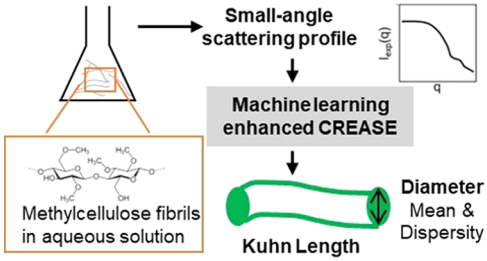
Accepted Manuscript
Version of record at: https://doi.org/10.1021/acs.macromol.2c02165

Machine Learning-Enhanced Computational Reverse-Engineering Analysis for Scattering Experiments (CREASE) for Analyzing Fibrillar Structures in Polymer Solutions

Zijie Wu and Arthi Jayaraman*

ABSTRACT: In this work, we present a machine learning (ML)-enhanced computational reverse-engineering analysis of scattering experiments (CREASE) approach to analyze the small-angle scattering profiles from polymer solutions with assembled semiflexible fibrils with dispersity in fibril diameters (e.g., aqueous solutions of methylcellulose fibrils). This work is an improvement over the original CREASE method [Beltran-Villegas, D. J.; et al. *J. Am. Chem. Soc.*, **2019**, *141*, 14916−14930], which identifies relevant dimensions of assembled structures in polymer solutions from their small-angle scattering profiles without relying on traditional analytical models. Here, we improve the original CREASE approach by incorporating ML for analyzing assembled semiflexible fibrillar structures with dispersity in fibril diameters. We first validate our CREASE approach without ML by taking as input the scattering profiles of in silico structures with known dimensions (diameter, Kuhn length) and reproducing as output those known dimensions within error. We then show how the incorporation of ML (specifically an artificial neural network, denoted as NN) within the CREASE approach improves the speed of workflow without sacrificing the accuracy of the determined fibrillar dimensions. Finally, we apply NN-enhanced CREASE to experimental small-angle X-ray scattering profiles from methylcellulose fibrils obtained by Lodge, Bates, and co-workers [Schmidt, P. W.; et al. *Macromolecules*, **2018**, *51*, 7767−7775] to determine fibril diameter distribution and compare NN-enhanced CREASE's output with their fibril diameter distribution fitted using analytical models. The diameter distributions of methylcellulose fibrils from NN-enhanced CREASE are similar to those obtained from analytical model fits, confirming the results by Lodge, Bates, and co-workers that methylcellulose form fibrils with consistent average diameters of ∼15−20 nm regardless of the molecular weight of methylcellulose chains. The successful implementation of NN-enhanced CREASE in handling experimental scattering profiles of complex macromolecular assembled structures with dispersity in dimensions demonstrates its potential for application toward other unconventional fibrillar systems that may not have appropriate analytical models.



I. INTRODUCTION

Small-angle scattering (SAS) is a widely used technique to probe structures in polymers and soft materials at multiple length scales without the need for excessive sample pretreatment such as crystallization or potentially, sample-altering preparation needed for other characterization techniques.[1−11] These other characterization methods, such as transmission electron microscopy (TEM),[12,13] cryogenic TEM,[13] or atomic force microscopy (AFM),[14] provide direct real-space imaging of the structure and commonly output the two-dimensional (2D) surface image or 2D projection of the inner three-dimensional (3D) structure representing a subset of the sample.[4,15] Some advanced techniques can also provide 3D imaging.[16,17] In contrast, SAS reveals spatial information about the overall 3D structure of the entire sample in reciprocal space. The sample preparation procedure for SAS is, in most cases, less involved as compared to that needed for electron microscopy methods,[18] and SAS provides structural information spanning a range of length scales from 1 to 100 nm,[3,19] or even microns if ultra-small-angle scattering (USAS) is used.[20]

In SAS experiments, the sample is subjected to a beam of X-ray in the case of small-angle X-ray scattering, SAXS, or neutrons in the case of small-angle neutron scattering, SANS, and the resulting elastic scattering results in a 2D scattering pattern. The 2D scattering pattern can be azimuthally averaged to obtain a one-dimensional (1D) scattering profile—a curve of scattering intensity ($I(q)$) as a function of the wavevector (q).[1,3,21] Analyzing the 2D or 1D SAS profile to interpret the 3D structure is not a trivial task. Direct interpretation based on shapes/slopes of the profile can provide several generic descriptors of the assembled structure, including the radius of gyration, molecular weight, and surface-to-volume ratio.[22,23] Without prior knowledge about the structure of interest, it is possible, albeit requiring

Accepted Manuscript
Version of record at: https://doi.org/10.1021/acs.macromol.2c02165

complicated procedures, to reconstruct an approximate 3D representation of the structure[24−29] from the scattering profile. However, the uniqueness of the reconstructed 3D structure cannot always be guaranteed, as multiple different structures can theoretically lead to similar SAS profiles. Prior knowledge about the formed structure is the only way to help eliminate degenerate solutions.[19,30] In a majority of applications, one has some knowledge about the structure of interest through microscopy or other imaging techniques and can use SAS as a follow-up to obtain more detailed 3D information about such structures using appropriate analytical models for those structures. For example, one can use TEM or SEM to identify the general shape of the assembled structure (e.g., spherical micelles, anisotropic shapes) and use SAS to determine the detailed dimensions of that 3D assembled shape using corresponding analytical models.[31] Analytical models are available for a range of conventional shapes, e.g., spheres,[1] micelles,[32] cylinders,[33] *etc.*, and are integrated into popular software packages such as SASView,[34] designed for users fitting SAS scattering profiles. However, analytical models may be too approximate or not exist for any arbitrary shape of interest or unconventional assembled structures obtained through novel polymer chemistries/processing techniques. In such cases, there is a need for other independent computational method(s) that does (do) not rely on analytical models to have confidence in the interpretation of the SAS profiles and understanding of the samples' structures.

To address this need for scattering analysis methods that are applicable to both conventional structures with existing analytical models and unconventional structures/chemistries that may not have good analytical models, we developed the "computational reverse-engineering analysis for scattering experiments" (CREASE) method.[35] The first step of CREASE, a genetic algorithm (GA), takes as input the SAS scattering profile and some knowledge about the general shape of the assembled structure from other imaging techniques. CREASE then uses the GA to optimize toward relevant dimensions that describe the assembled structure(s) whose computed scattering profile matches the input scattering profile. In the second step, taking as input those relevant dimensions that are output from the first step, molecular modeling and simulation are used to reconstruct the molecular-level packing of the assembly structure. As a proof-of-concept, CREASE has previously been applied to scattering profiles of spherical micelles,[36] cylindrical micelles,[35,37] and vesicles[38] in dilute amphiphilic polymer solutions. CREASE has also been extended to analyze scattering profiles from a concentrated binary mixture of polydisperse spherical nanoparticles to determine the extent of segregation/mixing of the two types of nanoparticles and the precise mixture composition.[39,40] In this paper, we extend CREASE's first step (GA) to analyze the scattering profiles from dilute solutions of semiflexible fibrils with dispersity in fibril dimensions and demonstrate NN-enhanced CREASE-GA's applicability to experimental small-angle X-ray scattering (SAXS) profiles (taken from the literature) from methylcellulose fibrils in aqueous solutions.

Methylcellulose (MC) is a derivative of cellulose in which some or all of the hydrogen atoms on each anhydroglucose unit (repeating unit of cellulose) are replaced by methyl groups.[41] This partial methylation promotes hydrophobicity of the chains and disrupts both the intrachain hydrogen bonds that contribute to the stiffness of the cellulose chain and the interchain hydrogen bonding network that stabilizes cellulose crystallinity, resulting in different structures and phase behavior of MC compared to unmodified cellulose. Due to the abundance of cellulose as a raw material and its benign physicochemical properties, especially solubility in water, MC is used in a wide range of industrial applications such as food, cosmetics, pharmaceutical products, and construction materials.[41] MC is also considered as a potential candidate in self-healing materials,[42] drug delivery,[43] foaming/emulsifying agents,[44] and biodegradable packaging materials.[45]

Past research on MC has shown that the solubility of MC in aqueous/organic solvents is different from unmodified cellulose and is dependent on the chemical structure of the chains themselves.[41,46] MC chains can have a degree of substitution (DS) between 0 and 3 depending on how many of the three hydroxyl groups on each anhydroglucose unit (on average) have their hydrogens replaced by methyl groups. The degree and regularity of the substitutions along the MC chains alter the hydrophilicity/hydrophobicity ratio of the chains and determine the solubility of MC in water. Commercial MCs with DS ≈ 1.7−2.2 and irregular substitution patterns are soluble in water at room temperature.[47] If the DS is too low, the interchain hydrogen bonding network is not disrupted to a sufficient extent; if the DS is too high, the chains become too hydrophobic to be dissolved in water and are more soluble in organic solvents.[41,48] Besides the degree of substitution, the solubility of MC in water is also strongly affected by temperature.[31,43,49−66] Commercial MC with DS ≈ 1.8[41] is known to undergo a two-stage thermoreversible gelation at elevated temperatures in an aqueous solution, where the chains form local, loose bundles in the temperature range from 25 °C to about 42 °C[50] and form a gel and phase-separate from the solution at a temperature above ∼42 °C. Previous studies have attempted to describe the molecular driving force of gelation and the structure of the MC chain assembly after gelation.[47] Kato et al.[67] first argued that gelation occurs by the association of highly substituted, hydrophobic blocks along the chain backbones, as supported by studies of Kobayashi et al.,[68] Li et al.,[58] and others. A broad range of techniques has been used to probe the final MC structure after gelation, including rheological measurements,[31,58] small-angle scattering measurements,[31,61,65] cryo-TEM,[31,61,63,65] and molecular dynamics simulation.[69−73]

Recent studies by Lodge, Bates, and co-workers[31,46,65] provide conclusive evidence about the structure of MC assemblies in water at elevated temperatures. Their TEM images clearly demonstrate that MC chains form fibrils, and using that fibrillar shape as guidance, the authors successfully fit the SAXS profile of the fibrils to an analytical model of a flexible cylinder.[74] Their fitting results suggest that MC fibrils have an average diameter of about 15−20 nm with significant dispersity.[31,71] The thermodynamic/kinetic driving force for chains to maintain the 15−20 nm diameter regardless of the MC chain length remains to be explained.[46] In this work, we analyze the SAXS profiles of MC fibrils using CREASE-GA and compare CREASE-GA output dimensions with the dimensions obtained upon fitting with the analytical model by Lodge, Bates, and co-workers. This comparison can serve both as an independent confirmation of their reported dimension of the fibrils and a validation of CREASE-GA's ability to analyze fibrillar structures in dilute solutions with dispersity in relevant dimensions. MC fibrils are considerably more complex than structures previously studied using CREASE,[35−38,75] with the semiflexibility of the fibril leading to randomness in the shape contour, as well as significant dispersity in dimensions. Successful application of machine learning-enhanced CREASE-GA to study small-angle

Accepted Manuscript
Version of record at: https://doi.org/10.1021/acs.macromol.2c02165

scattering profiles from methylcellulose fibrils serves as an important validation of CREASE's ability to analyze the complex, real-world experimental scattering data and demonstrates its potential to study other unconventional shapes for which no analytical fitting results exist.

II. APPROACH

II.I. Overview of CREASE-GA. In this work, we use genetic algorithm (GA),[76] a heuristic-based optimization technique inspired by the natural evolution process, to search for the "best" semiflexible fibril dimensions that will produce a computed scattering profile ($I_{comp}(q)$) that most closely matches the input experimental scattering profile ($I_{exp}(q)$). In principle, one could use other more sophisticated optimization techniques to accomplish the same outcome as we do with GA. We favor the use of GA for its ease of implementation, which in turn helps overcome barriers for the users of CREASE-GA in experimental groups without significant computational expertise.

In Figure 1 and Table 1, we present a schematic of the workflow of CREASE-GA.

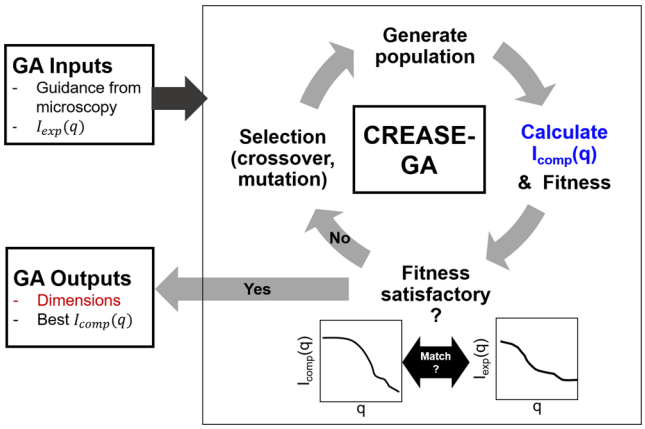


Figure 1. CREASE's genetic algorithm workflow. The parts highlighted in blue and red are varied in the two GA approaches, as described in the text below and listed in Table 1.

Table 1. Comparison between Debye-GA and NN-GA Workflows

GA method	Debye-GA	NN-GA
means to calculate $I_{comp}(q)$	Debye equation (eqs 1a, 1b, and 2)	pretrained neural network
GA outputs	contour length (L) diameter (D) Kuhn length (KL) background scattering intensity ($I_{background}$)	mean diameter (D_{mean}) dispersity index of D (PD_D) Kuhn length (KL) background scattering intensity ($I_{background}$)

We use two different GA approaches in this work:

(1) "Traditional" Debye-equation-based GA (abbreviated as Debye-GA in this paper): this Debye-GA is similar to many of our previous studies[35,39,40,75,77] and is used here to calculate $I_{comp}(q)$ that is compared to the $I_{exp}(q)$ for monodisperse semiflexible fibrils generated from in silico experiments (i.e., computationally generated configurations) with known dimensions. By comparing the dimensions whose $I_{comp}(q)$ "best" matches the $I_{exp}(q)$ to those known dimensions, we validate the efficacy of GA for correctly analyzing scattering profiles from semiflexible fibril structures. This Debye-GA is described in more detail in Section II.II.

(2) "Neural network-evaluated" GA (abbreviated as NN-GA): this GA incorporates a neural network (NN) to avoid the time-consuming evaluation of $I_{comp}(q)$ using Debye-GA. The Debye-GA can be time-consuming, especially when analyzing real experimental $I_{exp}(q)$ from methylcellulose samples with significant dispersity in diameter, such as those collected from SAXS measurements by Lodge, Bates, and co-workers.[31] This NN-GA is described in more detail in Section II.III. When applying NN-GA to experimental SAXS profiles collected by Lodge, Bates, and co-workers, we also adjust the GA outputs (see Table 1) according to our experience during the validation of Debye-GA. These adjustments are discussed in Section III.

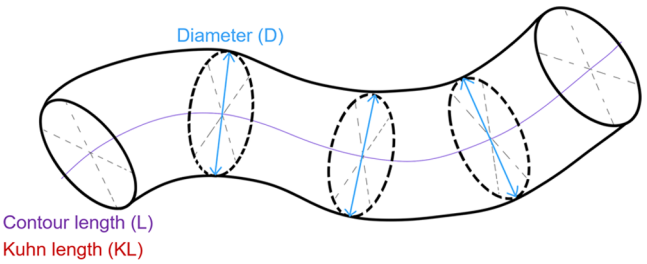


Figure 2. Depiction of the relevant "genes" that capture the structure of a semiflexible cylindrical fibril—Kuhn length (KL), contour length (L), and diameter (D).

II.II. Debye-GA for Monodisperse Semiflexible Fibrils. As depicted in Figure 2, the "genes" used in the CREASE-GA are the relevant dimensions of the semiflexible cylindrical fibril— Kuhn length (KL), contour length (L), and diameter (D). In principle, one could define additional "genes" that express dispersity in "D" along the fibril or dispersity in KL along the fibril or dimensions related to other cross-sections (e.g., ellipse, rectangle). To demonstrate that this CRE ASE -GA approach works, we choose to have genes similar to parameters in existing analytical models that we will compare our results to in the later part of this paper.

For the cases where the input $I_{exp}(q)$ comes from a semiflexible fibril with no dispersity in any dimensions, the traditional Debye-GA starts from an initial generation of 80 fibril configurations (a configuration is referred to as an "individual" following GA terminology) and each individual has a randomly generated set of "genes": contour length (L) (between 20 and 600 nm), diameter (D) (between 5 and 30 nm), and Kuhn length (KL) (between 2.4 and 240 nm). We then use these three values of L, D, and KL to create a contour for each individual and place point scatterers within that contour. In Supporting Information SI Section S.I, we provide more details about how we convert KL to another parameter (σ_θ) that is used to create the configuration of a contour and how we place scatterers within that contour. Each scatterer can be considered as part of an MC monomer, an entire MC monomer, or a collection of MC monomers whose collective electron density contributes to the computed scattering intensity. Using the coordinates of these scatterers, we calculate a computational scattering profile

Accepted Manuscript
Version of record at: https://doi.org/10.1021/acs.macromol.2c02165

($I_{comp}(q)$) for each individual using the Debye scattering equation (eq 1a).

$$I_{comp}(q) = F_M^2(q)S_{MM}(q) + I_{background} \quad (1a)$$

where $S_{MM}(q)$ is the inter-fibril structural factor and $I_{background}$ is the background intensity. For this study, we consider systems at dilute concentrations so that inter-fibril structure factor $S_{MM}(q) = 1$. $F_M^2(q)$ represents the scattering contribution of a single, isolated fibril.

$$F_M^2(q) = f^2 \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\sin(qr_{ij})}{qr_{ij}} \quad (1b)$$

where f is the scattering length density (SLD) of each scatterer (which is a constant in this case since all of the scatterers model the same chemical entity), r_{ij} is the distance between scatterers i and j, and N is the total number of scatterers.

As a result, eqs 1a and 1b can be combined and simplified to

$$I_{comp}(q) \sim \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\sin(qr_{ij})}{qr_{ij}} + I_{background} \quad (2)$$

The instrumental smearing effect,[78] a common source of deviation from the ideal I(q) curve, is not considered in this particular work as the methylcellulose SAXS profiles of interest from Lodge, Bates, and co-workers[31] are believed to be obtained with high enough resolution and have been successfully fit to analytical models without considering smearing. To compare scattering profiles with different absolute (unnormalized) intensities, we normalize the entire scattering profile so that the $I(q)$ value at the smallest q value of interest (q_{min}) is equal to 1. In practice, we first calculate the $I_{comp}(q)$ without $I_{background}$, divide the entire $I_{comp}(q)$ by $I_{comp}(q_{min})$ so that the scattering profile starts at $I_{comp}(q_{min}) = 1$, and add a uniform $I_{background}$ to the entire scattering profile as the last step.

Having calculated the $I_{comp}(q)$, we then assign a "fitness" for each individual based on the extent of match between the individual's $I_{comp}(q)$ and the target $I_{exp}(q)$; the $I_{exp}(q)$ is also normalized so that $I_{exp}(q_{min}) = 1$. We use the weighted sum of log squared error (SSE) as the fitness metric

$$SSE = \sum_{q_i} w_i \left[\log\left(\frac{I_{exp}(q_i)}{I_{comp}(q_i)}\right)\right]^2 \quad (3)$$

where $w_i = \log(q_i/q_{i-1})$, effectively assigning a higher weight to q points further away from each other. We note that SSE gives similar importance to all q points regardless of the magnitude of $I(q)$ by taking the ratio between $I_{exp}(q)$ and $I_{comp}(q)$. One could also use χ^2 ("chi-squared"[2] error instead of SSE here) as long as the scattering intensity at different q values is weighed similarly. χ^2 is a desirable choice to incorporate the uncertainty of scattering measurement when the information representing such uncertainty is available. After calculating the SSE values for all individuals in a generation, we rescale the SSE values for the individuals in that generation to obtain the fitness for each individual using

$$\text{fitness} = X(SSE_{max} - SSE) + Y \quad (4a)$$

where

$$X = (cs - 1)\frac{\max(SSE_{max} - SSE)}{\max(SSE_{max} - SSE) - \text{average}(SSE_{max} - SSE)} \quad (4b)$$

$$Y = (1 - X)\text{average}(SSE_{max} - SSE) \quad (4c)$$

where cs is a constant set to 10 and SSE_{max} is the highest SSE in the generation coming from the individual whose $I_{comp}(q)$ matches $I_{exp}(q)$ the least. After rescaling, the individual with the lowest SSE (i.e., the individual whose $I_{comp}(q)$ most closely matches $I_{exp}(q)$) has the highest fitness value. The rescaling of SSE values reduces the absolute difference between the high-fitness and low-fitness individuals so that the low-fitness individuals have a realistic chance of being selected for the next generation. This approach prevents the overly aggressive elimination of low-fitness individuals, which can lead to premature convergence of the generation before a global optimum is identified.

The fitness values of each individual in the current generation guide the GA to select individuals for the next generation, with the "fitter" individuals [those whose $I_{comp}(q)$s more closely match $I_{exp}(q)$] having a higher chance of being selected. The next generation also consists of 80 individuals, with one individual guaranteed to be the highest-fitness (lowest SSE) individual of the current generation. Such retention of the "best" individual (or one of the "best" individuals) for the next generation is called "elitism" in GA terminology. The other 79 individuals are randomly selected from the current generation, with the possibility for an individual to be selected being proportional to their (rescaled) fitness. Before continuing the next iteration of GA, the selected individuals also undergo genetic operations of "crossover" and "mutation", which aim to maintain an appropriate level of diversity within a generation. We describe these two genetic operations in SI Section S.II. This GA cycle—calculating fitness for the individuals, selection of individuals based on fitness, and genetic operations—is repeated for 150 generations or until the fitness plateaus. We have found that using 150 generations is long enough for the fitness of the best individual in a generation to plateau. At the end of a GA run, the "genes" (values of KL, L, and D) of the fittest individual in the entire GA run are reported. We note that the fittest individual may not necessarily come from the last generation as $I_{comp}(q)$ can have minor fluctuations for the same genes due to randomness in both flexible cylinder contour and scatterer placement.

II.III. NN-GA for Realistic Methylcellulose Fibrils with Dispersity in Diameter. The experimental SAXS profiles for methylcellulose (MC) fibrils obtained by Lodge, Bates, and co-workers come from samples with dispersity in both contour length and diameter of the fibril, as stated in ref 31. To have GA handle scattering from polydisperse samples, we need to adopt an additional step—the $I_{comp}(q)$ of each individual in a generation is the weighted average of $I_{comp}(q)$s calculated for multiple fibrils of different D values sampled from a distribution of D values. If we perform multiple $I_{comp}(q)$ calculations using the Debye equation to have dispersity in D accounted for in each individual's $I_{comp}(q)$, the time needed to finish a single Debye-GA run (Section II.II) until convergence becomes unrealistically long. This has motivated us to use a pretrained artificial neural network (NN), instead of the Debye equation, to evaluate the $I_{comp}(q)$ for a given set of "genes."

Accepted Manuscript
Version of record at: https://doi.org/10.1021/acs.macromol.2c02165

Table 2. Workflow of CREASE-GA vs Analytical Model Fitting of the Small-Angle Scattering Profile[a]

	CREASE-GA	Analytical model fitting
Initial assumptions	General shape of the morphology [semiflexible cylinder]	
Calculating scattering profile of a structure	Through "genes" representing the parameter • Nontrivial for unconventional shapes [how to model Kuhn length?]	Through analytical model • Use preexisting models [flexible cylinder model] • Or develop a new one
Choice of optimization algorithm	Genetic algorithm (GA)	Fitting algorithms provided by software
Extra preparation for NN-GA	• Collect training data • train NN to evaluate I(q) from parameters	
Fitting the I(q) – other user inputs	• GA hyperparameters (No. of generations, No. of individuals per generation, scatterer density)	• Scattering length density • Fitting algorithm hyperparameters

Fitting the I(q) - Caveats	• Reduce dimensionality of the parameters
	  • Remove insignificant parameters [fibril length]
	  • Group interdependent parameters [SLD grouped with intensity scale for single-component chemistry]
	• Dispersity in parameters
	• Instrumental smearing	

Computational resource consumption	• Traditional GA – scales quadratically with scatterer density • NN-eval GA – scatterer density is irrelevant. Usually takes < 1 hour.	• Takes minutes to hours

[a]Manifestation of the steps in this work is highlighted in blue.

The general architecture of NN deployed for the NN-GA is shown in Figure S4. The NN is trained to output $I_{comp}(q)$ for input genes—L, D, and KL from monodisperse fibrils—and background scattering intensity. The NN is implemented with TensorFlow[79] and consists of two fully connected hidden layers, each of 128 neurons. This NN architecture is determined by a grid search of hyperparameters. We use rectified linear units (ReLU) as the activation function and mean squared error (MSE) of output in logarithm ($\log_{10} I(q)$) as the loss metric. The logarithm treatment weighs $I(q)$ similarly regardless of the absolute scale of the $I(q)$, preventing overemphasis on high $I(q)$ values at small q values. We train the NN with 191,840 q–$I(q)$ pairs from 4360 Debye equation-evaluated scattering profiles of monodisperse fibrils randomly generated in the same dimension ranges as specified for Debye-GA in Section II.II. The Adam optimizer is used to optimize the weights and biases in the NN. Similar to the approach taken by Heil et al. on CREASE applied to mixtures of nanoparticles,[40] it is possible to train the NN with q normalized by a dimensionless length scale of interest (in our case, qD) so that the trained model can be applied to fibrils of different length scales from those selected in the training set. In this work, however, we choose to keep q and D as separate inputs with real units due to the difficulty in correctly managing the fibril stiffness (KL) in a dimensionless manner.

As highlighted in Figure 1, there are two key differences for NN-GA applied to polydisperse fibrils and the Debye-GA applied to monodisperse fibrils besides the method used to evaluate $I_{comp}(q)$. First, the dimensions that GA aims to predict

Accepted Manuscript
Version of record at: https://doi.org/10.1021/acs.macromol.2c02165

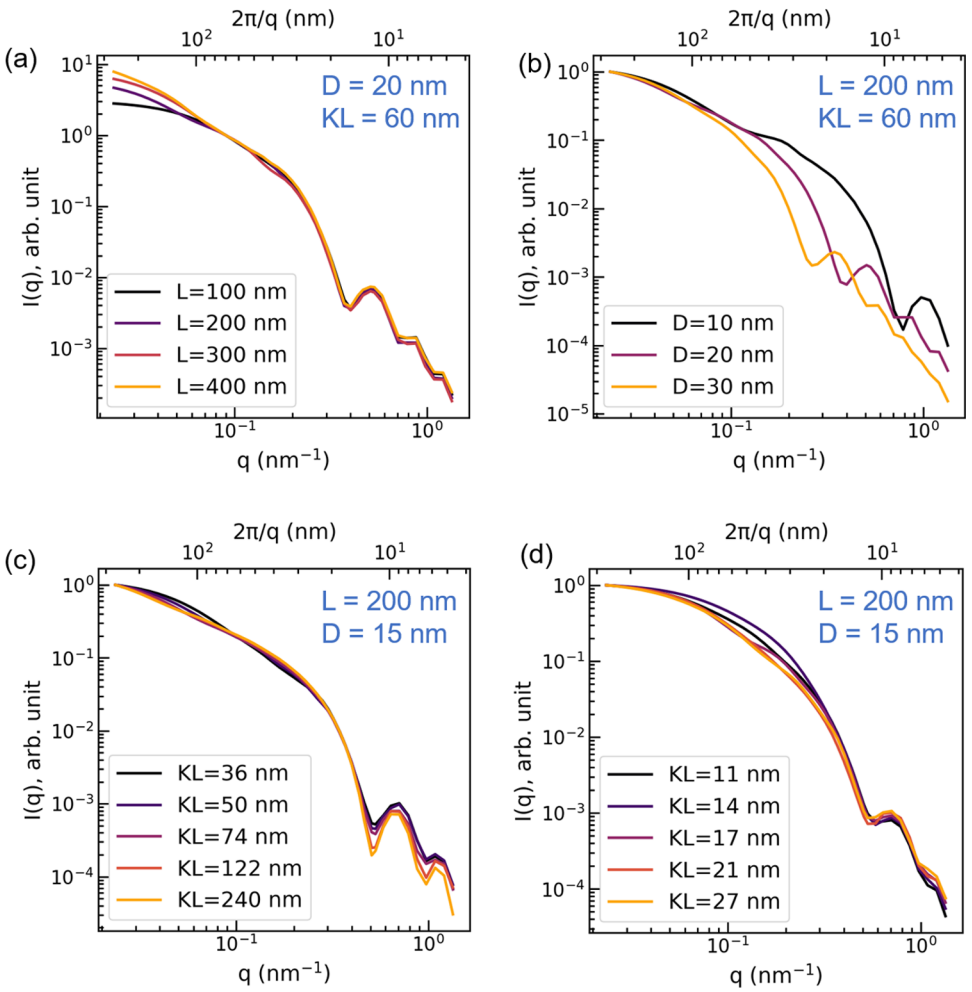


Figure 3. $I_{comp}(q)$ calculated with the Debye equation (eqs 1a,1b and 2) for varying dimensions shown in the legends, at constant values of the other dimensions shown in the blue text. All scattering profiles are normalized to 1 at the minimum q value, q_{min} except for Figure 3a.

become fibril average diameter (D_M), dispersity index of diameter (PD_D), Kuhn Length (KL, transformed into σ_θ as outlined in SI Section S.I), and negative $\log_{10}(I_{background})$. The contour length ($L$) of the fibril is removed from the searched dimensions because, from our experience, the actual contour lengths of the fibrils are too long to have an impact on the $I(q)$ in our q range of interest (see detailed discussion in Section III.I). Instead, we set the L of all individuals uniformly to 400 nm. Second, to evaluate an individual with dispersity in diameters, we first use NN to evaluate $I_{NN,}D(q)$ for 26 monodisperse fibrils with D from 5 to 30 nm (in increments of 1 nm), KL specified by the individual, and L = 400 nm. Then, the total $I_{comp}(q)$ for that polydisperse individual is calculated as

$$I_{comp}(q) = \frac{\sum_{D=5}^{30} I_{NN,D} P_{schulz}(D, D_M, PD_D)}{\sum_{D=5}^{30} P_{schulz}(D, D_M, PD_D)} \quad (5)$$

where $P_{schulz}(D, D_M, PD_D)$ is the probability density function at D under a Schulz distribution of mean D_M and dispersity index PD_D (i.e., the standard deviation of $D_M \times PD_D$).

II.IV. CREASE-GA versus Analytical Fitting Methods. Both the CREASE-GA and analytical fitting methods begin with an assumption of the morphology of the sample (semiflexible cylinder for MC in this work). However, the CREASE-GA is not limited to the number of parameters, as is the case for a predefined analytical model. This can be a decisive advantage for CREASE-GA when dealing with unconventional chemistries or morphologies, which may need additional relevant dimensions to describe the shape that may not be included/possible to include in an analytical model. When dealing with unconventional morphologies, users of CREASE-GA can decide on the relevant parameters that describe that morphology and choose the "genes" to be those parameters. Even though determining the "genes" (i.e., the mathematical descriptors of those relevant morphological parameters) may not be trivial in some cases, representing the "genes" in the context of the morphology should be a much less daunting task than developing a physically meaningful analytical model whose terms contain the parameters relevant to the morphology from a theoretical basis.

In principle, one can couple the GA optimization workflow with any method of evaluating $I_{comp}(q)$ from a set of dimensions, including analytical models. Further, as stated earlier, one could also choose other optimization methods, e.g., Bayesian Optimization. Regardless of the optimization method chosen, the user needs to select the "genes" (i.e., morphologically relevant parameters). The user must balance keeping the number of "genes" as low as possible to avoid "the curse of dimensionality" when searching in high-dimensional parameter space and maintaining enough genes to fully describe the morphologies. In this work, for example, we set L to a constant value for NN-GA upon realizing that it has the minimum impact on $I_{comp}(q)$ in our q range of interest. In the analytical model

Accepted Manuscript
Version of record at: https://doi.org/10.1021/acs.macromol.2c02165

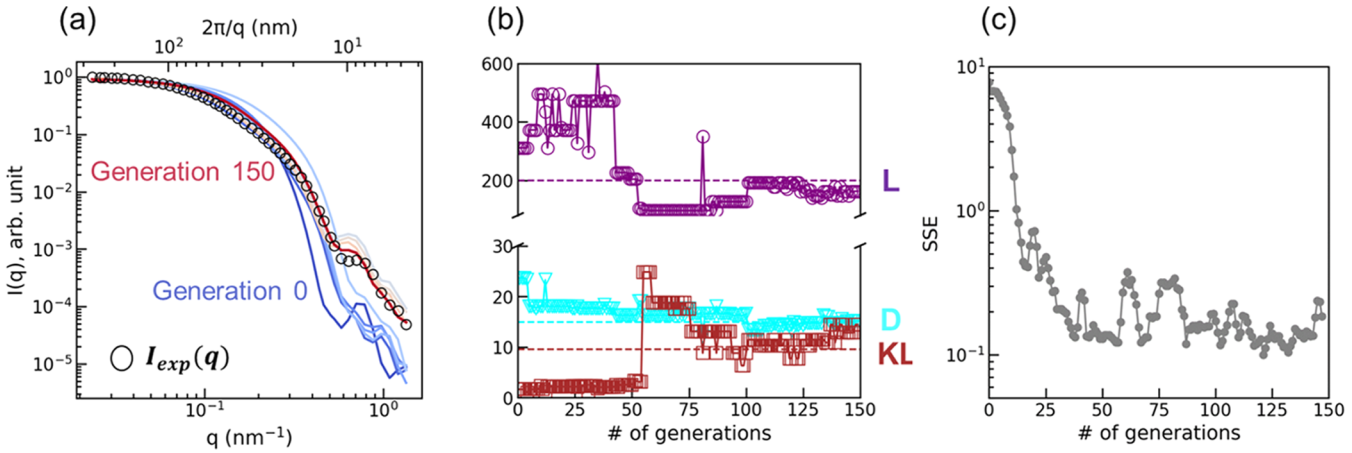


Figure 4. Evolution of the fittest individual in each generation in a single Debye-GA run. (a) $I_{comp}(q)$, (b) contour length (L), diameter (D), and Kuhn length (KL), and (c) sum of squared errors (SSE) of the fittest individual in each generation for a Debye-GA run on in silico $I_{exp}(q)$ from monodisperse fibril with target values of L = 200 nm, D = 15 nm, and KL = 9.6 nm. In (a), an $I_{comp}(q)$ curve is plotted every ten generations from 0th to 150th generation. Dotted lines in (b) mark the target values of the three dimensions. SSE data in (c) are the running average of 5 consecutive generations to smooth out the noise in the data.

fitting, to reproduce results from Lodge, Bates, and co-workers,[31] we also set L to an arbitrarily large value and set SLD of fibrils to 1 and only fit for the scattering intensity as these two parameters are inversely correlated. Therefore, to identify the "genes", first, in our results section, we describe a sensitivity analysis of the scattering profile to changing values of the different genes before demonstrating the CREASE-GA analysis on in silico and experimental scattering profiles.

In Table 2, we summarize the similarity and differences of interpreting the SAS profile using CREASE-GA vs analytical model fitting with common prepackaged software.

III. RESULTS AND DISCUSSION

III.I. Sensitivity of $I(q)$ to the Dimensions of Semi-flexible Fibrils.

We first study the sensitivity of $I(q)$ in the q range of interest based on the reported q range by Lodge, Bates, and co-workers[31] to changes in the three dimensions of the semiflexible fibril, namely, contour length (L), diameter (D), and Kuhn length (KL). This informs us on which "genes" have a more significant impact on $I(q)$ and thus are more easily detected and differentiated by CREASE-GA.

In Figure 3, we compare the $I_{comp}(q)$ curves when one of the three dimensions is changed as the other two dimensions are kept fixed. In all four plots, a secondary axis of $2\pi/q$ (top of the plots) is provided to indicate the length scale in real units to which the different q values in the x-axis correspond to. To quantify how different any two curves in Figure 3 are, we calculate the SSE between pairs of $I_{comp}(q)$s in Figure 3 and present the data in SI Table S1. In Figure 3a, for fixed values of D and KL, as L is increased from 100 to 400 nm, the $I_{comp}(q)$ in general remains unchanged except for small q (q < 0.06 nm^{-1}). In Figure 3b, for fixed values of L and KL as D increases, we see that the $I_{comp}(q)$ shifts toward lower q values while maintaining similar slopes in the intermediate q value range. Both changes in Figure 3a,b are expected as larger length scales correspond to lower q values in a scattering profile. However, the significant changes in $I_{comp}(q)$ with D varying from 10 to 30 nm are more discernible than the changes in $I_{comp}(q)$ for L varying from 200 to 400 nm; the SSE values in SI Table S1 confirm this. In Figure 3c,d, we see a significantly smaller effect of KL values on the computed scattering profile despite large variations in KL values from 11 to 240 nm. The change in the slope of $I_{comp}(q)$ at the intermediate q range is visible only when KL is increased from 11 to 27 nm. The slope of $I_{comp}(q)$ remains mostly unchanged when KL is increased from 36 to 240 nm. The decreasing sensitivity of $I(q)$ to KL with increasing KL is expected since with L kept constant, there are less Kuhn segments in the fibril to capture the stiffness as KL increases. When KL ≈ L, it becomes difficult to assign a definitive value to KL based on the fibril contour.

Based on the results in Figure 3, we expect that for a given $I_{exp}(q)$, CREASE-GA will perform the best in identifying values of D that lead to the $I_{exp}(q)$. According to the experimental scattering measurements by Schmidt et al.[31] of the methyl-cellulose fibrils, the L values lie in the range of 80−400 nm with (most likely) high dispersity. This suggests that CREASE-GA will perform poorly in identifying the L values that lead to the $I_{exp}(q)$. In addition, L mainly affects $I(q)$ in the low q region (q < 0.05 nm^{-1}), which we know from our experience is likely affected by interfibrillar interactions [the $S(q)$] and generally ignored (at dilute concentrations) during the fitting to obtain the form of the assembled structure. As a result, we remove L from the list of "genes" output by CREASE-GA when analyzing experimentally generated scattering profiles (discussed later in Section III.III) and instead simply set L to be a large value (400 nm). We retain KL as one of the "genes" that GA identifies for a given $I_{exp}(q)$ but do not expect GA to provide quantitatively accurate predictions on the value of KL unless the KL of MC fibrils are lower than ~25 nm (maximum KL in Figure 3d).

III.II. Analysis of "In Silico" Monodisperse Semiflexible Fibrils Using Debye-GA.

Before we apply CREASE-GA to $I_{exp}(q)$ of MC fibrils, we first validate CREASE-GA using as input the in silico $I_{exp}(q)$ generated using the Debye equation with known target dimensions and without any dispersity in dimensions.

In Figure 4a, we show how $I_{comp}(q)$ of the fittest individuals become similar in the later generations of the GA, signaling the gradual convergence of fitness of "best" individuals in the later generations of the GA. The only observable minor difference is the $I(q)$ values at q = (5−6) × 10^{-1} nm^{-1}. Correspondingly, in Figure 4b, the D and KL values determined at the end of this GA run approach the target dimensions. Even though, in this specific case, the L also approaches the target dimension, we show in SI

Accepted Manuscript
Version of record at: https://doi.org/10.1021/acs.macromol.2c02165

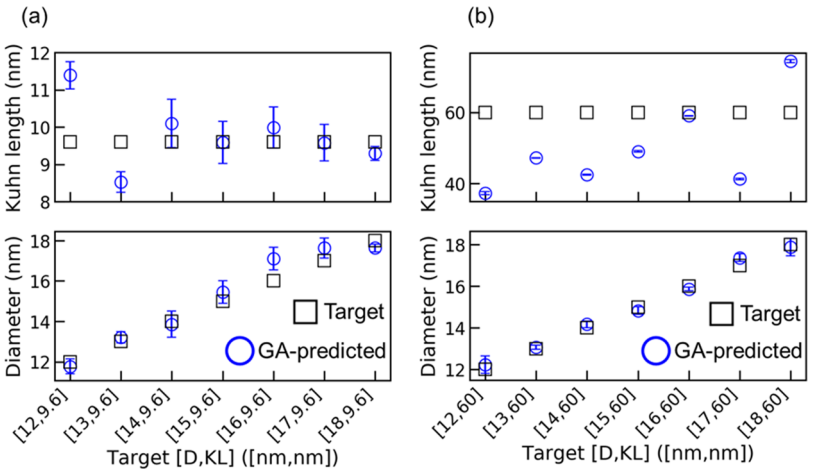


Figure 5. Debye-GA determined dimensions for target diameter (D) = 12−18 nm and Kuhn length (KL) = (a) 9.6 nm and (b) 60 nm. Black squares represent target dimensions and blue circles represent Debye-GA-predicted dimensions of the fittest individual. Error bars for the GA-predicted dimensions indicate standard deviation from the predicted dimensions of the fittest individuals from three independent GA runs for each target system.

Figure S5 that such quantitative accuracy of L is not reproducible across multiple repeated runs and varying target dimensions. Figure 4c shows that SSE decreases rapidly at early generations but eventually stabilizes and fluctuates around 0.1 after ∼120th generation, signaling the fittest individual in each generation has stopped improving. These results in Figure 4 indicate that the GA can make quantitatively relevant predictions for D and KL of a monodisperse semiflexible fibril.

Next, we evaluate Debye-GA on scattering profiles generated from in silico, monodisperse fibrils with a range of target D and KL and summarize the performance of traditional Debye-GA in Figure 5. In Figure 5, the error bars indicate the standard deviation between three independent Debye-GA runs on the same set of dimensions. We run multiple GA runs to ensure that we can identify multiple fit individuals whose $I_{comp}(q)$ may closely match the $I_{exp}(q)$. We choose these target D values (12−18 nm) to match the range of MC fibril diameters determined by Schmidt et al.[31] from TEM and analytical fitting to their SAXS data (∼18 nm on average, ∼12 nm at the thinnest, most densely packed region of the fibril).

At both the low (9.6 nm) and high (60 nm) target KL values, the GA quantitatively reproduces the target value of D, with errors rarely exceeding ±1 nm. As expected, based on Figure 3, the determinations of KL are not as accurate as D, with the error in KL prediction being in the range of 20−50%. KL values are more reliably predicted by the GA when the target value of KL is low (corresponding to the range of KL values where we observed changes in the slope of $I(q)$ in Figure 3d). In SI Figure S5, we present Debye-GA prediction of L on the same two series of target D and KL dimensions as in Figure 5 and observe poor agreement with the target L values.

Overall, the performances of CREASE-GA on predicting D, KL, and L are as expected based on Section III.I's conclusion of sensitivity of $I(q)$ to these three parameters. CREASE-GA predicts fibril D with quantitative accuracy, offers a qualitative estimation of KL, and performs poorly with L.

The results in this section validate GA as a reliable tool to determine key dimensions of interest for semiflexible fibrils, especially the fibril diameter, which is of most interest in the context of MC solution studies by Lodge, Bates, and co-workers.[31] However, a critical difference between the in silico fibrils in this section and the MC fibrils we eventually aim to apply CREASE-GA to is that MC fibrils are known to carry significant dispersity in diameters[31] with thin, densely packed crystalline regions and thick, loosely-packed semicrystalline or even amorphous regions. In the next section, we demonstrate an improvement in the traditional Debye-GA using neural networks (NN) and use this improved NN-GA for analyzing SAXS profiles from MC samples with high dispersity in diameter provided by Lodge, Bates, and co-workers.[31] Our aim is not only to predict the diameter of the fibrils but also directly determine the extent of dispersity in the diameter.

III.III. Comparison of NN-GA to Debye-GA on Scattering Profiles from In Silico, Polydisperse Fibrils. To illustrate the limitations of Debye-GA, we use Debye-GA on $I_{exp}(q)$ generated from in silico fibril samples with dispersity in diameter. To generate such a sample, we use the Debye equation to compute $I(q)$s of monodisperse fibrils of L = 200 nm, D = 5, 6, 7, ⋯, 30 nm at a constant KL (9.6 or 60 nm) and then calculate the weighted average of all of these $I(q)$s with weights based on the normal distribution of mean diameter = 15 nm, standard deviation = 1 nm. The averaged $I(q)$ then becomes the in silico $I_{exp}(q)$ representing a polydisperse fibril sample of mean diameter 15 nm, standard deviation 1 nm under the normal distribution. The choice of a normal distribution is arbitrary here as we are using in silico fibril samples to evaluate Debye-GA. When extending CREASE for analysis of experimental fibril samples, one should choose the type of distribution (e.g., Schulz distribution) that mimics reality. We also adjust which "genes" of Debye-GA are fixed (not optimized) and which "genes" are optimized for output, according to the lessons we learned in Section III.II. We set L of all individuals in Debye-GA to 200 nm since we have shown in Section III.II that L does not significantly impact the scattering profile in our q range of interest, and use Debye-GA to determine the genes of the GA—the mean diameter (D_{mean}), dispersity index (PD_D = standard deviation of diameter/mean diameter, assuming normal distribution), KL, and background intensity. To incorporate dispersity in Debye-GA, for each individual with genes of [D_{mean}, PD_D, KL, $I_{background}(q)$], we calculate using the Debye equation the $I(q)$ of 7 monodisperse fibrils with the diameter $(1 - 1.5PD_D) \times D_{mean}$, $(1 - PD_D) \times D_{mean}$, $(1 - 0.5PD_D) \times D_{mean}$, ⋯, $(1 + 1.5PD_D) \times D_{mean}$. Then, we use their weighted average based on the corresponding normal distribution defined by (D_{mean}, PD_D)

Accepted Manuscript
Version of record at: https://doi.org/10.1021/acs.macromol.2c02165

to arrive at the $I_{comp}(q)$ of the entire polydisperse individual. While ideally, more than 7 data points should be included for the weighted average to better represent the normal distribution, the time needed for a Debye-GA run will also significantly increase. In Figure 6, we show the performance of Debye-GA in reproducing the dimensions of those in silico fibril samples.

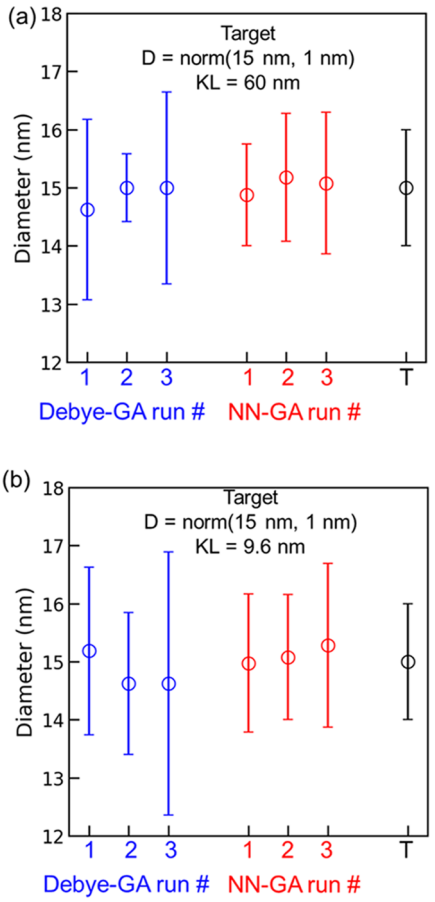


Figure 6. Fibril diameter distribution determined by Debye-GA (in blue) and NN-GA (in red) vs corresponding target values (in black) of in silico fibrils, whose target diameter distributions are represented by a normal distribution of mean 15 nm, standard deviation of 1 nm, and whose Kuhn lengths are (a) 60 nm or (b) 9.6 nm. Circles indicate mean values and error bars indicate standard deviations of the normal distribution representing fibril diameter.

In Figure 6, the predicted mean diameters (blue circle) and dispersity in diameter (depicted via blue error bars calculated as standard deviations of the normal distribution representing D) from three independent Debye-GA runs are compared to the target diameter distributions (black circle and black error bars). We see that at both high KL (Figure 6a) and low KL (Figure 6b), the predicted distribution of D is consistently similar to the target D distribution. Figure S6a,b compares the KL determined by Debye-GA against the target KL values. We notice a trend like what we saw in Section III.II for monodisperse fibrils that for these polydisperse fibrils at high KL, there is a considerable difference between the GA-determined KL and the target value. At low KL, the target values are consistently reproduced with quantitative relevance.

Even though the accuracy of Debye-GA determined dimensions is good in Figure 6, the time needed to do these Debye-GA calculations hinders its applicability to fibrils with dispersity in dimension. Despite using a low number of data points to sample the normal distribution (7) and a decreased scatterer density to fill the contour (0.272 nm^{-3} compared to 0.544 nm^{-3} originally used for Debye-GA in Section III.II), both aiming to speed up the GA at the expense of introducing error to calculated $I_{comp}(q)$, a full Debye-GA run of 150 generations still takes almost a week on one 32-core node (AMD EPYC 7002 processor) on the University of Delaware-based DARWIN supercomputing cluster [https://dsi.udel.edu/core/computational-resources/darwin/]. Such slow speed impedes Debye-based CREASE-GA's application as a general tool for researchers with limited access to computational resources and less time to wait to interpret their scattering results. This has motivated us to replace the rate-limiting step of the workflow—the Debye equation calculation—with a neural network (NN) that rapidly outputs the $I_{comp}(q)$ for a given set of genes. We show the NN-GA's results (red symbols) along with the Debye-GA's results (blue symbols) and target values (black symbols) in Figure 6a,6b. We discuss this comparison in more detail after we present key details about the NN-GA approach, namely, the NN model training and $I_{comp}(q)$ calculation, next.

As described in Section II.III, we train our NN using scattering profiles computed using the Debye equation of more than 4000 fibrils generated in silico with random combinations of L, D, and KL. The use of NN accelerates the $I_{comp}(q)$ calculation during GA by moving away from scatterer placements and intensive pair-wise Debye equation calculations to a machine learning model (NN) that predicts $I_{comp}(q)$ directly for given input dimensions. To demonstrate that the NN model correctly predicts the $I_{comp}(q)$ for a given set of dimensions, in Figure 7, we compare the NN-evaluated $I_{comp}(q)$ against the Debye-evaluated $I_{comp}(q)$ for those fibril dimensions. We show this comparison for varying values of D from 10 to 30 nm and KL values from 3.75 to 60 nm; we fix L = 200 nm. The $I_{comp}(q)$s predicted by the trained NN model (shown in lines in Figure 7) overlap with the $I_{comp}(q)$s calculated using the Debye equation (shown with symbols in Figure 7). There are deviations between the two curves in certain cases, marked by arrows in Figure 7a–c. However, such detailed features tend to disappear for polydisperse fibrils due to the averaging between $I(q)$s from fibrils of different diameters, so we consider these deviations acceptable.

Having shown in Figure 7 that the $I_{comp}(q)$ from NN matches the $I_{comp}(q)$ from the Debye equation for the same set of monodisperse fibril parameters, we compare the accuracy of NN-GA against Debye-GA with the same in silico $I_{exp}(q)$ from fibrils with dispersity in diameter as deployed in Figure 6. Because the evaluation of $I_{comp}(q)$ using the NN takes little time compared to the computationally intensive Debye equation, we are able to use 26 diameter values from 5, 6, ⋯, to 30 nm to represent the normal distribution of diameters, compared to only 7 data points for Debye-GA.

Figure 6a,b represents the predicted mean diameter and dispersity in diameter from three independent NN-GA runs (shown in red) on each in silico $I_{exp}(q)$ compared to the target diameter distribution (shown in black) and the Debye-GA prediction (shown in blue), and Figure S6a,b represents the same comparison for predicted KL. Compared to the respective predictions made with Debye-GA, NN-GA predictions either perform as well as Debye-GA or, in some cases, better with a higher level of accuracy than Debye-GA. In Figure S7, we show a comparison of Debye-GA predicted dimensions vs NN-GA predicted dimensions on targets with higher dispersity in

Accepted Manuscript
Version of record at: https://doi.org/10.1021/acs.macromol.2c02165

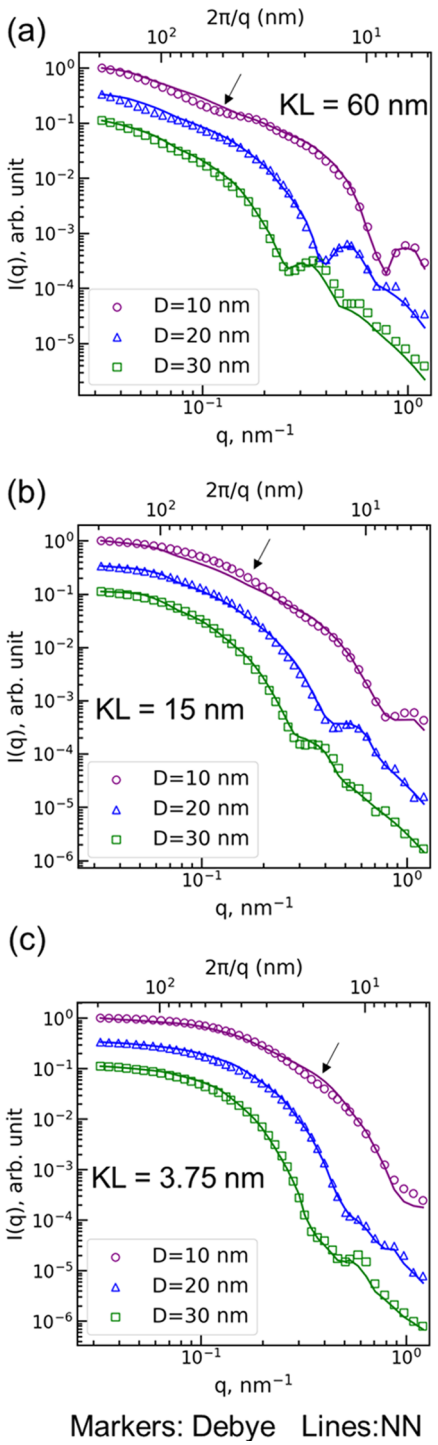


Figure 7. $I_{comp}(q)$ evaluated using the Debye equation calculation (markers) and NN (lines) for dimensions of L = 200 nm and (a) KL = 60 nm, (b) KL = 15 nm, and (c) KL = 3.75 nm. $I_{comp}(q)$s from Debye equation calculation are generated using a scatterer density of 0.544 nm^{-3}. Arrows mark regions of deviations.

diameter, cases where NN-GA performs better than Debye-GA. The NN-GA predicted diameter distributions are indeed significantly *closer* to the target than Debye-GA predictions (Figure S7a,b).

Overall, these comparisons in Figures 6, S6, and S7 suggest that NN-GA is faster than Debye-GA and can perform either as well as or better than Debye-GA in reproducing the target dimensions. This improved performance of NN-GA over Debye-GA is surprising at first because the NN is supposed to be a proxy for the Debye equation mimicking the "genes"-to-$I_{comp}(q)$ relationship based on data from Debye equation. However, the procedure of calculating $I_{comp}(q)$ in Debye-GA relies on (1) generating a random contour of a flexible cylinder and (2) generating randomly placed scatterers within the contour, both steps involving randomness that can cause fluctuations in the computed $I_{comp}(q)$. As a result, individuals carrying "genes" that are further from the target parameters can coincidentally produce an $I_{comp}(q)$ with lower SSE (closer match) than the $I_{comp}(q)$ produced by the candidate carrying genes closer to the "target" parameters, effectively preventing the GA from converging to the true optimal solution. This issue can be alleviated by using multiple replicate configurations to produce an average $I_{comp}(q)$ or more monodisperse fibrils to represent the individual, which has a disperse distribution of diameter, but both would further slowdown the workflow. This is against our desire to make CREASE widely accessible to the general research community regardless of access to sophisticated computational resources. Unlike the Debye-GA, the $I_{comp}(q)$ calculation using the NN is based on training given to the NN on computed scattering profiles of over 4000 fibrils generated across the parameter space. This inherently averages out the random fluctuations arising from the Debye-based $I_{comp}(q)$ calculation by minimizing the total loss to all scattering profiles in the training set. Further, NN-GA uses more diameter values (26 values) than the Debye-GA (7 values) to represent the normal distribution of diameter, further reducing errors in the calculated $I_{comp}(q)$. As a result, NN-GA is free of the random fluctuation in $I_{comp}(q)$ that the Debye-GA is plagued with and can more consistently converge to the optimal candidate than the Debye-GA.

Having demonstrated that the NN-GA can output target dimensions, we discuss the improvement in speed brought about by the NN-GA in more detail. In Figure 8, we compare the typical time needed to complete a 150-generation Debye-GA vs NN-GA run. For Debye-GA, the time shown in this figure is specific to a run of 80 individuals/generation, 150 generations, scatterer density of 0.544 nm^{-1}, 7 monodisperse $I(q)$s to represent the distribution of diameter for each polydisperse individual, on one 32-core node on University of Delaware's

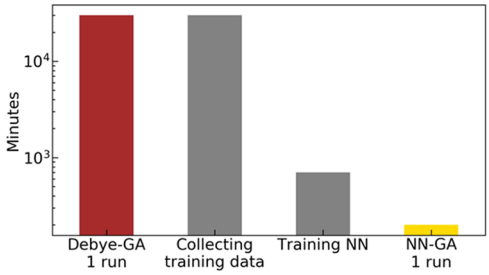


Figure 8. Estimated time needed for collecting training data and training the NN for the NN-GA, a complete run of typical Debye-GA, and a complete run of typical NN-GA. The time estimation of Debye-GA refers to a run of 80 individuals/generation, 150 generations, scatterer density of 0.544 nm^{-1}, 7 monodisperse $I(q)$s to represent the distribution of diameter per individual, on one node on the UD DARWIN supercomputing cluster. The time estimation of the NN-GA refers to a run of 80 individuals/generation, 150 generations, and 26 monodisperse $I(q)$s to represent the distribution of diameter per individual on a 4-core laptop.

Accepted Manuscript
Version of record at: https://doi.org/10.1021/acs.macromol.2c02165

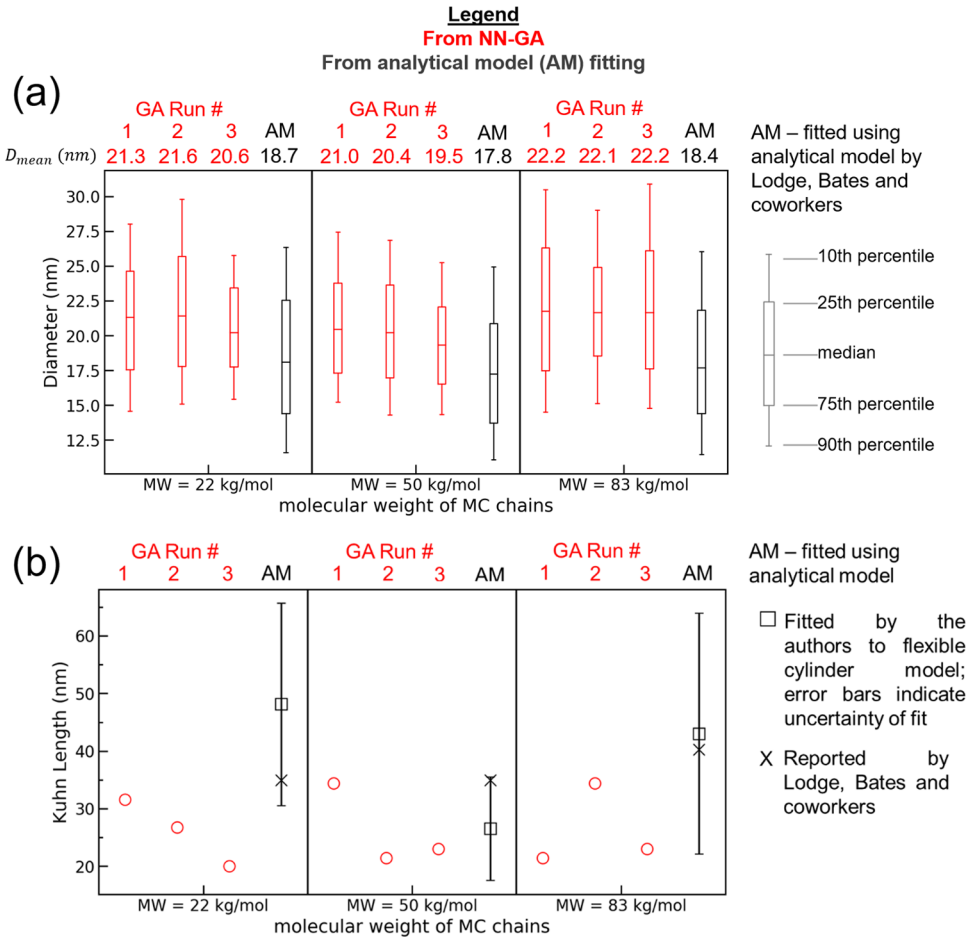


Figure 9. (a) Fibril diameter distribution and (b) Kuhn length determined by three independent NN-GA runs (red) and analytical fitting using the flexible cylinder model (dark gray) for three methylcellulose fibrils formed by chains at molecular weights (MW) 22, 50, and 83 kg/mol and at 0.3 wt %, assuming a constant number of methylcellulose monomers packed in fibrils with dispersity in diameter ("1st assumption"). The box plots in (a) represent the 10th, 25th, 50th, 75th, and 90th percentile of the Schulz distribution based on the mean diameter (D_{mean}) and dispersity index (PD_D) determined by each method. Determined D_{mean} values are presented on top of the figure.

DARWIN supercomputing cluster. For NN-GA, the time shown in Figure 8 is specific to a run of 80 individuals/generation, 150 generations, and 26 monodisperse $I(q)$s to represent the distribution of diameter per individual on a mid-tier 4-core laptop (Lenovo ThinkPad T490 carrying Intel Core i5 CPUs). These time estimations in Figure 8 can be different as one varies the number of individuals in a generation, the number of $I_{comp}(q)$ evaluations for each individual, scatterer density for the Debye-GA, or the computer hardware. Regardless, the time needed for an NN-GA on a mediocre laptop is negligible (~1 h) compared to a Debye-GA run for days to weeks on a sophisticated supercomputer.

We note that there are other methods developed to compute $I_{comp}(q)$ from the collection of scatterers that are faster than deploying the full-scale Debye equation.[80] However, they all scale to some degree with the number of beads/voxels used to represent the structure, while NN does not depend on any beads/voxels and is consistently fast. It is this acceleration enabled through the use of NN-GA that allows us to fully incorporate dispersity in D at high resolution and do 26 $I_{comp}(q)$ evaluations (one for each integral D value from 5−30 nm) for each individual in the GA with ease, in turn helping to improve the accuracy of prediction. Before one can use an NN-GA, there is indeed time and labor necessary for the collection of training data and training the neural network with optimal neural network architecture. In our case, the total computational time to collect about ~4000 individual scattering profiles is equivalent to about 600 h on a single node of an 18-core AMD EPYC 7002 processor on the Caviness Supercomputing Cluster at the University of Delaware.[81] However, this seemingly significant i nitial t ime i nvestment c an b e a lleviated b y the common approach of distributing the workload to several nodes and only needs to be done once before the trained NN model can be applied to all of the scattering profiles f rom similar morphologies. We also note that we generate a large amount of training data at high scattering density for the optimal performance of the NN-GA, and this level of resolution can be unnecessary if one merely aims to use CRE ASE -GAfor less demanding tasks such as a qualitative characterization or quick initial scanning of the scattering data.

Having shown the NN-GA's superiority in both accuracy and speed over the Debye-GA, we next use this NN-GA to determine the relevant dimensions for the systems that generated the experimental SAXS profiles obtained by Lodge, Bates, and co-workers.

III.IV. NN-GA Analysis of SAXS Data from Methylcellulose Fibrils with Dispersity in Diameter. In this section, we consider scattering profiles obtained from SAXS experiments conducted by Lodge, Bates, and co-workers and published in ref 31. The scattering profiles come from three

Accepted Manuscript
Version of record at: https://doi.org/10.1021/acs.macromol.2c02165

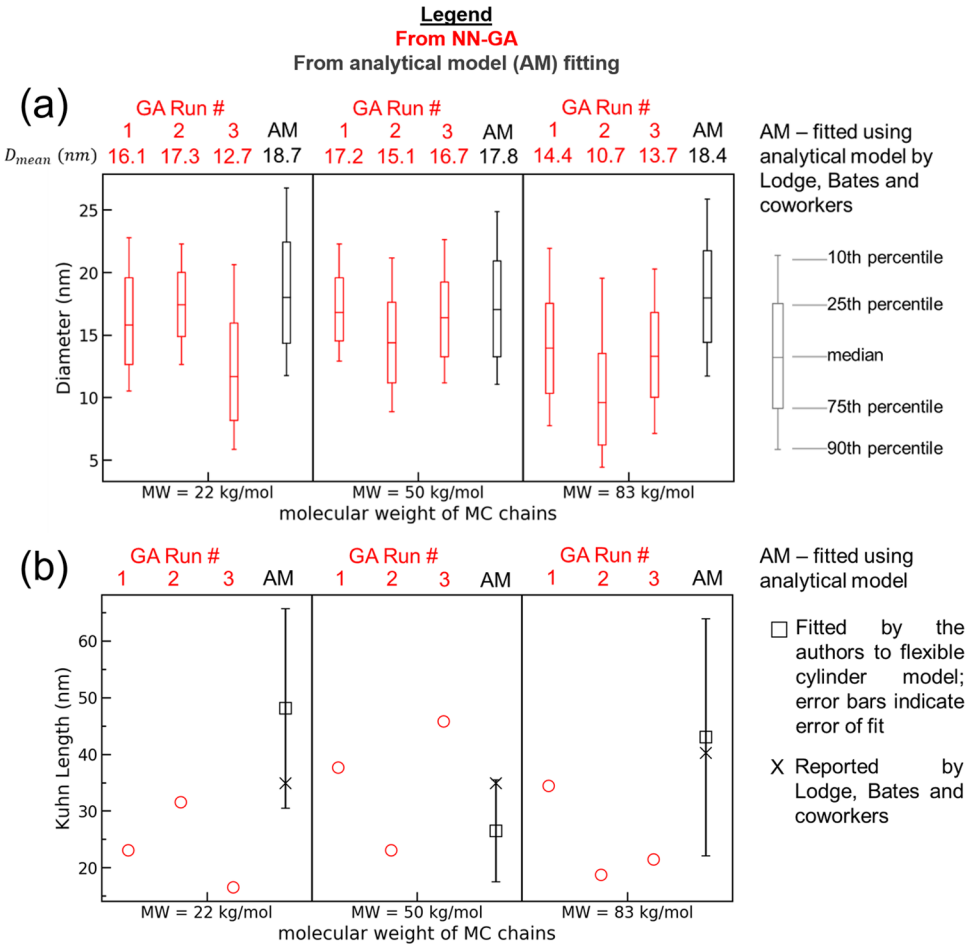


Figure 10. (a) Fibril diameter distribution and (b) Kuhn length determined by three independent NN-GA runs (red) and analytical fitting using the flexible cylinder model (dark gray) for three methylcellulose fibrils formed by chains at molecular weights (MW) 22, 50, and 83 kg/mol and at 0.3 wt %, assuming constant packing density of methylcellulose monomers packed in fibrils with dispersity in diameter ("2nd assumption"). The box plots in (a) represent the 10th, 25th, 50th, 75th, and 90th percentile of the Schulz distribution based on the mean diameter (D_{mean}) and dispersity index (PD_D) determined by each method. Determined D_{mean} values are presented on top of the figure.

methylcellulose (MC) solutions with MC chain molecular weights (22, 50, and 83 kg/mol) at a 0.3 wt % concentration. We run NN-GA on these input $I_{exp}(q)$ and compare the NN-GA's predicted dimensions—*mean D* (D_{mean}), dispersity index in *D* (PD_D), and KL—to those fitted using the analytical flexible cylinder model.[82,83] The results of the analytical model fit are published in the paper by Lodge, Bates, and co-workers.[31]

It is important to note that it is not known how the packing density of MC chains in the fibrils changes as the fibril diameter increases. Lodge, Bates, and co-workers have hinted that the crystallinity of the fibril is low in the fibrillar regions with larger diameters,[31,46] implying that the packing density of MC chains is smaller in regions of the fibril with large diameters. This change in the packing density of the chains should affect the scattering contribution of fibrils with different diameters to the averaged overall scattering. For example, if the packing density of chains were similar across different fibril diameters, thicker fibrils would have more methylcellulose monomers packed within their contour and make a larger contribution to the scattering than thinner fibrils due to a higher absolute (unnormalized) scattering intensity. To reflect this effect in our work, we consider two scenarios:

(1) fibrils, regardless of their diameters, always have the same number of MC monomers packed in the fibril (i.e., thicker fibrils have lower packing density), represented by similar absolute scattering intensity and

(2) fibrils, regardless of their diameters, have the same packing density of MC monomers, i.e., thicker fibrils have proportionately more MC monomers (more "scatterers" in the Debye equation), leading to their scattering intensity scaled by the second order against their available volume.

We expect the scenario in experiments to be somewhere between the above two cases: as the fibril diameter increases, the number of MC monomers packed in the contour should increase; however, the packing density should also decrease.

In Figure 9a, we present the predicted D_{mean} and PD_D together with the box plot representing the entire Schulz distribution[84] defined by these two parameters taking the first scenario—a constant number of MC monomers packed in fibrils regardless of diameter. We choose the Schulz distribution to describe the dispersity in fibril diameter following the work of Lodge, Bates, and co-workers.[46] We perform three independent NN-GA runs for each sample to study the possible variation in predictions between the GA runs. The NN-GA output is consistent between the three independent runs in both the value and spread of the *D* distribution and predicts a slightly higher *D* than the analytical model fitting in all cases (determined D_{mean} shown on top of the

Accepted Manuscript
Version of record at: https://doi.org/10.1021/acs.macromol.2c02165

figure). Despite the quantitative difference, the two methods—NN-GA and analytical model fit—agree that the average D of the fibrils is around 20 nm regardless of the molecular weight, corroborating the key conclusion of Lodge, Bates, and co-workers[31] that the molecular weight of MC chains does not impact the D of the formed fibril.

For the first scenario, in Figure 9b, we present the determined KLs of NN-GA and KLs from the analytical model fitting of the MC fibrils. We note that Lodge, Bates, and co-workers only reported a maximum likelihood estimation for KL in ref 31 without the related prediction uncertainty (error of prediction). From our own experience in fitting these $I_{exp}(q)$s of MC fibrils with the same analytical model, the fitted value of KLs has significant uncertainty. In fact, the reported KL by Lodge, Bates, and co-workers through analytical model fitting bears significant fluctuation across different MC volume fractions and chain lengths.[31] For example, the KLs of MC fibrils from 50 kD chains fluctuate from 9 to ∼110 nm with varying MC chain volume fractions. Since it is unlikely that the stiffness of fibrils changes significantly with the MC chain volume fraction, this high fluctuation is likely an indication of high uncertainty like what we see. In Figure 9b, we capture this uncertainty by plotting average KL values and error bars from our own fits using the analytical model. When we do our own analytical model fit, we fix all other dimensions (besides KL) to values reported in ref 31 for the best possible comparison to Lodge, Bates, and co-workers' results and scale the q points by $\sqrt{I(q)}$. In Figure 9b, we also indicate the KL values fitted by Lodge, Bates, and co-workers as reported in ref 31 with crosses. As expected, there are significant discrepancies both among the three NN-GA trials for each sample as well as between NN-GA predictions and analytical model fitting results. This is not surprising as we know from our study on monodisperse fibrils that there is high uncertainty in CREASE-GA's predicted KLs for semiflexible fibrils at this level of stiffness. Assuming that the KL value is not affected by MC chain molecular weight, the NN-GA output in Figure 9b suggests that the KL for MC fibrils is between 20 and 35 nm.

In Figure 10a, we present the predicted D_{mean} and PD_D still assuming Schulz distribution under the second scenario—constant packing density of monomers within fibrils of disperse diameter. We see that the predictions between 3 independent runs for each MW are consistently and slightly below the analytical model fit diameters by Lodge, Bates, and co-workers, predicting a mean diameter of about 15 nm in most cases. In Figure 10b, we present the predicted KL of MC fibrils; similar to those predicted under the first scenario (Figure 9b), the determined KLs fluctuate around 20−35 nm.

We cannot say which of the two sets of results under the two scenarios—(1) fibrils, regardless of their diameters, always have the same number of MC monomers packed in the fibril or (2) fibrils, regardless of their diameters, have the same packing density of MC—is more accurate. However, we know that the two sets of predictions are the two extremes, and the realistic case lies somewhere in between the two scenarios. Regardless, the NN-GA predicted mean diameter of 15−20 nm agrees with the analytical fitted results by Lodge, Bates, and co-workers and serves as validation for our NN-enhanced CREASE-GA method.

IV. CONCLUSIONS

In this study, we presented a machine learning-enhanced CREASE (computational reverse-engineering analysis for scattering experiments) method for analyzing scattering results for generic fibrillar structures in solution and applied CREASE specifically to small-angle X-ray scattering (SAXS) results from methylcellulose fibrils obtained by Lodge, Bates, and co-workers (SAXS data published in ref 31).

To reduce the time needed to run a Debye-equation-based genetic algorithm (GA) optimization, we developed a neural network-evaluated GA (NN-GA) method to evaluate $I_{comp}(q)$ for comparison to $I_{exp}(q)$. For both the Debye-equation-based $I_{comp}(q)$ calculation and NN-evaluated $I_{comp}(q)$, we presented a number of results where we used $I_{exp}(q)$ from in silico structures as input to CREASE-GA and demonstrated that CREASE-GA is able to output dimensions of fibrils in the in silico structures with quantitative accuracy. After this validation on in silico structures, we applied CREASE-GA to experimental SAXS profiles obtained by Lodge, Bates, and co-workers.[31] The CREASE-GA-determined average diameters are similar to those fitted by Schmidt et al. (∼15-20 nm) using an analytical model and confirm their result of consistent diameters of methylcellulose fibrils across different chain molecular weights.

Following this successful determination of MC fibril dimensions using CREASE-GA, which is the first step of CREASE, our ongoing effort focuses on the second step of CREASE—the molecular reconstruction step—where we study the internal packing of MC fibrils using molecular modeling and simulation. These results may also provide information on how the packing density of MC monomers changes with variations in the fibril diameter and elucidate which of the two scenarios we considered in CREASE-GA—(1) fibrils, regardless of their diameters, always have the same number of MC monomers packed in the fibril or (2) fibrils, regardless of their diameters, have the same packing density of MC—is more accurate.

Through this demonstration of the application of CREASE to experimental scattering profiles and comparison of the CREASE-GA results to published results from analytical model fits, we prove that CREASE can predict the dimensions of the fibrils with both accuracy and efficiency. This gives us the confidence to apply it to other experimentally relevant morphologies with more sophisticated structures and "genes" representing those structural dimensions where there is no good and appropriate analytical model to interpret small-angle scattering data. Our incorporation of the neural network into GA also demonstrates its efficiency and accuracy in dealing with dispersity in dimensions, expanding CREASE's applicability as a tool for scattering profile interpretation in the broader scattering community. Even though in this work, the CREASE-GA results were in agreement with analytical model results, in the past, we have shown that the CREASE method can sometimes perform better than analytical fitting for predicting certain dimensions of other assembled structures that exhibit size dispersity (e.g., solution of vesicles[77]). As shown in the last part of this study, CREASE can also prove useful when the researcher must consider various scenarios of molecular packing that impact scattering. In such cases, developing new analytical models for every scenario can be difficult, and CREASE may be easier to use. Finally, CREASE will be powerful in interpreting scattering results from concentrated MC solutions where the structure factor will be coupled with the form factor, and the existing analytical structure factor models may not be applicable or too approximate.

The code in this work has been incorporated into our open-source CREASE-GA python package [https://github.com/arthijayaraman_lab/crease-ga] and is open for view and download for readers interested in learning more about and/or using our CREASE-GA.

ASSOCIATED CONTENT

Supporting Information

Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.macromol.2c02165.

Additional details on the implemented flexible cylinder model and CREASE-GA architecture; additional details on the NN architecture; and additional results (PDF)

AUTHOR INFORMATION

Corresponding Author

Arthi Jayaraman − *Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, Delaware 19716, United States; Department of Materials Science and Engineering, University of Delaware, Newark, Delaware 19716, United States;* orcid.org/0000-0002-5295-4581; Email: arthij@udel.edu

Author

Zijie Wu − *Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, Delaware 19716, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.macromol.2c02165

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge financial support from the National Science Foundation, DMR CMMT Grant #2105744. The authors are grateful to Professor Timothy Lodge, Dr. Lucy L. Solomon, and Professor Frank Bates for graciously providing raw experimental SAXS data and useful feedback. The computational work in this paper is supported using the Caviness supercomputing cluster at the University of Delaware and DARWIN supercomputing cluster at the University of Delaware, the latter supported by the National Science Foundation, Grant #1919839.

REFERENCES

(1) Guinier, A.; Fournet, G.; Yudowitch, K. L.*Small-Angle Scattering of X-rays*; John Wiley & Sons, Inc.: New York, 1955.
(2) Blanchet, C. E.; Svergun, D. I. Small-Angle X-Ray Scattering on Biological Macromolecules and Nanocomposites in Solution. *Annu. Rev. Phys. Chem.* **2013**, *64*, 37−54.
(3) Da Vela, S.; Svergun, D. I. Methods, development and applications of small-angle X-ray scattering to characterize biological macromolecules in solution. *Curr. Res. Struct. Biol.* **2020**, *2*, 164−170.
(4) Welborn, S. S.; Detsi, E. Small-angle X-ray scattering of nanoporous materials. *Nanoscale Horiz.* **2020**, *5*, 12−24.
(5) Gosecka, M.; Gosecki, M. Characterization methods of polymer core−shell particles. *Colloid Polym. Sci.* **2015**, *293*, 2719.
(6) Liu, D.; Song, K.; Chen, W.; Chen, J.; Sun, G.; Li, L. Review: Current progresses of small-angle neutron scattering on soft-matters investigation. *Nucl. Anal.* **2022**, *1*, No. 100011.
(7) Hyland, L. L.; Taraban, M. B.; Yu, Y. B. Using small-angle scattering techniques to understand mechanical properties of biopolymer-based biomaterials. *Soft Matter* **2013**, *9*, 10218−10228.
(8) Wei, Y.; Hore, M. J. A. Characterizing polymer structure with small-angle neutron scattering: A Tutorial. *J. Appl. Phys.* **2021**, *129*, No. 171101.
(9) Guilbaud, J.-B.; Saiani, A. Using small angle scattering (SAS) to structurally characterise peptide and protein self-assembled materials. *Chem. Soc. Rev.* **2011**, *40*, 1200−1210.
(10) Kikhney, A. G.; Svergun, D. I. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett.* **2015**, *589*, 2570−2577.
(11) Qian, S.; Sharma, V. K.; Clifton, L. A. Understanding the Structure and Dynamics of Complex Biomembrane Interactions by Neutron Scattering Techniques. *Langmuir* **2020**, *36*, 15189.
(12) Franken, L. E.; Boekema, E. J.; Stuart, M. C. A. Transmission Electron Microscopy as a Tool for the Characterization of Soft Materials: Application and Interpretation. *Adv. Sci.* **2017**, *4*, No. 1600476.
(13) Friedrich, H.; Frederik, P. M.; de With, G.; Sommerdijk, N. A. J. M. Imaging of Self-Assembled Structures: Interpretation of TEM and Cryo-TEM Images. *Angew. Chem., Int. Ed.* **2010**, *49*, 7850.
(14) Nguyen-Tri, P.; Ghassemi, P.; Carriere, P.; Nanda, S.; Assadi, A. A.; Nguyen, D. D. Recent applications of advanced atomic force microscopy in polymer science: A review. *Polymers* **2020**, *12*, 1142.
(15) Mourdikoudis, S.; Pallares, R. M.; Thanh, N. T. Characterization techniques for nanoparticles: comparison and complementarity upon studying nanoparticle properties. *Nanoscale* **2018**, *10*, 12871.
(16) Schryvers, D.; Cao, S.; Tirry, W.; Idrissi, H.; Van Aert, S. Advanced three-dimensional electron microscopy techniques in the quest for better structural and functional materials. *Sci. Technol. Adv. Mater.* **2013**, *14*, No. 014206.
(17) Nan, N.; Wang, J. FIB-SEM three-dimensional tomography for characterization of carbon-based materials. *Adv. Mater. Sci. Eng.* **2019**, *2019*, No. 8680715.
(18) Agbabiaka, A.; Wiltfong, M.; Park, C. Small Angle X-Ray Scattering Technique for the Particle Size Distribution of Nonporous Nanoparticles. *J. Nanopart.* **2013**, *2013*, No. 640436.
(19) Pauw, B. R. Everything SAXS:small-angle scattering pattern collection and correction. *J. Phys.: Condens. Matter* **2014**, *26*, No. 239501.
(20) Schaefer, D. W.; Agamalian, M. M. Ultra-small-angle neutron scattering: a new tool for materials research. *Curr. Opin. Solid State Mater. Sci.* **2004**, *8*, 39−47.
(21) Glatter, O.; Kratky, O.; Kratky, H.*Small Angle X-ray Scattering*; Academic Press, 1982.
(22) Rambo, R. P.; Tainer, J. A. Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. *Biopolymers* **2011**, *95*, 559−571.
(23) Feigin, L.; Svergun, D. I.*Structure Analysis by Small-Angle X-ray and Neutron Scattering*; Springer: 1987; Vol. *1*.
(24) Chacón, P.; Morán, F.; Díaz, J. F.; Pantos, E.; Andreu, J. M. Low-resolution structures of proteins in solution retrieved from X-ray scattering with a genetic algorithm. *Biophys. J.* **1998**, *74*, 2760−2775.
(25) Svergun, D. I. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.* **1999**, *76*, 2879−2886.
(26) Walther, D.; Cohen, F. E.; Doniach, S. Reconstruction of low-resolution three-dimensional density maps from one-dimensional small-angle X-ray solution scattering data for biomolecules. *J. Appl. Crystallogr.* **2000**, *33*, 350−363.
(27) Tóth, G. Simultaneous Monte Carlo Determination of Particle Size Distribution and Pair-Correlation Function of Spherical Colloids from a Diffraction Experiment. *Langmuir* **1999**, *15*, 6718−6723.
(28) Svergun, D. I.; Stuhrmann, H. New developments in direct shape determination from small-angle scattering. 1. Theory and model calculations. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1991**, *47*, 736−744.
(29) Svergun, D. I.; Petoukhov, M. V.; Koch, M. H. Determination of domain structure of proteins from X-ray solution scattering. *Biophys. J.* **2001**, *80*, 2946−2953.
(30) Volkov, V. V.; Svergun, D. I. Uniqueness of ab initio shape determination in small-angle scattering. *J. Appl. Crystallogr.* **2003**, *36*, 860−864.
(31) Schmidt, P. W.; Morozova, S.; Owens, P. M.; Adden, R.; Li, Y.; Bates, F. S.; Lodge, T. P. Molecular Weight Dependence of Methylcellulose Fibrillar Networks. *Macromolecules* **2018**, *51*, 7767−7775.

(32) Pedersen, J. S. Form factors of block copolymer micelles with spherical, ellipsoidal and cylindrical cores. *J. Appl. Crystallogr.* **2000**, *33*, 637−640.
(33) Pedersen, J. S. Analysis of small-angle scattering data from colloids and polymer solutions: modeling and least-squares fitting. *Adv. Colloid Interface Sci.* **1997**, *70*, 171.
(34) Doucet, M.; Cho, J. H.; Alina, G.; Bakker, J.; Bouwman, W.; Butler, P.; Campbell, K.; Gonzales, M.; Heenan, R.; Jackson, A.*SasView Version 4.1. Zenodo*. http://www.sasview.org, 2017.
(35) Beltran-Villegas, D. J.; Wessels, M. G.; Lee, J. Y.; Song, Y.; Wooley, K. L.; Pochan, D. J.; Jayaraman, A. Computational Reverse-Engineering Analysis for Scattering Experiments on Amphiphilic Block Polymer Solutions. *J. Am. Chem. Soc.* **2019**, *141*, 14916.
(36) Wessels, M. G.; Jayaraman, A. Machine Learning Enhanced Computational Reverse Engineering Analysis for Scattering Experiments (CREASE) to Determine Structures in Amphiphilic Polymer Solutions. *ACS Polym. Au* **2021**, *1*, 153−164.
(37) Wessels, M. G.; Jayaraman, A. Computational Reverse-Engineering Analysis of Scattering Experiments (CREASE) on Amphiphilic Block Polymer Solutions: Cylindrical and Fibrillar Assembly. *Macromolecules* **2021**, *54*, 783.
(38) Ye, Z.; Wu, Z.; Jayaraman, A. Computational Reverse Engineering Analysis for Scattering Experiments (CREASE) on Vesicles Assembled from Amphiphilic Macromolecular Solutions. *JACS Au* **2021**, *1*, 1925.
(39) Heil, C. M.; Jayaraman, A. Computational Reverse-Engineering Analysis for Scattering Experiments of Assembled Binary Mixture of Nanoparticles. *ACS Mater. Au* **2021**, *1*, 140.
(40) Heil, C. M.; Patil, A.; Dhinojwala, A.; Jayaraman, A. Computational Reverse-Engineering Analysis for Scattering Experiments (CREASE) with Machine Learning Enhancement to Determine Structure of Nanoparticle Mixtures and Solutions. *ACS Cent. Sci.* **2022**, *8*, 996−1007.
(41) Nasatto, L. P.; Pignon, F.; Silveira, L. M. J.; Duarte, E. R. M.; Noseda, D. M.; Rinaudo, M. Methylcellulose, a Cellulose Derivative with Original Physical Properties and Extended Applications. *Polymers* **2015**, *7*, 777−803.
(42) Zhang, H.; Xie, Y.; Tang, Y.; Ni, S.; Wang, B.; Chen, Z.; Liu, X. Development and characterization of thermo-sensitive films containing asiaticoside based on polyvinyl alcohol and Methylcellulose. *J. Drug Delivery Sci. Technol.* **2015**, *30*, 133−145.
(43) Chrai, S. S.; Robinson, J. R. Ocular evaluation of methylcellulose vehicle in albino rabbits. *J. Pharm. Sci.* **1974**, *63*, 1218−1223.
(44) Wollenweber, C.; Makievski, A. V.; Miller, R.; Daniels, R. Adsorption of hydroxypropyl methylcellulose at the liquid/liquid interface and the effect on emulsion stability. *Colloids Surf., A* **2000**, *172*, 91−101.
(45) Tunç, S.; Duman, O. Preparation and characterization of biodegradable methyl cellulose/montmorillonite nanocomposite films. *Appl. Clay Sci.* **2010**, *48*, 414−424.
(46) Schmidt, P. W.; Morozova, S.; Ertem, S. P.; Coughlin, M. L.; Davidovich, I.; Talmon, Y.; Reineke, T. M.; Bates, F. S.; Lodge, T. P. Internal Structure of Methylcellulose Fibrils. *Macromolecules* **2020**, *53*, 398−405.
(47) Coughlin, M. L.; Liberman, L.; Ertem, S. P.; Edmund, J.; Bates, F. S.; Lodge, T. P. Methyl cellulose solutions and gels: fibril formation and gelation properties. *Prog. Polym. Sci.* **2021**, *112*, No. 101324.
(48) Hirrien, M.; Desbrières, J.; Rinaudo, M. Physical properties of methylcelluloses in relation with the conditions for cellulose modification. *Carbohydr. Polym.* **1996**, *31*, 243−252.
(49) Kobayashi, K.; Huang, C.; Lodge, T. P. Thermoreversible Gelation of Aqueous Methylcellulose Solutions. *Macromolecules* **1999**, *32*, 7070.
(50) Haque, A.; Morris, E. R. Thermogelation of Methylcellulose. Part I: Molecular Structures and Processes. *Carbohydr. Polym.* **1993**, *22*, 161.
(51) Sarkar, N. Thermal Gelation Properties of Methyl and Hydroxypropyl Methylcellulose. *J. Appl. Polym. Sci.* **1979**, *24*, 1073.
(52) Sarkar, N.; Walker, L. C. Hydration Dehydration Properties of Methylcellulose and Hydroxypropylmethylcellulose. *Carbohydr. Polym.* **1995**, *27*, 177.
(53) Sarkar, N. Kinetics of Thermal Gelation of Methylcellulose and Hydroxypropylmethylcellulose in Aqueous-Solutions. *Carbohydr. Polym.* **1995**, *26*, 195.
(54) Chevillard, C.; Axelos, M. A. V. Phase Separation of Aqueous Solution of Methylcellulose. *Colloid Polym. Sci.* **1997**, *275*, 537.
(55) Nishinari, K.; Hofmann, K. E.; Moritaka, H.; Kohyama, K.; Nishinari, N. Gel-Sol Transition of Methylcellulose. *Macromol. Chem. Phys.* **1997**, *198*, 1217.
(56) Desbrières, J.; Hirrien, M.; Rinaudo, M. A calorimetric study of methylcellulose gelation. *Carbohydr. Polym.* **1998**, *37*, 145−152.
(57) Desbrières, J.; Hirrien, M.; Ross-Murphy, S. B. Thermogelation of methylcellulose: rheological considerations. *Polymer* **2000**, *41*, 2451−2461.
(58) Li, L.; Thangamathesvaran, P. M.; Yue, C. Y.; Tam, K. C.; Hu, X.; Lam, Y. C. Gel Network Structure of Methylcellulose in Water. *Langmuir* **2001**, *17*, 8062.
(59) Li, L. Thermal Gelation of Methylcellulose in Water: Scaling and Thermoreversibility. *Macromolecules* **2002**, *35*, 5990.
(60) Funami, T.; Kataoka, Y.; Hiroe, M.; Asai, I.; Takahashi, R.; Nisbinari, K. Thermal Aggregation of Methylcellulose with Different Molecular Weights. *Food Hydrocolloids* **2007**, *21*, 46.
(61) Bodvik, R.; Dedinaite, A.; Karlson, L.; Bergström, M.; Bäverbäck, P.; Pedersen, J. S.; Edwards, K.; Karlsson, G.; Varga, I.; Claesson, P. M. Aggregation and network formation of aqueous methylcellulose and hydroxypropylmethylcellulose solutions. *Colloids Surf., A* **2010**, *354*, 162−171.
(62) Fairclough, J. P. A.; Yu, H.; Kelly, O.; Ryan, A. J.; Sammler, R. L.; Radler, M. Interplay between Gelation and Phase Separation in Aqueous Solutions of Methylcellulose and Hydroxypropylmethylcellulose. *Langmuir* **2012**, *28*, 10551.
(63) Lott, J. R.; McAllister, J. W.; Arvidson, S. A.; Bates, F. S.; Lodge, T. P. Fibrillar Structure of Methylcellulose Hydrogels. *Biomacromolecules* **2013**, *14*, 2484.
(64) Arvidson, S. A.; Lott, J. R.; McAllister, J. W.; Zhang, J.; Bates, F. S.; Lodge, T. P.; Sammler, R. L.; Li, Y.; Brackhagen, M. Interplay of Phase Separation and Thermoreversible Gelation in Aqueous Methylcellulose Solutions. *Macromolecules* **2013**, *46*, 300.
(65) McAllister, J. W.; Schmidt, P. W.; Dorfman, K. D.; Lodge, T. P.; Bates, F. S. Thermodynamics of Aqueous Methylcellulose Solutions. *Macromolecules* **2015**, *48*, 7205.
(66) McAllister, J. W.; Lott, J. R.; Schmidt, P. W.; Sammler, R. L.; Bates, F. S.; Lodge, T. P. Linear and Nonlinear Rheological Behavior of Fibrillar Methylcellulose Hydrogels. *ACS Macro Lett.* **2015**, *4*, 538.
(67) Kato, T.; Yokoyama, M.; Takahashi, A. Melting Temperatures of Thermally Reversible Gels IV. Methyl Cellulose−Water Gels. *Colloid Polym. Sci.* **1978**, *256*, 15.
(68) Kobayashi, M.; Yoshioka, T.; Kozasa, T.; Tashiro, K.; Suzuki, J.; Funahashi, S.; Izumi, Y. Structure of Physical Gels Formed in Syndiotactic Polystyrene Solvent Systems Studied by Small-Angle Neutron-Scattering. *Macromolecules* **1994**, *27*, 1349.
(69) Huang, W.; Dalal, I. S.; Larson, R. G. Analysis of Solvation and Gelation Behavior of Methylcellulose Using Atomistic Molecular Dynamics Simulations. *J. Phys. Chem. B* **2014**, *118*, 13992−14008.
(70) Huang, W.; Ramesh, R.; Jha, P. K.; Larson, R. G. A Systematic Coarse-Grained Model for Methylcellulose Polymers: Spontaneous Ring Formation at Elevated Temperature. *Macromolecules* **2016**, *49*, 1490−1503.
(71) Ginzburg, V. V.; Sammler, R. L.; Huang, W.; Larson, R. G. Anisotropic Self-Assembly and Gelation in Aqueous Methylcellulose—Theory and Modeling. *J. Polym. Sci., Part B: Polym. Phys.* **2016**, *54*, 1624.
(72) Yang, Y.; Wu, W.; Liu, H.; Xu, H.; Zhong, Y.; Zhang, L.; Chen, Z.; Sui, X.; Mao, Z. Aggregation behaviors of thermo-responsive methylcellulose in water: A molecular dynamics simulation study. *J. Mol. Graphics Modell.* **2020**, *97*, No. 107554.

(73) Sethuraman, V.; Dorfman, K. D. Simulating precursor steps for fibril formation in methylcellulose solutions. *Phys. Rev. Mater.* **2019**, *3*, No. 055601.
(74) Pedersen, J. S.; Schurtenberger, P. Scattering Functions of Semiflexible Polymers with and without Excluded Volume Effects. *Macromolecules* **1996**, *29*, 7602.
(75) Wessels, M. G.; Jayaraman, A. Machine learning enhanced computational reverse engineering analysis for scattering experiments (crease) to determine structures in amphiphilic polymer solutions. *ACS Polym. Au* **2021**, *1*, 153.
(76) Mitchell, M.*An Introduction to Genetic Algorithms*; MIT Press, 1998.
(77) Ye, Z.; Wu, Z.; Jayaraman, A. Computational Reverse Engineering Analysis for Scattering Experiments (CREASE) on Vesicles Assembled from Amphiphilic Macromolecular Solutions. *JACS Au* **2021**, *1*, 1925−1936.
(78) Wignall, G. D. Instrumental resolution effects in small-angle scattering. *J. Appl. Crystallogr.* **1991**, *24*, 479−484.
(79) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; Kudlur, M.; Levenberg, J.; Monga, R.; Moore, S.; Murray, D. G.; Steiner, B.; Tucker, P.; Vasudevan, V.; Warden, P.; Wicke, M.; Yu, Y.; Zheng, X. In *TensorFlow: A System for Large-Scale Machine Learning*, Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation; USENIX Association: Savannah, GA, USA, 2016; 265−283.
(80) Schneidman-Duhovny, D.; Kim, S. J.; Sali, A. Integrative structural modeling with small angle X-ray scattering profiles. *BMC Struct. Biol.* **2012**, *12*, 17.
(81) Caviness (caviness.hpc.udel.edu). https://www.hpc.udel.edu/systems/caviness (accessed August 26, 2022).
(82) Pedersen, J. S.; Schurtenberger, P. Scattering Functions of Semiflexible Polymers with and without Excluded Volume Effects. *Macromolecules* **1996**, *29*, 7602−7612.
(83) Chen, W.-R.; Butler, P. D.; Magid, L. J. Incorporating Intermicellar Interactions in the Fitting of SANS Data from Cationic Wormlike Micelles. *Langmuir* **2006**, *22*, 6539−6548.
(84) Kotlarchyk, M.; Chen, S. H. Analysis of small angle neutron scattering spectra from polydisperse interacting colloids. *J. Chem. Phys.* **1983**, *79*, 2461−2469.