JOURNAL OF APPLIED CRYSTALLOGRAPHY

ISSN 1600-5767

Machine-learning-informed scattering correlation analysis of sheared colloids

Lijie Ding,[a]‡ Yihao Chen[b]*‡ and Changwoo Do[a]

[a]Neutron Scattering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA, and [b]Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA. *Correspondence e-mail: ychen258@sas.upenn.edu

Received 20 December 2024
Accepted 11 April 2025

Edited by E. P. Gilbert, Australian Centre for Neutron Scattering, ANSTO, Australia

This article is part of a collection of articles related to the 19th International Small-Angle Scattering Conference (SAS2024) in Taipei, Taiwan.

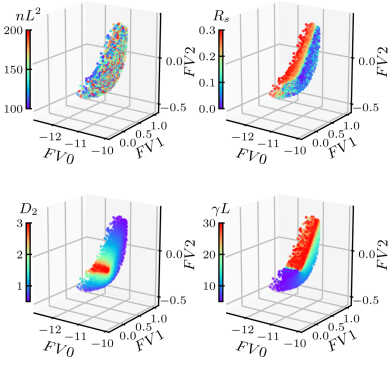
‡ Equal contribution. YC conceived this work and carried out theoretical analysis. LD carried out Monte Carlo simulations and machine-learning analysis.

Keywords: small-angle scattering; machine learning; Gaussian process regression; Monte Carlo simulations; colloids.

We have carried out theoretical analysis, Monte Carlo simulations and machine-learning analysis to quantify microscopic rearrangements of dilute dispersions of spherical colloidal particles from coherent scattering intensity. Both monodisperse and polydisperse dispersions of colloids were created and underwent a rearrangement consisting of an affine simple shear and non-affine rearrangement using the Monte Carlo method. We calculated the coherent scattering intensity of the dispersions and the correlation function of intensity before and after the rearrangement and generated a large data set of angular correlation functions for varying system parameters, including number density, polydispersity, shear strain and non-affine rearrangement. Singular value decomposition of the data set shows the feasibility of machine-learning inversion from the correlation function for the polydispersity, shear strain and non-affine rearrangement using only three parameters. A Gaussian process regressor is then trained on the data set and can retrieve the affine shear strain, non-affine rearrangement and polydispersity with relative errors of 3%, 1% and 6%, respectively. Altogether, our model provides a framework for quantitative studies of both steady and non-steady microscopic dynamics of colloidal dispersions using coherent scattering methods.

1. Introduction

Quantification of the microscopic dynamics of materials made of nanometre- to micrometre-scale constituents is vital in understanding the origins of macroscopic mechanical properties and designing novel functional materials for pharmaceutical, environmental and other industrial applications (Wu *et al.*, 2020). While traditional optical microscopy can provide real space information, it is limited by its resolution and the opacity of the materials (Badon *et al.*, 2017). Scattering techniques (Murphy *et al.*, 2020; Guinier *et al.*, 1955) like X-ray photon correlation spectroscopy (Chu & Hsiao, 2001; Shpyrko, 2014; Leheny *et al.*, 2015; Madsen *et al.*, 2020), dynamic light scattering (Goldburg, 1999; Aime & Cipelletti, 2019) and small-angle neutron scattering (Shibayama, 2011; Chen, 1986) provide great opportunities to probe the microscopic information of such materials, and they have been deployed to study the microstructural dynamics of colloidal (Chen *et al.*, 2020*a*; Donley *et al.*, 2023), polymeric (Ruocco *et al.*, 2013) and atomic materials (Lüttich *et al.*, 2018). However, the challenge of scattering techniques is the quantification of microscopic rearrangement in real space back from the scattering patterns (Fourier space) which, most of the time, are only available in a limited range of two dimensions (in the case of an area detector) and sometimes even one dimension (in the case of a photon-counting device). Previous efforts have investigated steady shear (Burghardt *et al.*, 2012), diffusion



(Leitner *et al.*, 2009), localization of particles (Chen *et al.*, 2020*b*) *etc.* Most of the schemes characterize the temporal correlation function of the scattering intensity and require an average of the intensity or/and the correlation function of the intensity in a time interval assuming steady dynamics. On the other hand, non-steady microscopic dynamics, like non-affine rearrangement, are widely observed in real space and play an important role in the non-linear properties of soft materials (Keim & Arratia, 2015; Wen *et al.*, 2012), like yield (Jana & Pastewka, 2019) and memory effects (Galloway *et al.*, 2022). However, they are much less studied in coherent scattering experiments due to their non-linear and transient nature, and their interpretation often requires intensive modeling and computation (Ma *et al.*, 2014; Banetta *et al.*, 2022; He *et al.*, 2024).

To address these issues, we developed a generalized theoretical and machine-learning (ML) framework (Murphy, 2012; Carleo *et al.*, 2019) to quantify affine and non-affine rearrangements of dilute, both monodisperse and polydisperse, colloidal dispersions using the correlation function of coherent scattering intensity. Recently, various ML frameworks have been developed to interpret scattering experiments, such as SCAN [SCattering Ai aNalysis (Tomaszewski *et al.*, 2021)], which automates structural analysis using predefined particle shape models, and CREASE [computational reverse-engineering analysis for scattering experiments (Anker *et al.*, 2023; Lu & Jayaraman, 2024; Akepati *et al.*, 2024; Heil *et al.*, 2023; Wu & Jayaraman, 2022; Wessels & Jayaraman, 2021)], which leverages genetic algorithms and surrogate ML methods (*e.g. XGBoost*) to reconstruct 3D structural features, including domain size, shape, orientation and spatial distributions from scattering profiles. In particular, CREASE-2D (Akepati *et al.*, 2024) enabled analysis of 2D scattering patterns, avoiding the traditional approximate analysis of scattering intensity profiles. Other ML methods have also been applied for particle tracking and ordered structures in soft materials (Clegg, 2021) and surface scattering analysis (Hinderhofer *et al.*, 2023). Our approach is distinct from these methods in that it directly correlates coherent scattering intensity with physical parameters such as shear strain, non-affine rearrangement and polydispersity in dynamically sheared colloidal systems. Unlike generalizable tools, our ML approach utilizes Monte Carlo (MC) simulation (Krauth, 2006) to generate particle configurations and rearrangements in two dimensions. The coherent scattering intensity of the particles and the correlation function of the intensity were calculated before and after the rearrangement, and using singular value decomposition (SVD) we extracted three essential features of the correlation function that reliably recover the magnitude of both affine and non-affine rearrangements and polydispersity. Similar approaches have been applied to other soft-matter systems including colloids (Chang *et al.*, 2022; Huang *et al.*, 2023; Tung *et al.*, 2022; Tung *et al.*, 2024*a*), lamellae (Tung *et al.*, 2024*b*; Tung *et al.*, 2025) and polymers (Tung *et al.*, 2023; Ding *et al.*, 2024*a*; Ding *et al.*, 2024*b*; Ding *et al.*, 2025). We then used Gaussian process regression (GPR) (Williams & Rasmussen, 2006) to map from scattering data to the system parameters including polydispersity, shear strain and non-affine rearrangement. We also tested our trained GPR using simulation data aside from the training data; good agreement between the ML-extracted system parameters and the MC references was achieved, showing the accuracy of our approach. Our model can be easily applied to coherent scattering experiments to extract microscopic rearrangements between two scattering patterns, which is especially useful for studies of non-steady and transient dynamics.

The rest of this paper is organized as follows: Section 2 introduces our colloidal systems, the theoretical analysis of coherent scattering intensity, MC simulation and a brief summary of the GPR; our results are presented in Section 3 to illustrate the scattering intensity and correlation function of the colloidal dispersions under rearrangements, validate the feasibility for ML inversion of system parameters using SVD of the correlation and show the application of ML analysis for scattering data using GPR; finally, we summarize our paper and discuss potential future directions following this work in Section 4.

2. Methods

2.1. Coherent scattering and rearrangement transformation

The normalized scattering intensity of a polydisperse dispersion of N spherical particles with a configuration of 2D coordinates $\mathcal{S} = \{(x, y)\}$ is given by (Chen, 1986)

$$I(\mathbf{q}; \mathcal{S}) = \frac{\left[\sum_{i=1}^{N} V_i F_i(\mathbf{q}) \exp(-i\mathbf{q}\cdot\mathbf{r}_i)\right]\left[\sum_{i=1}^{N} V_i F_i^{\dagger}(\mathbf{q}) \exp(i\mathbf{q}\cdot\mathbf{r}_i)\right]}{\sum_{i=1}^{N} V_i^2}, \tag{1}$$

where $V_i = (4\pi/3)R_i^3$ is the volume of particle i with radius R_i, $\mathbf{q}$ is the scattering wavevector and $\mathbf{r}_i = (x_i, y_i)$ is the position of particle i. $F_i(\mathbf{q})$ is the form factor amplitude of the ith particle such that (Guinier *et al.*, 1955)

$$F_i(\mathbf{q}) = 3\frac{\sin(qR_i) - qR_i\cos(qR_i)}{(qR_i)^3}, \tag{2}$$

where $q = |\mathbf{q}|$ is the magnitude of the scattering wavevector.

The rearrangement transformation of the particle positions consists of an affine simple shear deformation along the x axis (shear gradient runs along the y axis) with a shear strain γ, as shown in Fig. 1(*a*), and a non-affine rearrangement where the particles are randomly displaced by δx_i and δy_i, which follow a Gaussian distribution with zero mean and a standard deviation D_2. Such a transformation Γ is expressed as

$$\Gamma\begin{pmatrix} x_i \\ y_i \end{pmatrix} = \begin{pmatrix} x_i + \gamma y_i + \delta x_i \\ y_i + \delta y_i \end{pmatrix}. \tag{3}$$

In homodyne scattering, the average translation of all the particles has no effect on the scattering intensity, so we choose the form of equation (3) to have a fixed origin for the affine shear.

The correlation function of two scattering intensities before and after the transformation Γ is

$$g(\mathbf{q}) = \frac{\langle I(\mathbf{q}; \mathcal{S}) I[\mathbf{q}; \Gamma(\mathcal{S})]\rangle_{\mathcal{S}}}{\langle I(\mathbf{q}; \mathcal{S})\rangle_{\mathcal{S}}^2}, \tag{4}$$

where $\langle \ldots \rangle_{\mathcal{S}}$ is the average over different realizations of the configuration.

In the case of dilute dispersions of monodisperse spherical particles, including materials doped with a dilute amount of monodisperse tracer particles, the positions of different particles and, thus, their rearrangements are uncorrelated, and the correlation function $g(\mathbf{q})$ associated with the transformation Γ is predicted as [similar to that in Aime & Cipelletti (2019)]

$$g(\mathbf{q}) = 1 + \mathrm{sinc}^2\left(\frac{q\gamma L \cos\theta}{2}\right) \exp(-q^2 D_2^2), \tag{5}$$

where $\mathrm{sinc}(x) = \sin(x)/x$, L is the size the scattering beam, and θ is the angle between $\mathbf{q}$ and the x axis.

2.2. Monte Carlo simulation

We sampled the positions of particles $\mathcal{S}$ in a $[-L, L]^2$ square with a number density n. The radius of the particles follows a log-normal distribution such that $\ln R_i \sim \mathcal{N}(\ln R_0, R_s)$, where the polydispersity index of the dispersion, PDI $= \langle V_i^2\rangle_i / \langle V_i\rangle_i^2 = \langle R_i^6\rangle_i / \langle R_i^3\rangle_i^2 = \exp(9R_s^2)$, is controlled by R_s (Kotz *et al.*, 2019).

We calculated the scattering intensity $I(\mathbf{q}) = I(q_x, q_y) = I[q\cos(\theta), q\sin(\theta)]$ in the polar coordinates for all the particles inside the box of $[-0.5L, 0.5L]^2$ before and after the Γ transformation, and then calculated the correlation function $g(\mathbf{q})$. The values of $I(\mathbf{q})$ and $g(\mathbf{q})$ are averaged over 2×10^4 samples of $\mathcal{S}$ for each set of system parameters (nL^2, R_s, D_2, γL). We also calculated the radial and angular average of the correlation function such that $g(q) = \langle g(\mathbf{q})\rangle_\theta$ and $g(\theta) = \langle g(\mathbf{q})\rangle_q$, where $\langle \ldots \rangle_\theta$ and $\langle \ldots \rangle_q$ are averages over all measured θ and q, respectively. Without loss of generality, we use the natural unit $R_0 = 1$ for the size of particles and the beam size $L = 800R_0$. We measured $I(\mathbf{q})$ and $g(\mathbf{q})$ with respect to 100 different values of $q \in [0.5, 5]$ uniformly distributed on a log scale, and 101 different values of $\theta \in [0, \pi]$ uniformly distributed on a linear scale, and we note that $I(q, \theta + \pi) = I(q, \theta)$ due to $\pm\mathbf{q}$ symmetry of equation (1). The choice of R_0, L and the range of q is to mirror experimental conditions: a colloidal dispersion of particles with a radius of 10 nm, a synchrotron X-ray beam size of 8 µm and an area detector at a small-angle scattering setup that covers a wavevector ranging from 0.05 to 0.5 nm^{-1}.

2.3. Gaussian process regression

To obtain the inverted mapping from the scattering correlation function $\mathbf{x} = g(\mathbf{q})$ to the system parameters or inversion targets $\mathbf{y} = (nL^2, R_s, D_2, \gamma L)$, we trained a GPR using the data generated by MC simulation. In the context of GPR, the prior on the regression function is a Gaussian process, $g(\mathbf{x}) \sim GP[m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')]$, where $m(\mathbf{x})$ is the prior mean function and $k(\mathbf{x}, \mathbf{x}')$ is the covariance function or kernel. The goal of the GPR is to find the optimized posterior $p(\mathbf{Y}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{Y})$ of the function output $\mathbf{y}$. The joint distribution of the Gaussian process is (Williams & Rasmussen, 2006)

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{Y}_* \end{pmatrix} \sim \mathcal{N}\left\{ \begin{bmatrix} m(\mathbf{X}) \\ m(\mathbf{X}_*) \end{bmatrix}, \begin{bmatrix} k(\mathbf{X}, \mathbf{X}) & k(\mathbf{X}, \mathbf{X}_*) \\ k(\mathbf{X}_*, \mathbf{X}) & k(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix} \right\}, \tag{6}$$

where a constant function is used for the prior mean $m(\mathbf{x})$, and the kernel consists of a Radial basis function and white noise. $k(\mathbf{x}, \mathbf{x}') = \exp\left[-|\mathbf{x} - \mathbf{x}'|^2/(2l)\right] + \sigma\delta(\mathbf{x}, \mathbf{x}')$, where δ is the Kronecker delta function. l and σ denote the hyperparameters corresponding to the correlation length and variance of observational noise, which can be obtained by training on simulation data. In practice, we use the *scikit-learn* Gaussian process library (Pedregosa *et al.*, 2011) for convenience of implementation and efficiency.

To investigate the distribution of $\mathbf{X}$, we define the pair distance distribution function (PDDF) for $\mathbf{X}$,

$$p(z) = \frac{1}{M^2} \sum_{i,j=1}^{M} \delta\big(|\mathbf{x}_i - \mathbf{x}_j| - z\big), \tag{7}$$

where $M = |\mathbf{X}|$ is the number of data points. In addition, to help determine the appropriate range of hyperparameter l when initiating then optimization process, we calculated the auto-correlation function (ACF) for feature μ (Chang *et al.*, 2022),

$$C_\mu(z) = \frac{\langle \mu(\mathbf{x})\mu(\mathbf{x} + z)\rangle_{\mathbf{x}} - \langle \mu(\mathbf{x})\rangle_{\mathbf{x}}^2}{\langle \mu^2(\mathbf{x})\rangle_{\mathbf{x}} - \langle \mu(\mathbf{x})\rangle_{\mathbf{x}}^2}, \tag{8}$$

where the $\langle \ldots \rangle_{\mathbf{x}}$ is averaged over all data points in $\mathbf{X}$.

3. Results

3.1. Scattering function of the dispersion

Fig. 1(*a*) shows an example of the configuration $\mathcal{S}$ of monodisperse particles before and after the rearrangement transformation Γ, where $\gamma L = 30$ and $D_2 = 0.5$. Figs. 1(*b*) and 1(*c*) show the corresponding coherent scattering patterns of configurations $\mathcal{S}$ and $\Gamma(\mathcal{S})$, respectively. Speckles are clearly observed in the scattering patterns. Fig. 1(*d*) shows the product of the two instances of scattering intensity shown in Figs. 1(*b*) and 1(*c*).

Fig. 2(*a*) shows an example of the scattering pattern $I(\mathbf{q})$ of monodisperse particles after averaging over 20000 configurations of $\mathcal{S}$. The scattering intensity $I(\mathbf{q})$ is isotropic and reflects the form factor $F^2(\mathbf{q})$ of the spherical particles. Fig. 2(*b*) shows the correlation $g(\mathbf{q})$ averaged over 20000 samples. $g(\mathbf{q})$ is highly anisotropic with a high correlation in the y axis ($\theta = \pi/2$). The non-affine rearrangement is isotropic and causes the decay of structural correlation in all directions. For the affine shear rearrangement [equation (3)], the y coordinates of the particles remain unchanged, so the affine shear only leads to the decay of correlation in other directions except the y axis (gradient of the affine shear). The pattern of $g(\mathbf{q})$ is captured very well by the theoretical prediction of equation (5) as shown in Fig. 2(*c*), and the mean percentage

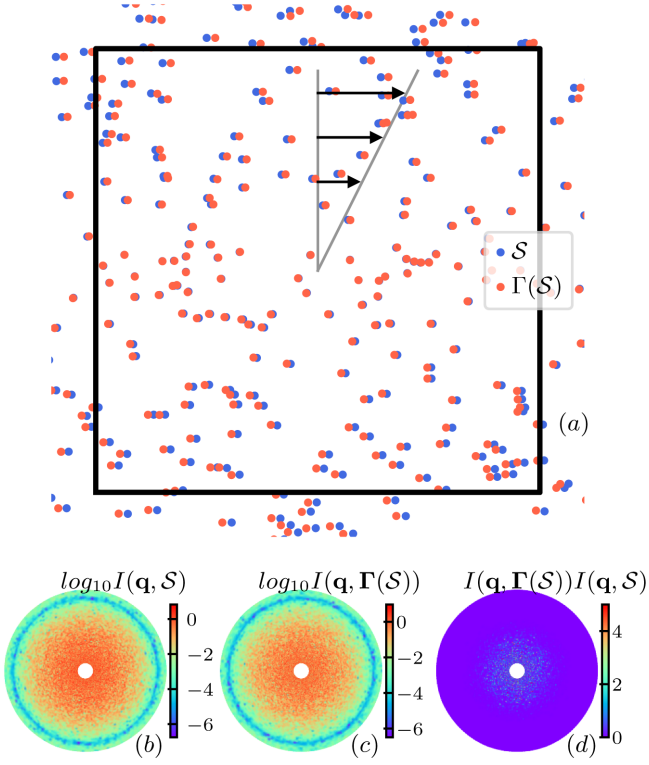


Figure 1
Illustration of a single configuration of particles undergoing transformation Γ, in which $(nL^2, R_s, D_2, \gamma L) = (150, 0, 0.5, 30)$. The black frame indicates the region the beamline shines on. (*a*) Spatial distribution of particles before ($\mathcal{S}$) and after [$\Gamma(\mathcal{S})$] the transformation. For better visualization, the particles are not to scale. Scattering intensity of the configuration (*b*) before and (*c*) after the transformation Γ. (*d*) Product of the correlation function between the scattering intensity of transformed and non-transformed configurations.

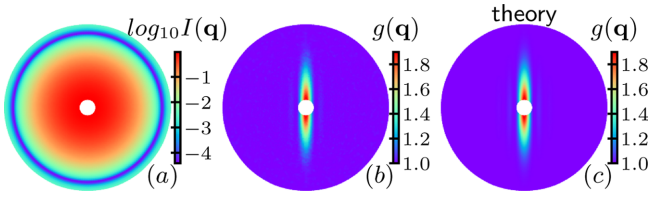


Figure 2
Scattering function $I(\mathbf{q})$ and scattering correlation function $g(\mathbf{q})$ of the particles with $(nL^2, R_s, D_2, \gamma L) = (150, 0, 0.5, 10)$. The range of the scattering wavevector is $q = |\mathbf{q}| \in [0.5, 5]$, plotted in linear scale. (*a*) Averaged scattering intensity. (*b*) Averaged correlation function. (*c*) Theoretical predicted correlation function for the monodisperse system as in equation (5).

error between the MC calculation $g(\mathbf{q})$ and the theoretical prediction $g_{\text{theo}}(\mathbf{q})$ is $\text{Err} = \langle[g(\mathbf{q}) - g_{\text{theo}}(\mathbf{q})]/g_{\text{theo}}(\mathbf{q})\rangle_{\mathbf{q}} = 0.9\%$, where $\langle\ldots\rangle_{\mathbf{q}}$ is the average over all $\mathbf{q}$.

To further quantify the effect of polydispersity, shear strain and non-affine transformation on the correlation function, we used $(R_s, D_2, \gamma L) = (0, 1, 10)$ as a baseline to demonstrate the effect of the three system parameters $(R_s, D_2, \gamma L)$ on $g(q)$ and $g(\theta)$. Fig. 3 shows the radial and angular averaged correlation function $g(q)$ and $g(\theta)$ with different values of $(R_s, D_2, \gamma L)$, which alter the correlation function in different ways. The

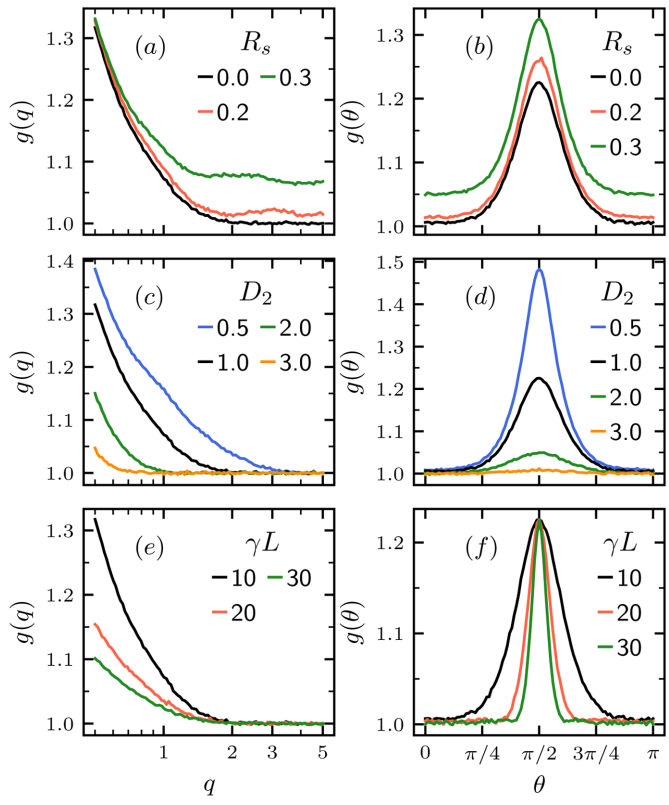


Figure 3
Radial (left column) and angular (right column) correlation functions, $g(q)$ and $g(\theta)$, for various polydispersity R_s, non-affine rearrangement D_2 and affine shear γL with a reference of $(n, R_s, D_2, \gamma L) = (150, 0, 1, 10)$ represented by the black lines. (*a*) and (*b*) Variable R_s, (*c*) and (*d*) variable D_2, and (*e*) and (*f*) variable γL.

radial correlation function, $g(q)$, peaks at the lowest q and decays at higher q. The variation of particle size R_s controls the height of the plateau of $g(q)$ at the high-q limit, which increases with larger R_s. Affine shear strain γL and non-affine rearrangement D_2 affect the peak of $g(q)$ at the lowest q in a similar way, where increasing both shear strain γL and D_2 lowers the peak. The effect of the three system parameters on the angular correlation function $g(\theta)$ is more distinguishable, as shown in Figs. 3(*b*), 3(*d*) and 3(*f*). The angular correlation function $g(\theta)$ has a peak at $\theta = \pi/2$ (affine shear gradient direction), and γL, D_2 and R_s affect the width, height and baseline of the peak. Therefore, we focus on the angular correlation function for the rest of the work. However, radial correlation functions also show significant enough feature differences, which can be used for the same inversion analysis presented here.

3.2. Feasibility of machine-learning inversion

We generated a data set of 6000 angular scattering correlation functions, $\mathbf{F} = \{g(\theta)\}$, whose corresponding system parameters $\mathbf{Y} = \{(nL^2, R_s, D_2, \gamma L)\}$ are randomly distributed such that $nL^2 \in U(100, 200)$, $R_s \in U(0, 0.3)$, $D_2 \in U(0.5, 3)$ and $\gamma L \in U(5, 30)$, where $U(a, b)$ is a uniform distribution in the interval $[a, b]$. Note that $g(\theta)$ is measured for 101 θ uniformly in $[0, \pi]$, so $\mathbf{F}$ is a 6000 × 101 matrix. Following a

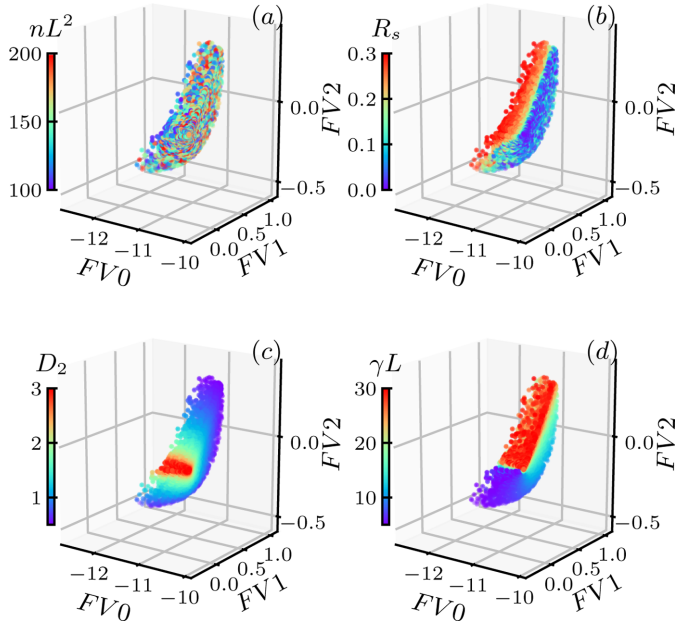


Figure 4
Distribution of the system parameters in the singular value space. (*a*) Number density nL^2. (*b*) Variation of particle size R_s. (*c*) Non-affine transformation D_2. (*d*) Shear rate γL.

similar framework to the literature (Chang *et al.*, 2022; Ding *et al.*, 2024*b*), we carried out a principle component analysis of the data set matrix **F** using the SVD $\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$, where **U**, **Σ** and **V** are matrices of size 6000 × 6000, 6000 × 101 and 101 × 101, respectively. Matrix **V** consists of the singular vectors, and the entries of $\mathbf{\Sigma}^2$ are the corresponding coefficients of the projection of **F** onto the principal vectors in **V**.

Projecting each $g(\theta)$ in the data set **F** onto the three singular vectors (*V*1, *V*2, *V3*) yields three corresponding projection values (*FV*0, *FV*1, *FV*2), which can be considered as a dimension reduction of the original $g(\theta)$. This converts each $g(\theta)$, as well as the system parameter $\mathbf{Y} = \{(nL^2, R_s, D_2, \gamma L)\}$ associated with it, in the data set **F** to a point in 3D space spanned by the three projection values (*FV*0, *FV*1, *FV*2). Fig. 4 shows the distribution of $(nL^2, R_s, D_2, \gamma L)$ in such a space. The distribution shows the feasibility of mapping the features of $g(\theta)$ back to these system parameters. From the color distribution, we note that the values of $(R_s, D_2, \gamma L)$ are well spread out in the (*FV*0, *FV*1, *FV*2) space, indicating a smooth and continuous mapping from the projection values of $g(\theta)$ to these corresponding system parameters. However, the distribution of number density nL^2 in (*FV*0, *FV*1, *FV*2) is rather random, implying it is not suitable for the inverted mapping. The inability to extract the number density information from our scattering correlation function is not surprising, as we are working in the limit of dilute dispersions where the number density of particles does not play a role in the microscopic structure or the rearrangement.

3.3. Inference of the system parameters

For the ML inversion of system parameters $(R_s, D_2, \gamma L)$ from angular scattering correlation function $g(\theta)$, we split the data set $\mathbf{F} = \{g(\theta)\}$ into two groups, a training set $\mathbf{F}_{\mathrm{train}} = \{g(\theta)\}_{\mathrm{train}}$ consisting of 70% of **F**, and a test set $\mathbf{F}_{\mathrm{test}} = \{g(\theta)\}_{\mathrm{test}}$ consisting of the remaining 30%. We used the training set to optimize the GPR, especially the hyperparameters (l, σ) for each system parameter individually by maximizing the log marginal likelihood using gradient descent (Williams & Rasmussen, 2006). We then used the trained GPR to predict the system parameters of the test set and compare the GPR predicted system parameters with those actually used for the MC simulations.

Table 1
Optimized hyperparameters for each feature, obtained from the maximum log marginal likelihood.

	l	σ
R_s	1.338×10^{-1}	2.673×10^{-3}
D_2	2.419×10^{-1}	2.717×10^{-3}
γL	1.405×10^{-1}	1.927×10^{-2}

Fig. 5 shows the determination of the log marginal likelihood contour in the (l, σ) space for each system parameter $(R_s, D_2, \gamma L)$. To gauge the appropriate range of l for each system parameter, as shown in Fig. 5(*a*), we first analyzed the PDDF of $g(\theta) \in \mathbf{F}$ and then investigated the ACF for $(R_s, D_2, \gamma L)$, which gave us a rough range in which we could search for the optimized l. The resulting log marginal likelihood contours are shown in Figs. 5(*b*)–5(*d*), and the values of optimized (l, σ) are shown in Table 1.

We applied the trained GPR with the optimized hyperparameters on the test set $\mathbf{F}_{\mathrm{test}}$ to infer ML inverted system parameters $(R_s, D_2, \gamma L)$ and compared the inferred results with MC references. Fig. 6 shows the comparison of system

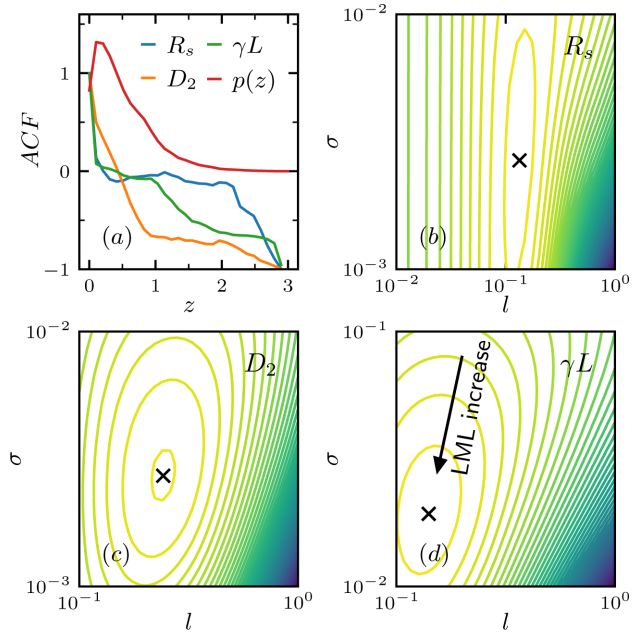


Figure 5
Determining the hyperparameters l and σ for system parameters. (*a*) PDDF $p(z)$ of the data set **F** and ACF for three system parameters. Log marginal likelihood of hyperparameters for system parameters: (*b*) R_s, (*c*) non-affine rearrangement D_2 and (*d*) affine shear strain γL.

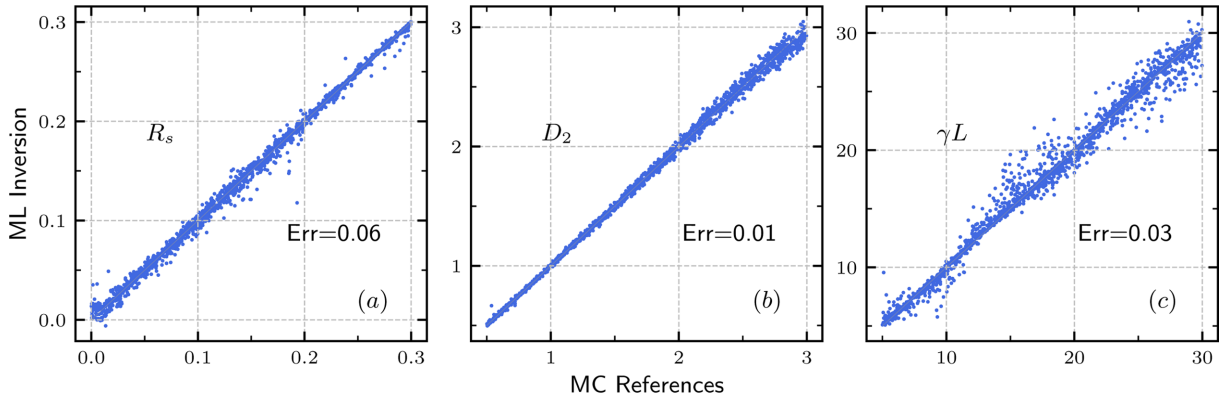


Figure 6
Comparison between the system parameters extracted from the angular scattering correlation function $g(\theta)$ using the GPR and their corresponding MC reference used for generating scattering data. Averaged relative error Err is indicated in each plot. (*a*) Variation of particles size R_s. (*b*) Non-affine rearrangement D_2. (*c*) Affine shear strain γL.

parameters (R_s, D_2, γL). Almost all of the data points lie around the diagonal line, indicating a good estimation. For each system parameter μ, the relative error between the MC reference μ_{MC} and ML inversion μ_{ML} is estimated by $\text{Err} = \langle|\mu_{MC} - \mu_{ML}|/\max(\mu_{MC}, \mu_{ML})\rangle$, where $\langle\ldots\rangle$ here is averaged over all data points. The relative error is labeled for each system parameter in each panel of Fig. 6 and shows a very high precision: 1% for D_2, 3% for γL and 6% for R_s. The precise quantification of system parameters demonstrates the power of our ML approach for analyzing and extracting microscopic rearrangement from coherent scattering data.

4. Summary

We have presented an ML-informed analysis framework that successfully recovers the polydispersity and microscopic rearrangements, including both affine simple shear and non-affine transformation, with high precision from the correlation function $g(\mathbf{q})$ of coherent scattering intensity of the dilute dispersions of spherical particles.

Our simulated colloidal systems and scattering intensity aim to mirror real synchrotron scattering setups, including the beam size, particle size and range of the detectable wavevector; therefore, our SVD features and GPR models can easily be compared and adopted to analyze real experimental data. The direction of affine shear is not necessarily always in the x direction for experimental data. However, the direction is easy to identify by the high correlation strip in the $g(\mathbf{q})$ pattern; one can rotate the scattering data before applying our model. Currently, our approach applies to dilute dispersions of spherical colloids. If the system deviates significantly from these assumptions, such as involving anisotropic particles or nonlinear turbulent flow, our model would require retraining with simulations that capture the corresponding scattering behavior. Additionally, incorporating experimental data sets labeled with known microscopic rearrangements, such as shear-driven viscous laminar fluids (Aime & Cipelletti, 2019), could further enhance its applicability. By leveraging such data, the model could generalize better to real-world experiments, including those with sparse or incomplete data sets. If the experimental scattering data are taken on a different grid of wavevector $\mathbf{q}$, interpolated data can be utilized to feed into our GPR model. Alternatively, we can also use generative models such as Kolmogorov–Arnold networks (Liu *et al.*, 2024*b*; Liu *et al.*, 2024*a*) to obtain $I(\mathbf{q})$ as a continuous function of the system parameters and use it to directly fit the experimental data.

Further, similar methods can be deployed to study the microscopic rearrangement of disordered colloidal systems, like glasses and gels, where ML-assisted quantification methods have a great potential to overcome the challenges imposed by the out-of-equilibrium nature (Schoenholz *et al.*, 2016; Horwath *et al.*, 2024). However, for such highly concentrated systems, the MC direct sampling method often suffers from a high rejection rate. As a result, alternative techniques like Brownian dynamics or molecular dynamics simulations are better suited for capturing the intricate dynamics and interactions in these systems.

APPENDIX *A*

Singular value decomposition of the data set

Fig. 7(*a*) shows the singular value versus its rank, where the rapid decay of the singular value indicates that the significance

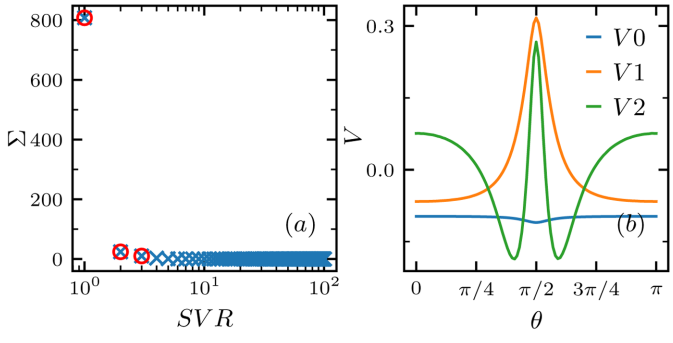


Figure 7
SVD of the angular scattering correlation data set. (*a*) Singular value Σ versus singular value rank (SVR); the top three ranked values are highlighted with red circles. (*b*) Singular vectors corresponding to the first three singular values.

of the projection onto higher-rank singular vectors quickly becomes negligible. Therefore, decomposition of $g(\theta)$ into the top three singular vectors will provide a good approximation of the whole $g(\theta)$. Fig. 7(*b*) shows the singular vectors ($V1$, $V2$, $V3$) corresponding to the first three singular values.

Acknowledgements

YC thanks R. L. Leheny and A. G. Yodh for helpful discussions.

Funding information

This research was performed at the Spallation Neutron Source, which is a US Department of Energy (DOE) Office of Science User Facility operated by Oak Ridge National Laboratory (ORNL). This research was sponsored by the Laboratory Directed Research and Development Program of ORNL, managed by UT-Battelle, LLC, for the US DOE. The ML aspects were supported by by the US DOE Office of Science, Office of Basic Energy Sciences Data, Artificial Intelligence and Machine Learning at DOE Scientific User Facilities Program (award No. 34532). Monte Carlo simulations and computations used resources of the Oak Ridge Leadership Computing Facility, which is supported by the US DOE Office of Science (contract No. DE-AC05-00OR22725), and resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy under contract No. DE-AC02-05CH11231. YC acknowledges support by the National Science Foundation (NSF) (grant No. DMR-2003659) and by the NSF Penn Materials Research Science and Engineering Center (grant No. DMR-2309043).

References

Aime, S. & Cipelletti, L. (2019). *Soft Matter* **15**, 200–212.
Akepati, S. V. R., Gupta, N. & Jayaraman, A. (2024). *JACS Au* **4**, 1570–1582.
Anker, A. S., Butler, K. T., Selvan, R. & Jensen, K. M. (2023). *Chem. Sci.* **14**, 14003–14019.
Badon, A., Boccara, A. C., Lerosey, G., Fink, M. & Aubry, A. (2017). *Opt. Express* **25**, 28914–28934.
Banetta, L., Leone, F., Anzivino, C., Murillo, M. S. & Zaccone, A. (2022). *Phys. Rev. E* **106**, 044610.
Burghardt, W. R., Sikorski, M., Sandy, A. R. & Narayanan, S. (2012). *Phys. Rev. E* **85**, 021402.
Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L. & Zdeborová, L. (2019). *Rev. Mod. Phys.* **91**, 045002.
Chang, M., Tung, C., Chang, S., Carrillo, J. M., Wang, Y., Sumpter, B. G., Huang, G., Do, C. & Chen, W. (2022). *Commun. Phys.* **5**, 46.
Chen, S.-H. (1986). *Annu. Rev. Phys. Chem.* **37**, 351–399.
Chen, Y., Rogers, S. A., Narayanan, S., Harden, J. L. & Leheny, R. L. (2020*a*). *Phys. Rev. Mater.* **4**, 035602.
Chen, Y., Rogers, S. A., Narayanan, S., Harden, J. L. & Leheny, R. L. (2020*b*). *Phys. Rev. E* **102**, 042619.
Chu, B. & Hsiao, B. S. (2001). *Chem. Rev.* **101**, 1727–1762.
Clegg, P. S. (2021). *Soft Matter* **17**, 3991–4005.
Ding, L., Tung, C.-H., Cao, Z., Ye, Z., Gu, X., Xia, Y., Chen, W.-R. & Do, C. (2024*a*). *arXiv*, 2411.00134.
Ding, L., Tung, C.-H., Carrillo, J.-M. Y., Chen, W.-R. & Do, C. (2025). *arXiv*, 2501.14647.
Ding, L., Tung, C.-H., Sumpter, B. G., Chen, W.-R. & Do, C. (2024*b*). *arXiv*, 2410.05574.
Donley, G. J., Narayanan, S., Wade, M. A., Park, J. D., Leheny, R. L., Harden, J. L. & Rogers, S. A. (2023). *Proc. Natl Acad. Sci. USA* **120**, e2215517120.
Galloway, K., Teich, E., Ma, X., Kammer, C., Graham, I., Keim, N., Reina, C., Jerolmack, D., Yodh, A. & Arratia, P. (2022). *Nat. Phys.* **18**, 565–570.
Goldburg, W. I. (1999). *Am. J. Phys.* **67**, 1152–1160.
Guinier, A., Fournet, G., Walker, C. B. & Yudowitch, K. L. (1955). *Small-angle scattering of X-rays*. Wiley.
He, H., Liang, H., Chu, M., Jiang, Z., de Pablo, J. J., Tirrell, M. V., Narayanan, S. & Chen, W. (2024). *Proc. Natl Acad. Sci. USA* **121**, e2401162121.
Heil, C. M., Ma, Y., Bharti, B. & Jayaraman, A. (2023). *JACS Au* **3**, 889–904.
Hinderhofer, A., Greco, A., Starostin, V., Munteanu, V., Pithan, L., Gerlach, A. & Schreiber, F. (2023). *J. Appl. Cryst.* **56**, 3–11.
Horwath, J. P., Lin, X.-M., He, H., Zhang, Q., Dufresne, E. M., Chu, M., Sankaranarayanan, S. K., Chen, W., Narayanan, S. & Cherukara, M. J. (2024). *Nat. Commun.* **15**, 5945.
Huang, G., Tung, C., Porcar, L., Wang, Y., Shinohara, Y., Do, C. & Chen, W. (2023). *Macromolecules* **56**, 6436–6443.
Jana, R. & Pastewka, L. (2019). *J. Phys. Mater.* **2**, 045006.
Keim, N. C. & Arratia, P. E. (2015). *Soft Matter* **11**, 1539–1546.
Kotz, S., Balakrishnan, N. & Johnson, N. L. (2019). *Continuous multivariate distributions*, Vol. 1, *Models and applications*. John Wiley & Sons.
Krauth, W. (2006). *Statistical mechanics: algorithms and computations*, Vol. 13. OUP.
Leheny, R. L., Rogers, M. C., Chen, K., Narayanan, S. & Harden, J. L. (2015). *Curr. Opin. Colloid Interface Sci.* **20**, 261–271.
Leitner, M., Sepiol, B., Stadler, L.-M., Pfau, B. & Vogl, G. (2009). *Nat. Mater.* **8**, 717–720.
Liu, Z., Ma, P., Wang, Y., Matusik, W. & Tegmark, M. (2024*a*). *arXiv*, 2408.10205.
Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y. & Tegmark, M. (2024*b*). *arXiv*, 2404.19756.
Lu, S. & Jayaraman, A. (2024). *Prog. Polym. Sci.* **153**, 101828.
Lüttich, M., Giordano, V. M., Le Floch, S., Pineda, E., Zontone, F., Luo, Y., Samwer, K. & Ruta, B. (2018). *Phys. Rev. Lett.* **120**, 135504.
Ma, L., Zhang, F., Allen, A. & Levine, L. (2014). *Acta Cryst.* A**70**, 338–347.
Madsen, A., Fluerasu, A. & Ruta, B. (2020). *Synchrotron light sources and free-electron lasers: accelerator physics, instrumentation and science applications*, pp. 1989–2018. Springer.
Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.
Murphy, R. P., Riedel, Z. W., Nakatani, M. A., Salipante, P. F., Weston, J. S., Hudson, S. D. & Weigandt, K. M. (2020). *Soft Matter* **16**, 6285–6293.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. (2011). *J. Mach. Learn. Res.* **12**, 2825–2830.
Ruocco, N., Dahbi, L., Driva, P., Hadjichristidis, N., Allgaier, J., Radulescu, A., Sharp, M., Lindner, P., Straube, E., Pyckhout-Hintzen, W. & Richter, D. (2013). *Macromolecules* **46**, 9122–9133.
Schoenholz, S. S., Cubuk, E. D., Sussman, D. M., Kaxiras, E. & Liu, A. J. (2016). *Nat. Phys.* **12**, 469–471.
Shibayama, M. (2011). *Polym. J.* **43**, 18–34.
Shpyrko, O. G. (2014). *J. Synchrotron Rad.* **21**, 1057–1064.
Tomaszewski, P., Yu, S., Borg, M. & Rönnols, J. (2021). *Proceedings of the Swedish workshop on data science (SweDS)*, pp. 1–6. IEEE.
Tung, C., Chang, S., Chang, M., Carrillo, J., Sumpter, B. G., Do, C. & Chen, W. (2023). *Carbon Trends* **10**, 100252.

Tung, C.-H., Chang, S.-Y., Chen, H.-L., Wang, Y., Hong, K., Carrillo, J. M., Sumpter, B. G., Shinohara, Y., Do, C. & Chen, W.-R. (2022). *J. Chem. Phys.* **156**, 131101.
Tung, C.-H., Ding, L., Chang, M.-C., Huang, G.-R., Porcar, L., Wang, Y., Carrillo, J.-M. Y., Sumpter, B. G., Shinohara, Y., Do, C. & Chen, W. R. (2024*a*). *arXiv*, 2412.15474.
Tung, C.-H., Ding, L., Huang, G.-R., Porcar, L., Shinohara, Y., Sumpter, B. G., Do, C. & Chen, W.-R. (2025). *J. Appl. Cryst.* **58**, 523–534.
Tung, C. H., Hsiao, Y. J., Chen, H. L., Huang, G. R., Porcar, L., Chang, M. C., Carrillo, J. M., Wang, Y., Sumpter, B. G., Shinohara, Y., Taylor, J., Do, C. & Chen, W. R. (2024*b*). *J. Colloid Interface Sci.* **659**, 739–750.
Wen, Q., Basu, A., Janmey, P. A. & Yodh, A. G. (2012). *Soft Matter* **8**, 8039–8049.
Wessels, M. G. & Jayaraman, A. (2021). *Macromolecules* **54**, 783–796.
Williams, C. K. & Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, Vol. 2. MIT Press.
Wu, Q., Miao, W., Zhang, Y., Gao, H. & Hui, D. (2020). *Nanotechnol. Rev.* **9**, 259–273.
Wu, Z. & Jayaraman, A. (2022). *Macromolecules* **55**, 11076–11091.