Title

Combining small angle X-ray scattering (SAXS) with protein structure predictions to characterize conformations in solution

ISBN

9780323991810

Authors

Chinnam, Naga Babu

Syed, Aleem

Hura, Greg L

et al.

Publication Date

2023

DOI

10.1016/bs.mie.2022.09.023

Peer reviewed

Published in final edited form as:
Methods Enzymol. 2023 ; 678: 351–376. doi:10.1016/bs.mie.2022.09.023.

Applying Small Angle X-Ray Scattering (SAXS) To Protein Structure Predictions To Characterize Conformations In Solution

Naga Babu Chinnam[1], **Aleem Syed**[1], **Greg Hura**[2,3], **Michal Hammel**[2], **John A. Tainer**[1,4], **Susan E. Tsutakawa**[2,#]

[1]Department of Molecular and Cellular Oncology, The University of Texas M.D. Anderson Cancer Center, Houston, TX, United States

[2]Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley National Laboratory, Berkeley, California, United States

[3]Chemistry and Biochemistry Department, University of California Santa Cruz, Santa Cruz, CA, United States

[4]Department of Cancer Biology, University of Texas MD Anderson Cancer Center, Houston, TX 77030 USA.

Abstract

Accurate protein structure predictions, enabled by recent advances in machine learning algorithms, provide an entry point to probing structural mechanisms and to integrating and querying many types of biochemical and biophysical results. Limitations in such protein structure predictions can be reduced and addressed through comparison to experimental Small Angle X-ray Scattering (SAXS) data that provides protein structural information in solution. SAXS data can not only validate computational predictions, but can improve conformational and assembly prediction to produce atomic models that are consistent with solution data and biologically relevant states. Here, we describe how to obtain protein structure predictions, compare them to experimental SAXS data and improve models to reflect experimental information from SAXS data. Furthermore, we consider the potential for such experimentally-validated protein structure predictions to broadly improve functional annotation in proteins identified in metagenomics and to identify functional clustering on conserved sites despite low sequence homology.

Keywords

CASP-SAXS; protein structure prediction; protein flexibility; hybrid method; FoXS; BILBOMD; metagenomics

1. INTRODUCTION

Deciphering a protein's functional mechanisms is enabled by accurate protein structure(s) and knowledge of the active site and interaction interfaces. In the past, the availability

[#]Corresponding author, setsutakawa@lbl.gov.

of structures was a major bottleneck. Scientists were limited by proteins whose structures could be experimentally determined at high resolution by X-ray crystallography, nuclear magnetic resonance (NMR), and, more recently, cryo-electron microscopy (Cryo-EM). Protein structure predictions, particularly for proteins > 100 residues without a known orthologous structure, were not reliable, even for obtaining the correct topology [1]. After the 2020 Critical Assessment of Protein structure predictions (CASP14), it was clear that machine learning algorithms improved the accuracy of protein structure predictions to the level that predictions could be reliably used [2–8]. Scientists can input the amino acid sequence into a machine learning server and receive a protein structure prediction that is likely to be highly accurate. Among many factors that contribute to the structure prediction accuracy, we highlight several factors in the top protein structure prediction algorithm, AlphaFold2. 1) Use of evolutionary co-variance, which is the concept that residues that vary together in evolution are more likely than not to be close in tertiary structure, first implemented by Deborah Marks and Chris Sanders [9,10]. By default, evolutionary covariance depends on the quality and depth (how many sequences) in the multiple sequence alignments of the target. 2) Use of inter-residue probability distribution distograms that are a spreadsheet of probabilities of any one residue being within distance to other residues in the protein. First applied in CASP13 by the Deepmind team, this opened up the algorithms to better incorporate long-distance relationships, such as N- and C-capping of alpha helices. 3) Application of attention machine learning algorithms that prioritize parts of the input as being more significant, was an impactful change in CASP14 from the traditional convolutional neural networks applied in CASP13. 4) Inclusion of the 3D model within the machine learning algorithm, allowed a confidence assessment in the B-factor column of the coordinates. In our and others' experience, low confidence regions often coincided with loop and other flexible regions and may be a means to efficiently identify such regions, as we show in our example below.

Yet, current protein structure prediction algorithms have limitations. (1) Exposed regions. Although globular regions with substantial contacts and strong evolutionary conservation are likely to be highly accurate in predicting, regions with fewer contacts or flexibility are likely to be mispredicted. (2) Proteins with few family members. The number of homologous sequences matters. Accuracy depends partly on the depth of the multiple sequence alignment and derived evolutionary covariances from sets of residues that mutate simultaneously. Thus, proteins with low sequence homology to other proteins will have a lower probability of obtaining an accurate structure. (3) Protein-protein, protein-DNA, and protein-RNA complexes. The latest machine learning algorithms have not been tested in CASP for the accurate prediction of complexes. Also, given that computational time is limited for many publicly available servers, the calculation for complexes is more complicated than for individual proteins. For complexes with RNA or DNA, machine learning algorithms cannot predict RNA and DNA conformations sculpted by protein interactions. (4) Bias towards deposited structures. Accuracy is a relative term as it depends on the standard. In its assessment of protein structure prediction algorithms, the CASP committee uses high resolution structures determined mainly by X-ray crystallography. However, we and others have shown that proteins in solution can be conformationally distinct from proteins trapped in crystal lattices and that sometimes may not adopt biologically-relevant conformations

[11,12]. As the machine learning algorithms use the PDB as their training database, their predictions are influenced by the presence of co-factors, nucleic acids, or ligands, in addition to the salt and aggregating conditions to form ordered crystallographic protein arrays. With the target being biologically relevant conformations, we and others anecdotally have found predictions to be closer to crystal structures than to the solution structures. (5) Small molecules and ligands. Current machine learning algorithms cannot predict the structural impact of bound ligands, which can be functionally important for allosteric binders and induced fits. (6) Impact of mutations. Machine learning algorithms minimize disagreements between multiple inputs but, by necessity, remove outliers. Thus, current algorithms are challenged to predict the structural consequences of a mutation.

Although there are limitations in protein structure predictions, comparison to experimental data, even if sparse, has the potential to test, validate, and guide protein structure predictions. In particular, Small Angle X-ray Scattering (SAXS) can provide global structural restraints. SAXS data contains information on the distances between all electron pairs. Unlike NMR, these distances are not assigned, and it is currently not sufficient to derive an atomic model. However, all distances are included, and it is possible to calculate the SAXS curves from atomic models – allowing a quantitative comparison between atomic models and experimental data [13–18]. One step further, atomic models can be altered and/or or used as an ensemble of multiple atomic models, to test against the SAXS data and gain information on the conformation(s) and dynamics in solution [19,17,16,15,14]. Here, we focus on how to use atomic models from protein structure algorithms with experimental SAXS data to validate prediction and/or to characterize the structure in solution. A match between the prediction and the SAXS data will support the validity of the protein structure prediction. We will show how to compare atomic models to experimental SAXS data. Similarity but with slight differences between calculated and SAXS data suggest that the solution structure is different, either in conformation, dynamics, or stoichiometry. We will show servers that can alter the predicted model structure and that can test ensembles of these altered atomic models.

The rationale for using SAXS data to validate structure predictions is that it has a major advantage compared to other structural methods such as X-ray crystallography. SAXS data collection is straightforward with minimal sample preparation and is essentially available without cost through synchrotron SAXS beamlines to any academic scientist with purified protein [20–23]. SAXS synchrotron data collection for research to be published is typically free. No labelling is required. The particle size limitation at our SIBYLS beamline in the Advanced Light Source Synchrotron ranges from 8 kD to 600 kD, within the range of most protein samples [24]. Importantly, oligomeric states [25,26], protein flexibility [27–29], and shape [30] can be calculated from SAXS data. In our experience, the stoichiometry is particularly important as it is often not recognized. Our structural genomics study of 50 proteins, mostly from *Pyrococcus furiosus* showed that over half of the proteins studied formed homo-oligomers [20]. It is estimated that over half of the proteins in the PDB are multimerizing [31]. Although we focus primarily on protein models, SAXS data can be collected on protein, DNA, and RNA individually or as complexes and compared to relevant atomic models. The impact of mutations and small molecules can also be probed through SAXS [32,33].

SAXS data can effectively validate protein structure prediction since it contains structural information on the distribution of electron pair distances of proteins in solution [34,35]. Analogous to X-ray crystallography, SAXS data are collected in reciprocal space, with the Intensity (I) of the scattered X-rays plotted as a function of the scattering angle q. The scattering angle is typically calculated as momentum transfer q to incorporate beamline parameters and X-ray wavelength. The units can be in $Å^{-1}$ or nm^{-1}. Like X-ray crystallography, SAXS data can be converted to real space through the Fourier transform X-ray scattering relationship between reciprocal and real space [34,36]. In real space, SAXS data are represented as a histogram of relation proportion P of electron pairs at distance r, also known as P(r). D_{max} is the longest electron pair distance in the P(r). SAXS data can be compared to protein structures in real or reciprocal space and is also called SAXS curves or SAXS profiles. Molecular envelopes can be calculated *ab initio* from experimental SAXS curves [30,37]. Although useful, *ab initio* envelopes are limited by the difficulty to differentiate between closely related models, so we find comparing SAXS data to atomic models directly more quantitative [38].

Here, we focus on comparing atomic models to SAXS data after the data has been collected and analysed. This comparison allows us to validate a protein structure prediction and/or alter the prediction to better fit the solution data. A single SAXS curve for each condition has been generated from available SAXS data. Recent reviews provide detailed protocols for SAXS sample preparation, data collection, and analysis [39,40]. At our SIBYLS beamline at the Advance Light Source synchrotron, SAXS data can be collected in two modes: in a sample cell in high throughput (HT) SAXS (30 μl of a 0.5–2 mg/ml) or via Size-Exclusion Chromatography-coupled (SEC)-Multi-Angle Light Scattering (MALS)-SAXS (50–100 μl 5–20 mg/ml). HT-SAXS provides the best signal-to-noise. To compare SAXS to the atomic model, the sample ideally must be well behaved, have no aggregation, and be stoichiometrically monodisperse. SEC-MALS-SAXS will separate heterogeneity and provide a monodisperse sample for difficult samples but dilute the sample > 4-fold [41,42]. Making it a relatively inexpensive and available structural method, protein sample(s) can be mailed in at the SIBYLS beamline; the SAXS data are collected by experienced staff, and the SAXS data are made available to users.

The basic process for SAXS validation of protein structure predictions is to obtain atomic models from a publicly accessible server, predict the SAXS curves from the atomic models, compare these predicted SAXS curves to experimental SAXS data, and, if necessary, modify the models (e.g. change oligomerization state, change the protein conformation) to improve fit to the SAXS data (see Figure 1A).

2. OBTAINING A PROTEIN STRUCTURE PREDICTION.

2.1 Overview.

Currently, of the many servers in protein structure prediction available, we are most familiar with RosettaFold [8] created by David Baker’s group, and AlphaFold2 [4] by the Deepmind Technologies company and recommend these programs here. However, accurate structure predictions can be obtained from other servers as well. We recommend experimental SAXS validations for any atomic model, regardless of source including crystallographic models [39].

In terms of accuracy of your protein structure prediction, it is our experience and documented in CASP14 that the accuracy of individual domains for most protein predictions and, in particular, confidence for their topology are high [2]. However, caution should be taken for domain-domain orientation, for proteins with low sequence homology and for regions that are predicted to be disordered. For the latter, likely due to their lack of representation in the PDB training database for the machine learning algorithms, there is currently failure for regions predicted to be disordered and these regions are predicted simply as random coil. Although SAXS-based atomic analyses can be informative with these type of proteins [42,43,24,44–47,29], care should be taken in conclusions as these regions adopt multiple conformations. It is also unclear how loop regions involved in domain swaps will be predicted, but they can play critical mechanistic roles as found for p62 in the TFIIH core 7-subunit complex that extends from the XPD ATPase site to its DNA binding groove [48]. Another cautionary note for protein structure predictions is the possible presence of conformational changes, induced by small molecules or intrinsic to the protein domain. SAXS later on in the chapter can assess conformational flexibility and guide selection of alternative conformations.

Structures for homo- and heteromeric complexes can also be predicted using these protein structure prediction machine learning algorithms but have not yet been assessed by CASP [49,50] (See Note 1). Predictions of complexes can be tested against experimental SAXS data similar to single chain models. For nucleic acid complexes, RNA prediction servers will also be helpful for this category of predictions and that SAXS will be of use for validation and tweaking the atomic RNA models to be more consistent with solution data.

2.2 Protein structure prediction Servers

Follow the instructions for each server to obtain the prediction. Be sure that the sequence exactly matches the SAXS sample, including tags. Tags can also be added to already predicted models in the databases in modelling programs such as Modeller within CHIMERA, as previously described [19]. Follow the instructions for each site. These programs will typically output multiple models. All models can be compared to the SAXS data. As mentioned earlier, potential pitfalls are for proteins with little or no sequence homology to other proteins, as sometimes found for viral or bacterial proteins.

2.2.1 https://robetta.bakerlab.org/ —This site hosts the RosettaFold protein structure prediction server. Homo-oligomeric and heteromeric complexes can also be predicted as described in RosettaFold's frequently asked questions (FAQ) [49].

2.2.2 https://alphafold.ebi.ac.uk/ —In collaboration with EMBL-European Bioinformatics Institute (EMBL-EBI), this is a database of AlphaFold2 protein

Note 1.As with any atomic model for complexes, interfaces should be assessed for high sequence conservation and charge/hydrophobic complementarity. The PISA server (https://www.ebi.ac.uk/pdbe/pisa/) provides an automated assessment for the latter [61]. Even in crystal structures, it is not unprecedented for interfaces to be mis-assigned, as we found for the repulsive interface between PCNA and sumo [47]. It should be noted that SAXS structure predictions for 20 model organisms including *Homo sapiens, Escherichia coli, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Mus musculus, Rattus norvegicus, Caenorhabditis elegans, Danio rerio*, and *Drosophila melanogaster*. Alternatively, these models can be obtained on https://www.uniprot.org/ under individual protein entries. Besides the predicted structure in PDB format, the database also provides a per-residue confidence score (pLDDT) for each residue between 0 and 100. The regions below 50 pLDDT may be unstructured in isolation. Consequently, these values can be used to guide conformational sampling, as described in section 5. We note that the Swiss-Model also adds in structure predictions for complexes [51].

2.2.3. ColabFold.—AlphaFold2 can be run on this installation on a Google Colaboratory site. There are limitations in runtime, so large proteins may not be accurately predicted. AlphaFold-Multimer can be run at this site for atomic models of complexes, with the same time limit restrictions [50]. https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

2.3. Pitfalls to look out for.

Atomic models should have secondary and tertiary structure in their predictions for ordered regions. If the prediction server has stopped in the middle because of a time limit given for the prediction (not uncommon occurrence for large proteins or complexes), these predicted-to-be-ordered regions will look garbled. To assess if these regions are predicted to be disordered or ordered, there are a number of highly accurate disorder prediction servers publicly available including PONDR and DisProt [52–55] that predict disordered regions within the sequence. If the protein is homo-oligomerizing, there may be regions that overlap within the oligomer and that will need to be adjusted.

3. PREDICTION OF SAXS CURVE FROM AN ATOMIC MODEL.

3.1 Overview.

To validate structure predictions, we compare experimental SAXS data to the SAXS curve calculated from the structure prediction. Multiple programs can calculate a SAXS curve from atomic models of protein, DNA, and/or RNA [13–17,56]. We recommend FoXS, that uses the Debye formula to explicitly compute all inter-atomic distances including a predicted hydration layer and taking into account the displaced solvent [17] (Figure 2). When input with experimental SAXS data, there are additionally two constants, c1 and c2, that adjust the total excluded volume and the density of the hydration layer, respectively, to better fit the experimental data (see note 2 for potential issues with c1 and c2). The FoXS server is readily available and can be conveniently input with multiple atomic models uploaded as a compressed folder. A recent Methods paper gives a detailed review [19].

Note 2.For the prediction of SAXS from atomic models, it is best to input the protein structure predictions with the experimental data, to improve the estimation of the hydration layer. The FoXS program uses two fitting constants, c1 and c2, to fit the experimental SAXS data. Note that a c2 constant close to 4, indicates overfitting to the experimental SAXS data. Often this occurs when the protein is adopting multiple conformations in solution and is reduced by analysing against an ensemble of multiple structures (via MultiFoXS).

3.2 Server for Prediction of SAXS curve from an atomic model.

The server is located at https://modbase.compbio.ucsf.edu/foxs/

3.3 Inputs for FoXS server.

Use default parameters.

3.3.1 Atomic model(s) input for FoXS.

—The sequence should match the SAXS sample. Tags and missing sequence could be added to the atomic model by creating a homology model based on the protein structure prediction with Modeller implemented in Chimera. Atomic model(s) can be uploaded as a single structure or multiple structures as a zip-compressed file.

3.3.2 Experimental SAXS data input for FoXS.

—Only one SAXS profile can be uploaded. The SAXS profile should have three columns: q, I(q), and error, where q is the momentum transfer. $q=4\pi\sin(\theta/2)/\lambda$ in Å^{-1}, θ is the scattering angle, λ is the X-ray wavelength, and I is scattering intensity as a function of q. Typically, most beamlines provide the SAXS curve calculated with q and considering beamline-specific parameters, including X-ray wavelength and distance from detector. One potential pitfall is that the scattering angle should be in Å^{-1} and not in nm^{-1}. Our SIBYLS beamline provides the scattering data in Å^{-1}.

3.4 Output from the FoXS server.

Graphs showing fit of atomic models to experimental SAXS data. Below the graphs is a table of atomic models, calculated χ^2, c1, and c2 constants, ranked by the goodness of fit χ^2. The predicted SAXS curve for the atomic model can be obtained by clicking on the link on the right.

3.5 Processing of reciprocal space SAXS curves predicted from structure predictions.

The output fit file has columns 1 to 4 for scattering angle q, I_{exp}, $error_{exp}$, and I_{model}. To use the I_{model}, the experimental column (I_{exp}) need to be removed. One can use the Linux command "awk '{print $1,$4}' fox.fit >model.dat" where foxs.fit is the FoXS-generated .fit file, $1 and $4 refer to the column # from left to right, and output file model.dat. Note that some programs output the model Intensity in column 3, so it is advised to check the output fit file to double check. In a basic text editor that does not add any hidden text to the file, remove unnecessary headers and tail text. The new file called model.dat can be compared to experimental data by V_R implemented in the SAXS similarity analysis (see below).

3.6 Generation of real space SAXS curves predicted from structure predictions.

To generate the experimental and model SAXS curves in real space, use GNOM in the ATSAS package [36,57], as previously described [39]. For the model, the experimental data must be removed, such that q and I_{model} are in the first two columns, as shown for the model.dat file in the previous section. Dmax must be input into the GNOM program but may differ for experimental and predicted SAXS data. See Note 3, for issues in experimental data.

4. COMPARISON OF EXPERIMENTAL AND PREDICTED SAXS CURVES.

4.1 Overview.

SAXS curves can be compared in both reciprocal space and real space. Quantitative comparison of experimental and predicted SAXS curves is an ongoing area of development in SAXS analyses, and we expect further improvements in this area.

4.2 Experimental SAXS data.

SAXS data can be collected readily from any synchrotron, as described in other methods reviews [39,40]. Data should be of high quality. Our SIBYLS beamline provides users preliminary assessment on data quality. Experimental data pitfalls include uninterpretable data caused by aggregation, bubbles in the sample cell, poor buffer subtraction (likely culprit when data are cut off at low q), too low protein concentration, contribution of detergent micelles, or inter-particle repulsion. We show examples of these pitfalls at our beamline web page. (https://bl1231.als.lbl.gov/htsaxs/statistics) [23]. In contrast, stoichiometric and conformational flexibility are amenable for the SAXS analysis programs using ensemble methods and can be dealt with using ensembles of atomic models. For example, in a study of SARS-CoV-2 replication machinery, we could detect multiple and distinct oligomerization states of Nsp7, Nsp8, and RNA [44]. Although mixtures do lower the resolution of what can be concluded from a SAXS curve, information can still provide valuable insights – in this example, we determined that the assembly of the viral replication/transcription complex was not linear and was RNA-dependent.

4.3 Comparison metrics (Figure 2).

4.3.1. χ^2 in reciprocal space.—The most common comparison metric and one that is implemented in FoXS is χ^2, a statistical test that quantitates the difference between the experimental and predicted data as a function of q, divided by the experimental noise. The lower the number, the better the model matches the experimental data. (See note 4 for caveats in χ^2 applications in SAXS analysis). It is automatically generated in many of the SAXS modeling programs.

4.3.2. V_R in reciprocal space.—A volatility ratio, originally developed for stock market analyses, uses a ratio to normalize the curves and spreads the weight of the comparison through the whole curve [38]. V_R has been implemented in https://bl1231.als.lbl.gov/saxs-similarity/ and is described in detail in Chapter 14. As many as 100 SAXS curves (experimental or calculated) can be uploaded by drag-and-drop into the input box, and the web server automatically compares the similarity of each curve. As described

Note 3. In the GNOM analysis, aggregation of the SAXS sample may cause an artificial increase in Dmax. SEC-SAXS data should not have any aggregation; sometimes, there's been an accumulation of damaged protein on the sample cell window. Processing of the SEC-SAXS data with RAW can help minimize this type of aggregation [62].

Note 4. As applied in SAXS, the χ^2 comparison metric does not evenly compare the reciprocal space curves from low to high q but is biased towards curves at the low q region. As SAXS is an exponentially decaying function with a 100 to 1000-fold difference in signal over the typical curve, the difference between curves of Intensity as a function of q ($I_{exp} - I_{pred}$) will be more significant at low q compared to high q, where a 1% difference would be 2–3 magnitudes smaller. How the curves are scaled will also bias χ^2. Another critical factor is that experimental noise or uncertainties of the data is in the denominator of χ^2, making noisier data appear to have a lower and better χ^2 fit (higher χ^2 scores) than better quality data with less noise. Thus, different models can be compared to one experimental data, but χ^2 fits to different experimental data to one model cannot be compared.

on the web page, SAXS .dat files (reciprocal space) must be in plain text and simple file names are required. The SAXS .dat files must contain two columns, q (Å^{-1}) and Intensity. A third column containing the experimental errors is optional and can be used for χ^2 calculations on the SAXS-Similarity server. The default q range analysed is 0.015 – 0.2 Å^{-1}, but the q range can be adjusted. A pairwise similarity plot is output with colors representing the Volatility ratio. Clicking on single cells shows the pairwise overlay and relative radius of gyration (Rg). An alternative view as a force plot is useful for visualizing curves grouped based on similarity.

4.3.3. SAXS data can also be compared in real space visually.—Plots are normalized based on the area under the curve. Quantitative metrics have been tested [58], but none have been established. Nonetheless, it is useful to compare reciprocal and real space (see note 5). In the real space curve, shoulders after the prominent peak indicate separated domains (beads on a string with varying distances).

4.4 A cautionary note is the possibility for SAXS to validate a completely inaccurate protein structure prediction.

In CASP13, atomic models were distorted to fit to the SAXS data in reciprocal space, with completely incorrect protein topology and geometry [12]. However, in comparison of real space curves, there was clear differences between experimental SAXS data and the structure prediction. Thus, it is good practice to consider SAXS-validated structure predictions as a testable atomic model, look at plots in both reciprocal and real space and to make sure that geometry makes sense.

5. FITTING OF THE PROTEIN STRUCTURE PREDICTION(S) TO THE EXPERIMENTAL SAXS DATA.

5.1 Overview.

Flexibility is a common component in protein mechanisms, and so many proteins will adopt multiple conformations. Flexibility can be disordered regions, particularly at tails or shifts in domains relative to each other. Based on our experiences, it is also more likely than not that the protein structure prediction is in a different oligomeric state, conformational state, or requires an ensemble of structures consistent with what is occurring in solution. Several programs will generate alternative conformations or complexes and score the output structures on their agreement to input SAXS data. More detail on these programs can be found in a recent methods paper [19]. A limitation is that these servers for these programs only work on proteins; small molecules and nucleic acids cannot be readily input. However, advanced molecular dynamic users can adapt these programs as needed. It should also be noted that one program, AllosMod, can include glycosylation.

[Note 5.] In CASP13, computational scientists used experimental SAXS data to fit protein structure predictions[12]. At that time, prediction scientists had the wrong topology for large proteins with no known structural templates, the most challenging targets. Their programs squashed the incorrect topology to fit the overall shape, and the reciprocal space fits were almost perfect. Yet, a comparison of these models in real space showed significant differences, showing that comparison in both reciprocal and real space is of value.

5.2 Software to adjust protein model conformations or stoichiometry.

5.2.1 BILBOMD (https://bl1231.als.lbl.gov/bilbomd/) is a stand-alone web server that will generate a large population of conformally diverse models [59,15]. It performs all the multistate modeling stages: conformational sampling, SAXS profile calculation, and multistate models enumeration. The conformational sampling is based on the minimal molecular dynamics (MD) simulation using CHARMM. SAXS profile calculation and enumeration of multi-state models use FoXS [17,16] and MultiFoXS [15] programs, respectively. The entire protocol is fully automated and does not require user interaction [19]. More generally, BILBOMD explores conformational space based on molecular dynamics by keeping domains as rigid bodies and allowing flexible regions to move. Bond distances are maintained, but the temperature is set high, and bond angles are allowed to change to generate the most conformationally diverse population efficiently. As required input into BILBOMD, each segment or polypeptide is uploaded separately, plus the experimental SAXS data (see note 6 for potential pitfalls). A Rg range is input, typically 10–20 Å below and above the experimentally defined Rg. The program uses the minimal ensemble algorithm and will output an ensemble of conformations most consistent with the experimental SAXS data.

5.2.2 AllosMod-FoXS is a combination of AllosMod and FoXS servers [60] and generates a population of conformation by altering non-bonded distances in the input pdb and compares the altered atomic model to the experimental SAXS data. It differs from BilboMD, where defined regions are kept rigid. Instead, AllosMod uses molecular dynamics of the entire protein to generate an energy landscape of conformations around the atomic model. Practically, most conformational changes are in regions that have few contacts (e.g. linkers, thin regions) more than in well-folded globular regions, providing a means to generate conformational changes within the domain region itself. AllosMod has different sampling options, which can be tested. The higher temperature levels offer the most exploration of conformational space, but which is best is based on the individual protein structures. The AllosMod server is at the Sali lab website (https://modbase.compbio.ucsf.edu/allosmod-foxs/). The initial output is a list of models, ranked based on fit to the input experimental data and a link to a compressed folder with all models. An option for ensemble analysis by MultiFoXS (see below) is available on this output page. A potential pitfall is that the protein's geometry can become overly distorted; loss of secondary structure elements indicates that this has occurred.

5.2.3 MultiFoXS is built on the minimal ensemble algorithm in BILBOMD [59] and is effective when an ensemble of conformations or different assembly states (*e.g.* a mix of monomer, dimer) fit the scattering data [15]. MultiFoXS uses a minimalist approach and identifies the minimum number of conformations needed to agree to the experimental SAXS data and their relative population in the sample.

Note 6. For BilboMD, keep file names short and with no special characters. Input files may have to have unneeded headers and tails removed. Numbering is reset to the number one for the first residue in the file, so the residue number for the rigid domains may need to be adjusted accordingly. There is a 999 residue limit in the BILBOMD server.

These conformations are meant to serve as representative members of a larger population and should not be over-interpreted as "the true conformations" in the solution. The easiest way to use MultiFoXS is to input the multiple structures as a zip-compressed file in FoXS and then select further MultiFoXS analysis in the FoXS output window. If FoXS outputs high c2 constants near 4 for single structures, then use of ensembles in MultiFoXS can reduce the c2 constant, indicating that the protein is adopting multiple conformations in solution. The Porod exponent, determined from the reciprocal space SAXS curve, can provide additional corroboration for the presence of multiple conformations and flexibility [39,29,27,28].

5.2.4. FoXSDOCK is a valuable program for two-component systems [15]. Two structures can be rotated and translated around each other. Models are scored based on the interface, on the agreement to the experimental SAXS data, and/or a composite of the two. FoXSDOCK is available at the Sali lab website (https://modbase.compbio.ucsf.edu/foxsdock/). FoXSDOCK does not consider sequence conservation, so localization of conserved residue at the interface is expected and can serve as an additional check on accuracy (see Note 1).

5.2.5. Chimera offers a manual structure adjustment for proteins with multiple domains or oligomers. For beginners, this method leads to a more intuitive understanding of domain placement and the consequence to the scattering curve. Chimera can be downloaded from UCSF (https://www.cgl.ucsf.edu/chimera/download.html). The SAXS calculator can be found in the Tools->Higher Order Structure->Small Angle X-Ray Profile.

6. EXAMPLE: XRCC1 SOLUTION STATE

Here we provide an example of how to determine the solution state of the transient dimer of scaffold-like protein XRCC1 [42] by integrating the AlphaFold2 model, crystal structure of XRCC1-BRCT2 domain, and SAXS fitting.

6.1. AlphaFold2 model of human XRCC1.

Initially, we obtained the AlphaFold2 model of human XRCC1 monomer from the database (https://alphafold.ebi.ac.uk/) in PDB format. (Figure 3A). We used the per-residue confidence scores (pLDDT) below 50 to suggest regions unstructured in isolation (Figure 3A) and to guide conformational sampling implemented in BILBOMD [59,19].

6.2. Conformational sampling of XRCC1 monomer by BILBOMD.

The XRCC1 rigid and flexible regions that drive conformational sampling were assigned based on pLDDT value taken from AlphaFold2 initial model. While residues with pLDDT < 50 are flexible, the residues with pLDDT > 50 were keep rigid (Figure 3A). The BILBOMD server include an app that allowed interactive selection of regions within the AlphaFold2 model that create a constraints file, "const.inp" for the control of the conformational sampling. In the web app, rigid bodies are displayed as circles with the circle size proportional to the number of residues. At the same time, the flexible regions are shown as lines connecting to the circles (Figure 3B). BILBOMD outputs contain top-scoring multistate models for 1–4 states delivered by email. Additional output includes foxs_rg.out the file with the list of Rg and maximal dimension (Dmax) values for all generated models. In the XRCC1 monomer, the χ^2 value of the best-scoring one-state model is 22.3 (Figure 3C), and the best multistate model (4-states) is 2.9. As can be validated by residual of the SAXS fit, none of the single-models (Figure 3C) or multistate models satisfactory represent the solution state. This is most likely due to the presence of XRCC1 dimeric state in solution [42] therefore, we performed the subsequent conformational sampling of XRCC1-dimer.

6.3. Conformational sampling of XRCC1 dimer by BILBOMD.

SEC-MALS shows that the XRCC1 coexists as a monomer and dimer in solution [42]. Thus, we perform conformational sampling of dimeric state using the BILBOMD approach as described in step 7.2. The initial XRCC1 dimer model was built by superimposing two AlphaFold2 models at its BRCT2 interface. In the XRCC1 dimer, we would like to maintain the dimer interface at the BRCT2 domains (the XRCC1-BRCT2 crystal structure PDB 3PC8, [42]) as one rigid body during conformational sampling. As expected, the XRCC1 dimer single state or multistate model does not fit experimental SAXS data well (Figure 3C). Therefore, we fit the SAXS data with a mixture of monomer and dimer models.

6.4. Determining XRCC1 monomer/dimer solution state by MultiFOXS.

In the final step, we determined a multistate model for XRRC1 that represents the coexistence of monomeric and dimeric states in solution. As shown in Figure 3D, the XRCC1 is a disordered scaffold-like protein and adopts a mixture of monomers and dimers through its BRCT1-BRCT1 domains interaction [42]. We initially compress (zip file) four monomer and four dimer models found in the BILBOMD outputs' four-state model from steps 6.2 and 6.3. We upload the zip file into the FoXS web server [17] and fit experimental SAXS data used in steps 6.2 and 6.3, respectively. Initially, we obtained fit for all eight models with the χ^2 values between 17.6 and 59.0 (Figure 3 D). By performing multistate model fit using the MultiFoXS option in the FoXS webserver [15], we obtain an excellent fit to the SAXS data with the χ^2 values of the three-state model 1.6 (Figure 3C and 3D).

7. SUMMARY AND CONCLUSIONS

Machine learning algorithms have just started to solve protein structure prediction problems. Current limitations for homomeric and heteromeric complexes and complexes with nucleic acids and small molecules (ATP, drugs) will likely be addressed in the near future. Thus, it is important that SAXS can contribute and guide development of experimental models for conformations occurring in solution. Opening up the possibility for a SAXS-guided model being used as an AlphaFold2 template, Thomas Terwilliger has found that input of a crystallographic density-modified template into AlphaFold2 leads to a significantly improved output model that better fits the experimental density and shifts the register of a loop region [63]. Accurate solution structures will provide atomic models for testable hypotheses on substrate specificity, catalytic mechanisms, and regulation.

Although it is currently considered that an atomic structure cannot be determined from SAXS data alone, is this the point in history that we can consider SAXS validation of structure predictions as saying that an atomic structure can be determined by SAXS? With the exception of the highest resolution data, we could not derive atomic structures from crystallography, NMR, and cryo-EM data, without geometric constraints from knowledge of bond angles, bond length, secondary structures, etc. Indeed, low resolution crystallography and cryo-EM data is greatly aided by prior atomic models, either from other techniques and/or more recently, structure prediction algorithms. Given this use of structure predictions in other structural fields, can we now say that an atomic structure can be derived from SAXS data alone?

The potential application for accurate protein structure predictions goes beyond individual proteins with known functions. Multiple sequence alignments have been transformative in identifying critical residues but are dependent on sequence conservation. Although currently requiring deep sequence conservation, accurate protein structure predictions have the potential to provide a semi-orthogonal perspective for alignments. Proteins with only low sequence identity are challenging for sequence alignments, but if those critical residues co-localize in similar patterns, then better functional assignment and annotation can be achieved. High throughput structural techniques such as SAXS and SANS could validate predictions of proteins with shallow sequence depth. We have found co-localization of significant Evolutionary Tracing Scores mapped onto structures can pinpoint active sites and allosteric sites [64,65]. This will be especially critical in the metagenomic sequence space that is growing exponentially every year [66]. The Integrated microbial genomes and microbiomes (IMG) database holds ~ 70 billion genes, at the time of submission (https://img.jgi.doe.gov/cgi-bin/mer/main.cgi). Many viral and bacterial proteins currently have no or little sequence identity to other proteins, making accurate annotation difficult; anecdotally as much as 80% of a viral genome could not be annotated. As there are intensive efforts to mine the microbiome and viral metagenome for bioenergy, biomedicine, and environmental applications [67], opening up this area of genome science with accurate protein structure predictions, validated by experimental SAXS data, could accelerate finding solutions for global and health-related problems. It is also feasible that accurate and validated protein structure predictions for microbes or viruses associated with desired activity, could be run through metabolite docking programs to help identify proteins with that activity. The sensitivity of SAXS to induced domain movements o nnotation in the metagenome through functional clustering of conserved sites.

ACKNOWLEDGEMENTS

We thank long-time collaborators and colleagues in the SAXS field, including Dmitry Svergun, Robert P. Rambo, Andrej Sali, Dina Schneidman-Duhovny, Jesse Hopkins, and Daniel Rosenberg for their many insights and contributions, including useful programs for SAXS data collection and analysis. Work is supported by NCI P01 CA092584 (to S.E.T., G.H., M.H., J.A.T.) and 1R01GM137021 (To S.E.T. and G.H.). The SIBYLS beamline's efforts are supported by DOE-BER IDAT under contract DE-AC02-05CH11231.

REFERENCES

1. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A (2018) Critical assessment of methods of protein structure prediction (CASP)-Round XII. Proteins 86 Suppl 1:7–15. doi:10.1002/prot.25415 [PubMed: 29082672]

2. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J (2021) Critical assessment of methods of protein structure prediction (CASP)-Round XIV. Proteins 89 (12):1607–1617. doi:10.1002/prot.26237 [PubMed: 34533838]

3. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Applying and improving AlphaFold at CASP14. Proteins 89 (12):1711–1721. doi:10.1002/prot.26257 [PubMed: 34599769]

4. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596 (7873):583–589. doi:10.1038/s41586-021-03819-2 [PubMed: 34265844]

5. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Zidek A, Bridgland A, Cowie A, Meyer C, Laydon A, Velankar S, Kleywegt GJ, Bateman A, Evans R, Pritzel A, Figurnov M, Ronneberger O, Bates R, Kohl SAA, Potapenko A, Ballard AJ, Romera-Paredes B, Nikolov S, Jain R, Clancy E, Reiman D, Petersen S, Senior AW, Kavukcuoglu K, Birney E, Kohli P, Jumper J, Hassabis D (2021) Highly accurate protein structure prediction for the human proteome. Nature 596 (7873):590–596. doi:10.1038/s41586-021-03828-1 [PubMed: 34293799]

6. Jumper J, Hassabis D (2022) Protein structure predictions to atomic accuracy with AlphaFold. Nature methods 19 (1):11–12. doi:10.1038/s41592-021-01362-6 [PubMed: 35017726]

7. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Zidek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic acids research 50 (D1):D439–D444. doi:10.1093/nar/gkab1061 [PubMed: 34791371]

8. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millan C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ, Baker D (2021) Accurate prediction of protein structures and interactions using a three-track neural network. Science 373 (6557):871–876. doi:10.1126/science.abj8754 [PubMed: 34282049]

9. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C (2011) Protein 3D structure computed from evolutionary sequence variation. PloS one 6 (12):e28766. doi:10.1371/journal.pone.0028766 [PubMed: 22163331]

10. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences of the United States of America 108 (49):E1293–1301. doi:10.1073/pnas.1111471108 [PubMed: 22106262]

11. Andrec M, Snyder DA, Zhou Z, Young J, Montelione GT, Levy RM (2007) A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. Proteins 69 (3):449–465. doi:10.1002/prot.21507 [PubMed: 17623851]

12. Ogorzalek TL, Hura GL, Belsom A, Burnett KH, Kryshtafovych A, Tainer JA, Rappsilber J, Tsutakawa SE, Fidelis K (2018) Small angle X-ray scattering and cross-linking for data assisted protein structure prediction in CASP 12 with prospects for improved accuracy. Proteins. doi:10.1002/prot.25452

13. Svergun D, Barberato C, Koch MHJ (1995) CRYSOL– a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. Journal of applied crystallography 28 (6):768–773. doi:10.1107/S0021889895007047

14. Schneidman-Duhovny D, Kim SJ, Sali A (2012) Integrative structural modeling with small angle X-ray scattering profiles. BMC structural biology 12:17. doi:10.1186/1472-6807-12-17 [PubMed: 22800408]

15. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A (2016) FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. Nucleic acids research 44 (W1):W424–429. doi:10.1093/nar/gkw389 [PubMed: 27151198]

16. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A (2013) Accurate SAXS profile computation and its assessment by contrast variation experiments. Biophysical journal 105 (4):962–974. doi:10.1016/j.bpj.2013.07.020 [PubMed: 23972848]

17. Schneidman-Duhovny D, Hammel M, Sali A (2010) FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic acids research 38 (Web Server issue):W540–544. doi:10.1093/nar/gkq461 [PubMed: 20507903]

18. Rambo RP, Tainer JA (2013) Super-resolution in solution X-ray scattering and its applications to structural systems biology. Annual review of biophysics 42:415–441. doi:10.1146/annurev-biophys-083012-130301

19. Schneidman-Duhovny D, Hammel M (2018) Modeling Structure and Dynamics of Protein Complexes with SAXS Profiles. Methods in molecular biology 1764:449–473. doi:10.1007/978-1-4939-7759-8_29 [PubMed: 29605933]

20. Hura GL, Menon AL, Hammel M, Rambo RP, Poole FL 2nd, Tsutakawa SE, Jenney FE Jr., Classen S, Frankel KA, Hopkins RC, Yang SJ, Scott JW, Dillard BD, Adams MW, Tainer JA (2009) Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). Nature methods 6 (8):606–612. doi:10.1038/nmeth.1353 [PubMed: 19620974]

21. Classen S, Hura GL, Holton JM, Rambo RP, Rodic I, McGuire PJ, Dyer K, Hammel M, Meigs G, Frankel KA, Tainer JA (2013) Implementation and performance of SIBYLS: a dual endstation small-angle X-ray scattering and macromolecular crystallography beamline at the Advanced Light Source. Journal of applied crystallography 46 (Pt 1):1–13. doi:10.1107/S0021889812048698 [PubMed: 23396808]

22. Classen S, Rodic I, Holton J, Hura GL, Hammel M, Tainer JA (2010) Software for the high-throughput collection of SAXS data using an enhanced Blu-Ice/DCS control system. Journal of synchrotron radiation 17 (6):774–781. doi:10.1107/S0909049510028566 [PubMed: 20975223]

23. Dyer KN, Hammel M, Rambo RP, Tsutakawa SE, Rodic I, Classen S, Tainer JA, Hura GL (2014) High-throughput SAXS for the characterization of biomolecules in solution: a practical approach. Methods in molecular biology 1091:245–258. doi:10.1007/978-1-62703-691-7_18 [PubMed: 24203338]

24. Hammel M, Tainer JA (2021) X-ray scattering reveals disordered linkers and dynamic interfaces in complexes and mechanisms for DNA double-strand break repair impacting cell and cancer biology. Protein science : a publication of the Protein Society 30 (9):1735–1756. doi:10.1002/pro.4133 [PubMed: 34056803]

25. Rambo RP, Tainer JA (2013) Accurate assessment of mass, models and resolution by small-angle scattering. Nature 496 (7446):477–481. doi:10.1038/nature12070 [PubMed: 23619693]

26. Korasick DA, Tanner JJ (2018) Determination of protein oligomeric structure from small-angle X-ray scattering. Protein science : a publication of the Protein Society 27 (4):814–824. doi:10.1002/pro.3376 [PubMed: 29352739]

27. Rambo RP, Tainer JA (2011) Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. Biopolymers 95 (8):559–571. doi:10.1002/bip.21638 [PubMed: 21509745]

28. Reyes FE, Schwartz CR, Tainer JA, Rambo RP (2014) Methods for using new conceptual tools and parameters to assess RNA structure by small-angle X-ray scattering. Methods in enzymology 549:235–263. doi:10.1016/B978-0-12-801122-5.00011-8 [PubMed: 25432752]

29. Brosey CA, Tainer JA (2019) Evolving SAXS versatility: solution X-ray scattering for macromolecular architecture, functional landscapes, and integrative structural biology. Current opinion in structural biology. doi:10.1016/j.sbi.2019.04.004

30. Svergun DI, Petoukhov MV, Koch MH (2001) Determination of domain structure of proteins from X-ray solution scattering. Biophysical journal 80 (6):2946–2953. doi:10.1016/S0006-3495(01)76260-1 [PubMed: 11371467]

31. Marsh JA, Teichmann SA (2015) Structure, dynamics, assembly, and evolution of protein complexes. Annual review of biochemistry 84:551–575. doi:10.1146/annurev-biochem-060614-034142

32. Pratt AJ, Shin DS, Merz GE, Rambo RP, Lancaster WA, Dyer KN, Borbat PP, Poole FL 2nd, Adams MW, Freed JH, Crane BR, Tainer JA, Getzoff ED (2014) Aggregation propensities of superoxide dismutase G93 hotspot mutants mirror ALS clinical phenotypes. Proceedings of the National Academy of Sciences of the United States of America 111 (43):E4568–4576. doi:10.1073/pnas.1308531111 [PubMed: 25316790]

33. Brosey CA, Ho C, Long WZ, Singh S, Burnett K, Hura GL, Nix JC, Bowman GR, Ellenberger T, Tainer JA (2016) Defining NADH-Driven Allostery Regulating Apoptosis-Inducing Factor. Structure 24 (12):2067–2079. doi:10.1016/j.str.2016.09.012 [PubMed: 27818101]

34. Putnam CD, Hammel M, Hura GL, Tainer JA (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Quarterly reviews of biophysics 40 (3):191–285. doi:10.1017/S0033583507004635 [PubMed: 18078545]

35. Tsutakawa SE, Hura GL, Frankel KA, Cooper PK, Tainer JA (2007) Structural analysis of flexible proteins in solution by small angle X-ray scattering combined with crystallography. Journal of structural biology 158 (2):214–223. doi:10.1016/j.jsb.2006.09.008 [PubMed: 17182256]

36. Svergun DI (1992) Determination of the Regularization Parameter in Indirect-Transform Methods Using Perceptual Criteria. Journal of applied crystallography 25:495–503. doi:Doi 10.1107/S0021889892001663

37. Grant TD (2018) Ab initio electron density determination directly from solution scattering data. Nature methods 15 (3):191–193. doi:10.1038/nmeth.4581 [PubMed: 29377013]

38. Hura GL, Budworth H, Dyer KN, Rambo RP, Hammel M, McMurray CT, Tainer JA (2013) Comprehensive macromolecular conformations mapped by quantitative SAXS analyses. Nature methods 10 (6):453–454. doi:10.1038/nmeth.2453 [PubMed: 23624664]

39. Chinnam NB, Syed A, Burnett KH, Hura GL, Tainer JA, Tsutakawa SE (2022) Universally Accessible Structural Data on Macromolecular Conformation, Assembly, and Dynamics by Small Angle X-Ray Scattering for DNA Repair Insights. Methods in molecular biology 2444:43–68. doi:10.1007/978-1-0716-2063-2_4 [PubMed: 35290631]

40. Tully MD, Tarbouriech N, Rambo RP, Hutin S (2021) Analysis of SEC-SAXS data via EFA deconvolution and Scatter. J Vis Exp (167). doi:10.3791/61578

41. Zhou Y, Millott R, Kim HJ, Peng S, Edwards RA, Skene-Arnold T, Hammel M, Lees-Miller SP, Tainer JA, Holmes CFB, Glover JNM (2019) Flexible Tethering of ASPP Proteins Facilitates PP-1c Catalysis. Structure. doi:10.1016/j.str.2019.07.012

42. Hammel M, Rashid I, Sverzhinsky A, Pourfarjam Y, Tsai MS, Ellenberger T, Pascal JM, Kim IK, Tainer JA, Tomkinson AE (2021) An atypical BRCT-BRCT interaction with the XRCC1 scaffold protein compacts human DNA Ligase IIIalpha within a flexible DNA repair complex. Nucleic acids research 49 (1):306–321. doi:10.1093/nar/gkaa1188 [PubMed: 33330937]

43. Hammel M, Rosenberg DJ, Bierma J, Hura GL, Thapar R, Lees-Miller SP, Tainer JA (2021) Visualizing functional dynamicity in the DNA-dependent protein kinase holoenzyme DNA-PK complex by integrating SAXS with cryo-EM. Prog Biophys Mol Biol 163:74–86. doi:10.1016/j.pbiomolbio.2020.09.003 [PubMed: 32966823]

44. Wilamowski M, Hammel M, Leite W, Zhang Q, Kim Y, Weiss KL, Jedrzejczak R, Rosenberg DJ, Fan Y, Wower J, Bierma JC, Sarker AH, Tsutakawa SE, Pingali SV, O'Neill HM, Joachimiak A, Hura GL (2021) Transient and stabilized complexes of Nsp7, Nsp8, and Nsp12 in SARS-CoV-2 replication. Biophysical journal 120 (15):3152–3165. doi:10.1016/j.bpj.2021.06.006 [PubMed: 34197805]

45. Brosey CA, Yan C, Tsutakawa SE, Heller WT, Rambo RP, Tainer JA, Ivanov I, Chazin WJ (2013) A new structural framework for integrating replication protein A into DNA processing machinery. Nucleic acids research 41 (4):2313–2327. doi:10.1093/nar/gks1332 [PubMed: 23303776]

46. Tsutakawa SE, Van Wynsberghe AW, Freudenthal BD, Weinacht CP, Gakhar L, Washington MT, Zhuang Z, Tainer JA, Ivanov I (2011) Solution X-ray scattering combined with computational modeling reveals multiple conformations of covalently bound ubiquitin on PCNA. Proceedings of the National Academy of Sciences of the United States of America 108 (43):17672–17677. doi:10.1073/pnas.1110480108 [PubMed: 22006297]

47. Tsutakawa SE, Yan C, Xu X, Weinacht CP, Freudenthal BD, Yang K, Zhuang Z, Washington MT, Tainer JA, Ivanov I (2015) Structurally distinct ubiquitin- and sumo-modified PCNA: implications for their distinct roles in the DNA damage response. Structure 23 (4):724–733. doi:10.1016/j.str.2015.02.008 [PubMed: 25773143]

48. Yan C, Dodd T, He Y, Tainer JA, Tsutakawa SE, Ivanov I (2019) Transcription preinitiation complex structure and dynamics provide insight into genetic diseases. Nature structural & molecular biology 26 (6):397–406. doi:10.1038/s41594-019-0220-3

49. Baek M, Anishchenko I, Park H, Humphreys IR, Baker D (2021) Protein oligomer modeling guided by predicted interchain contacts in CASP14. Proteins 89 (12):1824–1833. doi:10.1002/prot.26197 [PubMed: 34324224]

50. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Žídek A, Bates R, Blackwell S, Yim J, Ronneberger O, Bodenstein S, Zielinski M, Bridgland A, Potapenko A, Cowie A, Tunyasuvunakool K, Jain R, Clancy E, Kohli P, Jumper J, Hassabis D (2021) Protein complex prediction with AlphaFold-Multimer. bioRxiv:2021.2010.2004.463034. doi:10.1101/2021.10.04.463034

51. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T (2018) SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic acids research 46 (W1):W296–W303. doi:10.1093/nar/gky427 [PubMed: 29788355]

52. Hatos A, Hajdu-Soltesz B, Monzon AM, Palopoli N, Alvarez L, Aykac-Fas B, Bassot C, Benitez GI, Bevilacqua M, Chasapi A, Chemes L, Davey NE, Davidovic R, Dunker AK, Elofsson A, Gobeill J, Foutel NSG, Sudha G, Guharoy M, Horvath T, Iglesias V, Kajava AV, Kovacs OP, Lamb J, Lambrughi M, Lazar T, Leclercq JY, Leonardi E, Macedo-Ribeiro S, Macossay-Castillo M, Maiani E, Manso JA, Marino-Buslje C, Martinez-Perez E, Meszaros B, Micetic I, Minervini G, Murvai N, Necci M, Ouzounis CA, Pajkos M, Paladin L, Pancsa R, Papaleo E, Parisi G, Pasche E, Barbosa Pereira PJ, Promponas VJ, Pujols J, Quaglia F, Ruch P, Salvatore M, Schad E, Szabo B, Szaniszlo T, Tamana S, Tantos A, Veljkovic N, Ventura S, Vranken W, Dosztanyi Z, Tompa P, Tosatto SCE, Piovesan D (2020) DisProt: intrinsic protein disorder annotation in 2020. Nucleic acids research 48 (D1):D269–D276. doi:10.1093/nar/gkz975 [PubMed: 31713636]

53. Piovesan D, Tabaro F, Micetic I, Necci M, Quaglia F, Oldfield CJ, Aspromonte MC, Davey NE, Davidovic R, Dosztanyi Z, Elofsson A, Gasparini A, Hatos A, Kajava AV, Kalmar L, Leonardi E, Lazar T, Macedo-Ribeiro S, Macossay-Castillo M, Meszaros A, Minervini G, Murvai N, Pujols J, Roche DB, Salladini E, Schad E, Schramm A, Szabo B, Tantos A, Tonello F, Tsirigos KD, Veljkovic N, Ventura S, Vranken W, Warholm P, Uversky VN, Dunker AK, Longhi S, Tompa P, Tosatto SC (2017) DisProt 7.0: a major update of the database of disordered proteins. Nucleic acids research 45 (D1):D219–D227. doi:10.1093/nar/gkw1056 [PubMed: 27899601]

54. Quaglia F, Meszaros B, Salladini E, Hatos A, Pancsa R, Chemes LB, Pajkos M, Lazar T, Pena-Diaz S, Santos J, Acs V, Farahi N, Ficho E, Aspromonte MC, Bassot C, Chasapi A, Davey NE, Davidovic R, Dobson L, Elofsson A, Erdos G, Gaudet P, Giglio M, Glavina J, Iserte J, Iglesias V, Kalman Z, Lambrughi M, Leonardi E, Longhi S, Macedo-Ribeiro S, Maiani E, Marchetti J, Marino-Buslje C, Meszaros A, Monzon AM, Minervini G, Nadendla S, Nilsson JF, Novotny M, Ouzounis CA, Palopoli N, Papaleo E, Pereira PJB, Pozzati G, Promponas VJ, Pujols J, Rocha ACS, Salas M, Sawicki LR, Schad E, Shenoy A, Szaniszlo T, Tsirigos KD, Veljkovic N, Parisi G, Ventura S, Dosztanyi Z, Tompa P, Tosatto SCE, Piovesan D (2022) DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. Nucleic acids research 50 (D1):D480–D487. doi:10.1093/nar/gkab1082 [PubMed: 34850135]

55. Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. Biochim Biophys Acta 1804 (4):996–1010. doi:10.1016/j.bbapap.2010.01.011 [PubMed: 20100603]

56. Forster F, Webb B, Krukenberg KA, Tsuruta H, Agard DA, Sali A (2008) Integration of small-angle X-ray scattering data into structural modeling of proteins and their assemblies. Journal of molecular biology 382 (4):1089–1106. doi:10.1016/j.jmb.2008.07.074 [PubMed: 18694757]

57. Manalastas-Cantos K, Konarev PV, Hajizadeh NR, Kikhney AG, Petoukhov MV, Molodenskiy DS, Panjkovich A, Mertens HDT, Gruzinov A, Borges C, Jeffries CM, Svergun DI, Franke D (2021) ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. Journal of applied crystallography 54 (1):343–355. doi:doi:10.1107/S1600576720013412 [PubMed: 33833657]

58. dos Reis MA, Aparicio R, Zhang Y (2011) Improving protein template recognition by using small-angle x-ray scattering profiles. Biophysical journal 101 (11):2770–2781. doi:10.1016/j.bpj.2011.10.046 [PubMed: 22261066]

59. Pelikan M, Hura GL, Hammel M (2009) Structure and flexibility within proteins as identified through small angle X-ray scattering. General physiology and biophysics 28 (2):174–189

60. Weinkam P, Pons J, Sali A (2012) Structure-based model of allostery predicts coupling between distant sites. Proceedings of the National Academy of Sciences of the United States of America 109 (13):4875–4880. doi:10.1073/pnas.1116274109 [PubMed: 22403063]

61. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. Journal of molecular biology 372 (3):774–797. doi:10.1016/j.jmb.2007.05.022 [PubMed: 17681537]

62. Hopkins JB, Gillilan RE, Skou S (2017) BioXTAS RAW: improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. Journal of applied crystallography 50 (Pt 5):1545–1553. doi:10.1107/S1600576717011438 [PubMed: 29021737]

63. Terwilliger TC, Poon BK, Afonine PV, Schlicksup CJ, Croll TI, Millán C, Richardson JS, Read RJ, Adams PD (2022) Improved AlphaFold modeling with implicit experimental information. bioRxiv:2022.2001.2007.475350. doi:10.1101/2022.01.07.475350

64. Lees-Miller JP, Cobban A, Katsonis P, Bacolla A, Tsutakawa SE, Hammel M, Meek K, Anderson DW, Lichtarge O, Tainer JA, Lees-Miller SP (2021) Uncovering DNA-PKcs ancient phylogeny, unique sequence motifs and insights for human disease. Prog Biophys Mol Biol 163:87–108. doi:10.1016/j.pbiomolbio.2020.09.010 [PubMed: 33035590]

65. Tsutakawa SE, Bacolla A, Katsonis P, Bralic A, Hamdan SM, Lichtarge O, Tainer JA, Tsai CL (2021) Decoding Cancer Variants of Unknown Significance for Helicase-Nuclease-RPA Complexes Orchestrating DNA Repair During Transcription and Replication. Front Mol Biosci 8:791792. doi:10.3389/fmolb.2021.791792 [PubMed: 34966786]

66. Chen IA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, Hajek P, Ritter S, Varghese N, Seshadri R, Roux S, Woyke T, Eloe-Fadrosh EA, Ivanova NN, Kyrpides NC (2021) The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. Nucleic acids research 49 (D1):D751–D763. doi:10.1093/nar/gkaa939 [PubMed: 33119741]

67. Iglesias A, Latorre-Perez A, Stach JEM, Porcar M, Pascual J (2020) Out of the Abyss: Genome and Metagenome Mining Reveals Unexpected Environmental Distribution of Abyssomicins. Front Microbiol 11:645. doi:10.3389/fmicb.2020.00645 [PubMed: 32351480]

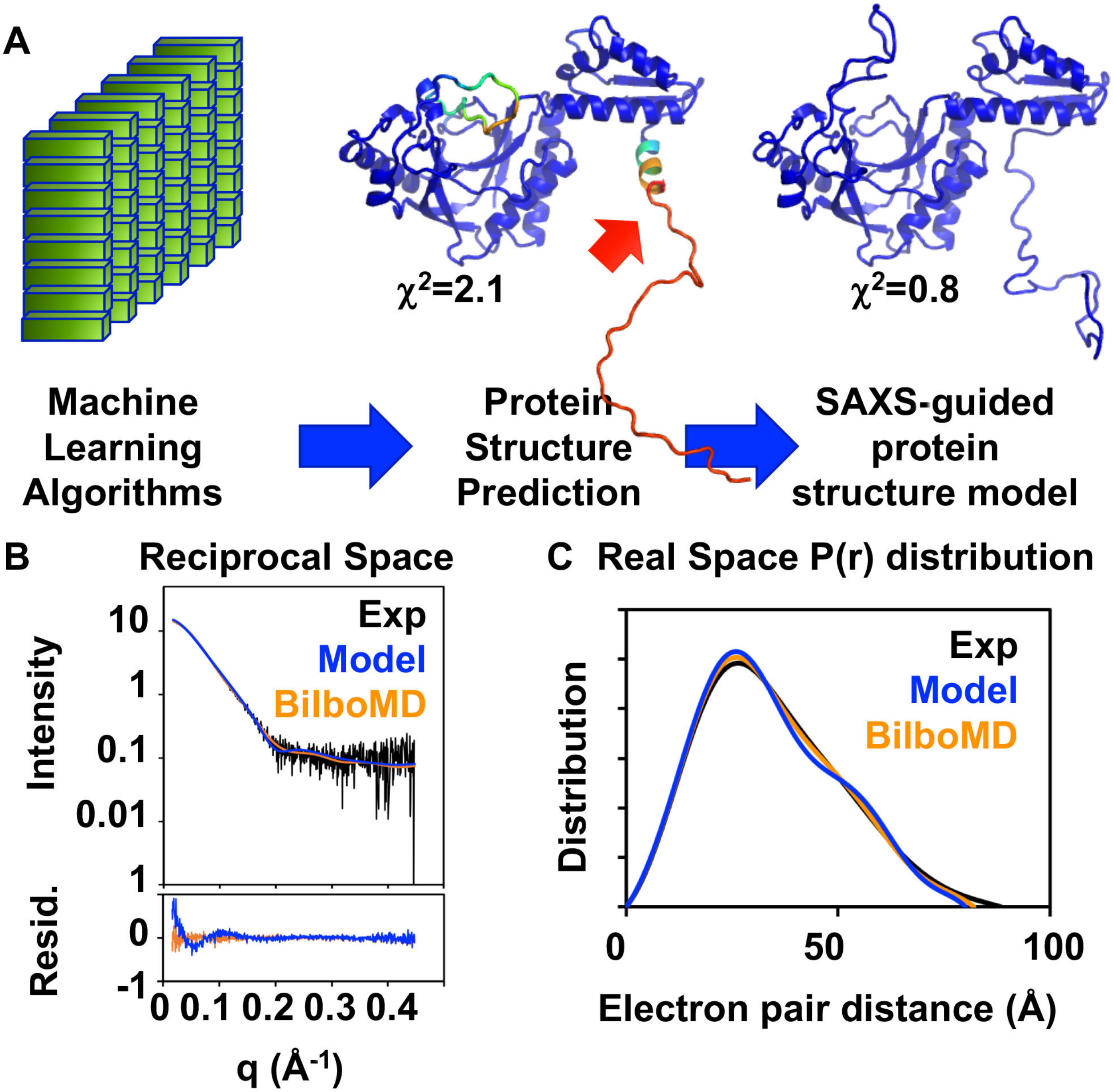


Figure 1. From predictions to solution structures.

A. Schematic showing overall strategy. Shown are representative protein predictions from AlphaFold2 and modification of that model using BILBOMD. The AlphaFold2 prediction is colored from blue to red based on confidence. Note that BILBOMD improved the fit to the experimental SAXS data and suggests that a low confidence helix (see red arrow) does not occur significantly in solution. B and C. Corresponding SAXS data in reciprocal and real space, respectively for experimental, AlphaFold2 protein structure prediction, and BILBOMD model.

Debye Formula for calculating SAXS curve from atomic model

I(q) = $\sum_{i=1}^{N}\sum_{j=1}^{N} f_i(q)f_j(q)\,\frac{sin(qd_{jn})}{qd_{jn}}$, where I is intensity as a function of q is a function of scattering angle θ and X-ray wavelength λ; q = 4π(sin (θ/2))/ λ; q = 4π(sin (θ/2))/ λ, d is the distance between atoms i and j, and N is the number of atoms. The form factor takes into account the displaced solvent and the density of the hydration layer (c1 and c2), which are adjusted when experimental data is input. $f_i(q)=f_v(q)-c_1f_s(q)+c_2s_if_w(q)$, where $f_v(q)$ is the atomic form factor *in vacuo*, $f_s(q)$ is the form factor of the displaced solvent, s_i is the fraction of solvent accessible surface of the atom *i*, and $f_w(q)$ is the water form factor.

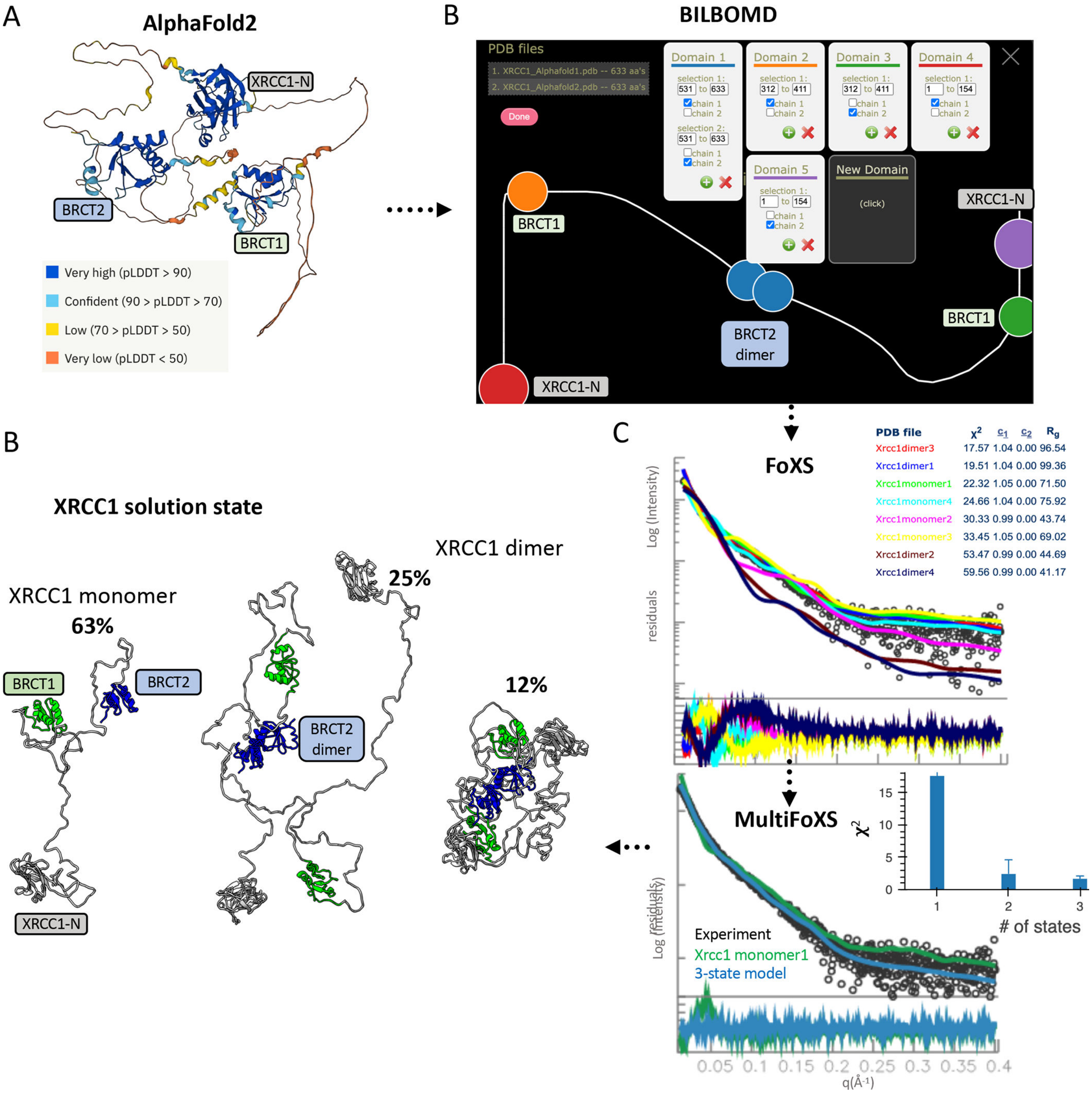
ChiSquare for the differential calculation of two SAXS curves, taking into account experimental error

$\chi^2 = \sum_{i=1}^{N} \frac{\left(\frac{I_1(q_i)}{m_1} - \frac{I_2(q_i)}{m_2}\right)^2}{\left(\frac{\sigma_1(q_i)}{m_1} + \frac{\sigma_2(q_i)}{m_2}\right)^2}$, where I1 and I2 are the intensities of the two curves, m1 and m2 are the Intensity means and σ1 and σ2 are the errors for the two SAXS measurements to be compared. L is the number of data points. The ideal is 1. Inclusion of error can drop the value < 1, so the minimum number would be considered the best fit. This is a difference-based metric, with bias towards low q and how the curves are scaled.

Volatility Ratio

$V_R = \sum_{i=1}^{N} abs\left[\frac{R(q_i)-R(q_{i+1})}{(R(q_i)+R(q_{i+1}))/2}\right]$, where R is the ratio of the intensities at q_i when ratio is normalized so the average over the range is 1 and where the ratio is binned at a minimal frequency (Δq = π/d where q is a function of scattering angle θ and X-ray wavelength λ; q = 4π(sin (θ/2))/ λ. By assuming a maximum dimension d < 40 nm, the number of bins N is 25 over a q range, q < 0.2Å^{-1}. This is a ratio-based metric, with weight distributed through the curve.

Figure 2. Equations used for comparison of SAXS curves in reciprocal space.



PDB files

1. XRCC1_Alphafold1.pdb -- 633 aa's
2. XRCC1_Alphafold2.pdb -- 633 aa's

Done

Domain 1
selection 1: 531 to 633
☑chain 1
□chain 2
selection 2: 531 to 633
□chain 1
☑chain 2

Domain 2
selection 1: 312 to 411
☑chain 1
□chain 2

Domain 3
selection 1: 312 to 411
□chain 1
☑chain 2

Domain 4
selection 1: 1 to 154
☑chain 1
□chain 2

Domain 5
selection 1: 1 to 154
□chain 1
☑chain 2

New Domain
(click)

PDB file	χ^2	c_1	c_2	R_g
Xrcc1dimer3	17.57	1.04	0.00	96.54
Xrcc1dimer1	19.51	1.04	0.00	99.36
Xrcc1monomer1	22.32	1.05	0.00	71.50
Xrcc1monomer4	24.66	1.04	0.00	75.92
Xrcc1monomer2	30.33	0.99	0.00	43.74
Xrcc1monomer3	33.45	1.05	0.00	69.02
Xrcc1dimer2	53.47	0.99	0.00	44.69
Xrcc1dimer4	59.56	0.99	0.00	41.17

Figure 3. XRCC1 solution state as identified by integrating AlphaFold2 protein structure prediction and SAXS modeling.

(A) AlphaFold2 model colored by per-residue confidence score (pLDDT) values. (B) Building of restraining file that controls BILBOMD conformational sampling of XRCC1 dimer by selecting rigid domain regions based on pLDDT values. (C) FoXS web server output shows multiple SAXS fit, residual, χ^2 values, and Radius of gyration for four XRCC1 monomers and four dimer models derived from BILBOMD modeling. Bottom panel-MultiFoXS implementation in FoXS web server show χ^2 values for one-, two- and three state model. The plot show comparison of SAXS fit and fit-residual for the one- and three-state model. (D) Three state model is shown together with the percentage of each model used to fit the SAXS data shown in panel C. Rigid body domains are highlighted.