Visualizing how inclusion of higher reciprocal space in SWAXS data analysis improves shape restoration of biomolecules: case of lysozyme

Ashish\textsuperscript{a,b}

\textsuperscript{a}CSIR-Institute of Microbial Technology, Chandigarh, India; \textsuperscript{b}Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India

Communicated by Ramaswamy H. Sarma

\section*{Introduction}

Advancement in optics, electronics and data acquisition/processing programs/protocols has allowed collection of high quality X-ray scattering data from small to wider angles with ease from dilute protein solution samples (Hura et al., 2009; Jacques & Trewella, 2010; Virtanen et al., 2011). In absence of diffraction quality crystals or when limited by the molecular size or monodisperse nature of the biomolecule to be studied by NMR, Small/Wide Angle X-ray Scattering (SWAXS) is a reliable technique to gain insight into structural dynamics of biomolecules. Improvement in sample preparation protocols, technology involved in data collection and downstream processing has substantially increased use of Cryo-Electron Microscopy and some other imaging methods, yet responsiveness to added excipients or experimental conditions cannot be readily studied (Banerjee et al., 2021; Glaeser, 2021; Schaffer et al., 2021). Immobilization or long-term stabilization of biomolecules in structure friendly medium or support films also remains a challenge and continuous development are underway (Crook & Powers, 2020; Glaeser, 2021; Letertre et al., 2021). Side-stepping these limitations, with relative ease SWAXS can provide information as a response to varying physico-chemical conditions including ligands or binding partners and/or temperature/pH (Badmalia et al., 2017; Dhiman et al., 2020; Jacques & Trewella, 2010; Mallik et al., 2012; Pandey et al., 2014). Scattering data is acquired as variation in Intensity, \( I(q) \) as a function of momentum transfer vector, \( q \) defined as in Eq. (1). Here, \( \theta \) and \( \lambda \) are scattering angle away from original path and wavelength of incident X-rays, respectively. If \( q \) is defined as in Eq. (1), then the corresponding length of the vector, \( r \) in real space represented by that particular \( q \) is defined by Eq. (2). Units of \( q \) are in reciprocal space, usually \( \text{Å}^{-1} \) or \( \text{nm}^{-1} \), and accordingly, \( r \) is in \( \text{Å} \) or \( \text{nm} \). Usually, SAXS data from dilute solutions of biomolecules decay close to background noise by \( q \) value of 0.12 – 0.2 \( \text{Å}^{-1} \) and in very few instances there is signal above noise in \( q \) beyond 0.3 \( \text{Å}^{-1} \) or so. By using orthogonal data from biophysical and/or biochemical experiments, we published few examples where large scale conformational rearrangements or associations could be delineated in proteins and their complexes by systematically designed SAXS experiments (Badmalia et al., 2017; Dhiman et al. 2020; Pandey et
al., 2014; Pandey et al., 2014; Sharma et al., 2020). Even shape restorations using limited resolution information coupled with molecular modeling has been able to provide new molecules of functional relevance (Sagar et al., 2017; Sagar et al., 2017; Singh et al., 2020). Big impetus to shape restoration was aided by the research of Svergun and BioSAXS group at EMBL Hamburg (Svergun, 1999), especially the integrated suite of programs and online support to researchers (https://www.embl-hamburg.de/biosaxs/atlas-online/). Scattering data beyond q value of ~0.6 Å⁻¹ is considered as WAXS, and rarely one captures information from dilute protein solutions. Sometimes one can observe unambiguous peak-like profiles known as Bragg peaks due to regular spacing or orderliness of solutes. They usually arise in systems composed of strongly interacting particles which leads to nanoscale organization in solutions.

Amongst the well-characterized proteins by structural methods, hen egg white lysozyme is one protein which is highly positively charged for its size at low pH (4 or below). Since this protein has a high innate propensity to crystallize into diffraction competent forms, lysozyme has been overtly studied experimentally by refining the diffraction data under varied conditions with about 800+ structures of monomeric form and as heat induced dimer and trimeric state (Sharma et al., 2016; Xu et al., 2018). Additionally, monomeric state of the same protein has been studied by analyzing multidimensional NMR (PDB ID 1EBL) which provided insight into conformers accessible to this protein and inherent motions in the molecule in solution (Schwalbe et al., 2001). Earlier, using relatively higher concentrations of lysozyme solutions, SAXS data was used to interpret presence of ‘clusters’ in solutions of this protein based on observed deviation of the classical profile at low q of the datasets (Stradner et al., 2004). This study was conceptually interesting as it tried to address presence of short-range attraction and long-range repulsions leading to formation of small clusters in solution as a pre-requisite to inherent tendency to crystallize in diffraction competent forms. This work was challenged for inconsistency and possible errors arising due to improper processing of SAXS data (Shukla et al., 2008), and was later subject of many follow-up studies (Bergman et al., 2019; Cametti et al., 2013; Cardinaux et al., 2011; Stradner et al., 2006). Any case, SWAXS technique allied with other experimental and/or theoretical or in silico simulation methods still remains promising to reliably track structural features associated with phase transitions of lysozyme from solution to crystalline state. The primary aspect towards this effort depends on how far in q or finer in resolution; SWAXS data can be utilized with certainty and what information additional resolution provides in terms of biomolecular state? To acquire quality WAXS data from dilute solutions, primary advancements have been increase in the flux of photons effectively incident on sample and use of highly sensitive detectors at wider angles. Both these increments also have practical limitations as intensity beyond a certain value induces radiolysis and degradation of biomolecules, and over sensitive detectors pick up higher noise. Reliable signal over noise also depends on the structural properties of the biomolecule being studied and solvent/buffer conditions. Overcoming these limitations, rare cases are known where reliable S/WAXS data could be acquired and used to understand size and shape properties of proteins and their complexes (Makowski et al., 2008; Mallik et al., 2012; Phan-Xuan et al., 2020; Virtanen et al., 2011).

As a precedence to this specific work, three papers need to be cited, where data in WAXS range or higher resolution was addressed arising from lysozyme (Kofinger & Hummer, 2013; Makowski et al., 2008; Virtanen et al., 2011). In 2008, Makowski and coworkers published influence molecular crowding effects on inherent motions in lysozyme by studying WAXS data of concentrated solutions (Makowski et al., 2008). They showed WAXS data from lysozyme solutions varying in protein concentration from 1 to 300 mg/ml. In their case, these authors considered momentum transfer, s as in Eq. (3). (Considering Eqs. (1) and (3), the relationship between their and our work can be

\[ s = 2 \sin(\theta)/\lambda, \]  

\[ q = 2\pi/s, \]  

seen in Eq. (4). In their figures, Bragg peaks can be seen at s values around 0.05–0.06, 0.09–0.1, 0.20–0.25 and 0.26–0.29 Å⁻¹ which would convert to q values of 0.31–0.38, 0.57–0.63, 1.38–1.51, and 1.63–1.82 Å⁻¹. Their focus remained in the error values between different measurements arising from concentrated and dilute solutions of protein. These authors observed higher standard deviations in the intensity values (y-axis) in dilute solutions and concluded that higher variation arise in dilute solutions since protein molecules could undergo higher degree of molecular motions compared to concentrated solutions. Another work discussed SWAXS profile from lysozyme solution(s) mainly with a perspective on modeling hydration layer around proteins using MD simulations and solvation models (Virtanen et al., 2011). They used the datasets published earlier (Makowski et al., 2008) and presented it in l(q) vs. q format, where q was defined as in Eq. (1). It was interesting to note that their SWAXS profiles from experiment and simulations of solvation layers agreed very well for lysozyme. Particularly, their SWAXS profiles showed Bragg peaks around 0.35, 0.6 and 1.5 Å⁻¹. The publication discussed different ways to compute SAXS profiles from atomic resolution structures using CRYSOl program (Svergun et al., 1995). They also compared calculated pair-distance distribution curves/function (PDDF), \( P(r) \) using calculated SWAXS profiles from modeling and experimental data. They considered q range of only 0.06–0.6 Å⁻¹ which provided \( P(r) \) curves with peak around 22 Å and subtle shoulders around 5, 15 and 30 Å, but only in profiles computed from SWAXS profile from solvation simulations.

Another publication addressed how SWAXS information can be applied to extract atomic-resolution structural details about protein molecules in solution (Kofinger & Hummer, 2013). They developed a mathematical method to efficiently compute PDDF as well as SWAXS profiles from MD simulations. While, they presented experimental profile from lysozyme up to q value of 1 Å⁻¹ only, they computed and plotted theoretical SWAXS profile from lysozyme up to 3 Å⁻¹. Like earlier publications cited above, they computed Bragg
peaks around 0.3, 0.6, 1.4 and 1.7 Å⁻¹ for structures of lysozyme. Additionally, they showed $P(r)$ curves computed from theoretical SWAXS profiles by progressively considering higher $q$ ranges. The $P(r)$ curves computed from SWAXS profiles of solvated lysozyme structures showed that besides main peak at 2.2 Å, there were clear multiple peaks around 5, 10 and 15 Å. These peaks were more prominent than previous publication (Virtanen et al., 2011) and became more pronounced as $q_{\text{max}}$ being considered was increased from 1 to 3 Å⁻¹. It was mentioned that these peaks arise from secondary structural elements when calculating high-precision SWAXS measurements. It was discussed that the electron density contrast of proteins against aqueous solution is less in the WAXS region or corresponding to smaller vectors or finer structure. Thus, in experiments with nearly matched contrast protein and solvent/buffer, the WAXS data contains information from finer structure which leads to additional peaks in computed $P(r)$ profiles, particularly in the lower $r$ region. Together, these three papers reported SWAXS profiles from lysozyme in solution both from experiment and theory, and more importantly agreed well with each other. Recently, taking lysozyme as a test case, SWAXS profiles were analyzed to decipher how protein shape undergoes distortion upon drying from solution state (Phan-Xuan et al., 2020). Analysis of the peaks in SWAXS zone indicated that native structure of lysozyme was recovered when hydration levels were more than 35% by weight. Findings were unique and such analysis should help in understanding shape profile of proteins during storage in lyophilized form, particularly with relevance in biopharma market.

Good quality SWAXS data from lysozyme to date was mainly applied to optimization of methods to decode inherent breathing in molecules and/or simulate hydration layer around protein molecules. It remains to be explored how analysis of experimental SWAXS data with higher $q$ data aids in solving or affects re-refining known structural models? The working hypothesis is that such analysis in correlation with other biophysical methods should improve our understanding of biomolecular structural dynamics of proteins in solution. In this work, we acquired SWAXS datasets from a series of dilute lysozyme solutions and progressively increased the $q$ range during data processing and model refinement steps. Along with traditional dummy residue modeling, we performed normal mode analysis of a representative crystal structure and then compared results with structures from crystallography and NMR which suggested that: 1) neither crystallography nor NMR completely captures structural information about lysozyme molecules in solution as seen in SWAXS profile, and 2) inclusion of higher $q$ data in refinement induces more changes in structural models than performing refinement with only low resolution data.

**Materials and methods**

**Proteins for SWAXS experiments**

Lysozyme solutions were prepared as published earlier (Sharma et al., 2016). Only differential step was that eluted protein samples from gel filtration chromatography were not concentrated. Different portions of single peak were collected, pooled and protein concentrations were estimated using extinction coefficient of 2.5 for 1 mg/ml of lysozyme (mass over mass dilutions were done to have final absorbance values between 0.5 to 1 units at 280 nm). Buffer eluting from columns when no proteins were eluting were used as matched buffer for SWAXS experiments. All samples and matched buffer were stored at 10 °C till data collection.

**SWAXS data acquisition and processing**

Scattering experiments mentioned here were done at X9 Beamline of National Synchrotron Light Source, Brookhaven National Lab NY USA under PASS proposal system in May 2014 to PI for different solution SAXS experiments. X-rays with wavelength 1.29Å were utilized, and samples or buffer were flown through a capillary at flow rate of 60 μl/min. The sample stage holding the capillary was maintained at 10 °C using connected water circulator. Three exposures of 30 s were simultaneously captured on two CCD detectors, one for SAXS and another (tilted) for WAXS data. Using Python based programs, the SAXS and WAXS files were averaged, scaled, and merged to obtain SWAXS datasets. Taking example of data for 2 mg/ml sample (as it represents average concentration of protein samples), sequential steps involved in processing, i.e. averaging, scaling, merging and buffer subtraction have been shown in Supplementary Figure S1. SAXS and WAXS data collected was from 0.008 to 0.24 Å⁻¹ and 0.1 to 1.99 Å⁻¹, respectively, and during merging, overlap of data in the $q$ range from 0.1 to 0.24 Å⁻¹ was utilized. Again taking files for sample with 2 mg/ml concentration and its matched buffer, differences in intensity values at $q$ values used during averaging of three frames for SAXS or WAXS data and subsequent merging have been shown in Supplementary Figure S2. In all datasets including buffer, the scale factor applied to WAXS data was about 0.025–0.03 keeping SAXS portion of data at unity. Post-processing, we obtained variation in SWAXS Intensity profiles, $I(q)$ as a function of $q$ where $q$ is defined as in Eq. (1) and units are in 1/A. Kratky analysis, Guinier analysis and residual mapping of the fits were done using ATSAS suite of programs version 3.0.1. Computed or estimated intensity profiles were compared with experimental data using same program. Distance distribution profiles of the interatomic vectors, $P(r)$ were estimated using GOM program integrated within ATSAS suite (Franke et al., 2017; Svergun, 1992). Deconvolution of the multiple peaks in the estimated $P(r)$ profiles was done using Origin Software v 5.0.

All SWAXS datasets presented in this work are available at Small Angle Scattering Biological Data Bank (www.sasbd-b.org). Access codes for datasets are: SASDMC2; SASDMD2; SASDME2; SASDMF2; SASDMG2; SASDMH2; SASDMJ2; SASDMK2; SASDML2.

**Structure restoration and refinement processes**

Using the estimated parameters and vector distribution profiles were used to compute dummy residue models using DAMMIF program (Franke & Svergun, 2009) at online portal...
of ATSAS (https://www.embl-hamburg.de/biosaxs/atsas-online/). 20 models were computed for different ranges of \( q_{\text{max}} \) used for \( P(r) \) analyses without any bias, aligned and averaged using DAMAVER suite of programs (Volkov & Svergun, 2003). The averaged models were aligned over PDB structures solved using crystal diffraction or NMR data by using plugin for SUPREF or SUPCOMB program (Kozin & Svergun, 2001) integrated in PyMol program or offline version. Resultant fits vs. experimental data were plotted using ATSAS analysis program. SREFLLEX program (Panjivich & Svergun, 2016) was considered to employ normal mode analysis on PDB structure to search for models which better fitted the experimental SWAXS data. MOLMOL program v 1.0 was used to compute mean structure of the 50 conformers of lysozyme from NMR data (PDB ID 1E8L). The RMSD variation between \( C^\alpha \) residues was computed using structure alignment option in PyMol program and plotted using same graphics program.

Results and discussion

SWAXS data from dilute solution of lysozyme

Five SWAXS datasets from lysozyme solutions varying in concentration from 0.4 to 5 mg/ml are presented in Figure 1A and Supplementary Figure S3. Double Log plot of the datasets in the \( q \) range of 0.008 – 1.72 Å\(^{-1} \) showed complete lack of any upward or downward trend of data points as \( q \to 0 \) implying aggregation or interparticulate effect in samples under study (Supplementary Figure S3A). Similar to publications reported before (Kofinger & Hummer, 2013; Makowski et al., 2008; Virtanen et al., 2011), our Log \( I(q) \) vs. \( q \) plots of the datasets also showed presence of three Bragg peaks at \( q \to 0.325, 0.65 \) and 1.4 Å\(^{-1} \) in all samples (Figure 1A). Considering Eqs. (1) and (2), these peaks arise from systematic organization of about 19.3, 9.7 and 4.5 Å in real dimensions of Lysozyme molecules. These peaks could be overtones of same organization, but need to be evaluated in more detail later. To confirm that the SWAXS profiles of Lysozyme samples do not alter significantly in the concentration range studied, we scaled all the datasets which confirmed no such deviation (Supplementary Figure S3B). Expectedly, noise was highest for the sample with lowest protein concentration, i.e. 0.4 mg/ml, so before averaging, this dataset was removed (Figure 1A and Supplementary Figure S3C and D). All further analyses were done using this averaged SWAXS profile of lysozyme (Figure 1B). Dimensionless Kratky plot showed the first peak at \( q_R \) value of 1.7 confirming classical globular nature of the molecules in solution (Rambo & Tainer, 2011; Sagar et al., 2020) (right panel in Figure 1C). Corresponding linear fit to the Guinier approximation and residuals of the fit are shown in left panels of Figure 1C.

Emergence of peaks across vectors shorter than \( R_g \) value in distance distribution profiles

Using the averaged SWAXS profile of lysozyme, probability distribution profiles of the pairwise vectors inside the shape of lysozyme were computed. Probability of finding a vector, \( r \) of 0 Å in dimension or equal to longest dimension (\( R_{\text{max}} \)) of lysozyme molecule were constrained to be zero. During search, input \( q_{\text{max}} \) was varied as mentioned in Figure 2 and Supplementary Figure S4. \( R_{\text{max}} \) values were gradually increased till the estimated probability distribution profile or curve, \( P(r) \) was considered to ‘land’ smoothly on the \( x \)-axis. Simultaneously, estimated SWAXS intensity profile representing the corresponding \( P(r) \) profile and its residuals were monitored (Figure 2). Interestingly, estimated \( R_{\text{max}} \) \( R_g \) and \( I_0 \) values remained unchanged at 41.93 Å, 14.06 Å and 248.70 a.u., respectively as these estimations relied mainly on SAXS portion of the data and are less affected by WAXS profile. At same time, automatically estimated Porod volume dropped from about 14000 Å\(^3 \) to about 9300 Å\(^3 \) upon consideration of the structural peaks in the WAXS data. A careful analysis of the Porod plot, i.e. \( I(q) \times q^4 \) vs. \( q \) showed that the curve increases consistently and becomes parallel to \( x \)-axis around \( q \) close to 0.17 Å\(^{-1} \). Around this point, the Porod volume estimated to be around 17104 Å\(^3 \). Further increment in \( q \) value (above 0.17 Å\(^{-1} \)) in the Porod plot led to negative slope of the plot which reduced estimated Porod volume to about 9300 Å\(^3 \). This observation suggests that for reliable results, Porod volume should be estimated within the Porod region or \( q \) value where the first plateau of the plot appears. Possibly, over-consideration of WAXS region led to erroneous volume estimations and thus mass calculations. Molecular masses of lysozyme molecule estimated by considering increased \( q \) values provided lesser than expected values (please see data in Supplementary Figure S4). In practice, one should consider values till Porod region or use estimated \( I_0 \) value in conjunction with standard or characterized samples to estimate molecular mass of the species in solution, as done earlier (Badmalia et al., 2017; Sharma et al., 2020).

Quality estimation of the fits was also comparable, but the profile of the \( P(r) \) curves clearly changed (Figure 2 and Supplementary Figure S4). Here, regularization parameter, ALPHA needs to be addressed here which is a measure of the quality of smoothness of the computed fit (Svergun, 1992). GNOM program optimizes the ALPHA value so that a balance is achieved between the smoothness of the distribution function and fitting to the experimental data. The ALPHA values for the estimations were 42, 17, 97 and 31 as the input \( q_{\text{max}} \) was 0.28, 0.51, 1.01 and 1.72 Å\(^{-1} \), respectively. Higher ALPHA value indicates that during search more emphasis is made on solving a smoother distance distribution function than fitting to experimental data and vice versa. Apart from \( P(r) \) solved using \( q \) range of 0.008–1.01 Å\(^{-1} \), ALPHA values for other solutions were comparable. While considering data points up to 0.28 and 0.51 Å\(^{-1} \), the computed \( P(r) \) profile showed a single Gaussian peak like profile. Despite consideration of the first Bragg peak at 0.35 Å\(^{-1} \), the \( P(r) \) curve remained single peak with maximum close to 20 Å. Differing from this trend, consideration of the second peak at 0.65 Å\(^{-1} \) induced an appearance of an additional peak below the earlier observed one. Going further, usage of data points up to 1.7 Å\(^{-1} \), led to emergence of more peaks at smaller \( r \) values. Usually, one observes peaks or peak-n-shoulder profiles appearing on the \( r \) values which are bigger than...
Figure 1. SWAXS datasets from concentration series of lysozyme protein are shown here. A. Left panel shows the datasets from solutions varying in protein concentration of 5 to 0.4 mg/ml in Log $I(q)$ vs. $q$ mode. Legends mention the color code used for the respective dataset. (Double Log plot of same datasets are shown in Supplementary Figure S3A). B. This plot shows the averaged dataset generated as described in text and shown in Supplementary Figure S3B-D. C. Left upper panel shows the Guinier plot of averaged dataset. Lower left panel shows the residuals of the linear fit. Right panel has the plot of dimensionless Kratky of the same dataset with crosshair at $qR_g = 1.7$. 
the main peak of the \( P(r) \) curve. Such shoulders in the higher \( r \) side of plots have been interpreted to be arising from domains which are positioned away in space from the bulk mass of the protein shape either by rigid or flexible linker(s) (Ashish et al., 2007; 2008; Badmalia et al., 2017; Solanki et al. 2014). Observation of distinct peaks on right side of the main peak in the \( P(r) \) curve computed using \( q_{\text{max}} \) of 1.7 Å\(^{-1}\) implies that there might be sub-domains within the single
lobed shape of this protein. Interestingly, similar peaks were reported by Köfinger and Hummer when they computed \( P(r) \) functions from SWAXS profiles calculated from atomic-resolution models using \( q_{\text{max}} \) values 1–3 Å\(^{-1}\) (Köfinger & Hummer, 2013). Besides, main peak at 18–20 Å for lysozyme, their profiles had peaks around 2, 5, 10–12 and 15 Å, and were attributed to secondary structures in the structure of lysozyme. Moreover, they showed that these peaks remained conserved in the \( P(r) \) profiles computed for GB3 (third IgG-binding domain of Protein G) and Ubiquitin as well.

To delineate peak positions in our \( P(r) \) curves, we attempted deconvolution of the solved profiles by considering them to be composed of multiple Gaussian peaks. The \( P(r) \) curves computed using increasing \( q_{\text{max}} \) value in data processing have been stacked in Figure 3A (right) and corresponding fits over the experimental data have been shown in Figure 3A (left). Deconvolution of the \( P(r) \) profiles solved using \( q_{\text{max}} \) values up to 0.65 and 1.7 Å\(^{-1}\) estimated contributing peaks at \( r \) values 10.2 and 19.2 Å, and 1.7, 5.6, 10.2, 13.8 and 19.3 Å, respectively. Importantly, these calculations showed that our experimental results and \( P(r) \) estimation agree well with predictions made by Köfinger and Hummer (Köfinger & Hummer, 2013). We feel that these lower order peaks may arise from secondary structures or non-uniform distribution of vectors inside volume of the protein molecules.

**Shape restoration using dummy residues in uniform density mode**

Using the SWAXS profiles and estimated \( P(r) \) curves for different \( q \) range of data, 20 models were calculated using dummy residues, aligned and averaged. The average of the models and their variation are presented in Figure 4. The final resultant \( I(q) \) profile of each individual model is overlaid on the experimental data in the right column of Figure 4. \( \chi^2 \) of SAXS profiles of computed models to experimental data was about 1.2, 0.9, 0.7 and 0.9 while using the upper limit of \( q \) in calculations to be 0.28, 0.51, 1.01 and 1.72 Å\(^{-1}\). Between the models solved for each case, the normalized spatial disposition (NSD) or numerical estimation of similarity in shape profile was calculated to be 0.56, 0.93, 1.07 and 1.47 with increase in used \( q_{\text{max}} \) value which indicated best agreement between models when using \( q_{\text{max}} \) of 1.01 Å\(^{-1}\). Additionally, the ensemble resolution was in range of 18–22 Å, which was about half the value of the \( R_{\text{M}} \) value computed for lysozyme. It is important to point out here that upon using \( q_{\text{max}} \) value up to 1.72 Å\(^{-1}\), the models started to show ‘holes’ in the average models (please see red arrow in Figure 4). The holes or gaps also appeared in individual models solved with \( q_{\text{max}} \) up to 1.01 Å\(^{-1}\), but during averaging, the holes or gaps in the final model are not distinctly visible. This suggested that the use of multi-peak \( P(r) \) profile solves models which are sub-divided as interconnected smaller domains or possibly traces rigid parts in the structure. As seen from crystal diffraction or NMR data based models of lysozyme, there is a cavity in the structure of protein which is also its active site. This protein is actually shaped like an asymmetric ‘kidney bean’ which in low resolution can be interpreted as two domain system. Since the protein is held together by four disulphide linkages, which accounts for its unusually high stability, there lies a possibility that these intermolecular tethering can be sensed in the SWAXS data and it leads to non-smooth distribution of frequency of vectors. This non-uniform behavior possibly gets reflected as holes or gaps when computing models with such \( P(r) \) profiles. Such holes may never occur in structure of protein molecules, but possibly represent how inner flexibility varies across different portions within the overall shape of protein. In future, we plan to map different vectors inside protein structures via MD simulations and compare with SWAXS data presented here and interpretations (please see conclusions also).

To assess comparability of the SWAXS data based model with atomic-definition models solved by crystal diffraction or NMR data, we opted for the ensemble solved by NMR data of lysozyme. Here, PDB submission ID 1E8L was considered which has 50 conformers of lysozyme, all agreeing to the distance constraints and dipolar couplings in multinuclear NMR data (Schwalbe et al., 2001). This work reported that the 50 low-energy conformers in this ensemble showed a backbone RMSD of only 0.5 Å to their mean structure, and an RMSD of 1.5 Å to crystal structure of lysozyme. From SWAXS data perspective, we collect scattering in time and rotation averaged mode from all atoms in molecule except hydrogen. Additionally, minor contribution might also occur from associated hydration layer as well. Thus, instead of comparing our SWAXS data based models with backbone of the structure solved by crystal diffraction data, we opted to superimpose the SWAXS data based models with conformers solved using NMR data (all 50 of them) (Figure 4). The calculated \( \chi^2 \) value between \( I(q) \) profiles of the experimental SWAXS data vs. the theoretical SWAXS profile from PDB 1E8L was found to be about 4.27 to 5.7 indicating significant variance between scattering data and NMR based models. Automated alignment of models indicated that the averaged dummy residue model mainly represents the central core of the protein shape, and more mobile side-chains emerge out of the core, and are within the volume of the variation between the 20 dummy residue models. The average dummy residue model solved using \( P(r) \) profile and SAXS data up to 0.28 Å\(^{-1}\) was single lobed comparing well with the NMR data based models. The kidney bean shape of lysozyme could not be appreciated in this SWAXS data based dummy residue models. Increasing the \( q_{\text{max}} \) value to 0.51 Å\(^{-1}\) changed the solved dummy residue shape from single globe to kidney bean type shape as expected for lysozyme. Further increase in \( q_{\text{max}} \) value used did not aid in shape profiles of the dummy residue averaged model implying additional resolution does not aid in deriving information from uniform density models.

**Re-arranging coordinates of crystal structure using SWAXS data**

Since usage of information from distance distribution functions solved using WAXS data and dummy residue modeling protocols did not provide clearer differences/advantage,
some queries were raised in our mind: whether approximations applied during $P(r)$ estimation from SWAXS profiles, and/or assumptions considered during dummy residue modeling and averaging lead to some limitations in visualizing shape profile seen from crystallography or NMR data? Earlier, when dummy residue models have been limited in providing structural information about protein and some structure was available for the system, we applied normal mode analysis coupled with experimental SAXS data to deduce new structures [recent examples are (Ansari et al., 2020; Sagar et al., 2017)]. To explore if considering atomic resolution structure with primary to tertiary structure details specific to lysozyme
will provide new information, we used SWAXS datasets increasing in q_{max} value and a crystal structure of lysozyme with SREFLEX program. This program computes low frequency vibrations accessible to the structure and searches for altered structures whose theoretical SAXS profile better agree with experimental scattering data. We considered PDB ID 4D9Z, a crystal structure of lysozyme in monomeric state solved by us at 318 K or 45°C (Sharma et al., 2016).

### Figure 4
Uniform density shape models restored using SWAXS data with increasing q_{max} value committed into modeling (mentioned above each set of results). Left column shows rotated views of the average and variation in the 20 uniform density models. All models were superimposed on the possible structures of lysozyme solved using NMR data (PDB ID 1E8L; orange lines). Red arrow in model solved using q_{max} of 1.72 Å⁻¹ indicates the cavity in the averaged model of dummy residues. 2nd column shows superimposition of average, variation and NMR structures. 3rd column shows only average and NMR structures, and last column shows only the NMR structures of the same superimposition. For all sets, the models have been rotated in the axis of the paper. Variation between the computed SAXS profiles of the final averaged model to NMR structures in PDB ID 1E8L are mentioned as \( \chi^2 \) values. Right column of tabular presentation shows the final SAXS profiles of the 20 models solved using dummy residues (black lines) over the experimental data (circles).

### Table

| q_{max} used (corresponding r) | Average of 20 Uniform Density Models | Fit and Statistics |
|--------------------------------|--------------------------------------|-------------------|
| 0.28 Å⁻¹ (22.4 Å) | ![Image of 20 uniform density models](image1) | \( \chi^2 \) of models to data 1.223±0.34 |
| 0.51 Å⁻¹ (12.3 Å) | ![Image of 20 uniform density models](image2) | Ensemble Resolution 18 ± 2 Å NSD 0.560 ± 0.012 |

\( \chi^2 \) (models to PDB 1E8L): 4.276

\( \chi^2 \) (models to PDB 1E8L): 5.771
structure was solved at resolution of 1.71 Å in P212121 space group with \( a, b \) and \( c \) values = 30.0, 55.9, and 72.5 Å, respectively with \( \alpha, \beta, \) and \( \gamma = 90^\circ \). Before appreciating if normal mode analysis can provide identifiable changes in structure, we first compared crystal structure in PDB ID 4D9Z with three more crystal structures, i.e. PDB IDs 6LYZ, 3RT5, and 3SP3 and average of NMR data based conformers in PDB ID 1E8L. 6LYZ was solved as monomeric lysozyme from diffraction data at resolution of 2.0 Å; space group of P43212; \( a, b, c \) values = 79.1, 79.1 and 37.9 Å; \( \alpha, \beta, \) and \( \gamma = 90^\circ \). Taking a different case, lysozyme in 30% propanol, we considered crystal diffraction refinement based structure of monomeric lysozyme (PDB ID 3RT5) solved at resolution of 1.75 Å; space group of P4_212_1; \( a, b, c \) values = 77.9, 77.9 and 36.9 Å; \( \alpha, \beta, \) and \( \gamma = 90^\circ \).

### Table 1

| Structure | Resolution | Space Group | \( a, b, c \) (Å) |
|-----------|------------|-------------|-------------------|
| 4D9Z      | 1.71 Å     | P212121     | 30.0, 55.9, 72.5  |
| 6LYZ      | 2.0 Å      | P43212      | 79.1, 79.1, 37.9  |
| 3RT5      | 1.75 Å     | P4_212_1    | 77.9, 77.9, 36.9  |

### Figure 4. Continued.

\( \chi^2 \) of models to data 0.723±0.19
Ensemble Resolution 23 ± 2 Å
NSD 1.075 ± 0.019

\( \chi^2 \) of models to data 0.934±0.23
Ensemble Resolution 22 ± 2 Å
NSD 1.471 ± 0.028
and $\gamma = 90^\circ$. Third crystal structure was of monomeric lysozyme in 20% sucrose (PDB ID 3SP3) solved at resolution of 1.8 Å; space group of $P4_12_12_1$; $a$, $b$, $c$ values = 76.8, 76.8 and 37.6 Å; $\alpha$, $\beta$, and $\gamma = 90^\circ$. One can see that though the crystals were set-up under very different conditions, the reference being used here for SREFLEX program, i.e. PDB ID 4D9Z was different from other three crystal structures, while latter three appeared to be similar in crystal packing parameters. This could be seen in similar computed RMSD values amongst $C^\alpha$ of the PDB ID 4D9Z vs. all three crystal structures. The RMSD values were 0.42 to 0.48 over 115 to 123 $C^\alpha$ atoms between the pairs of structures (Figure 5). At same time, RMSD was only 0.288 (over 118 $C^\alpha$ atoms) and 0.289 (over 120 $C^\alpha$ atoms) between pairs of structures 3RT5:6LYZ and 3SP3:6LYZ, respectively indicating higher level of similarity in these structures. The fourth structure for comparison was the mean structure from 50 conformers solved for lysozyme protein using NMR data (PDB ID: 1E8L). The RMSD of 4D9Z structure to the mean structure was 1.283 (over 117 $C^\alpha$ atoms) suggesting significant deviation is there in the information deduced from crystalline to solution state. (Other three crystal structures showed an RMSD value of 1.45, in correlation with values reported by the authors who solved the NMR data based models (Schwalbe et al., 2001)). The differences in the coordinate positions of respective $C^\alpha$ atoms are shown as sticks protruding from the reference structure, i.e. PDB 4D9Z and are spread across the whole structure (Figure 5).

Next, we used SWAXS profiles with increasing $q_{\text{max}}$ data and structure of lysozyme (PDB ID 4D9Z) to perform normal...
mode calculations and search for structure which fits SWAXS data better than starting PDB structure. Final fits of SWAXS profiles and computed $\chi^2$ value of the best model deduced with no breaks in structure are shown in second row of Figure 6. Though initial $\chi^2$ values between experimental SWAXS data and theoretical profiles from crystal structure
4D9Z was about 2 to 2.6, post normal mode analysis based search reduced the $\chi^2$ value initially, but later increased with inclusion of higher $q$ data. $\chi^2$ values increased steadily from 0.8 to 3.3 with increase in used $q_{\text{max}}$ value from 0.28 to 1.72 Å$^{-1}$ which was clear from the observed deviation between the experimental vs. calculated scattering profiles.

Interestingly, RMSD values between 4D9Z vs. best solution decreased from 3.0 to 1.533 over 100 to 104 C$^\alpha$ atoms. As shown in Figure 5, the extent and location of C$^\alpha$ atom which is differentially placed in space between the starting structure and best solution are shown as sticks in the third row of Figure 6. The last row shows the RMSD between final SWAXS data referenced solution to the mean structure of 50 conformers in PDB ID 1E8L as black sticks. Interestingly, the computed RMSD value decreased from 3.6 to 1.6 over 102–108 C$^\alpha$ atoms with increase in $q_{\text{max}}$ value to 1.72 Å$^{-1}$. Another point of observation was that when using $q_{\text{max}}$ till 0.28 Å$^{-1}$, most of the variations were around the center of mass of the lysozyme structure. On the other hand, using higher $q_{\text{max}}$ data points into search process yielded solutions where the changes were solved pan structure. Yet the summation of the variations in RMSD values was lower upon using higher $q_{\text{max}}$ value during search process suggesting better convergence between the structural models.

**Conclusion**

This work was performed to evaluate if usage of high $q$ data points in analyzing structural properties of the protein in solution improves final outcome? We used lysozyme SWAXS data as test case, since its structure has been deciphered under varying conditions both in crystalline and solution state. A few experimental and theoretical S/WAXS studies have also been published. By using matched buffer, we could obtain good quality SAXS and WAXS data from protein and buffer which were merged reliably. Additionally, we used a buffer composition which had pH 3.8 and 150 mM NaCl to keep lysozyme devoid of any aggregation or random association. The low $q$ data profile in our SWAXS datasets showed no aggregation or interparticulate effect over the concentration range studied. In WAXS range, we observed three Bragg peaks around $q$ values close to 0.325, 0.65 and 1.4 Å$^{-1}$ which appeared like overtones arising from real space vectors of about 19.3, 9.7 and 4.5 Å in dimensions. Importantly, these peak positions and relative heights compared well with SWAXS profiles published before by experiment and theoretical methods (Kofinger & Hummer, 2013; Makowski et al., 2008; Virtanen et al., 2011). Also, $P(r)$ curves computed using higher $q$ data from averaged SWAXS dataset showed peaks amongst lower $r$ values. The peak positions matched very well with those published before using SWAXS profiles computed from atomic resolution models (Kofinger & Hummer, 2013). Agreement between our results assures us that $P(r)$ curves computed using higher $q$ data and GNOM program can also be reliable as the newer PDDF estimation methods being developed. Since $R_{\text{max}}$ of the protein is about 42 Å, it is clear that these peaks at lower $r$ values arise due to some kind of organization within the shape of the molecule itself. Previous work interpreted these peaks arise from secondary structures in protein structures (Kofinger & Hummer, 2013). Using information from higher $q$ data, we attempted dummy residue modeling which brought out hollowness in individual models and holes in averaged model. To the best of our knowledge, such calculations using high $q$ data and dummy residue models have not been reported earlier, and it will be too premature to write them off as artifacts. More calculations have to be done with other proteins, and complement results with MD simulations or other biochemical experiments. ‘Holes’ in our dummy residue models further imply that the internal density of the lysozyme molecules may not be uniform and some local structures are ‘tightened’ or held apart. As reported earlier, this may be due to four disulphide bonds and/or secondary structures. New methods are being developed to use MD simulations in conjunction with S/WAXS data to filter out structures which better represent protein structures or their conformational properties in solution (Cordeiro et al., 2017). Particularly, use of protocols like HyPred and WAXSiS offer new ways to work with atomic resolution structures to compute hydration layers around proteins and compare with experimental SWAXS data (Knight & Hub, 2015; Virtanen et al., 2011). Additionally, new methods need to be developed to compute interatomic tensors or vectors within protein structure to correlate with observed peaks or holes in $P(r)$ curves or shape restoration when considering WAXS data. (Please note that datasets shown here can be downloaded from SASBDB database and/or additional files can be requested by email by those interested in mapping dynamics of lysozyme molecule or refining new methods of data analysis). Overall, our work showed that good quality SWAXS data can be collected from dilute solutions, and higher $q$ data aids in refining structures from other methods. Even normal mode analysis can provide models about the original structure by moving coordinates till their SAXS profiles match with experimental data.

**Acknowledgements**

This work was mainly supported by CSIR 12 Five Year Plan Project UNSEEN from CSIR INDIA to Ashish. Samples were prepared as routine measure to characterize beam intensity for experiments at different beamlines or for in-house experiments. Research scholars at that time, Yogendra S. Rathore and Ashish K. Solanki are acknowledged for performing some parts of sample preparation. Consistent support of faculty and staff of CSIR-IMTECH in pursuing research activities is acknowledged. Use of the National Synchrotron Light Source, Brookhaven National Laboratory, was supported by the United States Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract DE-AC02-98CH10886. This is IMTECH communication number 026/2021.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**References**

Ansari, M. Y., Batra, S. D., Ojha, H., Dhiman, K., Ganguly, A., Tyagi, J. S., & Mande, S. C. (2020). A novel function of Mycobacterium tuberculosis...
concentrated lysozyme solutions. Proceedings of the National Academy of Sciences of the United States of America, 105(13), 5075–5080. https://doi.org/10.1073/pnas.0711928105

Singh, P., Khurana, H., Yadav, S. P., Dhiman, K., Ashish, Singh, R., & Sharma, D. (2020). Biochemical characterization of ClpB protein from Mycobacterium tuberculosis and identification of its small-molecule inhibitors. International Journal of Biological Macromolecules, 165, 375–387. https://doi.org/10.1016/j.ijbiomac.2020.09.131

Solanki, A. K., Rathore, Y. S., Badmalia, M. D., Dhole, R. R., Nath, S. K., & Nihalani, D., & Ashish. (2014). Global shape and ligand binding efficiency of the HIV-1-neutralizing Antibodies differ from those of antibodies that cannot neutralize HIV-1. Journal of Biological Chemistry, 289, 34780–34800. https://doi.org/10.1074/jbc.M114.563486

Stradner, A., Cardinaux, F., & Schurtenberger, P. (2006). A small-angle scattering study on equilibrium clusters in lysozyme solutions. The Journal of Physical Chemistry. B, 110(42), 21222–21231. https://doi.org/10.1021/jp0639804

Stradner, A., Sedgwick, H., Cardinaux, F., Poon, W. C., Egelhaaf, S. U., & Schurtenberger, P. (2004). Equilibrium cluster formation in concentrated protein solutions and colloids. Nature, 432(7016), 492–495. https://doi.org/10.1038/nature03109

Svergun, D. I. (1992). Determination of the regularization parameter in indirect-transform methods using perceptual criteria. Journal of Applied Crystallography, 25(4), 495–503. https://doi.org/10.1107/S0021889892001663

Svergun, D. I. (1999). Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. Biophysical Journal, 76(6), 2879–2886. https://doi.org/10.1016/S0006-3495(99)77443-6

Svergun, D., Barberato, C., & Koch, M. H. J. (1995). CRYSOL – A program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. Journal of Applied Crystallography, 28(6), 768–773. https://doi.org/10.1107/S0021889995007047

Virtanen, J. J., Makowski, L., Sosnick, T. R., & Freed, K. F. (2011). Modeling the hydration layer around proteins: Applications to small- and wide-angle X-ray scattering. Biophysical Journal, 101(8), 2061–2069. https://doi.org/10.1016/j.bpj.2011.09.021

Volkov, V. V., & Svergun, D. I. (2003). Uniqueness of ab initio shape determination in small-angle scattering. Journal of Applied Crystallography, 36(3), 860–864. https://doi.org/10.1107/S0021889803000268

Xu, H., Lebrette, H., Yang, T., Srinivas, V., Hovmöller, S., Hogbom, M., & Zou, X. (2018). A rare lysozyme crystal form solved using highly redundant multiple electron diffraction datasets from micron-sized crystals. Structure (London, England: 1993), 26(4), 667–675.e663. https://doi.org/10.1016/j.str.2018.02.015