



OPEN

# An automated data processing and analysis pipeline for transmembrane proteins in detergent solutions

Dmitry S. Molodenskiy, Haydyn D. T. Mertens & Dmitri I. Svergun<sup>✉</sup>

The application of small angle X-ray scattering (SAXS) to the structural characterization of transmembrane proteins (MPs) in detergent solutions has become a routine procedure at synchrotron BioSAXS beamlines around the world. SAXS provides overall parameters and low resolution shapes of solubilized MPs, but is also meaningfully employed in hybrid modeling procedures that combine scattering data with information provided by high-resolution techniques (*eg.* macromolecular crystallography, nuclear magnetic resonance and cryo-electron microscopy). Structural modeling of MPs from SAXS data is non-trivial, and the necessary computational procedures require further formalization and facilitation. We propose an automated pipeline integrated with the laboratory-information management system ISPyB, aimed at preliminary SAXS analysis and the first-step reconstruction of MPs in detergent solutions, in order to streamline high-throughput studies, especially at synchrotron beamlines. The pipeline queries an ISPyB database for available *a priori* information via dedicated services, estimates model-free SAXS parameters and generates preliminary models utilizing either *ab initio*, high-resolution-based, or mixed/hybrid methods. The results of the automated analysis can be inspected online using the standard ISPyB interface and the estimated modeling parameters may be utilized for further in-depth modeling beyond the pipeline. Examples of the pipeline results for the modelling of the tetrameric alpha-helical membrane channel Aquaporin0 and mechanosensitive channel T2, solubilized by n-Dodecyl  $\beta$ -D-maltoside are presented. We demonstrate how increasing the amount of *a priori* information improves model resolution and enables deeper insights into the molecular structure of protein-detergent complexes.

Membrane proteins (MPs) are crucial for the normal functioning of an organism as they are involved in essential processes including signalling, nutrient and ion transport, maintaining biological membrane structure and integrity<sup>1–4</sup>. They are encoded by about 30% of the human genome<sup>5</sup> and are major targets for modern therapeutic drugs. This makes structural studies of MP solutions extremely important both scientifically and medically<sup>6–10</sup>.

The defining structural feature of an integral MP is the presence of a large contiguous apolar surface, facilitating the insertion of the protein into the hydrophobic lipid alkyl chain interior of a biological membrane. In order to characterize MPs *in vitro*, one needs to substitute the natural phospholipid bilayer with an artificial amphiphilic environment that supports the functional state. At present the majority of investigators studying MPs utilize readily available and inexpensive surfactant systems or detergents (small amphiphilic molecules)<sup>6,8</sup>. However, a common problem observed is the destabilization of protein structure following exposure to detergent molecules<sup>11,12</sup>. Fortunately, the development of milder detergents and detergent stabilized bilayer systems (bicelles) has helped to overcome this problem<sup>7,8,13,14</sup>. Importantly, the solution behaviour of many commercially available detergents has been well characterised using a range of techniques, including small angle X-ray scattering (SAXS)<sup>15</sup>. Micellar solutions of such detergents have been shown to adopt characteristic sizes, which can be reliably separated from detergent-protein complexes of interest using chromatographic techniques, *eg.* high performance liquid chromatography (HPLC).

Recently, MPs have become increasingly popular objects for small angle scattering investigations<sup>16–19</sup> thanks to the availability of improved online size-exclusion chromatography (SEC-SAXS) set-ups at dedicated SAXS

European Molecular Biology Laboratory (EMBL) Hamburg Unit, DESY, Notkestrasse 85, 22607, Hamburg, Germany.  
<sup>✉</sup>e-mail: [svergun@embl-hamburg.de](mailto:svergun@embl-hamburg.de)

synchrotron beamlines. Integrated HPLC facilitates simultaneous sample purification and X-ray data acquisition, separating the scattering signal from protein, detergent and protein-detergent complexes (PDCs) that appear in the elution profile. Direct measurement of the mobile phase (*ie.* buffer) and elution peaks corresponding to the macromolecule(s) of interest enable a rapid and robust sample characterization following a well-defined sequence of data reduction procedures<sup>20</sup>. These procedures are amenable to automation and can be further coupled to advanced analysis and modeling approaches for MP systems, such as those detailed below.

Reconstruction of protein shape from a detergent solubilised MP is hampered by the strong scattering signal of bound detergent, typically forming a corona around the lipophilic trans-membrane region. The corona consists of hydrophobic and hydrophilic components, with significantly different scattering/electron densities. Thus the results of standard SAXS modelling procedures are, in such cases, non-realistic and ambiguous. Using complementary data from small-angle neutron scattering (SANS) including contrast matching approaches to isolate scattering from each solution component, the ambiguity of models calculated from detergent solubilised MPs can be reduced<sup>21–24</sup>. It was recently shown that SANS data recorded at several contrast points together with an *a priori* estimate of detergent aggregation number, allow one to reliably reconstruct the shapes of MPs in detergent solutions<sup>24</sup>. In this publication it was suggested that a minimum of two SANS curves, collected under different contrast conditions, are required to obtain a stable solution. This requirement is necessary to isolate the signals from the two principal contributions to the scattering intensity, from the protein itself and from the hydrophilic part of the detergent corona. For this purpose the author developed an *ab initio* algorithm with a four-phase bead model for a protein: (1) protein, (2) detergent heads, (3) detergent tails, and (4) solvent<sup>24</sup>. An additional set of geometrical and symmetrical constraints was also required.

In comparison to shape reconstruction from multi-contrast SANS data, reconstruction of MPs from a single SAXS curve is significantly more challenging as the contributions from different phases are difficult to separate. Nevertheless, in view of the relative ease of access to X-ray facilities and lower sample consumption, compared to that of neutron sources, SAXS-based approaches to MP characterization are increasingly popular. A recent work on the tetrameric water-channel protein Aquaporin0<sup>25</sup> showed that if an atomic structure of the protein is available, it is possible to reconstruct the entire complex by combining online SEC-SAXS with refractometry measurements. In that case, it was essential to use the available *a priori* information about the PDC in order to provide sufficient restraints on the model for a meaningful reconstruction. In the present paper we further formalize this approach and develop an automatic pipeline to build models of PDC utilizing *a priori* information taken from a dedicated database.

Given the rapidly growing interest in the applications of SAXS to study MPs, systematic storage of the relevant information is urgently required. The ISPyB information system<sup>26</sup>, a database dedicated to biological SAXS (BioSAXS) and crystallographic experiments on synchrotron beamlines, offers an appropriate framework. ISPyB is already employed at major European synchrotrons (Petra-III, ESRF, SOLEIL, DIAMOND, MAX IV) to facilitate the data flow beginning with the sample tracking and ending with its structural characterization and visualization of the generated models.

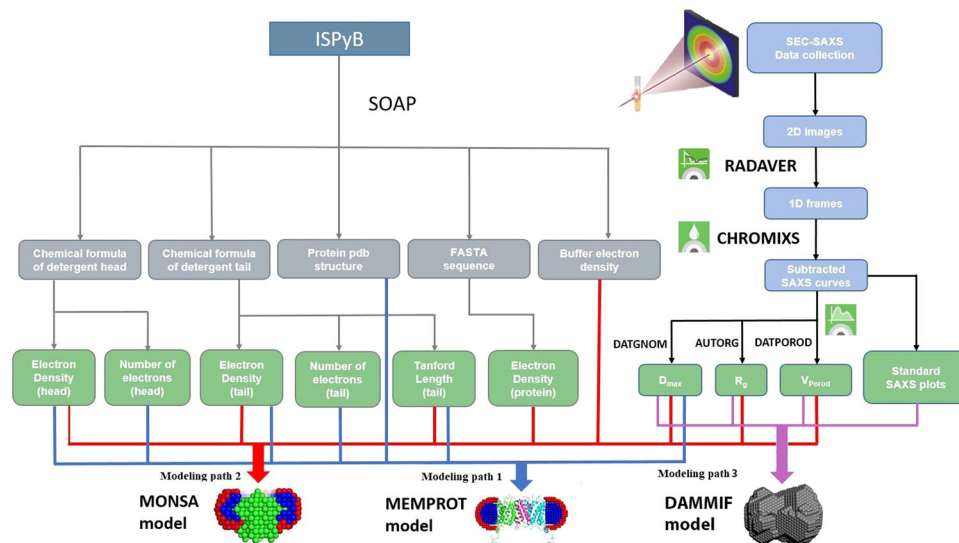
In order to be used in this project, the ISPyB database was extended to store information on the available high-resolution structures of MPs (e.g. homologs or crystallographic models), amino-acid sequences in FASTA format and chemical formulae of the given detergent tails and heads. The latter are used for the calculation of electron densities and approximate sizes of the detergent corona in the modelling. The electron density values can also be provided directly by the user for the cases when the automatic calculations are impossible or the values are not considered to be sufficiently accurate. Depending on the amount of *a priori* information, the pipeline decides whether it is possible to automatically reconstruct the model and, if so, builds it utilizing either the multiphase *ab initio* program MONSA<sup>27</sup> or the hybrid modelling program MEMPROT<sup>28</sup>. The starting parameters for the relevant fitting procedures are evaluated based on the available data stored in ISPyB, with the hybrid approach<sup>25</sup> executing upon detection of a high resolution protein component. In the absence of a high resolution protein structure the *ab initio* reconstruction is executed utilizing information about the amino-acid sequence of the MP, chemical formula of the detergent and additional data on the other solvent/solute components to assess the shape of the PDC.

The automated membrane protein pipeline (AMPP) described herein, is implemented at the P12 beamline of the EMBL at Petra-III storage ring (synchrotron DESY, Hamburg, Germany) as an extension of a standard SEC-SAXS pipeline SASFLOW. It is integrated into the control software BECQUEREL<sup>29</sup> operating the beamline. If the MP-related fields in the ISPyB database are lacking, BECQUEREL runs a standard pipeline mode for soluble proteins, using the program DAMMIF<sup>30</sup> for a single-phase *ab initio* shape determination. The MP-related extension starts by default if the connection with ISPyB is established by the control software, and the membrane-specific fields in the ISPyB database are filled. The system has been optimized through extensive testing on different MPs in detergent solutions, and several examples of its applications are presented below.

## Materials and Methods

**Materials.** The published SAXS data of the tetrameric  $\alpha$ -helical membrane channel, Aquaporin0 solubilized by n-Dodecyl  $\beta$ -D-Maltoside<sup>25,28</sup> was used for the pipeline testing. This sample corresponds to a reliably reconstituted PDC with known geometrical parameters of a detergent corona providing a test case with the modeling program MEMPROT<sup>28</sup>. The model of the Aquaporin0-detergent complex<sup>24,25,28</sup> was used as a reference for the models generated by AMPP. Additional pipeline tests were conducted using the experimental SEC-SAXS data for the mechanosensitive T2 channel solubilized n-Dodecyl  $\beta$ -D-Maltoside<sup>31</sup> downloaded from the SASBDB database ([www.sasbdb.org](http://www.sasbdb.org), ID SASDCY6)<sup>32</sup>.

**Computational methods.** The AMPP is implemented as an extension of SASFLOW<sup>33</sup>, an automated data reduction and analysis pipeline currently running at P12. AMPP utilizes several modules and modeling programs



**Figure 1.** Workflow of the pipeline extension for MPs. The pipeline utilizes *a priori* data from the ISPyB database, calculates overall SAXS parameters and decides, which software to employ for the subsequent analysis.

available in the ATSAS package<sup>34</sup> and, in addition, calls the program MEMPROT<sup>28</sup>. The primary data analysis is executed directly upon completion of a SEC-SAXS data acquisition and proceeds in the following workflow (Fig. 1):

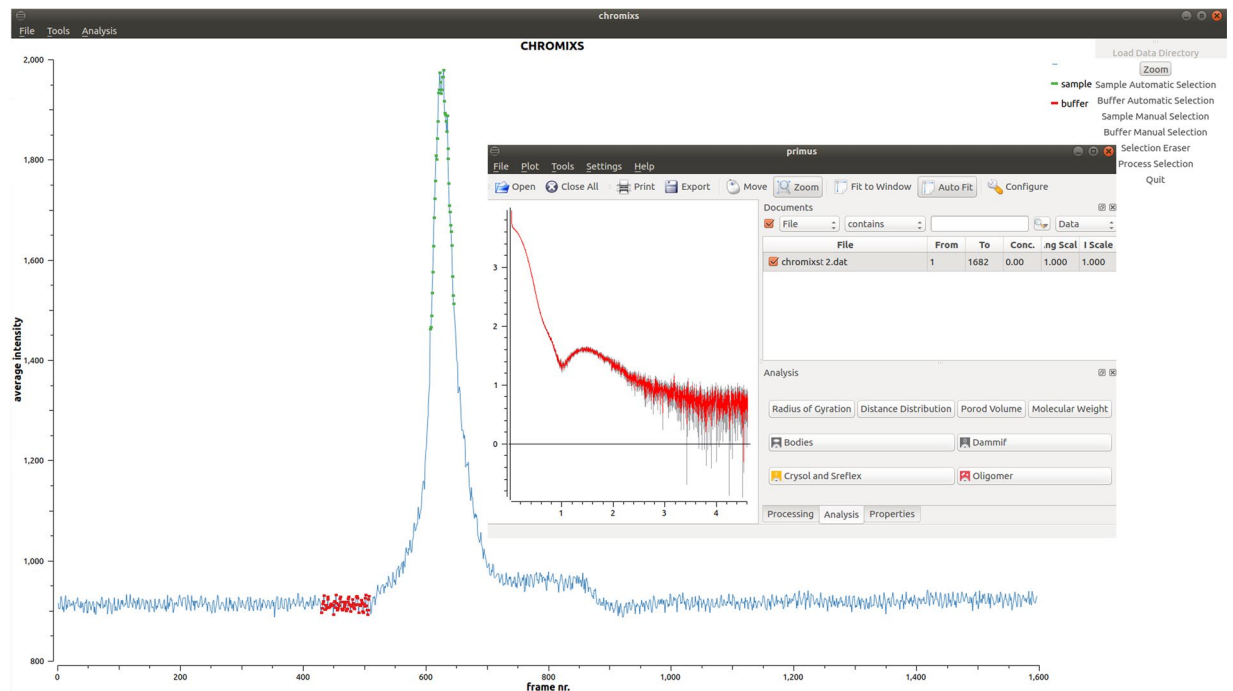
- (1) Radial integration of 2D SAXS images by RADDAVER<sup>33</sup> (typically 1000–3000 frames are collected for a standard SEC-SAXS run set with one second exposure per image/frame).
- (2) Determination of the buffer and sample ranges from the elution profile using CHROMIXS<sup>35</sup>, an important analysis component of the pipeline specifically designed for SEC-SAXS analysis. CHROMIXS automatically identifies both buffer and sample regions in the elution profile, and performs frame averaging and subtraction to produce the SAXS curve of the purified component for subsequent analysis (Fig. 2). It is worth noting that CHROMIXS is capable of identifying multiple elution peaks, providing the users with a set of distinct SAXS curves from individual fractions, if present. Here we consider the simple case, where a single peak corresponding to PDC is identified automatically by CHROMIXS.
- (3) Assessment of the overall quality of the SEC-SAXS data and useability of the sample. Reliable SEC-SAXS data from a PDC suitable for further analysis should fulfill several requirements: i) the main peak in the CHROMIXS elution profile should correspond to a monodisperse fraction only, yielding similar scattering profiles across the entire sample region of the chromatogram; ii) the final buffer subtracted 1D SAXS profile from the PDC must satisfy the standard Guinier criteria (eg.  $0.5 < sR_g < 1.3$ )<sup>36,37</sup>. These important points have been integrated into a single figure-of-merit parameter  $K$ :

$$K = 100\% \cdot c_v^{aver} \cdot \langle \chi^2 \rangle \quad (1)$$

Here  $c_v^{aver}$  is the coefficient of variation ( $c_v = \frac{\sigma}{\mu}$ , where  $\mu$  is the mean value and  $\sigma$  is the standard deviation) of the  $R_g$  value derived from the curve generated by CHROMIXS (averaged and buffer subtracted) employing Guinier approximation;  $\langle \chi^2 \rangle$  denotes the mean square weighted deviation of the frames across the chromatographic peak with respect to the peak frame with maximum integral intensity:

$$\langle \chi^2 \rangle = \frac{1}{m} \sum_{l=1}^m \frac{1}{n-1} \sum_{k=1}^n \left( \frac{I_p(s_k) - c_l \cdot I_l(s_k)}{\sqrt{\sigma_p^2(s_k) + c_l^2 \cdot \sigma_l^2(s_k)}} \right)^2 \quad (2)$$

where  $n$  is the number of points in the frame and  $m$  is the number of frames in the peak.  $I_p(s)$  and  $\sigma_p(s)$  denote the intensity and standard deviation of the frame with maximal integral intensity, and  $I_l(s)$  and  $\sigma_l(s)$  are intensities and standard deviations of all the other frames across the candidate monodisperse region;  $c_l$  is the scaling coefficient;  $s = \frac{4\pi \sin \theta}{\lambda}$  (where  $\theta$  is the scattering angle and  $\lambda$  is the wavelength) is the scattering vector. For the statistically identical scattering profiles, the  $\langle \chi^2 \rangle$  value tends to unity, whereas the  $c_v^{aver}$  value for the good-quality averaged SAXS profiles tends to zero. The  $K$  value integrates the two contributions into a single criterion to characterize the overall quality of a SEC-SAXS dataset. In practice,  $K$  values  $\leq 10\%$  indicate reliable data quality;  $K \geq 30\%$  indicates that the data is of low quality and should be treated with caution. AMPP generates a corresponding notification and/or warning message based on this analysis to aid the user in data interpretation. In the case of the elution of multiple peaks, for example those corresponding to the PDC and also that of empty detergent micelles, a  $K$  value is determined for each SAXS profile generated.



**Figure 2.** An elution profile of T2 membrane protein, built by CHROMIXS from 1600 consequent SEC-SAXS data sets (average intensity versus frame number). The frames corresponding to the buffer and sample are automatically identified (marked by red and green, respectively), averaged and subtracted one from the other. The resulting curve is displayed in PRIMUS interface (inset).

- (4) Calculation of the overall parameters (radius of gyration  $R_g$ , maximum size  $D_{max}$ , excluded volume  $V_{Porod}$ , molecular weight  $MW$ ) and the generation of relevant plots (Guinier plot:  $\ln I(s) \cdot s^2$  vs  $s$ ; and the real-space distance distribution function:  $P(r)$  vs  $r$ ) for all subtracted curves. An example of these plots for the mechanosensitive T2 channel data is shown in the generated summary table (Fig. 3). These parameters are employed for the evaluation of the search volume to conduct the *ab initio* reconstruction using MONSA or for the hybrid MEMPROT fitting procedures.
- (5) Querying and collecting the available *a priori* information from the ISPyB database via dedicated Simple Object Access Protocol (SOAP) web services. Depending on the results, one of the three modeling paths is executed for the subsequent modeling (Fig. 1).

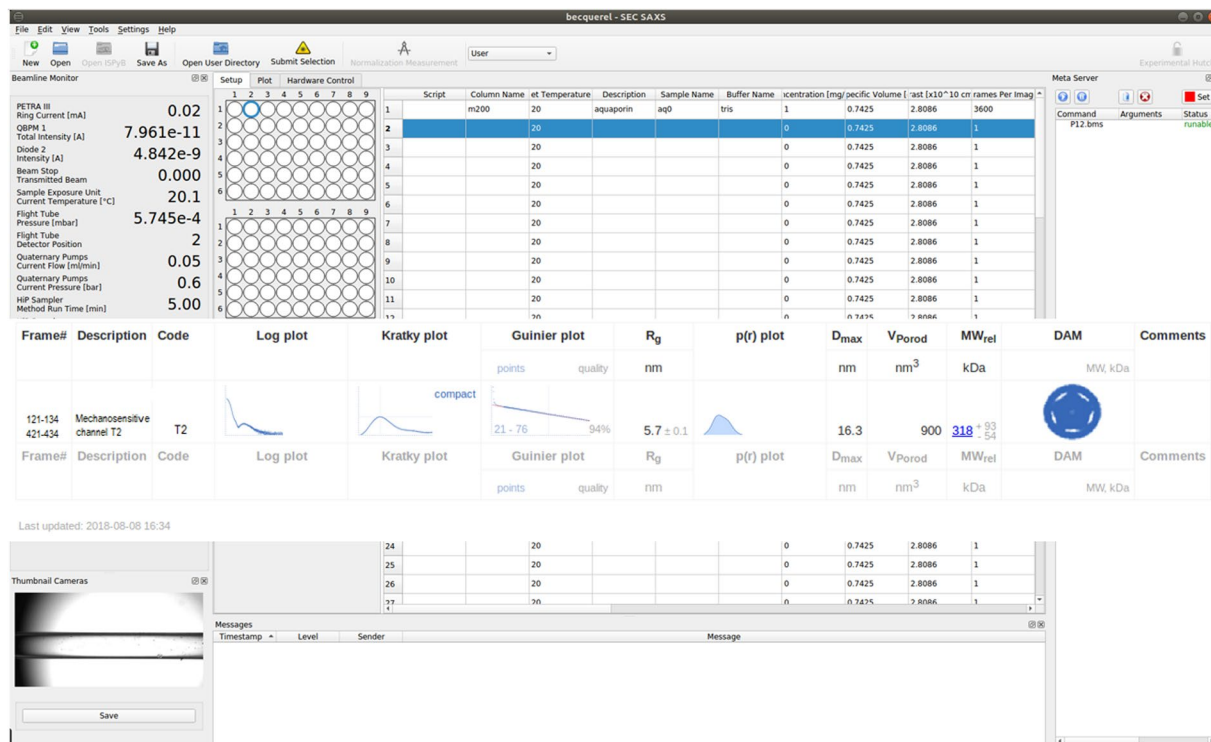
**Modeling Path 1: High resolution model of the protein and chemical formulae of the detergent heads and tails are available.** Initial estimates of the length of the detergent tail and head components are done using the Tanford formula for the detergent tail length:  $l_c = 1.5 + 1.265n_c$ , where  $n_c$  is the total number of carbon atoms. The starting parameters for the evaluation of the corona geometry using MEMPROT<sup>28</sup> are determined as follows:

For a spherical (ellipticity,  $e = 1$ ) detergent torus with the height ( $a$ ) and cross-sectional axes (minor =  $b/e$ , major =  $b \cdot e$ ) equal to the length of a detergent tail (Tanford<sub>tail</sub>), the geometrical parameters are set as:  $a = b = t = \text{Tanford}_{tail}$ <sup>28</sup> (Fig. 4), where  $t$  is the thickness of the hydrophilic detergent head group. Although the hydrophobic tail of a detergent is typically slightly longer than its hydrophilic head, this assumption appears to be a reasonable first approximation for the further refinement.

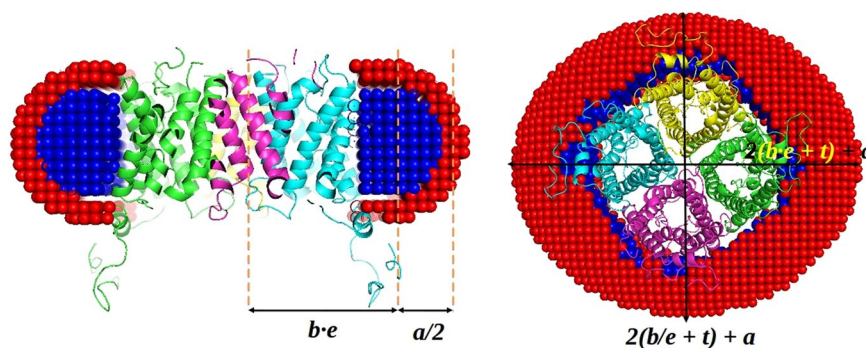
AMPP employs the MEMPROT minimization procedure, refining the starting parameters within a limited range, estimated empirically based on several tests conducted on different MPs. The  $a$ ,  $b$  and  $t$  parameters are searched in the  $\pm 1/3(\text{Tanford}_{tail})$  range, while ellipticity and rotation parameters are kept fixed in order to speed up the computations. The hydrophobic radius of the model is taken as  $a + t = 2 \cdot \text{Tanford}_{tail}$ , which is in a good agreement with the statistical mean value of transmembrane thickness  $D_{Transmemb} \approx 30 \pm 3 \text{ \AA}$ , obtained from the OPM database<sup>38</sup> (<https://opm.phar.umich.edu/types/1>). The maximum value of  $D_{Transmemb}$  is checked during the modelling such that it does not exceed 45 Å.

**Modeling Path 2: Electron densities of a protein and chemical formulae of detergent heads/tails are known, but the high resolution model of the protein is not available.** The electron densities can be either provided directly *via* the ISPyB interface or calculated from the protein primary sequence (in FASTA format) or from the chemical formula using the built-in look-up tables<sup>39,40</sup>. Knowledge of the detergent chemical formula is mandatory as it provides additional information to constrain the geometry of the MP model. When the high resolution models are not available, *ab initio* low resolution bead modeling is performed



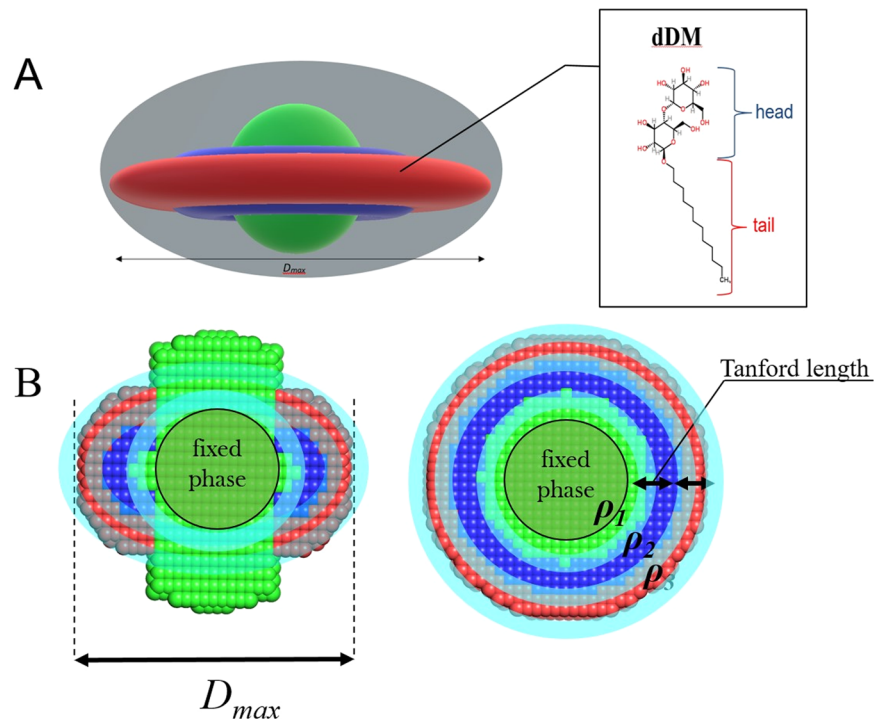


**Figure 3.** An example of the automatic SAXS analysis of T2 in detergent solution at the P12 beamline (PETRA-III). The sample and buffer ranges of frames on SEC-SAXS chromatogram as determined automatically by CHROMIXS are shown in the Frame# field. In this study, a high resolution model was available in the ISPyB database and the modeling was performed by MEMPROT. A thumbnail of the obtained model is displayed in the DAM field.



**Figure 4.** Starting geometrical parameters of a corona for MEMPROT modeling are calculated based on the SAXS curve and *a priori* information from ISPyB. The hydrocarbon tails of the detergent are modelled as an elliptical hollow torus of the height  $a$  and cross-sectional minor and major axes  $b/e$  and  $be$ , where  $e$  is an ellipticity of the torus (adapted from [Perez *et al.*, 2015]). The figure is generated using the PyMol Molecular Graphics System, version 1.7.4. Schrödinger, LLC ([www.pymol.org](http://www.pymol.org)).

by MONSA<sup>27</sup>, allowing for three phases with distinct contrasts corresponding to the protein (phase 1), detergent tails (phase 2) and detergent heads (phase 3). For each phase the contrast is calculated and the search space confined to a cylinder and two coaxial spherical hemi-tori (Fig. 5A), thus avoiding contacts between the hydrophobic regions and solvent in the final model. A spherical region is fixed for the protein phase in the vicinity of the origin of coordinates, while the rest of the cylindrical region is free to become either protein or solvent. Penalties for the model discontinuity and looseness are applied during the simulated annealing procedure as described in the original paper<sup>27</sup>. Three additional boundary regions between the phases (Fig. 5A) have a thickness of a few Angstrom allowing MONSA to select the optimum phase assignment during the fitting. The boundary beads (eg. between phases 1 and 2) are allowed to be assigned to either phase 1 or phase 2 (Fig. 5A). The model is sought within a spherical search volume of diameter  $d = D_{max}$ , where the protein can be surrounded by two-stacked tori



**Figure 5.** (A) Three-phase grid with boundary regions between phases (marked by cyan) and a model that is sought taking into account phase volumes and radii of gyration. (B)  $D_{max}$  is considered to be orthogonal to transmembrane region as depicted at the picture. Radius of sphere, exterior and interior radii of tori are calculated from  $D_{max}$  and the Tanford length. The figure is generated using the PyMol Molecular Graphics System, version 1.7.4. Schrödinger, LLC ([www.pymol.org](http://www.pymol.org)).

(detergent tails and heads) with the lengths evaluated from the Tanford formula (Fig. 5B). The transmembrane diameter is calculated as:

$$D_{memb} = D_{max} - 2 \cdot Length_{(head)} - 2 \cdot Tanford_{(tail)} \quad (3)$$

where  $Length_{(head)}$  is considered to be the same as the  $Tanford_{(tail)}$ , given that for the most common detergent molecules it is typically smaller or equal to that of the tail. The expected values for toroidal volumes and  $R_g$  values of detergent phases are calculated on the fly from geometrical considerations<sup>41</sup> and serve as additional constraints while modelling. The total computed volume and  $R_g$  of the particle is then compared to the experimental  $V_{Porod}$  and  $R_g$ , obtained from step 2. If available, information from complementary methods (e.g. phase volumes or phase  $R_g$ ) can be used to improve the reliability of the final model. Using a spherical 40 beads-per-diameter MONSA grid, together with the above mentioned geometrical restrictions, low resolution models were obtained in a reasonable processing time about 10–15 minutes on a desktop PC (Intel Xeon CPU 3.7 GHz 8 Cores, 12.5 GB RAM). This algorithm is generally applicable to the reconstruction of both transmembrane and membrane-associated proteins. In the case of membrane-associated proteins one can simply shift the fixed “protein” phase to the edge/surface of the detergent disc and extend the protein/solvent phase cylinder outside the entire detergent disc region.

### Modeling Path 3: *A priori* information from ISPyB is not sufficient to generate the required constraints.

In this case neither MONSA nor MEMPROT modeling is employed. The pipeline works in the standard SEC-SAXS mode and the low-resolution model is built using a single-phase approximation (phase corresponding to either PDC or solvent). The pipeline reconstructs low-resolution models from the low- $s$  region (typically  $s_{max} < 1.0 \text{ nm}^{-1}$ ) of the experimental SAXS data *ab initio*, minimizing the contribution from the inhomogeneous internal structure and describing only the overall particle shape. The value of  $s_{max}$  for the initial reconstruction is automatically estimated from the first local minimum of the experimental SAXS curve. A standard *ab initio* modeling procedure is then executed using DAMMIF (as described e.g. in<sup>33,34</sup>): DAMMIF generates 10 independent models and the normalized spatial discrepancy (NSD) is computed to find the model with the lowest variance in NSD with respect to the other members of the ensemble. This model is then used for both alignment/superposition of the ensemble and for the computation of an average model volume; finally, the pipeline employs this averaged model as input for a final refinement in DAMMIN. Therefore, the output of the pipeline in this regime is a single DAMMIN model that fits the data up to an automatically determined  $s_{max}$  value. We would like to stress that this approach has its limitations and the model represents a low-resolution approximation of the molecule’s overall shape and possible high-resolution details on the surface should not be taken as

experimental evidence. With these caveats, such *ab initio* modelling still offers a good low resolution estimate of the overall size and shape of the PDCs of interest in the absence of additional information.

- (6) As soon as the overall SAXS parameters and MP models have been computed, the pipeline builds a final summary file in XML format, gathering together all the results in a compact and human/machine-readable form (Fig. 3). For the models generated, if the quality of the achieved fit is poor ( $\chi^2 > 2.0$ ), the pipeline flags this with a warning. This provides the user with a measure of the reliability of the on-line model(s) generated.

Upon completion the pipeline compresses the calculated 1D SEC-SAXS profiles into a single HDF5 file and reports them back together with the final model into the ISPyB database. The data from the buffer and sample regions, subtracted curves and overall SAXS parameters for each detected peak in the elution profile are displayed in a convenient tabular form (Fig. 3). It is worth noting that the summary table generated is applicable to all SEC-SAXS runs conducted at the P12 beamline and is not restricted to the measurement and analysis of MPs. For example, it is often the case that the sample exists in solution as a mixture of different states (eg. multiple oligomeric states of a protein or complex), which are separated chromatographically. In this case, AMPP will determine the shapes of individual components in the mixture using the modeling path 3. When chromatographic peaks corresponding to unloaded/empty detergent micelles are analyzed, their overall hollow appearance will be reconstructed by the DAMMIF single-phase modeling procedure, whereas the multi-phase distance-restricted MONSA model will fail to fit the data (see Supplementary Fig. S1). The pipeline results can therefore provide a hint to users who mistakenly measure empty/unloaded micelles at the beamline.

The results generated can be inspected online using the ISPyB web-server<sup>26</sup>. The ISPyB database and web services were extended to accommodate the information on chemical formulae and electron densities of detergents, as well as FASTA amino acid sequences and electron densities of the MPs and buffers. The SEC-SAXS elution profile and the final model can be visualized in the existing ISPyB GUI, as the pipeline output files are compatible with the required format.

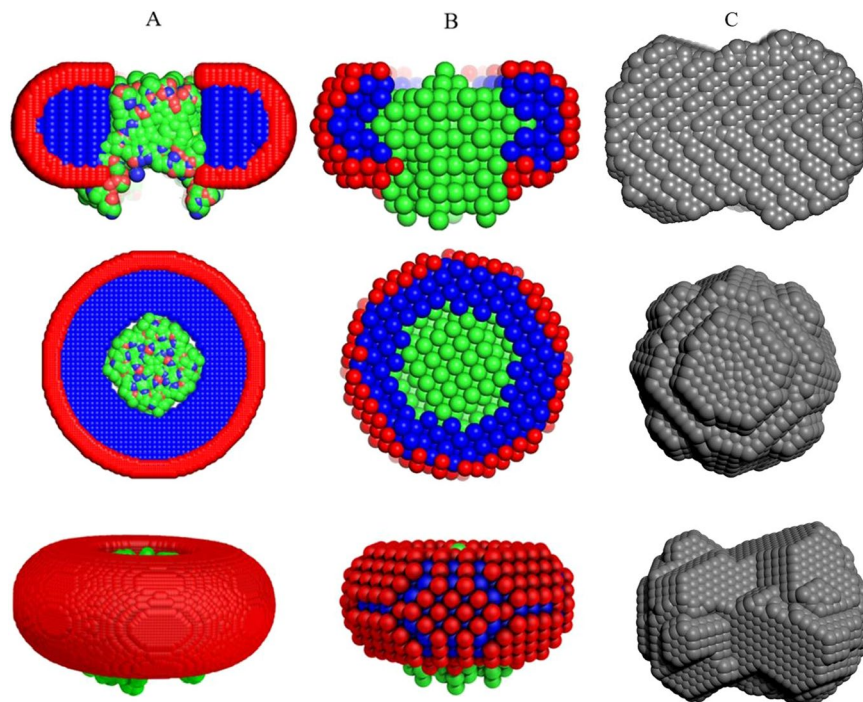
## Results and Discussion

To test the modelling components of the AMPP, we performed an automatic structural characterization of a tetrameric  $\alpha$ -helical membrane channel Aquaporin0 and mechanosensitive channel T2 in DDM solution. The AMPP was utilized in all three pipeline regimes: i) with the high resolution model available ii) with detergent composition and the protein primary structure (FASTA sequence) known iii) plain *ab initio* modelling (no *a priori* information). The estimated *K*-factor (Eq. 1) for the T2 sample was  $8.7\% < 10\%$  indicating a good data quality. The subtracted published SAXS curve was taken for Aquaporin0 and thus no *K*-factor was determined. The models generated from these three modelling trajectories are displayed in Fig. 6 (as the protein was known to be tetrameric, P4 symmetry was utilized in MONSA and DAMMIF).

Expectedly, the automatically generated Aquaporin0 models display more adequate structural details with increasing amount of *a priori* information. The single phase based model generated *ab initio* by DAMMIF (Fig. 6C) provides only the overall shape roughly approximating that of the protein-detergent complex. This reflects the limitation of the *ab initio* shape determination procedure, which must assume constant density inside the particle. The more information-rich multi-phase MONSA and MEMPROT models (Fig. 6B,A) display a good level of similarity, with the hybrid model generated using the latter program being more detailed and arguably having a higher resolution. For the multi-phase MONSA modeling path the starting cylindrical search volume used for protein-solvent phases was considerably larger than the final model and contains a spherical core with the fixed protein phase (green area shown in Fig. 5B). However, after the minimization procedure the volume occupied by the protein phase becomes only slightly larger than the central fixed spherical volume, and the MONSA models do display protrusions resembling the four  $\alpha$ -helical “legs” expected for Aquaporin0. These models are similar to those generated using the hybrid modeling path with MEMPROT, where the protein structure is assumed to be known.

The automatically generated models using the three modeling paths of the pipeline provide good fits to the experimental data at low angles, thus the overall particle shapes coincide and are well described by the data. A numeric evaluation of the proximity of the obtained models was performed using the measure of normalized spatial discrepancy (NSD)<sup>42</sup>, and the NSD comparisons indicate that the models are similar at low resolution (Table 1). The deviations observed at higher angles (Fig. 7) reflect contributions from the detergents including the complex contrast situation, which is not fully captured by the simplified modeling procedures used. In particular, one of the simplifications we implemented for the MEMPROT modeling was to assume a horizontal detergent torus with a circular geometry, and a fixed ellipticity  $e = 1$ . Overall, the models generated by the pipeline do adequately describe the gross appearance of PDCs but caution is of course required regarding the internal structure. Further analysis is recommended in order to obtain more accurate and detailed models, and the initial parameters suggested by the pipeline can serve as a good starting point.

Similar results were obtained from the mechanosensitive channel T2 experimental data (Figs. 7b and 8). There is a high degree of similarity between the three types of models generated (Table 1), and interestingly, MONSA does reconstruct models with a void in the protein phase (green beads in Fig. 8). This void is maintained also when an increased penalty for the bead discontinuity during minimization is introduced. Upon inspection, the high resolution model of a homologous structure, the MscS ion channel of *T. tengcongensis* (PDB: 3T9N) also displays a void in the extracellular domain of the structure. Thus the multiphase *ab initio* reconstruction reliably generates features consistent with those expected for the T2 ion channel. The MONSA model of the T2 channel



**Figure 6.** Examples of automatically reconstructed Aquaporin0 PDC models by the AMPP pipeline utilizing MEMPROT (hybrid modeling with known protein model (PDB:2B6P) and unknown detergent belt) (A), MONSA (multi-phase *ab initio*, the model was built from precalculated searching volume; fourfold symmetry around *z*-axis is imposed) (B) and DAMMIF (C) (single phase *ab initio* also with P4 symmetry applied). The first row shows the side view cross-section of the PDC components that is represented with a section inside the detergent corona; protein beads are green for *ab initio* models and colored according to atom type for the hybrid model, detergent tails and heads are blue and red, respectively. The second and third rows represent the top and the side view of the models. The figure is generated using the PyMol Molecular Graphics System, version 1.7.4. Schrödinger, LLC ([www.pymol.org](http://www.pymol.org)).

Sample	Modeling Software	NSD
Aquaporin 0	MEMPROT-MONSA	1.55
	MONSA-DAMMIF	1.13
	MEMPROT-DAMMIF	2.43
T2	MEMPROT-MONSA	1.79
	MONSA-DAMMIF	1.45
	MEMPROT-DAMMIF	3.24

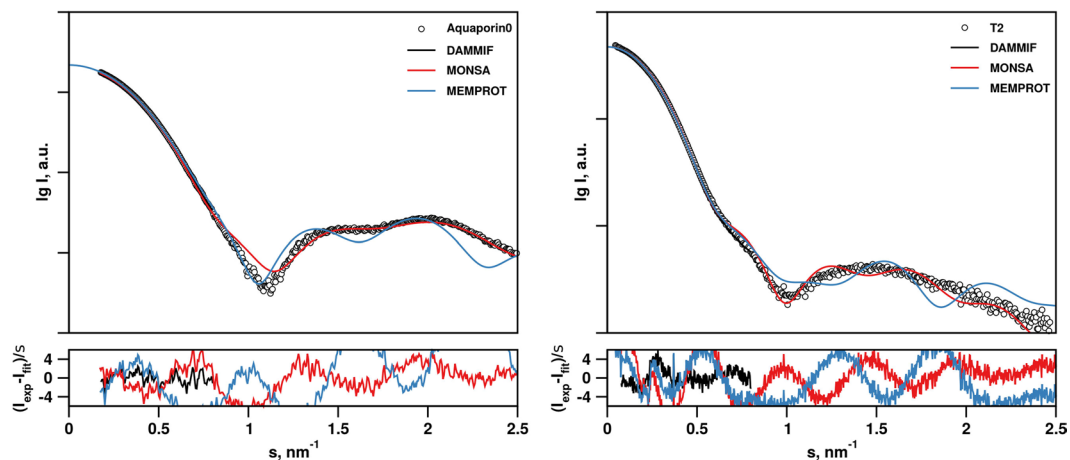
**Table 1.** Standard normalized spatial discrepancy (NSD) between the reconstructed models with differing levels of *a priori* information.

protein fits the data reasonably well up to  $s = 0.25 \text{ \AA}^{-1}$ . The on-the-fly generated MEMPROT model provides a satisfactory fit up to  $s = 0.1 \text{ \AA}^{-1}$ , corresponding to a real space resolution  $2\pi/s \sim 60 \text{ \AA}$ .

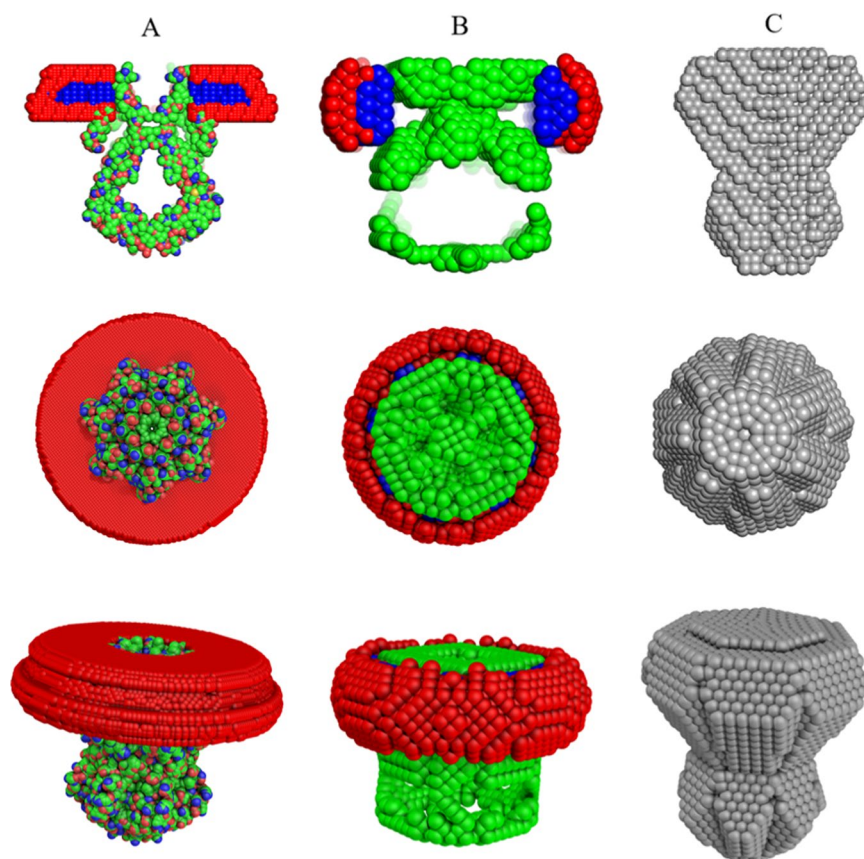
It is also worth noting that atomistic models are often incomplete, missing *eg.* density for flexible loops which could not be resolved at high resolution. In general, to reliably reconstruct a PDC it is recommended to utilize both methods (full *ab initio* with MONSA, and hybrid with MEMPROT) and compare the results. Thus, the *ab initio* modelling is an important component of the automated MP modelling procedure, facilitating a direct comparison with the hybrid modelling and giving an insight into the possible structural organization of the PDC in solution.

In summary, an automated SAXS-based data analysis pipeline for multi-contrast systems has been developed and implemented at the P12 beamline of the EMBL (Petra-III, Hamburg). The pipeline facilitates the determination of the overall parameters and automated construction of tentative low-resolution models of solubilized PDCs. Given that the interpretation of the SAXS data from membrane proteins is by far non-trivial task, these results and models are expected to be further refined interactively. However, as demonstrated in the provided examples, AMPP offers at least a good starting point for the subsequent analysis. The relevant data and workflows are fully integrated into the ISPyB data curation system for SAXS. Future development of the pipeline is planned





**Figure 7.** SAXS data from Aquaporin0 (A) and T2 (B) samples and the corresponding DAMMIF, MONSA and MEMPROT model fits.



**Figure 8.** Automatically reconstructed T2 models by the AMPP pipeline utilizing MEMPROT (hybrid modeling: 3T9N pdb model with found shape of detergent corona) (A), MONSA (multi-phase *ab initio* with P7 symmetry) (B) and DAMMIF (single phase *ab initio*, P7 symmetry) (C). The first row shows the side view cross-section of the PDC components that is represented with a section inside the detergent corona; protein beads are green for *ab initio* models and colored according to atom type for the *hybrid* model, detergent tails and heads are blue and red, respectively. The second and third rows represent the top and the side view of the models. For T2, the  $K$  value was calculated from the reduced SEC-SAXS chromatographic data as follows:  $K = 100\% \cdot C_v^{\text{aver}} \cdot \langle \chi^2 \rangle = 100 \cdot (0.2/5.65) \cdot (2.46) = 8.7 < 10\%$  (Eq. 1).  $K$  was not determined for Aquaporin0 in the present case, as the published data does not include the chromatographic SEC-SAXS data frames. The figure is generated using the PyMol Molecular Graphics System, version 1.7.4. Schrödinger, LLC ([www.pymol.org](http://www.pymol.org)).

to include handling of other more complicated membrane assemblies such as nanodiscs<sup>19,43</sup> and amphipols<sup>44</sup> and also analysis of inhomogeneous samples such as mixed micelles.

### Data availability

AMPP is integrated into SASFLOW, an ATSAS data analysis pipeline, which is available by request (atsas@embl-hamburg.de).

Received: 26 July 2019; Accepted: 20 April 2020;

Published online: 15 May 2020

### References

- Okamoto, T., Schlegel, A., Scherer, P. E. & Lisanti, M. P. Caveolins, a Family of Scaffolding Proteins for Organizing “Preassembled Signaling Complexes” at the Plasma Membrane. *J. Biol. Chem.* **273**, 5419–5422 (1998).
- Koebnik, R., Locher, K. P. & Van Gelder, P. Structure and function of bacterial outer membrane proteins: barrels in a nutshell. *Mol. Microbiol.* **37**, 239–253 (2000).
- Terlau, H. & Kirchhoff, F. Ion Channels/Excitable Membranes. in *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine* 913–916, [https://doi.org/10.1007/3-540-29623-9\\_5640](https://doi.org/10.1007/3-540-29623-9_5640), (Springer Berlin Heidelberg, 2001).
- Shi, Y. & Massagué, J. Mechanisms of TGF- $\beta$  Signaling from Cell Membrane to the Nucleus. *Cell* **113**, 685–700 (2003).
- Wallin, E. & Heijne, G. V. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **7**, 1029–1038 (2008).
- Tanford, C. & Reynolds, J. Characterization of membrane proteins in detergent solutions. *Biochimica Biophysica Acta* **457**, 133–170 (1976).
- Bordier, C. Phase separation of integral membrane proteins in Triton X-114 solution. *J. Biol. Chem.* **256**, 1604–7 (1981).
- Seddon, A. M., Curnow, P. & Booth, P. J. Membrane proteins, lipids and detergents: not just a soap opera. *Biochim. Biophys. Acta - Biomembr.* **1666**, 105–117 (2004).
- von Heijne, G. Membrane-protein topology. *Nat. Rev. Mol. Cell Biol.* **7**, 909–918 (2006).
- Verchère, A., Broutin, I. & Picard, M. Reconstitution of Membrane Proteins in Liposomes. In *Methods in Molecular Biology* 259–282, [https://doi.org/10.1007/978-1-4939-7151-0\\_14](https://doi.org/10.1007/978-1-4939-7151-0_14) (2017).
- Underhill, C. B., Chi-Rosso, G. & Toole, B. P. Effects of detergent solubilization on the hyaluronate-binding protein from membranes of simian virus 40-transformed 3T3 cells. *J. Biol. Chem.* **258**, 8086–8091 (1983).
- Privé, G. G. Detergents for the stabilization and crystallization of membrane proteins. *Methods* **41**, 388–397 (2007).
- Schuck, S., Honsho, M., Ekroos, K., Shevchenko, A. & Simons, K. Resistance of cell membranes to different detergents. *Proc. Natl. Acad. Sci.* **100**, 5795–5800 (2003).
- Dos Santos Morais, R. *et al.* Contrast-Matched Isotropic Bicelles: A Versatile Tool to Specifically Probe the Solution Structure of Peripheral Membrane Proteins Using SANS. *Langmuir* **33**, 6572–6580 (2017).
- Oliver, R. C. *et al.* Dependence of Micelle Size and Shape on Detergent Alkyl Chain Length and Head Group. *PLoS One* **8**, e62488 (2013).
- Kozma, D., Simon, I. & Tusnády, G. E. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res.* **41**, D524–D529 (2012).
- Kynde, S. A. R. *et al.* Small-angle scattering gives direct structural information about a membrane protein inside a lipid environment. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **70**, 371–383 (2014).
- Maric, S. *et al.* Stealth carriers for low-resolution structure determination of membrane proteins in solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **70**, 317–328 (2014).
- Midtgaard, S. R. *et al.* Self-assembling peptides form nanodiscs that stabilize membrane proteins. *Soft Matter* **10**, 738–752 (2014).
- Brennich, M. E., Round, A. R. & Hutin, S. Online Size-exclusion and Ion-exchange Chromatography on a SAXS Beamline. *J. Vis. Exp.* 1–9, <https://doi.org/10.3791/54861> (2017).
- Johns, A. *et al.* Modular Structure of Solubilized Human Apolipoprotein B-100. *J. Biol. Chem.* **281**, 19732–19739 (2006).
- Cardoso, M. B., Smolensky, D., Heller, W. T. & O’Neill, H. Insight into the Structure of Light-Harvesting Complex II and Its Stabilization in Detergent Solution. *J. Phys. Chem. B* **113**, 16377–16383 (2009).
- Nogales, A. *et al.* Three-dimensional Model of Human Platelet Integrin  $\alpha$ IIb $\beta$ 3 in Solution Obtained by Small Angle Neutron Scattering. *J. Biol. Chem.* **285**, 1023–1031 (2010).
- Koutsoubas, A. Low-Resolution Structure of Detergent-Solubilized Membrane Proteins from Small-Angle Scattering Data. *Biophys. J.* **113**, 2373–2382 (2017).
- Berthaud, A., Manzi, J., Pérez, J. & Mangenot, S. Modeling Detergent Organization around Aquaporin-0 Using Small-Angle X-ray Scattering. *J. Am. Chem. Soc.* **134**, 10080–10088 (2012).
- De Maria Antolinos, A. *et al.* ISPyB for BioSAXS, the gateway to user autonomy in solution scattering experiments. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **71**, 76–85 (2015).
- Svergun, D. I. Restoring Low Resolution Structure of Biological Macromolecules from Solution Scattering Using Simulated Annealing. *Biophys. J.* **76**, 2879–2886 (1999).
- Pérez, J. & Koutsoubas, A. Memprot: a program to model the detergent corona around a membrane protein based on SEC-SAXS data. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **71**, 86–93 (2015).
- Hajizadeh, N. R., Franke, D. & Svergun, D. I. Integrated beamline control and data acquisition for small-angle X-ray scattering at the P12 BioSAXS beamline at PETRAIII storage ring DESY. *J. Synchrotron Radiat.* **25**, 906–914 (2018).
- Franke, D. & Svergun, D. I. DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *J. Appl. Crystallogr.* **42**, 342–346 (2009).
- Flayhan, A. *et al.* Saposin Lipid Nanoparticles: A Highly Versatile and Modular Tool for Membrane Protein Research. *Structure* **26**, 345–355.e5 (2018).
- Valentini, E., Kikhney, A. G., Previtali, G., Jeffries, C. M. & Svergun, D. I. SASBDB, a repository for biological small-angle scattering data. *Nucleic Acids Res.* **43**, D357–D363 (2015).
- Franke, D., Kikhney, A. G. & Svergun, D. I. Automated acquisition and analysis of small angle X-ray scattering data. *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.* **689**, 52–59 (2012).
- Franke, D. *et al.* ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Crystallogr.* **50**, 1212–1225 (2017).
- Panjikovich, A. & Svergun, D. I. CHROMIXS: automatic and interactive analysis of chromatography-coupled small-angle X-ray scattering data. *Bioinformatics* **34**, 1944–1946 (2018).
- Guinier, A. & Fournet, G. *Small-angle scattering of X-rays*. New York: Wiley (1955).
- Svergun, D. I., Koch, M. H. J., Timmins, P. A. & May, R. P. *Small angle X-ray and neutron scattering from solutions of biological macromolecules*. Oxford University Press **19**, (2013).

38. Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I. & Lomize, A. L. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.* **40**, D370–D376 (2012).
39. Fraser, R. D. B., MacRae, T. P. & Suzuki, E. An improved method for calculating the contribution of solvent to the X-ray diffraction pattern of biological molecules. *J. Appl. Crystallogr.* **11**, 693–694 (1978).
40. Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. The packing density in proteins: standard radii and volumes 1 Edited by J. M. Thornton. *J. Mol. Biol.* **290**, 253–266 (1999).
41. Kawaguchi, T. Radii of gyration and scattering functions of a torus and its derivatives. *J. Appl. Crystallogr.* **34**, 580–584 (2001).
42. Kozin, M. B. & Svergun, D. I. Automated matching of high- and low-resolution structural models. *J. Appl. Crystallogr.* **34**, 33–41 (2001).
43. Denisov, I. G., Grinkova, Y. V., Lazarides, A. A. & Sligar, S. G. Directed self-assembly of monodisperse phospholipid bilayer Nanodiscs with controlled size. *JACS* **126**(11), 3477–3487 (2004).
44. Tribet, C., Audebert, R. & Popot, J. L. Amphipols: polymers that keep membrane proteins soluble in aqueous solutions. *PNAS* **93**(26), 15047–15050 (1996).

## Acknowledgements

This work was supported by iNEXT, grant number 653706, funded by the Horizon 2020 programme of the European Commission.

## Author contributions

D.S. and H.M. proposed the original idea; D.M. developed and tested the AMPP software; D.M., H.M., and D.S. contributed equally to the writing of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-64933-1>.

**Correspondence** and requests for materials should be addressed to D.I.S.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020