Machine Learning Methods for X-ray Scattering Data Analysis from Biomacromolecular Solutions

D. Franke[1,2], Cy M. Jeffries[1], D. I. Svergun[1]

[1] European Molecular Biology Laboratory, Hamburg Outstation Notkestrasse 85, D22607 Hamburg, Germany.

[2] Correspondence to: franke@embl-hamburg.de

Abstract

Small Angle X-ray scattering (SAXS) of biological macromolecules in solutions is a widely employed method in structural biology. SAXS patterns include information about the overall shape and low-resolution structure of dissolved particles. Here we describe how to transform experimental SAXS patterns to feature vectors and how a simple k-nearest neighbour approach is able to retrieve information on overall particle shape, maximum diameter (D_{max}) as well as Molecular Mass (MM) directly from experimental scattering data. Based on this transformation we develop a rapid multi-class shape-classification ranging from compact, extended and flat categories to hollow and random-chain like objects. This classification may be employed e.g. as a decision block in automated data analysis pipelines. Further we map protein structures from the protein data bank (PDB) into the classification space and, in a second step, use this mapping as a data source to obtain accurate estimates for structural parameters (D_{max} , MM) of the macromolecule under study based on the experimental scattering pattern alone, without inverse Fourier transform for D_{max}. All methods presented are implemented in a Fortran binary DATCLASS, part of the ATSAS data analysis suite, available on Linux, Mac, Windows and free for academic use (https://www.embl- hamburg.de/biosaxs/software.html).

Introduction

Small angle X-ray scattering (SAXS) is an increasingly popular method in structural biology that usefully complements high-resolution structural techniques such as X-ray crystallography (MX), nuclear magnetic resonance spectroscopy (NMR) and electron microscopy (EM). SAXS does not require crystals, labelling or isolated particles at cryogenic temperatures and its applications extend to the determination of structural parameters, e.g. the Radius of Gyration (R_g) maximum extend (D_{max}) and the molecular mass (MM), obtaining the low-resolution shapes of macromolecules and rigid body modelling of complexes, quantitative characterization of flexibility, as well as time-resolved conformational changes (1). The scattering intensity $I(q)$ is recorded as a function of the scattering vector q, where the momentum transfer $q = 4\pi sin\ \theta/\lambda$ where θ corresponds to half of the angle between incoming and scattered photons. To determine the scattering of the macromolecule under study, the background scattering, including sample holder and solvent, typically an aqueous buffer, has to be subtracted.

Over time many methods have been developed to extract relevant information directly from the experimental scattering intensities, exclusively working with the experimentally obtained data. In contrast, in this manuscript we consider the application of Data Mining and Machine Learning (2) to extract structural information from SAXS data. In short, we shall evaluate the idea that, if there were a way to locate *similar* macromolecules with known structural parameters, the parameter values of these similar structures could be used to approximate the parameter values of the specimen under study. It is to note that in this context *similarity* shall refer to similarity in scattering patterns, with the assumption that similar scattering pattern implies similar overall structure, and not necessarily similar higher resolution detail; the latter may not be the case (3).

For each of the major methods in structural biology curated data banks invite researchers to deposit models as well as raw data, in particular: the Protein Data Bank (PDB) (4), the Biological Magnetic Resonance Data Bank (BMRB) (5), the Electron Microscopy Data Bank (EMDB) (6), and the Small Angle Scattering Biological Data Bank (SASBDB) (7), respectively. Here, a large number of records on structural parameters, sequences, shapes, models, and more, have been accumulated. Using tools like CRYSOL (8) or FoXS (9), theoretical scattering patterns of atomic models may be readily calculated.

Finally we bring the initial idea and available data together by describing methods on how to make large amounts of data accessible for *Knowledge Discovery*. In particular, in the context of Data Mining and Machine Learning any measurable property of the specimen under study may be considered a *feature*. Features describe the input for a machine learning method and may be concrete values or abstract concepts. In SAXS, the experimental R_g, the calculated forward scattering $I(0)$ as well as the individual experimental intensities at each q and any function thereof may be considered potential features. In this manuscript we shall describe how to represent the overall shape of a protein, e.g. compact, flat, extended or random chain, with only three shape-related features. Here, random-chains are a mixture of conformations ranging from a compact to fully extended chains, while extended only refers to preferred extended particles in solution. Further, to predict structural parameters, a fourth, size-related, feature may be included in the *feature vector*. The advantage of describing a complex SAXS pattern in a feature vector of only a few components becomes apparent if one assumes a form of distance relationship between feature vectors. If two points in the feature space are close together in the Euclidian sense, then their properties, i.e. shape and/or structural parameters, should be similar. Conversely, if they are far apart, their properties should be significantly different. In order to predict properties of an unknown entity, one may look up its closest neighbour(s) in the feature space and apply known properties of the neighbour to the unknown entity. However, the larger the number of components in the feature vector, i.e. the more dimensions are considered, the more likely are sparsely populated regions in the underlying data source that could reduce predictive power, a problem also known as the "Curse of Dimensionality" (10).

Here we present a framework of data transformation and feature selection for a fast and selective lookup of structural neighbours in the space of SAXS patterns. Based on the proposed feature selection and the source data of the database, different information may be inferred. In the case of geometrical bodies (11), simple shapes may be determined quickly, e.g. for use as a proto-shape for *ab initio* modelling, in the case of the PDB (4), structural parameters such as D_{max} and *MM* of the immediate neighbours, as discussed in this work, but also others of interest, may be looked up and be used as a starting point for further analysis and refinement.

Methods

Shape classification

Data Simulation

The command-line program BODIES (11) was modified to simplify the automated simulation of large amounts of SAXS patterns derived from geometrical objects with uniform scattering length density of compact spheres, flat discs, extended rods, compact-hollow cylinders, hollow spheres and flat rings (Figure 1a). The corresponding dimensions of the geometric bodies, i.e. inner and outer radius, height, length and width, etc., were uniformly and independently sampled in ranges from 10 to 500 Ångstrom, respectively. Classification labels were generated based on the extent of the object, in short: proportions more or less extreme than 1:4 were considered to define compact, extended and flat objects, in addition a inner cavity of more than 25% of the outer radius generally indicates a hollow object. Based on this, 460 000 scattering patterns of various compact, flat, extended, filled and hollow geometric objects were generated. While clearly limited a selection of body types, enumerating an exhaustive list of geometrical body shapes would be, at least, very difficult, especially considering the lack of analytical form factors. As shown later in the text, classification with *k*-nearest neighbours extends somewhat outside the boundaries of the mapped class volumes, thus smoothing out any gaps between geometric objects (Figure 1d). Further, in order to allow identification of intrinsically-disordered proteins, we employed EOM (12) to generate an additional 560 000 simulations of random chains subsequently averaged in groups of 20 repetitions to simulate mixtures of flexible proteins. The lengths of the random chains were selected to follow the size distribution of amino acid sequences of asymmetric units in the PDB. In total, 488 000 scattering patterns were created across all geometric classes to be used as training data set for machine-learning classification that encompass basic geometric objects and disordered polymer chains (Figure 1d).

Data Transformation

To normalize for the varying size of objects, R_g and forward scattering $I(0)$ were required. As the generated data is ideal and free of noise, the R_g was obtained from the slope of the Guinier plot (ln$I(q)$ vs q^2) of the first ten computed points and $I(0)$ was directly available from the data due to simulation. With these two parameters, the data was transformed to the dimensionless Kratky scale (13):

$$\left(qR_g\right)^2 I\left(qR_g\right)/I(0)\ vs.\ qR_g$$

Following this, the normalised Porod invariant, or integral Q', of the dimensionless Kratky plot was calculated up to qR_g=3, qR_g=4 and qR_g=5, respectively, and expressed as a *normalized apparent volume*, or V' (14), i.e.

$$V' = \frac{2\pi^2}{Q'}\ \ where\ \ Q' = \int_0^{qR_g} \left(qR_g\right)^2 I\left(qR_g\right)/I(0) dqR_g.$$

Each scattering pattern was therefore reduced to three features and its associated class label (Figure 1b,c). The qR_g upper bounds were chosen as they provide a trade-off between contained shape information and the limitations of the assumption of uniform scattering length density; larger qR_g values would separate the point clouds in unrealistic ways (not shown). That said, with the present selection, the corresponding three-dimensional scatter plot of the simulated data shows a V'-space with good separation of the different shape classes (Figure 1c,d).

Learning, prediction, validation

As Figure 1d depicts a well-defined point cloud within the three-dimensional V'-space, we added 25 000 randomized points with *unknown* class label to the space prior to learning. This helped to facilitate compactness of the resulting predictions, otherwise a query point outside this well-defined V' would still have, far away, neighbours and would thus be grouped to a class it does not belong. It is to note that this random point cloud is not shown in Figure 1d as it would obscure the actual data of interest.

To classify the shape of an unknown entity, its feature vector has to be computed and the k-nearest-neighbours in the three-dimensional V'-space are determined by kd-tree search (15) across the whole training set. Here we chose k=9, partly to avoid *unknown* classification of the randomly distributed cases, but also to facility a majority vote classification where classes overlap., The classes of the neighbours are then weighted by empirical class weights (Supp. Tab. 3) and the class with the maximum sum of weights is selected as label for the unknown entity.

To evaluate the performance of this approach, we used leave-one-out cross-validation, i.e. we removed each of the 488 000 structures from the source data in turn and used the remaining data points to predict the class of the removed one. Cross-validated performance of this multi-class classifier was evaluated by F1-Measure and Matthew's Correlation Coefficient (MCC) (16).

Prediction of structural parameters

Data Generation

A snapshot of more than 220 000 asymmetric units and biological assemblies was taken from the PDB (4). From these we discarded duplicates (i.e. biological assemblies identical to asymmetric units), entries with nucleotides as well as peptides with less than 50 amino acids. Entries with more than one MODEL were discarded unless the models were very similar, in which case we used the first one listed in the atomic coordinate file. Metals, inorganic molecules and other post-translational additions where filtered out from all structures. No filtering was applied with respect to sequence identity, as similarity in sequence does not always imply similarity in structure (17). From the remaining 165 982 unique atomic structures, we calculated scattering patterns using CRYSOL (8) using 30 spherical harmonics and 1001 equidistant points up to a q_{max} of 0.6A^{-1}. Besides the calculated scattering pattern CRYSOL also reports a variety of structural parameters, in particular R_g, D_{max} and MM, which we recorded for later use.

Learning, prediction and validation

Similar to the geometric bodies, the V' values were computed for the atomic structures. Given that for the estimation of structural parameters not only the shape, but also the size of the molecule is important, R_g was included as a size feature in addition to the three V' shape features; here, R_g was chosen over D_{max} as the former can be directly obtained from the experimental data whereas the latter can usually only indirectly be estimated.

To assess the structural parameters of an unknown entity, the feature vector is computed and the k-nearest structural neighbours, here k=5, in a four-dimensional space combining the three dimensions of V' along with R_g, are determined by kd-tree search (15). Here, the parameter k=5 was chosen to minimize the relative prediction error. From this, the parameters, i.e. D_{max} and MM, are estimated as weighted mean of D_{max} and MM of the neighbours, were the weights correspond to the normalized inverse Euclidean distance to the unknown entity, i.e. the closer the neighbour, the more important its contribution to the prediction.

To evaluate the performance of this approach, we used leave-one-out cross-validation, i.e. we removed each of the 165 982 structures from the source data in turn and used the remaining structures to predict D_{max} and MM of the removed structure.

Application of shape classification and prediction of structural parameters to experimental data

The classifier was further applied to the 401 public experimental SAXS data sets without nucleotides, available from SASBDB (7) at the time of writing. As random-chain classifications may potentially indicate modular, flexible or unfolded proteins, we also collected experimental SAXS data on folded and chemically-modified unfolded ribonuclease A and folded and denatured Lipase B at the EMBL P12 SAXS beam line at PETRA-III (18), DESY, Hamburg, Germany, to compare the results of the random-chain classification with those from traditional biophysical methods, i.e. CD spectropolarimetry and tryptophan fluorescence spectroscopy. See Supplementary Material for details on their preparation.

To study the effects of experimental noise on shape classification and prediction of structural parameters, we further collected experimental data of 100 repetitions of 50 ms exposures of bovine serum albumin (BSA) in 50 mM HEPES, pH 7.5 buffer. After subtracting 100 buffers from 100 samples, the resulting 100 data sets were identical up to noise as evaluated by CorMap (19).

All experimental data was submitted to SASBDB for reference. The following accession codes were assigned: SASDDK3 (Lipase B), SASDDL3 (folded ribonuclease A), SASDDM3 (chemically unfolded ribonuclease A) and SASDDN3 (100 repetitions of bovine serum albumin; buffers, samples and subtracted data were deposited).

Results

Shape Classification

Appropriate evaluation of multi-class classification systems is itself a topic of ongoing research. In this work we follow the recommendations of Powers (2011) and report F1 score and MCC for each shape category (Table 1). Here, F1 is a measure that considers precision and recall of the classifier with a range between 0.0 and 1.0, correspondingly MCC determines the correlation between expected and predicted classes with a range from -1.0 to 1.0. In both cases, larger (positive) values are associated with better performance. In addition, Supp. Figure 3 details the confusion matrix, i.e. the actual counts of expected and predicted classes of the Leave-One-Out cross validation, together with recall and precision percentages in the margins. Overall accuracy of classification across all shapes is reported as 96.5%.

	F1 score	MCC
Unknown	0.991	99.1 %
Compact	0.962	95.1 %
Extended	0.969	95.8 %
Flat	0.957	94.7 %
Ring	0.980	97.8 %
Compact-hollow	0.938	93.3 %
Hollow-sphere	0.997	99.7 %
Random-chain	0.964	96.2 %

Table 1: F1 score and Matthew's Correlation Coefficient (MCC) for k-nearest neighbours multi-class classification results of the individual shape categories.

Further, we predicted the shape classification of the 165 982 unique atomic structures of the PDB and visualized the resulting point cloud in *V'*-space (Figure

2a). It is immediately apparent that the overall shape of the distribution of proteins (opaque circles) is very similar to that obtained by geometric objects (transparent background), with only 25 structures considered outside the volume mapped by the geometric objects and thus being assigned an “unknown” class label (open circles). Interestingly, most (~90%) of the PDB structures are classified as compact/globular, while, for example, more extended proteins are much less represented (~3%). A different picture arises from experimental data deposited in SASBDB (Figure 2b). Here the distribution (Supp. Tab. 4) tends more towards the extended, flat and random-chain area (>50%), reflecting the fact that solution scattering is often employed for systems that do not easily crystallize. Indeed, the shape classification of experimental SAXS data may also be done to describe protein solution state or solution state transitions when the high-resolution structure is not available or obtainable. For example, Figure 2c,d show the V'-space point cloud positions of SAXS data obtained from native ribonuclease A compared to a final-state completely denatured protein highlighting the shift from compact to random/flexible shape categories. SAXS data collected from Lipase B samples that underwent systematic chemical denaturation shows the ‘denaturation trace’ through V'-space as the protein populations unfold at ever-increasing concentration of guanidine hydrochloride.

Prediction of Structural Parameters

Figure 3a,c summarize the results of the Leave-One-Out cross validation for prediction of structural parameters of the PDB. As the values of the parameters are derived from the atomic structures, a good agreement may be expected; in about 90% of the cases the estimate is within 10% of the true value. The evaluation of experimental data as deposited in SASBDB (Figure 3b,d) is not as straightforward as the deposited values depend on sample quality, experimental conditions and the data analysis of the respective researcher. Interestingly, compared to the results of the PDB, there seems to be a tendency to obtain somewhat larger D_{max} values in manual analysis (Figure 3b), which may, for example, be explained by the influence of the hydration shell.

Effects of Experimental Noise

Figure 4 elucidates the effect of experimental noise on 100 repetitions of bovine serum albumin; all frames were found similar to each other up to noise as per CorMap test (19). As depicted in panel (a), the mapped locations of the 100 frames are slightly spread out, but still close together. Histograms of the estimated structural parameters D_{max} and MM are shown in Figure 4(b) and 4(c) respectively. Again, a spread may be observed, however, the width of the distributions most likely correlates strongly with the amount of noise present in the data (not evaluated). Both distributions are centered on values somewhat larger than what one may expect from strictly monomeric BSA (~100 Ångstrom and ~67kDa, respectively), but this may be attributed to the presence of a fraction of dimers in solution (20).

Discussion

Rapid shape classification as presented in this work is a unique approach in the field of biological SAXS. However, it is obvious that accurate estimates of R_g and $I(0)$ are key for appropriate transformation of experimental SAXS data to V'-space. Interestingly, misspecification of these parameters will often result in a data point outside the body of shape space as depicted by Figure 1(d) and consequently lead to an “unknown” classification; therefore the shape classification may also be used as an initial validation of R_g and $I(0)$. Further, it has applications as a building block for automated data analysis (21, 22, 23), e.g. to decide whether *ab initio* shape modeling or ensemble optimization should be applied. In addition, shape modeling applications may use the initial classification as a starting point for their models; DAMMIF (24) has already been modified to not only use a start model based on the classification, but also to adapt the search and annealing parameters, e.g. by enabling anisometry penalties for extended or flat objects.

Similarly, at present D_{max} may only be obtained by inverse Fourier transform of the experimental scattering pattern, which may be difficult to determine accurately (25, 26). The presented method provides an independent D_{max} estimate from similar entries in the PDB based on experimental data alone. Consequently, this approach may be applied to obtain a starting estimate of D_{max} for the indirect Fourier transform, or, as a tool for quality assessments during data deposition procedures, e.g., to SASBDB, whereby the automated D_{max} estimates may be compared to submitted values for validation purposes (Figure 3a).

In the past, multiple concentration-independent methods to determine the *MM* of biological macromolecules from SAXS data have been established (14, 27, 28), each with their own respective strengths and weaknesses. In this manuscript we report the results of the size-and-shape based database lookup method (Figure 3b) without attempting to directly compare with any of the established methods. The interested reader may find a thorough, comprehensive and quantitative comparison of all four methods elsewhere (29).

It is to note that some details of the presented method were empirically determined, e.g. the qR_g integration limits for V': while the general magnitude is appropriate, e.g. on the lower end integration to qR_g=1 corresponds to the Guinier range and on normalized scale the integral is a constant up to rounding errors. Consequently, on the higher end qR_g=10 would correspond to wider-angle i.e. higher resolution information that is not easy to rationalize in terms of overall parameters. Thus the selected qR_g values of 3, 4 and 5 are reasonable, but not necessarily optimal. For example, we chose N=3 integration limits also for the ease of display. A different selection of limits in number and magnitude might result in an improved predictive performance. Along the same line of argument one may observe that in many machine-learning applications it is required to normalize, scale or transform the training data prior to learning and prediction to achieve a good predictive result. Here, we used the data “as-is”, however, it is possible that there is a transformation function that minimizes the relative error and/or (root) mean square error of the prediction. Potential avenues of investigation for the *k* nearest neighbours method include: (a) selection of *k* and the applied distance weights (b) arbitrary linear and non-linear data scaling and transformation prior to learning; (c) metric selection and metric learning (30); (d) and, of course, any other learning method as regression functions, Support Vector Machines, Neural Networks, Deep Learning, etc. As in this manuscript we focus on outlining and introducing a novel approach, we did not exhaustively investigate all these options; however, the classifier as presented here is already on par with established methods (29).

Conclusion

In this manuscript we present a conceptually new approach to rapidly analyze the scattering patterns in biological SAXS, not as an isolated data point, but in the context of all known biological macromolecules. We have outlined and described a simple data transformation that combines large amounts of SAXS data into a few numbers that suggest themselves as coordinates in a feature space for machine learning. This space simplifies and improves lookup of similar scattering patterns in a large dataset. The presented approach of integrating the intensities has a strong advantage over the methods based on actual (normalized) intensity values. Our method is independent of the spacing of the available data points, obviating the need for interpolation to a common grid and fluctuations of individual intensities have less of an effect for lookup due to the integration, thus also avoiding the "Curse of Dimensionality".

The techniques described here allow for rapid shape classification and provide estimates of MM and D_{max} with good accuracy. It is to note that so far D_{max} was only available indirectly through inverse Fourier transform, but with the new approach it is now also accessible from experimental data directly. Further, the general approach as described easily extends to additional parameters of interest extracted from source data as labels may be assigned arbitrarily.

The method has been implemented in the program DATCLASS, integral part of the ATSAS data processing and analysis suite (31) which is freely available for academic users (https://www.embl- hamburg.de/biosaxs/software.html).

Author Contributions

The initial idea was conceived of and all developments were done by D.F. Experimental data was collected by C.M.J.. D.F, C.M.J and D.I.S. participated in critical discussion and wrote the manuscript.

Acknowledgements

This work was supported by iNEXT, project number 653706, funded by the Horizon 2020 programme of the European Union, Bundesministerium für Bildung und Forschung [grant TTSAS number 05K2016], and the Human Frontier Science Program [grant number RGP0017/2012].

Supporting Citations

References (32-34) appear in the Supporting Material.

References

1. Svergun, D. I., M. H. J. Koch, P. A. Timmins and R. P. May. 2013. Small Angle X-Ray and Neutron Scattering from Solutions of Biological Macromolecules Oxford University Press.
2. Fayyad, U., G. Piatetsky-Shapiro and P. Smyth. 1996. From Data Mining to Knowledge Discovery in Databases. AI Magazine. 17:37-54.

3. Petoukhov, M. V. and D. I. Svergun. 2015. Ambiguity assessment of small-angle scattering curves from monodisperse systems. Acta Cryst D. 71:1051-1058.
4. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne. 2000. The Protein Data Bank. Nucleic Acids Res. 28:235-242.
5. Ulrich, E. L., H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. K. Wenger, H. Yap and J. L. Markley. 2008. BioMagResBank. Nucleic Acids Res. 36:D402-D408.
6. Lawson, C. L., M. L. Baker, C. Best, C. Bi, M. Dougherty, P. Feng, G. van Ginkel, B. Devkota, I. Lagerstedt, S. J. Ludtke, R. H. Newman, T. J. Oldfield, I. Rees, G. Sahni, R. Sala, S. Valankar, J. Warren, J. W. Westbrook, K. Henrick, G. J. Kleywegt, H. M. Berman and W. Chiu. 2011. EMDataBank.org: unified data resource for CryoEM. Nucleic Acids Res. 39:D456-D464.
7. Valentini, E., A. G. Kikhney, G. Previtali, C. M. Jeffries and D. I. Svergun. 2014. SASBDB, a repository for biological small-angle scattering data. Nucleic Acids Res.
8. Svergun, D. I., C. Barberato and M. H. J. Koch. 1995. CRYSOL – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. J. Appl. Cryst. 28:768-773.
9. Schneidman-Duhovny, D., M. Hammel and A. Sali. 2010. FoXS: A Web server for Rapid Computation and Fitting of SAXS Profiles. NAR. 38:W540-544.
10. Bellman, R. E. 1957. Dynamic Programming Princeton University Press.
11. Konarev, P. V., M. V. Petoukhov, V. V. Volkov and D. I. Svergun. 2006. ATSAS 2.1, a program package for small-angle scattering data analysis. J Appl Cryst. 39:277-286.
12. Tria, G., H. D. T. Mertens, M. Kachala and D. I. Svergun. 2015. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. IUCrJ. 2:207-217.
13. Durand, D., C. Vivès, D. Cannella, J. Pérez, E. Pebay-Peyroula, P. Vachette and F. Fieschi. 2010. NADPH oxidase activator p67phox behaves in solution as a multidomain protein with semi-flexible linkers. J Struct Biol. 169:45-53.
14. Fischer, H., M. de Oliveira Neto, H. B. Napolitano, I. Polikarpov and A. F. Craievich. 2010. Determination of the molecular weight of proteins in solution from a single small-angle X-ray scattering measurement on a relative scale. J Appl Cryst. 43:101-109.
15. Bentley, J. L. 1975. Multidimensional Binary Search Trees Used for Associative Searching. Comm ACM. 18:509-517.
16. Powers, D. M. W. 2011. Evaluation: from Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies. 2:37-63.
17. Kosloff, M. and R. Kolodny. 2008. Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins. 71:891-902.
18. Blanchet, C., S. Spilotros, F. Schwemmer, M. A. Graewert, A. G. Kikhney, C. M. Jeffries, D. Franke, D. Mark, R. Zengerle, F. Cipriani, S. Fiedler, M. Roessle and

D. I. Svergun. 2015. Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA-3, DESY). J Appl Cryst. 48:431-443.

19. Franke, D., C. M. Jeffries and D. I. Svergun. 2015. Correlation Map, a method to quantitatively assess systematic differences for the analysis of one dimensional spectra. Nat Meth. 12:419-422.
20. Jeffries, C. M., M. A. Graewert, C. E. Blanchet, D. B. Langley, A. E. Whitten and D. I. Svergun. 2016. Preparing monodisperse macromolecular samples for successful biological small-angle X-ray and neutron-scattering experiments. Nat Prot. 11:2122-2153.
21. Brennich, M., J. Kieffer, G. Bonamis, A. De Maria Antolinos, S. Hutin, P. Pernot and A. Round. 2016. Online data analysis at ESRF BioSAXS beamline BM29. J Appl Cryst. 49:203-212.
22. Franke, D., A. G. Kikhney and D. I. Svergun. 2012. Automated acquisition and analysis of small angle X-ray scattering data. Nucl Inst Meth A. 689:52-59.
23. Shkumatov, A. V. and S. V. Strelkov. 2015. DATASW, a tool for HPLC-SAXS data analysis. Acta Cryst. D71:1347-1350.
24. Franke, D. and D. I. Svergun. 2009. DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. J Appl Cryst. 42:342-346.
25. Glatter, O. & Kratky, O. eds. 1982. Small-angle X-ray Scattering Academic Press. London.
26. Svergun, D. I. 1992. Determination of the Regularization Parameter in Indirect-Transform Methods Using Perceptual Criteria. J Appl Cryst. 25:495-503.
27. Porod, G. 1951. Die Roentgenkleinwinkelstreuung von dichtgepackten kolloidalen Systemen, 1. Teil. Kolloid Z. 124:83-114.
28. Rambo, R. P. and J. A. Tainer. 2013. Accurate assessment ofmass, models and resolution by small-angle scattering. Nature. 496:477-481.
29. Hajizadeh, N. R., D. Franke, C. M. Jeffries and D. I. Svergun. 2018. Consensus Bayesian assessment of protein molecular mass from solution X-ray scattering data. Sci Rep. In print.
30. Xing, E. P., A. Y. Ng, M. I. Jordan and S. Russel. 2003. Distance metric learning, with application to clustering with side-information. Advances in Neural Information Processing Systems 15:505-512.
31. Franke, D., M. V. Petoukhov, P. V. Konarev, A. Panjkovich, A. Tuukkanen, H. D. T. Mertens, A. G. Kikhney, N. R. Hajizadeh, J. M. Franklin, C. M. Jeffries and D. I. Svergun. 2017. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. J Appl Cryst. 50:1212-1225.

Supporting References

32. Gasteiger, E., C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel and A. Bairoch. 2005. The Proteomics Protocols Handbook Humana Press.
33. Micsonai, A., F. Wien, L. Kernya, Y. H. Lee, Y. Goto, M. Réfrégiers and J. Kardos. 2015. Accurate secondary structure prediction and fold recognition for

circular dichroism spectroscopy. Proc. Natl. Acad. Sci. U.S.A.

34. Wang, Y., J. Trewhella and D. P. Goldenberg. 2008. Small-Angle X-ray Scattering of Reduced Ribonuclease A: Effects of Solution Conditions and Comparisons with a Computational Model of Unfolded Proteins. J Mol Biol. 377:1576-1592.

1

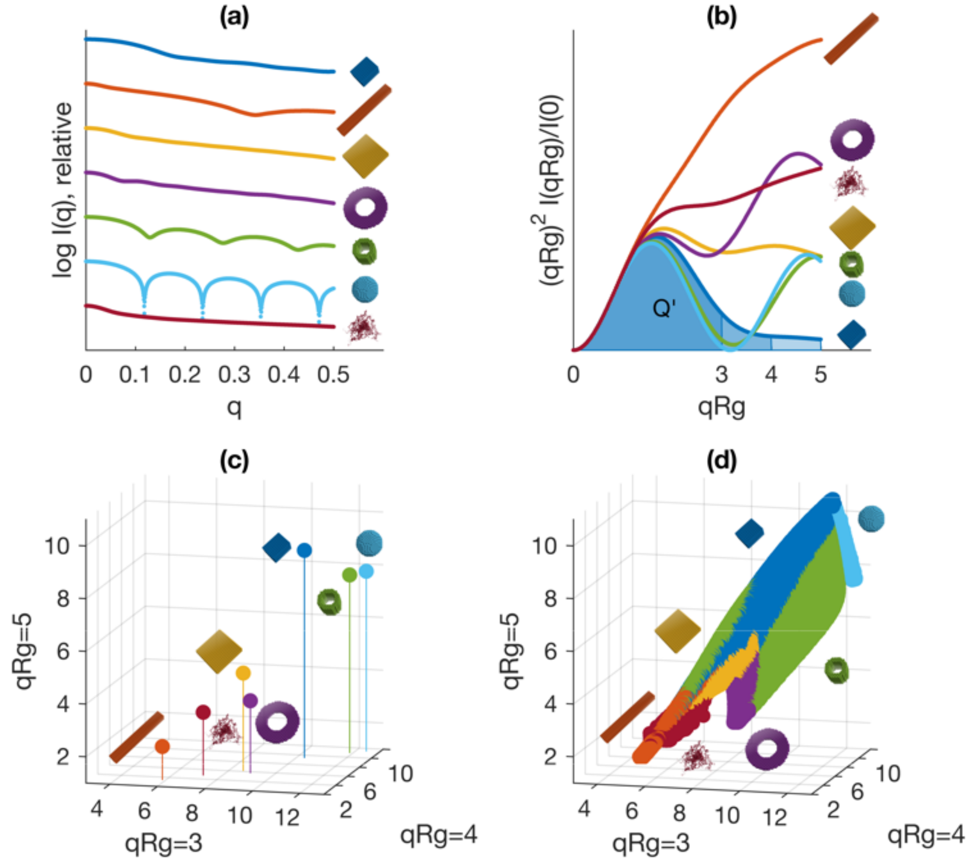


Figure 1: Transformation of scattering patterns of geometric objects and random-chain on arbitrary log scale (a) via integration of the normalized Kratky Plot (b) to *V*'-space (c, d). Panels (a,b,c) depict a randomly selected member of each object class, while panel (d) shows the locations of all 488.000 scattering patterns generated. Color assignments are identical in all panels: compact (dark blue), extended (orange), flat (yellow), ring (violet), compact-hollow (green), hollow-sphere (light blue) and random-chain (dark red), also indicated by corresponding pictograms.

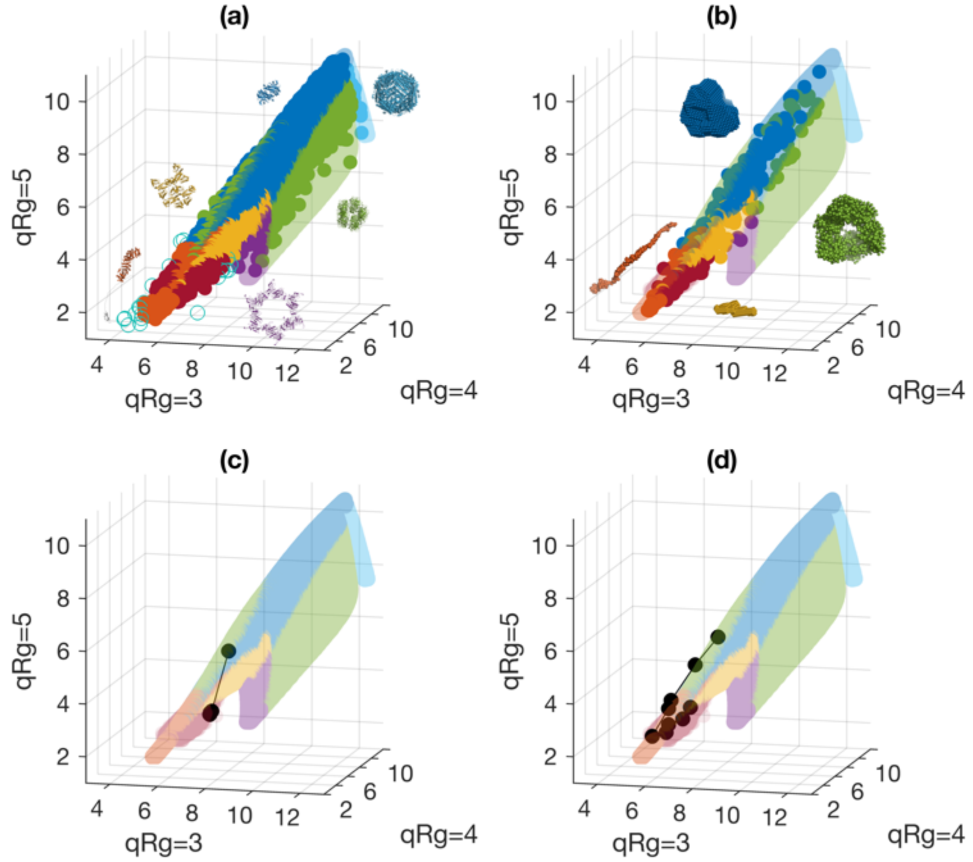
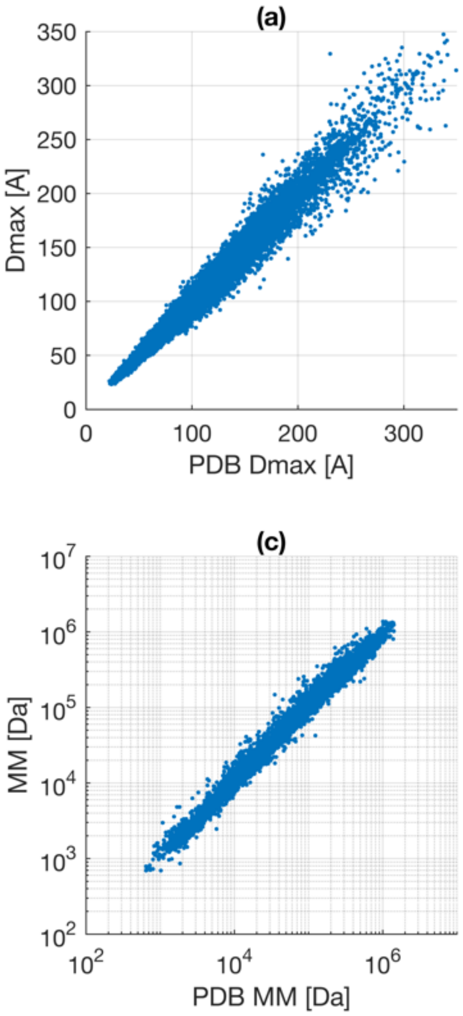


Figure 2: Distribution of (a) atomic structures of the PDB and (b) experimental scattering data from SASBDB (opaque) indicating a good agreement of the *V*'-space mapped out by shapes (transparent) and that covered by atomic structures and experimental data. The open circles in (a) depict classifications with an 'unknown' class label; structures and models displayed in panels (a) and (b) were randomly chosen and placed for the purpose of illustration (PDB: 12as (compact), 1v18 (extended), 3oei (flat), 3h3w (ring), 4avt (compact hollow), 3a68 (hollow sphere), and 2kzw (unknown); SASBDB: SASDA52 (compact), SASDA57 (extended), SASDAY4 (flat), SASDBD7 (compact hollow)). Panels (c) and (d) show the locations of experimental data of chemically unfolded ribonuclease A and lipase B, respectively. The *V*'-space trace for ribonucleaseA shows the position of the native, folded protein (compact) compared to the chemically unfolded final state (random/flexible). The trace for Lipase B shows the effect of systematically unfolding the protein population through a denaturation gradient of guanidine hydrochloride, from compact to extended until a random-chain conformation is reached (see Supplementary Methods for details). Color assignments are identical to those of Figure 1.



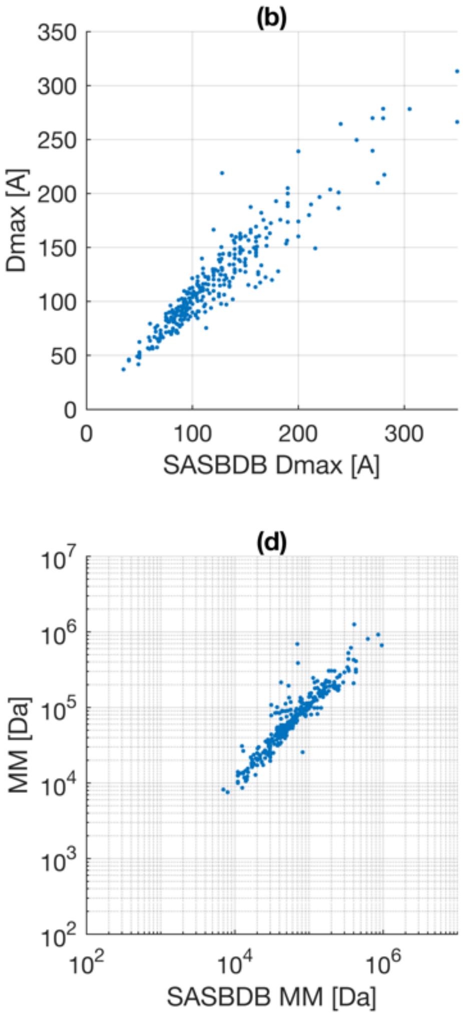


Figure 3: Estimates of D_{max} (a, b) and *MM* (c, d) for entries of PDB (a, c) and SASBDB (b, d). In the case of the PDB the expected values are known and a good agreement can be observed, in about 90% of the cases the estimate is within 10% of the expected value (a,c). No such claim can be made in the case of SASBDB as the expected values obtained depend on the type of the experiment, the sample quality, and the data analysis of the submitter.

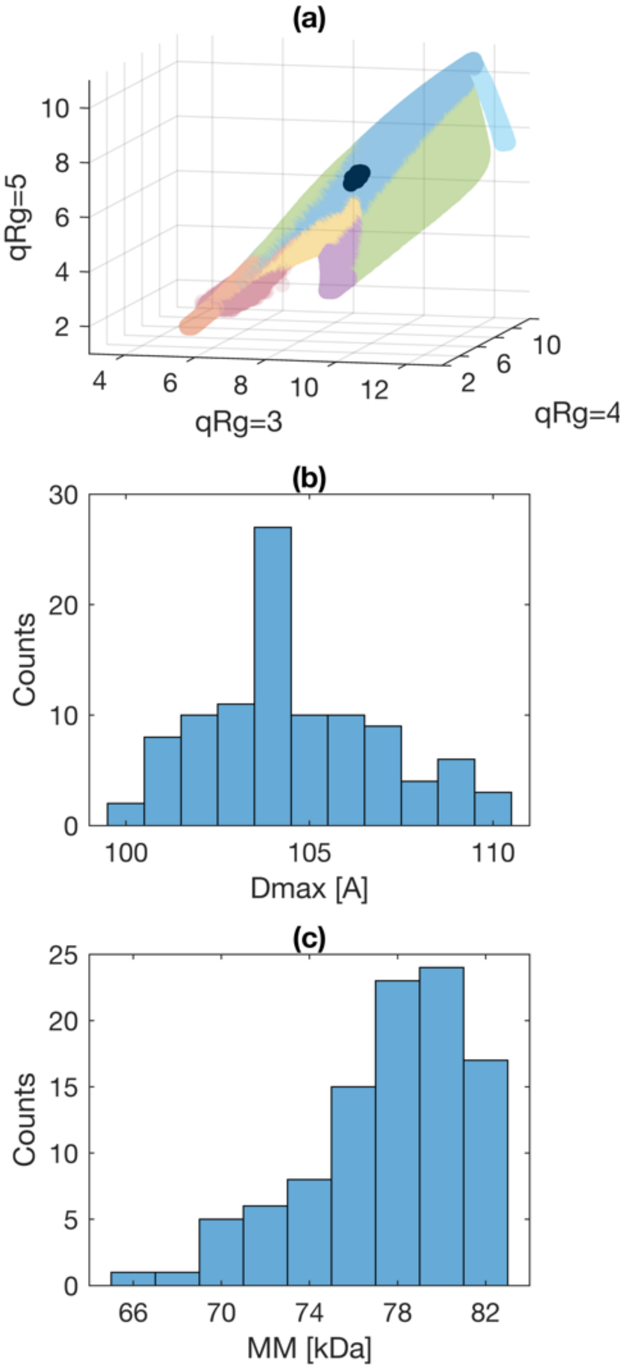


Figure 4: Locations of shape classification in V'-space (a) and histograms of structural parameters (b,c) of 100 repetitions of bovine serum albumin which are identical up to noise. Although affected by the experimental noise, all frames map closely together in V' space (a), the estimates of D_{max} vary from 100 to 110A (b) and *MM* from 66 kDa to 82 kDa (c).