JOURNAL OF APPLIED CRYSTALLOGRAPHY
ISSN 1600-5767

Quantifying dispersity in size and shape of nanoparticles from small-angle scattering data using machine learning based CREASE

Rohan S. Adhikari,[a] Sri Vishnuvardhan Reddy Akepati,[b] Matthew R. Carbone,[c] Asritha Polu,[d] Hyeong Jin Kim,[e] Yugang Zhang[e] and Arthi Jayaraman[a,b]*

[a]Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, Delaware 19716, USA, [b]Department of Materials Science and Engineering, University of Delaware, Newark, Delaware 19716, USA, [c]Computing and Data Sciences Directorate, Brookhaven National Laboratory, Upton, New York 11973, USA, [d]Data Science Program, University of Delaware, Newark, Delaware 19716, USA, and [e]Center for Functional Nanomaterials, Brookhaven National Laboratory, Upton, New York 11973, USA. *Correspondence e-mail: arthij@udel.edu

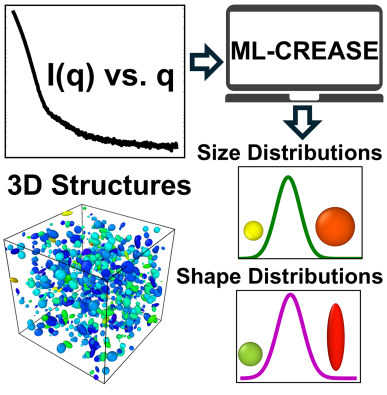
Received 18 April 2025
Accepted 24 June 2025

Edited by J. Ilavsky, Argonne National Laboratory, USA

Keywords: SAXS; small-angle X-ray scattering; machine learning; CREASE; particle size dispersity; particle shape dispersity.

Supporting information: this article has supporting information at journals.iucr.org/j

We use machine learning (ML) enhanced computational reverse engineering analysis of scattering experiments (CREASE) to interpret small-angle X-ray scattering (SAXS) data obtained from a system of nanoparticles without *a priori* knowledge of their exact shapes (*e.g.* spheres or ellipsoids), sizes (0.5–50 nm) and distributions. The SAXS measurements yielded three categories of scattering profiles exhibiting 'strong', 'weak' and 'no' features. Diminishing features (*e.g.* broadening or disappearing peaks) in scattering profiles have always been attributed to the presence of significant dispersity in the system. Such featureless SAXS data are not suitable for traditional analysis using analytical models. If one were to fit a relevant analytical model (*e.g.* the *lmfit* analytical model for polydisperse spheres) to these 'weak' and 'no' SAXS profiles from our nanoparticle systems, one would obtain non-unique interpretations of the data. Relying on electron microscopy to identify the distributions of nanoparticle shapes and sizes is also unfeasible, especially in high-throughput synthesis and characterization loops. In such situations, to identify the distributions of particle sizes and shapes that could be present in the sample, one must rely on methods like ML-CREASE to interpret the data quickly and output all relevant interpretations about the structure present in the system. The ML-CREASE optimization loop takes the experimental scattering profile as input and outputs multiple candidate solutions whose computed scattering profiles match the SAXS profile input. The ML-CREASE method outputs distributions of relevant structural features, such as the volume fraction of the nanoparticles in the system and the mean and standard deviation of the particle size and aspect ratio, assuming a type of distribution (*e.g.* normal, log-normal) for size and aspect ratio. We find that, for the SAXS profiles analyzed here, accounting for the shape dispersity along with size dispersity of the nanoparticles using ML-CREASE improved the match between the computed scattering profiles and input experimental profiles.



1. Introduction

Advances in particle synthesis methods have led to the availability of nanoparticles in a variety of shapes (Glotzer & Solomon, 2007). A few notable examples include the thin film stretching method (Ho *et al.*, 1993) to produce ellipsoidal particles, achieving aspect ratios between two and five, and the particle replication in non-wetting templates or PRINT method (Rolland *et al.*, 2005) to produce disc-like and rod-like nanoparticles (Kinnear *et al.*, 2017; Liu *et al.*, 2022). Compared with spherical nanoparticles, anisotropy in particle shape provides additional degrees of freedom for tuning the

interactions between the nanoparticles and their targets, which can then be leveraged in various applications, *e.g.* drug delivery (Shi *et al.*, 2017; Beach *et al.*, 2024), energy harvesting (Thorkelsson *et al.*, 2015) and chemical sensing (Zheng *et al.*, 2021). Shape anisotropy can also be used for directing assembly into complex structures with spatial arrangements for use in applications requiring certain optical and photonic properties (Wu & Pauly, 2022; Wang *et al.*, 2020). Advances in particle synthesis methods have increased the need for improved characterization techniques that identify the shapes and sizes of nanoparticles more accurately. Most researchers rely on microscopy techniques, *e.g.* transmission electron microscopy (TEM) (Carter & Williams, 2016; Fultz & Howe, 2012), scanning electron microscopy (SEM) (Goldstein *et al.*, 2017; Ul-Hamid, 2018) and atomic force microscopy (Voigtländer, 2019; Eaton & West, 2010), or scattering techniques, *e.g.* small-angle X-ray scattering (SAXS) (Lindner & Oberdisse, 2024; Narayanan *et al.*, 2017), to identify the size and shape distributions of nanoparticles (Modena *et al.*, 2019; Mourdikoudis *et al.*, 2018). Although microscopy techniques (SEM and TEM) provide high-resolution scans of the nanoparticles, these 2D images lack depth-related information (Li *et al.*, 2016; Dawadi *et al.*, 2021) and sample preparation can be tedious, making it less practical in high-throughput automated synthesis and characterization loops. On the other hand, scattering techniques measure the ensemble-averaged information about the structure across multiple length scales and are also amenable to high-throughput automation loops (Quek *et al.*, 2023; Dyer *et al.*, 2014; Rodríguez-Ruiz *et al.*, 2017).

While scattering techniques are suitable for high-throughput characterization, the interpretation of scattering data can be non-trivial (Jeffries *et al.*, 2021; Yager *et al.*, 2023). Structural information obtained through scattering techniques is in reciprocal space (*i.e.* the intensity of the scattered wave versus the magnitude of the wavevector). Interpreting these data generally requires analytical model fitting, computational methods and/or machine learning algorithms. In analytical model fitting, the user selects relevant theoretical models, *e.g.* the hard sphere (Blum & Stell, 1979; Salacuse & Stell, 1982) or sticky hard sphere models (Menon *et al.*, 1991), to fit experimental data by finding model parameters that minimize the difference between the experimental scattering profile and the analytical model. Analytical models for traditional geometries of soft materials and their assemblies have been collated into user-friendly packages such as *SASfit* (Breßler *et al.*, 2015), *Irena* (Ilavsky & Jemian, 2009), *McSAS* (Bressler *et al.*, 2015) and *ATSAS* (Petoukhov *et al.*, 2012; Manalastas-Cantos *et al.*, 2021; Franke *et al.*, 2025). Other model-agnostic methods include Guinier analysis, which is used to determine the radius of gyration of particles in a sample by plotting the logarithm of the scattering intensity against the square of the magnitude of the scattering vector (Guinier, 1955). Similarly, Porod analysis (Porod, 1951), applied to the high-q region of the scattering profile, provides information about the surface area and interface roughness of the particles (Schmidt, 1988). Users can also use Fourier transform methods to convert scattering data from reciprocal space to real space, providing direct information about the size and shape of the scattering objects (Schmidt-Rohr, 2007; Röding *et al.*, 2022). Computational methods like reverse Monte Carlo (RMC) (McGreevy & Pusztai, 1988) simulations have also been used to iterate towards structures whose computed scattering patterns match the experimental data (McGreevy, 1995; McGreevy, 2001). Studies that utilize the RMC technique for the interpretation of scattering profiles include the analysis of ultra-small-angle scattering data to obtain the representative 3D configurations of silica nanoparticles in a rubber matrix (Hagita *et al.*, 2018), analysis of 2D small-angle scattering data to obtain the representative 3D configurations and orientations of magnetic nanoparticles (Barnsley *et al.*, 2022), analysis of X-ray scattering data to obtain atomistic configurations of liquid mercury near its critical point (Hagita *et al.*, 2010), and analysis of neutron scattering data to obtain representative 3D configurations of polymer grafted nanoparticles using the *MONSA* program for RMC (Luo *et al.*, 2018). RMC alleviates some of the drawbacks of analytical model fits by providing a representative 3D structure as an output but suffers from low computational efficiency when the density of soft materials or their assemblies is expected to be high.

To accelerate and potentially automate the interpretation of scattering profiles, researchers have begun to turn to machine learning (ML) methods; we encourage readers to consult the references cited in the recent review articles on this topic (Anker *et al.*, 2023; Lu & Jayaraman, 2024). One such method is the ML enhanced computational reverse engineering analysis of scattering experiments (ML-CREASE) which has been successfully used to interpret structures from scattering profiles for a variety of materials – polymer solutions (Wu & Jayaraman, 2022; Ye *et al.*, 2021; Wessels & Jayaraman, 2021*b*; Beltran-Villegas *et al.*, 2019), surfactant-coated particles (Heil *et al.*, 2023*a*), nanoparticle mixtures (Heil *et al.*, 2023*b*; Heil *et al.*, 2022; Heil & Jayaraman, 2021), biomolecular networks (Lee *et al.*, 2020) and dipeptide solutions (Gupta *et al.*, 2025). In many of these cases, traditional analytical model fits had failed either because the models available were too approximate for the system at hand or because the material's structure had significant dispersity in dimensions for which the analytical models perform poorly. While earlier implementations of CREASE were used to analyze azimuthally averaged scattering profiles, recent extension of this method has led to the CREASE-2D method, which interprets the entire 2D scattering profiles without any azimuthal angle averaging to assess the extent of structural anisotropy in addition to other relevant information about the structure (Akepati *et al.*, 2024; Gupta *et al.*, 2025).

ML-CREASE's interpretation of the scattering profiles provides a detailed understanding of the form and structure of the assembled materials as distributions of mathematical parameters that describe relevant structural features. ML-CREASE also provides as output representative 3D real-space structures for various structural features which in turn can be used for other analyses (*e.g.* structure-induced property calculation) (Heil *et al.*, 2023*b*; Patil *et al.*, 2022*a*; Patil *et al.*,

2022*b*). This would be impossible with analytical models that do not provide real-space candidate 3D structures. While RMC simulations can also provide 3D structures as output, in that approach the user is optimizing one real-space structural configuration at a time, which can lead to a single locally optimized (perhaps not globally optimized) 3D representation as the interpretation. The use of genetic algorithms in the ML-CREASE approach provides multiple possible structural interpretations of the scattering profiles (*i.e.* degenerate solutions) (*e.g.* Lee *et al.*, 2020; Wessels & Jayaraman, 2021*a*). These multiple real-space structural interpretations can then be compared with information from other measurements (structure, property) or molecular simulations to identify which of the multiple answers that ML-CREASE outputs are physically possible and which are numerically correct but unphysical.

In this work, we use the ML-CREASE method to analyze the distribution of sizes and shapes of gold nanoparticles from azimuthally averaged 1D SAXS profiles. Gold nanoparticles were synthesized using an automated fluidic platform based on the Turkevich method (Wuithschick *et al.*, 2015) and subsequently characterized *in situ* by SAXS to probe their nanoscale morphology at the National Synchrotron Light Source II (NSLS-II) at Brookhaven National Laboratory. The resulting nanoparticle SAXS profiles were then classified into distinct categories based on the presence or absence of pronounced features (peaks) as 'strong', 'weak' and 'no' profiles. In our work presented in this paper, we find that the ML-CREASE method is capable of predicting the size and shape distributions of nanoparticles even when the scattering profiles lack any pronounced features; as noted before, the analysis of such featureless scattering profiles has proven difficult with traditional fitting approaches using analytical models. The predictions from ML-CREASE include the extent of dispersity in sizes (*e.g.* spherical volume radius) as well as dispersity in shapes (aspect ratio) ranging from spherical (aspect ratio $\sim$1.0) to ellipsoidal (mean aspect ratio >1.0). Our results also suggest that selecting analytical models assuming all the nanoparticles are spherical is too restrictive and including shape dispersity leads to improvement in fits to SAXS profiles. This work shows the power of the ML-CREASE method to predict the distributions in sizes and shapes of nanoparticles from their 1D scattering profiles without *a priori* knowledge of the shape and size distributions, which are only accessible via imaging techniques that are incompatible with high-throughput SAXS characterization.

The article is structured as follows. First, the synthesis of nanoparticles, SAXS characterization and steps involved in ML-CREASE are presented in Section 2. Next, the distributions of nanoparticle shapes and sizes predicted by the ML-CREASE approach for the SAXS inputs are discussed in Section 3. Finally, in Section 4 we conclude by summarizing the capabilities of the ML-CREASE method and how it can be broadly applied to problems within the community. All code used in this work is hosted on GitHub (https://github.com/arthijayaraman-lab/CREASE_Size_Shape_Dispersity) and is freely available for use by the scientific community.

2. Methods

2.1. Experiments

2.1.1. Materials and preparation of gold nanoparticles

Sodium citrate (NCit), hydrogen chloride (HCl), sodium hydroxide (NaOH), Tween 20 and chloroauric acid (HAu) were purchased from Sigma–Aldrich, and used as received without further purification. Reagent solutions – 16 m*M* NCit, 0.01 wt% Tween in deionized water, 10 m*M* HCl, 10 m*M* NaOH and 2 m*M* HAu – were precisely injected using automated syringes (precision <1 µL) through selection valves into the main flow path, before undergoing well mixing by static mixer. The reaction was conducted at 100°C, and the as-synthesized products were analyzed after 10 min of reaction time.

2.1.2. Protocol for small-angle scattering experiments

Samples containing nanoparticles in solution were subjected to SAXS characterization at the Soft Matter Interfaces (SMI, 12-ID) and Complex Materials Scattering (CMS, 11-BM) beamlines at NSLS-II. At the SMI beamline, SAXS data were collected using a beam energy of 16.1 keV and beam size of 200 × 30 µm with a Pilatus 1M area detector (Dectris, Switzerland). The detector, consisting of 0.172 mm square pixels in a 981 × 1043 array, was placed 5 m downstream from the sample position. At the CMS beamline, SAXS data were collected using a beam energy of 13.5 keV and beam size of 200 × 200 µm with a Pilatus 2M area detector (Dectris, Switzerland). The detector, comprising 0.172 mm square pixels in a 1475 × 1679 array, was positioned 5 m downstream from the sample. Scattering patterns from each detector angle were stitched together using custom-developed software. Typical exposure times were 1 s at the SMI beamline and 15 s at the CMS beamline. The 2D SAXS patterns, collected continuously during synthesis, were reduced to 1D scattering intensity, $I(q)$, through real-time circular averaging. Here, q represents the wavevector transfer, $q = (4\pi/\lambda)\sin(\theta)$, where $\lambda = 0.77$ Å is the X-ray wavelength and 2θ is the scattering angle. Scattering angles were calibrated using silver behenate as the standard.

2.1.3. Description of the SAXS data

The scattering intensities were azimuthally averaged after background subtraction to obtain a total of 30 profiles with $I(q)$ as a function of q in the range of [0.02–0.18] Å^{-1}. On the basis of the presence or absence of characteristic peaks (features) in these profiles, the 30 scattering profiles are manually classified as profiles with 'strong', 'weak' and 'no' features. Representative examples of the 'strong', 'weak' and 'no' scattering profiles are shown in Fig. 1.

2.2. ML-CREASE

Taking one SAXS profile at a time as input, we aim to identify the corresponding shape and size distributions for the nanoparticles in the system via the ML-CREASE method.

We first assume that all nanoparticles are spherical [Figs. 2(*a*)–2(*c*)] with dispersity in size. In this case, the output from

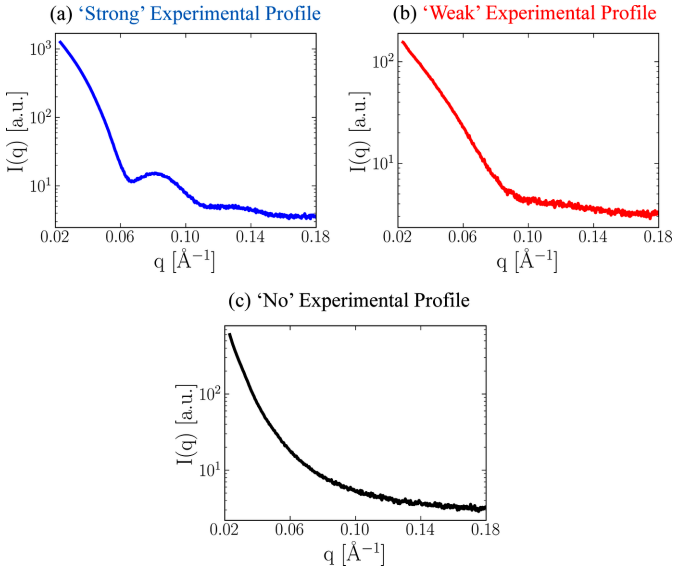


Figure 1
The three classes of experimental scattering profiles are manually labeled as 'strong', 'weak' and 'no' profiles according to the presence or absence of pronounced features (peaks) in that profile. Representative example of a (*a*) 'strong' experimental profile, (*b*) 'weak' experimental profile and (*c*) 'no' experimental profile.

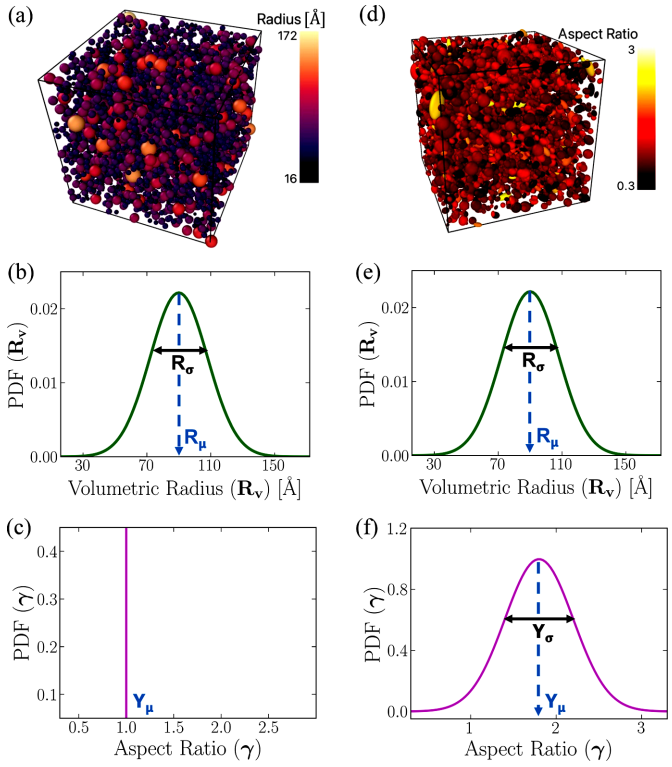


Figure 2
(*a*) Representative 3D structure for a nanoparticle system with dispersity only in particle size. (*b*) The probability density function (normal distribution) of the volumetric radii of the nanoparticles for a system with dispersity only in nanoparticle size. (*c*) The probability density function (Dirac delta distribution) of the aspect ratio of the nanoparticles for a system with dispersity only in nanoparticle size. (*d*) Representative 3D structure for a nanoparticle system with dispersity in both particle size and shape. (*e*) The probability density function (normal distribution) of the volumetric radii of the nanoparticles for a system with dispersity in both nanoparticle size and shape. (*f*) The probability density function (normal distribution) of the aspect ratio of the nanoparticles for a system with dispersity in both nanoparticle size and shape.

our analysis using ML-CREASE will be the mean and standard deviation for the radii of the spherical nanoparticles (assuming a normal distribution). Next, we relax the spherical assumption and expect dispersity in both particle size and particle shape [Figs. 2(*d*)–2(*f*)]. In this case, the output from our analysis using ML-CREASE will be the mean and standard deviation of the nanoparticle's size, denoted as 'equivalent sphere' radius and nanoparticle aspect ratio, assuming a normal distribution for both radius and aspect ratio. In addition to the above structural features (*e.g.* mean and standard deviation of size and aspect ratio), ML-CREASE outputs a few representative real-space 3D structures of the nanoparticles [as shown in Fig. 2(*a*) and Fig. 2(*d*)]. Users can obtain as many representative 3D structures as they wish using the previously published open-source *CASGAP* method (Gupta & Jayaraman, 2023) which takes as input structural features (nanoparticle size and shape distributions) and outputs a real-space 3D structure.

To use ML-CREASE to interpret scattering profiles, we have to follow these steps:

(i) Identify relevant structural features.

(ii) Create 3D real-space representations of structures for systematically varied values of the structural features; in this work we uniformly sample structural features within relevant pre-defined ranges.

(iii) Compute 1D scattering profiles for every 3D structural representation generated in step (ii).

(iv) Train and test an ML model on the dataset of input structural features and output 1D scattering profiles generated in step (iii).

(v) Embed the trained ML model in the genetic algorithm of the ML-CREASE method to identify sets of structural features whose computed scattering profiles match the experimental target.

2.2.1. Identifying relevant structural features

For the system of nanoparticles with dispersity in both particle size and particle shape, two structural features are required to describe the distribution of particle size (assuming a normal distribution in particle size), and two structural features are required to describe the distribution of particle shape (assuming a normal distribution in particle shape). Additionally, a structural feature is required to describe the degree of packing in the nanoparticle system.

For the case where we assume that the nanoparticles are spherical, with dispersity only in particle size, as discussed above, only the mean and standard deviation of the radii, and the volume fraction of nanoparticles in the system, are required as structural features. For the case where we expect dispersity in shape and size of the nanoparticles, we assume the nanoparticles can be spheres or other non-spherical ellipsoids. The volumetric radius of an ellipsoidal particle is

defined as the radius of a sphere that has the same volume as the ellipsoidal particle (Akepati *et al.*, 2024; Gupta & Jayaraman, 2023). For a sphere, the volumetric radius and the radius are interchangeable.

To define the shape of an ellipsoidal particle, one may choose three mutually perpendicular semi-axial lengths a, b and c. We assume that two of the three mutually perpendicular semi-axial lengths are equal ($a = b$). This restricts the shape anisotropy of the ellipsoids to be along one of the three semi-axial lengths (c). As such, the aspect ratio ($\gamma = c/a$) is used as a structural feature to describe shape. For spheres, the value of $\gamma = 1$.

For both cases, we use the volume fraction (ϕ), defined as a ratio of the total volume of all the nanoparticles in the system to the volume of the entire system, to describe the extent of crowding among the nanoparticles in the system.

In summary, for a system of spherical nanoparticles with dispersity only in particle size, the three structural features are the mean volumetric radius (R_μ), the standard deviation of the volumetric radii (R_σ) and the volume fraction (ϕ). For a system of nanoparticles with dispersity in both particle size and shape, the five structural features include R_μ, R_σ and ϕ, in addition to the mean aspect ratio (γ_μ) and the standard deviation of the aspect ratio (γ_σ).

2.2.2. Creating 3D structures for systematically varied values of the structural features

The next step is to generate a dataset of 3D representations for systematically varied values of structural features within defined ranges. We use the *CASGAP* method (Gupta & Jayaraman, 2023) to generate representative 3D structures corresponding to desired values of mean and standard deviation of radius and aspect ratio; the values of radius and aspect ratio are sampled from the truncated normal distribution for this study. The range of structural features can be defined by experiments (*e.g.* possible maximum and minimum sizes of nanoparticles) and to some extent by preliminary manual matching and sensitivity analysis procedures (*e.g.* to identify how values of structural features affect the scattering profile). For some examples of the manual matching analysis, see Section S1 in the supporting information. For some examples of the sensitivity analysis see Section S2.

For spherical nanoparticles with dispersity only in size, we obtain predictions from ML-CREASE for all three ('strong', 'weak' and 'no') classes of scattering profiles. We generate 3000 structures of nanoparticles with polydispersity only in size by systematically varying R_μ, R_σ and ϕ. Fig. 3(*a*) depicts the uniformly sampled values used to create these 3000 structures. R_μ was sampled in the ranges [5–20) Å, [20–100) Å and [100–500) Å (1000 samples each). All values of R_σ were sampled such that R_σ/R_μ was between 0 and 1. All values of ϕ were sampled from the range [0.05–0.15). We then use *CASGAP* to generate 3D real-space structures for all sampled values. The number of nanoparticles in a 3D structure generated using *CASGAP* must be large enough to capture the nanoparticle size and shape distributions well. At the same time, the number of nanoparticles must not be so large that the computation of scattering profiles [in step (iii) of the ML-CREASE method] becomes too slow. To obtain a reasonably sized 3D real-space structure that captures the size distributions of the spherical nanoparticles without requiring a computationally exhaustive scattering computation, we use three different box lengths to sample the 3000 real-space

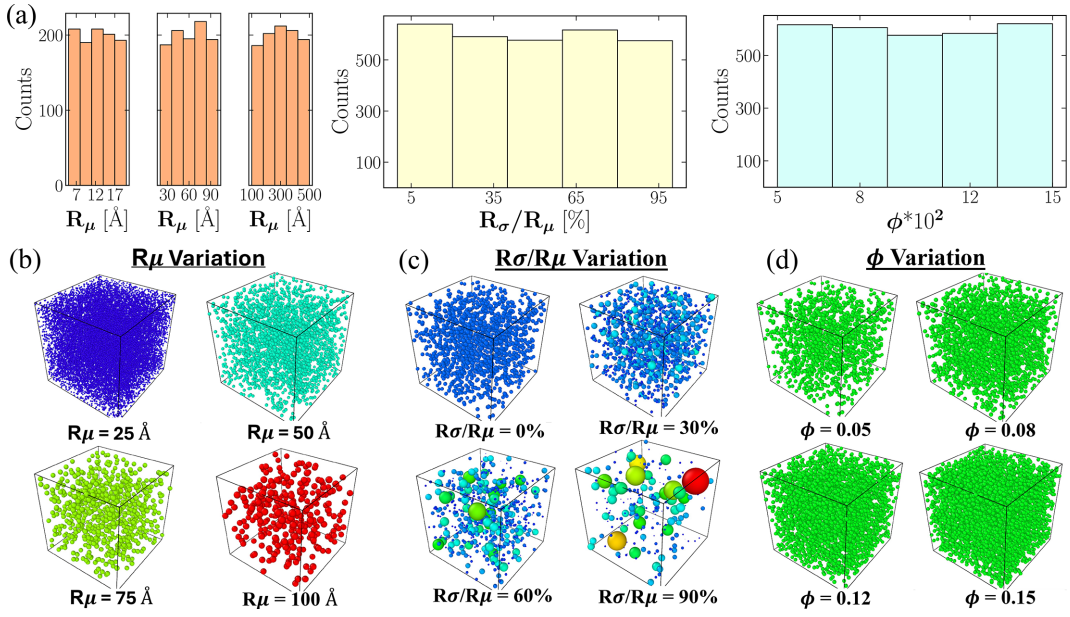


Figure 3
Structural features of the system [identified in step (i) of the ML-CREASE method] varied using a uniform distribution to generate 3D real-space structural representations. (*a*) Histograms for mean volumetric radius (R_μ), standard deviation of the volumetric radii (R_σ/R_μ) and volume fraction (ϕ) show the distribution of each structural feature in the 3000 sets of structural features studied. 3D real-space structural representations that vary (*b*) only in R_μ, (*c*) only in R_σ/R_μ and (*d*) only in ϕ. Nanoparticles in (*b–d*) are color-coded with respect to their radii.

structures of spherical nanoparticles with dispersity only in particle size. For R_μ in the range [5, 20), [20, 100) and [100, 500) Å, the box length in *CASGAP* was set to 600, 3000 and 15000 Å, respectively. In Figs. 3(*b*)–3(*d*) we present a few representative 3D structures.

We sample R_σ/R_μ from the range [0, 1) even though we find (during sensitivity analysis) that the R_σ structural feature is inconsequential to the computed scattering profile when it is greater than 50% of R_μ (see Section S2 for details). This range for R_σ/R_μ is chosen to make a one-to-one comparison between the predictions from ML-CREASE and the predictions from the *lmfit* analytical model for polydisperse spheres (Newville *et al.*, 2024). The *lmfit* model always predicts R_σ/R_μ greater than 0.5 for the 'no' scattering profiles. The consequences of including an unrestricted range for R_σ/R_μ on the predictions of ML-CREASE are discussed in the results section.

For nanoparticles with dispersity in size and shape, we restrict our analysis to the 'strong' and 'weak' classes of experimental profiles. After the manual matching procedure and sensitivity analysis, we generate another 3000 structures with systematic variation in the five structural features – mean volumetric radius (R_μ), standard deviation of the volumetric radii (R_σ), mean aspect ratio (γ_μ), standard deviation of the aspect ratio (γ_σ) and volume fraction (ϕ). R_μ is sampled from the range [20, 100) Å, R_σ/R_μ is sampled from the range [0, 0.5), γ_μ is sampled from the range [0.8, 4), γ_σ/γ_μ is sampled from the range [0, 0.5) and ϕ is varied in the range [0.05, 0.15). A box length of 3000 Å is used in *CASGAP* to generate all 3D representations for this case. The histogram in Fig. 4(*a*) depicts the variation of the two additional structural features that account for the dispersity in particle shape, mean aspect ratio (γ_μ) and standard deviation of the aspect ratio (γ_σ). The differences in the 3D representations as one of the structural features is varied while the other four structural features are held constant are shown in Figs. 4(*b*)–4(*c*).

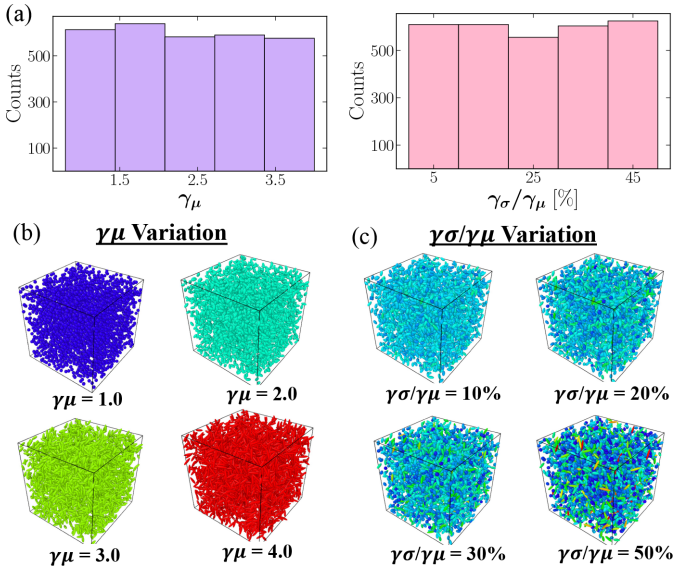


Figure 4
Mean aspect ratio and standard deviation of aspect ratio are additional structural features identified for nanoparticles with dispersity in both size and shape, beyond the three structural features identified for nanoparticles with dispersity only in particle size. (*a*) Histograms for mean aspect ratio (γ_μ) and standard deviation of the aspect ratio (γ_σ/γ_μ) show the distribution of the two structural features in the 3000 sets of structural features studied. (*b*) *CASGAP*-generated 3D representations that vary only in the structural feature γ_μ. (*c*) *CASGAP*-generated 3D representations that vary only in the structural feature γ_σ/γ_μ. Nanoparticles in (*b*–*c*) are color-coded with respect to their aspect ratios.

2.2.3. Computing 1D scattering profiles from 3D real-space structural representations

CASGAP outputs the coordinates and aspect ratios for each nanoparticle into a format that is convenient for visualization with programs such as *OVITO* (Stukowski, 2009). Once the coordinates and aspect ratios for all 3D representations have been collected, they can be used to calculate the scattering profile using the scattering equation (Guinier, 1955; Glatter, 1979; Brisard & Levitz, 2013).

The computation of scattering profiles is much faster when the complex scattering amplitudes $[A_{\rm comp}(\mathbf{q})]$ are first computed instead of directly computing the scattering intensities $[I_{\rm comp}(q)]$ using the Debye scattering equation (Akepati *et al.*, 2024; Brisard & Levitz, 2013). The equation we use to calculate the complex scattering amplitudes of the nanoparticles is

$$A_{\rm comp}(\mathbf{q}) = \sum_{n=1}^{N} \Delta\rho_n v_n f_n(\mathbf{q}) \exp(-i\mathbf{q}\cdot\mathbf{r}_n). \quad (1)$$

$A_{\rm comp}(\mathbf{q})$ can be understood as the Fourier transform of the scattering length density contrast ($\Delta\rho_n$) of the nanoparticles. $f_n(\mathbf{q})$ is the analytical form factor of the nanoparticles. For ellipsoids, $f_n(\mathbf{q})$ can be calculated using the analytical form factor expression reported in Pedersen's tabulation of analytical form factors (Pedersen, 1997). For a detailed explanation of the analytical form factor expression for ellipsoids and its implementation, we refer the reader to the work of Akepati *et al.* (2024).

Box length corrections are applied to the computed scattering amplitudes following the work of Brisard & Levitz (2013). Once the box length corrections are applied to the computed scattering amplitudes, the scattering intensities $[I_{\rm comp}(q)]$ of the nanoparticles can be computed in a straightforward manner, as shown in equation (2) where θ is azimuthal angle:

$$I_{\rm comp}(q) = \frac{1}{V}\left\langle \left|A_{\rm comp}(\mathbf{q})\right|^2 \right\rangle_\theta. \quad (2)$$

Another advantage of computing the scattering profiles using the complex amplitudes instead of the Debye scattering equation (Cantor & Schimmel, 1980; Svergun *et al.*, 2013) is that the full 2D projection of the scattering profile (as is measured experimentally) can be computed without azimuthal averaging. This becomes crucial for systems with orientational anisotropy, where the 2D scattering profile contains valuable information that could be lost through azimuthal averaging. The Debye scattering equation, on the other hand, can only compute azimuthally averaged scattering profiles. For

this work, as we do not consider orientational anisotropy and are working with 1D scattering data from experiments, we average the scattering profiles azimuthally [$\langle \ldots \rangle_\theta$, where θ is azimuthal angle in equation (2)] to obtain 1D scattering profiles.

Using the methods above, we calculate the computed scattering profiles for all the structures generated for each case in the previous step.

2.2.4. ML model training to compute $I(q)$ from the structural features

To create a surrogate ML model that links structural features to computed scattering profiles, we make use of the datasets generated in steps (ii) and (iii). We choose the eXtreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016) surrogate ML model for use in ML-CREASE. Although deep learning and neural networks are popular for image and language processing applications, the XGBoost method is generally known to work better for tabular data (Song *et al.*, 2020; Choi, 2019; Shwartz-Ziv & Armon, 2022). Furthermore, conventional deep learning models exhibit limited capacity to generate diverse structural solutions without extensive training data (>10^4 samples), a requirement prohibitive for many experimental scattering studies (Elasri *et al.*, 2022; Shrestha & Xie, 2023).

From our dataset containing 3000 sets of structural features and the corresponding scattering profiles, we set aside 80% of the dataset for training and 20% for testing. We repeat this process for the datasets collected for the first case of spherical nanoparticles with polydispersity in size and for the second

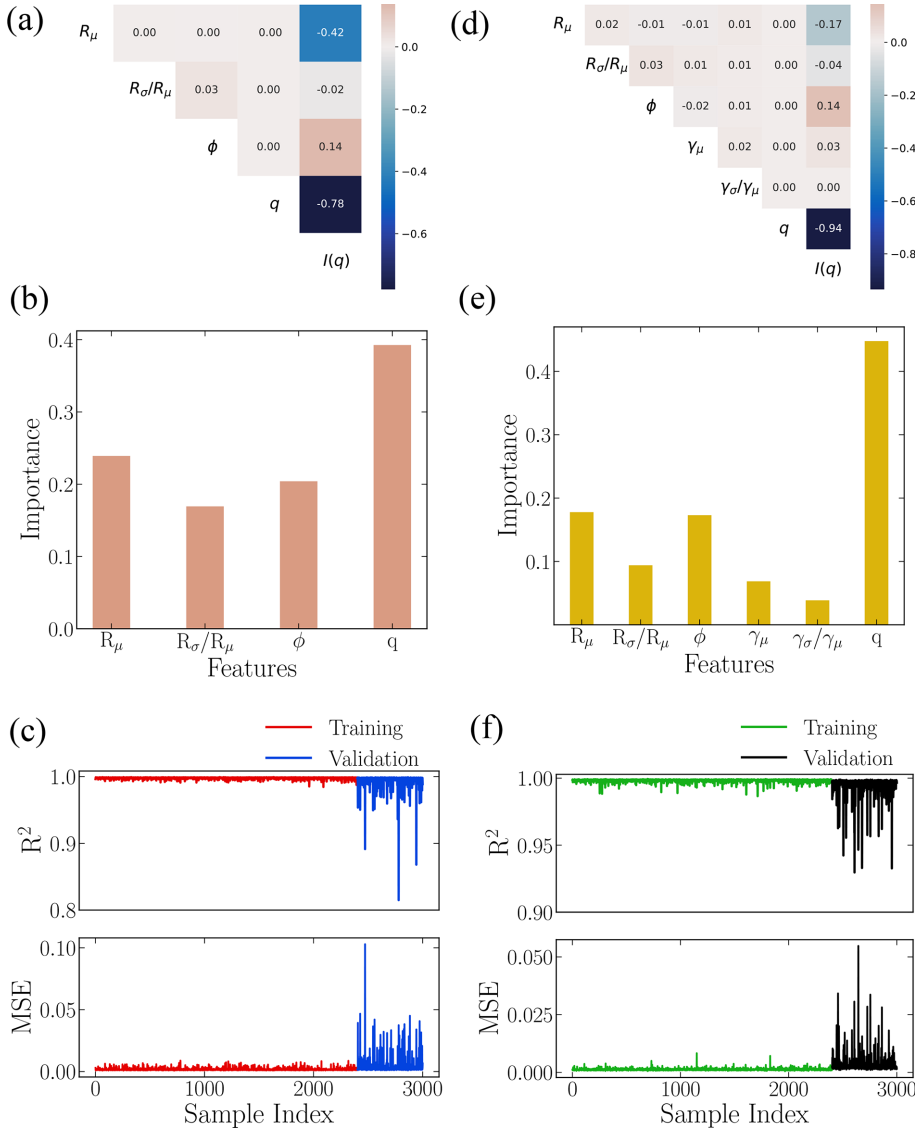


Figure 5
(*a*) Pearson correlation matrix of the dataset generated for spherical nanoparticles with dispersity only in particle size. (*b*) Feature importance values assigned by the XGBoost model from the training set for spherical nanoparticles with dispersity only in particle size. (*c*) Comparing the performance of the XGBoost ML model on the training and test sets using the R^2 and MSE metrics (spherical nanoparticles with dispersity in size). (*d*) Pearson correlation matrix of the dataset generated for nanoparticles with dispersity in particle size and shape. (*e*) Feature importance values assigned by the XGBoost ML model from the training set for nanoparticles with dispersity in particle size and shape. (*f*) Comparing the performance of the XGBoost ML model on the training and test sets using the R^2 and MSE metrics (nanoparticles with dispersity in size and shape).

Table 1
Comparison of the average R^2 and average MSE metrics during the XGBoost ML model training and testing for the two cases: nanoparticles with dispersity only in particle size and nanoparticles with dispersity in both particle size and shape.

The average R^2 and average MSE from the training sets are calculated from the 2400 training samples. The average R^2 and average MSE from the test sets are calculated from the 600 test samples.

Dataset	Average R^2 (training set)	Average R^2 (test set)	Average MSE (training set)	Average MSE (test set)
Nanoparticles with dispersity only in particle size	0.997	0.994	0.002	0.005
Nanoparticles with dispersity in both particle size and shape	0.998	0.994	0.001	0.003

case of nanoparticles with dispersity in both particle size and particle shape; these lead to two distinct ML models for the two cases. The training dataset for both cases is formatted into a tabular form that includes the values of structural features, the q value and the corresponding computed scattering profiles [$I(q)$]. The training dataset for spherical nanoparticles with dispersity only in size contains 2234400 rows (2400 training samples $\times$ 931 q values) and five columns [three structural features, q and $I(q)$ entries in every row]. The training dataset for nanoparticles with dispersity in size and shape also contains 2234400 rows (2400 training samples $\times$ 931 q values) and seven columns [five structural features, q and $I(q)$ entries in every row].

Using the respective tabular training sets, we employ Bayesian optimization (Thebelt *et al.*, 2022) to identify the optimal set of hyperparameters for the XGBoost regressor. During Bayesian optimization, the hyperparameters related to the architecture of the decision trees in the XGBoost model (*e.g.* learning rate, maximum depth of decision tress) are optimized to minimize the cross-validation error while also avoiding overfitting. The optimized hyperparameters for the two XGBoost ML models are shown in Section S3. After hyperparameter tuning, we train the two XGBoost ML models on the two training sets for the two cases – spherical nanoparticles with dispersity only in particle size and nanoparticles with dispersity in particle size and shape.

In Fig. 5(*a*), we show the Pearson correlation matrix analyzed using the dataset from spherical nanoparticles with dispersity only in particle size. The correlation matrix shows the strongest correlation between the intensity value (I) and q; this is not surprising as the scattering intensities are a function of q by definition. The R_σ/R_μ structural feature shows the weakest correlation with the intensity. As mentioned previously, during sensitivity analysis we find that with increasing values of R_σ/R_μ the scattering profiles become increasingly featureless, and when R_σ/R_μ is larger than 0.5, the effect on scattering profiles is minimal. Therefore, the weak correlation in Fig. 5(*a*) between R_σ/R_μ and intensity is not surprising. The importance assigned to the three structural features (R_μ, R_σ/R_μ and ϕ) and q by the trained XGBoost model for the prediction of intensity is shown in the feature importance plots in Fig. 5(*b*). In Fig. 5(*c*), the R^2 and mean squared error (MSE) scores are used to quantify the performance of the XGBoost ML model on the training and test sets. The R^2 and MSE metrics for the training and test samples are obtained by comparing the $I(q)$ predictions of the XGBoost ML model with the corresponding $I(q)$ computed from the scattering equation [step (iii) of the ML-CREASE method] for that sample. The 'Sample Index' in Fig. 5(*c*) is not the same as the 'Sample ID'. The 'Sample ID' for the training and test samples is chosen at random to obtain 2400 samples for training and 600 samples for testing. 'Sample Index' is used in Fig. 5(*c*) for the ease of differentiating the performance of the XGBoost model on the training and test samples. The average R^2 and average MSE of the XGBoost ML model (for spherical nanoparticles with dispersity only in particle size) during the training and testing are sufficiently close, as shown in Table 1.

In Fig. 5(*d*), we show the Pearson correlation matrix analyzed using the dataset from nanoparticles with dispersity in size and shape. The correlation matrix once again shows the strongest correlation between the intensity value $I(q)$ and q, as expected. The two additional structural features γ_μ and γ_σ/γ_μ, which relate to the shape distribution of the nanoparticles, show the weakest correlation with the intensity value, $I(q)$. This is because, in orientationally disordered systems with many particles, the information related to the shapes of the nanoparticles gets averaged out while computing the ensemble-averaged 1D scattering profiles. This does not imply that methods seeking to interpret 1D scattering profiles can choose to ignore the shape-related information and assume spherical nanoparticles. If anything, the averaging out of shape-related information in 1D scattering profiles drives scattering methods to consider multiple shapes and provide the user with many/all interpretations of the input profile. The feature importance assigned to the five structural features and q by the trained XGBoost model for the prediction of the intensity $I(q)$ is shown in Fig. 5(*e*). In Fig. 5(*f*), the R^2 and MSE scores are used to quantify the performance of the XGBoost ML model for the training and test samples. The average R^2 and average MSE of the XGBoost ML model (for nanoparticles with dispersity in both particle size and shape) during the training and testing are sufficiently close, as shown in Table 1.

After training, the XGBoost ML model can take the structural features (R_μ, R_σ/R_μ and ϕ) and the q values as input (in tabular form) and predict the corresponding scattering intensity values [$I(q)$]. Analogously, the XGBoost ML model for nanoparticles with dispersity in particle size and shape is capable of taking the structural features (R_μ, R_σ/R_μ, γ_μ, γ_σ/γ_μ and ϕ) and the q values as an input (in tabular form) and predicting the corresponding scattering intensity values [$I(q)$]. After establishing the forward mapping from structural features to $I(q)$ for the two cases, we move onto the next step of the ML-CREASE method, which is using the genetic algorithm (GA) optimization loop to identify sets of structural features whose computed scattering profiles closely match the input experimental profile.

2.2.5. CREASE's genetic algorithm (CREASE-GA)

With the ML model serving as a forward model linking structural features to the computed scattering profile, one can use a variety of optimization algorithms to solve the inverse problem of identifying the structural features for a given experimental scattering profile. We prefer the use of GAs in CREASE as they identify multiple optimal solutions (sets of structural features) for a given input scattering profile. By incorporating the trained XGBoost ML model into the CREASE-GA, the interpretation of the input scattering profiles (1D) can be achieved in much less time.

There are various types of GAs, such as the continuous-parameter GA and binary GA (Mitchell, 1998; Holland, 1992). In our work, we use a continuous-parameter GA which is better suited for the evolution of 'genes'. In ML-CREASE, each structural feature is mapped onto a 'gene' and the complete set of structural features identified in step (i) of the ML-CREASE method forms an 'individual'. For the case of spherical nanoparticles with dispersity in particle size, each 'individual' has an assigned value for the three structural features. For the case of nanoparticles with dispersity in both particle size and shape, each 'individual' has an assigned value for five structural features. In both cases, the GA loop has 100 'individuals' in each 'generation'.

For each individual, the XGBoost ML model takes as input the values of the structural features and predicts the computed scattering profile. The fitness of the individual is calculated on the basis of how closely the computed scattering profile matches the input experimental profile. We use the weighted sum of log squared errors (SSE) as implemented by Wu & Jayaraman (2022) to evaluate the fitness of each individual, as shown below:

$$\mathrm{SSE} = \sum_{n=1}^{N} w_i \left\{ \log\left[\frac{I_{\mathrm{exp}}(q_i)}{f I_{\mathrm{comp}}(q_i) + c} \right] \right\}^2, \tag{3}$$

where $w_i = \log(q_i/q_{i-1})$. Since the experimental profile is in arbitrary units, the computed scattering profiles predicted by the XGBoost model need to be uniformly scaled by a factor f during the fitness evaluation. In addition, experimental profiles obtained through background subtraction can have minor uncertainties. We use the parameter c to capture these uncertainties due to background subtraction. The values for f and c for a GA individual are obtained such that they minimize the SSE between the computed scattering profile for that individual and the experimental profile.

If the value of SSE is high, it implies a poor match and low fitness; alternatively, a low value of SSE denotes a good match and a high fitness for that individual. In each iteration of the CREASE-GA, the fitness of the 100 individuals in the generation is evaluated and the individuals are ranked according to their fitness. The 100 individuals for the next generation are obtained by the single-point crossover and adaptive mutation operations. The top 30 individuals ranked according to their fitness by the CREASE-GA ('parents') are randomly paired to obtain 70 new individuals ('children') for the next generation. From this new generation composed of 'parents' and 'children', the top two best-performing individuals ('elites') are retained as is for the next generation, the other 98 individuals undergoing an adaptive mutation operation. The adaptive mutation operation is necessary to ensure that the CREASE-GA does not converge quickly to a local minimum. For more details about the crossover and adaptive mutation operations in the context of ML-CREASE, we refer the reader to previous publications on CREASE (Beltran-Villegas *et al.*, 2019; Wu & Jayaraman, 2022; Akepati *et al.*, 2024). Finally, CREASE-GA converges when a generation is composed of individuals with similar fitness values. After analyzing the composition of individuals in successive generations, we conclude that 200 generations of the GA are sufficient to obtain convergence (see Section S4).

Another advantage of using GAs is that the final generation of individuals from ML-CREASE can be used to obtain a distribution of the identified values for every structural feature. We include or exclude an individual from the final identification of structural features and their ranges depending on how closely the computed scattering profile for that individual matches the input experimental profile. We enforce the following fitness criterion for the selection of individuals (shown below):

$$(\mathrm{SSE})_{\mathrm{ind}}^{\mathrm{acc}} < 10 \times (\mathrm{SSE})_{\mathrm{ind}}^{\mathrm{best}}. \tag{4}$$

The individual with the best fitness has the lowest SSE $[(\mathrm{SSE})_{\mathrm{ind}}^{\mathrm{best}}]$. In both cases, all of the individuals included in identifying a range of structural features from ML-CREASE have an SSE $[(\mathrm{SSE})_{\mathrm{ind}}^{\mathrm{acc}}]$ less than ten times that of the SSE for the best-performing individual $[(\mathrm{SSE})_{\mathrm{ind}}^{\mathrm{best}}]$. This fitness of the accepted individuals is a parameter in the CREASE-GA Python script and can be set to the desired value by the user.

Next, we describe *in silico* tests to ensure that ML-CREASE is working as expected and then use ML-CREASE to obtain distributions of size and shape from the 1D SAXS profiles.

3. Results

3.1. *In silico* inputs to validate the ML-CREASE approach

We take 600 test samples whose scattering profiles are not used in the ML model training and provide them as an input to ML-CREASE. At the end of the CREASE-GA run for each test sample, the best-ranked structural features of the final generation are selected and compared with the original structural features of that test sample. After this procedure is completed for all 600 test samples, we use a parity plot to compare the original values for each structural feature with the predictions from ML-CREASE. Both the ML-CREASE predictions and the original structural features are non-dimensionalized in the cross plots; a value of zero indicates the minimum value and one indicates the maximum value of that structural feature.

For spherical nanoparticles with dispersity only in particle size, the *in silico* test for the three structural features – R_μ,

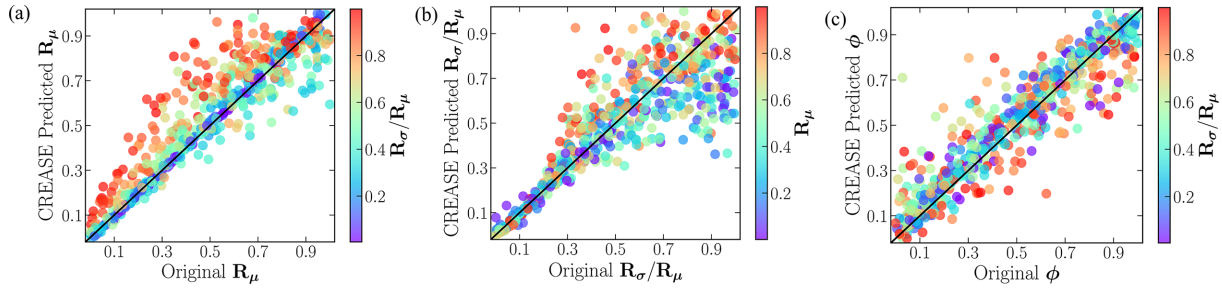


Figure 6
Validation of ML-CREASE using *in silico* tests for spherical nanoparticles with dispersity only in size. All values for structural features (R_μ, R_σ/R_μ and ϕ) shown in (*a*–*c*) are non-dimensionalized from [0–1], with 0 representing the minimum value of the structural features in the 600 test samples and 1 representing the maximum value of the structural features in the 600 test samples. (*a*) CREASE predictions for R_μ compared with the original value of R_μ for the 600 test samples. The data points are color-coded with respect to R_σ/R_μ. (*b*) CREASE predictions for R_σ/R_μ compared with the original value of R_σ/R_μ for the 600 test samples. The data points are color-coded with respect to R_μ. (*c*) CREASE predictions for ϕ compared with the original value of ϕ for the 600 test samples. The data points are color-coded with respect to R_σ/R_μ.

R_σ/R_μ and ϕ – is shown in Fig. 6. The mean volumetric radius [Fig. 6(*a*)] and the volume fraction [Fig. 6(*c*)] are color-coded with respect to the standard deviation of the volumetric radii (expressed as a fraction of the mean volumetric radius, R_σ/R_μ) to understand the role of dispersity in the performance of ML-CREASE (in the prediction of particle size and volume fraction). The standard deviation of the volumetric radii [Fig. 6(*b*)] is color-coded with respect to the mean volumetric radius to understand the effect of the particle size on the performance of ML-CREASE (in the prediction of size dispersity).

The ML-CREASE predicted R_μ is in good agreement with the original R_μ value when R_σ/R_μ is less than 0.5; the same trend is also observed when comparing ML-CREASE predicted and original values for R_σ/R_μ. When R_σ/R_μ is greater than 0.5, the ML-CREASE predictions deviate from the original, regardless of the value of R_μ of the system. During the sensitivity analysis procedure, we find that the R_σ/R_μ structural feature no longer influences the scattering intensities when it is greater than 0.5 (see Section S2). Hence the scatter in the ML-CREASE predictions for R_σ/R_μ when the original R_σ/R_μ is greater than 0.5 is to be expected. We conclude, from these results, that it is impossible for the CREASE-GA to correctly interpret the standard deviation of the volumetric radii when it is greater than 0.5 times the mean volumetric radius.

On the basis of the results for spherical nanoparticles (Fig. 6) with dispersity only in particle size, we restrict the ranges of R_σ/R_μ and γ_σ/γ_μ to [0–0.5) for the analysis of nanoparticles with dispersity in both size and shape. The ML-CREASE predicted structural features for systems of nanoparticles with dispersity in both size and shape are compared with the original structural feature values in Fig. 7. The ML-CREASE predictions of R_μ and ϕ are color-coded with respect to γ_σ/γ_μ for us to understand the effect of particle shapes on the quality of the ML-CREASE predictions. Analogously, the predicted values of γ_μ are color-coded with respect to R_μ to understand the effect of particle size on the ML-CREASE predictions.

The ML-CREASE predictions for R_μ and ϕ compare well with their original values, as shown in Figs. 7(*a*) and 7(*b*), respectively. Fig. 7(*a*) indicates a correlation between the deviations in the ML-CREASE predicted R_μ and the original mean aspect ratio (γ_μ) of the nanoparticle system. The best-performing ML-CREASE individual predicts a lower R_μ than the original for nanoparticle systems with a high aspect ratio and a higher R_μ than the original for nanoparticle systems with a low aspect ratio. We attribute this systematic trend as a consequence of defining an effective sphere radius (volumetric radius) for ellipsoidal particles. We use the volumetric radius (effective sphere radius) due to its intuitive definition. We conclude the systematic trend is not significant, since the

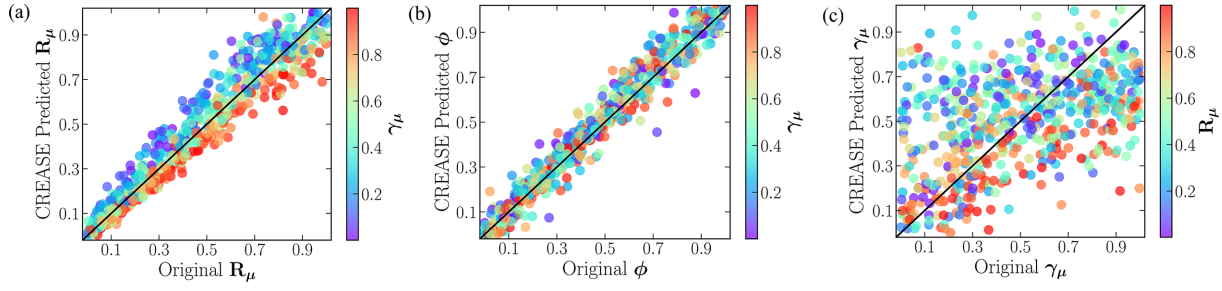


Figure 7
Validation of ML-CREASE using *in silico* tests for nanoparticles with dispersity in both size and shape. All values for structural features (R_μ, ϕ and γ_μ) shown in (*a*–*c*) are non-dimensionalized from [0–1], with 0 representing the minimum value of the structural features in the 600 test samples and 1 representing the maximum value of the structural features in the 600 test samples. (*a*) CREASE predictions for R_μ compared with the original value of R_μ for the 600 test samples. The data points are color-coded with respect to γ_μ. (*b*) CREASE predictions for ϕ compared with the original value of ϕ for the 600 test samples. The data points are color-coded with respect to γ_μ. (*c*) CREASE predictions for γ_μ compared with the original value of γ_μ for the 600 test samples. The data points are color-coded with respect to R_μ.

agreement between the R_μ of the best-performing ML-CREASE individual and the original R_μ is good for the range of aspect ratios considered in our work. The ML-CREASE predictions for γ_μ do not compare well with their original values [Fig. 7(*c*)] which is expected considering the low correlation between the γ_μ structural feature and the intensity values $I(q)$ obtained during the ML model training [Fig. 5(*d*)]. As discussed before, in the absence of orientational and positional order (as in amorphous liquid or non-crystalline materials), the information about the shapes of the nanoparticles can be averaged out in the scattering measurement.

For results of the *in silico* tests for R_σ/R_μ and γ_σ/γ_μ see Section S5. Since the range of R_σ/R_μ is restricted to [0, 0.5) for nanoparticles with dispersity in size and shape, the ML-CREASE predicted R_σ/R_μ compares well with the original R_σ/R_μ. ML-CREASE predictions for γ_σ/γ_μ do not compare well with the original γ_σ/γ_μ; this is expected as we find during the ML model training that there is no correlation between γ_σ/γ_μ and intensity values $I(q)$ [Fig. 5(*d*)].

In our *in silico* test, we are only comparing the predicted values from ML-CREASE's best-ranked individual with the original values; thus, the deviation between the ML-CREASE predictions and the original values of γ_μ is to be expected. Next, we analyze the variability in the ML-CREASE predictions among the individuals in the last 'converged' generation of the GA loop. The variability in the ML-CREASE predictions between the 100 CREASE-GA individuals is analyzed in Section S6. The deviation in R_μ of a CREASE-GA individual from the R_μ of the best-performing CREASE-GA individual gets larger as the individual is ranked lower. The R_μ value reaches its extrema for the worst-performing GA individuals. The variability in γ_μ between the CREASE-GA individuals does not follow the same pattern. The γ_μ value frequently reaches its extrema for individuals ranked in the top half by CREASE-GA. Hence the analysis of variability between the GA individuals suggests that ML-CREASE can identify multiple nanoparticle systems with disparate shapes as possible solutions to an input scattering profile.

On the basis of the *in silico* tests of ML-CREASE with only particle size dispersity, particularly the *in silico* tests for R_σ/R_μ, we expect good agreement between the predictions from ML-CREASE and those from the *lmfit* analytical model when the dispersity in nanoparticle size is low. We expect a disagreement between the predictions from ML-CREASE and those from the *lmfit* analytical model when the dispersity in the nanoparticle size is high. For the case of nanoparticles with dispersity in both size and shape, from the variability analysis (Section S6) for γ_μ, we expect ML-CREASE to be able to identify nanoparticle systems with distinct shape distributions for the input experimental profiles.

3.2. Interpretation of SAXS profiles

First, the 30 SAXS profiles – ten profiles each for the 'strong', 'weak' and 'no' classes – are input to ML-CREASE by assuming spherical nanoparticles with dispersity only in particle size. In Figs. 8(*a*) and 8(*b*) we compare the ML-CREASE identified R_μ and R_σ/R_μ values with the predictions from the *lmfit* analytical model for polydisperse spheres (Newville *et al.*, 2024). In Fig. 8(*a*), the ML-CREASE predictions for R_μ agree with the analytical model (polydisperse spheres) for the 'strong' profiles and 'weak' profiles. However, in contrast, the 'no' profiles do not agree for most cases. This is not surprising as the dispersity in size and shape is likely what gave rise to the lack of peaks in the scattering profiles for the 'no' class, as partly confirmed in the next figure. In Fig. 8(*b*), the ML-CREASE predictions for R_σ/R_μ from the 'strong', 'weak' and 'no' profiles are clustered. The R_σ/R_μ values identified by ML-CREASE for 'strong' profiles are the lowest. The R_σ/R_μ values identified by ML-CREASE for the 'weak' profiles are higher than those values identified for the 'strong' profiles and are generally lower than those values identified by ML-CREASE for the 'no' profiles. The R_σ/R_μ values identified by ML-CREASE for the 'no' profiles are generally the highest, supporting our reasoning for why the mean radius values are not in agreement with the analytical model.

We note that the results shown here correspond to a truncated normal distribution of the size dispersity in the spherical nanoparticles. It is straightforward to adapt the ML-CREASE method for different kinds of distributions to model dispersity. This is demonstrated in Section S7 by analyzing the dispersity in the spherical nanoparticles using a log-normal distribution.

In comparison with analytical models, there are some unique advantages of ML-CREASE. In particular, the ML-CREASE predictions have error bars associated with them. These denote the range of variation for a particular structural feature within the GA individuals that pass the acceptance criterion [shown in equation (4)] in the final generation of the CREASE-GA output. In Fig. 8(*c*) the computed scattering profile for the best-performing CREASE individual is compared with the experimental data for one representative case each of the 'strong', 'weak' and 'no' profiles. The shaded regions around the CREASE predictions represent the standard deviation between the ML-CREASE predicted $I(q)$s for the GA individuals that pass the fitness criterion [equation (4)]. Hence, the use of GAs in ML-CREASE allows for the quantification of uncertainty in the interpretation of small-angle scattering data. For a comparison of the predicted $I(q)$s and the interpreted features between ML-CREASE and analytical models, see Section S8.

Next, we interpret experimental scattering profiles by allowing for dispersity in particle shape in addition to particle size. Distributions of aspect ratios of ellipsoids are usually represented using the Cauchy or Lorentz distributions (Pabst & Berthold, 2007; Lai & Balakrishnan, 2009). Since the gold nanoparticles synthesized for this work are orientationally and positionally disordered, the information about the shape distribution of the nanoparticles gets averaged out. Hence, we choose to model the aspect ratios of the nanoparticles with the better understood normal distribution instead of the Cauchy or Lorentz distributions. We explain the adaptation of ML-CREASE to different kinds of distributions using an example in Section S7. Users can follow a similar procedure to adapt

ML-CREASE for more detailed structural feature distributions, where necessary.

As discussed previously, we restricted the R_σ and γ_σ structural features to the ranges [0, 0.5) of R_μ and γ_μ, respectively, while developing the ML-CREASE method for nanoparticles with dispersity in both size and shape. Since the ML-CREASE identified R_σ/R_μ for the 'no' experimental profiles are generally greater than 0.5 [Fig. 8(*b*)] when assuming spherical nanoparticles with dispersity only in particle size, we ignore the 'no' profiles for this analysis and only analyze the 'strong' and 'weak' profiles for the case of nanoparticles with dispersity in both size and shape.

All the 'strong', 'weak' and 'no' experimental profiles used in this study are shown in Section S9. The predictions from ML-CREASE for nanoparticles with dispersity in both particle size and shape are compared with the ML-CREASE predictions for spherical nanoparticles in Fig. 9.

The values of the structural features predicted by ML-CREASE for the ten samples in 'strong' and ten samples in 'weak' classes are shown in Figs. 9(*a*) and 9(*b*), respectively. For six out of the 20 experimental profiles analyzed, the ML-CREASE method with dispersity in both shape and size predicts sphere-like nanoparticles (γ_μ close to 1.0). This indicates that the synthesized nanoparticles in these cases may indeed be composed largely of sphere-like particles; these samples can be found where the pink and green symbols overlap in Fig. 9(*a*) and the purple and orange symbols overlap in Fig. 9(*b*), showing a value of γ_μ of 1.0. For the other 14 experimental profiles, the ML-CREASE method with dispersity in both shape and size shows that the γ_μ is higher than 1.0. Interestingly, the R_σ values identified by the ML-CREASE method for nanoparticles with dispersity in both size and shape are always lower than the R_σ predictions from ML-CREASE for spherical nanoparticles. This implies that when there is dispersity in shape the particle size dispersity is likely to be smaller than if we assumed all particles are spheres. This indicates that shape dispersity is also a viable interpretation for explaining the different classes of experimental profiles. In Section S10 we compare all the fitness values of the best individuals for the two cases – the spherical particles with dispersity in size and the particles with shape and size dispersity. For the 14 experimental profiles for which the two cases have dissimilar predictions for the mean aspect ratio of the nanoparticles, the comparison of the fitness between the two cases shows that accounting for shape dispersity in addition to size dispersity always improves the fitness of the best-performing GA individual. For the six experimental profiles for which the two cases have similar predictions for the mean aspect ratio of the nanoparticles, the fitness of the best-performing GA individuals from the two cases is also similar.

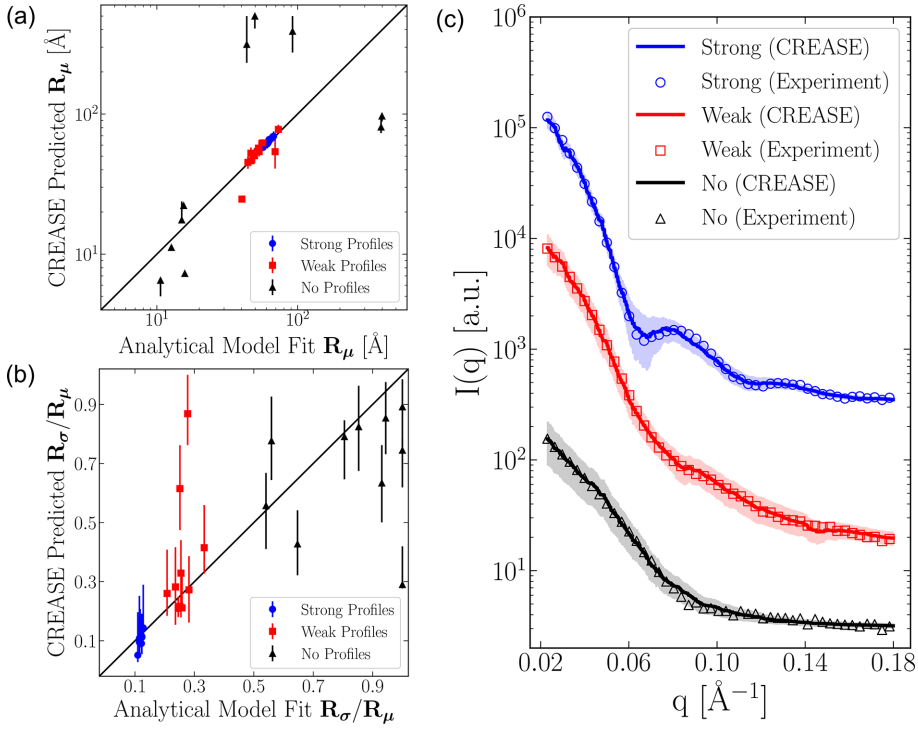


Figure 8
Comparison of ML-CREASE identified structural features from experimental scattering profiles of nanoparticles with the predictions from the *lmfit* analytical model (nanoparticles only have dispersity in particle size). The range of variation for the structural features R_μ and R_σ/R_μ in the last generation of the GA individuals is shown in (*a*) and (*b*), respectively. Symbols in (*a*) and (*b*) represent the ML-CREASE identified structural features with the best fitness for the 'strong' (blue circles), 'weak' (red squares) and 'no' (black triangles) experimental profiles. (*c*) The scattering profiles predicted by the XGBoost ML model for the selected GA individuals [that pass the fitness criterion shown in equation (4)] are compared with an experimental scattering profile for representative cases of the 'strong' (blue), 'weak' (red) and 'no' (black) profiles. Solid lines are the ML-CREASE predictions for the individuals with the best fitness, shaded regions are the standard deviation in $I(q)$ from the ML-CREASE predictions for the selected GA individuals, and symbols are the experimental data.

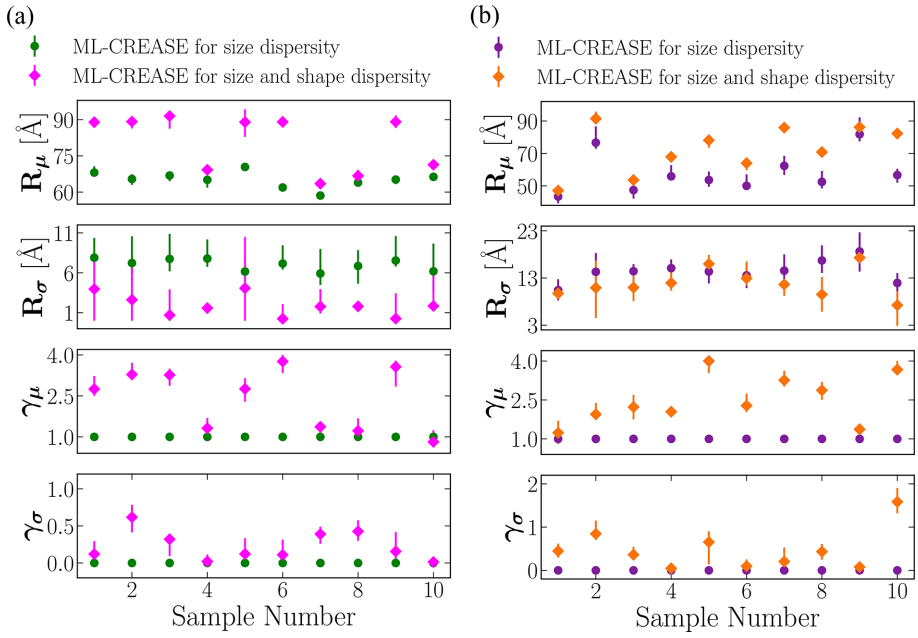


Figure 9
Comparison of ML-CREASE identified structural features for nanoparticles with dispersity only in particle size with the ML-CREASE identified structural features for nanoparticles with dispersity in both size and shape. (*a*) Comparison of ML-CREASE identified structural features for 'strong' experimental profiles. ML-CREASE identified structural features for spherical nanoparticles with dispersity only in particle size are shown as circles (green), ML-CREASE identified structural features for nanoparticles with dispersity in both size and shape are shown as diamonds (magenta). (*b*) Comparison of ML-CREASE identified structural features for 'weak' experimental profiles. ML-CREASE identified structural features for spherical nanoparticles with dispersity only in particle size are shown as circles (purple), ML-CREASE identified structural features for nanoparticles with dispersity in both size and shape are shown as diamonds (orange).

Additionally, in Section S11, TEM images are reported along with their corresponding SAXS profiles for one representative case each from the 'strong', 'weak' and 'no' classes. The TEM images show the lowest size and shape dispersity for the 'strong' profile, moderate size and shape dispersity for the 'weak' profile, and the highest size and shape dispersity for the 'no' profile. The TEM results are in good qualitative agreement with the ML-CREASE interpreted structural features from the corresponding scattering profiles.

4. Conclusions

In summary, citrate-stabilized gold nanoparticles were synthesized using an automated fluidic platform, and subsequently subjected to synchrotron-based *in situ* SAXS characterization. The data from these measurements produced three classes of SAXS profiles that were hand-labeled as 'strong', 'weak' and 'no' based on the presence or absence of features (peaks) in that profile. We then extended the ML-CREASE method to identify (*a*) the size dispersity of spherical nanoparticles and (*b*) the size and shape dispersity of nanoparticles. For the assumption of spherical nanoparticles with size dispersity [case (*a*)], the predictions from ML-CREASE were compared with the predictions from the *lmfit* analytical model. The ML-CREASE predictions agreed with the predictions from the *lmfit* model for 'strong' profiles, but the agreement worsened going from 'weak' profiles to 'no' profiles. Given that the approximations in analytical models become weaker with increasing polydispersity, the disagreements between ML-CREASE and the *lmfit* analytical model for the 'weak' and 'no' profiles were expected. When we repeated the five steps to extend ML-CREASE to identify dispersity in both size and shape of the nanoparticles (spheres to ellipsoids), the results suggested that the system has both spheres and aspherical (ellipsoidal) particles. For six out of the 20 SAXS profiles, both case (*a*) and case (*b*) results confirmed the presence of spheres only. For the remaining 14 SAXS profiles, ML-CREASE predictions from case (*b*) suggest that the particles are not all spherical; when the predicted dispersity in shape was high the dispersity in size was low. For these 14 SAXS profiles, the computed scattering profiles for the predicted CREASE structures from case (*b*) showed a better match to the input SAXS profile than the predicted structures from case (*a*).

The power of ML-CREASE is evident in the analysis of systems that exhibit dispersity in size and shape which usually result in featureless (*i.e.* broadened or flattened peaks) scattering profiles. Such systems with dispersity are often hard to analyze with analytical models. The capability of the ML-CREASE method to identify the size and shape distributions of nanoparticles without *a priori* knowledge is particularly consequential when such information is hard to obtain using imaging techniques. The use of a GA in ML-CREASE also allows for the identification of a range of structural features that describe the size and shape distributions of nanoparticles for an input SAXS profile.

Acknowledgements

We acknowledge the DARWIN and CAVINESS high-performance computing clusters at the University of Delaware for providing computational resources that supported this work.

Conflict of interest

The authors declare no competing financial interest.

Data availability

The codes used in this work are freely available from our laboratory's GitHub page (https://github.com/arthijayaraman-lab/CREASE_Size_Shape_Dispersity). The use of these codes is illustrated as a case study on the CREASE-GA website (https://crease-ga.readthedocs.io/en/latest/casestudy1.html).

Funding information

Funding via the Multi University Research Initiative (MURI) from the Office of Naval Research (ONR) (award No. N00014-23-1-2499) is acknowledged. The DARWIN high-performance computing cluster is supported by the NSF under grant No. 1919839. The experiments were supported by Brookhaven National Laboratory (BNL), Laboratory Directed Research and Development (LDRD) grants No. 22-059, 'Precision synthesis of multiscale nanomaterials through AI-guided robotics for advanced catalysts', and No. 24-004, 'Human-AI-facility integration for the multi-modal studies on high-entropy nanoparticles'. This research also used resources of the Center for Functional Nanomaterials, the CMS (11-BM) and SMI (12-ID) beamlines, and resources of the National Synchrotron Light Source II, which are US Department of Energy Office of Science User Facilities operated at Brookhaven National Laboratory under contract No. DE-SC0012704.

References

Akepati, S. V. R., Gupta, N. & Jayaraman, A. (2024). *JACS Au* **4**, 1570–1582.
Anker, S., Butler, A. T., Selvan, K. & Jensen, R. (2023). *Chem. Sci.* **14**, 14003–14019.
Barnsley, L. C., Nandakumaran, N., Feoktystov, A., Dulle, M., Fruhner, L. & Feygenson, M. (2022). *J. Appl. Cryst.* **55**, 1592–1602.
Beach, M. A., Nayanathara, U., Gao, Y., Zhang, C., Xiong, Y., Wang, Y. & Such, G. K. (2024). *Chem. Rev.* **124**, 5505–5616.
Beltran-Villegas, D. J., Wessels, M. G., Lee, J. Y., Song, Y., Wooley, K. L., Pochan, D. J. & Jayaraman, A. (2019). *J. Am. Chem. Soc.* **141**, 14916–14930.
Blum, L. & Stell, G. (1979). *J. Chem. Phys.* **71**, 42–46.
Breßler, I., Kohlbrecher, J. & Thünemann, A. F. (2015). *J. Appl. Cryst.* **48**, 1587–1598.
Bressler, I., Pauw, B. R. & Thünemann, A. F. (2015). *J. Appl. Cryst.* **48**, 962–969.
Brisard, S. & Levitz, P. (2013). *Phys. Rev. E* **87**, 013305.
Cantor, C. R. & Schimmel, P. R. (1980). *Biophysical chemistry. Part II: techniques for the study of biological structure and function*. Macmillan.
Carter, C. B. & Williams, D. B. (2016). *Transmission electron microscopy: diffraction, imaging, and spectrometry*. Springer.
Chen, T. & Guestrin, C. (2016). *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785–794, San Francisco, USA. Association for Computing Machinery.
Choi, D.-K. (2019). *Int. J. Precis. Eng. Manuf.* **20**, 129–138.
Dawadi, S., Katuwal, S., Gupta, A., Lamichhane, U., Thapa, R., Jaisi, S., Lamichhane, G., Bhattarai, D. P. & Parajuli, N. (2021). *J. Nanomaterials* **2021**, 1–23.
Dyer, K. N., Hammel, M., Rambo, R. P., Tsutakawa, S. E., Rodic, I., Classen, S., Tainer, J. A. & Hura, G. L. (2014). *Methods Mol. Biol. Clifton NJ* **1091**, 245–258.
Eaton, P. & West, P. (2010). *Atomic force microscopy*. Oxford University Press.
Elasri, M., Elharrouss, O., Al-Maadeed, S. & Tairi, H. (2022). *Neural Process. Lett.* **54**, 4609–4646.
Franke, D., Gräwert, T. & Svergun, D. I. (2025). *J. Appl. Cryst.* **58**, 1027–1033.
Fultz, B. & Howe, J. M. (2012). *Transmission electron microscopy and diffractometry of materials*. Springer Science & Business Media.
Glatter, O. (1979). *J. Appl. Cryst.* **12**, 166–175.
Glotzer, S. C. & Solomon, M. J. (2007). *Nat. Mater.* **6**, 557–562.
Goldstein, J. I., Newbury, D. E., Michael, J. R., Ritchie, N. W. M., Scott, J. H. J. & Joy, D. C. (2017). *Scanning electron microscopy and X-ray microanalysis*. Springer.
Guinier, A. (1955). *Small-angle scattering of X-rays*. John Wiley.
Gupta, N., Akepati, S. V. V. R., Bianco, S., Shah, J., Adams, D. J. & Jayaraman, A. (2025). *arXiv*, https://doi.org/10.48550/arXiv.2504.03869.
Gupta, N. & Jayaraman, A. (2023). *Nanoscale* **15**, 14958–14970.
Hagita, K., McGreevy, R. L., Arai, T., Inui, M., Matsuda, K. & Tamura, K. (2010). *J. Phys. Condens. Matter* **22**, 404215.
Hagita, K., Tominaga, T. & Sone, T. (2018). *Polymer* **135**, 219–229.
Heil, C. M. & Jayaraman, A. (2021). *ACS Mater. Au* **1**, 140–156.
Heil, C. M., Ma, Y., Bharti, B. & Jayaraman, A. (2023*a*). *JACS Au* **3**, 889–904.
Heil, C. M., Patil, A., Dhinojwala, A. & Jayaraman, A. (2022). *ACS Cent. Sci.* **8**, 996–1007.
Heil, C. M., Patil, A., Vanthournout, B., Singla, S., Bleuel, M., Song, J.-J., Hu, Z., Gianneschi, N. C., Shawkey, M. D., Sinha, S. K., Jayaraman, A. & Dhinojwala, A. (2023*b*). *Sci. Adv.* **9**, eadf2859.
Ho, C. C., Keller, A., Odell, J. A. & Ottewill, R. H. (1993). *Colloid Polym. Sci.* **271**, 469–479.
Holland, J. H. (1992). *Sci. Am.* **267**, 66–72.
Ilavsky, J. & Jemian, P. R. (2009). *J. Appl. Cryst.* **42**, 347–353.
Jeffries, C. M., Ilavsky, J., Martel, A., Hinrichs, S., Meyer, A., Pedersen, J. S., Sokolova, A. V. & Svergun, D. I. (2021). *Nat. Rev. Methods Primer* **1**, 1–39.
Kinnear, C., Moore, T. L., Rodriguez-Lorenzo, L., Rothen-Rutishauser, B. & Petri-Fink, A. (2017). *Chem. Rev.* **117**, 11476–11521.
Lai, C. D. & Balakrishnan, N. (2009). *Continuous bivariate distributions*. New York: Springer.
Lee, J. Y., Song, Y., Wessels, M. G., Jayaraman, A., Wooley, K. L. & Pochan, D. J. (2020). *Macromolecules* **53**, 8581–8591.
Li, T., Senesi, A. J. & Lee, B. (2016). *Chem. Rev.* **116**, 11128–11180.
Lindner, P. & Oberdisse, J. (2024). *Neutrons, X-rays, and light: scattering methods applied to soft condensed matter*. Elsevier.
Liu, Z., Liu, N. & Schroers, J. (2022). *Prog. Mater. Sci.* **125**, 100891.
Lu, S. & Jayaraman, A. (2024). *Prog. Polym. Sci.* **153**, 101828.
Luo, Z., Marson, D., Ong, Q. K., Loiudice, A., Kohlbrecher, J., Radulescu, A., Krause-Heuer, A., Darwish, T., Balog, S., Buonsanti, R., Svergun, D. I., Posocco, P. & Stellacci, F. (2018). *Nat. Commun.* **9**, 1343.
Manalastas-Cantos, K., Konarev, P. V., Hajizadeh, N. R., Kikhney, A. G., Petoukhov, M. V., Molodenskiy, D. S., Panjkovich, A., Mertens, H. D. T., Gruzinov, A., Borges, C., Jeffries, C. M., Svergun, D. I. & Franke, D. (2021). *J. Appl. Cryst.* **54**, 343–355.

McGreevy, R. L. (1995). *Nucl. Instrum. Methods Phys. Res. A* **354**, 1–16.
McGreevy, R. L. (2001). *J. Phys. Condens. Matter* **13**, R877–R913.
McGreevy, R. L. & Pusztai, L. (1988). *Mol. Simul.* **1**, 359–367.
Menon, S. V. G., Manohar, C. & Rao, K. S. (1991). *J. Chem. Phys.* **95**, 9186–9190.
Mitchell, M. (1998). *An introduction to genetic algorithms*. MIT Press.
Modena, M. M., Rühle, B., Burg, T. P. & Wuttke, S. (2019). *Adv. Mater.* **31**, 1901556.
Mourdikoudis, S. M., Pallares, R. & Thanh, K. (2018). *Nanoscale* **10**, 12871–12934.
Narayanan, T., Wacklin, H., Konovalov, O. & Lund, R. (2017). *Crystallogr. Rev.* **23**, 160–226.
Newville, M., Otten, R., Nelson, A., Stensitzki, T., Ingargiola, A., Allan, D., Fox, A., Carter, F., Michał, Osborn, R., Pustakhod, D., Weigand, S., lneuhaus, Aristov, A., Glenn, Mark, mgunyho, Deil, C., Hansen, A. L. R., Pasquevich, G., Foks, L., Zobrist, N., Frost, O., Stuermer, Jaskula, J.-C., Caldwell, S., Eendebak, P., Pompili, M., Nielsen, J. H. & Persaud, A. (2024). *lmfit/lmfit-py: 1.3.2*, https://doi.org/10.5281/zenodo.598352.
Pabst, W. & Berthold, C. (2007). *Part. Part. Syst. Charact.* **24**, 458–463.
Patil, A., Heil, C. M., Vanthournout, B., Bleuel, M., Singla, S., Hu, Z., Gianneschi, N. C., Shawkey, M. D., Sinha, S. K., Jayaraman, A. & Dhinojwala, A. (2022*a*). *Adv. Opt. Mater.* **10**, 2102162.
Patil, A., Heil, C. M., Vanthournout, B., Singla, S., Hu, Z., Ilavsky, J., Gianneschi, N. C., Shawkey, M. D., Sinha, S. K., Jayaraman, A. & Dhinojwala, A. (2022*b*). *ACS Mater. Lett.* **4**, 1848–1854.
Pedersen, J. S. (1997). *Adv. Colloid Interface Sci.* **70**, 171–210.
Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., Gorba, C., Mertens, H. D. T., Konarev, P. V. & Svergun, D. I. (2012). *J. Appl. Cryst.* **45**, 342–350.
Porod, G. (1951). *Kolloid-Z.* **124**, 83–114.
Quek, A. J., Cowieson, N. P., Caradoc-Davies, T. T., Conroy, P. J., Whisstock, J. C. & Law, R. H. P. (2023). *Int. J. Mol. Sci.* **24**, 14258.
Röding, M., Tomaszewski, P., Yu, S., Borg, M. & Rönnols, J. (2022). *Front. Mater.* **9**, 956839.
Rodríguez-Ruiz, I., Radajewski, D., Charton, S., Phamvan, N., Brennich, M., Pernot, P., Bonneté, F. & Teychené, S. (2017). *Sensors* **17**, 1266.
Rolland, J. P., Maynor, B. W., Euliss, L. E., Exner, A. E., Denison, G. M. & DeSimone, J. M. (2005). *J. Am. Chem. Soc.* **127**, 10096–10100.
Salacuse, J. J. & Stell, G. (1982). *J. Chem. Phys.* **77**, 3714–3725.
Schmidt, P. W. (1988). *Makromol. Chem. Macromol. Symp.* **15**, 153–166.
Schmidt-Rohr, K. (2007). *J. Appl. Cryst.* **40**, 16–25.
Shi, J., Kantoff, P. W., Wooster, R. & Farokhzad, O. C. (2017). *Nat. Rev. Cancer* **17**, 20–37.
Shrestha, R. & Xie, B. (2023). *arXiv*, https://doi.org/10.48550/arXiv.2312.13253.
Shwartz-Ziv, R. & Armon, A. (2022). *Inf. Fusion* **81**, 84–90.
Song, K., Yan, F., Ding, T., Gao, L. & Lu, S. (2020). *Comput. Mater. Sci.* **174**, 109472.
Stukowski, A. (2009). *Modell. Simul. Mater. Sci. Eng.* **18**, 015012.
Svergun, D. I., Koch, M. H. J., Timmins, P. A. & May, R. P. (2013). *Small angle X-ray and neutron scattering from solutions of biological macromolecules*. Oxford University Press.
Thebelt, A., Tsay, C., Lee, R., Sudermann-Merx, N., Walz, D., Shafei, B. & Misener, R. (2022). *Adv. Neural Inf. Process. Syst.* **35**, 37401–37415.
Thorkelsson, K., Bai, P. & Xu, T. (2015). *Nano Today* **10**, 48–66.
Ul-Hamid, A. (2018). *A beginners' guide to scanning electron microscopy*. Springer.
Voigtländer, B. (2019). *Atomic force microscopy*. Springer.
Wang, L., Hasanzadeh Kafshgari, M. & Meunier, M. (2020). *Adv. Funct. Mater.* **30**, 2005400.
Wessels, M. G. & Jayaraman, A. (2021*a*). *Macromolecules* **54**, 783–796.
Wessels, M. G. & Jayaraman, A. (2021*b*). *ACS Polym. Au* **1**, 153–164.
Wu, W. & Pauly, M. (2022). *Mater. Adv.* **3**, 186–215.
Wu, Z. & Jayaraman, A. (2022). *Macromolecules* **55**, 11076–11091.
Wuithschick, M., Birnbaum, A., Witte, S., Sztucki, M., Vainio, U., Pinna, N., Rademann, K., Emmerling, F., Kraehnert, R. & Polte, J. (2015). *ACS Nano* **9**, 7052–7071.
Yager, K. G., Majewski, P. W., Noack, M. M. & Fukuto, M. (2023). *Nanotechnology* **34**, 322001.
Ye, Z., Wu, Z. & Jayaraman, A. (2021). *JACS Au* **1**, 1925–1936.
Zheng, J., Cheng, X., Zhang, H., Bai, X., Ai, R., Shao, L. & Wang, J. (2021). *Chem. Rev.* **121**, 13342–13453.