, Mertens HD, Moulin M, et al. 2018b. Conformational States of ABC Transporter MsbA in a Lipid Environment Investigated by Small-Angle Scattering Using Stealth Carrier Nanodiscs. *Structure*. 26(8):1072-1079.e4

Joung IS, Cheatham TE. 2008. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B*. 112(30):9020–41

Kabanikhin SI. 2008. Definitions and examples of inverse and ill-posed problems. *J. Inverse Ill Posed Probl.* 16(4):

Kalambet Y, Kozmin Y, Mikhailova K, Nagaev I, Tikhonov P. 2011. Reconstruction of chromatographic peaks using the exponentially modified Gaussian function. *J. Chemom.* 25(7):352–56

Karplus M, McCammon JA. 2002. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* 9(9):646–52

Karplus PA, Diederichs K. 2015. Assessing and maximizing data quality in macromolecular crystallography. *Curr. Opin. Struct. Biol.* 34:60–68

Keogh EJ, Pazzani MJ. 2001. Derivative dynamic time warping

Keogh E, Ratanamahatana CA. 2005. Exact indexing of dynamic time warping. *Knowl. Inf. Syst.* 7(3):358–86

Kieffer J, Wright JP. 2013. PyFAI: a Python library for high performance azimuthal integration on GPU. *Powder Diffr.* 28(S2):S339–50

Kikhney AG, Svergun DI. 2015. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett.* 589(19 Pt A):2570–77

Korasick DA, Tanner JJ. 2018. Determination of protein oligomeric structure from small-angle X-ray scattering. *Protein Sci.* 27(4):814–24

Krishnamurthy H, Gouaux E. 2012. X-ray structures of LeuT in substrate-free outward-open and apo inward-open states. *Nature*. 481(7382):469–74

Krishnamurthy H, Piscitelli CL, Gouaux E. 2009. Unlocking the molecular secrets of sodium-coupled transporters. *Nature*. 459(7245):347–55

Kunji ERS, Harding M, Butler PJG, Akamine P. 2008. Determination of the molecular mass and dimensions of membrane proteins by size exclusion chromatography. *Methods*. 46(2):62–72

Lasdon L, Duarte A, Glover F, Laguna M, Martí R. 2010. Adaptive memory programming for constrained global optimization. *Comput. Oper. Res.* 37(8):1500–1509

Lee S, Mao A, Bhattacharya S, Robertson N, Grisshammer R, et al. 2016. How Do Short Chain Nonionic Detergents Destabilize G-Protein-Coupled Receptors? *J. Am. Chem. Soc.* 138(47):15425–33

Levantino M, Yorke BA, Monteiro DC, Cammarata M, Pearson AR. 2015. Using synchrotrons and XFELs for time-resolved X-ray crystallography and solution scattering experiments on biomolecules. *Curr. Opin. Struct. Biol.* 35:41–48

Lipfert J, Columbus L, Chu VB, Lesley SA, Doniach S. 2007. Size and shape of detergent micelles determined by small-angle X-ray scattering. *J. Phys. Chem. B*. 111(43):12427–38

Liu H, Hexemer A, Zwart PH. 2012. The *Small Angle Scattering ToolBox* (*SASTBX*): an open-source software for biomolecular small-angle scattering. *J. Appl. Crystallogr.* 45(3):587–93

Liu YD, Chen X, Enk JZ, Plant M, Dillon TM, Flynn GC. 2008. Human IgG2 antibody disulfide rearrangement in vivo. *J. Biol. Chem.* 283(43):29266–72

Liu YD, Wang T, Chou R, Chen L, Kannan G, et al. 2013. IgG2 disulfide isoform conversion kinetics. *Mol. Immunol.* 54(2):217–26

Loll PJ. 2014. Membrane proteins, detergents and crystals: what is the state of the art? *Acta Crystallogr. F Struct. Biol. Commun.* 70(Pt 12):1576–83

Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. 2012. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.* 40(Database issue):D370-6

Loose M, Malla S, Stout M. 2016. Real-time selective sequencing using nanopore technology. *Nat. Methods*. 13(9):751–54

Lo VL-X. 2011. *Iterative projection algorithms and applications in x-ray crystallography*. Doctoral dissertation thesis

López Cárdenas DC. 2016. *Systematic evaluation of ill-posed problems in model-based parameter estimation and experimental design*. Doctoral dissertation thesis

Lu Z, Chen X, Li Q, Zhang X, Zhou P. 2014. A Hand Gesture Recognition Framework and Wearable Gesture-Based Interaction Prototype for Mobile Devices. *IEEE Trans. Hum. Mach. Syst.* 44(2):293–99

Lyons J, Biswas N, Sharma A, Dehzangi A, Paliwal KK. 2014. Protein fold recognition by alignment of amino acid residues using kernelized dynamic time warping. *J. Theor. Biol.* 354:137–45

Mahajan S, Sanejouand Y-H. 2015. On the relationship between low-frequency normal modes and the large-scale conformational changes of proteins. *Arch. Biochem. Biophys.* 567:59–65

Malaby AW, Chakravarthy S, Irving TC, Kathuria SV, Bilsel O, Lambright DG. 2015. Methods for analysis of size-exclusion chromatography-small-angle X-ray scattering and reconstruction of protein scattering. *J. Appl. Crystallogr.* 48(Pt 4):1102–13

Mayerhöfer TG, Pipa AV, Popp J. 2019. Beer's Law-Why Integrated Absorbance Depends Linearly on Concentration. *ChemPhysChem*. 20(21):2748–53

McKinnon KIM. 1998. Convergence of the Nelder--Mead Simplex Method to a

Nonstationary Point. *SIAM J. Optim.* 9(1):148–58

Meisburger SP, Taylor AB, Khan CA, Zhang S, Fitzpatrick PF, Ando N. 2016. Domain Movements upon Activation of Phenylalanine Hydroxylase Characterized by Crystallography and Chromatography-Coupled Small-Angle X-ray Scattering. *J. Am. Chem. Soc.* 138(20):6506–16

Merzel F, Smith JC. 2002. SASSIM: a method for calculating small-angle X-ray and neutron scattering and the associated molecular envelope from explicit-atom models of solvated proteins. *Acta Crystallogr. D Biol. Crystallogr.* 58(Pt 2):242–49

Mo Y, Lee B-K, Ankner JF, Becker JM, Heller WT. 2008. Detergent-associated solution conformations of helical and beta-barrel membrane proteins. *J. Phys. Chem. B*. 112(42):13349–54

Mura S, Nicolas J, Couvreur P. 2013. Stimuli-responsive nanocarriers for drug delivery. *Nat. Mater.* 12(11):991–1003

Myers C, Rabiner L, Rosenberg A. 1980. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Trans. Acoust.* 28(6):623–35

Møller M, Nielsen SS, Ramachandran S, Li Y, Tria G, et al. 2013. Small angle X-ray scattering studies of mitochondrial glutaminase C reveal extended flexible regions, and link oligomeric state with enzyme activity. *PLoS ONE*. 8(9):e74783

Nadarajah S. 2011. Making the Cauchy work. *Braz. J. Probab. Stat.* 25(1):99–120

Narayanan T, Wacklin H, Konovalov O, Lund R. 2017. Recent applications of synchrotron radiation and neutrons in the study of soft matter. *Crystallogr. Rev.* 23(3):160–226

Nash SG. 2000. A survey of truncated-Newton methods. *Journal of Computational and Applied Mathematics*. 124(1–2):45–59

Neale C, Ghanei H, Holyoake J, Bishop RE, Privé GG, Pomès R. 2013. Detergent-mediated protein aggregation. *Chem. Phys. Lipids*. 169:72–84

Neal B. 2019. *On the Bias-Variance Tradeoff: Textbooks Need an Update*. Undergraduate thesis thesis

Oliver RC, Lipfert J, Fox DA, Lo RH, Doniach S, Columbus L. 2013. Dependence of micelle size and shape on detergent alkyl chain length and head group. *PLoS ONE*. 8(5):e62488

Orellana L. 2019. Large-Scale Conformational Changes and Protein Function: Breaking the in silico Barrier. *Front. Mol. Biosci.* 6:117

Orr C. 2019. *Structure function relationship of anti-CD40 monoclonal antibodies* . Doctoral dissertation thesis

Orthaber D, Bergmann A, Glatter O. 2000. SAXS experiments on absolute scale with Kratky systems using water as a secondary standard. *J. Appl. Crystallogr.* 33(2):218–25

O'Rourke J. 1998. *Computational Geometry in C*. Cambridge University Press

Panjkovich A, Svergun DI. 2016. Deciphering conformational transitions of proteins by small angle X-ray scattering and normal mode analysis. *Phys. Chem. Chem. Phys.* 18(8):5707–19

Pearson AR, von Stetten D, Huse N. 2015. If You Can Get a Crystal Structure, Why Bother with Anything Else? *Synchrotron Radiat. News*. 28(6):10–14

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

Pérez J, Koutsioubas A. 2015. Memprot: a program to model the detergent corona around a membrane protein based on SEC-SAXS data. *Acta Crystallogr. D Biol. Crystallogr.* 71(Pt 1):86–93

Pernot P, Round A, Barrett R, De Maria Antolinos A, Gobbo A, et al. 2013. Upgraded ESRF BM29 beamline for SAXS on macromolecules in solution. *J. Synchrotron Radiat.* 20(Pt 4):660–64

Perutz MF. 1949. An X-ray study of horse methemoglobin. *Proc. R. Soc. Lond. A Math. Phys. Sci.* 195(1043):474–99

Petoukhov MV, Franke D, Shkumatov AV, Tria G, Kikhney AG, et al. 2012. New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* 45(Pt 2):342–50

Petoukhov MV, Konarev PV, Kikhney AG, Svergun DI. 2007. ATSAS 2.1 – towards automated and web-supported small-angle scattering data analysis. *J. Appl. Crystallogr.* 40(s1):s223–28

Petoukhov MV, Svergun DI. 2005. Global rigid body modeling of macromolecular complexes against small-angle scattering data. *Biophys. J.* 89(2):1237–50

Poitevin F, Orland H, Doniach S, Koehl P, Delarue M. 2011. AquaSAXS: a web server for computation and fitting of SAXS profiles with non-uniformally hydrated atomic models. *Nucleic Acids Res.* 39(Web Server issue):W184-9

Polyakova A. 2015. *Applying complementary structural techniques to elucidate structure-function relationships of the bacterial Na+-hydantoin transporter Mhp1*. Doctoral dissertation thesis. University of Leeds

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 2007. *Numerical Recipes 3rd Edition:*

The Art Of Scientific Computing. Cambridge, UK: Cambridge University Press. 3rd ed.

Purushothaman S, Ayet San Andrés S, Bergmann J, Dickel T, Ebert J, et al. 2017. Hyper-EMG: A new probability distribution function composed of Exponentially Modified Gaussian distributions to analyze asymmetric peak shapes in high-resolution time-of-flight mass spectrometry. *Int. J. Mass Spectrom.*

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(3):191–285

Ramachandran PL, Lovett JE, Carl PJ, Cammarata M, Lee JH, et al. 2011. The short-lived signaling state of the photoactive yellow protein photoreceptor revealed by combined structural probes. *J. Am. Chem. Soc.* 133(24):9395–9404

Rambo RP, Tainer JA. 2013. Accurate assessment of mass, models and resolution by small-angle scattering. *Nature*. 496(7446):477–81

Różycki B, Kim YC, Hummer G. 2011. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure*. 19(1):109–16

Ryan TM, Trewhella J, Murphy JM, Keown JR, Casey L, et al. 2018. An optimized SEC-SAXS system enabling high X-ray dose for rapid SAXS assessment with correlated UV measurements for biomolecular structure analysis. *J. Appl. Crystallogr.* 51(1):97–111

S. Djordjevic S. 2019. Unconstrained Optimization Methods: Conjugate Gradient Methods and Trust-Region Methods. In *Applied Mathematics*, ed. B Carpentieri. IntechOpen

Sakoe H, Chiba S. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust.* 26(1):43–49

Salomon-Ferrer R, Case DA, Walker RC. 2013. An overview of the Amber biomolecular simulation package. *WIREs Comput Mol Sci*. 3(2):198–210

Salvador S, Chan P. 2007. FastDTW: Toward accurate dynamic time warping in linear time and space. *IDA*. 11(5):561–80

Schindler CEM, de Vries SJ, Sasse A, Zacharias M. 2016. SAXS Data Alone can Generate High-Quality Models of Protein-Protein Complexes. *Structure*. 24(8):1387–97

Schmitz KS. 1990. *Introduction to Dynamic Light Scattering by Macromolecules*. Elsevier

Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A. 2013. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* 105(4):962–74

Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A. 2016. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res.* 44(W1):424–29

Seddon AM, Curnow P, Booth PJ. 2004. Membrane proteins, lipids and detergents: not just a soap opera. *Biochim. Biophys. Acta*. 1666(1–2):105–17

Shevchuk R, Hub JS. 2017. Bayesian refinement of protein structures and ensembles against SAXS data using molecular dynamics. *PLoS Comput. Biol.* 13(10):e1005800

Shevtsov M, Soupikov A, Kapustin A. 2007. Ray-Triangle Intersection Algorithm for Modern CPU Architectures

Shewchuk JR. 1994. An Introduction to the Conjugate Gradient Method Without the Agonizing Pain. Carnegie Mellon University

Srivastava SR, Zadafiya P, Mahalakshmi R. 2018. Hydrophobic mismatch modulates stability and plasticity of human mitochondrial VDAC2. *Biophys. J.* 115(12):2386–94

Stansfeld PJ, Goose JE, Caffrey M, Carpenter EP, Parker JL, et al. 2015. MemProtMD: Automated Insertion of Membrane Protein Structures into Explicit Lipid Membranes. *Structure*. 23(7):1350–61

Svergun DI. 1992. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J. Appl. Crystallogr.* 25(4):495–503

Svergun DI. 1994. Solution scattering from biopolymers: advanced contrast-variation data analysis. *Acta Crystallogr. A Found. Crystallogr.* 50(3):391–402

Svergun DI. 1999. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.* 76(6):2879–86

Svergun DI, Koch MHJ, Timmins PA, May RP. 2013. *Small Angle X-Ray and Neutron Scattering from Solutions of Biological Macromolecules*. Oxford University Press

Svergun DI, Petoukhov MV, Koch MH. 2001. Determination of domain structure of proteins from X-ray solution scattering. *Biophys. J.* 80(6):2946–53

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

Sweeny EA, Gupta K, Shorter J. 2013. Conformational Changes of Hsp104 Revealed through Small Angle X-Ray Scattering (SAXS). *Biophys. J.* 104(2):182a

Syberg F, Suveyzdis Y, Kötting C, Gerwert K, Hofmann E. 2012. Time-resolved Fourier transform infrared spectroscopy of the nucleotide-binding domain from the ATP-binding

Cassette transporter MsbA: ATP hydrolysis is the rate-limiting step in the catalytic cycle. *J. Biol. Chem.* 287(28):23923–31

Tan C-Y, Huang Y-X. 2015. Dependence of refractive index on concentration and temperature in electrolyte solution, polar solution, nonpolar solution, and protein solution. *J. Chem. Eng. Data*. 60(10):2827–33

Thompson MC, Barad BA, Wolff AM, Sun Cho H, Schotte F, et al. 2019. Temperature-jump solution X-ray scattering reveals distinct motions in a dynamic enzyme. *Nat. Chem.* 11(11):1058–66

Tian X, Vestergaard B, Thorolfsson M, Yang Z, Rasmussen HB, Langkilde AE. 2015. In-depth analysis of subclass-specific conformational preferences of IgG antibodies. *IUCrJ*. 2(Pt 1):9–18

Tirion MM. 1996. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* 77(9):1905–8

Toby BH. 2006. *R* factors in Rietveld analysis: How good is good enough? *Powder Diffr.* 21(1):67–70

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

Ujwal R, Bowie JU. 2011. Crystallizing membrane proteins using lipidic bicelles. *Methods*. 55(4):337–41

Vakili M, Merkens S, Gao Y, Gwozdz PV, Vasireddi R, et al. 2019. 3D Micromachined Polyimide Mixing Devices for in Situ X-ray Imaging of Solution-Based Block Copolymer Phase Transitions. *Langmuir*. 35(32):10435–45

Valentini E, Kikhney AG, Previtali G, Jeffries CM, Svergun DI. 2015. SASBDB, a repository for biological small-angle scattering data. *Nucleic Acids Res.* 43(Database issue):D357-63

van Duyl BY, Rijkers DTS, de Kruijff B, Killian JA. 2002. Influence of hydrophobic mismatch and palmitoylation on the association of transmembrane alpha-helical peptides with detergent-resistant membranes. *FEBS Lett.* 523(1–3):79–84

Velankar S, Alhroub Y, Alili A, Best C, Boutselakis HC, et al. 2011. Pdbe: protein data bank in europe. *Nucleic Acids Res.* 39(Database issue):D402-10

Vestergaard B. 2016. Analysis of biostructural changes, dynamics, and interactions - Small-angle X-ray scattering to the rescue. *Arch. Biochem. Biophys.* 602:69–79

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, et al. 2019. SciPy

1.0--Fundamental Algorithms for Scientific Computing in Python. *arXiv*

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*

Vonderheide RH, Glennie MJ. 2013. Agonistic CD40 antibodies and cancer therapy. *Clin. Cancer Res.* 19(5):1035–43

Wajant H. 2015. Principles of antibody-mediated TNF receptor activation. *Cell Death Differ.* 22(11):1727–41

Wales DJ, Doye JPK. 1997. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *J. Phys. Chem. A*. 101(28):5111–16

Ward A, Reyes CL, Yu J, Roth CB, Chang G. 2007. Flexibility in the ABC transporter MsbA: Alternating access with a twist. *Proc Natl Acad Sci USA*. 104(48):19005–10

Weyand S, Shimamura T, Yajima S, Suzuki S, Mirza O, et al. 2008. Structure and molecular mechanism of a nucleobase-cation-symport-1 family transporter. *Science*. 322(5902):709–13

Weyerich B, Brunner-Popela J, Glatter O. 1999. Small-angle scattering of interacting particles. II. Generalized indirect Fourier transformation under consideration of the effective structure factor for polydisperse systems. *J. Appl. Crystallogr.* 32(2):197–209

White AL, Chan HTC, French RR, Willoughby J, Mockridge CI, et al. 2015. Conformation of the human immunoglobulin G2 hinge imparts superagonistic properties to immunostimulatory anticancer antibodies. *Cancer Cell*. 27(1):138–48

Wlodawer A, Minor W, Dauter Z, Jaskolski M. 2008. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J.* 275(1):1–21

Wypych J, Li M, Guo A, Zhang Z, Martinez T, et al. 2008. Human IgG2 antibodies display disulfide-mediated structural isoforms. *J. Biol. Chem.* 283(23):16194–205

Xiang Y, Gong XG. 2000. Efficiency of generalized simulated annealing. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics*. 62(3 Pt B):4473–76

Yamniuk AP, Suri A, Krystek SR, Tamura J, Ramamurthy V, et al. 2016. Functional Antagonism of Human CD40 Achieved by Targeting a Unique Species-Specific Epitope. *J. Mol. Biol.* 428(14):2860–79

Yang Z, Wang C, Zhou Q, An J, Hildebrandt E, et al. 2014. Membrane protein stability can be compromised by detergent interactions with the extramembranous soluble domains.

Protein Sci. 23(6):769–89

Yershova A, Jain S, Lavalle SM, Mitchell JC. 2010. Generating uniform incremental grids on SO(3) using the hopf fibration. *Int. J. Rob. Res.* 29(7):801–12

Yershova A, LaValle SM. 2004. Deterministic sampling methods for spheres and SO(3)

Zaitsev-Doyle JJ, Puchert A, Pfeifer Y, Yan H, Yorke BA, et al. 2019. Synthesis and characterisation of α-carboxynitrobenzyl photocagedl -aspartates for applications in time-resolved structural biology. *RSC Adv.* 9(15):8695–99

Zernike F, Prins JA. 1927. Die beugung von röntgenstrahlen in flüssigkeiten als effekt der molekülanordnung. *Z. Physik*. 41(2–3):184–94

Zhang B, Harder AG, Connelly HM, Maheu LL, Cockrill SL. 2010. Determination of Fab-hinge disulfide connectivity in structural isoforms of a recombinant human immunoglobulin G2 antibody. *Anal. Chem.* 82(3):1090–99

Zhao C, Shukla D. 2018. SAXS-guided Enhanced Unbiased Sampling for Structure Determination of Proteins and Complexes. *Sci. Rep.* 8(1):17748

Zheng W, Tekpinar M. 2011. Accurate flexible fitting of high-resolution protein structures to small-angle x-ray scattering data using a coarse-grained model with implicit hydration shell. *Biophys. J.* 101(12):2981–91

List of Hazardous Chemicals Used

Reagent	GHS codes	GHS hazard and precaution statements
None	--	--