BNL-229267-2025-JAAM

Automated SAXS analysis for structural discovery in biologics and polymeric nanoparticles

C. Ramirez, J. Byrnes

To be published in "Biophysical Journal"

November 2025

Photon Sciences

Brookhaven National Laboratory

U.S. Department of Energy

USDOE Office of Science (SC), Basic Energy Sciences (BES). Scientific User Facilities (SUF)

DISCLAIMER

Automated SAXS Analysis for Structural Discovery in Biologics and Polymeric Nanoparticles

Cesar Ramirez[1], Elena Di Mare[1], James Byrnes[2], Eman Ahmed[1], Maria Pineiro-Goncalves[1], Cristian Lopez[1], N. Sanjeeva Murthy[3,†], Adam J. Gormley[1,†]

[1] Rutgers, The State University of New Jersey, Department of Biomedical Engineering, Piscataway, NJ 08854, USA

[2] Energy & Photon Sciences Directorate, National Synchrotron Light Source II, Brookhaven National Laboratory, Upton, NY, 11973, USA

[3] Rutgers, The State University of New Jersey, Laboratory for Biomaterials Research, Department of Chemistry and Chemical Biology, Piscataway, NJ 08854, USA

[†]Correspondence to:

Adam J. Gormley: adam.gormley@rutgers.edu

N. S. Murthy: nsmurthy@chem.rutgers.edu

Keywords: high-throughput analysis, machine learning, intrinsically disordered protein

Abbreviations: AIC: Akaike information criterion, BIC: Bayesian information criterion, BIFT: Bayesian indirect Fourier transform, BSA: Bovine serum albumin, D_{max}: Maximum intra-particle distance, DP: Degree of polymerization, GPA: Guinier peak analysis, GMM: Gaussian mixture model, IDP: Intrinsically disordered protein, MAE: Mean absolute error, mAb: Monoclonal antibody, ML: Machine learning, MLP: Multi-layer perceptron, PDDF: Pair distance distribution function, R_g: Radius of gyration, SASBDB: Small angle scattering biological data bank, SAXS: Small-angle X-ray scattering, SCNP: Single-chain polymer nanoparticle.

Abstract:

Small-angle X-ray scattering (SAXS) is a powerful technique for assessing macromolecular structure. High-throughput SAXS is limited by the time-consuming and, at times, subjective nature of SAXS data interpretation. We present SAXS Assistant, a Python-based script that streamlines SAXS data analysis to extract features for machine learning (ML) and key structural parameters, including the Guinier radius of gyration (R_g), pair distance distribution function (PDDF)-derived R_g, maximum particle dimension (D_{max}), and Kratky plots. The script builds upon BioXTAS RAW, and validates reliability via Guinier/PDDF R_g agreement, an important indicator of well-measured datasets. For assistance in D_{max} estimation, a multi-layer perceptron (MLP) regressor was trained with 1,940 data files from the small angle scattering biological data bank (SASBDB). The model achieved a test set performance R^2 = 0.90 and mean absolute error (MAE) = 11.7 Å. Training exclusively with experimental data translates analyses from researchers, including experts in the field, to the ML model, which helps assess D_{max} estimations from PDDF. Gaussian mixture model (GMM) clustering was implemented to classify profiles into structural classes based on entries in the SASBDB. Users may therefore assess the similarity between experimental samples and known biomolecular shapes within the mapped repository entries. This probabilistic clustering aids in quantifying information from Kratky and generating shape-descriptive features. SAXS Assistant

accelerates SAXS data analysis through enforced quality control, ML-ready outputs, and flags for low-confidence results. In addition to providing the ability to analyze large datasets at high-throughput, this tool is versatile and may serve researchers in both biological and synthetic polymer research fields.

Statement of Significance

Interpreting small-angle X-ray scattering (SAXS) data is often time-consuming and subjective. SAXS Assistant helps streamline this process, automating structural parameter extraction and integrating machine learning-derived insights. Models were developed using 3,322 researcher-evaluated SASBDB submissions, incorporating features previously shown to capture macromolecular shape. The tool enhances Kratky-based interpretation through unsupervised clustering, offering probabilistic insights into conformational composition. Moreover, it aids in assessing the quality of the pair-distance distribution function by providing a machine learning–predicted maximum dimension. SAXS Assistant is open-source, pip-installable, and accessible to users with minimal coding experience. By the generation of summary plots and structured outputs, it supports expert validation while broadening access to SAXS analysis for biological and synthetic macromolecules.

1. Introduction

Small-angle X-ray scattering (SAXS) is a technique often used to obtain basic structural information of macromolecules in solution, such as the radius of gyration (R_g), flexibility, and folding behavior (1-5). A typical SAXS dataset is obtained by subtracting the scattering profile of the buffer from that of the biomolecule in solution (6). The resulting profile contains intensity

values *I(q)* as a function of the scattering vector *q*, and error in the measured intensities presented as the standard deviation *σ(q)* (2,6,7). High-throughput SAXS experiments can produce hundreds of scattering profiles, substantially increasing the time and effort required for data analysis (1). In addition, SAXS data quality must be assessed prior to analysis to prevent fallacious structural parameter determination (1,2,5,8). While collecting scattering profiles at high-throughput is facile, accurate high-speed analysis of large datasets remains a major challenge.

Researchers continue to develop tools and scripts to streamline SAXS analysis workflows and to meet user-specific needs that may build on available pre-existing tools (3,7,9). There are a variety of software tools available for data analysis, reduction, and modeling (3,10,11). These include BioXTAS RAW, SASFIT, ScÅtter, and SasView (3,7-10). Among these, BioXTAS RAW is a popular platform due to its approachability and diverse functionalities, including shape reconstruction, buffer subtraction, and data reduction (3). ATSAS remains the most widely used analysis suite including tools for data preprocessing as well as advanced modeling (3,10,11). Despite advancements in analysis tools, aspects of SAXS data analysis still rely on user judgment. For example, the selection of the Guinier region is a largely subjective process, often determined through iterative calculations of the R_g and an assessment of fit quality metrics such as R^2 and residuals (1). Algorithms like *Auto Guinier* from BioXTAS RAW largely address this by automating region selection (3,10,11). Likewise, D_{max} selection through indirect Fourier transform (IFT) provides some metrics like the quality of the fit, but still requires additional considerations based on the characteristic appearance of the curve obtained by IFT such as smoothness to reduce termination effects (6,12). The advent of accessible machine learning (ML) toolkits and user guides for biomaterials research has the potential to further lower the barrier to entry for high-throughput analytical techniques (13). SAXS is well-suited for ML applications, due to its

compatibility with high-throughput data generation, the availability of large, curated repositories such as the small angle scattering biological data bank (SASBDB), and tools for the generation of synthetic data (6,14-16).

ML has been previously applied to overcome the challenges of expertise requirements to properly interpret SAXS data and the need for predefined scattering models. For example, computational reverse engineering analysis for scattering experiments (CREASE) integrates genetic algorithms and molecular simulations for the interpretation of scattering data without the need for pre-defined models (17-20). CREASE has previously been applied for the study of self-assembling amphiphilic block copolymers, revealing both micelle and monomer-level structural information (17). Other ML-based approaches have also been explored, such as the SWAXS-focused method by Chen and coworkers for RNA duplexes, which illustrates the potential of ML to obtain data insights from experimental scattering (21). Here, we introduce SAXS Assistant, a Python-based script designed to streamline and largely automate SAXS data analysis and dataset generation. The script extracts key structural parameters, including Guinier R_g, D_{max}, and IFT-derived R_g, and generates diagnostic plots for user inspection (Figure 1). It builds upon open-source components from RAW and processes folders of buffer-subtracted SAXS profiles (3). As a quality control step, the script requires agreement between R_g values derived from Guinier and the IFT curve to ensure dataset reliability (1,4,12). Files that fail to meet this agreement threshold are flagged and excluded from automatic reporting, minimizing the risk of incorrect structural assignment and reducing the need for user intervention. While initially developed to support high-throughput SAXS screening of random copolymers and the discovery of compact single-chain polymer nanoparticles (SCNPs), the script is broadly applicable to biological and synthetic macromolecules alike.

To lower the subjectivity during D_{max} and IFT evaluation, an ML component was incorporated to help evaluate the results obtained from the pair distance distribution function (PDDF). The ML-guided D_{max} prediction helps users evaluate the quality of their structural estimations by translating to the model analysis patterns learned from researchers. Our interest was in incorporating researcher-level intuition from human-determined results to enhance SAXS data analysis for specialists and newcomers alike (6,16). We also integrated an unsupervised clustering framework into the script as an additional structural interpretation aid. A Gaussian mixture model (GMM) was applied to identify structural classes based on global scattering features. This type of classification is particularly useful for biological samples which may contain regions with distinct conformations, enabling quantification of heterogeneity (8). The model quantifies structural composition with respect to distinct structural domains, enhancing the qualitative understandings of Kratky analysis through quantifiable metrics. This was also extended to probe the structure of polymer SCNPs from our group's on-going research to develop synthetic protein-mimetics from random copolymers. By integrating this clustering framework with automated databasing, we enhance the interpretability, organization, and ML readiness of high-throughput SAXS workflows.

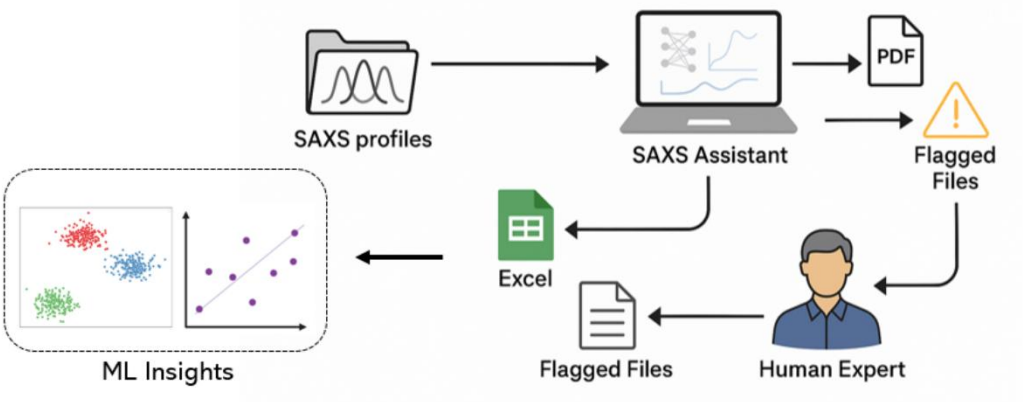

Figure 1. Schematic of SAXS Assistant. SAXS profiles are uploaded to the SAXS Assistant software, which analyzes good quality profiles, outputs key structural parameters in the form of a PDF and an Excel file, and flags problematic SAXS profiles for manual assessment. Created using

Biorender. A schematic workflow of data from input to output maybe found in the supplemental information.

2 Materials and Methods

Use of RAW Open-Source Code and Dependencies

Portions of open-source code from BioXTAS RAW were used to read profiles and the *Auto Guinier* and BIFT functions were used to obtain R_g and the PDDF, respectively (3,22). All code was written in Python. The SAXS Assistant source code and package is available through GitHub and PyPI via `pip install SAXS-Assistant`.

SAXS Assistant was developed in Python (version ≥3.7). The source code and installation package are available on GitHub and PyPI. All required dependencies are installed automatically via setup.py during pip installation. For users who wish to install packages manually, a complete list of dependencies is provided in the included requirements.txt file as stated on the project's PyPI page (https://pypi.org/project/saxs-assistant/).*Data from SASBDB*

A Python script gathered 3,577 SAXS profiles from the REST API of the SASBDB repository. Files with missing values for Guinier R_g, *I(0)*, PDDF R_g or *I(0)* were omitted and those flagged by SAXS Assistant were not used for comparison between methods or model training. The profiles from the SASBDB were truncated to $q < 0.25$ Å^{-1} and those with q reported in nm^{-1} were converted to Å^{-1}. Profiles of differing resolutions had discrepancies in the size of q-spacing increments; with a large portion being high-resolution data yielding small q-spacing increments. To allow for a closer comparison between profiles regardless of original profile resolution, we sampled each profile for q-values akin to those collected by our group, q_{min} near 0.005 Å^{-1} (23). This also speeds up analysis as PDDF calculation becomes slow with increase in data size. Data files that resulted

in errors during file reading were also omitted. A DataFrame containing Guinier and PDDF R_g along with sample identifier was created for future analyses.

SAXS Data Collection

SAXS/WAXS data were collected over the q range of 0.005-3.13 $Å^{-1}$, with a q range of 0.005-0.25 $Å^{-1}$ used for analysis. Both SAXS and WAXS data were collected simultaneously with two detectors, Pilatus 1M for SAXS and Pilatus 900K for WAXS, and 15.14 keV X-rays ($\lambda = 0.8189$ Å). Background subtraction was performed for each SAXS profile using the buffer corresponding to a given sample by scaling for the water scattering at q ~2 Å. All SAXS data collections by the group employing the high throughput (HT-SAXS) method at beamline 16-ID life science X-ray scattering (LIX), part of the National Synchrotron Light Source II (NSLS-II) at Brookhaven National Laboratory (Upton, NY) (24-26).

Guinier Peak Analysis for Data Visual Inspection

Guinier peak analysis (GPA) can confirm the existence of a Guinier region in a collected dataset, as well as inform on the quality of this region (i.e., how many points are present in the region) (4). GPA is achieved by transforming the Guinier region into a plot of $qI(q)$ versus q^2, where the presence of a peak in this region confirms the existence of a Guinier region (4). The GPA plot shows a rise in points within the Guinier region from q^2 near zero to $q^2_{max} = 1.5/R_g^2$.

Another variant of GPA is the dimensionless GPA plot, which is obtained from plotting $qR_gI(q)/I(0)$ versus (qR_g) (4). In this work, the y-axis is shown on a natural logarithm scale, consistent with the Scatter program implementation. Dimensionless GPA can be used to validate the obtained R_g and $I(0)$ values and examine the peak location (4). In the current script, GPA is plotted up to $q^2 \sim 0.0012$ $Å^2$. In the scope of this script, GPA is only used to provide visual insight into the existence of the

Guinier region and inform whether a lack of Guinier points led to the inability of the script to provide analysis for a given sample. The script plots the theoretical *I(q)* calculated from Eq. 1 alongside experimental data from the Guinier approximation to visualize the agreement (or lack thereof) between experimental and theoretical values using the dimensionless GPA plot (4,27).

$$I(q) \approx I(0) \exp\left(-\frac{q^2 R_g^2}{3}\right) \quad (1)$$

Determination of R_g

The script determines R_g through two complementary methods, a custom PDDF-informed method from the script (referred to as *PDDF-Informed* in the software) and the *Auto Guinier* function from RAW (3). *PDDF-Informed* is similar to RAW *Auto Guinier*, except that it requires agreement between PDDF and Guinier R_g by filtering out fits that do not fall within the range of 15% of R_g values calculated from the PDDF as q_{min} is varied. This combined Guinier/PDDF technique was implemented to mitigate Guinier R_g's sensitivity to minor sample aggregation, which is commonly observed in our polymer nanoparticle sample analysis. The results of these two methods are then evaluated to decide whether PDDF-Informed or *Auto Guinier* yields the best solution based on the mean residuals from the calculated *I(q)* from the Guinier approximation Eq. 1 to the data in the region of $qR_g < 2$. This range was chosen as this is the plotted range of the dimensionless GPA plot shown in summary plots. The results of each method are both displayed in the summary plots for the user to validate the script decision, including dimensionless GPA plots, Guinier fit, and residuals to fit.

SAXS Assistant selects whether *PDDF-Informed* or *Auto Guinier* yields a better fit by calculating the method with the lowest mean residuals from Eq. 1 and the collected data. A green box is added

to the summary plot GPA for the solution chosen, unless both *PDDF-Informed* and *Auto Guinier* yield the same result.

Training ML model to Predict D_{max} from SASBDB Data

An MLP regressor was trained to predict D_{max} using the data obtained from the SASBDB, using entries with $D_{max} \leq 300$ as majority of data was within this range. The MLP consisted of four hidden layers, each having 32 neurons, with the parameters set as follows: activation = 'relu', solver = 'adam', alpha = 1e-3, max_iter = 1000, and random_state = 42. First, features were extracted from the profiles by transforming the data into the dimensionless Kratky scale as shown by Franke et al (16). Subsequently, the normalized Porod invariant of the dimensionless Kratky plot was calculated up to qR_g = 3, 4 and 5 using normalized apparent volume, *V'*, as the final value (16).

$$Q' = \int_0^{qRg} (qR_g)^2 I(qR_g)/I(0) dqR_g \rightarrow \quad V' = \frac{2\pi^2}{Q'} \tag{2}$$

In addition to the features obtained from the above transformation, Guinier R_g and minimum qR_g (minimum q after sampling), were input features for the model. Data scaling was done using *StandardScaler()* from scikit-learn. The data was split into training and testing sets using a 60/40 split, as a large test set was preferred for robustness in model validation. The R^2 score and mean absolute error (MAE) were used as evaluation metrics and tested with 1,294 experimental profiles from the SASBDB (see Supplemental Data). The optimal architecture was identified by attempting different numbers of neurons and hidden layers to predict D_{max} (6). The model was tested for

overfitting by evaluating performance as the dataset size increased, in which case a significant decline in performance as data fraction increased would indicate overfitting (Fig. S1) (13).

To assess the relative influence of input features on D_{max} prediction, we performed a permutation importance analysis using the *permutation_importance* function from scikit-learn. The trained MLP regressor was evaluated on the test set, and baseline performance was established by calculating R^2. Each feature (Guinier R_g, *V'3*, *V'4*, *V'5*, and minimum qR_g) was then randomly permuted 100 times, and the resulting decrease in R^2 was recorded. The mean of the R^2 drop were computed to quantify the stability of feature importance estimates. This approach allowed us to rank features according to their impact on predictive performance, with larger R^2 decreases indicating greater importance.

Shape Mapping and Structure Classification Through Clustering

Unsupervised clustering was done using a GMM. The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) were calculated to evaluate optimal cluster selection for up to nine clusters (Fig. S2). The final selection of cluster number considered both the BIC and Kratky plots of high confidence samples in each cluster to determine if sufficient resolution was achieved within the clusters (28). The Kratky plots were used to evaluate shape separation considering the effect of dataset size on the evaluation metric, where BIC tends to favor lower model complexity, to prevent overfitting (29,30). Meanwhile the AIC score is not as strict at penalizing for model complexity (30). The two cluster numbers that lead to the lower BIC scores were examined along with the Kratky of high confidence samples for each cluster during optimal cluster number selection (Figs. S3 and S4). Samples were excluded if they had fewer than 100 neighbors within the top 10% pairwise distance threshold, as these were considered extreme outliers unlikely to

contribute meaningfully to clustering. In total, 6 out of 3,328 samples were excluded based on this criterion.

To assess the relative contribution of scattering features (*V3, V4, V5*) in GMM clustering, we performed a feature separation analysis. Following GMM fitting, cluster labels were extracted, and several complementary metrics of feature relevance were calculated. Between- and within-cluster variances were estimated from the GMM means and covariances, and their ratio reported as a Fisher-style separation score. Additional measures included the range of cluster means, one-way ANOVA F-statistics with associated p-values (using *f_classif* from scikit-learn), and mutual information between features and assigned clusters (*mutual_info_classif*). Together, these metrics quantify how strongly each feature discriminates cluster membership.

Polymeric SCNP Synthesis and SAXS Sample Preparation

Statistical copolymers were prepared via automation-assisted photoinduced electron/energy transfer reversible addition-fragmentation chain transfer (PET-RAFT) polymerization in 96-well plate format as previously described (31-33). Copolymer designs were selected by Latin hypercube sampling (LHS) to combine up to four monomers in any feed ratios with degrees of polymerization (DP) ranging from 100 to 400. The monomers in this library included 2-hydroxypropyl methacrylate (HPMA) as a neutral monomer, [2-(methacryloyloxy)ethyl]-trimethylammonium chloride (TMAEMA) and 3-sulfopropyl methacrylate (SPMA) as ionizable monomers, methyl methacrylate (MMA), 2-hydroxypropyl methacrylate (HPMA), diethylaminoethyl methacrylate (DEAEMA), trifluoroethyl methacrylate (TFEMA), and butyl methacrylate (BMA) as hydrophobic monomers, poly(ethylene glycol) methacrylate (PEGMA) and N-[3-(dimethylamino)propyl]methacrylamide (DMAPMA) as hydrophilic monomers, and [2-

(methacryloyloxy)ethyl]dimethyl-(3-sulfopropyl) methacrylate (SBMA) as a zwitterionic monomer.

A Python script prepared Excel-based synthesis sheets that directed the Hamilton Microlab STARlet liquid handling robot's reagent transfer steps based on polymer designs selected through LHS. These synthesis sheets control monomer feed ratios, DP, and well position. Stock aliquots prepared in DMSO of monomer (2 M), chain transfer agent (50 mM), and photoinitiator (2 mM) were loaded into the robot and automated reagent transfers were carried out in clear, flat-bottom 96-well polypropylene plates to a final well volume of 200 µL and a final monomer concentration of 1 M. While CTA:initiator ratios were fixed at 25:1, monomer:CTA ratios ranged from 50 to 200 to control DP. Following a mixing step, plates were covered with plate sealing tape and transferred to our custom 96-array LED lightbox, which offers per-well control of photoinitiation driven by an Arduino UNO R4 Minima microcontroller (34). After 16 hours of light exposure, the plates were removed from the lightbox for purification.

Samples were diluted 10x in ultrapure water, transferred to 3.5 MWCO cellulose dialysis tubing, and suspended in 4 L of ultrapure water for dialysis. After 72 hours and four water changes, samples were transferred from dialysis tubing to 15 mL centrifuge tubes, frozen and subsequently lyophilized until purified polymer remained. Lyophilized polymers were redissolved in 100 mM potassium phosphate buffer at pH 7.4 and prepared at multiple concentrations for SAXS analysis.

Results and Discussion

Analysis Efficiency

For each profile analyzed, the script generates a single-page PDF sheet populated with summary plots for rapid user inspection of the analysis results, along with the exported raw graph data for re-graphing. (Figs. 2, S5-S7) (35). In the summary plots PDF, the plots corresponding to Guinier and PDDF are displayed in a similar style to BioXTAS RAW to increase familiarity for experimentalists accustomed to BioXTAS RAW software (3). The plots in the summary sheet include: (i) the scattering profile as *log(I(q))* vs. *q*; (ii) GPA plot; (iii) Kratky plot; (iv) Guinier fit, residuals and dimensionless GPA plot with the fitted data and calculated *I(q)* from R_g and *I(0)* and assigned cluster probabilities by GMM; and (v) PDDF plot fit to the data and residuals, including an ML-derived prediction of D_{max}. Together, the summary plots provide a concise one-page overview of both raw as well as derived information to support automated analysis and user validation. The scattering profile, GPA, and Kratky plots allow assessment of overall data quality and folding state; the Guinier and PDDF panels allow direct evaluation of R_g and D_{max}. Meanwhile, the ML components provide insights derived from models trained experimental data from SASBDB where GMM probabilities provide a structural classification summary and D_{max} predictions can help users assess the validity of the IFT-derived D_{max} value and curve. If PDDF-Informed and *Auto Guinier* yield different R_g values, a green box surrounds the selected solution (Fig. 3 *a*-*b*) (36,37). We evaluated SAXS Assistant's ability to correlate PDDF and Guinier R_g values against a conventional workflow that used *Auto Guinier* from RAW followed by PDDF determination with BIFT configured to start from Guinier q_{min} (Fig. 3 *c* and Table 1).

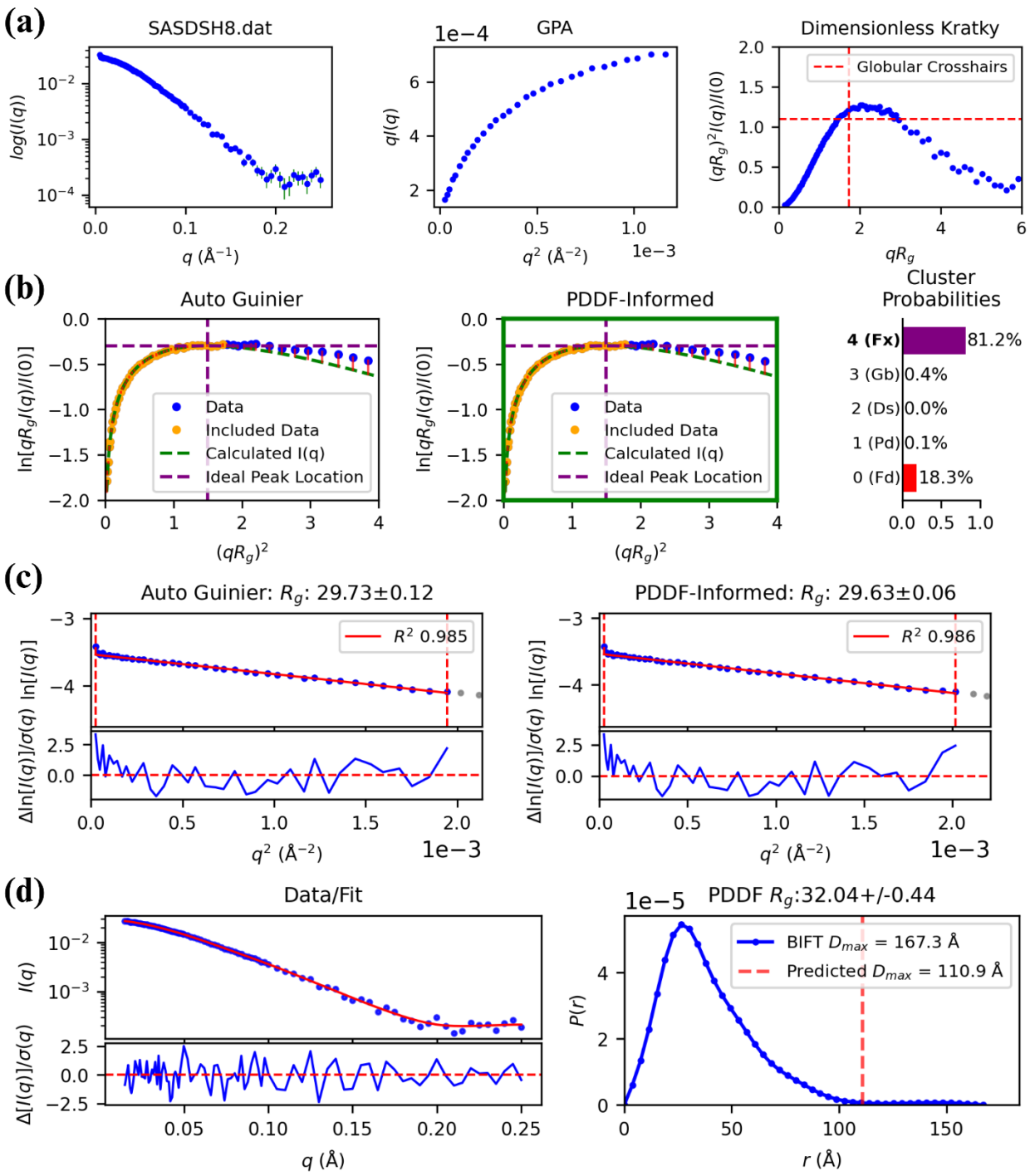

Figure 2. Summary plots for a solved file by SAXS Assistant. (a) Profile, GPA plot, and dimensionless Kratky. (b) Dimensionless GPA showing the fit of the data according to calculated *I(q)* from both R_g methods with the selected method surrounded by a green box, GMM-assigned probabilities abbreviated as Fx (Cluster 4, Flexible), Gb (Cluster 3, Globular), Ds (Cluster 2, Disordered), Pd (Cluster 1, Polydisperse), and Fd (Cluster 0, Folded). (c) Guinier fit window and residuals for both selection methods. (d) PDDF fit to the data, residuals, PDDF curve obtained from BIFT with a vertical line indicating the D_{max} predicted by ML, where red indicates more than 20% difference in D_{max} from IFT.

In SAXS Assistant, R_g values are selected through decision blocks, that first require agreement between reciprocal- and real space- derived R_g. This is followed by comparison of mean residuals between the two Guinier methods implemented (see Section 2, Determination of R_g). Files without such agreement are flagged as unsolved. Results from both the SAXS Assistant and the conventional approach (Auto Guinier + BIFT without decision blocks) were compared with the submitted Guinier and PDDF values for the same 3,170 SASBDB entries. Both workflows used identical input profiles and q-spacing, ensuring that differences in outputs arose only from the analysis logic. While the conventional method's mean discrepancy was 5.45% compared to SASBDB, SAXS Assistant reduced this to 1.77%, much closer to the reported values at 2.42%. Disagreement was much higher when *Auto Guinier* was used with BIFT in the absence of decision blocks as implemented by SAXS Assistant, with a 26% mean disagreement. A fraction of samples have significantly large differences in R_g from PDDF and Guinier, resulting in a large average percent difference as seen by the median values, however when dealing with large datasets this results in increased effort to identify and correct. We conclude that SAXS Assistant outperforms the conventional method and manual analysis in reporting R_g values that are in agreement with both Guinier and PDDF. The script performance is especially relevant when adapting high-resolution profiles by resampling to match the coarser q-spacing typical of our experiments ensuring comparability. Although higher-resolution data may perform better, the goal here is to enable fair comparison in the format of the lower-resolution data we routinely collect. This demonstrates that the SAXS Assistant pipeline minimizes the impact of problematic cases while maintaining consistency with reported to SASBDB values.

This is also validated when comparing R_g found by SAXS Assistant and conventional methods were compared to the corresponding reported values in SASBDB. Across the 3,170 entries, both

Auto Guinier and SAXS Assistant reproduced depositor Guinier R_g values within ~2% on average, demonstrating the robustness of *Auto Guinier* at identifying R_g (Table 2). This performance is further complemented in SAXS Assistant by the *PDDF-informed Guinier* method, which helps prevent outliers arising from discrepancies between real- and reciprocal-space R_g estimates. For PDDF-derived R_g, SAXS Assistant showed a mean percent deviation of 3.6% (median 1.8%), whereas BIFT alone showed a mean percent deviation of 24.0% despite a median of only 2.5%, reflecting the presence of severe outliers (Table 2). These findings highlight that the inflated average reported for BIFT is not representative of the majority of cases but rather driven by a minority of problematic datasets. SAXS Assistant mitigates this effect through its decision-block framework, ensuring improved robustness across large datasets thus increasing efficiency while minimizing trade-off between analysis quality. These findings are consistent with prior evaluations of BIFT in the BioXTAS RAW package, which reported that BIFT characteristically overestimates D_{max} relative to experimenter-reported values, which in turn can lead to corresponding overestimation of R_g from the PDDF (3).

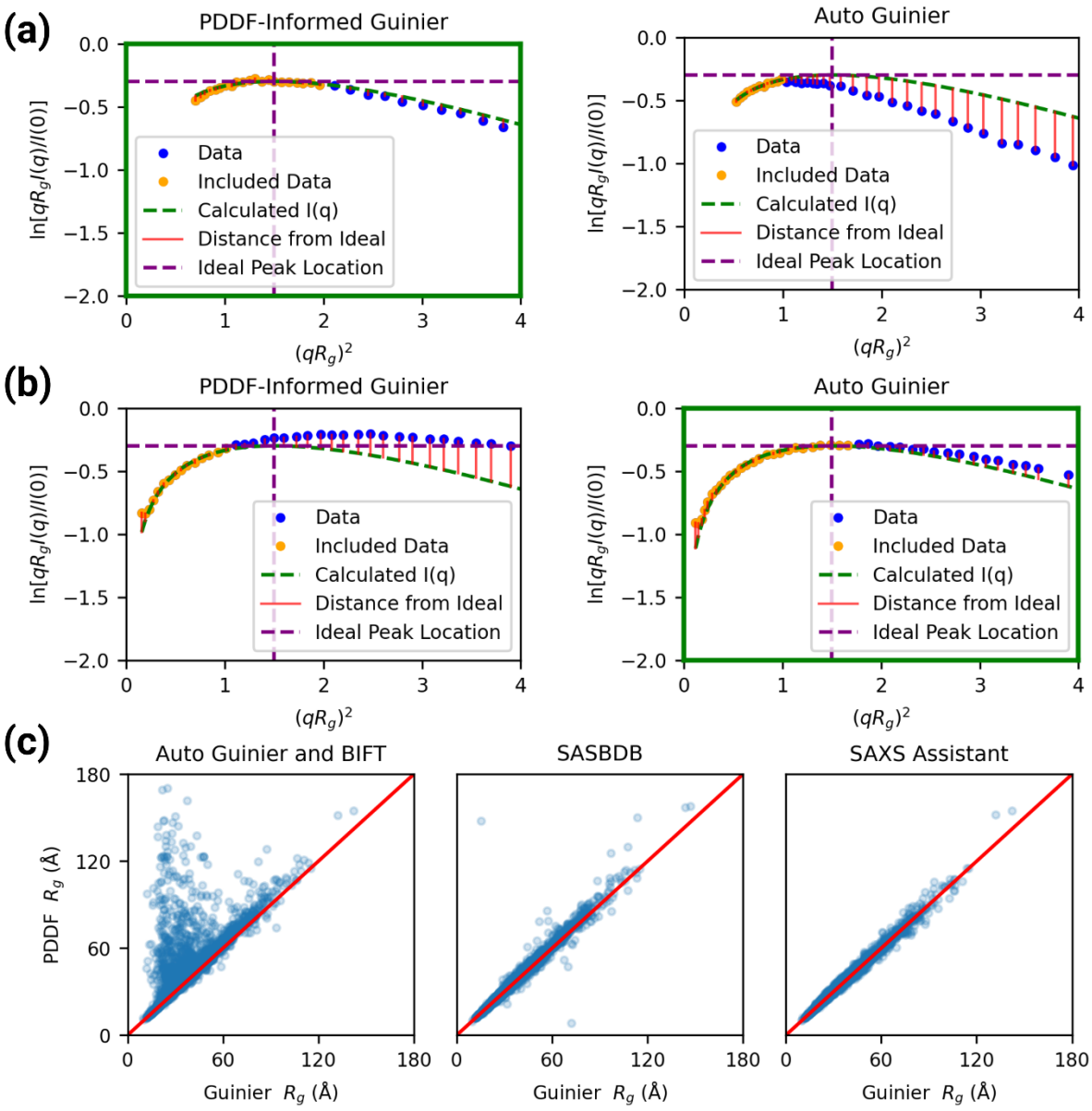

Figure 3. (a) *PDDF-Informed* provided a better agreement with calculated *I(q)*. *PDDF-Informed* R_g = 27.91 Å *Auto Guinier* R_g = 24.21 Å, SASBDB reported R_g =28.12 Å (SASDDG4). (b) Better agreement from *Auto Guinier* R_g = 38.0 Å. *PDDF-Informed* R_g = 43.76 Å, SASBDB-reported R_g = 36.4 Å (SASDA68). (c) Comparison of disagreement between Guinier and PDDF R_g.

Auto Guinier : x- axis RAW- derived . y-axis BIFT-derived w/o decision blocks

SASBDB: Guinier reported by experimentalist y- PDF Rg as reported by experimentalist

SAXS assistant: x- axis after constraining based GPA/PDF etc PDF-Rg is the arbiter followed by GPA. y-axis PDF-Rg. Adjusted for q min etc. heuristic approach chi value etc. log alpha etc.

Table 1. Comparison of performance at R_g validation demonstrates that SAXS Assistant outperforms the conventional method and is on par with analyses reported to the SASBDB.

Method	Mean % Disagreement	Median % Disagreement	75th Percentile
Auto Guinier and BIFT	26.00	5.45	13.51
SASBDB Submissions	3.72	2.42	4.70
SAXS Assistant	2.79	1.77	3.93

Table 2. Percent deviation of Guinier- and PDDF-derived R_g values from SASBDB-reported values for SAXS Assistant and conventional workflows without decision blocks as implemented by SAXS Assistant.

Method	Mean % Deviation	Median % Deviation	75th Percentile
SAXS Assistant Guinier R_g	1.57	0.87	2.07
Auto Guinier R_g	1.71	0.84	2.05
SAXS Assistant PDDF R_g	3.61	1.84	4.40
BIFT PDDF R_g	24.04	2.47	11.59

Estimating D_{max} with ML Assistance

The D_{max} of a scatterer is typically determined by the PDDF (6,8,12). The PDDF represents the distribution of electron pair distances in a scatterer, which, in addition to size information, provides structure and morphology insights based on the shape of the curve (6-8,12,38). Fig. S8 *a-b* shows the PDDF of three proteins from the SASBDB (SASDA32, SASDMX3, SASDJ92) of characteristically different structures and their corresponding bead models (14,39,40). BSA is a compact globular protein with a symmetrical, bell-shaped PDDF (8,12,41). S. epidermidis

extracellular binding protein (EmbP), which has an elongated structure, results in an asymmetrical PDDF with an extended tail (8,12). Immunoglobulin is a compact multi-domain protein, and this structure is reflected by two peaks in the PDDF (12,42). The PDDF can also be utilized for advanced characterization, such as *ab initio* model reconstructions (7,38).

Given the need for expert intuition in selection of an appropriate PDDF and D_{max}, automating PDDF selection is a challenging task (3,6). For this, we turn to ML models trained to predict D_{max}, leveraging large datasets of expert and researcher-selected PDDF curves. Our goal was the transfer of researcher-level intuition toward automated D_{max} determination (6,14,15). ML methods have been used to predict D_{max} without requiring an IFT, including the application of simulated data from geometrical models and of experimentally derived models for synthetic data generation for model training (6,16). To the best of our knowledge, SAXS Assistant is the first to use a large dataset of experimental profiles only. Here, we trained an MLP neural network to predict D_{max} with the architecture shown in Fig. 4 *a*. An MLP was selected for this application because they are well-suited for tasks involving non-linear data and are often used in applications of image recognition, natural language processing, forecasting, and pattern recognition (43,44). MLPs are neural networks composed of an input layer, hidden layers and an output layer, wherein each layer contains several neurons (6,43,44).

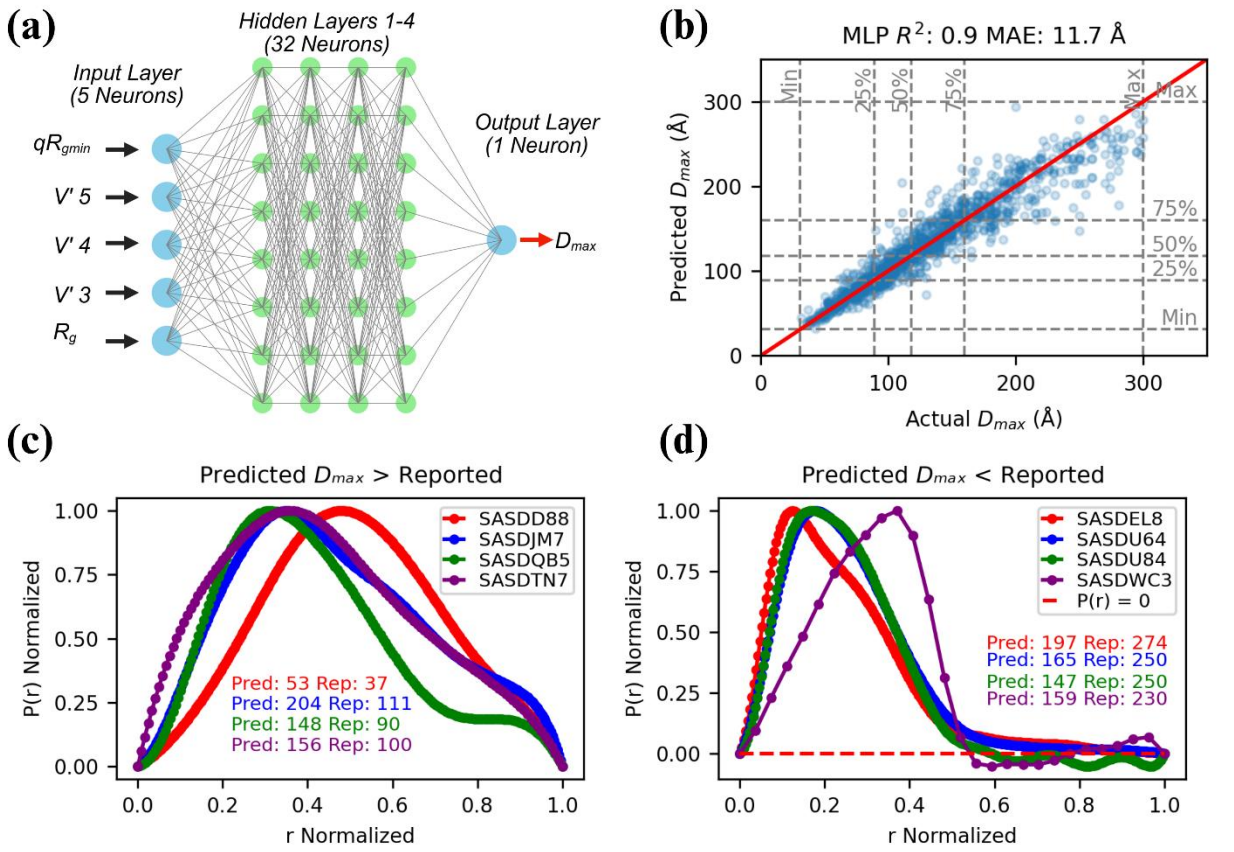

Figure 4. (a) Architecture of MLP. (b) Predicted versus actual D_{max} for the MLP model predicting on an unseen test set. (c) PDDFs corresponding to samples where the MLP predicted a D_{max} greater than the D_{max} reported to the SASBDB. Observation of the SASBDB PDDF demonstrates that the model's prediction was more appropriate and the SASBDB-reported D_{max} was an underestimation, leading to early truncation of the PDDF curve. (d) Samples reporting overestimated SASBDB D_{max} values made the model appear to underestimate D_{max}, though its predictions accurately reflected the expected curve shape. In (c) and (d), PDDFs were normalized by dividing P(r) by its maximum and r by the corresponding D_{max} (maximum r-value), such that both axes ranged up to 1 to highlight curve shapes independent of absolute scale.

Regression metrics were used to evaluate the model's performance when trained on 1,940 samples to predict D_{max} for 1,294 unseen test samples from SASBDB. The MLP yielded R^2 = 0.90 and MAE of 11.7 Å, demonstrating significant predictive power for D_{max} (Fig. 4 *b*). Only 115 files from the test set (9%) had a D_{max} deviating more than 20% from the reported SASBDB value. To further investigate the model's ability to predict D_{max}, we looked at samples for which the model predictions were > 20% from actual values. This threshold accounts for a potential 5-10%

uncertainty in D_{max} values submitted to SASBDB. Close examination of cases with predictions greater than the reported values often indicate a PDDF curve with underestimated D_{max}, not smoothly reaching zero (Fig. 4 *c*) (45,46). Similarly, when the model's predictions were less than the reported values, it was likely that the SASBDB PDDF had an overestimated D_{max} (Fig. 4 *d*) (47-49). Permutation importance analysis revealed that Guinier R_g was by far the most influential feature, with R^2 dropping by ~1.6 when permuted. By contrast, features such as *V'3*, *V'4*, and *V'5* had moderate contributions, while minimum qR_g showed relatively small influence compared to all other features. This confirms that while multiple features refine the predictions, the model strongly relies on Guinier R_g. The high importance of Guinier R_g is not surprising, as Franke et al. previously used Kratky-derived features for D_{max} prediction, in which the model required R_g as an input (16).

The MLP model's performance suggests it can intuitively predict D_{max} and assess the selected PDDF's appropriateness. The model was integrated in the script as a tool to assess the validity of the PDDFs obtained from BIFT. The model's predicted D_{max} value will be displayed on the PDDF obtained by BIFT. If D_{max} values differ by more than 20%, the prediction will be shown in red on the summary plots, else these will be shown in green (Figs. 2, S5 and S6). Through the SAXS Assistant pipeline, PDDFs consistent with Guinier-derived R_g and *I(0)* are obtained, and their D_{max} values are cross-checked using an MLP model trained on SASBDB data. This ML-based step does not replace conventional PDDF use in ab-initio reconstruction but rather provides a tool assisting in automated validation to flag potential over- or under-estimated D_{max} values.

Shape Mapping and Structure Classification

Previously, synthetic data of predefined shapes or classes were used as references to compare the structure of biological samples (16). In this work, rather than defining the structural landscape by pre-defined classes (shapes), our goal was to identify classes based solely on experimental SASBDB data. Typically, structure probing from SAXS data relies on qualitative analysis using Kratky plots. Dimensionless Kratky plots of globular proteins show a maximum peak at the crosshairs position of ($\sqrt{3}$, 1.1) (8,12,41). An intrinsically disordered or cylindrical biopolymer's Kratky curve will continue to increase as qR_g increases (8,12,41,50). Proteins with both elongated, unstructured regions and folded regions would display a peak beyond the characteristic globular peak position, with a curve that decreases at higher qR_g (8,12,41,50). An extended chain or experimentally unfolded protein shows an initial plateau followed by monotonic increase (8,12,41,50).

We aimed to complement Kratky analysis by incorporating ML-derived quantifiable metrics (12). This was accomplished by applying a GMM to cluster the mapped SASBDB entries into distinct classes. Algorithms like K-Means clustering assign data points to clusters based on their distance from a centroid, assuming circular or spherical thresholds for clusters (28,30). Hard thresholds are not necessarily aligned with real-world complexity, such as the structural composition of biologics (8,12). GMMs are probabilistic in nature, whereas this assumes that data points are from a mixture of finite Gaussian densities (28,30). GMMs perform soft clustering, determining the likelihood that a datapoint belongs to a cluster, preventing the loss of information for data displaying overlapping characteristics (28). GMM clustering can quantify complex morphologies, making it particularly useful for intrinsically disordered proteins (IDPs), which may contain both compact

and disordered or flexible regions (8,12,30). The entries mapped into 3D-space and clustered by GMM are shown in Fig. S9.

The Kratky plots of high-confidence samples within each cluster, or those with a probability greater than 0.90 of belonging to each cluster, are shown in Fig. S4. Cluster 0 (referred to as Folded) with 884 entries shows folded structures that did not display the characteristic globular peak (8). Both the BSA dimer and the alcohol dehydrogenase monomer samples fall within this cluster of folded but nonglobular morphologies. Cluster 1 (Polydisperse) contains 163 entries, most of which display a globular peak, with some showing an additional peak at higher qR_g. Many polydisperse systems are found in this cluster, including mixtures of different proteins where components may differ in size and shape as well as systems with heterogeneity. Samples found within Cluster 1 include ordered structures, core-shell structures like polysorbate micelles, HIV-1 envelope glycoproteins, and lipid nanoparticles. Cluster 2 (Disordered) had 697 entries with Kratky plots displaying characteristics of highly flexible biomolecules, Gaussian chain-like structures, unfolded proteins, as well as plots resembling those seen in studies of protein aggregation (41,51). Cluster 2 includes well-known IDPs like α-synuclein, which has random coil regions and is linked to Parkinson's disease through amyloid fibril aggregation in neurons (51). A second well-known IDP found in this cluster was the tau protein, which is associated with Alzheimer's disease through the formation of β-sheet intracellular aggregates (52). Cluster 3 (Globular) had 693 entries of globular structures including monomeric BSA, tetramer alcohol dehydrogenase, and lysozyme (8,12). Cluster 4 (Flexible) is the largest cluster with 885 samples with structures that resemble partially unfolded proteins or samples with flexible regions (8,12). An example of proteins in this cluster are a mutated dimeric aldehyde-alcohol dehydrogenase, in which the mutation resulted in a more extended conformation compared to the non-mutated

enzyme (53). Another entry in this cluster is chitinase, a plant enzyme from the class I glycoside hydrolase family 19. Chitinase's structure is known to consist of two domains connected by a flexible linker (54).

GMM clustering was applied to quantify the structural composition of samples by comparison of class assignment probabilities of monomer (SASDEE5), purified dimer (SASDFR8), and an unspecified mixture of BSA (SASDDN3) (16,55,56). Fig. 5 *a* shows the Kratky and cluster probabilities for these three different samples. For monomeric BSA, globular character dominates (Cluster 3, Globular). The dimer shows a primarily folded structure with some globular features (Cluster 0, Folded). The mixture shows an intermediate between the dimer and the monomer forms, with an increase of globular structural character compared to the dimer suggesting their potential presence. This mixture is from a dataset used by Franke et. al, the pioneers in SAXS data featurization through Kratky, who previously noted a potential presence of dimers in the solution (16). Their strategy was adopted in this work to quantify structural composition probabilistically, which validates their assumption on why this sample slightly deviated from the expected globular sample (16). We further validated the GMM clustering's ability to quantify structural composition with mAb, Palivizumab (SASDSU6), a nanobody (SASDSV6), and a complex of the two (SASDSW6) in Fig. 5 *b* (57). Antibodies are Y-shaped proteins whose structure consists of three functional components including two antigen binding domains (Fab), the fragment crystallizable region (Fc), and a linker providing conformational flexibility (42). The structural composition of the mAb is reflected in the assigned probabilities, with the highest probability belonging to Cluster 0 (Folded), followed by the Flexible and Disordered clusters. The nanobody showed a high probability of falling within the flexible/partially unfolded class (Cluster 4). The complex shows a shift in structural probabilities, forming an intermediate between the antibody and nanobody,

indicating a transition in overall structural behavior. The probability of a sample belonging to one of these clusters from classes discovered by the model may be used to enhance Kratky in understanding the structural composition of biologics. Considering that some biologics can exhibit structurally distinct regions, this probabilistic clustering approach offers a realistic quantification of molecular structure in reference to existing biological data.

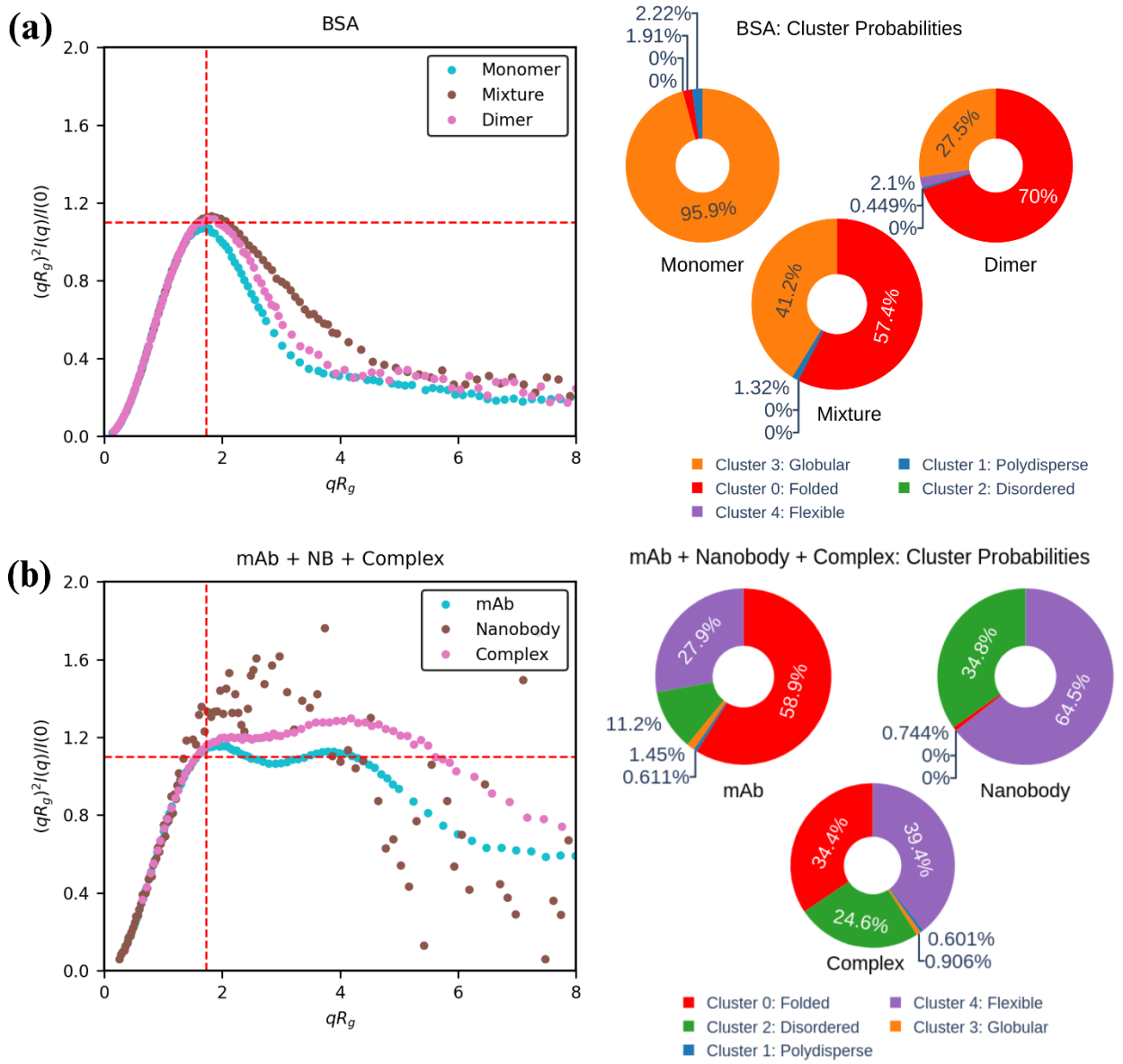

Figure 5. Dimensionless Kratky plots and corresponding cluster probability distributions. (a) BSA as a monomer (SASDEE5), purified dimer (SASDFR8), and in a mixture (SASDDN3). (b) Palivizumab (SASDSU6), its nanobody (SASDSV6), and their complex (SASDSW6). GMM

clustering assigns a probability of the samples belonging to the different structural classes: Cluster 0 (Folded), Cluster 1 (Polydisperse), Cluster 2 (Disordered/Extended), Cluster 3 (Globular), and Cluster 4 (Flexible / Partially Unfolded). Structural transitions across conditions are reflected by shifts in class probabilities. The mixture of BSA as well as the mAb-nanobody complex display shifts corresponding to the contribution of the individual components in each system.

Polymer Conformations

SCNPs are formed when a polymer chain collapses in solution due to a balance of intermolecular and intramolecular interactions between solvent and polymer, as well as between polymer chains (58). There is increasing interest in designing SCNPs that mimic protein folding for applications such as catalysis, nanoreactors, sensors, and biomedicine, using either covalent or non-covalent strategies (58-61). Non-covalent approaches rely on designing copolymer compositions that undergo hydrophobic collapse to form folded conformations (59,62). This principle aligns with the foldamer catalysis hypothesis, suggesting hydrophobic collapse can drive structural compaction, sometimes resulting in folded conformations yielding functional sites (63).

We applied the SAXS Assistant script for the group's ongoing research on polymeric SCNPs synthesized from statistical copolymers that result in self-assembled structures driven by hydrophobic collapse. Fig. 6 *a* shows one of these statistical copolymer-based SCNPs (SCNP No.1), which exhibits a concentration-dependent conformation. At higher concentrations, SCNP No.1 adopts a compact, globular conformation. At lower concentrations, it transitions to a modestly less globular, slightly more extended folded state, changing shape while maintaining its size, as shown by only a small increase in D_{max} at the lower concentration (Fig. S10). At higher concentrations, some proteins adopt more compact structures that enhance stability, in accordance with the excluded-volume theory (64). The crowding-responsive behavior of this SCNP supports the notion that polymeric SCNPs functionally resemble proteins, particularly IDPs (59,62).

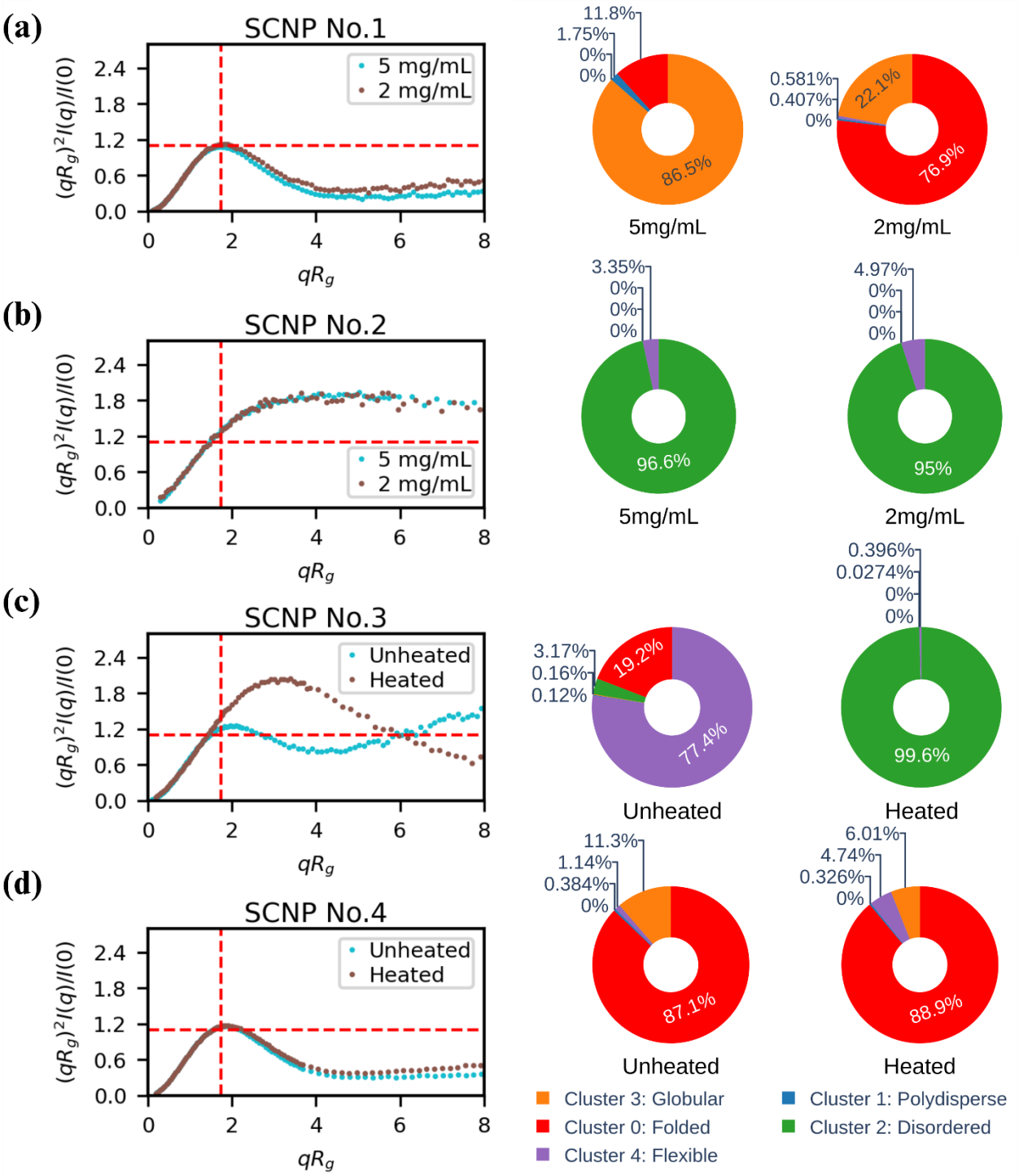

Figure 6. (a,b) Dimensionless Kratky plots and cluster probabilities of SCNPs showing protein-like responses to concentration. (a) SCNP No.1 exhibits folded globular behavior, losing globularity at lower concentrations. (b) SCNP No.2, the dominant polymeric SCNP form, shows persistent disorder regardless of concentration, like some IDPs. (c-d) SCNP responses to 80˚C heat stress for one hour. (c) SCNP No.3 shows major structural changes, consistent with protein-like behavior. (d) SCNP No.4 retains structural integrity, deviating from protein-like responses.

In the presence of crowders, IDPs can be categorized into three categories: foldable, non-foldable (remain unstructured), and unfoldable (undergo further unfolding). The majority of literature cites a lack of significant structure-forming effects of crowders on IDPs (64). This observation also validates the notion of similarity between SCNPs and IDPs. The limited literature on foldable IDPs may reflect their rarity, just as monomeric compositions leading to compact SCNPs are rare. In SCNP work by our group, most compositions result in unstructured and disordered conformations like the example seen in Fig. 6 *b*. Because compact SCNPs are much rarer than extended SCNPs, class imbalance will always favor structurally disordered SCNPs. This phenomenon points to the need for efficient optimization methods like ML supplemented with probabilistic structural characterization, which is capable of potentially addressing these class imbalance issues. Fig. 6 *c* shows SCNP No.3 which, similar to proteins, loses structural integrity after heat stress at 80 °C for one hour. In contrast, SCNP No.4 (Fig. 6 *d*) deviates from behavior typical of proteins, whereby under the same heat stress conditions, its structure remains intact. Proteins are known for their sensitivity to environmental factors like temperature, and overcoming this limitation has become a major research focus (32,65-67). Applying ML to uncover the mechanisms behind these differences will require identifying additional SCNPs with similar behaviors, enabling more confident conclusions about the compositional features responsible. Uncovering monomer compositions leading to different polymer conformations will shed deeper insights into structure–function relationships in synthetic systems, while also advancing our understanding of the mechanisms that govern protein folding and stability.

Feature importance in unsupervised learning like clustering remains an emerging area without standardized approaches. Our analysis of *V'3–V'5* relied on separation-based metrics, including between-/within-cluster variance ratios, ANOVA F-tests, and mutual information with cluster

labels (Table S1). These metrics are conceptually aligned with recent work proposing prototype-based feature importance (PBFI), which assess the contribution of features via differences in cluster prototypes (68). Consistent with this framework, we observed that *V'3* and *V'4* dominate cluster discrimination, while *V'5* contributes less (Table S1). Since *V'5* corresponds to a higher *q* region, this analysis suggests that clustering is primarily driven by low- and mid-*q* features from Kratky (*V'3*, and *V'4*). We therefore hypothesized that although background subtraction errors inevitably influence clustering, their impact is less pronounced for minor subtraction deviations, since these disproportionately affect the high-*q* region (69). To test this, we varied subtraction scaling factors and found that while absolute cluster probabilities shifted, the overall structural classification was preserved, supporting the robustness of the clustering results. Tables S2 and S3 show that for the samples in Fig. 6 *a*, even when subtraction was intentionally over- or under-scaled, the overall cluster identities followed the same trend. For instance, the 5 mg/mL sample generally maintained a higher probability of globular classification than folded, whereas the opposite remained true for the 2 mg/mL sample with the final cluster assignment remaining the same. These results indicate that while subtraction errors shift absolute probabilities, the relative classification trends remain robust under slight deviations. This is also not surprising considering the STDs of the features where *V'5* shows greater instability with a greater STD compared to the lower *q* features.

Limitations and Future Work

While SAXS Assistant performs well at automated Guinier- and PDDF-based parameter extraction and is well suited for high-throughput data analysis, several limitations should be noted. First, the accuracy of any analysis depends on the quality of the input data. Importantly, regarding appropriate subtractions as these can affect the validity of nearly all extracted parameters (70).

Therefore, while the clustering and regressor models capture features effectively to enable predictions, their sensitivity to background subtraction and thus high-q noise must be considered. Feature separation analysis on GMM indicated that clustering holds reduced sensitivity to minor subtraction errors, however poor data quality can still bias probability assignments. Since the SASBDB contains reported values which depend on depositor experience and analysis, thus may introduce some bias in model predictions. Second, the tool generalizes well across proteins and polymeric SCNP systems, however its applicability to concentrated or aggregated samples has not been extensively studied within this work. Particularly relevant, polydisperse samples were underrepresented in the GMM training set. In such cases, manual inspection or alternative modeling approaches (e.g., CREASE) may be more appropriate (18,20). Furthermore, SAXS Assistant resamples profiles by default to match our group's standard q spacing. While resampling had little effect on SASBDB evaluation results, it may not suit all users' preferences. Additionally, we note that for elongated or highly flexible scatterers, PDDF-derived R_g values are often systematically larger than Guinier R_g values. SAXS Assistant applies a 15% agreement threshold to balance robustness and automation, which means that in some cases the tool may preferentially report an R_g closer to the Guinier estimate even when a slightly larger PDDF R_g could be more appropriate (71). Lastly, as with any automated tool, results should not be taken as definitive without proper user evaluation. SAXS Assistant is intended to complement, not replace, expert judgment. The one-page summary plots are a central component of the workflow, enabling rapid validation and helping users decide when manual reanalysis is necessary. This distinction highlights that SAXS Assistant is best suited for large-scale, high-throughput analysis, with manual methods remaining preferable for specialized or borderline cases. These limitations also present opportunities for future work. Expanding training datasets to include more polydisperse

and aggregated systems, adding flexible resampling options, and developing improved background subtraction correction or detecting modules would further enhance the generalizability and usability of SAXS Assistant.

Conclusion

SAXS Assistant is an open-source analysis script that utilizes functionalities from open-source BioXTAS RAW program and integrates ML for prediction and structural analysis. Developed to extract R_g, D_{max}, Kratky, and shape information from subtracted SAXS data, organizes these parameters into an ML-compatible Excel format and user-friendly summary document for final validation. SAXS Assistant is available through GitHub, and PyPI via `pip install SAXS-Assistant`. Instructions can be found by running `from saxs_assistant import show_tutorial,show_supplemental` followed by the execution of `show_tutorial()`, and `show_supplemental()` which leads to folder with a .ipynb example to get started.

SAXS Assistant also provides tools to gain deeper insights into SAXS data from unsupervised ML models. Over 3,000 SASBDB entries were used for script validation and ML model training, which enabled learning from diverse biological data from researchers around the globe. The MLP model predicts D_{max} values from features extracted during analysis, aiding in assessing the appropriateness of PDDF curves. The script also uses GMM clustering on dimensionless Kratky-based features to classify samples into probabilistic classes. These classes can aid in quantifying the structural character in SAXS samples based on the different classes discovered through clustering of SASBDB data. This approach was applied to polymeric SCNPs to characterize their folding behavior, demonstrating how ML-derived structural classification can enhance the

qualitative interpretations from Kratky with quantitative metrics from the model. Together, these tools extend SAXS Assistant's utility beyond basic analysis to enable quantitative structural characterization of biologics and biomimetic synthetic systems. We believe this platform will prove to be useful to those dealing with large datasets and those new to SAXS data analysis, offering ML tools that provide insights from real biological data analyzed by researchers and experts.

Data and Code Availability

All supplemental data, including the SASBDB ID with structural parameters used for training regressor (xlsx), clustering model (xlsx) and evaluation of the script (xlsx), trained models (joblib), and analysis scripts (.py), are available on GitHub at: https://github.com/GormleyLab/SAXS-Assistant An installable version of SAXS Assistant is available in PyPI via `pip install SAXS-Assistant`. The script can be used to redirect to the supplemental data directly by running `from saxs_assistant import show_supplemental` followed by `show_supplemental()`.

Acknowledgements:

This research used the LiX beamline (16-ID) of the National Synchrotron Light Source II, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE by Brookhaven National Laboratory under Contract No. DE-SC0012704. The LiX beamline is part of the Center for BioMolecular Structure (CBMS), which is primarily supported by the National Institutes of Health, National Institute of General Medical Sciences (NIGMS) through a Center Core Grant

(P30GM133893), and by the DOE Office of Biological and Environmental Research (KP1607011). LiX also received additional support from NIH Grant S10 OD012331.

We thank Matthew Tamasi, Gabriella Tirado-Mansilla, Eugene Cheong, Alexander Suponya, Jordan Eckhoff, Eric Quartey, Dr. D. Christopher Radford and Dr. Prajakatta Mulay for their support as members of the research team. We also thank Brookhaven National Laboratory, where SAXS measurements were conducted, for providing access to their facilities and technical support. Lastly, we gratefully acknowledge all contributors to the SASBDB database for making their scattering data publicly available, which enabled the analyses presented in this work.

Funding

This work was funded by the National Institutes of Health (NIH), National Institute of General Medical Sciences (NIGMS) (R35GM138296) and the National Science Foundation (DMREF-2118860 and CBET-2309852). The SAXS data other than those in SASBDB were obtained at NSLS-II beamline 16-ID for life science x-ray scattering (LiX). The LiX beamline, part of the Center for BioMolecular Structure (CBMS), is primarily supported by NIGMS through a P30 Grant (P30GM133893), and by the DOE Office of Biological and Environmental Research (KP1605010). LiX also received additional support from NIH Grant S10 OD012331. As part of NSLS-II, a national user facility at Brookhaven National Laboratory, work performed at the CBMS was supported in part by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences Program under contract number DE-SC0012704.

Author Contributions

C.R. and A.G. conceived the project. C.R. developed the software and wrote the manuscript with input and contributions from A.G., N.M., J.B., E.D., E.A., C.L., M.P., and N.M. E.D. carried out

SCNP synthesis and purification for SAXS. E.D. and J.B. assisted with data acquisition, analysis and interpretation of results. J.B., N.M. and A.G. provided supervision and critical revisions. All authors contributed to the discussion of the results and approved the final version of the manuscript.

Declaration of Interests

The authors declare no competing interests.

Supporting Material

Supplemental Figures 1–10, are available in the Supplemental Information PDF document.

References

1. Grant, T. D., J. R. Luft, L. G. Carter, T. Matsui, T. M. Weiss, A. Martel, and E. H. Snell. 2015. The accurate assessment of small-angle X-ray scattering data. Acta Crystallogr D Biol Crystallogr. 71(Pt 1):45-56, doi: 10.1107/s1399004714010876.
2. Spill, Y. G., S. J. Kim, D. Schneidman-Duhovny, D. Russel, B. Webb, A. Sali, and M. Nilges. 2014. SAXS Merge: an automated statistical method to merge SAXS profiles using Gaussian processes. J Synchrotron Radiat. 21(Pt 1):203-208, doi: 10.1107/s1600577513030117.
3. Hopkins, J. 2024. BioXTAS RAW 2: new developments for a free open-source program for small-angle scattering data reduction and analysis. Journal of Applied Crystallography. 57(1):194-208, doi: doi:10.1107/S1600576723011019.
4. Putnam, C. 2016. Guinier peak analysis for visual and automated inspection of small-angle X-ray scattering data. Journal of Applied Crystallography. 49:1412-1419, doi: 10.1107/S1600576716010906.
5. Zabelskii, D. V., A. V. Vlasov, Y. L. Ryzhykau, T. N. Murugova, M. Brennich, D. V. Soloviov, O. I. Ivankov, V. I. Borshchevskiy, A. V. Mishin, A. V. Rogachev, A. Round, N. A. Dencher, G. Büldt, V. I. Gordeliy, and A. I. Kuklin. 2018. Ambiguities and completeness of SAS data analysis: investigations of apoferritin by SAXS/SANS EID and SEC-SAXS methods. Journal of Physics: Conference Series. 994(1):012017, doi: 10.1088/1742-6596/994/1/012017.
6. Molodenskiy, D. S., D. I. Svergun, and A. G. Kikhney. 2022. Artificial neural networks for solution scattering data analysis. Structure. 30(6):900-908.e902, doi: 10.1016/j.str.2022.03.011.
7. Schneidman-Duhovny, D., S. J. Kim, and A. Sali. 2012. Integrative structural modeling with small angle X-ray scattering profiles. BMC Structural Biology. 12(1):17, doi: 10.1186/1472-6807-12-17.
8. Tully, M., N. Tarbouriech, R. Rambo, and S. Hutin. 2021. Analysis of SEC-SAXS data via EFA deconvolution and Scatter. Journal of visualized experiments : JoVE.(167), doi: 10.3791/61578.

9. Liu, H., A. Hexemer, and P. H. Zwart. 2012. The Small Angle Scattering ToolBox (SASTBX): an open-source software for biomolecular small-angle scattering. Journal of Applied Crystallography. 45(3):587-593, doi: doi:10.1107/S0021889812015786.
10. 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. Journal of Applied Crystallography. 50(4):1212-1225, doi: doi:10.1107/S1600576717007786.
11. Petoukhov, M. V., D. Franke, A. V. Shkumatov, G. Tria, A. G. Kikhney, M. Gajda, C. Gorba, H. D. Mertens, P. V. Konarev, and D. I. Svergun. 2012. New developments in the ATSAS program package for small-angle scattering data analysis. J Appl Crystallogr. 45(Pt 2):342-350, doi: 10.1107/s0021889812007662.
12. Kikhney, A. G., and D. I. Svergun. 2015. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. FEBS Lett. 589(19 Pt A):2570-2577, doi: 10.1016/j.febslet.2015.08.027.
13. Meyer, T. A., C. Ramirez, M. J. Tamasi, and A. J. Gormley. 2023. A User's Guide to Machine Learning for Polymeric Biomaterials. ACS Polymers Au. 3(2):141-157, doi: 10.1021/acspolymersau.2c00037.
14. Kikhney, A. G., C. R. Borges, D. S. Molodenskiy, C. M. Jeffries, and D. I. Svergun. 2020. SASBDB: Towards an automatically curated and validated repository for biological scattering data. Protein Sci. 29(1):66-75, doi: 10.1002/pro.3731.
15. Valentini, E., A. G. Kikhney, G. Previtali, C. M. Jeffries, and D. I. Svergun. 2015. SASBDB, a repository for biological small-angle scattering data. Nucleic Acids Res. 43(Database issue):D357-363, doi: 10.1093/nar/gku1047.
16. Franke, D., C. M. Jeffries, and D. I. Svergun. 2018. Machine Learning Methods for X-Ray Scattering Data Analysis from Biomacromolecular Solutions. Biophysical Journal. 114(11):2485-2492, doi: https://doi.org/10.1016/j.bpj.2018.04.018.
17. Beltran-Villegas, D. J., M. G. Wessels, J. Y. Lee, Y. Song, K. L. Wooley, D. J. Pochan, and A. Jayaraman. 2019. Computational Reverse-Engineering Analysis for Scattering Experiments on Amphiphilic Block Polymer Solutions. Journal of the American Chemical Society. 141(37):14916-14930, doi: 10.1021/jacs.9b08028.
18. Ye, Z., Z. Wu, and A. Jayaraman. 2021. Computational Reverse Engineering Analysis for Scattering Experiments (CREASE) on Vesicles Assembled from Amphiphilic Macromolecular Solutions. JACS Au. 1(11):1925-1936, doi: 10.1021/jacsau.1c00305.
19. Heil, C. M., A. Patil, A. Dhinojwala, and A. Jayaraman. 2022. Computational Reverse-Engineering Analysis for Scattering Experiments (CREASE) with Machine Learning Enhancement to Determine Structure of Nanoparticle Mixtures and Solutions. ACS Central Science. 8(7):996-1007, doi: 10.1021/acscentsci.2c00382.
20. Akepati, S. V. R., N. Gupta, and A. Jayaraman. 2024. Computational Reverse Engineering Analysis of the Scattering Experiment Method for Interpretation of 2D Small-Angle Scattering Profiles (CREASE-2D). JACS Au. 4(4):1570-1582, doi: 10.1021/jacsau.4c00068.
21. Chen, Y. L., and L. Pollack. 2020. Machine learning deciphers structural features of RNA duplexes measured with solution X-ray scattering. IUCrJ. 7(Pt 5):870-880, doi: 10.1107/s2052252520008830.
22. Hansen, S. 2000. Bayesian estimation of hyperparameters for indirect Fourier transformation in small-angle scattering. Journal of Applied Crystallography. 33(6):1415-1421, doi: 10.1107/S0021889800012930.

23. Upadhya, R., E. Di Mare, M. J. Tamasi, S. Kosuri, N. S. Murthy, and A. J. Gormley. 2023. Examining polymer-protein biophysical interactions with small-angle x-ray scattering and quartz crystal microbalance with dissipation. J Biomed Mater Res A. 111(4):440-450, doi: 10.1002/jbm.a.37479.
24. DiFabio, J., S. Chodankar, S. Pjerov, J. Jakoncic, M. Lucas, C. Krywka, V. Graziano, and L. Yang. 2016. The life science x-ray scattering beamline at NSLS-II. AIP Conference Proceedings. 1741(1), doi: 10.1063/1.4952872.
25. Yang, L., E. Lazo, J. Byrnes, S. Chodankar, S. Antonelli, and M. Rakitin. 2021. Tools for supporting solution scattering during the COVID-19 pandemic. J Synchrotron Radiat. 28(Pt 4):1237-1244, doi: 10.1107/s160057752100521x.
26. Yang, L., S. Antonelli, S. Chodankar, J. Byrnes, E. Lazo, and K. Qian. 2020. Solution scattering at the Life Science X-ray Scattering (LiX) beamline. Journal of Synchrotron Radiation. 27(3):804-812, doi: doi:10.1107/S1600577520002362.
27. Zheng, W., and R. B. Best. 2018. An Extended Guinier Analysis for Intrinsically Disordered Proteins. Journal of Molecular Biology. 430(16):2540-2553, doi: 10.1016/j.jmb.2018.03.007.
28. Weber, C. M., D. Ray, A. A. Valverde, J. A. Clark, and K. S. Sharma. 2022. Gaussian mixture model clustering algorithms for the analysis of high-precision mass measurements. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. 1027:166299, doi: https://doi.org/10.1016/j.nima.2021.166299.
29. Gorbatovski, A., and S. Kovalchuk (2023). Bayesian Networks for Named Entity Prediction in Programming Community Question Answering. In J. Mikyška, C. de Mulatier, M. Paszynski, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. A. Sloot, eds. Computational Science – ICCS 2023. Springer Nature Switzerland.
30. Wan, H., H. Wang, B. Scotney, and J. Liu (2019). A Novel Gaussian Mixture Model for Classification. 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC).
31. Di Mare, E. J., A. Punia, M. S. Lamm, T. A. Rhodes, and A. J. Gormley. 2024. Data-Driven Design of Novel Polymer Excipients for Pharmaceutical Amorphous Solid Dispersions. Bioconjugate Chemistry. 35(9):1363-1372, doi: 10.1021/acs.bioconjchem.4c00294.
32. Tamasi, M., R. Patel, C. Borca, S. Kosuri, H. Mugnier, R. Upadhya, N. S. Murthy, M. Webb, and **A. J. Gormley**. 2022. Machine learning on a robotic platform for the design of polymer-protein hybrids. Advanced Materials. 34:2201809, doi: 10.1002/adma.202201809.
33. Tamasi, M., S. Kosuri, J. DiStefano, R. Chapman, and **A. J. Gormley**. 2020. Automation of controlled/living radical polymerization. Advanced Intelligent Systems. 2:1900126, doi: 10.1002/aisy.201900126.
34. Lee, J., P. Mulay, M. J. Tamasi, J. Yeow, M. M. Stevens, and **A. J. Gormley**. 2023. A fully automated platform for photoinitiated RAFT polymerization. Digital Discovery. 2:219-233, doi: 10.1039/D2DD00100D.
35. Kieslich, B., R. H. Weiße, J. Brendler, A. Ricken, T. Schöneberg, and N. Sträter. 2023. The dimerized pentraxin-like domain of the adhesion G protein-coupled receptor 112 (ADGRG4) suggests function in sensing mechanical forces. Journal of Biological Chemistry. 299(12), doi: 10.1016/j.jbc.2023.105356.
36. Mylonas, E., and D. I. Svergun. 2007. Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering. Journal of Applied Crystallography. 40(s1):s245-s249, doi: doi:10.1107/S002188980700252X.
37. Yuenyao, A., N. Petchyam, N. Kamonsutthipaijit, P. Chaiyen, and D. Pakotiprapha. 2018. Crystal structure of the flavin reductase of Acinetobacter baumannii p-

hydroxyphenylacetate 3-hydroxylase (HPAH) and identification of amino acid residues underlying its regulation by aromatic ligands. Archives of Biochemistry and Biophysics. 653:24-38, doi: https://doi.org/10.1016/j.abb.2018.06.010.

38. Liu, H., and P. H. Zwart. 2012. Determining pair distance distribution function from SAXS data using parametric functionals. Journal of Structural Biology. 180(1):226-234, doi: https://doi.org/10.1016/j.jsb.2012.05.011.
39. Belviso, B. D., G. F. Mangiatordi, D. Alberga, V. Mangini, B. Carrozzini, and R. Caliandro. 2022. Structural Characterization of the Full-Length Anti-CD20 Antibody Rituximab. Frontiers in Molecular Biosciences. Volume 9 - 2022, doi: 10.3389/fmolb.2022.823174.
40. Büttner, H., M. Perbandt, T. Kohler, A. Kikhney, M. Wolters, M. Christner, M. Heise, J. Wilde, S. Weißelberg, A. Both, C. Betzel, S. Hammerschmidt, D. Svergun, M. Aepfelbacher, and H. Rohde. 2020. A Giant Extracellular Matrix Binding Protein of Staphylococcus epidermidis Binds Surface-Immobilized Fibronectin via a Novel Mechanism. mBio. 11(5):10.1128/mbio.01612-01620, doi: 10.1128/mbio.01612-20.
41. Watanabe, Y. 2019. Size-exclusion chromatography combined with solution X-ray scattering measurement of the heat-induced aggregates of water-soluble proteins at low ionic strength in a neutral solution. Journal of Chromatography A. 1603:190-198, doi: https://doi.org/10.1016/j.chroma.2019.06.042.
42. Chiu, M. L., D. R. Goulet, A. Teplyakov, and G. L. Gilliland. 2019. Antibody Structure and Function: The Basis for Engineering Therapeutics. Antibodies (Basel). 8(4), doi: 10.3390/antib8040055.
43. Chan, K. Y., B. Abu-Salih, R. Qaddoura, A. M. Al-Zoubi, V. Palade, D.-S. Pham, J. D. Ser, and K. Muhammad. 2023. Deep neural networks in the cloud: Review, applications, challenges and research directions. Neurocomputing. 545:126327, doi: https://doi.org/10.1016/j.neucom.2023.126327.
44. Singh, J., and R. Banerjee (2019). A Study on Single and Multi-layer Perceptron Neural Network. 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC).
45. Khanam, T., M. Afsar, A. Shukla, F. Alam, S. Kumar, H. Soyar, K. Dolma, M. Pasupuleti, K. K. Srivastava, R. S. Ampapathi, and R. Ramachandran. 2020. M. tuberculosis class II apurinic/apyrimidinic-endonuclease/3'-5' exonuclease (XthA) engages with NAD+-dependent DNA ligase A (LigA) to counter futile cleavage and ligation cycles in base excision repair. Nucleic acids research. 48(8):4325-4343, doi: 10.1093/nar/gkaa188.
46. Zhan, B., Y. Gao, W. Gao, Y. Li, Z. Li, Q. Qi, X. Lan, H. Shen, J. Gan, G. Zhao, and J. Li (2022). Structural insights of the elongation factor EF-Tu complexes in protein translation of Mycobacterium tuberculosis. Commun Biol.
47. Lorenz, C., S. Ince, T. Zhang, T. Zhang, A. Cousin, R. Batra-Safferling, L. Nagel-Steger, C. Herrmann, and A. M. Stadler. 2020. Farnesylation of human guanylate-binding protein 1 as safety mechanism preventing structural rearrangements and uninduced dimerization. The FEBS journal. 287(3):496-514, doi: 10.1111/febs.15015.
48. Sagar, A., N. Peddada, V. Choudhary, Y. Mir, R. Garg, and Ashish. 2024. Visualizing the nucleating and capped states of f-actin by Ca²⁺-gelsolin: Saxs data based structures of binary and ternary complexes. International journal of biological macromolecules. 278(Pt 1):134556, doi: 10.1016/j.ijbiomac.2024.134556.
49. Morishima, K., R. Inoue, T. Nakagawa, M. Shimizu, R. Sakamoto, T. Oda, K. Mayumi, and M. Sugiyama. 2025. Size-exclusion chromatography-small-angle neutron scattering system optimized for an instrument with medium neutron flux. Journal of Applied Crystallography. 58(2):595-602, doi: 10.1107/S1600576725000779.

50. Burger, V. M., D. J. Arenas, and C. M. Stultz. 2016. A Structure-free Method for Quantifying Conformational Flexibility in proteins. Scientific Reports. 6(1):29040, doi: 10.1038/srep29040.
51. Dey, M., A. Gupta, M. D. Badmalia, Ashish, and D. Sharma. 2025. Visualizing gaussian-chain like structural models of human α-synuclein in monomeric pre-fibrillar state: Solution SAXS data and modeling analysis. Int J Biol Macromol. 288:138614, doi: 10.1016/j.ijbiomac.2024.138614.
52. Kolarova, M., F. García-Sierra, A. Bartos, J. Ricny, and D. Ripova. 2012. Structure and Pathology of Tau Protein in Alzheimer Disease. International Journal of Alzheimer's Disease. 2012(1):731526, doi: https://doi.org/10.1155/2012/731526.
53. Kim, G., L. Azmi, S. Jang, T. Jung, H. Hebert, A. J. Roe, O. Byron, and J.-J. Song. 2019. Aldehyde-alcohol dehydrogenase forms a high-order spirosome architecture critical for its activity. Nature Communications. 10(1):4527, doi: 10.1038/s41467-019-12427-8.
54. Sierra-Gómez, Y., A. Rodríguez-Hernández, P. Cano-Sánchez, H. Gómez-Velasco, A. Hernández-Santoyo, D. Siliqi, and A. Rodríguez-Romero. 2019. A biophysical and structural study of two chitinases from Agave tequilana and their potential role as defense proteins. Febs j. 286(23):4778-4796, doi: 10.1111/febs.14993.
55. Bucciarelli, S., S. R. Midtgaard, M. Nors Pedersen, S. Skou, L. Arleth, and B. Vestergaard. 2018. Size-exclusion chromatography small-angle X-ray scattering of water soluble proteins on a laboratory instrument. Journal of Applied Crystallography. 51(6):1623-1632, doi: doi:10.1107/S1600576718014462.
56. Graewert, M. A., S. Da Vela, T. W. Gräwert, D. S. Molodenskiy, C. E. Blanchet, D. I. Svergun, and C. M. Jeffries. 2020. Adding Size Exclusion Chromatography (SEC) and Light Scattering (LS) Devices to Obtain High-Quality Small Angle X-Ray Scattering (SAXS) Data. Crystals. 10(11):975.
57. Ettich, J., C. Wittich, J. M. Moll, K. Behnke, D. M. Floss, J. Reiners, A. Christmann, P. A. Lang, S. H. J. Smits, H. Kolmar, and J. Scheller. 2023. Respiratory syncytial virus–approved mAb Palivizumab as ligand for anti-idiotype nanobody-based synthetic cytokine receptors. Journal of Biological Chemistry. 299(11), doi: 10.1016/j.jbc.2023.105270.
58. Alqaisi, M., J. F. Thümmler, F. Lehmann, F.-J. Schmitt, L. Lentz, F. Rieder, D. Hinderberger, and W. H. Binder. 2024. Tuning nanoparticles' internal structure: fluorinated single-chain nanoparticles (SCNPs) generated by chain collapse of random copolymers. Polymer Chemistry. 15(29):2949-2958, doi: 10.1039/D4PY00355A.
59. Arbe, A., J. A. Pomposo, A. J. Moreno, F. LoVerso, M. González-Burgos, I. Asenjo-Sanz, A. Iturrospe, A. Radulescu, O. Ivanova, and J. Colmenero. 2016. Structure and dynamics of single-chain nano-particles in solution. Polymer. 105:532-544, doi: https://doi.org/10.1016/j.polymer.2016.07.059.
60. Sanchez-Sanchez, A., and J. A. Pomposo. 2014. Single-Chain Polymer Nanoparticles via Non-Covalent and Dynamic Covalent Bonds. Particle & Particle Systems Characterization. 31(1):11-23, doi: https://doi.org/10.1002/ppsc.201300245.
61. Zeng, Y., T. Xu, X.-F. Hou, J. Liu, C. Liu, Z. Chang, J. Fang, and D. Chen. 2023. Enzyme Stabilization and Catalytic Activity Enhancement by Single-Chain Nanoparticles of Fluorinated Zwitterionic Random Copolymers. ACS Applied Polymer Materials. 5(5):3777-3791, doi: 10.1021/acsapm.3c00390.
62. Barbee, M. H., Z. M. Wright, B. P. Allen, H. F. Taylor, E. F. Patteson, and A. S. Knight. 2021. Protein-Mimetic Self-Assembly with Synthetic Macromolecules. Macromolecules. 54(8):3585-3612, doi: 10.1021/acs.macromol.0c02826.

63. Dill, K. A., and L. Agozzino. 2021. Driving forces in the origins of life. Open Biology. 11(2):200324, doi: doi:10.1098/rsob.200324.
64. Fonin, A. V., A. L. Darling, I. M. Kuznetsova, K. K. Turoverov, and V. N. Uversky. 2018. Intrinsically disordered proteins in crowded milieu: when chaos prevails within the cellular gumbo. Cellular and Molecular Life Sciences. 75(21):3907-3929, doi: 10.1007/s00018-018-2894-9.
65. Romero, P. A., A. Krause, and F. H. Arnold. 2013. Navigating the protein fitness landscape with Gaussian processes. Proc Natl Acad Sci U S A. 110(3):E193-201, doi: 10.1073/pnas.1215251110.
66. Narayanan, H., F. Dingfelder, I. Condado Morales, B. Patel, K. E. Heding, J. R. Bjelke, T. Egebjerg, A. Butte, M. Sokolov, N. Lorenzen, and P. Arosio. 2021. Design of Biopharmaceutical Formulations Accelerated by Machine Learning. Mol Pharm. 18(10):3843-3853, doi: 10.1021/acs.molpharmaceut.1c00469.
67. Frokjaer, S., and D. E. Otzen. 2005. Protein drug stability: a formulation challenge. Nature Reviews Drug Discovery. 4(4):298-306, doi: 10.1038/nrd1695.
68. Nápoles, G., N. Griffioen, S. Khoshrou, and Ç. Güven (2024). Feature Importance for Clustering. In V. Vasconcelos, I. Domingues, and S. Paredes, eds. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer Nature Switzerland.
69. Gräwert, T. W., and D. I. Svergun. 2020. Structural Modeling Using Solution Small-Angle X-ray Scattering (SAXS). Journal of Molecular Biology. 432(9):3078-3092, doi: https://doi.org/10.1016/j.jmb.2020.01.030.
70. Graewert, M. A., and D. I. Svergun. 2022. Chapter One - Advanced sample environments and sample requirements for biological SAXS. In Methods in Enzymology. J. A. Tainer, editor. Academic Press, pp. 1-39.
71. Receveur-Brechot, V., and D. Durand. 2012. How random are intrinsically disordered proteins? A small angle scattering perspective. Curr Protein Pept Sci. 13(1):55-75, doi: 10.2174/138920312799277901.