Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 2075

Machine learning inversion from small-angle scattering for charged polymers

Lijie Ding, [a] Chi-Huan Tung,[a] Jan-Michael Y. Carrillo, [b] Wei-Ren Chen [a] and Changwoo Do *[a]

We develop Monte Carlo simulations for uniformly charged polymers and a machine learning algorithm to interpret the intra-polymer structure factor of the charged polymer system, which can be obtained from small-angle scattering experiments. The polymer is modeled as a chain of fixed-length bonds, where the connected bonds are subject to bending energy, and there is also a screened Coulomb potential for charge interaction between all joints. The bending energy is determined by the intrinsic bending stiffness, and the charge interaction depends on the interaction strength and screening length. All three contribute to the stiffness of the polymer chain and lead to longer and larger polymer conformations. The screening length also introduces a second length scale for the polymer besides the bending persistence length. To obtain the inverse mapping from the structure factor to these polymer conformation and energy-related parameters, we generate a large data set of structure factors by running simulations for a wide range of polymer energy parameters. We use principal component analysis to investigate the intra-polymer structure factors and determine the feasibility of the inversion using the nearest neighbor distance. We employ Gaussian process regression to achieve the inverse mapping and extract the characteristic parameters of polymers from the structure factor with low relative error.

Received 24th January 2025
Accepted 17th June 2025

DOI: 10.1039/d5dd00038f

rsc.li/digitaldiscovery

1 Introduction

Semiflexible charged polymers,[1] also known as polyelectrolytes,[2,3] represent an essential class of materials that are fundamental to both biological processes and technological applications.[4] Their unique behaviors arise from the interplay between molecular flexibility and electrostatic interactions, which are governed by the presence of ionizable groups along their chains. Notable natural examples include DNA,[4,5] RNA,[6] and proteins,[7] all of which play pivotal roles in cellular functions. Synthetic polyelectrolytes, on the other hand, have found extensive use in a variety of fields, including water treatment,[8] energy storage,[9] drug delivery,[10] and responsive materials.[11] The conformational and dynamic properties of charged polymers are shaped by factors such as charge density, ionic strength of the surrounding environment, and the intrinsic bending stiffness of the polymer chain. A thorough understanding of these properties is crucial for tailoring polyelectrolytes to meet the specific demands of diverse applications.

To understand the structure and behavior of the charged polymers, both experimental and theoretical approaches have been employed. Experimental techniques such as small-angle scattering[12] (SAS) including X-ray scattering[13] and neutron scattering[14,15] have proven indispensable for understanding these properties of the charged polymers.[16] Scattering methods provide insights into the nanoscale structure and dynamics of charged polymers, enabling the characterization of key conformational parameters such as radius of gyration, persistence length, and inter- and intra-molecular interactions. Theoretical and computational approaches, including analytical models[17,18] and computer simulations, complement experimental efforts by capturing the fundamental physics of charged polymer systems. Techniques such as molecular dynamics[19,20] (MD) and Monte Carlo[21,22] (MC) simulations have provided significant insights into polymer configurations, bending rigidity, and electrostatic interactions.

Despite the progress made on both the experimental and theoretical fronts, bridging the scattering function measured in SAS experiments with the polymer parameters used for modeling charged polymers in theory and simulations remains a significant challenge. The difficulties lie in extracting physical quantities about polymer conformation by decoding the scattering function. Recent advances in machine learning (ML) have opened new avenues in scattering analysis, enabling parameter extraction without requiring explicit analytical forms of the scattering function.[23] By training ML models on simulation-generated data, it becomes possible to establish an inverse

[a]*Neutron Scattering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. E-mail: doc1@ornl.gov*

[b]*Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*

mapping from the scattering function to the underlying model parameters. This approach has shown promise in a variety of systems, including colloids,[23–25] polymers,[26–29] and lamellar structures.[30,31] These applications demonstrate the potential of ML to bridge the gap between experimental scattering data and theoretical models, providing a robust framework for parameter extraction in complex systems.

In this work, we introduce such an inversion by the ML approach for the charged polymer system, where the data are generated using MC simulations. The polymer configuration is governed by the intrinsic bending stiffness, charge density and salt concentration of the surrounding medium. We first investigate the effects of these key variables on polymer conformation and then calculate the intra-polymer structure factor. To assess the feasibility of inversion, we perform principal component analysis of the scattering data and quantify the feasibility using the nearest neighbor distance of the polymer parameters in the structure factor space. Finally, we employ Gaussian process regression (GPR) to extract both the conformational and energy-related parameters of the polymers from the structure factor, demonstrating the accuracy and robustness of this approach.

2 Method

2.1 Charged polymer in an ionic fluid

We model the polymer as a chain of N connected bonds with fixed length l_b, such that the joint connecting bonds $i-1$ and i is $\mathbf{r}_i$ and the tangent of bond i is $\mathbf{t}_i \equiv (\mathbf{r}_{i+1} - \mathbf{r}_i)/l_b$. The polymer energy is given by:

$$E = \sum_{i=0}^{N-2} \frac{\kappa}{2} \frac{(\mathbf{t}_{i+1} - \mathbf{t}_i)^2}{l_b} - \sum_{i=0}^{N-1} \sum_{j\neq i}^{N-1} \frac{A}{r_{ij}} e^{-r_{ij}/\lambda_D} \quad (1)$$

where κ is the bending modulus, $\frac{A}{r_{ij}} e^{-r_{ij}/\lambda_D}$ is the Yukawa potential, or screened Coulomb potential,[20,32] that models the charge interaction, A is the interaction strength between charged monomers, λ_D is the Debye screening length,[33] and $r_{ij} = |\vec{\mathbf{r}}_i - \vec{\mathbf{r}}_j|$ is the distance between joints i and j. In addition, the self-avoidance of the polymer is enforced by adding hard sphere interaction of diameter l_b between different joints. The interaction strength $A = \frac{(\sigma_e l_b)^2}{4\pi\varepsilon}$ is directly related to the charge density of the polymer σ_e, where ε is the dielectric constant of the medium. The Debye screen length $\lambda_D = \sqrt{\frac{\varepsilon k_B T}{2e^2 I}}$, where k_B is the Boltzmann constant, T is the system temperature, e is elementary charge, and $I = \frac{1}{2}\sqrt{z_i^2 n_i}$ is the ionic strength, in which z_i and n_i are the charge number of the number density of ion species i, respectively.

2.2 Monte Carlo simulation

To calculate the conformational properties of the charged polymer at equilibrium, we sample the configuration space of the charged polymers using the off-lattice Markov Chain Monte Carlo (MCMC) method[34] we previously developed; this off-lattice method provides accurate calculation of the polymer conformation and overcomes the orientational bias rooted in the lattice model.[35] The polymer configuration $\{\vec{\mathbf{r}}_0, \vec{\mathbf{r}}_1 \ldots \vec{\mathbf{r}}_{N-1}\}$ is updated using two MC moves: continuous crankshaft and pivot. Crankshaft picks two random joints on the polymer chain and rotates all the bonds between them for a random angle within the interval $[-\phi_c, \phi_c]$. Pivot randomly selects one joint k on the chain and rotate the preceding sub-chain $(k, \ldots, N)$ within a cone of angle $\phi_p(k)$ centering at the original orientation. To improve the acceptance rate of these updates and thus boost the efficiency of the simulation, the crankshaft rotation angle is adjusted according to the bending modulus such that $\phi_c = \frac{2\pi}{3(1+\kappa)}$, and the pivot rotation angle $\phi_p = \phi_c$. Combining these two moves allows full exploration of the polymer configuration with the contour length fixed and the polymer conformation calculated using this algorithm has been benchmarked against theoretical calculations. More details on the MCMC simulation can be found in our previous paper.[34]

To better characterize and understand the conformation of the charged polymer, we calculate the radius of gyration, bond angle correlation and structure factor of the polymer. The radius of gyration square is $R_g^2 = \frac{1}{2}\langle r_{ij}^2 \rangle_{ij}$, where the $\langle \ldots \rangle_{ij}$ denotes the average of all pair of joints. The bond–bond correlation is $\langle \cos(\theta(s)) \rangle = \langle \hat{\mathbf{t}}_i \cdot \hat{\mathbf{t}}_{i+s} \rangle_i$ where $\langle \ldots \rangle_i$ denotes the average over all bonds and s represents the contour distance between two bonds along the polymer chain. Finally, the isotropic intra-polymer structure factor[12,14] is given by:

$$S(q) = 1 + \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j\neq i}^{N-1} \frac{\sin(q r_{ij})}{q r_{ij}} \quad (2)$$

where q is the magnitude of the scattering vector. When running the MCMC simulation, we first randomize the system by running 2000 MC sweeps at inverse temperature $\beta = 1/k_B T = 0$, then tempering the system for another 2000 MC sweeps while gradually decreasing the temperature to $\beta = 1$. We sample the polymer configuration and calculate the average of the conformation parameters for while running for another 4000 MC sweeps, each MC sweep consists of N crankshafts and N pivot updates. We use a natural unit in our simulation where energy is in unit of $k_B T = 1$ and length is in unit of $l_b = 1$ such that the polymer contour length $L = N l_b = N$. We use degree of discretization $L = 500$ for all of our simulations.

2.3 Principal component analysis

To study the relationship between structure factor $S(q)$ and the polymer parameters including radius of gyration R_g^2, end-to-end distance R^2, bending stiffness κ and interaction strength A for various screening distance λ_D, we generate a data set consisting of 4000 combinations of (κ, A, λ_D) and corresponding $\log S(q)$ and carry out principal component analysis for the data sets. The $S(q)$ is calculated for 100 $q \in [10^{-1}, 1]$, uniformly placed in the log scale, and $\kappa \sim U(5, 50)$, $A \sim U(0, 10)$ and $\lambda_D \sim U_d(1, 10)$, where $U(a, b)$ is the uniform distribution in the

interval $[a, b]$ and $U_d(a, b)$ is the discrete uniform distribution. Similar to previous work,[23] we use singular value decomposition (SVD) to find the three most important bases of the 4000 × 100 matrix $\mathbf{F} = \{\log S(q)\}$, such that $\mathbf{F} = \mathbf{U}\Sigma\mathbf{V}^T$. The diagonal entries of Σ^2 are proportional to the weight of the variance of the projection of $\mathbf{F}$ onto each principal vector of $\mathbf{V}$. Projecting $\mathbf{F}$ onto the first few bases provides a way to analyze $\mathbf{F}$ in a dimensionally reduced space. A useful tool to study the distribution of the polymer parameters $\mathbf{Y} = \{(\kappa, A, R_g^2/L^2, R^2/L^2)\}$ is to calculate the nearest-neighbor distance of $\zeta \in \{\kappa, A, R_g^2/L^2, R^2/L^2\}$ on the $\mathbf{F}$ manifold. For n-number of vectors, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, the first nearest neighbor is defined as $\mathrm{NN}_1(\mathbf{x}_i) = \mathrm{argmin}_{\mathbf{x}_j \neq \mathbf{x}_i} |\mathbf{x}_j - \mathbf{x}_i|$; similarly, the second nearest neighbor is $\mathrm{NN}_1(\mathbf{x}_i) = \mathrm{argmin}_{\mathbf{x}_j \neq \mathbf{x}_i, \mathrm{NN1}(\mathbf{x}i)} |\mathbf{x}_j - \mathbf{x}_i|$, and we define the normalized nearest neighbor distance D_{NN} for the $\zeta(\mathbf{x})$ as:

$$D_{\mathrm{NN}}(\zeta) = \frac{\langle 2\zeta(\mathbf{x}) - \zeta(\mathrm{NN}_1(\mathbf{x})) - \zeta(\mathrm{NN}_2(\mathbf{x}))\rangle_{\mathbf{x}}}{(\max_{\mathbf{x}}(\zeta) - \min_{\mathbf{x}}(\zeta))/2} \tag{3}$$

where $\langle \ldots \rangle_{\mathbf{x}}$ is the average overall $\mathbf{x}$. The nearest-neighbor distance helps quantify the feasibility of the parameter inversion from scattering, serving as a local sensitivity metric on the scattering $\mathbf{F}$ manifold. By measuring how a given parameter ζ changes when moving from one scattering signature to its two closest neighbors, $D_{\mathrm{NN}}(\zeta)$ tells us how well small differences in $\log S(q)$ can be traced back to unique changes in ζ. Concretely, large $D_{\mathrm{NN}}(\zeta)$ indicates that minor variances in $\log S(q)$ can map to large jumps in ζ, signaling regions where inversion is unstable or degenerate. Whereas small $D_{\mathrm{NN}}(\zeta)$ means that even significant noise in $\log S(q)$ produces only modest shifts in ζ, thus the inverse mapping remains well-conditioned and robust.

2.4 Gaussian process regression

To perform inverse mapping from the scattering function, $\mathbf{x} = \log S(q)$, to the system parameters, or inversion targets $\mathbf{y} = (\kappa, A, R_g/L^2, R^2/L^2)$, we employ a Gaussian Process Regression (GPR) model trained on data generated through Monte Carlo (MC) simulations. Under the framework of GPR,[36,37] the goal is to obtain the posterior distribution $p(\mathbf{Y}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{Y})$ for the function output $\mathbf{y}$. In this setup, the training and test sets are defined as $\mathbf{X} = \{\log S(q)\}_{\mathrm{train}}$ and $\mathbf{X}_* = \log S(q)_{\mathrm{test}}$, respectively, while $\mathbf{Y}$ and $\mathbf{Y}_*$ correspond to the inversion targets $(\kappa, A, R_g/L^2, R^2/L^2)$. GPR assumes a Gaussian process prior over the regression function, $g(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$, where $m(\mathbf{x})$ is the prior mean function and $k(\mathbf{x}, \mathbf{x}')$ is the covariance kernel. The joint distribution for the Gaussian process is expressed as follows:

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{Y}_* \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{X}) \\ m(\mathbf{X}_*) \end{bmatrix}, \begin{bmatrix} k(\mathbf{X}, \mathbf{X}) & k(\mathbf{X}, \mathbf{X}_*) \\ k(\mathbf{X}_*, \mathbf{X}) & k(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix} \right) \tag{4}$$

Here, we use a constant prior mean function $m(\mathbf{x})$, while the kernel function is modeled as a combination of a Radial Basis Function (RBF) and a white noise term:

$$k\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left(-\frac{\left|\mathbf{x} - \mathbf{x}'\right|^2}{2l} \right) + \sigma\delta\left(\mathbf{x}, \mathbf{x}'\right), \tag{5}$$

where l represents the correlation length, σ is the variance of the observational noise, and δ is the Kronecker delta function. These hyperparameters are optimized during training using the simulation data. In practice, we utilize the scikit-learn[38,39] Gaussian Process library due to its convenience and efficiency.

3 Results

We first study the effect of each polymer parameter on the conformation of the polymer, then investigate the scattering function of the charged polymer, where we also show the principal component analysis of our data set $\mathbf{F} = \{\log S(q)\}$. We then discuss the feasibility of inversion based on the SVD of $\mathbf{F}$. With the feasibility established, we finally test our trained GPR for the inversion.

3.1 Variation of polymer conformation

Both the local bond-to-bond bending and long-range charge interaction contribute to the stiffness of the entire polymer. Such stiffness will affect the overall size of the charged polymer, which can be captured by the radius of gyration R_g^2 and end-to-end distance R^2. Fig. 1(a) and (c) shows both R_g^2 and R^2 increase with screening length λ_D and bending stiffness κ, and intuitively, the effects of κ on both R_g^2 and R^2 are more significant when λ_D is small, as the R_g^2 and R^2 *versus* λ_D curves for different κ start to converge as the λ_D increases. In contrast, while R_g^2 and R^2 also increase with larger charge interaction strength A, these curves diverge as λ_D increases, which happens because the increasing screening length λ_D amplifies the effect of charge interaction.

When the polymer is only subjected to bending κ, or in the case of $A = 0$, the polymer is a classic semiflexible polymer

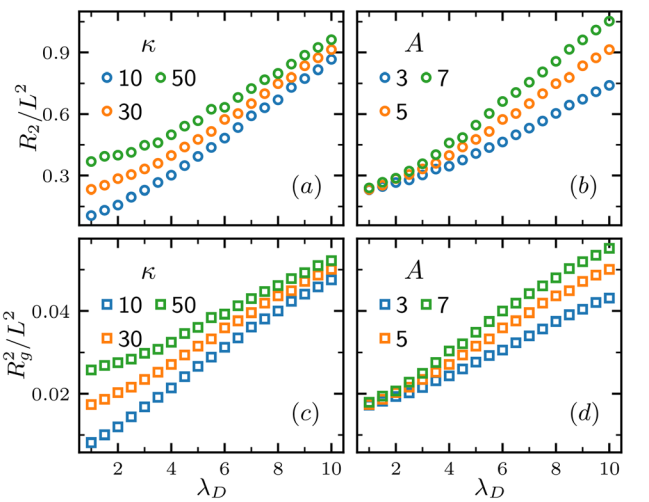

Fig. 1 Radius of gyration R_g^2 and end-to-end distance R^2 of the charged polymer *versus* various bending stiffness κ, charge interaction strength A and screen length λ_D. (a) Normalized end-to-end distance R^2/L^2 *versus* screen length λ_D for various bending stiffness κ. (b) R^2/L^2 *versus* screen length λ_D for various charge interaction strength A. (c) and (d), similar to (a) and (b), respectively, but for normalized radius of gyration R_g^2/L^2.

whose bond angle correlation can be described by a single exponential decay:

$$\langle \cos\theta(s)\rangle = e^{-s/\lambda_0} \tag{6}$$

where λ_0 is the persistent length. s is the bond–bond distance along the polymer contour. However, as pointed out in a previous study,[20] the charge interaction introduces new length scales, and as a result, the bond angle correlation can be described by:

$$\langle \cos\theta(s)\rangle = (1-\alpha)e^{-s/\lambda_1} + \alpha e^{-s/\lambda_2} \tag{7}$$

λ_1 and λ_2 correspond to two different length scales, and it is also notable that the effective bending rigidity can be calculated by $\lambda_e = \lambda_2/\alpha$.[20]

Fig. 2(a) shows the bond angle correlation function $\langle \cos\theta(s)\rangle$ for various screening length λ_D, and the fitted lines are calculated using the single scale model as in eqn (6). As λ_D increases, the single scale model fitting starts to diverge from the data point, indicating the necessity of switching to the double length scale model (eqn (7)); Fig. 2(b) shows such fitting results, and the two length scale model can still describe the decay of $\langle \cos\theta(s)\rangle$ at large λ_D.

Fig. 2(c) show all three length scales λ_0, λ_1 and λ_e *versus* screening length λ_D for various bending stiffness κ. At low λ_D the one length scale still fits the bond angle correlation data, and increases with increasing λ_D. When switching to the two length scale model, the long length scale λ_1 increases with increasing λ_D, while the short length scale λ_e decreases and deviates from λ_1 and then plateaus. The plateau value increases with bending stiffness κ. The switch from the one-length scale λ_0 to two-length scale (λ_1, λ_e) in the plot is determined by monitoring the divergence between the λ_0 and λ_1 when fitting the correlation function at low screening length λ_D. Fig. 2(d) shows a similar result but for various charge interaction strength A. Similar to its effect on the end-to-end distance and radius of gyration, A amplifies the effect of increasing λ_D, while the short length scale λ_e plateaus at a similar value for various A, confirming it corresponds to the bending stiffness κ.

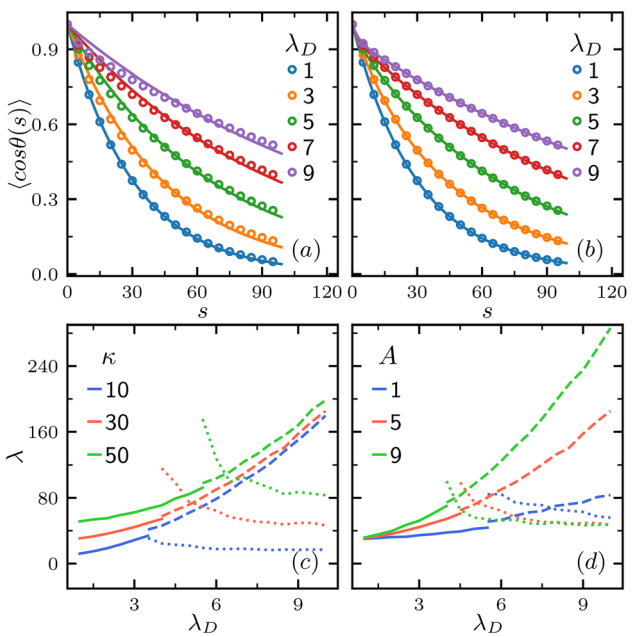

Fig. 2 Different length scales of the charged polymer, fitted using both single length scale and double length scale models. (a) Bond angle correlation $\langle \cos\theta(s)\rangle$ for various screening length λ_D with $\kappa = 30$, $A = 5$, solid lines are fitted using a single length scale (eqn (6)). (b) Similarly, but fitted using double length scale (eqn (7)). (c) Three persistent lengths, λ_0 for the solid line, λ_1 for the dashed line and λ_e for the dotted line, *versus* screening length λ_D for various κ with $A = 5$. (d) Similar to (c), but for various A with $\kappa = 30$.

3.2 Scattering factor of the polymers

We then turn to the inter-polymer structure factor. For comparison, we also calculate the structure factor of a solid rod, whose polymer configuration is $\vec{\mathbf{r}}_i = i\hat{\mathbf{x}}$, with all bonds pointing to the same direction. Fig. 3(a) shows the variation of structure factor $S(q)$ for various bending stiffness κ. Compared to the solid rod, the polymer structure factor shows a bump at a structure vector q range comparable to its radius of gyration. Fig. 3(b) shows the structure factor of the polymer divided by the rod $S(q)/S_{\mathrm{rod}}(q)$, where the bump is better shown. As the bending stiffness κ increases, the peak in $S(q)/S_{\mathrm{rod}}(q)$ lowers and the corresponding q value also decreases, indicating an increase of the characteristic length. Fig. 3(c) and (d) shows the $S(q)/S_{\mathrm{rod}}(q)$ for various charge interaction strength A and screening length λ_D, and both show similar effects on the structure factor of the polymer as they make the polymer more extended and stiff.

To better analyze the structure factor of the charged polymer, we carry out principal component analysis described in Sec. 2.3. By decomposing the $\mathbf{F} = \{\log S(q)\}$ into $\mathbf{F} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}$, we find that the singular value Σ decays rapidly *versus* its rank, as shown in

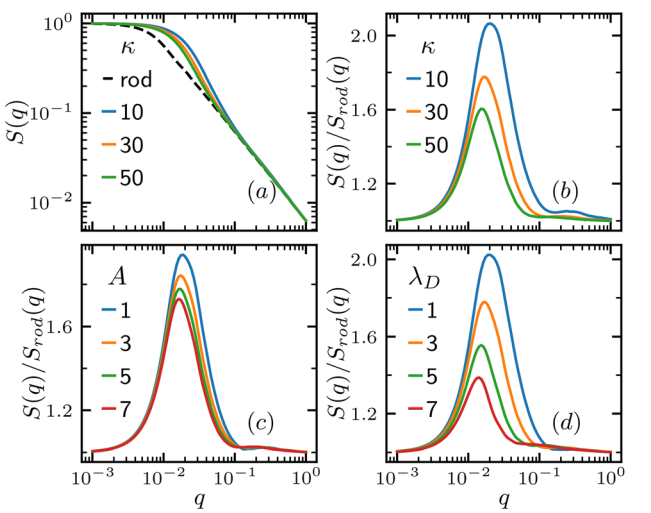

Fig. 3 Variation of the structure factor of the charged polymer. (a) Structure factor $S(q)$ for various bending stiffness κ with $\lambda_D = 3$, $A = 5$ and rod effectively representing the $\kappa = \infty$ case. (b) Structure factor $S(q)$ normalized by the rod's structure factor $S_{\mathrm{rod}}(q)$ for various κ. (c) $S(q)/S_{\mathrm{rod}}(q)$ for various charge interaction strength A with $\kappa = 30$, $\lambda_D = 3$. (d) $S(q)/S_{\mathrm{rod}}(q)$ for various screening length λ_D with $\kappa = 30$, $A = 5$.

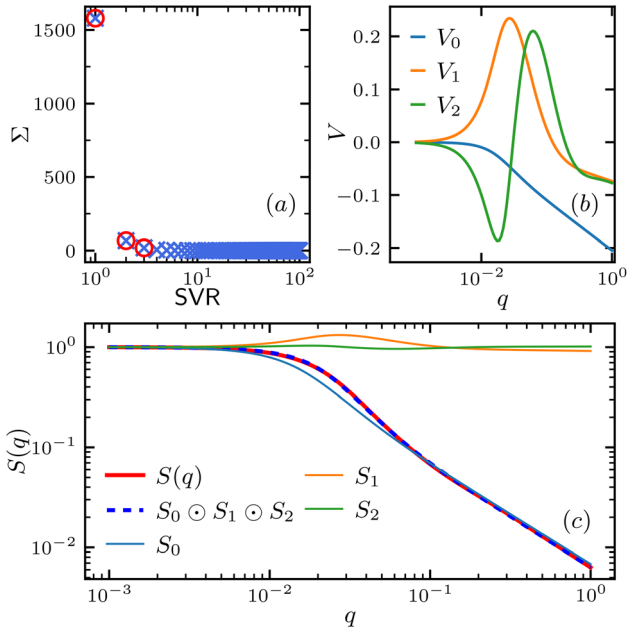

Fig. 4 Singular value decomposition of the structure factor data set $\mathbf{F} = \{\log S(q)\}$. (a) Singular value Σ *versus* Singular Value Rank (SVR), with the top 3 ranks highlighted in a red circle. (b) First 3 singular vectors V_0, V_1 and V_2. (c) Decomposition of the log $S(q)$ with $\kappa = 10$, $A = 5$, and $\lambda_D = 3$; log(S_0), log(S_1) and log(S_2) are projections of log $S(q)$ onto V_0, V_1 and V_2, respectively.

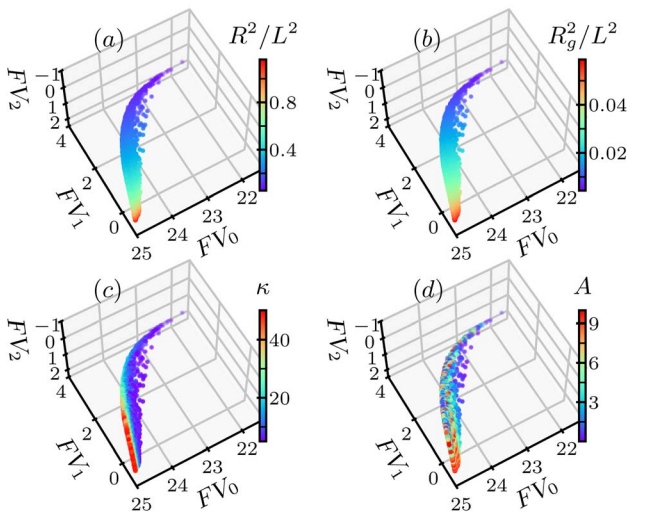

Fig. 5 Distribution of the polymer parameters (R^2/L^2,$R_g{}^2/L^2$,κ,A) in the SVD space spanned by (V_0, V_1, V_2). (a) End-to-end distance divided by length square R^2/L^2, (b) Radius of gyration square divided by length square $R_g{}^2/L^2$. (c) Bending stiffness κ. (d) Charge interaction strength A.

Fig. 4(a), indicating we can represent the log $S(q) \in \mathbf{F}$ using few bases. Fig. 4(b) shows the first 3 singular vectors, and Fig. 4(c) shows the projection of a structure factor $S(q)$ onto each basis, and the reconstruction from only the 3 bases closely matches the original $S(q)$. This decomposition will allow us to further determine the feasibility of extracting these polymer parameters from the structure factor.

3.3 Feasibility for machine learning inversion

While it is straightforward to calculate the structure factor $S(q)$ from the polymer parameters, including length L, bending stiffness κ, charge interaction strength A and screening length λ_D, and calculate the end-to-end distance R^2 and radius of gyration $R_g{}^2$ using MC simulation, the feasibility of doing the inversion is to be further assessed. Fig. 5 shows the distribution of (R^2/L^2,$R_g{}^2/L^2$,κ,A) in the structure factor space. This mapping is achieved by projecting all of the structure factor log $S(q) \in \mathbf{F}$ into the space spanned by the first 3 singular vectors (V_0, V_1, V_2), and the corresponding 3 coefficients of each log $S(q)$ correspond to a single point in the $\mathcal{R}^3$ space. As shown in Fig. 5(a–c), the end-to-end distance R^2/L^2, radius of gyration $R_g{}^2/L^2$ and bending stiffness κ are all well spread out on in the FV manifold, indicating they are eligible to be extracted from the structure factor. Fig. 5(d) shows the distribution of charge interaction strength A and it is unclear if it can be extracted due to some randomness in the distribution.

Intuitively, when then screening length λ_D is very small, the effect of the charge interaction becomes negligible, preventing it from having a meaningful impact on the structure factor $S(q)$, thus it is not expected to have A feasible for extraction from the $S(q)$ at low λ_D. To quantify this feasibility, we slice the structure factor data set $\mathbf{F} = \{\log S(q)\}$ into different slices for different screening lengths λ_D, and calculate the nearest neighbor distance for each slice. As shown in Fig. 6(a), we plot 3 slices of the charge interaction strength A distribution, and the randomness reduces as the screening length λ_D increases. Quantitatively, Fig. 6(b) shows the nearest neighbor distance D_{NN} for each polymer parameter and $D_{NN}(A)$ is much larger than that of the others when the screening length λ_D is small, and then it decays to lower value as the λ_D increases, leading to a more significant impact of the charge interaction strength A

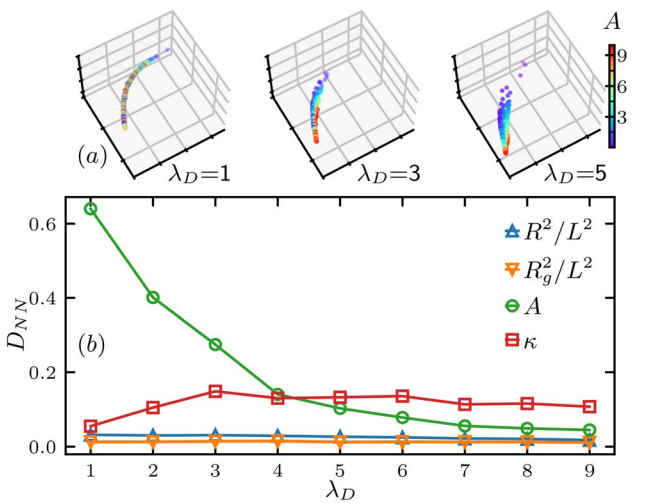

Fig. 6 Nearest neighbor distance analysis of the charge interaction strength A. (a) Value distribution of A in the SVD space for various slices of screening length λ_D, the axes are the same as in Fig. 5. (b) Nearest neighbor distance D_{NN} for various polymer parameters *versus* different slices of the data $\mathbf{F}$ separated by the λ_D value.

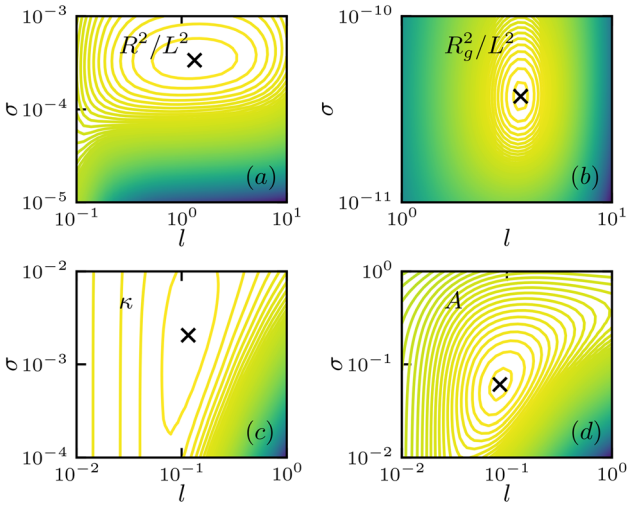

Fig. 7 Log marginal likelihood contour of hyperparameters correlation length l and noise level σ for various polymer parameters, with the optimized value marked with a black cross. (a) End-to-end distance R^2/L^2. (b) Radius of gyration R^2/L^2. (c) Bending stiffness κ. (d) Charge interaction strength A.

on the polymer conformation. This indicates the charge interaction strength A, which is directly related to the charge density of the polymer, is still extractable if the screening length is large enough.

3.4 Extraction of the polymer parameters

With the feasibility for inversion and corresponding conditions established for the polymer parameter $(R^2/L^2,R_g{}^2/L^2,\kappa,A)$, we train the GPR using 70% of the entire data set $\mathbf{F} = \{\log S(q)\}$as the training set $\{\log S(q)\}_{\text{train}}$, and then test the trained GPR using the remaining 30% data $\{\log S(q)\}_{\text{test}}$ by comparing the actual polymer parameters with the ones extracted from the structure factor $S(q)$. The split between the training and testing data is random. To obtain the trained regressor, we need to find the optimized hyperparameters (l, σ) for each inversion target, or polymer parameters. We search for the (l, σ) that maximize the log marginal likelihood,[36] which are shown in Fig. 7.

Fig. 8 shows a comparison between polymer parameters $((R^2/L^2,R_g{}^2/L^2,\kappa,A))$ obtained from ML inversion and the corresponding reference used in or calculated through MC simulation. We note that due to the high nearest neighbor distance $D_{\text{NN}}(A)$ of charge interaction strength at low screening length λ_{D}, we only used data with $\lambda_{\text{D}} \geq 4$ for the inversion of A. Nevertheless, the data agree well, and lie closely along the diagonal line, with relatively low error, and for polymer parameter ζ, the relative error between MC reference ζ_{MC} and ML inversion ζ_{ML} is estimated by $\text{Err} = \langle|\zeta_{\text{MC}} - \zeta_{\text{ML}}|/\max(\zeta_{\text{MC}},\zeta_{\text{ML}})\rangle$, where $\langle\ldots\rangle$ is the average over all data points. The relative error is annotated on each panel of Fig. 8 and shows very high precision for $((R^2/L^2,R_g{}^2/L^2,\kappa)$ and good precision for A. While the errors for end-to-end distance R^2, radius of gyration $R_g{}^2$ and bending modulus κ are very small, the error for charge density is relatively large as we are including data with all screening length $\lambda \geq 4$. In practice, the screening length can be estimated based on the solvent conditions; a reduced range of λ_{D} will lead to better accuracy in the extraction of charge density A.

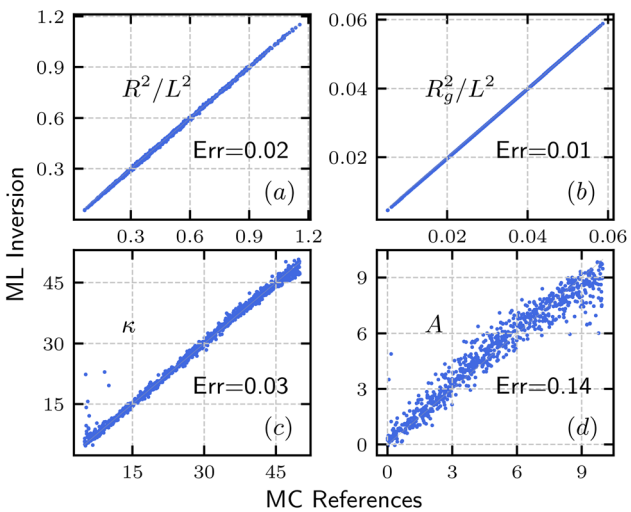

Fig. 8 Comparison between the polymer parameter extracted from structure factor and input or direct calculation from MC simulation. (a) End-to-end distance R^2/L^2. (b) Radius of gyration R^2/L^2. (c) Bending stiffness κ. (d) Charge interaction strength A. (a–c) Utilized all range of $\mathbf{F}$ and (d) only used data with $\lambda_{\text{D}} \geq 4$.

4 Conclusions

In this work, we apply the off-lattice MC simulation for a semi-flexible polymer to study the charged polymers, and investigate the ML inversion from scattering for such a polymer. We model the polymer using a chain of connected bonds, and the polymer energy consists of both bending energy and screened Coulomb interaction, which are proportional to the bending stiffness κ and charge interaction strength A, respectively. The charge interaction range is determined by the screen length λ_{D}. We first study the polymer conformation, where the polymer size, quantified by the end-to-end distance R^2 and radius of gyration $R_g{}^2$, increases with κ, A and λ_{D}. The bond angle correlation function transits from the single length scale to double length scale as the screening length λ_{D} increases. We calculate the intra-polymer structure factor $S(q)$ of the charged polymer, compare it to that of the solid rod, and show the $S(q)$ is sensitive to all three polymer parameters κ, A and λ_{D}. We calculate the $S(q)$ for a wide range of κ, A and λ_{D}, then carry out principal component analysis using singular value decomposition to find the singular vectors, which allows us to do dimension reduction of the structure factor. In addition, we investigate the feasibility for inversion from scattering for both the conformation parameters: end-to-end distance R^2 and radius of gyration $R_g{}^2$, and the energy parameters: bending stiffness κ and charge interaction strength A. We quantify the feasibility using nearest neighbor distance D_{NN}, and find that R^2, $R_g{}^2$ and κ are eligible for a wide range of screening lengths λ_{D} and the charge interaction strength A is eligible for inversion from structure factor when the λ_{D} is large enough. Finally, we use GPR to obtain the inverse mapping from structure factor $S(q)$ to polymer

parameters (R^2,$R_g{}^2$,κ,A) by optimizing the hyperparameters using a training data set, apply the inversion GPR to extract polymer parameters from structure factor for a test data set, and compare the ML extracted value to the MC reference; they agree well, and low relative errors are achieved.

Our approach provides a unique method to obtain the bending stiffness and the charge density σ_e, which is directly related to the charge interaction strength $A = \frac{(\sigma_e l_b)^2}{4\pi\varepsilon}$ using the scattering data. A natural next step would be to carry out a SANS experiment for some charged polymer sample, and apply our approach on the experimentally measured SANS data. In practice, this approach assumes the experimental data falls within the range of training data, and a procedure of trial and error maybe required based on the fitting results, in which the training set needs to be expanded as needed. In addition, experimental data naturally come with noise, for which a denoising procedure[40] can be helpful, and the analysis of noisy data will naturally provide uncertainties by the GPR.[28] Moreover, the effect of noise for the GPR prediction can be systematically studied by measuring the accuracy of the inversion when different levels of noise are added to the testing data. Finally, this framework can be expanded to the study of more complicated charged polymer systems including charge-patterned polypeptides,[41] alternating copolymers[42] and zwitterionic patterned polymers.[43] To study these systems, it is required to model the polymer energy accordingly. It is natural to introduce variable charge interaction strength A for different monomer segments to model the charge pattern and polarity, and a screened dipole–dipole interaction can be used for modeling the zwitterionic polymer.

Data availability

The code and data for this work are available at the GitHub repository: **https://github.com/ljding94/Charged_Polymer** with DOI: **https://doi.org/10.5281/zenodo.15624816**.

Author contributions

L. D. conceived this work, carried out MC simulation and ML analysis, and drafted the manuscript. C. H. T. discussed the results and reviewed the manuscript. J. M. Y. C. conceived this work, discussed the results, and reviewed the manuscript. W. R. C. discussed the results and reviewed the manuscript. C. D. conceived this work, discussed the results and revised the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This research was performed at the Spallation Neutron Source and the Center for Nanophase Materials Sciences, which are DOE Office of Science User Facilities operated by Oak Ridge National Laboratory. This research was sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U. S. Department of Energy. The ML aspects were supported by the U.S. Department of Energy Office of Science, Office of Basic Energy Sciences Data, and Artificial Intelligence and Machine Learning at DOE Scientific User Facilities Program under Award Number 34532. Monte Carlo simulations and computations used resources of the Oak Ridge Leadership Computing Facility, which is supported by the DOE Office of Science under Contract DE-AC05-00OR22725.

Notes and references

1 R. R. Netz and D. Andelman, *Phys. Rep.*, 2003, **380**, 1–95.
2 A. V. Dobrynin and M. Rubinstein, *Prog. Polym. Sci.*, 2005, **30**, 1049–1118.
3 S. Förster and M. Schmidt, *Phys. Prop. Polym.*, 2005, 51–133.
4 G. S. Manning, *Q. Rev. Biophys.*, 1978, **11**, 179–246.
5 S. Lameh, L. Ding and D. Stein, *Phys. Rev. Appl.*, 2020, **14**, 054042.
6 V. A. Bloomfield, D. M. Crothers and I. Tinoco, *Nucleic acids: Structures, properties, and functions*, 2000.
7 C. Tanford and M. L. Huggins, *J. Electrochem. Soc.*, 1962, **109**, 98C.
8 B. Bolto and J. Gregory, *Water Res.*, 2007, **41**, 2301–2324.
9 M. Winter and R. J. Brodd, *Chem. Rev.*, 2004, **104**, 4245–4270.
10 W. B. Liechty, D. R. Kryscio, B. V. Slaughter and N. A. Peppas, *Annu. Rev. Chem. Biomol. Eng.*, 2010, **1**, 149–173.
11 M. A. C. Stuart, W. T. Huck, J. Genzer, M. Müller, C. Ober, M. Stamm, G. B. Sukhorukov, I. Szleifer, V. V. Tsukruk, M. Urban, *et al.*, *Nat. Mater.*, 2010, **9**, 101–113.
12 P. Lindner and T. Zemb, *Neutrons, x-rays and light: scattering methods applied to soft condensed matter*, 2002.
13 B. Chu and B. S. Hsiao, *Chem. Rev.*, 2001, **101**, 1727–1762.
14 S.-H. Chen, *Annu. Rev. Phys. Chem.*, 1986, **37**, 351–399.
15 M. Shibayama, *Polym. J.*, 2011, **43**, 18–34.
16 M. Nierlich, C. Williams, F. Boué, J. Cotton, M. Daoud, B. Famoux, G. Jannink, C. Picot, M. Moan, C. Wolff, *et al.*, *J. Phys.*, 1979, **40**, 701–704.
17 R. R. Netz and H. Orland, *Eur. Phys. J. E:Soft Matter Biol. Phys.*, 2003, **11**, 301–311.
18 A. V. Dobrynin, R. H. Colby and M. Rubinstein, *Macromolecules*, 1995, **28**, 1859–1871.
19 M. J. Stevens and K. Kremer, *J. Chem. Phys.*, 1995, **103**, 1669–1690.
20 A. Gubarev, J.-M. Y. Carrillo and A. V. Dobrynin, *Macromolecules*, 2009, **42**, 5851–5860.
21 F. Carlsson, P. Linse and M. Malmsten, *J. Phys. Chem. B*, 2001, **105**, 9040–9049.
22 P. Chodanowski and S. Stoll, *J. Chem. Phys.*, 1999, **111**, 6069–6081.
23 M.-C. Chang, C.-H. Tung, S.-Y. Chang, J. M. Carrillo, Y. Wang, B. G. Sumpter, G.-R. Huang, C. Do and W.-R. Chen, *Commun. Phys.*, 2022, **5**, 46.

24 C.-H. Tung, S.-Y. Chang, M.-C. Chang, J.-M. Carrillo, B. G. Sumpter, C. Do and W.-R. Chen, *Carbon Trends*, 2023, **10**, 100252.
25 L. Ding, Y. Chen and C. Do, *Appl. Crystallogr.*, 2025, **58**(3), 992–999.
26 C.-H. Tung, S.-Y. Chang, H.-L. Chen, Y. Wang, K. Hong, J. M. Carrillo, B. G. Sumpter, Y. Shinohara, C. Do and W.-R. Chen, *J. Chem. Phys.*, 2022, **156**, 131101.
27 L. Ding, C.-H. Tung, B. G. Sumpter, W.-R. Chen and C. Do, *arXiv*, 2024, preprint, arXiv:2410.05574, DOI: **10.48550/arXiv.2410.05574**.
28 L. Ding, C.-H. Tung, Z. Cao, Z. Ye, X. Gu, Y. Xia, W.-R. Chen and C. Do, *Digital Discovery*, 2025, **4**, 1570–1577.
29 L. Ding, C.-H. Tung, B. G. Sumpter, W.-R. Chen and C. Do, *J. Chem. Theory Comput.*, 2025, **21**, 4176–4182.
30 C.-H. Tung, Y.-J. Hsiao, H.-L. Chen, G.-R. Huang, L. Porcar, M.-C. Chang, J.-M. Carrillo, Y. Wang, B. G. Sumpter, Y. Shinohara, *et al.*, *J. Colloid Interface Sci.*, 2024, **659**, 739–750.
31 C.-H. Tung, L. Ding, M.-C. Chang, G.-R. Huang, L. Porcar, Y. Wang, J.-M. Y. Carrillo, B. G. Sumpter, Y. Shinohara and C. Do, *J. Chem. Phys.*, 2025, **162**, 074106.
32 J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids: with Applications to Soft Matter*, Academic press, 2013.
33 J. N. Israelachvili, *Intermolecular and Surface Forces*, Academic press, 2011.
34 L. Ding, C.-H. Tung, B. G. Sumpter, W.-R. Chen and C. Do, *J. Chem. Theory Comput.*, 2024, **20**, 10697–10702.
35 C.-H. Tung, L. Ding, G.-R. Huang, Y. Wang, J.-M. Y. Carrillo, B. G. Sumpter, Y. Shinohara, C. Do and W.-R. Chen, *J. Chem. Phys.*, 2024, **161**, 224107.
36 C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*, MIT press Cambridge, MA, 2006, vol. 2.
37 J. Wang, *Comput. Sci. Eng.*, 2023, 4–11.
38 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
39 L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt and G. Varoquaux, *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.
40 C.-H. Tung, S. Yip, G.-R. Huang, L. Porcar, Y. Shinohara, B. G. Sumpter, L. Ding, C. Do and W.-R. Chen, *J. Colloid Interface Sci.*, 2025, 137554.
41 J. Dinic and M. V. Tirrell, *Biomacromolecules*, 2024, **25**, 2838–2851.
42 C. Yi, Y. Yang and Z. Nie, *J. Am. Chem. Soc.*, 2019, **141**, 7917–7925.
43 L. Zheng, H. S. Sundaram, Z. Wei, C. Li and Z. Yuan, *React. Funct. Polym.*, 2017, **118**, 51–61.