





A machine learning inversion scheme for determining interaction from scattering

Ming-Ching Chang¹, Chi-Huan Tung ^{2,3}, Shou-Yi Chang ², Jan Michael Carrillo⁴, Yangyang Wang⁴, Bobby G. Sumpter ⁴✉, Guan-Rong Huang³, Changwoo Do³ & Wei-Ren Chen ³✉

Small angle scattering techniques have now been routinely used to quantitatively determine the potential of mean force in colloidal suspensions. However the numerical accuracy of data interpretation is often compounded by the approximations adopted by liquid state analytical theories. To circumvent this long standing issue, here we outline a machine learning strategy for determining the effective interaction in the condensed phases of matter using scattering. Via a case study of colloidal suspensions, we show that the effective potential can be probabilistically inferred from the scattering spectra without any restriction imposed by model assumptions. Comparisons to existing parametric approaches demonstrate the superior performance of this method in accuracy, efficiency, and applicability. This method can effectively enable quantification of interaction in highly correlated systems using scattering and diffraction experiments.

¹Department of Computer Science, University at Albany—State University of New York, Albany, NY 12222, USA. ²Department of Materials Science and Engineering, National Tsing Hua University, Hsinchu 300044, Taiwan. ³Neutron Scattering Division, Oak Ridge National Laboratory, Oak Ridge, TEN 37831, USA. ⁴Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TEN 37831, USA. ✉email: sumpterb@ornl.gov; chenw@ornl.gov

Measuring interaction between particles in condensed matter has been of paramount interest since it provides a starting point for describing the statistical properties of the system under consideration. Elastic scattering and diffraction techniques have played an important role in this continued endeavor: From the measured spectra, extensive effort has been devoted to inferring the nature of forces that govern the properties of a variety of highly correlated disordered systems including dense atomic liquids^{1–4}, ionic liquids⁵ such as electrolytes⁶ and molten salts^{7,8}, molecular fluids^{9,10}, suspensions of colloids, micelles and emulsions^{11–13}, solutions of polymers^{14,15}, and polyelectrolytes^{16,17}. Nonetheless, potential inversion based on this experimental protocol is often hampered by the difficulty of precisely modeling the two-point static correlation functions, the quantity of interest in radiation scattering experiments, in terms of the relevant parameter.

In this report, we present a non-parametric strategy to circumvent the intrinsic limitation of existing approaches and demonstrate its feasibility by a case study of charged colloidal suspensions, a representative soft matter system. When the long-range electrostatic repulsion dominates overall distances, the effective interaction between charged colloids can be described by a hard sphere with a Yukawa tail of screened Coulomb repulsion¹⁸:

$$\beta V_{HSY}(r) = \begin{cases} \infty, & \text{if } r < D \\ A \frac{\exp[-\kappa(r-D)]}{r}, & \text{otherwise} \end{cases} \quad (1)$$

where $\beta \equiv \frac{1}{k_B T}$ is the Boltzmann factor, A the coupling parameter defined as $\frac{Z^2 e^2}{\epsilon(1+\kappa D)^2}$, D the colloidal diameter, Z the charge number, e the electric charge, ϵ the dielectric constant of the solvent, and κ the Debye screening constant. Eq. (1) has been extensively used to model the electrostatic interaction in a wide variety of charged colloidal systems including ionized nanoparticles¹¹, self-assemblies¹⁹, and biological systems²⁰. The relevant correlation function is the interparticle structure factor $S(Q)$ in reciprocal Q space. One well-adopted approach to determine $V_{HSY}(r)$ from the measurable $S(Q)$ is through the Ornstein-Zernike (OZ) integral equation^{21–23}:

$$h(r_{12}) = c(r_{12}) + n_p \int c(r_{13})h(r_{23})d^3r_3, \quad (2)$$

where $h(r_{12}) \equiv g(r_{12}) - 1$ and $g(r)$ is the pair distribution function, $c(r_{12})$ is the direct correlation function, and n_p is the particle number density. $S(Q)$ is the Fourier transform of $g(r)$. Since both $h(r)$ and $c(r)$ are unknown, a second closure equation is required to solve Eq. (2). Several closures have been developed as parametric^{12,22,24} or non-parametric²⁵ approximations. Despite the popularity of the integral equation approach, existing studies have indicated its limitations: Because the accuracy of a given closure is not known a priori, its validity in any specific phase region needs to be justified computationally^{26–33}. Moreover, the convergence behavior of each closure is found to depend on the complexity of adopted numerical procedures³⁴. For highly charged systems, the extraction of $V_{HSY}(r)$ from the corresponding $S(Q)$ is also known to be compounded by the strong electrostatic interactions³⁵. As alluded to above, no ideal closure is currently available.

Results and discussion

The position we take here is that developing another closure does not necessarily provide the most effective solution. Instead of dealing with the integral equations, iterative approaches including Boltzmann inversion¹ and force matching³⁶ are also developed to solve the potential inversion problem. Based on the predictor-

corrector scheme, these approaches have proved to be able to determine the effective pair potentials without assuming their mathematical form. However, a significant drawback emerged as in each step of the iteration an independent simulation is required to provide the connection between the predictor potential and its corresponding structural feature, which leads to the low efficiency. Another iterative scheme based on test particle insertion³⁷ was developed to circumvent the limitation of heavy computational load at a cost of the prerequisite of detailed individual particle trajectories instead of coarse-grained two-point correlation as input. These iterative approaches are also unable to provide a broad view of the relation between varying interaction potential and equilibrium structure owing to the fact that the criteria of convergence only focused on the neighborhood around the target pair correlation function.

From a Bayesian perspective³⁸ we instead sought to solve this inversion problem via a machine learning (ML) approach based on Gaussian process³⁹, which defines a distribution over functions as a conceptual extension of the familiar Gaussian distribution. Using Eq. (1) we computationally generated an extensive library of $S(Q)$ from the equilibrium fluid phase (see Methods), which is defined as the training set $\{S_{train}(Q)\}$ in this study. In the vector space of $S(Q)$, we treated the A , κ , and n_p of $\{S_{train}(Q)\}$ as a collection of normally distributed random variables and accordingly determined the statistical relationship of each variable in the ML process (see Methods). Using the optimized covariance matrix designed to quantify data correlation, we were able to probabilistically infer the values of A , κ , and n_p from a given $S(Q)$ without having to rely on a prescribed parametric equation, such as OZ equation and a complementary closure, to specify this mathematical relationship deterministically. The uncertainties associated with the related parameters naturally emerge during the inference process embedded in the spectral analysis procedure.

Feasibility of potential inversion. Before implementing the ML process, it is critical to first examine the feasibility of our proposed approach. To uniquely determine the potential parameters from a given $S(Q)$, a necessary condition to meet is the separability of the dataset in the vector space of $S(Q)$ of dimension 80, the sampled Q points in our computational trajectory analysis. Here we used a principal component analysis (PCA) by the singular value decomposition (SVD)⁴⁰ to extract relevant information of the data distribution in this high dimensional vector space (see Methods). From the results of singular value analysis⁴¹ presented in Fig. 1a, b, it is confirmed that the variance of original data is mostly retained by the first three statistically significant singular value ranks. The vector space \mathbb{R}^3 spanned by these three singular vectors, denoted as SVD0, SVD1, and SVD2 in Fig. 1c, is therefore sufficiently expressive in capturing essential features of the correlated data.

In this \mathbb{R}^3 vector space, each $S(Q)$ in $\{S_{train}(Q)\}$ is represented by a point and the distribution of a gridded dataset is given in Fig. 2. As demonstrated by Fig. 2a, it is seen to be narrowly distributed along one dimension of a twisted three-dimensional manifold which resembles the shape of a half fern leaf: The elongated central axis represents the data of hard-sphere fluids with different volume fraction ϕ defined as $\frac{\pi}{6} n_p D^3$. Each pinna bursting forth from the central stalk consists of data points with the same ϕ but different κ and A . We thoroughly examined the data distribution presented in Fig. 2a and no inseparable overlapping was found. This observation, which reflects the one-to-one mapping between $S(Q)$ and a set of ϕ , $\frac{1}{\kappa D}$, and βA in the colloidal fluid phase, provides initial support for a viable framework capable of inversely extracting potential parameters

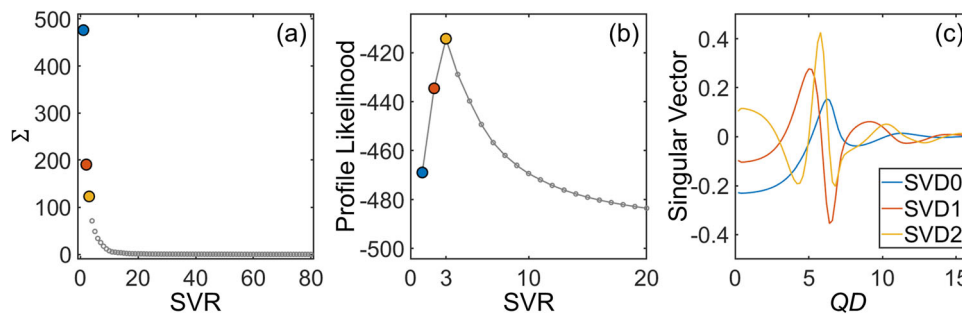


Fig. 1 Principal component analysis of the $S(Q)$ dataset. **a** The eigenvalues of principal components. SVR stands for singular value rank and the singular value is represented by Σ . As demonstrated in panel **b**, the results of maximum profile likelihood analysis⁴¹ show that the linear subspace spanned by the first three singular vectors given in Panel **c** is sufficiently expressive to illustrate the features of the correlated data points. SVD0, SVD1, and SVD2 are presented as a function of dimensionless unit QD.

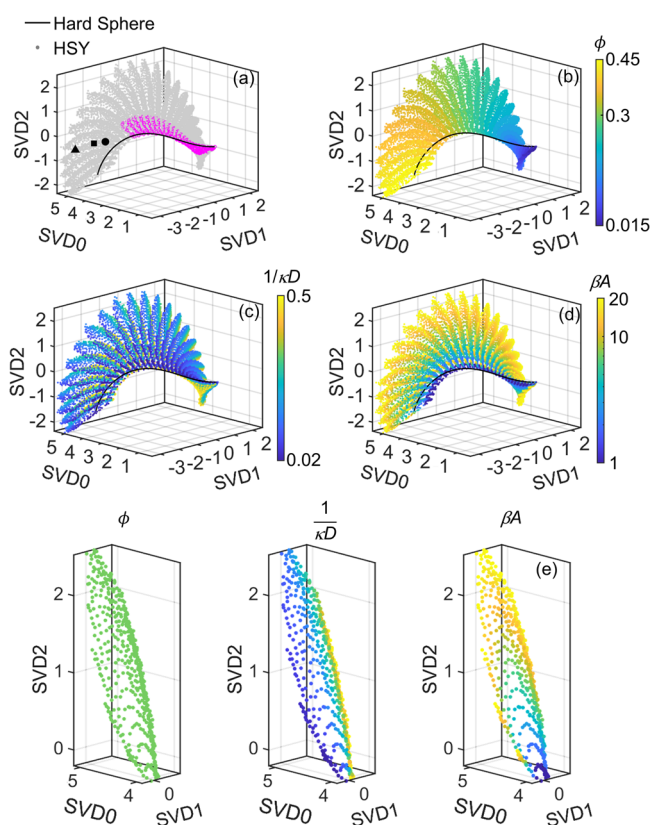


Fig. 2 Distributions of potential parameters in the space spanned by leading singular vectors. **a** The distributions of training data of hard-sphere Yukawa (HSY) in the \mathbb{R}^3 vector space spanned by SVD0, SVD1, and SVD2 given in Fig. 1c. The magenta symbols represent the fluid phase regions where the effective interaction can be precisely determined by solving the OZ equation with MPB-RMSA closure^{32,33}. The black symbols represent the three highly correlated charged colloidal suspensions given in Fig. 4. Panels **b-d** give the distributions of ϕ , $\frac{1}{\kappa D}$ and βA . The detailed distributions for $\phi = 0.3$ are given as an example in **e**.

from measured scattering functions. The data points collected from the fluid phase characterized by $A < 3 k_B T$ are further marked by magenta color in Fig. 2a. In this phase region, the effective interaction of charged colloidal suspensions can be precisely determined by OZ equation complemented by the MPB-RMSA closure proposed by Heinen and coworkers^{32,33}, a sophisticated approach among the continued efforts^{42,43} aiming

at improving the MSA closure⁴⁴. Within probed phase regions, its quantitative accuracy is found to be equivalent to that of the Rogers-Young closure⁴⁵ but the computational efficiency is significantly improved. One can further label the data points presented in Fig. 2a with the numerical values of ϕ , $\frac{1}{\kappa D}$, and βA to examine the characteristics of their distributions. Judging from the results given in Fig. 2b–d, the distributions of ϕ , $\frac{1}{\kappa D}$ and βA are all seen to vary smoothly. This observation suggests that the data points of these three parameters not only are self-avoiding but also highly correlated with a certain length scale. One can therefore uniquely extract the potential parameters from a given $S(Q)$ in its vector space.

Having verified the feasibility of our proposed approach for spectral inversion, we can now readily demonstrate its numerical reliability. For this purpose it is instructive to further investigate the difference in the heterogeneity of distribution for ϕ , $\frac{1}{\kappa D}$, and βA . We found that the distribution of ϕ is a more slowly varying function in comparison to that of $\frac{1}{\kappa D}$ and βA . As exemplified by Fig. 2e, for the pinna of $\phi = 0.3$, both $\frac{1}{\kappa D}$ and βA are seen to vary characteristically within $\Delta x \sim 1$ and $\Delta x \sim 2$, respectively, where $\Delta x = |x - x'|$ and x, x' are the coordinates in the vector space of $S(Q)$. In addition, the corresponding gradient vectors of $\frac{1}{\kappa D}$ and βA are seen to point perpendicularly across and axially along the blade respectively. Note that the absolute magnitude of gradients expressed as $\lim_{\Delta x \rightarrow 0} \left| \frac{\Delta y}{\Delta x} \right|$ where Δy is the difference of corresponding potential parameters are analogous to the sensitivity metrics⁴⁶ defined as the ratio of the separation between structural features and that between their corresponding interaction potentials. For example, in the high-temperature limit of $\beta A \rightarrow 0$, $S(Q)$ of distinct $\frac{1}{\kappa D}$ distributed in a narrow region around the central stalk, such distribution corresponds to an extremely high gradient and thus implies low sensitivity in this region. These observations provided a comprehensive aspect of the feasibility of potential inversion. Following the model of Yukawa potential we can express the discrepancy between interparticle potentials solely in terms of the related physical quantities, which benefits the systematic exploration of structure-interaction relations in the phase diagram. Clearly, the numerical accuracy of potential extraction depends on how well this observed distributional heterogeneity is addressed.

Gaussian process regression. In the context of Gaussian Process (GP)^{38,39,47}, we developed a covariance matrix, as the beating heart of our non-parametric inversion method, to quantitatively describe the statistical relationships of ϕ , $\frac{1}{\kappa D}$ and βA in the vector space of $S(Q)$ (see Methods). A function f relating $S(QD)$ and A, κ and n_p can be formulated as $f \sim \mathcal{GP}(\mu, k)$ in terms of a prior

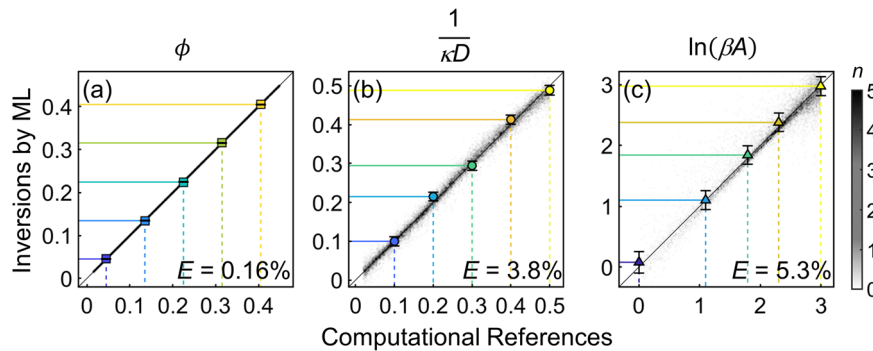


Fig. 3 Comparison of potential parameters in simulations and those inverted by machine learning. The comparison of the extracted **a** ϕ , **b** $\frac{1}{\kappa D}$ and **c** $\ln(\beta A)$ and their corresponding computational references. The numerical accuracy of reconstruction is quantified by the relative error E given in each panel.

mean function μ and a prior covariance function k . Given a training set (X, Y) , where X represents the n sets of $S(QD)$ in $\{S_{train}(Q)\}$ and Y represents the corresponding n sets of regression targets of A , κ and n_p , the purpose of the ML process is to determine μ and k from the knowledge of training data. Following the standard procedure of GP³⁹, a constant function is used to specify μ . The $n \times n$ covariance matrix K_{XX} specifies the correlations between training data pairs modeled by the radial basis function (RBF) kernel (see Methods). The Gaussian observational noise term is added to the kernel matrix $K = K_{XX} + \sigma^2 I$ where σ describes the variance of observational noise, and I denotes identity matrix. The hyperparameters l and σ are determined by maximizing the log marginal likelihood using a gradient descent algorithm during training^{39,48}. Given a test point X_* , the goal of GPR is to estimate $Y_* = f(X_*)$, where the covariance matrix of the test set is denoted as $K_{X_*X_*}$, and K_{XX_*} consists of entries measuring the correlations between training and test points. For simplicity, $K \equiv K_X$, $K_* \equiv K_{X_*}$, and $K_{**} \equiv K_{X_*X_*}$. For a $S(QD)$ measured from the equilibrium fluid phase, the relevancy of its A , κ and n_p with those in the training sets should follow the same correlation patterns deduced from the training process. Therefore, regression targets can be determined from this joint distribution:

$$\begin{bmatrix} Y \\ Y_* \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mu(X) \\ \mu(X_*) \end{bmatrix}, \begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix} \right), \quad (3)$$

where T indicates matrix transposition and \mathcal{N} denotes normal distribution.

Another extensive set of $S(Q)$ termed as $\{S_{test}(Q)\}$ was simulated separately from the equilibrium fluid phase to gauge the numerical accuracy of our proposed method. Figure 3 presents the comparison of input parameters in simulations and those inverted from $\{S_{test}(Q)\}$. All three extracted parameters are in remarkable accord with their computational inputs, but ϕ is in closer quantitative agreement as indicated by the relative error E . The origin of this varying degree of uncertainty is worth exploring: As illustrated in Fig. 2e, both $\frac{1}{\kappa D}$ and βA exhibit extreme changes within a relatively short Euclidean distance in comparison to the large-scale variation of ϕ . This observation suggests that the susceptibility of $S(Q)$ towards the variation of potential variables is inherently determined by their distributions in the vector space of $S(Q)$. The difference in E can therefore be viewed as a reflection of landscape heterogeneity for different parameters.

In Fig. 4a–c we give the comparison of $V_{HSY}(r)$ calculated from the ML-extracted parameters specified by the solid lines in Fig. 3 and their corresponding references used in the simulation. To demonstrate the numerical accuracy of our approach, the $S(Q)$

of three highly correlated systems, as indicated by the black symbols in Fig. 2a, were calculated and compared with their references in $\{S_{train}(Q)\}$. Accordingly the corresponding coherent intensities of small-angle neutron scattering (SANS) $I(Q)$ were also calculated assuming the system consists of monodisperse densely-packed spherical particles (see Methods). From the results presented in Fig. 4d–f, it is clearly seen that the differences between the reference $I(Q)$ and $S(Q)$ (blue dashed curves) and the reconstructed $I(Q)$ and $S(Q)$ (red solid curves) are indeed indistinguishable on the scale of our plots. This observation demonstrates that the structural variation of suspending charged particles caused by the change of $V_{HSY}(r)$ with the magnitude of E given in Fig. 3a–c cannot be detected by elastic scattering techniques. The insets in Fig. 4d–f also show that the difference in $I(Q)$ is within the range of statistical variation of resolution of general SANS instruments. For the three strongly correlated systems presented in Fig. 4d–f, integral equation theories are often unable to determine ϕ , κ , and A from $S(Q)$ in a numerically precise manner³⁵. The validity of our proposed ML approach for quantitatively extracting $V_{HSY}(r)$ from the scattering spectra of charged colloidal suspensions over a wide range of ϕ in the equilibrium fluid phase is therefore verified.

Conclusions

In conclusion, we have developed a ML inversion method, based on the framework of Gaussian process, to inversely determine the effective interaction of colloidal suspensions from their scattering spectra. By treating the probability distributions of the relevant potential parameters in the vector space of the scattering function, our non-parametric approach circumvents the mathematical constraints inherent to deterministic models for spectral analysis. We demonstrated that our approach offers several advantages over the existing parametric approaches from the standpoint of numerical accuracy, computational efficiency (see Methods), and general applicability. While the ML method present in this work relied on the model of Yukawa potential unlike the aforementioned model-free iterative approaches, this method still provided crucial physical insights in determining the coarse-grained interactions as it directly yielded the related physical quantities. Complemented by computer simulations, our method can be systematically extended for solving the inversion scattering problems of various colloidal systems characterized by different effective interactions^{49–52}. Moreover, one recognized challenge in the analysis of scattering data of interacting systems is to take polydispersity into account adequately⁵³. In this pursuit, our ML strategy can enable quantitative characterization of highly interacting systems characterized by significant variations of interaction potential and particle size which cannot be addressed precisely by existing decoupling approximations^{18,54,55}. Another

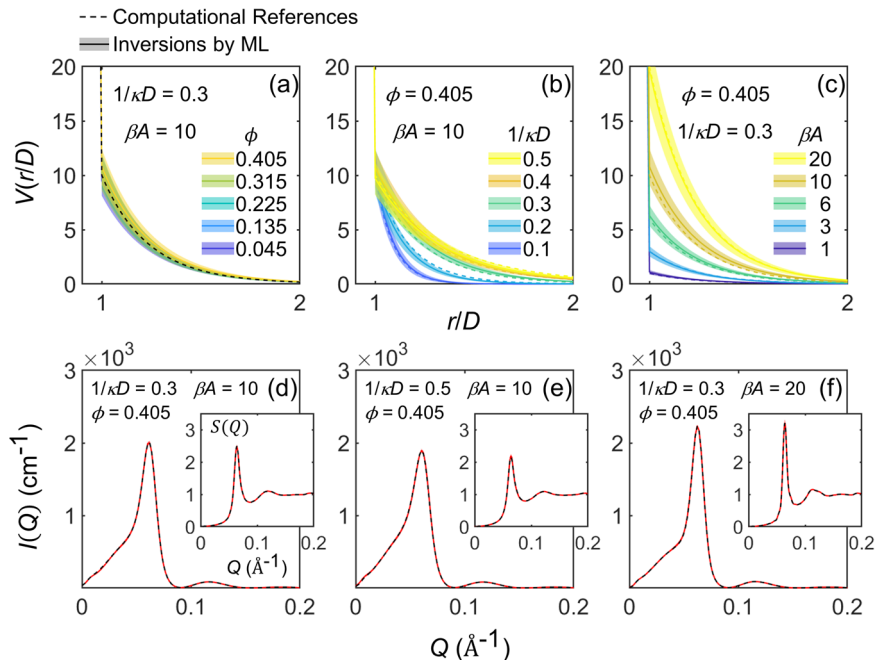


Fig. 4 Reconstructed interparticle potentials and scattering functions. Panels **a–c** present examples of $V_{HSY}(r)$ calculated based on ϕ , κ , and A given Fig. 3. The references used in computer simulations and those determined by ML inversion are presented in dashed and solid curves, respectively. The magnitude of uncertainties calculated based on the errors in Fig. 3 are represented by the shaded regions. The color scheme is the same as that of Fig. 3. Panels **d–f** give the reconstructed $S(Q)$ (insets) and $I(Q)$ calculated from three highly correlated systems (yellow curves in panels **a–c**) and upper right most points in Fig. 3. Both computationally generated reference $S(Q)$ and $I(Q)$ (black dashed curves) and their ML reconstructions (red solid curves) are indistinguishable within the scales used in these plots.

merit of our strategy is those parameters affect the scattering spectra yet independent of the pair interaction, for example, the bulk number density can also be inferred probabilistically from ML. In contrast, the conventional iterative approaches require the knowledge of the number density of particles in advance, which is in general unavailable in SANS experiments, especially for the micellar solutions.

In principle, the applicability of this method is not restricted to structural studies of colloidal systems. The essential idea underlying this method of spectral analysis, which does not suffer from the drawbacks of explicit modeling, allows quantitative extraction of relevant parameters based on which the targeted systems are computationally constructed. We are optimistic that our approach will provide a useful toolbox to facilitate the progress in many important inversion problems of radiation scattering, diffraction, and imaging^{56–59} experiments from strongly correlated systems which traditionally have been difficult and time-consuming to solve.

Methods

Molecular dynamics (MD) simulation. Using molecular dynamics (MD) simulation, 27,000 $S(QD)$ samples of charged colloidal suspensions in their equilibrium fluid phase were simulated based on Eq. (1). The ranges of parameters are chosen to be $0 < \beta A < 20$, $0 < \phi < 0.45$, and $0 < \frac{1}{\kappa D} < 0.5$ respectively, so that no splitting in the second peak⁶⁰ of $S(QD)$, which signifies the formation of the non-ergodic glassy state is observed. Based on the form of $\mu \equiv \{\phi, \kappa, A\}$ we separate the simulated $S(QD)$ into two main subsets:

- Grid15K contains grid sample points with 30 unique ϕ values: $\phi \in (0.015, 0.03, 0.045, \dots, 0.45)$, 25 unique κ values: $\frac{1}{\kappa D} \in (0.02, 0.04, 0.06, \dots, 0.5)$, and 20 unique A values: $\beta A \in (1, 2, 3, \dots, 20)$, such that there are totally $30 \times 25 \times 20 = 15,000$ sample points. It is the $\{S_{train}(Q)\}$.
- Rand12K covers the specified μ range with 12,000 uniform distributed random data points. It is the $\{S_{test}(Q)\}$.

The simulation cell contains 16,384 particles with an initial number density of $n_p = \frac{3}{4\pi} \phi \delta^{-3}$. Canonical (NVT) ensemble simulation was performed, where the temperature was maintained at $T = 1.0$ via a Nose-Hoover thermostat^{61,62} and

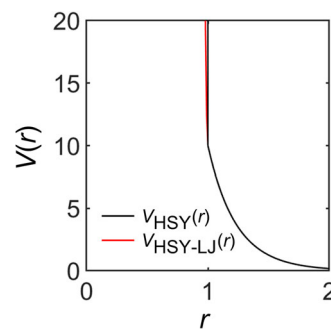


Fig. 5 Potential model used in the simulation. The hard-sphere Yukawa potential (black curve) and the approximated potential used in our simulations (red curve) by replacing the hardcore with truncated-shifted Lennard-Jones interaction.

integrated using the velocity-Verlet algorithm⁶³ with integration timestep of 0.001τ . Here δ , $\tau = \delta(\beta m)^{\frac{1}{2}}$, and $m = 1$ are the standard LJ reduced units for distance, time, and mass respectively. The excluded hardcore of $D = 1\delta$ in $V_{HSY}(r)$ used in our simulation is modeled by the repulsive component of 12-6 truncated-shifted Lennard-Jones (LJ) model²¹:

$$V_{LJ}(r) = \begin{cases} 4e \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^6 + \frac{1}{4} \right], & \text{if } r < 2^{\frac{1}{2}}\sigma \\ 0, & \text{if } r \geq 2^{\frac{1}{2}}\sigma \end{cases} \quad (4)$$

where $\beta\epsilon = 500$ and $\sigma = 2^{-\frac{1}{2}}\delta$ in reduced units.

As displayed in Fig. 5, only marginal difference between the model potential $V_{HSY}(r)$ and $V_{HSY-LJ}(r)$ is observed. Each simulation proceeded for 5τ where the simulation box size is changed to specify a value of ϕ , and another 25τ where the final 5τ was used to calculate $S(QD)$ using the existing procedure^{64,65}. Simulations were performed at the OLCF-4 Summit supercomputer at Oak Ridge National Laboratory (ORNL) using LAMMPS⁶⁶ with graphical processing unit (GPU) acceleration.

Examples of simulated $S(QD)$ are given in Fig. 6. Upon changing volume fraction $\phi = \frac{3}{4\pi} n_p D^3$ and A , its evolution is consistent with documented results (for example see refs. 11,12,32,33).

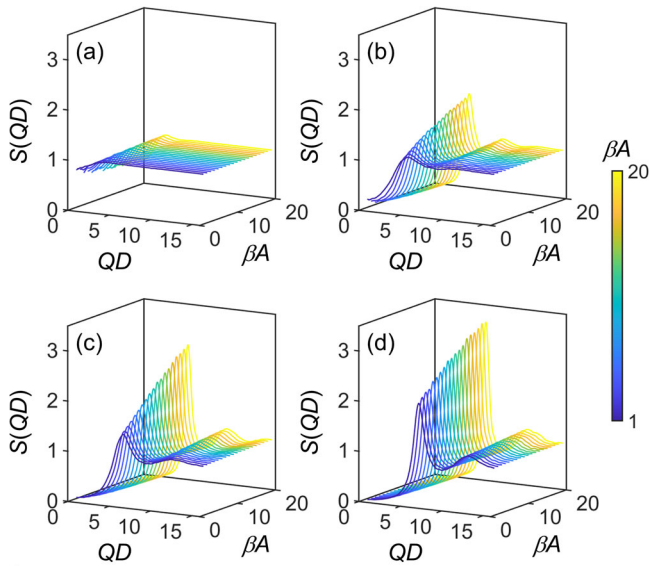


Fig. 6 Examples of simulated $S(QD)$. **a** $\phi = 0.015$, **b** $\phi = 0.15$, **c** $\phi = 0.30$, and **d** $\phi = 0.45$. The value of $\frac{1}{\kappa D} = 0.3$ and $0 < \beta A < 20$.

SVD and PCA analysis of MD data. We re-sampled each $S(QD)$ for $QD \in (0.2, 0.4, 0.6, \dots, 16.0)$ such that 80 QD sample points in each simulated $S(QD)$ are kept. This way, the dimension of the vector space of $S(QD)$ is 80. Each $S(QD)$ is therefore represented by a point in this \mathbb{R}^{80} vector space. Visualization of data distribution can be facilitated by dimensionality reduction: The data was arranged into a $80 \times 27,000$ matrix F . Using Singular Value Decomposition (SVD)⁴⁰, F can be decomposed into $F = U\Sigma V^T$.

The diagonal entries of Σ^2 are proportional to the percentages of the variance of the original data projected onto each corresponding principal axis. This Principal Component Analysis (PCA)⁴⁰ allows us to re-express the data as a set of new orthogonal variables to extract their intrinsic correlations. To remove the subjectivity involved in deciding the correct number of component axes to retain, we use the analysis of maximum profile likelihood⁴¹ to identify the statistical gap where the singular values begin to level off, as in Fig. 1a.

Determining the hyperparameters in machine learning from the data distributions. In the framework of the Gaussian Process, the covariance matrix K_{XX} specifies the correlations between training data pairs modeled by the radial basis function (RBF) kernel. Specifically, for $x, x' \in X$, kernels $k(x, x')$ as entries of K_{XX} are formulated by the following squared exponential expression:

$$k(x, x') = \exp\left[-\frac{(x - x')^2}{2l^2}\right], \quad (5)$$

where l is the correlation length which measures the similarity between training data points. The Gaussian observational noise term is added to the kernel matrix $K = K_{XX} + \sigma^2 I$ where σ describes the variance of observational noise, and I denotes identity matrix. To facilitate the optimization process of the hyperparameters l and σ , we first investigate the packing pattern of data in the manifold using the pair distance distribution function (PDDF)⁶⁷:

$$p(z) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \delta(|x_i - x_j| - z), \quad (6)$$

where δ is the Dirac delta function. The shape of PDDF (black curve) given in Fig. 7a suggests a nearly evenly distribution of data points⁶⁷.

Because no heterogeneous clustering of data is observed, the expectation values of the potential parameters associated with any given data point can be determined by its correlation between other data points through the optimized kernel functions used in GP. Information about the distributions of ϕ , $\frac{1}{\kappa D}$ and βA can be further deduced from their autocorrelation functions (ACF)⁶⁸:

$$C_\mu(z) = \frac{\langle \mu(x)\mu(x+z) \rangle_x - \langle \mu(x) \rangle_x^2}{\langle [\mu(x)]^2 \rangle_x - \langle \mu(x) \rangle_x^2}, \quad (7)$$

where $\langle \cdot \rangle_x$ represents the spatial average over position x of all sample points in the vector space. From the results given in Fig. 7(a) $C_\phi(z)$ is characterized by a slow relaxation with a characteristic length that is larger than the scale of the probed spatial domain. This observed decaying behavior is consistent with the smooth variation of ϕ revealed by Fig. 2(b). On the contrary, the self-correlations of κ and A are no longer retained when $z > 2$. This result gives a quantitative measure regarding the average length of their distributions shown in Fig. 2e. Because the length scale estimated by Equation (7) is a spatial range beyond which each

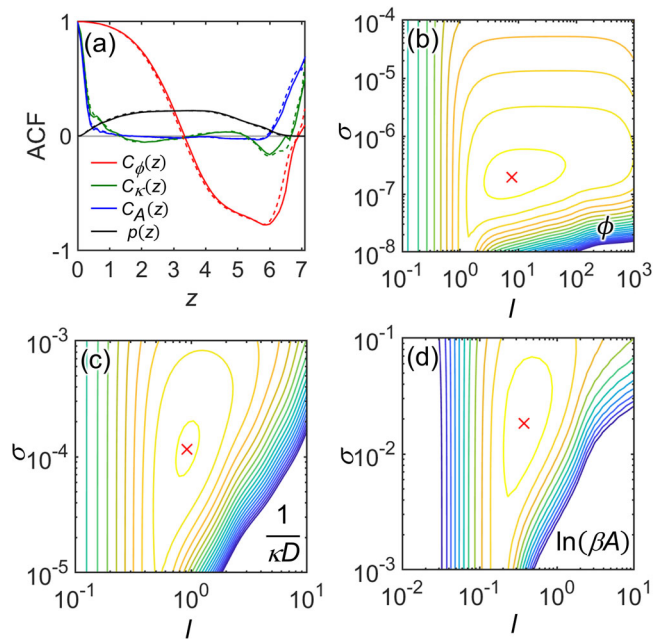


Fig. 7 The log marginal likelihood surfaces of l and σ . **a** ϕ , **b** $\frac{1}{\kappa D}$, and **c** $\ln(\beta A)$. The optimal values are marked by the red crosses. Panel **d** gives the PDDF $p(z)$ of $\{S_{train}(Q)\}$ (black curve), and the auto correlations of ϕ ($C_\phi(z)$, red curve), κ ($C_\kappa(z)$, green curve), and A ($C_A(z)$, blue curve). Due to the sparsity of available pairs, the upturns observed in the range of $z > 6$ lack statistical significance.

Table 1 The parameters l and σ of the radial basis function (RBF) kernel determined by gradient descent optimization.

θ	l	σ
ϕ	7.74	1.94×10^{-7}
$\frac{1}{\kappa D}$	0.922	1.16×10^{-4}
$\ln(\beta A)$	0.373	1.83×10^{-2}

l and σ are used to quantitatively represent the correlation and self-dispersion of the related potential parameters in the vector space of $S(Q)$.

parameter is allowed to vary randomly, it therefore defines the upper limit of the correlated spatial range of each parameter. Figure 7b–d presents the marginal likelihood surfaces over the two relevant parameters, the correlation length l and the variance of observational noise σ^2 , the central ingredients of the covariance matrix describing the correlation of training data on the framework of GP. For ϕ , $\frac{1}{\kappa D}$ and $\ln(\beta A)$, the contours are seen to be unimodal and convex with different degrees of skewness which reflects the characteristic patterns of their distributions in the vector space of $S(Q)$. The determination of the optimal l and σ^2 for ϕ , $\frac{1}{\kappa D}$ and βA is greatly facilitated by these monotonic features. The optimized l and σ^2 obtained from gradient descent^{39,48} are marked by red cross symbols in Fig. 7(b–d). It is instructive to indicate that the optimized l for each parameter is indeed less than its correlation length determined by ACF analysis given in Fig. 7a. In this paper, we adopt the `sklearn` `GaussianProcessRegressor` library⁶⁹ due to its efficiency and convenience of implementation. We give the results of optimization in Table 1. It is found that $l_\phi = 7.74$, $l_\kappa = 0.922$, $l_A = 0.373$, and $\sigma_\phi = 1.94 \times 10^{-7}$, $\sigma_\kappa = 1.16 \times 10^{-4}$, $\sigma_A = 1.83 \times 10^{-2}$, respectively. The numerical accuracy of these values is cross-validated. It is also noted that they are consistent with the calculations of ACF given in Fig. 7a.

The efficiency of our ML inversion approach. The computational demand of our method is extremely lightweight when compared to existing integral equation approaches. Our method does not require large-scale GPU cloud. The whole training process can be completed in hours, and testing can be completed in minutes or seconds. On a standard i7 laptop computer without GPU, our implementation of our ML process for 15,000 data points takes about 3 h to finish and a few minutes for testing all the other 12,000 data points. In general, solving an inversion problem from one $S(Q)$ curve takes less than 10^{-3} s. In comparison, it takes approximately 0.1 s to generate a $S(Q)$ from a specific combination of ϕ , κ

and A using the state-of-art integral equation approach^{32,33}. Moreover, for a general curve fitting routine, hundreds of iterations are often required to obtain satisfactory agreement between parametric models and experimental data. Overall the efficiency is improved significantly by our non-parametric ML approach in comparison to the existing parametric integral equation approaches.

Data availability

The datasets generated during the current study are available from both corresponding authors on request.

Code availability

The computer codes used for simulations and analyses are available from both corresponding authors upon request.

Received: 18 April 2021; Accepted: 1 December 2021;

Published online: 28 February 2022

References

- Schommers, W. Pair potentials in disordered many-particle systems: a study for liquid gallium. *Phys. Rev. A* **28**, 3599 (1983).
- Aers, G. C. & Dharma-wardana, M. W. C. Analysis of the structure factor of dense krypton gas: bridge contributions and many-body effects. *Phys. Rev. A* **29**, 2734 (1984).
- Levesque, D., Weis, J. J. & Reatto, L. Pair interaction from structural data for dense classical liquids. *Phys. Rev. Lett.* **54**, 451 (1985).
- Lyubartsev, A. P. & Laaksonen, A. Calculation of effective interaction potentials from radial distribution functions: a reverse Monte Carlo approach. *Phys. Rev. E* **52**, 3730 (1995).
- March, N. H. & Tosi, M. P. *Coulomb Liquids*. (Academic Press, 1984).
- Lyubartsev, A. P. & Laaksonen, A. Effective potentials for ion-DNA interactions. *J. Chem. Phys.* **111**, 11207 (1999).
- Babu, C. S. & Ichiye, T. New integral equation theory for primitive model ionic liquids: from electrolytes to molten salts. *J. Chem. Phys.* **100**, 9147 (1994).
- Caccamo, C. Integral equation theory description of phase equilibria in classical fluids. *Phys. Rep.* **274**, 1 (1996).
- Soper, A. K. Empirical potential Monte Carlo simulation of fluid structure. *Chem. Phys.* **202**, 295 (1996).
- Yamaguchi, T., Hidaka, K. & Soper, A. The structure of liquid methanol revisited: a neutron diffraction experiment at -80°C and $+25^\circ\text{C}$. *Mol. Phys.* **96**, 1159 (1999).
- Schmitz, K. S. *Macroions in Solution and Colloidal Suspension* (Wiley-VCH, 1992).
- Nagele, G. Long-time dynamics of charged colloidal suspensions: hydrodynamic interaction effects. *Phys. Rep.* **272**, 215 (1996).
- Likos, C. N. Effective interactions in soft condensed matter physics. *Phys. Rep.* **348**, 207 (2001).
- Yethiraj, A. & Schweizer, K. S. Self consistent polymer integral equation theory: Comparisons with Monte Carlo simulations and alternative closure approximations. *J. Chem. Phys.* **97**, 1455 (1992).
- Schweizer, K. S. & Curro, J. G. PRISM theory of the structure, thermodynamics, and phase transitions of polymer liquids and alloys. *Adv. Polym. Sci.* **116**, 319 (1994).
- Yethiraj, A. & Shew, C.-Y. Structure of polyelectrolyte solutions. *Phys. Rev. Lett.* **77**, 3937 (1996).
- Qiu, X. et al. Measuring inter-DNA potentials in solution. *Phys. Rev. Lett.* **96**, 138101 (2006).
- Pusey, P. *Liquids, Freezing and Glass Transition* (Les Houches 1989 Session LI) (eds Hansen, J. P., Levesque, D. & Zinn-Justin, J.) 763 (North-Holland, Amsterdam, 1991).
- Zemb, T. & Charpin, P. Micellar structure from comparison of X-ray and neutron small-angle scattering. *J. Phys.* **46**, 249 (1985).
- Tardieu, A. et al. Proteins in solution: from X-ray scattering intensities to interaction potentials. *J. Cryst. Growth* **196**, 193 (1999).
- Hansen, J.-P. & McDonald, I. *Theory of Simple Liquids*, 2nd edn., 97 (Academic Press, 1986).
- Schmitz, K. S. *Macroions in Solution and Colloidal Suspension*, 17 (Wiley-VCH, 1992).
- Nagele, G. *The Physics of Colloidal Soft Matter*, Lectures Notes Vol. 14 (Institute of Fundamental Technological Research/Polish Academy of Sciences, 2004).
- Klein, R. & D'Aguzzo, B. *Light Scattering: Principles and Development*, 30 (ed. Brown, W.) (Clarendon Press, 1996).
- Goodall, R. E. et al. Data-driven approximations to the bridge function yield improved closures for the Ornstein-Zernike equation. *Soft Matter* **17**, 5393 (2021).
- Belloni, L. Inability of the hypernetted chain integral equation to exhibit a spinodal line. *J. Chem. Phys.* **98**, 8080 (1983).
- Beresford-Smith, B., Chan, D. Y. C. & Mitchell, D. J. The electrostatic interaction in colloidal systems with low added electrolyte. *J. Colloid Interface Sci.* **105**, 216 (1985).
- Belloni, L. Attraction of electrostatic origin between colloids. *Chem. Phys.* **99**, 43 (1985).
- Belloni, L. *Neutron, X-Ray and Light Scattering: Introduction to an Investigative Tool for Colloidal and Polymeric Systems* (eds Zemb, Th. & Lindner, P.) 135 (North-Holland, Amsterdam, 1991).
- Fritz, G., Bergmann, A. & Glatter, O. Evaluation of small-angle scattering data of charged particles using the generalized indirect Fourier transformation technique. *J. Chem. Phys.* **113**, 9733 (2000).
- Banchio, A. J. & Nagele, G. Short-time transport properties in dense suspensions: from neutral to charge-stabilized colloidal spheres. *J. Chem. Phys.* **128**, 104903 (2008).
- Heinen, M., Holmqvist, P., Banchio, A. J. & Nagele, G. Pair structure of the hard-sphere Yukawa fluid: an improved analytic method versus simulations, Rogers-Young scheme, and experiment. *J. Chem. Phys.* **134**, 044532 (2011).
- Heinen, M., Holmqvist, P., Banchio, A. J. & Nagele, G. Erratum: Pair structure of the hard-sphere Yukawa fluid: an improved analytic method versus simulations, Rogers-Young scheme, and experiment. *J. Chem. Phys.* **134**, 044532 (2011).
- Hus, M., Zalar, M. & Urbic, T. Correctness of certain integral equation theories for core-softened fluids. *J. Chem. Phys.* **138**, 224508 (2013).
- Anta, J. A. & Lago, S. Self-consistent effective interactions in charged colloidal suspensions. *J. Chem. Phys.* **116**, 10514 (2002).
- Noid, W. et al. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **128**, 244114 (2008).
- Stones, A. E., Dullens, R. P. & Aarts, D. G. Model-free measurement of the pair potential in colloidal fluids using optical microscopy. *Phys. Rev. Lett.* **123**, 098002 (2019).
- Murphy, K. P. *Machine Learning: A Probabilistic Perspective*, 151 (The MIT Press, 2012).
- Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning*, 1 (The MIT Press, 2006).
- Strang, G. *Introduction to Linear Algebra*, 5th edn, 363 (Wellesley-Cambridge Press, 2016).
- Zhu, M. & Ghodsi, A. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Comput. Stat. Data Anal.* **51**, 918 (2006).
- Hansen, J.-P. & Hayter, J. B. A rescaled MSA structure factor for dilute charged colloidal dispersions. *Mol. Phys.* **46**, 651 (1982).
- Snook, I. K. & Hayter, J. B. Static structure of strongly interacting colloidal particles. *Langmuir* **8**, 2880 (1992).
- Hayter, J. B. & Penfold, J. An analytic structure factor for macroion solutions. *Mol. Phys.* **42**, 109 (1981).
- Rogers, F. J. & Young, D. A. New, thermodynamically consistent, integral equation for simple fluids. *Phys. Rev. A* **30**, 999 (1984).
- Wang, H., Stillinger, F. H. & Torquato, S. Sensitivity of pair statistics on pair potentials in many-body systems. *J. Chem. Phys.* **153**, 124106 (2020).
- Bishop, C. M. *Pattern Recognition and Machine Learning*, Ch. 6 (Springer, 2006).
- Wang, K. et al. In *Advances in Neural Information Processing Systems* Vol 32, pp 146–148 (Curran Associates, Inc., 2019).
- Asakura, S. & Oosawa, J. F. On interaction between two bodies immersed in a solution of macromolecules. *J. Chem. Phys.* **22**, 1255 (1954).
- Witten, T. A. & Pincus, P. A. Colloid stabilization by long grafted polymers. *Macromolecules* **19**, 2509 (1986).
- Liu, Y., Chen, W.-R. & Chen, S.-H. Cluster formation in two-Yukawa fluids. *J. Chem. Phys.* **122**, 044507 (2005).
- Bianchi, E., Blaak, R. & Likos, C. N. Patchy colloids: state of the art and perspectives. *Phys. Chem. Chem. Phys.* **13**, 6397 (2011).
- Salgi, P. & Rajagopalan, R. Polydispersity in colloids: implications to static structure and scattering. *Adv. Colloid Interface Sci.* **43**, 169 (1993).
- Kotlarchyk, M. & Chen, S.-H. Analysis of small angle neutron scattering spectra from polydisperse interacting colloids. *J. Chem. Phys.* **79**, 2461 (1983).
- Hayter, J. B. & Penfold, J. Determination of micelle structure and charge by neutron small-angle scattering. *Colloid Polym. Sci.* **261**, 1022 (1983).
- Carbajal-Tinoco, M. D., Castro-Román, F. & Arauz-Lara, J. L. Static properties of confined colloidal suspensions. *Phys. Rev. E* **53**, 3745 (1996).
- Rajagopalan, R. & Rao, K. S. Interaction forces in charged colloids: inversion of static structure factors. *Phys. Rev. E* **55**, 4423 (1997).
- Behrens, S. H. & Grier, D. G. Pair interaction of charged colloidal spheres near a charged wall. *Phys. Rev. E* **64**, 050401(R) (2001).

59. Quesada-Perez, M., Moncho-Jordá, A., Martínez-Lopez, F. & Hidalgo-Ivarez, R. Probing interaction forces in colloidal monolayers: inversion of structural data. *J. Chem. Phys.* **115**, 10897 (2001).
60. van de Waal, B. W. On the origin of second-peak splitting in the static structure factor of metallic glasses. *J. Non Cryst. Solids* **189**, 118 (1995).
61. Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **81**, 511 (1984).
62. Hoover, W. G. Canonical dynamics: equilibrium phase-space distributions. *Phys. Rev. A* **31**, 1695 (1985).
63. Verlet, L. Computer “Experiments” on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* **159**, 98 (1967).
64. Carrillo, J.-M. & Dobrynin, A. V. Polyelectrolytes in salt solutions: molecular dynamics simulations. *Macromolecules* **44**, 5798 (2011).
65. Frigo, M. A fast Fourier transform compiler. In *Proceedings of the ACM SIGPLAN 1999 conference on Programming language design and implementation*. 169–180 (1999).
66. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117**, 1 (1995).
67. Glatter, O. The interpretation of real-space information from small-angle scattering experiments. *J. Appl. Crystallogr.* **12**, 166 (1979).
68. Debye, P., Anderson Jr, H. R. & Brumberger, H. Scattering by an inhomogeneous solid. II. The correlation function and its application. *J. Appl. Phys.* **28**, 679 (1957).
69. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825 (2011).

Acknowledgements

We thank Y. Shinohara, T. Egami, P. Falus, L. Porcar, and Y. Liu for helpful discussions. This research was performed at the Spallation Neutron Source and the Center for Nanophase Materials Sciences, which are US Department of Energy (DOE) Office of Science User Facilities operated by ORNL. MD simulations used resources of the Oak Ridge Leadership Computing Facility, which is supported by DOE Office of Science under Contract DE-AC05-00OR22725. C.-H.T. thanks the financial support from the Shull Wollan Center during his stay at ORNL. Y.W. is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Early Career Research Program Award KC0402010, under Contract DE-AC05-00OR22725. B.G.S. acknowledges support by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences Data, Artificial Intelligence and Machine Learning at DOE Scientific User Facilities Program under Award Number 34532. M.-C.C. thanks the support provided by the University at Albany—SUNY.

Author contributions

M.-C.C. proposed the regression analysis of Gaussian process. C.-H.T. and S.-Y.C. conducted the trajectory analysis and machine learning processes. J.M.C. performed MD simulations and developed analysis code. Y.W. discussed the results and helped with preparation of manuscript. C.D. discussed methods of data analysis and results. G.-R.H. discussed the machine learning and integral equation approaches. B.G.S. performed data analysis and interpretation, helped with the overall project development, and contributed to manuscript preparation and revision. W.-R.C. developed the mathematical framework of spectral inversion based on machine learning and prepared the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Bobby G. Sumpter or Wei-Ren Chen.

Peer review information *Communications Physics* thanks Lorenzo Rovigatti and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© UT-Battelle, LLC 2022