Biophysical *Journal*
Computational Tool

Biophysical Society

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

Daniel Franke,[1,*] Cy M. Jeffries,[1] and Dmitri I. Svergun[1]
[1]European Molecular Biology Laboratory, Hamburg, Germany

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 neighbor approach is able to retrieve information on overall particle shape and maximal diameter (D_{max}) as well as molecular mass directly from experimental scattering data. Based on this transformation, we develop a rapid multiclass 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 into the classification space and, in a second step, use this mapping as a data source to obtain accurate estimates for the structural parameters (D_{max}, molecular mass) 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, and Windows and free for academic use.

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, nuclear magnetic resonance spectroscopy, and electron microscopy. SAXS does not require crystals, labeling, or isolated particles at cryogenic temperatures, and its applications extend to the determination of structural parameters, e.g., the radius of gyration (R_g), maximal extend (D_{max}), and the molecular mass (MM), obtaining the low-resolution shapes of macromolecules and rigid body modeling of complexes, quantitative characterization of flexibility, and time-resolved conformational changes (1). The scattering intensity $I(q)$ is recorded as a function of the scattering vector q, with the momentum transfer $q = 4\pi \sin\theta/\lambda$, where θ corresponds to half of the angle between incoming and scattered photons, and λ corresponds to the wavelength. 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 should be noted 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 (5), the Electron Microscopy Data Bank (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

Submitted December 19, 2017, and accepted for publication April 12, 2018.
*Correspondence: franke@embl-hamburg.de
Editor: Jill Trewhella.
https://doi.org/10.1016/j.bpj.2018.04.018

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 $I(0)$, and 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 compact to fully extended chains, whereas 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. To predict properties of an unknown entity, one may look up its closest neighbor(s) in the feature space and apply known properties of the neighbor 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 neighbors 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 modeling in the case of the PDB (4), and structural parameters such as D_{max} and *MM* of the immediate neighbors as discussed in this work—but also other parameters of interest—may be looked up and used as a starting point for further analysis and refinement.

Materials and 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 (Fig. 1 *a*). The corresponding dimensions of the geometric bodies, i.e., inner and outer radius,

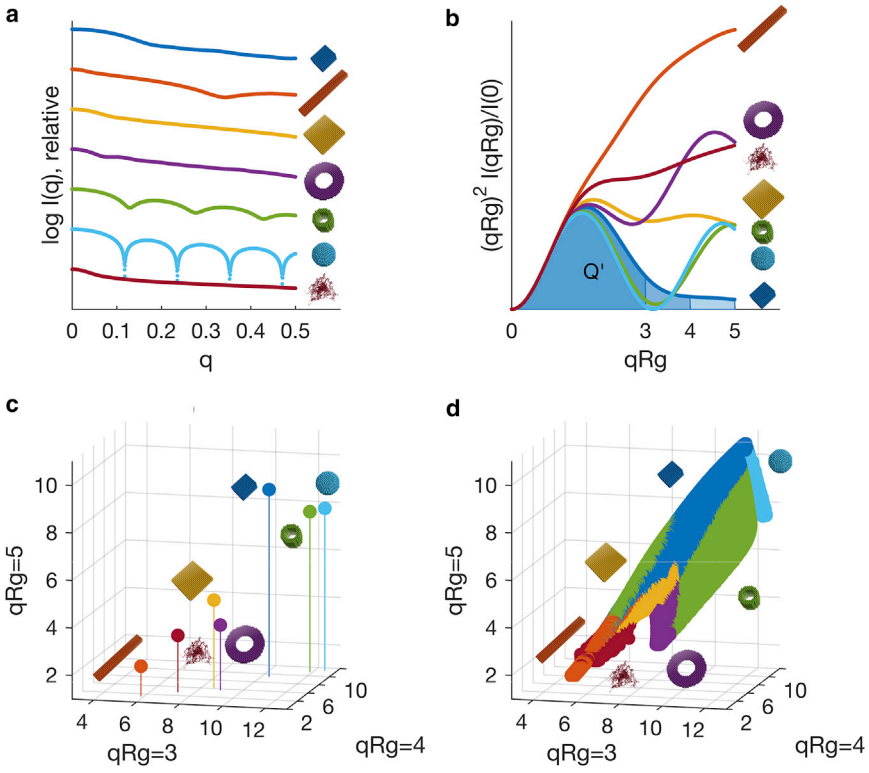


FIGURE 1 Transformation of scattering patterns of geometric objects and random-chain on arbitrary log scale (*a*) via integration of the normalized Kratky Plot (*b*) to *V'*-space (*c* and *d*). (*a–c*) depict a randomly selected member of each object class, whereas (*d*) shows the locations of all 488,000 scattering patterns generated. The 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.

height, length, and width, etc., were uniformly and independently sampled in ranges from 10 to 500 Å, 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, and in addition an 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. Although clearly limited, a selection of body types enumerating an exhaustive list of geometrical body shapes would be, at least, very difficult to obtain, especially considering the lack of analytical form factors. As shown later in the text, classification with *k*-nearest neighbors extends somewhat outside the boundaries of the mapped class volumes, thus smoothing out any gaps between geometric objects (Fig. 1 *d*). Further, to allow the identification of intrinsically disordered proteins, we employed Ensemble Optimization Method (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 a training data set for machine-learning classification that encompass basic geometric objects and disordered polymer chains (Fig. 1 *d*).

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 10 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):

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

After this, the normalized 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.,

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

Each scattering pattern was therefore reduced to three features and its associated class label (Fig. 1, *b* and *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 (data not shown). That said, with the selection presented here, the corresponding three-dimensional scatter plot of the simulated data shows a V'-space with good separation of the different shape classes (Fig. 1, *c* and *d*).

Learning, prediction, validation. As Fig. 1 *d* 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 before learning. This helped to facilitate compactness of the resulting predictions; otherwise, a query point outside this well-defined V' would still have far-away neighbors and would thus be grouped to a class it does not belong. It should be noted that this random point cloud is not shown in Fig. 1 *d*, 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-neighbors in the three-dimensional V'-space are determined by *k*-d-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 facilitate a majority vote classification in which classes overlap. The classes of the neighbors are then weighted by empirical class weights (Table S3), and the class with the maximal 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 Matthews 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, and 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 posttranslational additions were 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 with CRYSOL (8) using 30 spherical harmonics and 1001 equidistant points up to a q_{max} of 0.6 $Å^{-1}$. 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 neighbors (here $k = 5$) in a four-dimensional space combining the three dimensions of V' along with R_g are determined by k-d-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 the weighted mean of D_{max} and MM of the neighbors, where the weights correspond to the normalized inverse Euclidean distance to the unknown entity—i.e., the closer the neighbor, 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 the 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 European Molecular Biology Laboratory 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., circular dichroism spectropolarimetry, and tryptophan fluorescence spectroscopy. See Supporting Materials and Methods 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 were 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 BSA; buffers, samples, and subtracted data were deposited).

Results

Shape classification

Appropriate evaluation of multiclass classification systems is itself a topic of ongoing research. In this work, we follow the recommendations of Powers (16) and report the 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, and 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, Fig. S3 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. The overall accuracy of classification across all shapes is reported as 96.5%.

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 (Fig. 2 *a*). 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, whereas, for example, more extended proteins are much less represented (∼3%). A different picture arises from experimental data deposited in SASBDB (Fig. 2 *b*). Here, the distribution (Table S4) tends more toward 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, Fig. 2, *c* and *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 show the "denaturation trace" through V'-space as the protein populations unfold at ever-increasing concentrations of guanidine hydrochloride.

TABLE 1 F1 Score and MCC for *k*-Nearest Neighbors Multiclass Classification Results of the Individual Shape Categories

	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

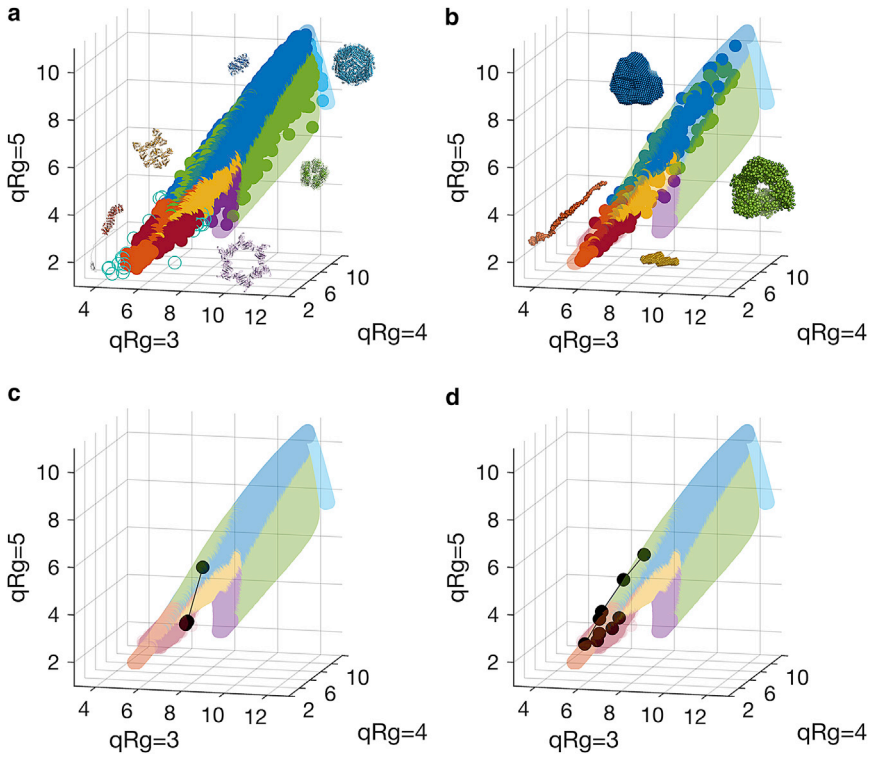


FIGURE 2 Distribution of (*a*) atomic structures of the PDB and (*b*) experimental scattering data from SASBDB (opaque) indicating a good agreement of the *V'*-space mapped out by shapes (transparent) and that covered by atomic structures and experimental data. The open circles in (*a*) depict classifications with an "unknown" class label; structures and models displayed in (*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*), and SASDBD7 (*compact hollow*)). (*c* and *d*) show the locations of experimental data of chemically unfolded ribonuclease A and lipase B, respectively. The *V'*-space trace for ribonuclease A 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 Supporting Materials and Methods for details). Color assignments are identical to those of Fig. 1.

Prediction of structural parameters

Fig. 3, *a* and *c* summarize the results of the leave-one-out cross-validation for the 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 ~90% of the cases, the estimate is within 10% of the true value. The evaluation of experimental data as deposited in SASBDB (Fig. 3, *b* and *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 (Fig. 3 *b*), which may, for example, be explained by the influence of the hydration shell.

Effects of experimental noise

Fig. 4 elucidates the effect of experimental noise on 100 repetitions of BSA; all frames were found similar to each other up to noise as per CorMap test (19). As depicted in Fig. 4 *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 Fig. 4, *b* and *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 Å and ~67 kDa, 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 Fig. 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 (21–23), 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

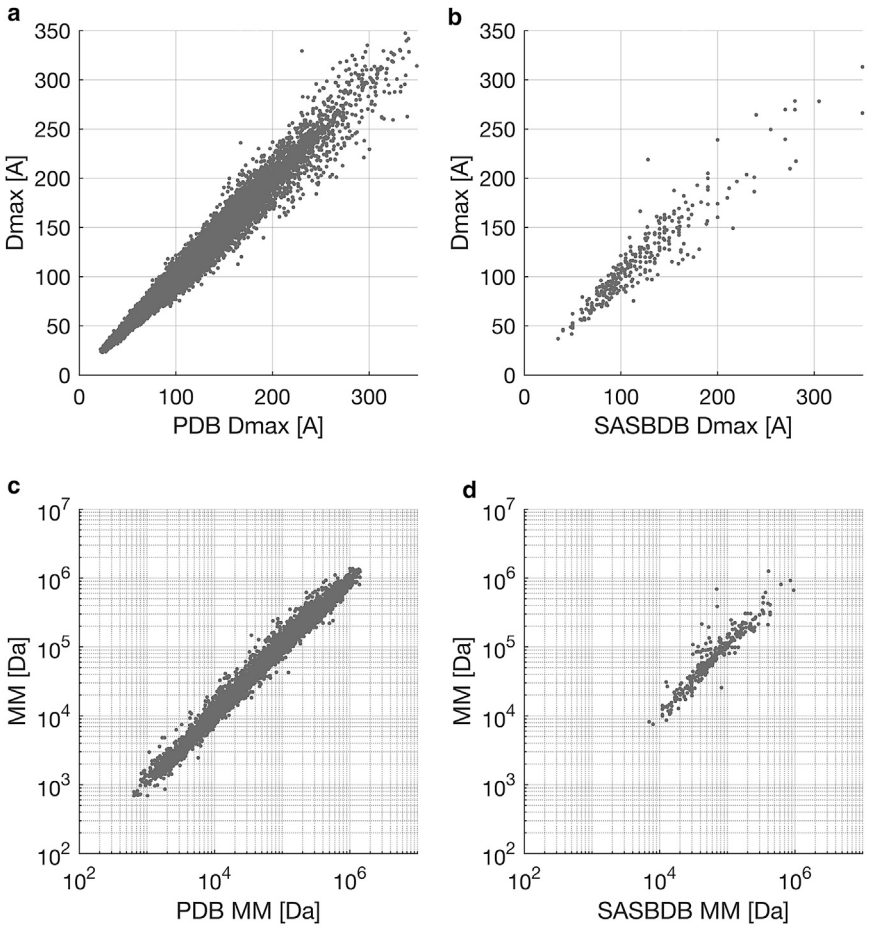


FIGURE 3 Estimates of D_{max} (*a* and *b*) and *MM* (*c* and *d*) for entries of PDB (*a* and *c*) and SASBDB (*b* and *d*). In the case of the PDB, the expected values are known, and a good agreement can be observed; in ~90% of the cases, the estimate is within 10% of the expected value (*a* and *c*). 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.

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 (Fig. 3 *b*).

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 (Fig. 3 *b*) 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 should be noted that some details of the presented method were empirically determined, e.g., the qR_g integration limits for V'; although the general magnitude is appropriate, e.g., on the lower end, integration to $qR_g = 1$ corresponds to the Guinier range, and on a 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 before 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 neighbors method include the following: 1) selection of k and the applied distance weights; 2) arbitrary linear and nonlinear data scaling and transformation before learning; 3) metric selection and metric learning (30); and, of course, 4) any other learning method such as regression functions, support vector machines, neural networks, deep learning, etc. As in this manuscript we focus on outlining and introducing, to our knowledge, 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).

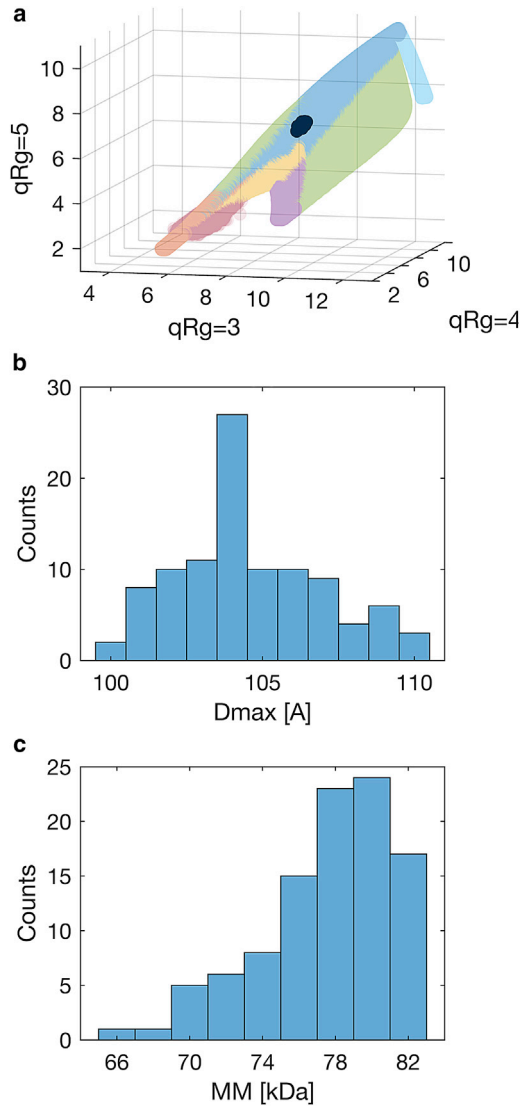


FIGURE 4 Locations of shape classification in *V'*-space (*a*) and histograms of structural parameters (*b* and *c*) of 100 repetitions of BSA that are identical up to noise. Although affected by the experimental noise, all frames map closely together in *V'*-space (*a*); the estimates of D_{max} vary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

Conclusions

In this manuscript, we present what is, to our knowledge, 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 data set. 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 because of 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 should be noted 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).

SUPPORTING MATERIAL

Supporting Materials and Methods, three figures, and four tables are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(18)30464-8.

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

This work was supported by iNEXT (grant number 653706, funded by the Horizon 2020 program 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, …, R. P. May. 2013. Small Angle X-Ray and Neutron Scattering from Solutions of Biological Macromolecules. Oxford University Press, Oxford, UK.
2. Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. 1996. From data mining to knowledge discovery in databases. *AI Mag.* 17:37–54.
3. Petoukhov, M. V., and D. I. Svergun. 2015. Ambiguity assessment of small-angle scattering curves from monodisperse systems. *Acta Crystallogr. D Biol. Crystallogr.* 71:1051–1058.
4. Berman, H. M., J. Westbrook, …, P. E. Bourne. 2000. The protein data bank. *Nucleic Acids Res.* 28:235–242.
5. Ulrich, E. L., H. Akutsu, …, J. L. Markley. 2008. BioMagResBank. *Nucleic Acids Res.* 36:D402–D408.
6. Lawson, C. L., M. L. Baker, …, W. Chiu. 2011. EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res.* 39:D456–D464.

7. Valentini, E., A. G. Kikhney, …, D. I. Svergun. 2015. SASBDB, a repository for biological small-angle scattering data. *Nucleic Acids Res.* 43:D357–D363.
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. *Nucleic Acids Res.* 38:W540–W544.
10. Bellman, R. E. 1957. Dynamic Programming. Princeton University Press, Princeton, NJ.
11. Konarev, P. V., M. V. Petoukhov, …, 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. Mertens, …, D. I. Svergun. 2015. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ.* 2:207–217.
13. Durand, D., C. Vivès, …, F. Fieschi. 2010. NADPH oxidase activator p67(phox) behaves in solution as a multidomain protein with semi-flexible linkers. *J. Struct. Biol.* 169:45–53.
14. Fischer, H., M. de Oliveira Neto, …, 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. *Commun. ACM.* 18:509–517.
16. Powers, D. M. W. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *J. Mach. Learn. Technol.* 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. E., A. Spilotros, …, D. I. Svergun. 2015. Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY). *J. Appl. Cryst.* 48:431–443.
19. Franke, D., C. M. Jeffries, and D. I. Svergun. 2015. Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra. *Nat. Methods.* 12:419–422.
20. Jeffries, C. M., M. A. Graewert, …, D. I. Svergun. 2016. Preparing monodisperse macromolecular samples for successful biological small-angle X-ray and neutron-scattering experiments. *Nat. Protoc.* 11:2122–2153.
21. Brennich, M., J. Kieffer, …, A. Round. 2016. Online data analysis at the 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. Phys. Res. Sec. A.* 689:52–59.
23. Shkumatov, A. V., and S. V. Strelkov. 2015. DATASW, a tool for HPLC-SAXS data analysis. *Acta Crystallogr. D Biol. Crystallogr.* 71:1347–1350.
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. O. Glatter, and O. Kratky, eds 1982. Small-Angle X-ray Scattering. Academic Press, London, UK.
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 of mass, models and resolution by small-angle scattering. *Nature.* 496:477–481.
29. Hajizadeh, N. R., D. Franke, …, D. I. Svergun. 2018. Consensus Bayesian assessment of protein molecular mass from solution X-ray scattering data. *Sci Rep* 10.1038/s41598-018-25355-2.
30. Xing, E. P., A. Y. Ng, …, S. Russel. 2003. Distance metric learning, with application to clustering with side-information. *Adv. Neural Inf. Process. Syst.* 15:505–512.
31. Franke, D., M. V. Petoukhov, …, 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.
32. Gasteiger, E., C. Hoogland, …, A. Bairoch. 2005. The Proteomics Protocols Handbook. Humana Press, New York.
33. Micsonai, A., F. Wien, …, J. Kardos. 2015. Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proc. Natl. Acad. Sci. USA.* 112:E3095–E3103.
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.

Biophysical Journal, Volume *114*

Supplemental Information

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

Daniel Franke, Cy M. Jeffries, and Dmitri I. Svergun

Supplementary Materials and Methods

Preparation of native ribonuclease A (RNAse) and carboxyamidomethylated ribonuclease A (cam-RNAse).

Lyophilised bovine pancreatic ribonuclease A (Sigma) was resuspended in phosphate buffered saline, pH 7.0, (PBS) and dialysed overnight at 4 °C against the same buffer to obtain a sample of natively-folded RNAse. The final sample concentration was 5.24 mg/ml (determined at $Abs_{280\ nm}$ using an $E_{0.1\%}$ = 0.71 ml/mg calculated from the amino acid sequence (1). The post dialysis buffer was used as an exact solvent blank for the SAXS measurements.

The preparation of disulfide-reduced and carboyxamidomethylated RNAse (*cam*-RNAse) followed the procedure as described by Wang, Trewhella, & Goldenberg (2008). Briefly, lyophilised RNASe powder (approximately 8–10 mg) was dissolved in 1 ml of 6 M guanidine hydrochloride (Gdn.HCl), 10 mM ethylenediaminetetraacetic acid (EDTA), 10 mM dithiothreitol (DTT) and 100 mM Tris, with a final (combined) pH of 8.0. The solution was incubated for 1.5 hr at room temperature with gentle mixing to effect protein unfolding and disulphide reduction. At the completion of the high-pH denaturation step, fresh iodoacetamide (180 mM stock in H_2O) was added to a final concentration of 30 mM and the system left for approximately 45 min to effect sulfhydryl and histidine alkylation. Concentrated HCl (1 M in H_2O) was then added to the RNAse with rapid mixing to a final concentration of 100 mM. The protein solution was dialysed overnight at 4 °C against 10 mM HCl in water. To two individual aliquots of post-dialysis *cam*-RNAse were removed and a stock solution of 8 M urea in 10 mM glycine (combined pH = 2.5) was added to final concentrations of 1 or 2 M, respectively. The protein concentrations of the *cam*-RNAse samples were: 5.97 mg/ml (10 mM HCl), 5.27 mg/ml (10 mM HCl, 1 M urea) and 4.58 mg/ml (10 mM HCl, 2 M urea). For the SAXS measurements, the post-dialysis 10 mM HCl solution was used as an exact solvent blank, with the addition of an equivalent mass of 8 M urea solution (+/- 1 mg) as used for the 1 M and 2 M urea *cam*-RNAse samples. The corresponding SAXS data of folded and unfolded RNAse can be located in the SASBDB entries SASDDL3 and SASDDM3, respectively.

Preparation of Candida antarctica lipase B.

A solution of *Candida antarctica* lipase B (Hampton Research) was dialysed overnight at room temperature against 100 mM NaCl, 20 mM Na_2HPO_4, containing 0, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5 or 6 M Gdn.HCl (combined pH = 6.0, adjusted using HCl/NaOH). In all instances, the respective post-dialysis buffer was used as the solvent blank for the SAXS measurements. A set of lipase B samples were also prepared under reducing conditions. DTT (1 M stock in H_2O) was added to each lipase B sample/solvent blank to a final concentration of 10 mM immediately prior to SAXS. The final concentrations of the lipase B samples were assessed using an $Abs_{280\ nm}$ $E_{0.1\%}$ = 1.239 ml/mg (1) and are summarised in Table 2. The SAXS data of both Lipase B both with and without DTT and the Gdn.HCl unfolding series can be located in the SASBDB entries SASDDJ3 and SASDDK3.

Circular dichroism spectropolarimetry: native RNAse and unfolded *cam*-RNAse.

Circular dichroism (CD) measurements were performed at room temperature using a Chirascan (Applied Photophysics) spectropolarimeter with a quartz cell pathlength (l) of 1 mm. The RNAse samples used for SAXS (described above) were diluted 50-fold in their respective supporting solvents. The approximate protein concentrations, C, in mg/ml used for the CD measurements are reported in Table 2.

The CD spectra were acquired across 175–280 nm using a time constant of 0.5 s at 1 nm wavelength intervals (1 nm bandwidth). The presented data (Supp. Fig. 1) represent the solvent-subtracted average of these scans for each sample, quoted as mean residue ellipticity, θ in deg.cm^2/dmol versus wavelength, λ in nm (where the molecular weight, MW, of RNAse = 13690 Da, and the number of amino acids, N= 124). The conversion from machine units (mdeg) to θ followed:

$$\theta = (\text{mdeg}*\text{MW})/(Nl\text{C}).$$

Those data with unduly high absorbance at low wavelength were discarded to produce the final spectra for:

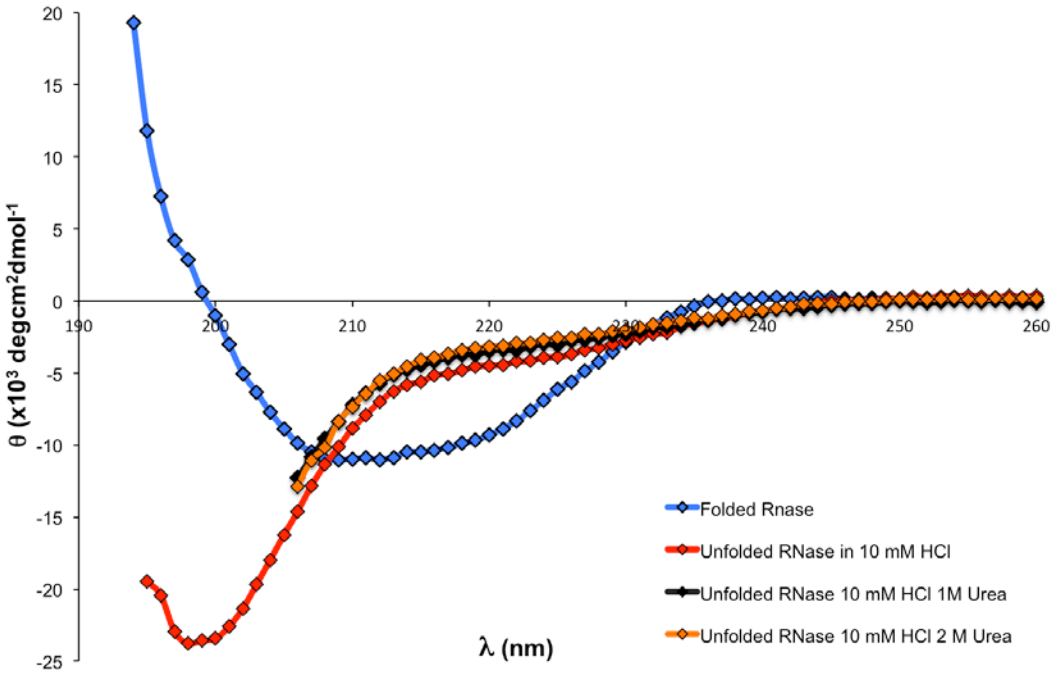
Native RNAse; 197–260 nm,
cam-RNAse in 10 mM HCl; 197–260 nm,
cam-RNAse in 10 mM HCl, 1 M urea; 205–260 nm,
cam-RNAse in 10 mM HCl, 2 M urea; 206–260 nm.

Secondary structure analysis was performed using the online BeStSel single spectrum analysis and fold recognition server, http://bestsel.elte.hu/ (2). Spectra were converted into absorption units, i.e., as the differential molar extinction coefficient, $\Delta\varepsilon$ ($M^{-1}.cm^{-1}$) vs λ, (where $\Delta\varepsilon = \theta/3298.2$) and analysed for secondary structure content using the 200–250 nm option of the BeStSel server. Only those spectra for native RNAse and *cam*-RNAse in 10 mM HCl access sufficiently low wavelengths for secondary structure analysis (Table 1), and consequently the secondary structure content of the remaining *cam*-RNASe samples were not assessed. The experimental results were compared to the secondary structure content extracted from the X-ray crystal structure of RNAse A (PDB: 3MZQ) and that reported for RNAse in the Protein Circular Dichroism Data Bank (http://pcddb.cryst.bbk.ac.uk/home.php, PCDDBID: CD0000063000).

Tryptophan fluorescence spectroscopy: native and unfolded lipase B.

Intrinsic tryptophan fluorescence spectroscopy measurements from lipase B and denatured lipase B in Gdn.HCl or Gdn.HCL plus 10 mM DTT (Table 2) were performed using a Tecan Infinite M1000 spectrometer. Scans were performed at 25 °C using an excitation wavelength of 295 nm, with the emission spectra recorded from 310–600 nm using an emission wavelength step size of 1 nm (flash frequency, 400 Hz; 50 flashes per nm). The fluorescence yields were normalised to protein concentration and the wavelength corresponding to the maximum flourescence yield for each scan was recorded to qualitatively assess red-shifts in the tryptophan emission spectra (Supp. Fig. 2).

Supplementary Figures and Tables



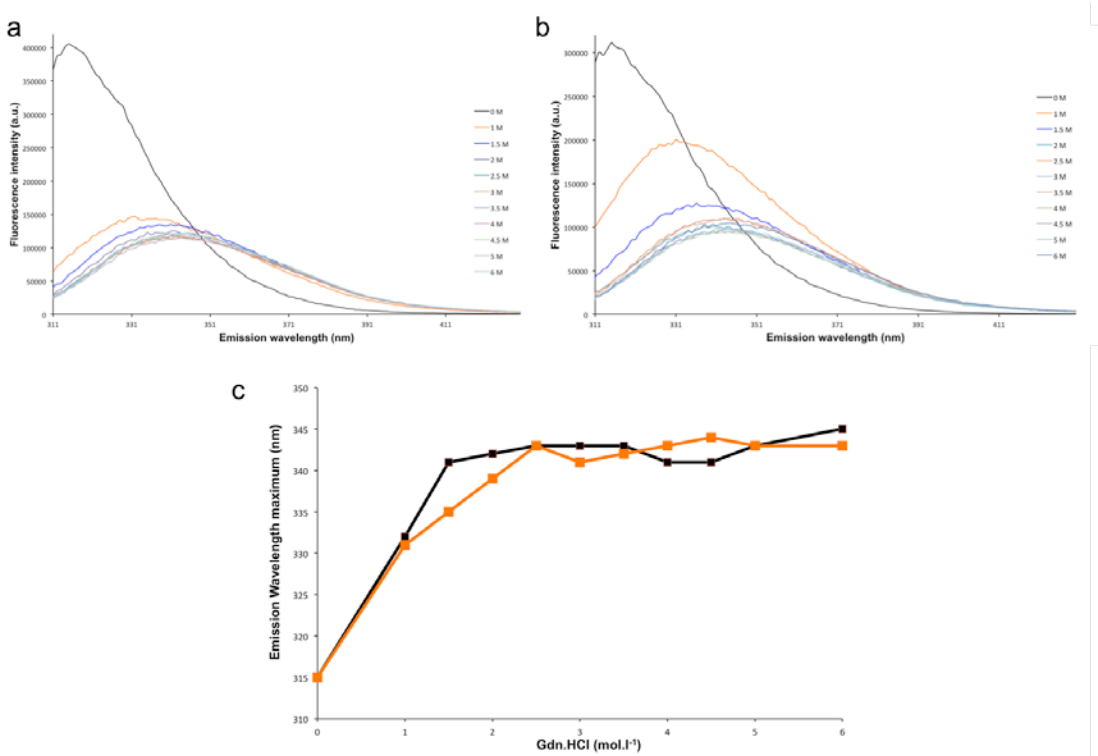
Supplementary Figure 1. CD spectra of native folded RNase (blue), unfolded *cam*-RNase (red) and *cam*-RNase in the presence of 1 M or 2M urea (black and orange, respectively)

RNAse A *CD Samples*	Protein Concentration (mg/ml)	*RNAse A* *% Secondary structure* α-helix	β	turn+other
Native RNAse	0.105	20.1	34.7	45.2
cam-RNAse 10 mM HCl	0.119	0	23.9	76
cam-RNAse 10 mM HCl, 1 M Urea	0.105	-	-	-
cam-RNAse 10 mM HCl, 2 M Urea	0.092	-	-	-
	PDB: 3MZQ	21.0	33.1	46.0
	PCDDBID: CD0000063000	20.9	33.1	45.9

Table 1: Sample protein concentrations and secondary structure analysis derived from CD measurements of native and *cam*-RNAse. Included is a comparison with secondary structure content extracted from the X-ray crystal structure of RNAse (PDB 3MZQ) and from CD spectra deposited in the Circular Dichroism Data Bank (CD0000063000).

Lipase B SAXS Samples		Lipase B fluorescence spectroscopy samples		
Gdn.HCl concentration (M)	Protein Concentration (mg/ml)	Gdn.HCl concentration (M)	Protein Concentration (mg/ml)	Protein Concentration (mg/ml), plus DTT
0	4.65	0	0.48	0.5
1	4.66	1	0.43	0.38
1.5	4.59	1.5	0.37	0.42
2	4.52	2	0.4	0.38
2.5	4.48	2.5	0.37	0.41
3	4.19	3	0.39	0.37
3.5	4.06	3.5	0.39	0.44
4	4.09	4	0.38	0.39
4.5	4.12	4.5	0.41	0.37
5	4.02	5	0.38	0.39
6	4.19	6	0.41	0.38

Table 2: Concentration of Gdn.HCl and Lipase B used for SAXS and fluorescence spectroscopy measurements. Note: the SAXS samples for lipase B under reducing conditions were prepared by adding 1 μl of 1 M DTT to 99 μl of protein. Therefore, within pipetting and spectrophotometric error, it is expected that the reduced lipase B sample concentrations will not differ significantly from the concentrations quoted here for the lipase B SAXS samples in Gdn.HCl.



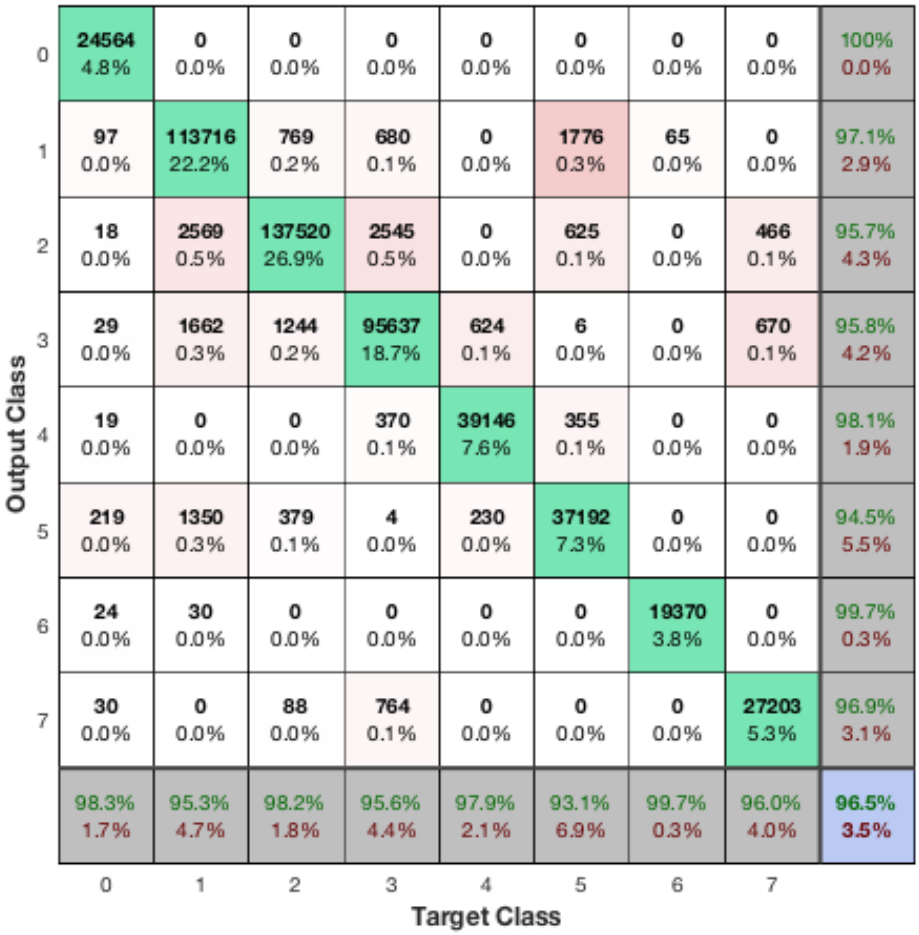
Supplementary Figure 2. a. Tryptophan fluorescence intensities vs emission wavelengths through a Gdn.HCl concentration gradient (0-6 M) for Lipase B with no DTT present in solution. b. With additional 10 mM DTT added to solution. c. The shift in emission wavelength maximum of Lipase B as a function of Gdn.HCl concentration.

Preparation and SAXS data of bovine serum albumin.

Lyophilised bovine serum albumin (Sigma: # 05470) was dissolved in 50 mM HEPES, pH 7.5, and 0.22 micron pore spin-filtered. The final sample concentration was 2.25 mg/ml evaluated at $Abs_{280\ nm}$ using an $E_{0.1\%}$ = 0.646 ml/mg (Gasteiger, et al., 2005). An aliquot of 0.22 micron filtered HEPES buffer was used as the solvent blank for the SAXS measurements. The subsequent SAXS data and collection parameters of both un-subtracted and subtracted SAXS data frames can be found in SASBDB entry SASDBK3.

Class Label	Class Weight
Unknown	1
Compact	2
Extended	2
Flat	2
Ring	2
Compact-hollow	4
Hollow-sphere	2
Random-chain	2

Table 3: Empirical class weights for k-nearest-neighbour shape classification.



Supplementary Figure 3: Leave-One-Out cross validation results for shape classification with recall and precision percentages in the margins. Class labels are (0) unknown, (1) compact, (2) extended, (3) flat, (4) ring, (5) compact-hollow, (6) hollow-sphere, (7) random-chain.

Class Label	PDB		SASBDB	
Unknown	25	0. 02 %	2	0.05 %
Compact	122.913	74.05 %	149	37.16 %
Extended	5.382	3.24 %	36	8.98 %
Flat	9.734	5.86 %	119	29.68 %
Ring	154	0.09 %	3	0.08 %
Compact hollow	26.909	16.21 %	25	6.23 %
Hollow sphere	125	0.08 %	0	0.00 %
Random Chain	740	0.45 %	67	16.71 %
Total	165.982	100.00 %	401	100.00 %

Table 4: Absolute and relative shape counts as depicted in main Figure 2(a) and 2(b).

References

1. Gasteiger, E., C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel and A. Bairoch. 2005. The Proteomics Protocols Handbook Humana Press.
2. 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.
3. 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.