Electronic Supplementary Information for:

Computational Reverse-Engineering Analysis for Scattering Experiments for Form Factor and Structure Factor Determination ('P(q) and S(q) CREASE')†

Christian M. Heil,[1] Yingzhen Ma,[2] Bhuvnesh Bharti,[2] and Arthi Jayaraman[1,3,*]

[1]Department of Chemical and Biomolecular Engineering, University of Delaware, 150 Academy St., Newark, DE 19716. USA.

[2]Cain Department of Chemical Engineering, Louisiana State University, 3307 Patrick F. Taylor Hall, Baton Rouge, LA 70803. USA

[3]Department of Materials Science and Engineering, University of Delaware, 201 DuPont Hall, Newark, DE 19716. USA.

*Corresponding author arthij@udel.edu

†Electronic supplementary information (ESI) available.

I. Methods:

A. Details of the machine learning (ML) model development

The direct evaluation of the computed I(q) profile of concentrated macromolecule solutions using the Debye equation is computationally intensive requiring high performance computing resources and significant time to complete. We implement machine learning (ML) models to directly link the low dimensional description of the structure using a set of genes to its corresponding computed scattering as our previous work[1] showed ~95% reduction in CREASE analysis time using such ML models. We seek to make the ML approach generic and as widely applicable to many different macromolecule solutions by taking advantage of the decomposition of I(q) into P(q) times S(q).[2]

$$I(q) \propto P(q) * S(q) \tag{S1}$$

Thus, instead of using a single ML model to link an individual's gene set to the I(q) scattering profiles, we develop separate ML models for the P(q) and S(q) prediction. This approach allows a user to readily swap out/in the P(q) ML model while retaining the S(q) ML model as long as the macromolecule is spherical and is disordered (*e.g.*, the macromolecule cannot pack into regularly ordered or anisotropic structures). As the S(q) just describes the spatial arrangement of the macromolecules, it is more widely applicable. The P(q) ML model must be tailored for the macromolecule of interest (*e.g.*, micelle, vesicle, polymer coated nanoparticles), and it should be trained to predict all relevant scattering profiles (*e.g.*, for a micelle P(q) model, $I_{micelle}$(q), I_{shell}(q), and I_{core}(q)).

For this work, we utilize artificial neural networks (ANNs) as the ML model because ANNs are more accessible to experimentalists with many common ML software packages including ANNs. An ANN is comprised of an input layer, output layer, and at least one intermediate or hidden layer with each layer potentially consisting of a different number of fully connected nodes.[3] The individual nodes' biases and weights used by their activation function (we use Rectified Linear Unit, ReLU) are optimized (we use the Adam optimization algorithm) to minimize the set loss function (we use mean square error of the training data and ANN prediction). For the concentrated micelle system for which this ML model was developed, we need to specify the architecture of the P(q) and S(q) models.

For the micelle P(q) model, the output layer was 3 nodes corresponding to the negative $\log_{10}$ of the $P_{micelle}(q)$, $P_{shell}(q)$, and $P_{core}(q)$ at the input q value (negative $\log_{10}$ used to account for logarithmic nature of the form factor profile). The input layer was 3 nodes corresponding to the micelle diameter, micelle dispersity, and core-micelle size ratio. Similar to previous work,[1] we non-dimensionalize the micelle diameter (D) with the input q value as $\log_{10} qD/2\pi$. This non-dimensionalization was shown to allow a similarly trained ANN to predict for diameters other than ones in the training set because the ANN instead was able to predict for $\log_{10} qD/2\pi$. To determine the ideal ANN architecture for the hidden layer(s) and nodes, we explore a variety of number of hidden layers (1-3) and number of nodes per hidden layer (32, 64, 128, 256, 512) and determine which combination provides the lowest validation error. The training data we used is obtained from the Debye equation evaluation of random scatterers placed in the micelles (as mentioned in the manuscript Methods section or detailed in previous work[4]). We utilized 45,000 different combinations of micelle diameter, micelle dispersity, and core-micelle size ratio; though this was due to the plethora of available data and does not represent the minimum training set. During the training of the P(q) ANN, we split the data into 70% for ANN training and 30% for validation. We found that a single hidden layer with 512 nodes provided the minimum validation loss.

For the S(q) model, the input layer was 6 nodes corresponding to the total diameter, diameter dispersity, concentration, and genes related to spatial organization as described previously.[1] As before, we non-dimensionalize the diameter (D) with the input q value as $\log_{10} qD/2\pi$. The output layer was 1 node corresponding to the predicted S(q) which has to be unnormalized by the training data average S(q) and S(q) standard deviation at that q point. Unlike the P(q) or I(q) which are easily normalized using a logarithm, the S(q) ANN model was best trained when normalized by the average S(q) and S(q) standard deviation from the training data set. To determine the ideal ANN architecture for the hidden layer(s) and nodes, we explore a variety of number of hidden layers (1-3) and number of nodes per hidden layer (32, 64, 128, 256, 512) and determine which combination provides the lowest validation error. The training data we used is obtained from the Debye equation evaluation of spheres placed in space (as mentioned in the manuscript Methods section or detailed in previous work[4]). We utilized 45,000 different combinations of diameter, size dispersity, concentration, and genes related to the spatial arrangement of the spheres. Similar to before, this was due to the plethora of available data and does not represent the minimum training set. During the training of the S(q) ANN, we split the data into 70% for ANN training and 30% for validation. We found that a single hidden layer with 512 nodes provided the minimum validation loss.

Similar to previous work,[1] we utilize the ML models in a multi-step approach to first optimize primarily the P(q) (on a smaller q range with q greater than ~1.5x the micelle diameter), then primarily the S(q) holding the P(q) constant (over the entire q range but setting micelle diameter, micelle dispersity, and core-micelle ratio from the initial run), and finally optimize both together to fine-tune the result (allowing converged genes to vary ~10%). We note that a user could instead optimize both P(q) and S(q) in one run of CREASE especially if *a priori* knowledge of the system allows for smaller limits for the genes (to facilitate easier optimization). The multi-step approach allows each run of CREASE to have fewer genes/parameters to optimize and converge. Regardless of whether a user optimizes the gene set over a single run or with a multi-step approach, the converged 'best' set of genes are converted into the 3D structure to obtain the structural information (RDF, domain sizes, aggregation number).

B. Points to consider before using ‘P(q) and S(q) CREASE’

In this section we provide a list of general guidelines for prospective users to consider before they begin to apply/adapt ‘P(q) and S(q) CREASE’ to their specific systems of interest.

- As of this paper, ‘P(q) and S(q) CREASE’ has been developed for analysis of systems that are comprised of spherical ‘primary particles’
    - Past work on CREASE handling P(q) (*e.g.*, amphiphilic polymer solutions at dilute conditions) have tackled fibrils (which are not spherical); see References[5-7]
    - Ongoing development for ‘P(q) and S(q) CREASE’ is focused on extending the method for anisotropic ‘primary particles’; the reader is encouraged to look out for future publications.
- CREASE in general has been designed to handle structures that are amorphous and in soft materials (*e.g.*, non-crystalline soft materials)
    - Consider alternative, previously developed analysis techniques for interpreting ordered superlattices comprised of any arbitrarily shaped particles[8, 9] or proteins with ordered secondary structures.[10-12]
- Step 1: User should use knowledge about the system at hand (e.g., underlying physics, chemistry) to create genes that will store information about the ‘primary particles’

    - Genes must be designed to hold all pertinent information that the user cares about (*e.g.,* the dimensions or ratios of dimensions or packing densities) and that CREASE will require to generate the structure and calculate the P(q)
    - For example, in the core-shell CREASE version in this work with independent dispersity in the core and shell sizes, the genes store the core size, the core size dispersity, the shell thickness, the shell thickness dispersity,
    - Questions to consider that can impact the number of genes to create:
        - How many layers are in the 'primary particle'? (This core-shell work had one core and one shell layer, but there could be particles with multiple shells)
        - Which layers, if any, are chemically similar? (this will be important for calculating the P(q) for the 'primary particle')
        - Are the dimensions (thickness, dispersity in thickness) of different layers related or distinct? (An example from this core-shell work is that one version we tested assumed the shell thickness is a constant fraction of the core size.)
        - Is the 'primary particle' 'hard' (no overlap between neighboring particles) or 'soft' (some overlap allowed between neighboring particles)?
- Step 2: User should use knowledge about the system at hand (e.g., underlying physics, chemistry) to adapt CREASE for generating the 3D structure of their 'primary particles'
    - In the machine learning version of CREASE one only creates the structure towards the end of the genetic algorithm (GA) cycle after the optimization has converged. In the non-machine learning version of CREASE in each GA step for every

individual, the code needs to generates an explicit structure (x,y,z positions) for all 'primary particles'. Our version of CREASE utilizes coarse-grained molecular dynamics simulations using the LAMMPS package[13] to generate the explicit 3D structure from the genes; others may consider stochastic approaches to do the same. Regardless of deterministic/stochastic approaches one uses, the user must write the code inserted into CREASE be able to convert the relevant genes (relating to size and size dispersity) into a total particle size that will then be used in the 3D structure creation simulation.

- To generate a 3D structure, the user must define the total 'primary particle' size based solely on the values of the genes for that specific individual (the gene values will change during the CREASE run, so this conversion cannot be static) and a hypothesis about the nature of the 'primary particle'.
    - For a hard 'primary particle', the total particle size would be based on the overall size of the 'primary particle'. For an example of a hard core-shell particle, the total particle diameter would be the summation of the core diameter (a gene) and twice the shell thickness (a gene).
    - For a soft 'primary particle' that allows overlap between neighboring 'primary particles', the total particle size would be based on both the overall size of the 'primary particle' and the fraction of overlap allowed. For an example of a soft core-shell particle, the total particle diameter would be the summation of the

core diameter (a gene) and twice the shell thickness (a gene) times the fraction of overlap allowed (a gene).

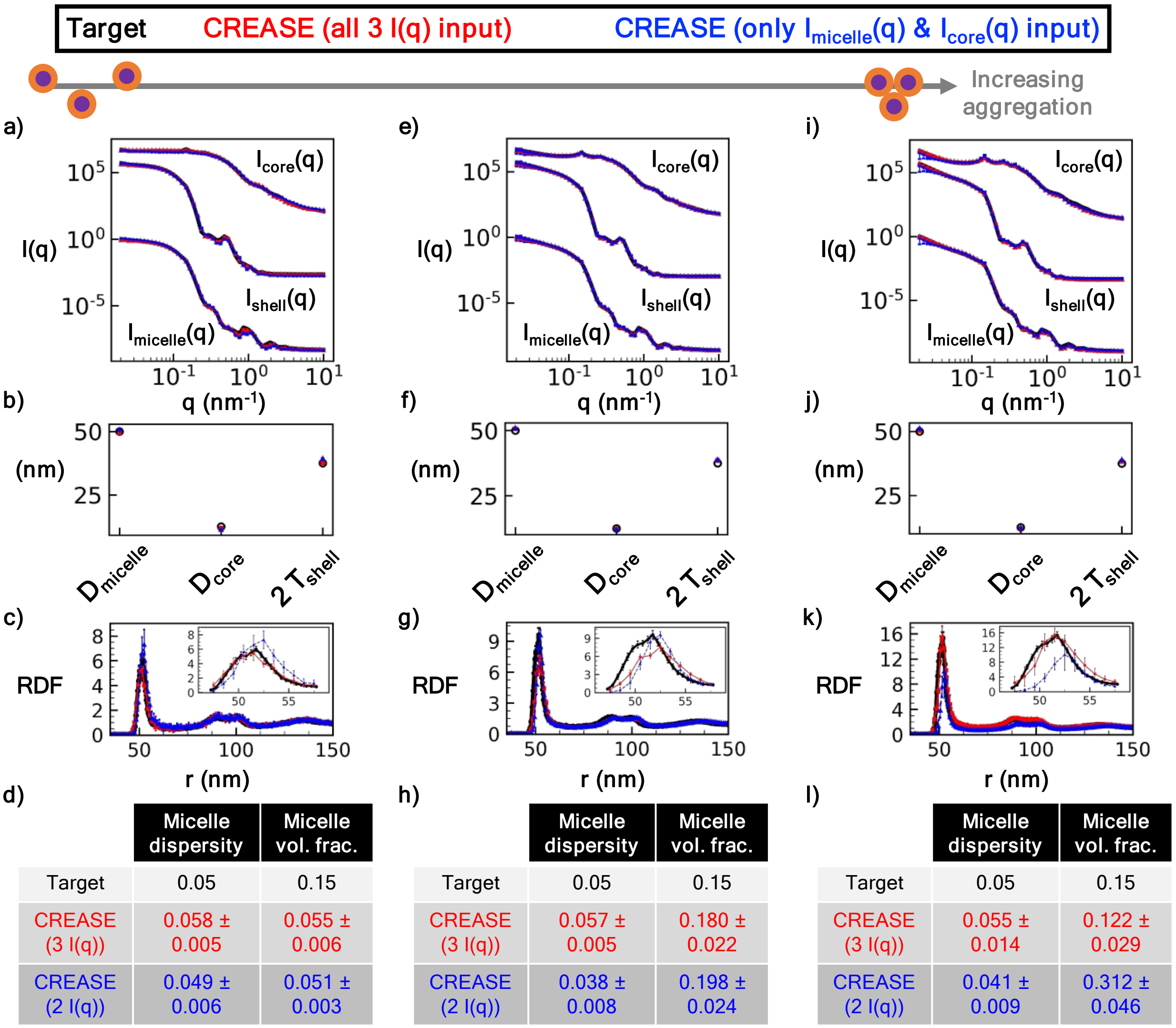
- Step 3: After generating the 3D structure of 'primary particles', CREASE must calculate the I(q) for that structure. The S(q) part of the calculation does not require modification from Reference[1] (for systems conforming to the assumptions listed at the beginning of this set of guidelines), only the P(q) calculation (specifically the F(q)) requires some modification as described in Step 4.
- Step 4: User should also use knowledge about the system at hand (e.g., underlying physics, chemistry) to develop idea(s) for the 'primary particle' form factor F(q) calculation:
    - The genes that were developed to store information about the 'primary particle' must be utilized to direct CREASE to calculate the scattering from the 'primary particle'
    - CREASE determines the 'primary particle' scattering by placing scatterers throughout the 'primary particle' and labelling each scatterer's 'identity' based on which domain within the 'primary particle' the scatterer resides in (x,y,z position). The scatterer's 'identity' enables the domain's SLD to be applied. References[4, 5, 7, 14] provide detailed explanation of placing scatterers and calculating the 'primary particle' scattering.
- CREASE now has been adapted to use the gene-based representation of the 'primary particle' to generate the explicit 3D structure and numerically calculate the I(q) accounting for both the S(q) and P(q). At this point, CREASE can now optimize, using its inbuilt genetic algorithm, for what values of genes gives a structure and 'primary particle' dimensions with the calculated I(q) that best matches the input experimental I(q).

II. Validation of P(q) and S(q) CREASE method on *in silico* concentrated micelle solutions

We validate our P(q) and S(q) CREASE method against computationally generated concentrated micelle solutions with various micelle size dispersity, core:micelle ratios, micelle volume fraction, and degree of aggregation. **ESI Table S1** provides the various parameters examined, and **ESI Figure S1-S12** provide the validation of P(q) and S(q) CREASE on those systems.

Table S1*: Design space that the P(q) and S(q) CREASE method is validated across*

	Values considered
Micelle size dispersity (lognormal)	0.05 & 0.15
Core-micelle size ratio	0.25, 0.50, & 0.75
Micelle volume fraction	0.15 & 0.40
Degree of aggregation	Varying contact peaks in micelle center-center radial distribution functions (RDF)



d)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.15
CREASE (3 I(q))	0.058 ± 0.005	0.055 ± 0.006
CREASE (2 I(q))	0.049 ± 0.006	0.051 ± 0.003

h)

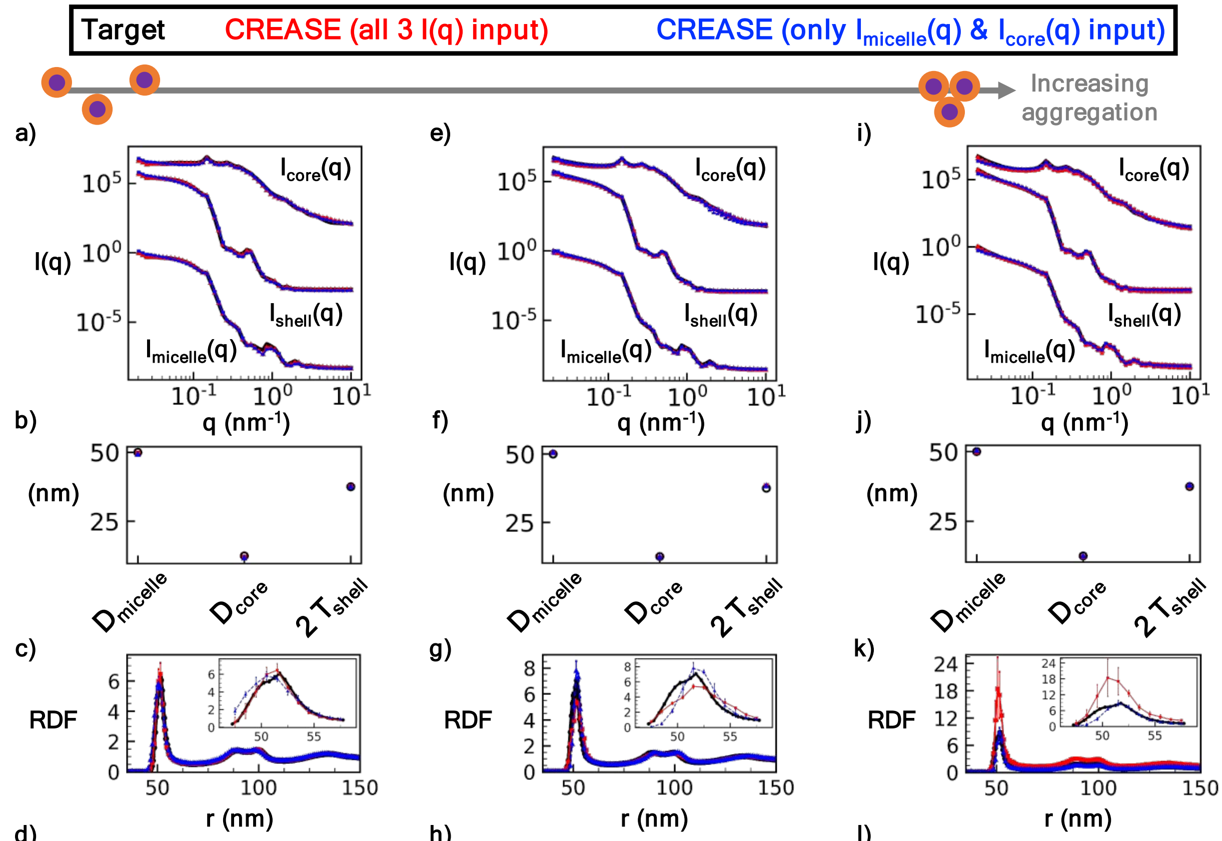
	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.15
CREASE (3 I(q))	0.057 ± 0.005	0.180 ± 0.022
CREASE (2 I(q))	0.038 ± 0.008	0.198 ± 0.024

l)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.15
CREASE (3 I(q))	0.055 ± 0.014	0.122 ± 0.029
CREASE (2 I(q))	0.041 ± 0.009	0.312 ± 0.046

***Figure S1**: P(q) and S(q) CREASE applied to a concentrated micelle solution with a 50 nm average diameter, 0.05 micelle size dispersity, 0.25 core-micelle size ratio, and 0.15 micelle volume fraction to simultaneously identify the micellular information and the structural arrangement. a) scattering intensity, I(q), for the target structure (black) with disperse micelle arrangement, CREASE with all three I(q) used as inputs (red), and CREASE with only the $I_{micelle}$(q) and I_{core}(q) used as inputs (blue). While the blue case only receives two I(q) curves as inputs, we calculate the I_{shell}(q) from the output structure for comparison. b) micelle diameter ($D_{micelle}$), core diameter (D_{core}), and twice the shell thickness (2 T_{shell}) for the target and CREASE outputs. c)*

micelle structural arrangement is quantified using the radial distribution function (RDF) comparing the target and CREASE outputs. The inset image provides a magnification of the primary RDF peak. d) additional micellular information on the micelle size dispersity and micelle volume fraction that the target possesses and the CREASE methods converge to. e-h) are the same as a-d) with increasing aggregation to a weakly aggregating target system. i-l) are the same as a-d) with increasing aggregation to a strongly aggregating target system. The error bars are the standard deviation of the average of 3 independent runs of the P(q) and S(q) CREASE method.



d)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.40
CREASE (3 I(q))	0.044 ± 0.007	0.350 ± 0.005
CREASE (2 I(q))	0.058 ± 0.006	0.339 ± 0.012

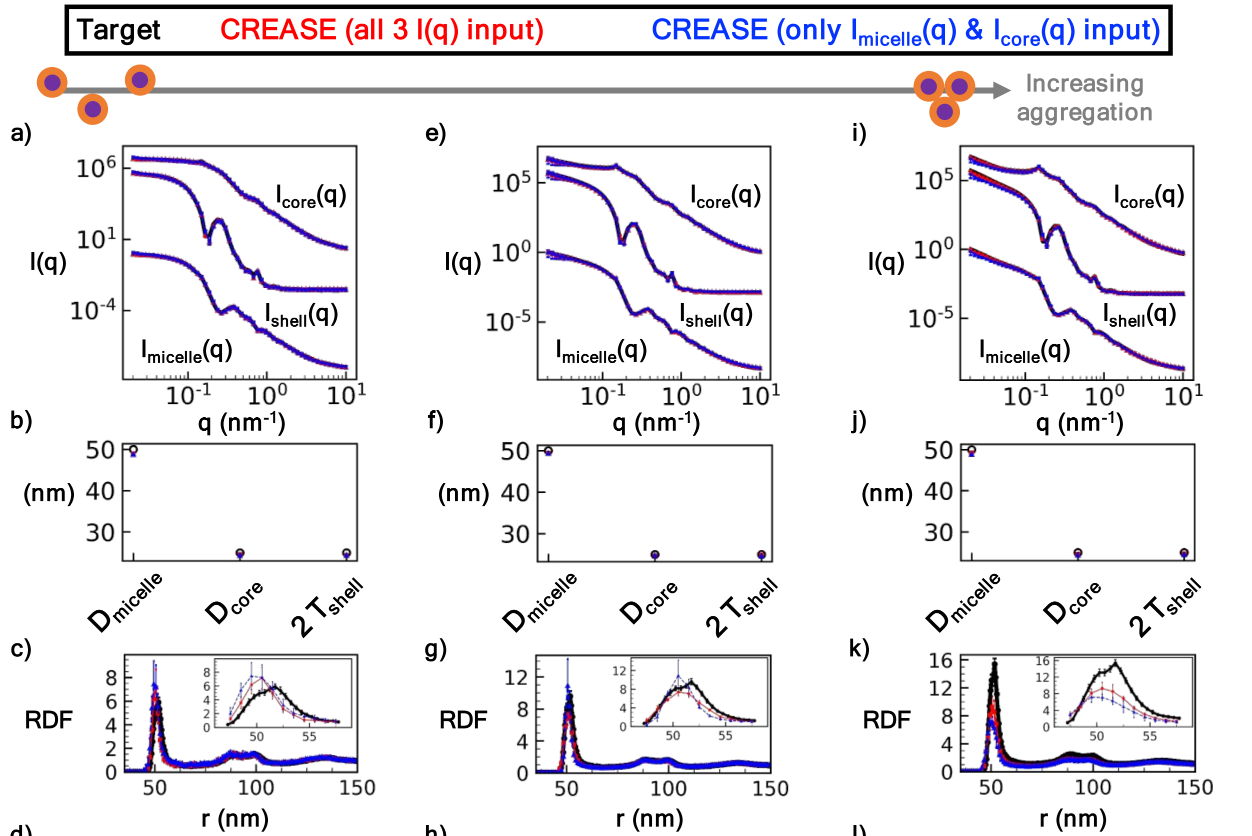
h)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.40
CREASE (3 I(q))	0.066 ± 0.001	0.442 ± 0.016
CREASE (2 I(q))	0.040 ± 0.006	0.431 ± 0.039

l)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.40
CREASE (3 I(q))	0.044 ± 0.014	0.174 ± 0.063
CREASE (2 I(q))	0.038 ± 0.005	0.383 ± 0.014

Figure S2*: Same as Figure S1 with a 50 nm average diameter, 0.05 micelle size dispersity, 0.25 core:micelle size ratio, and* ***0.40 micelle volume fraction****.*



d)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.15
CREASE (3 I(q))	0.040 ± 0.009	0.055 ± 0.001
CREASE (2 I(q))	0.047 ± 0.015	0.060 ± 0.010

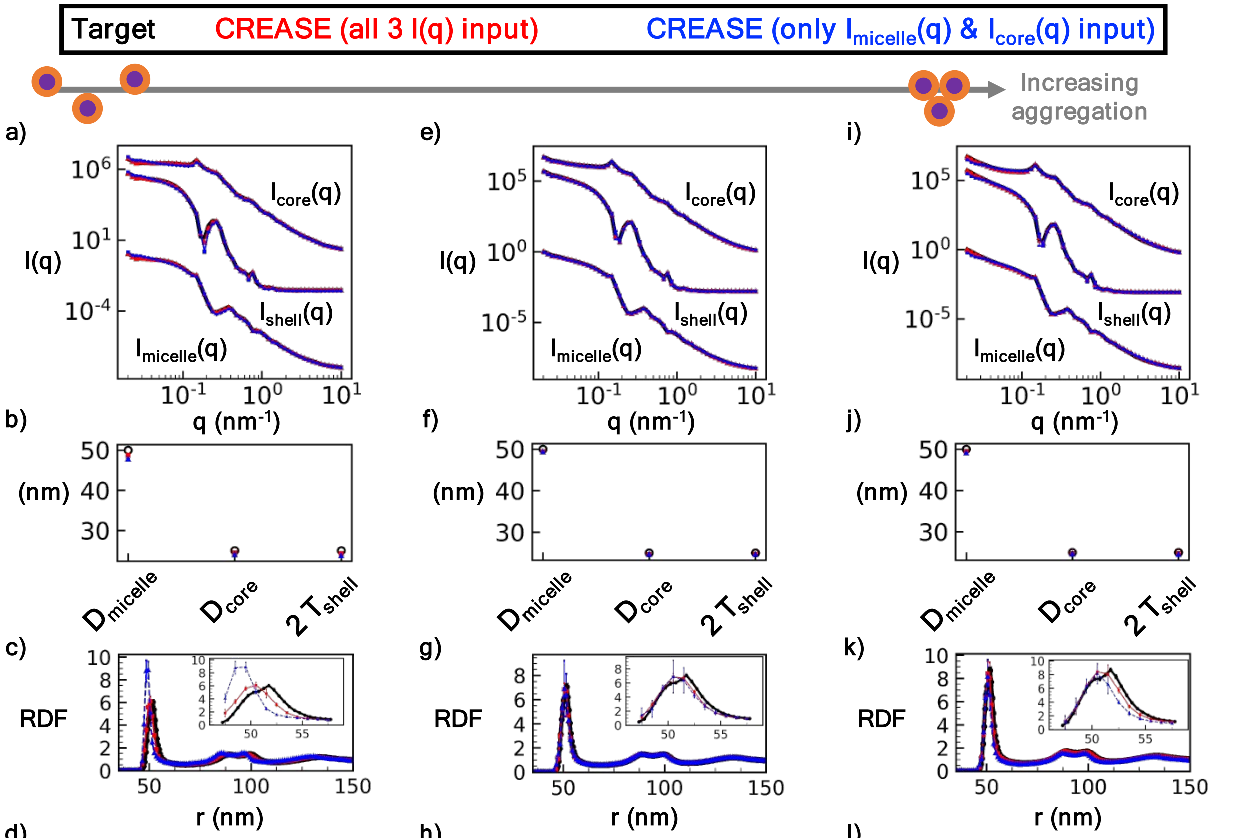
h)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.15
CREASE (3 I(q))	0.051 ± 0.006	0.244 ± 0.004
CREASE (2 I(q))	0.032 ± 0.006	0.217 ± 0.024

l)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.15
CREASE (3 I(q))	0.056 ± 0.008	0.245 ± 0.035
CREASE (2 I(q))	0.061 ± 0.007	0.332 ± 0.051

Figure S3*: Same as Figure S1 with a 50 nm average diameter, 0.05 micelle size dispersity,* ***0.50 core:micelle size ratio****, and 0.15 micelle volume fraction.*



d)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.40
CREASE (3 I(q))	0.048 ± 0.004	0.372 ± 0.031
CREASE (2 I(q))	0.032 ± 0.005	0.330 ± 0.008

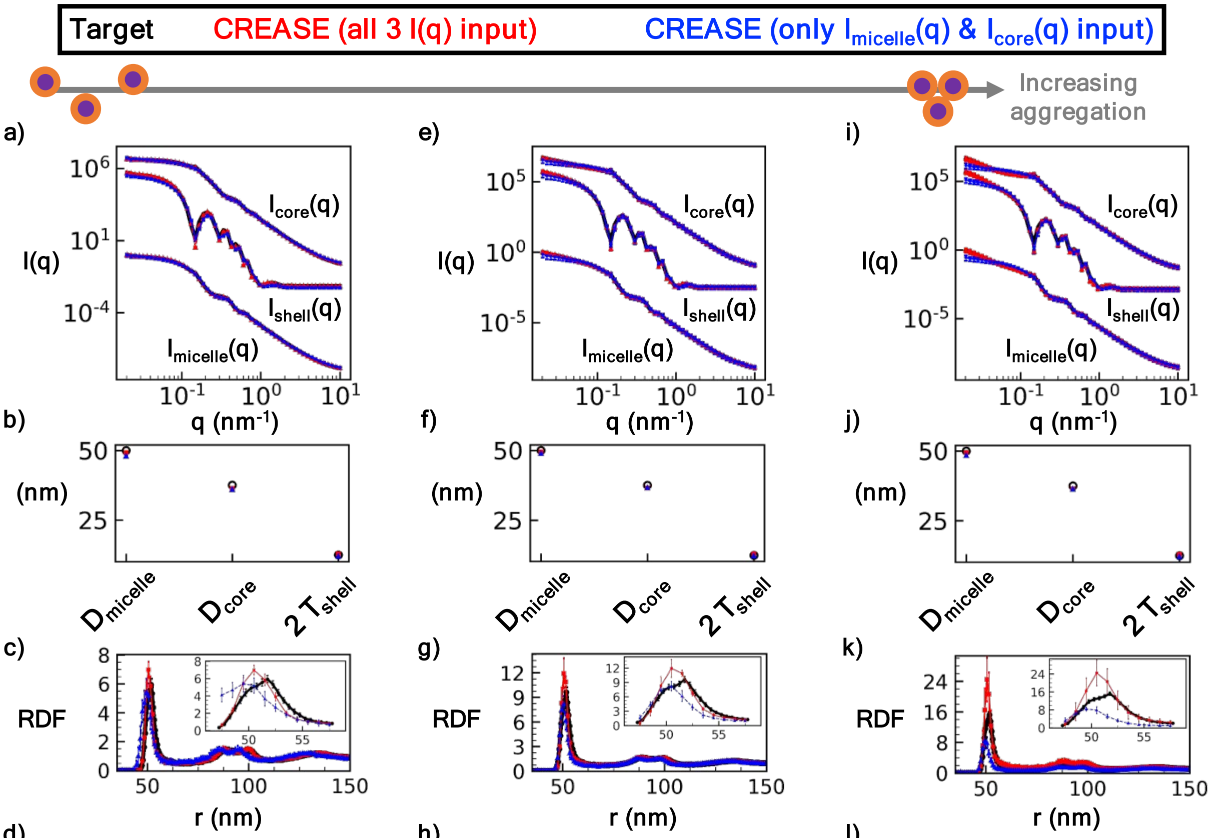
h)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.40
CREASE (3 I(q))	0.046 ± 0.004	0.424 ± 0.009
CREASE (2 I(q))	0.050 ± 0.017	0.443 ± 0.013

l)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.40
CREASE (3 I(q))	0.047 ± 0.009	0.438 ± 0.001
CREASE (2 I(q))	0.043 ± 0.013	0.423 ± 0.001

***Figure S4**: Same as Figure S1 with a 50 nm average diameter, 0.05 micelle size dispersity, **0.50 core:micelle size ratio**, and **0.40 micelle volume fraction**.*



d)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.15
CREASE (3 I(q))	0.038 ± 0.002	0.081 ± 0.003
CREASE (2 I(q))	0.062 ± 0.009	0.071 ± 0.016

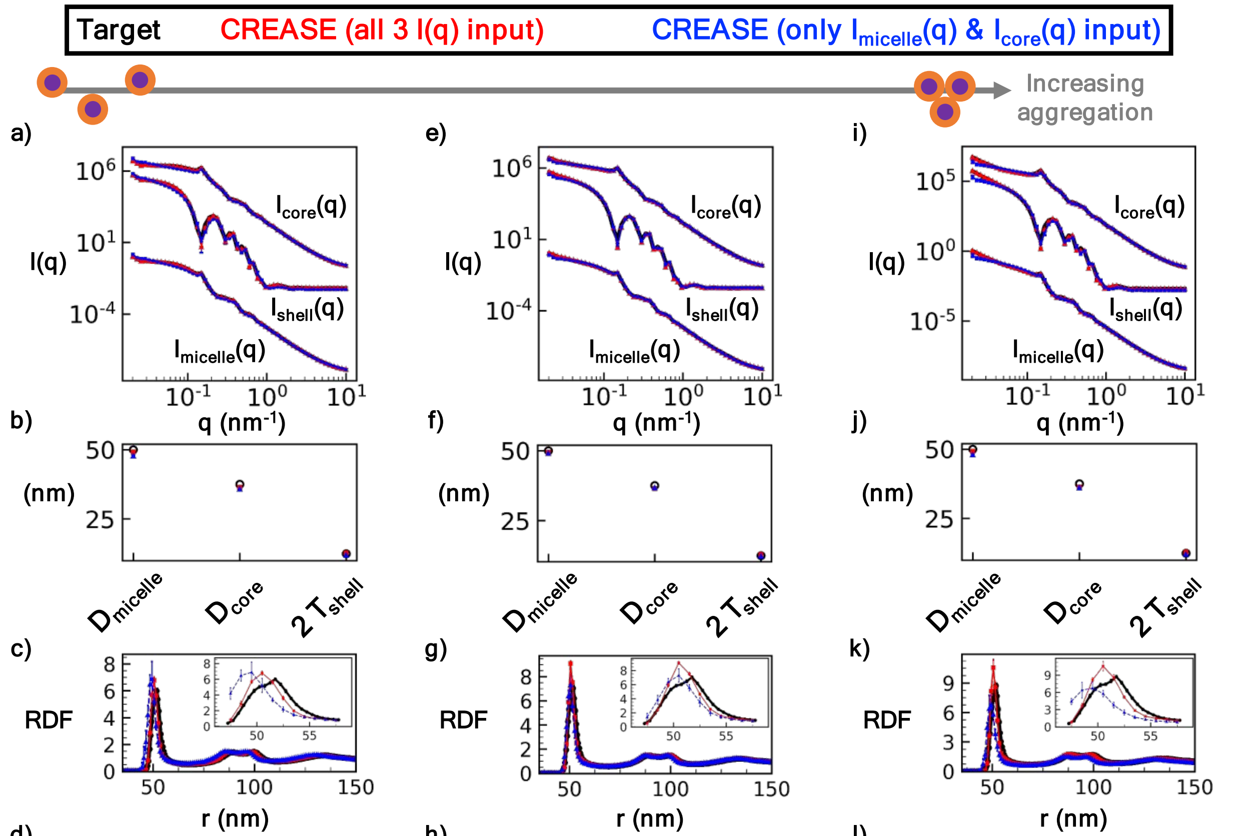
h)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.15
CREASE (3 I(q))	0.036 ± 0.001	0.144 ± 0.033
CREASE (2 I(q))	0.043 ± 0.004	0.135 ± 0.041

l)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.15
CREASE (3 I(q))	0.0371 ± 0.004	0.112 ± 0.038
CREASE (2 I(q))	0.045 ± 0.005	0.107 ± 0.034

***Figure S5**: Same as Figure S1 with a 50 nm average diameter, 0.05 micelle size dispersity, **0.75 core:micelle size ratio**, and 0.15 micelle volume fraction.*



d)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.40
CREASE (3 I(q))	0.040 ± 0.001	0.354 ± 0.018
CREASE (2 I(q))	0.053 ± 0.013	0.364 ± 0.013

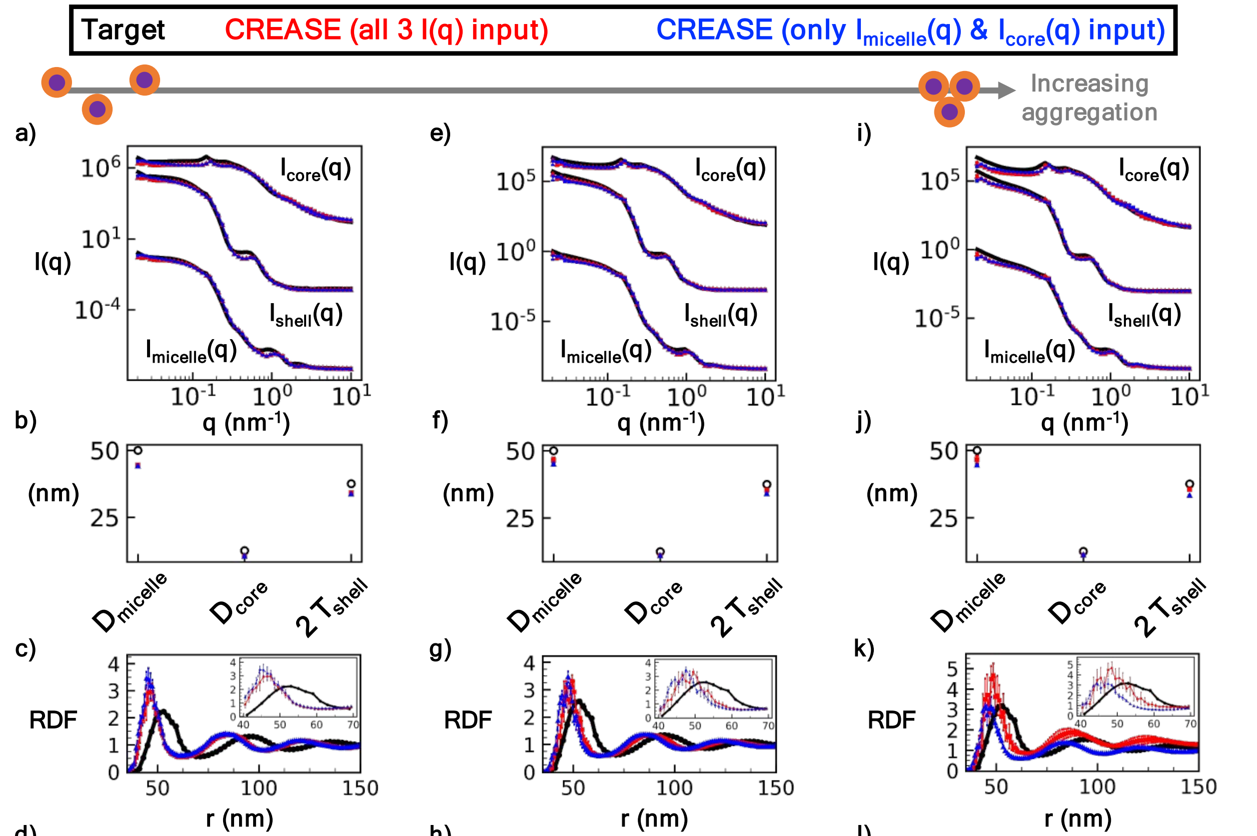
h)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.40
CREASE (3 I(q))	0.037 ± 0.001	0.357 ± 0.021
CREASE (2 I(q))	0.042 ± 0.005	0.365 ± 0.040

l)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.40
CREASE (3 I(q))	0.039 ± 0.002	0.363 ± 0.010
CREASE (2 I(q))	0.052 ± 0.009	0.321 ± 0.063

***Figure S6**: Same as Figure S1 with a 50 nm average diameter, 0.05 micelle size dispersity**, 0.75 core:micelle size ratio**, and **0.40 micelle volume fraction**.*



d)

	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.119 ± 0.001	0.275 ± 0.019
CREASE (2 I(q))	0.122 ± 0.004	0.251 ± 0.009

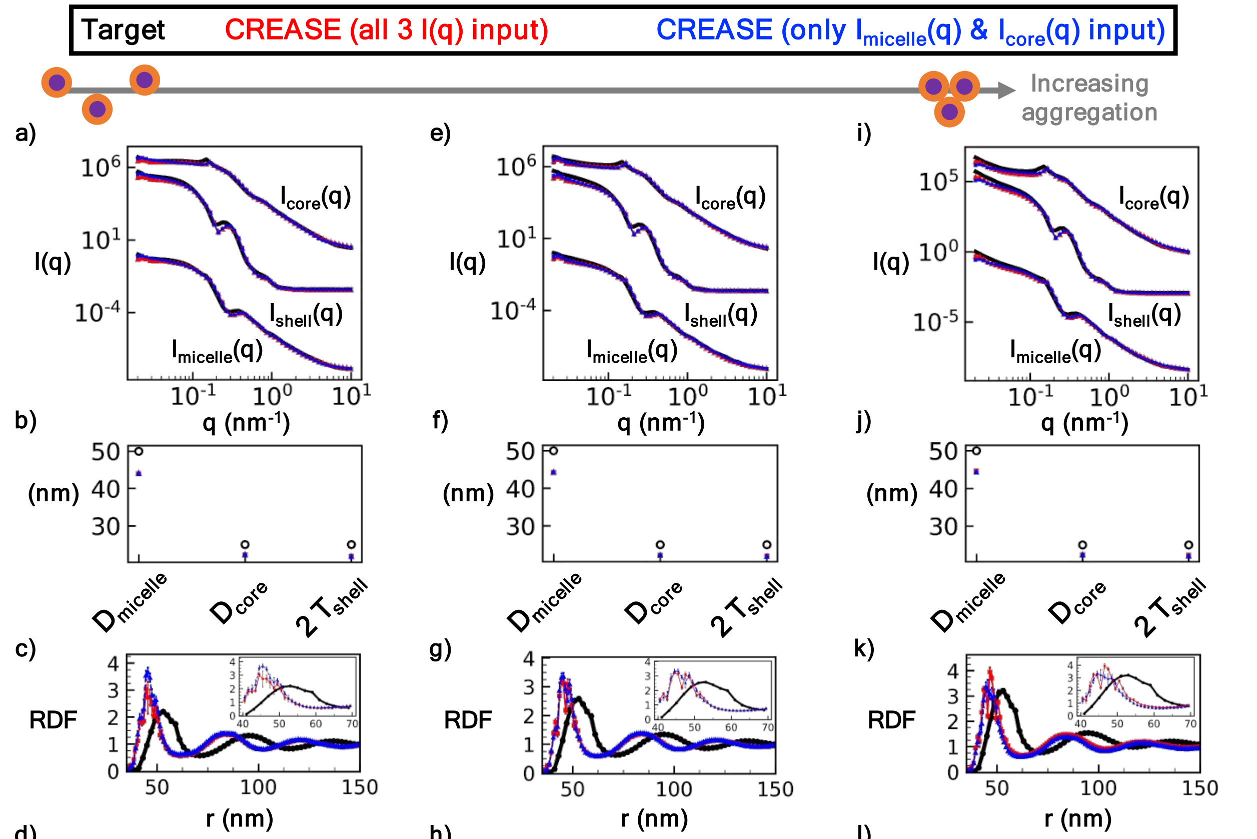
h)

	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.121 ± 0.002	0.499 ± 0.030
CREASE (2 I(q))	0.119 ± 0.001	0.495 ± 0.025

l)

	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.123 ± 0.004	0.473 ± 0.004
CREASE (2 I(q))	0.123 ± 0.001	0.520 ± 0.012

Figure S7*: Same as Figure S1 with a 50 nm average diameter,* ***0.15 micelle size dispersity****, 0.25 core:micelle size ratio, and* ***0.40 micelle volume fraction****.*



d)

	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.130 ± 0.012	0.247 ± 0.005
CREASE (2 I(q))	0.120 ± 0.001	0.250 ± 0.007

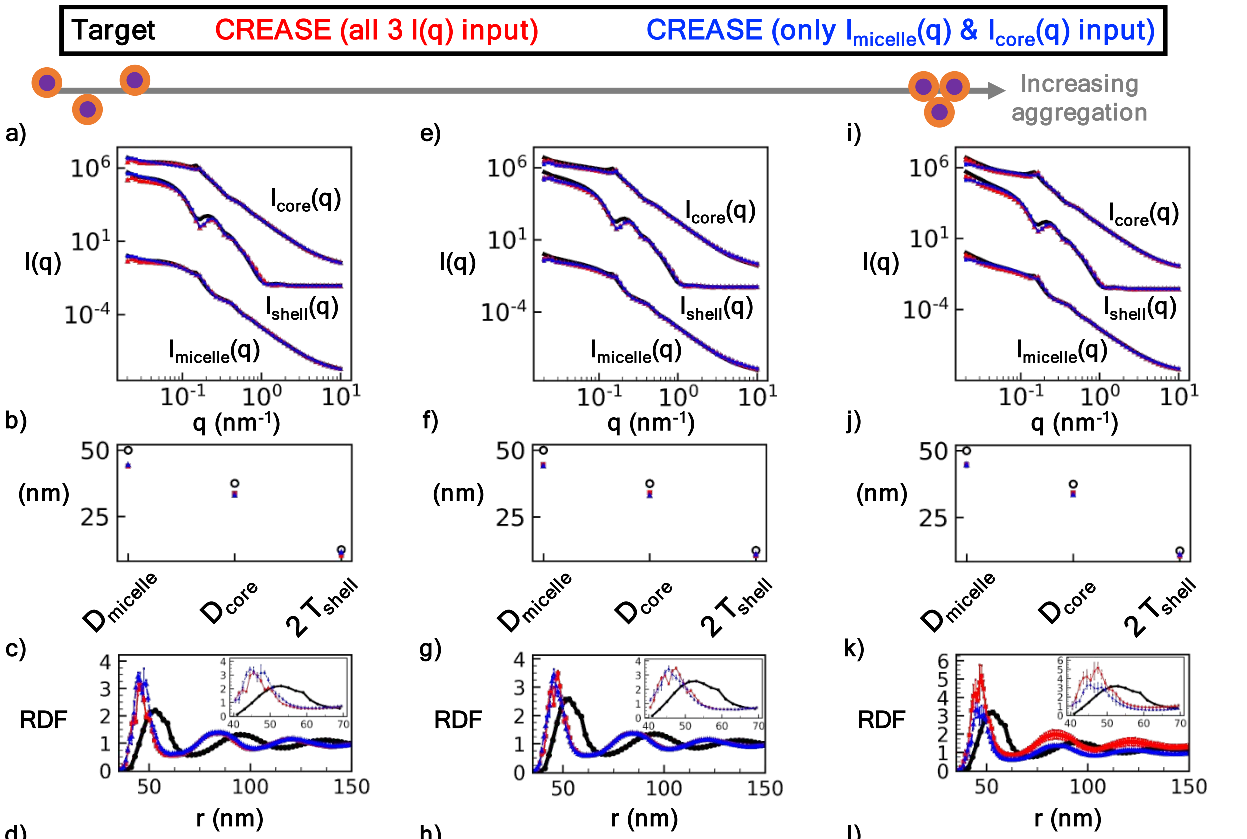
h)

	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.127 ± 0.001	0.466 ± 0.016
CREASE (2 I(q))	0.121 ± 0.001	0.467 ± 0.027

l)

	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.129 ± 0.002	0.528 ± 0.008
CREASE (2 I(q))	0.120 ± 0.001	0.467 ± 0.011

***Figure S8**: Same as Figure S1 with a 50 nm average diameter, **0.15 micelle size dispersity**, **0.50 core:micelle size ratio**, and **0.40 micelle volume fraction**.*



d)

	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.119 ± 0.001	0.274 ± 0.003
CREASE (2 I(q))	0.119 ± 0.001	0.275 ± 0.002

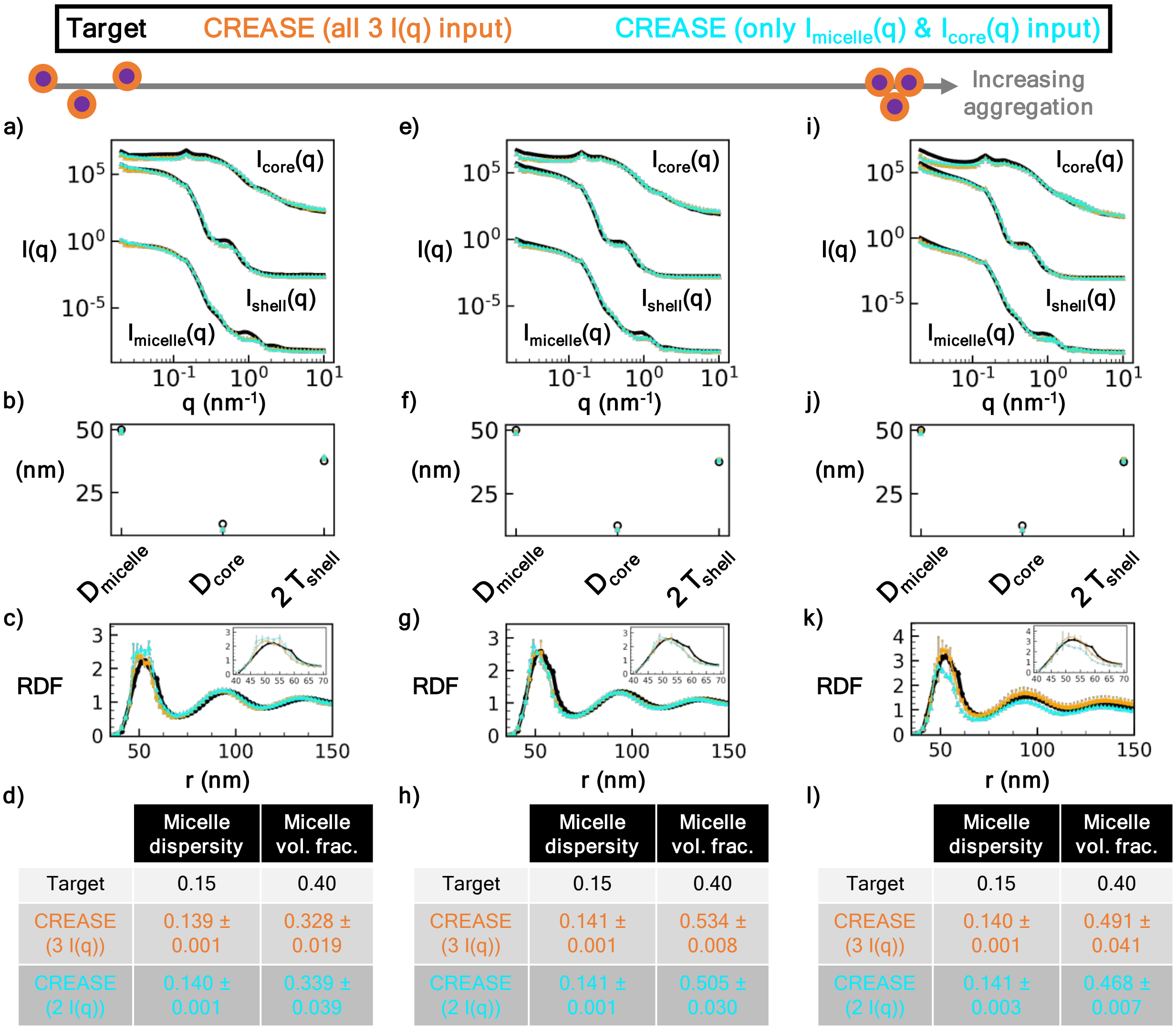
h)

	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.120 ± 0.001	0.448 ± 0.010
CREASE (2 I(q))	0.119 ± 0.001	0.352 ± 0.049

l)

	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.119 ± 0.001	0.386 ± 0.028
CREASE (2 I(q))	0.119 ± 0.001	0.387 ± 0.025

***Figure S9**: Same as Figure S1 with a 50 nm average diameter, **0.15 micelle size dispersity**, **0.75 core:micelle size ratio**, and **0.40 micelle volume fraction**.*



d)

	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.139 ± 0.001	0.328 ± 0.019
CREASE (2 I(q))	0.140 ± 0.001	0.339 ± 0.039

h)

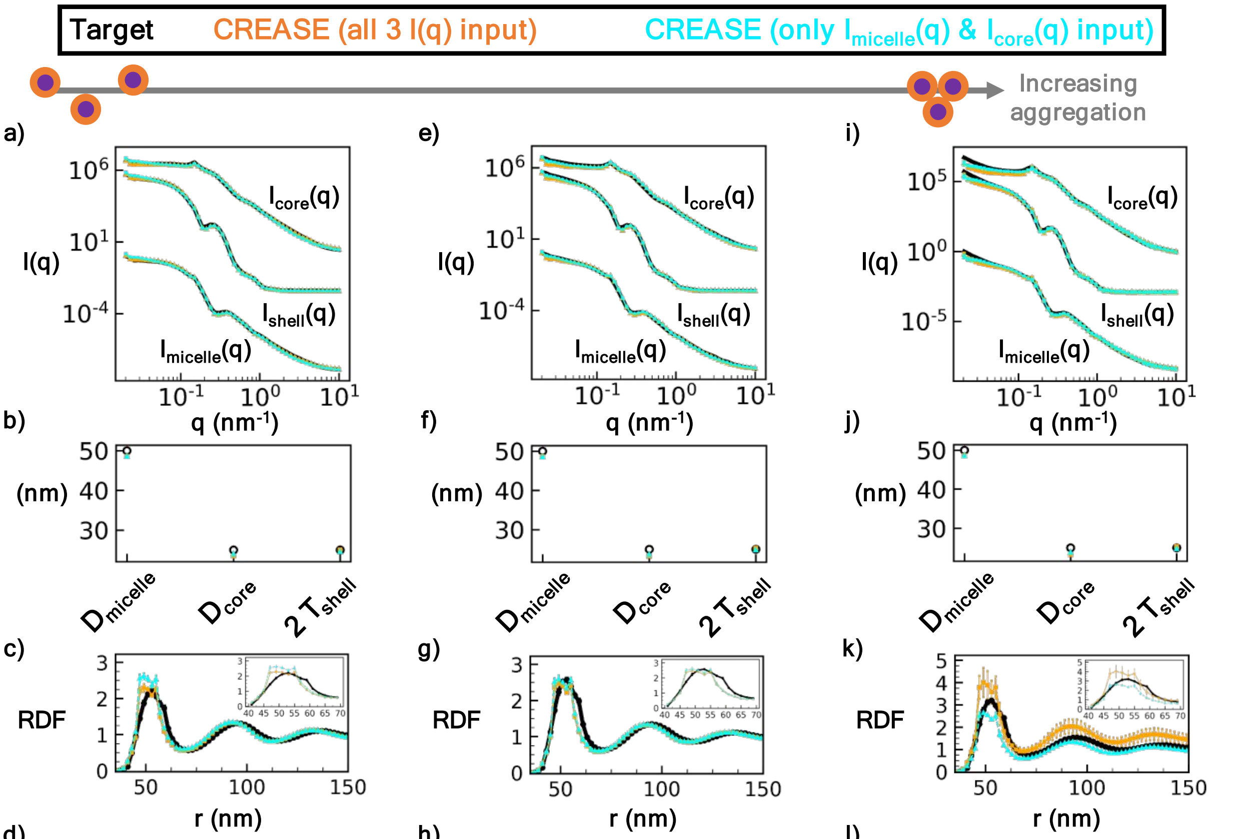
	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.141 ± 0.001	0.534 ± 0.008
CREASE (2 I(q))	0.141 ± 0.001	0.505 ± 0.030

l)

	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.140 ± 0.001	0.491 ± 0.041
CREASE (2 I(q))	0.141 ± 0.003	0.468 ± 0.007

Figure S10*: Same as Figure S1 with a 50 nm average diameter,* ***0.15 micelle size dispersity****, 0.25 core:micelle size ratio, and* ***0.40 micelle volume fraction****. For this Figure, P(q) and S(q) CREASE is provided a smaller range of micelle diameter and micelle size dispersity to demonstrate how the inclusion of addition information into P(q) and S(q) CREASE improves its performance at high size dispersity. Experimentally, one could perform cryo-TEM imaging to obtain an approximate micelle diameter and size dispersity. For both P(q) and S(q) CREASE, we set the micelle diameter as the target value (50 nm) ± 1 nm and the micelle size dispersity as the target value (0.15) ± 0.01.*

We highlight this difference by plotting the P(q) and S(q) CREASE with all three I(q) inputs in orange and the P(q) and S(q) CREASE with only two I(q) inputs in cyan.



d)

	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.140 ± 0.001	0.294 ± 0.005
CREASE (2 I(q))	0.139 ± 0.001	0.322 ± 0.007

h)

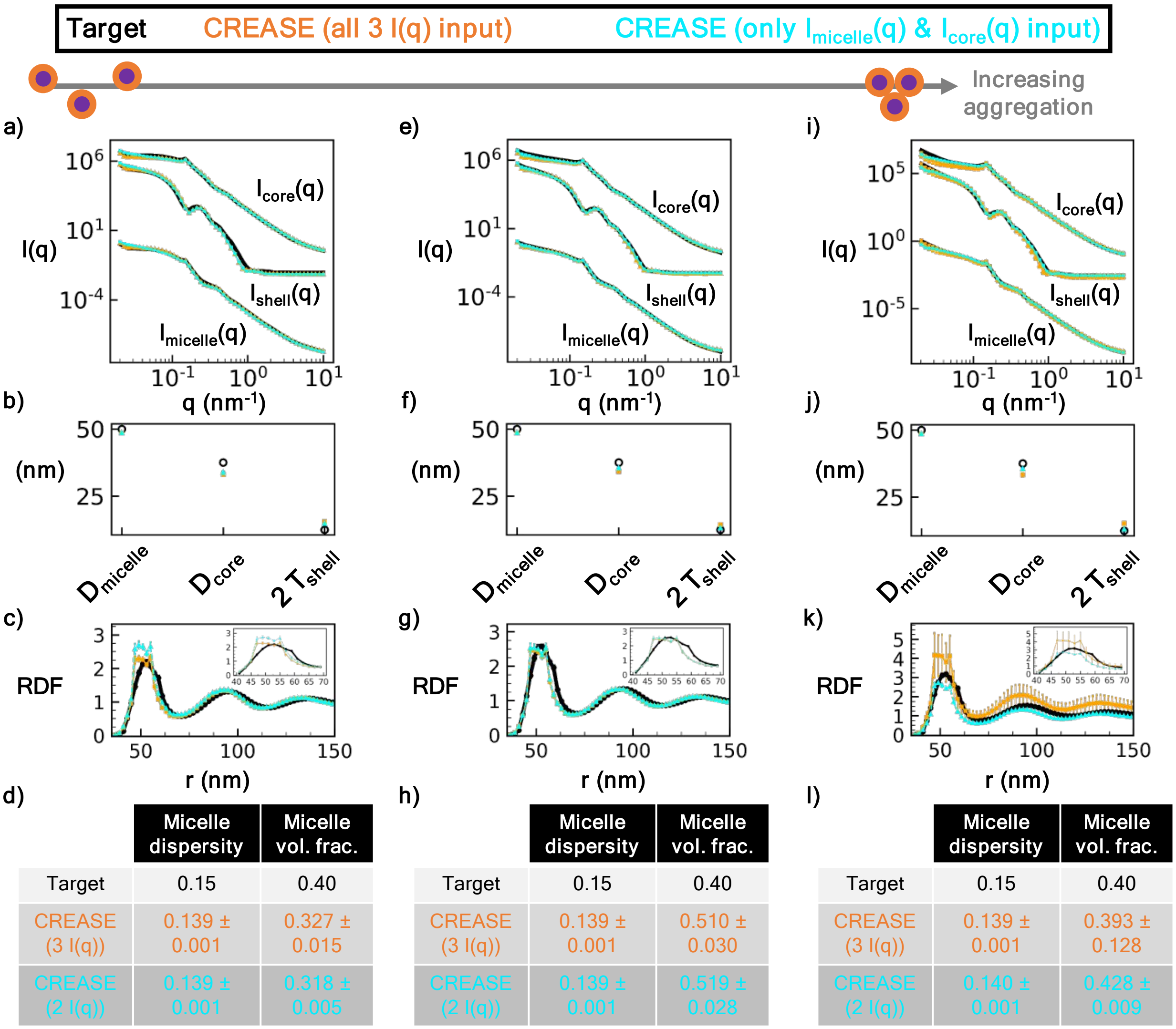
	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.139 ± 0.001	0.534 ± 0.006
CREASE (2 I(q))	0.139 ± 0.001	0.540 ± 0.002

l)

	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.140 ± 0.001	0.418 ± 0.080
CREASE (2 I(q))	0.139 ± 0.001	0.429 ± 0.094

Figure S11*: Same as Figure S1 with a 50 nm average diameter,* ***0.15 micelle size dispersity****,* ***0.50 core:micelle size ratio****, and* ***0.40 micelle volume fraction****. For this Figure, P(q) and S(q) CREASE is provided a smaller range of micelle diameter and micelle size dispersity to demonstrate how the inclusion of addition information into P(q) and S(q) CREASE improves its performance at high size dispersity. Experimentally, one could perform cryo-TEM imaging to obtain an approximate micelle diameter and size dispersity. For both P(q) and S(q) CREASE, we set the micelle diameter as the target value (50 nm) ± 1 nm and the micelle size dispersity as the target value (0.15) ± 0.01.*

We highlight this difference by plotting the P(q) and S(q) CREASE with all three I(q) inputs in orange and the P(q) and S(q) CREASE with only two I(q) inputs in cyan.



d)

	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.139 ± 0.001	0.327 ± 0.015
CREASE (2 I(q))	0.139 ± 0.001	0.318 ± 0.005

h)

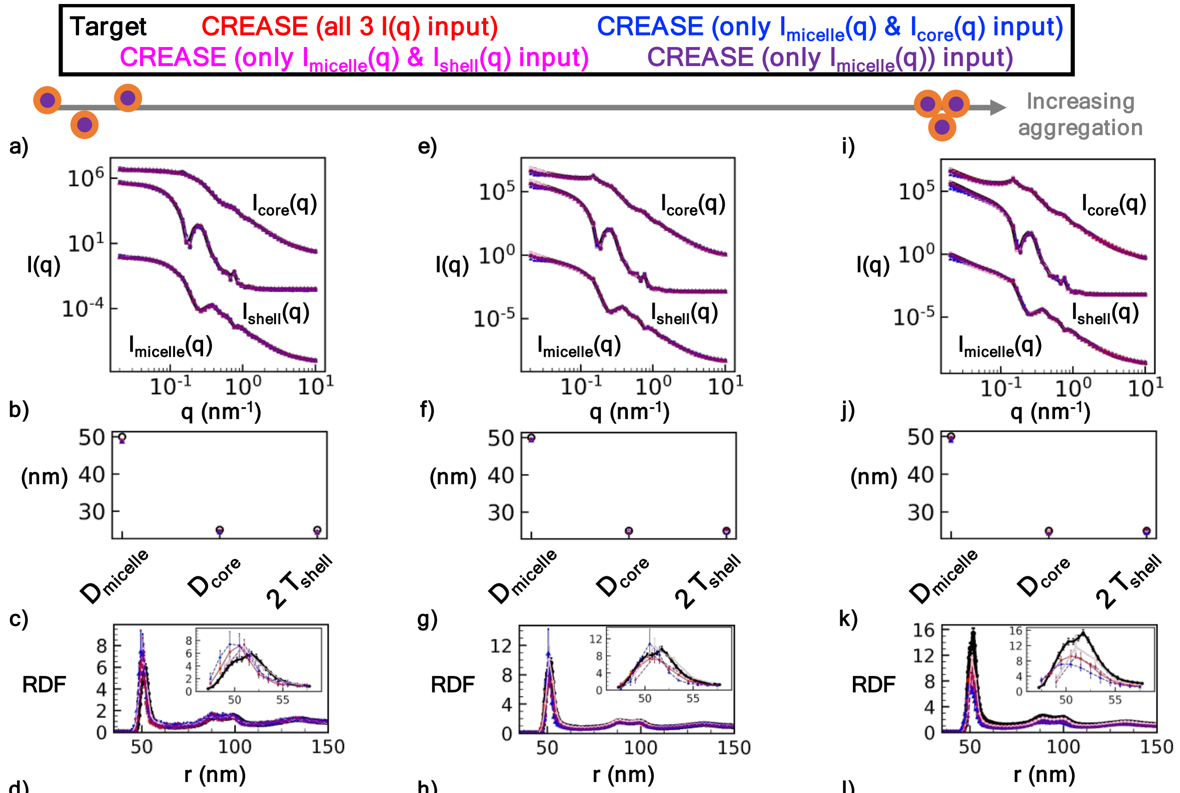
	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.139 ± 0.001	0.510 ± 0.030
CREASE (2 I(q))	0.139 ± 0.001	0.519 ± 0.028

l)

	Micelle dispersity	Micelle vol. frac.
Target	0.15	0.40
CREASE (3 I(q))	0.139 ± 0.001	0.393 ± 0.128
CREASE (2 I(q))	0.140 ± 0.001	0.428 ± 0.009

Figure S12*: Same as Figure S1 with a 50 nm average diameter,* ***0.15 micelle size dispersity****,* ***0.50 core:micelle size ratio****, and* ***0.40 micelle volume fraction****. For this Figure, P(q) and S(q) CREASE is provided a smaller range of micelle diameter and micelle size dispersity to demonstrate how the inclusion of addition information into P(q) and S(q) CREASE improves its performance at high size dispersity. Experimentally, one could perform cryo-TEM imaging to obtain an approximate micelle diameter and size dispersity. For both P(q) and S(q) CREASE, we set the micelle diameter as the target value (50 nm) ± 1 nm and the micelle size dispersity as the target value (0.15) ± 0.01.*

We highlight this difference by plotting the P(q) and S(q) CREASE with all three I(q) inputs in orange and the P(q) and S(q) CREASE with only two I(q) inputs in cyan.



d)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.15
CREASE 3 I(q)	0.040 ± 0.009	0.055 ± 0.001
CREASE $I_M(q)+I_{core}(q)$	0.047 ± 0.015	0.060 ± 0.010
CREASE $I_M(q)+I_{shell}(q)$	0.047 ± 0.002	0.054 ± 0.001
CREASE $I_M(q)$	0.029 ± 0.003	0.052 ± 0.002

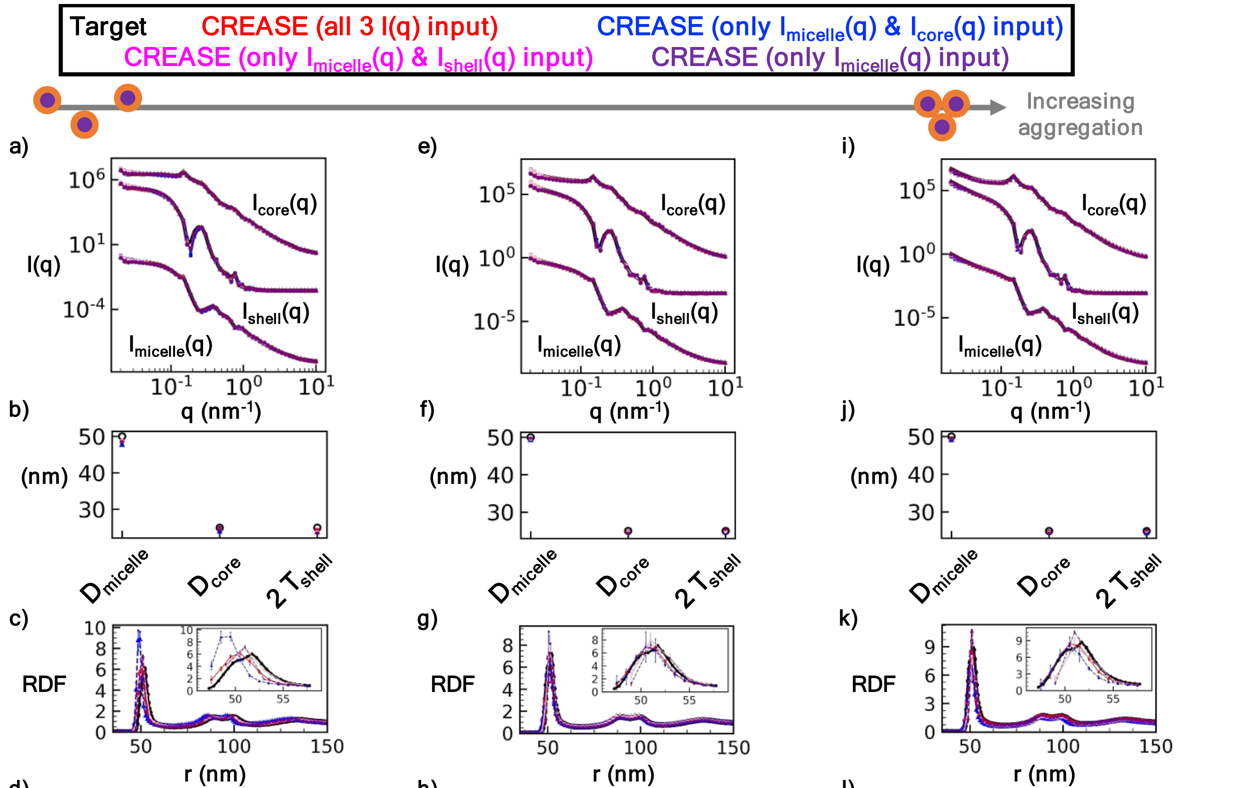
h)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.15
CREASE 3 I(q)	0.051 ± 0.006	0.244 ± 0.004
CREASE $I_M(q)+I_{core}(q)$	0.032 ± 0.006	0.217 ± 0.024
CREASE $I_M(q)+I_{shell}(q)$	0.046 ± 0.009	0.0197 ± 0.018
CREASE $I_M(q)$	0.041 ± 0.009	0.268 ± 0.025

l)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.15
CREASE 3 I(q)	0.056 ± 0.008	0.245 ± 0.035
CREASE $I_M(q)+I_{core}(q)$	0.061 ± 0.007	0.332 ± 0.051
CREASE $I_M(q)+I_{shell}(q)$	0.042 ± 0.013	0.333 ± 0.012
CREASE $I_M(q)$	0.036 ± 0.007	0.355 ± 0.022

Figure S13*: Same as Figure S1 with a 50 nm average diameter, 0.05 micelle size dispersity,* ***0.50 core:micelle size ratio****, and 0.15 micelle volume fraction. We apply CREASE with all three I(q) used as inputs (red), CREASE with only the $I_{micelle}(q)$ and $I_{core}(q)$ used as inputs (blue), CREASE with only the $I_{micelle}(q)$ and $I_{shell}(q)$ used as inputs (light pink), and CREASE with only the $I_{micelle}(q)$ used as input (purple). While the blue, light pink, and purple cases receive fewer than all three I(q) curves as inputs, we calculate the missing I(q) curve(s) from the output structure for comparison.*



d)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.40
CREASE 3 I(q)	0.048 ± 0.004	0.372 ± 0.031
CREASE $I_M(q)+I_{core}(q)$	0.032 ± 0.005	0.330 ± 0.008
CREASE $I_M(q)+I_{shell}(q)$	0.043 ± 0.001	0.320 ± 0.019
CREASE $I_M(q)$	0.027 ± 0.002	0.310 ± 0.010

h)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.40
CREASE 3 I(q)	0.046 ± 0.004	0.424 ± 0.009
CREASE $I_M(q)+I_{core}(q)$	0.050 ± 0.017	0.443 ± 0.013
CREASE $I_M(q)+I_{shell}(q)$	0.041 ± 0.004	0.390 ± 0.010
CREASE $I_M(q)$	0.033 ± 0.003	0.451 ± 0.021

l)

	Micelle dispersity	Micelle vol. frac.
Target	0.05	0.40
CREASE 3 I(q)	0.047 ± 0.009	0.438 ± 0.001
CREASE $I_M(q)+I_{core}(q)$	0.043 ± 0.013	0.423 ± 0.001
CREASE $I_M(q)+I_{shell}(q)$	0.039 ± 0.006	0.449 ± 0.004
CREASE $I_M(q)$	0.026 ± 0.001	0.463 ± 0.012

***Figure S14**: Same as Figure S1 with a 50 nm average diameter, 0.05 micelle size dispersity, **0.50 core:micelle size ratio**, and **0.40 micelle volume fraction**. We apply CREASE with all three I(q) used as inputs (red), CREASE with only the $I_{micelle}(q)$ and $I_{core}(q)$ used as inputs (blue), CREASE with only the $I_{micelle}(q)$ and $I_{shell}(q)$ used as inputs (light pink), and CREASE with only the $I_{micelle}(q)$ used as input (purple). While the blue, light pink, and purple cases receive fewer than all three I(q) curves as inputs, we calculate the missing I(q) curve(s) from the output structure for comparison.*

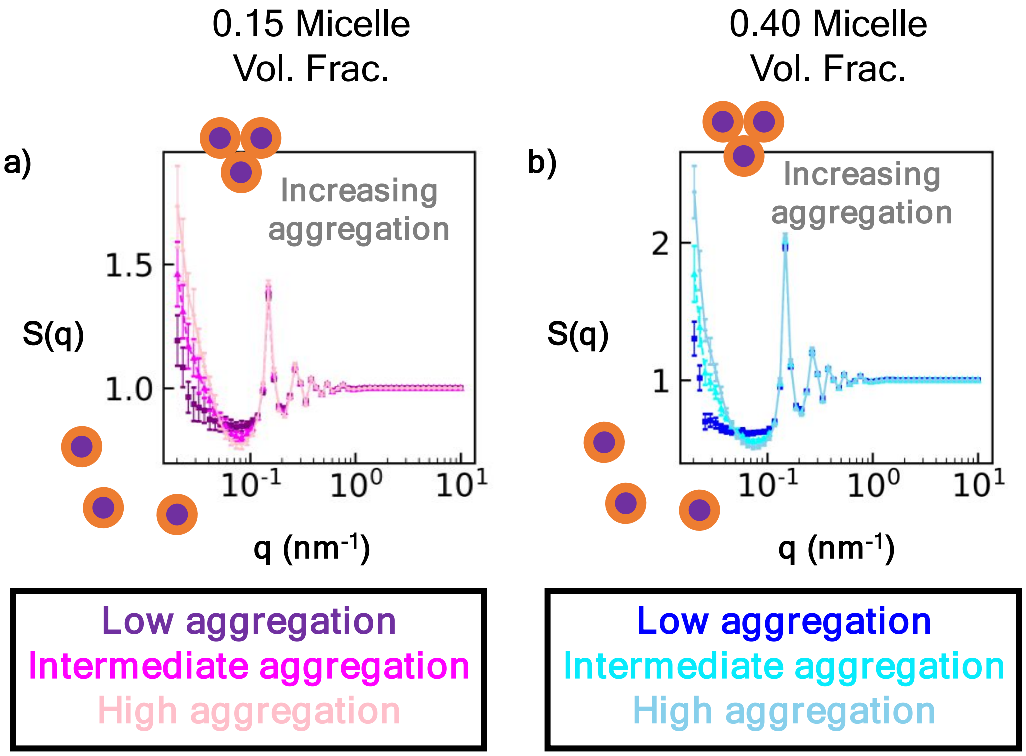


***Figure S15**: Structure factor, S(q), for the in silico target systems with a 50 nm average diameter, 0.05 micelle size dispersity, and 0.50 core:micelle size ratio at a) 0.15 micelle volume fraction and b) 0.40 micelle volume fraction. The S(q) was extracted from the 0.50 core:micelle size ratio system, but it is representative of the other core:micelle size ratios as that parameter does not influence the S(q).*

III. Application of P(q) and S(q) CREASE method on surfactant adsorbed nanoparticle solutions

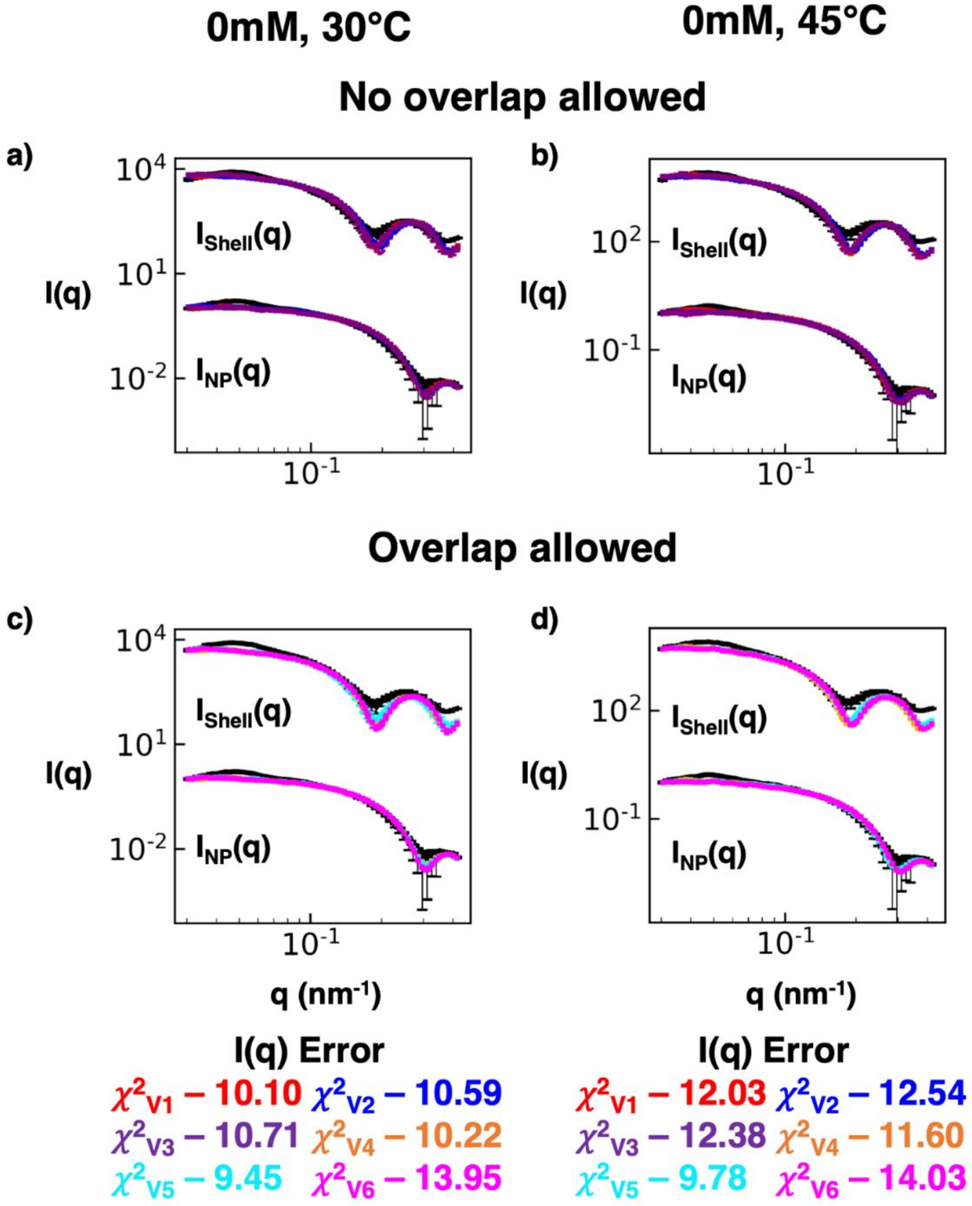


***Figure S16**: ‘P(q) and S(q) CREASE’ applied to experimental small angle neutron scattering of surfactant adsorbed nanoparticles to simultaneously identify the nanoparticle and surfactant shell dimensions and the structural arrangement in solution for various solution. a-d) Scattering*

intensity, $I(q)$, from the surfactant shell, $I_{shell}(q)$, and from the nanoparticle, $I_{NP}(q)$. The black line is the experimental contrast matched small angle neutron scattering profile, and the colored lines are the scattering profile results from various versions of 'P(q) and S(q) CREASE'. The red line is 'P(q) and S(q) CREASE' assuming a constant surfactant shell thickness; the blue line is 'P(q) and S(q) CREASE' assuming the surfactant shell thickness scales with the nanoparticle size; the purple line is 'P(q) and S(q) CREASE' assuming the surfactant shell has an average thickness and thickness dispersity independent of the nanoparticle; the orange line is 'P(q) and S(q) CREASE' assuming a constant surfactant shell thickness and allowing overlap between neighboring coated nanoparticles; the cyan line is 'P(q) and S(q) CREASE' assuming the surfactant shell thickness scales with the nanoparticle size and allowing overlap between neighboring coated nanoparticles; the magenta line is 'P(q) and S(q) CREASE' assuming the surfactant shell has an average thickness and thickness dispersity independent of the nanoparticle and allowing overlap between neighboring coated nanoparticles. a) and b) are the surfactant coated nanoparticle scattering showing the 'P(q) and S(q) CREASE' variations that do not allow surfactant shell overlap (red, blue, purple lines) at 30°C or 45°C, respectively. c) and d) are the surfactant coated nanoparticle scattering showing the 'P(q) and S(q) CREASE' variations that allow surfactant shell overlap (orange, cyan, magenta lines) at 30°C or 45°C, respectively. The χ^2 value is a quantitative measure of the scattering matches between the target (black) and 'P(q) and S(q) CREASE' variation with a lower value indicating a closer fit. The error bars are the experimental SANS standard deviation and the standard deviation of the average of 3 independent runs of the 'P(q) and S(q) CREASE'.

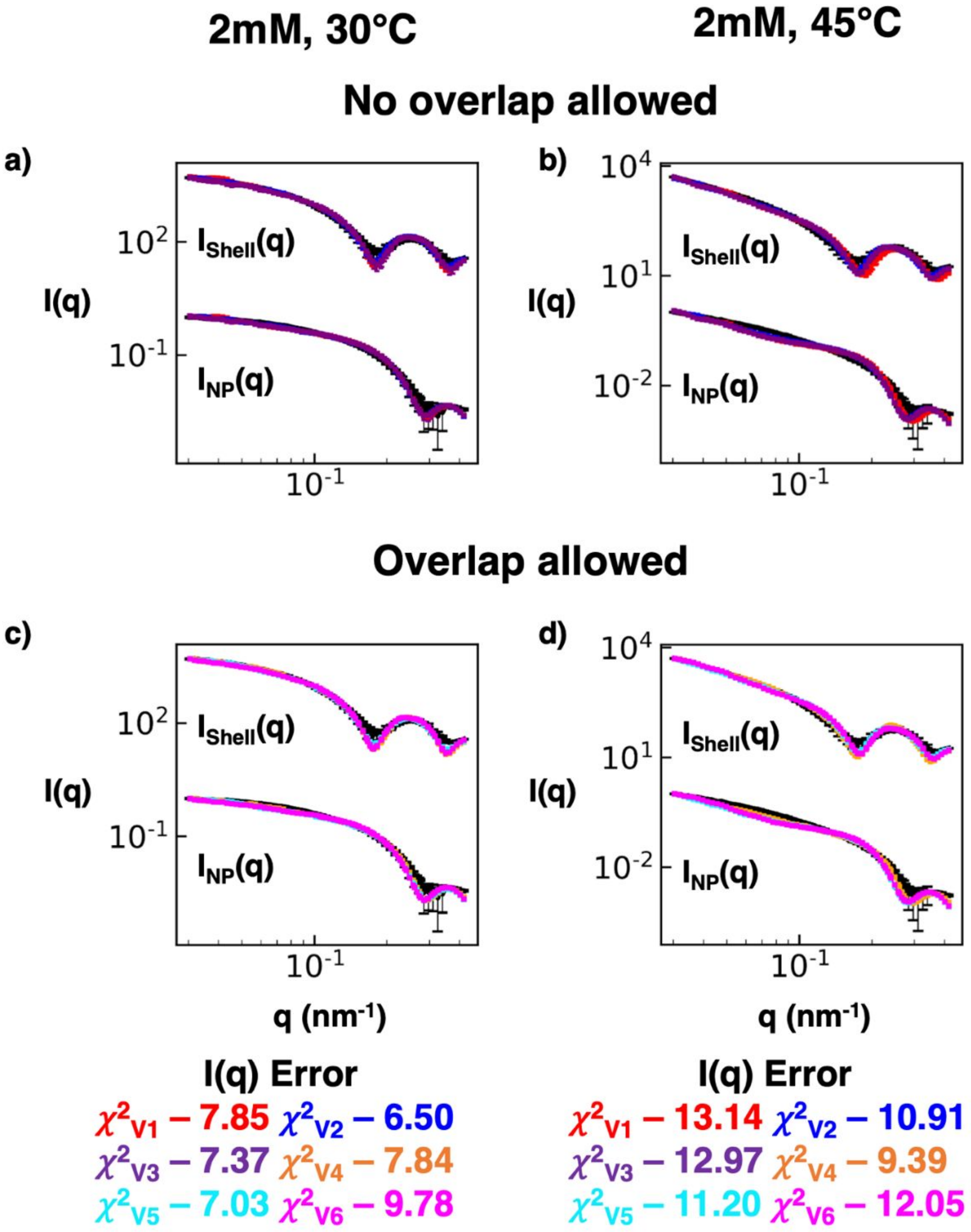


Figure S1*7: Same as Figure S16* ***for 2mM salt concentration****.*

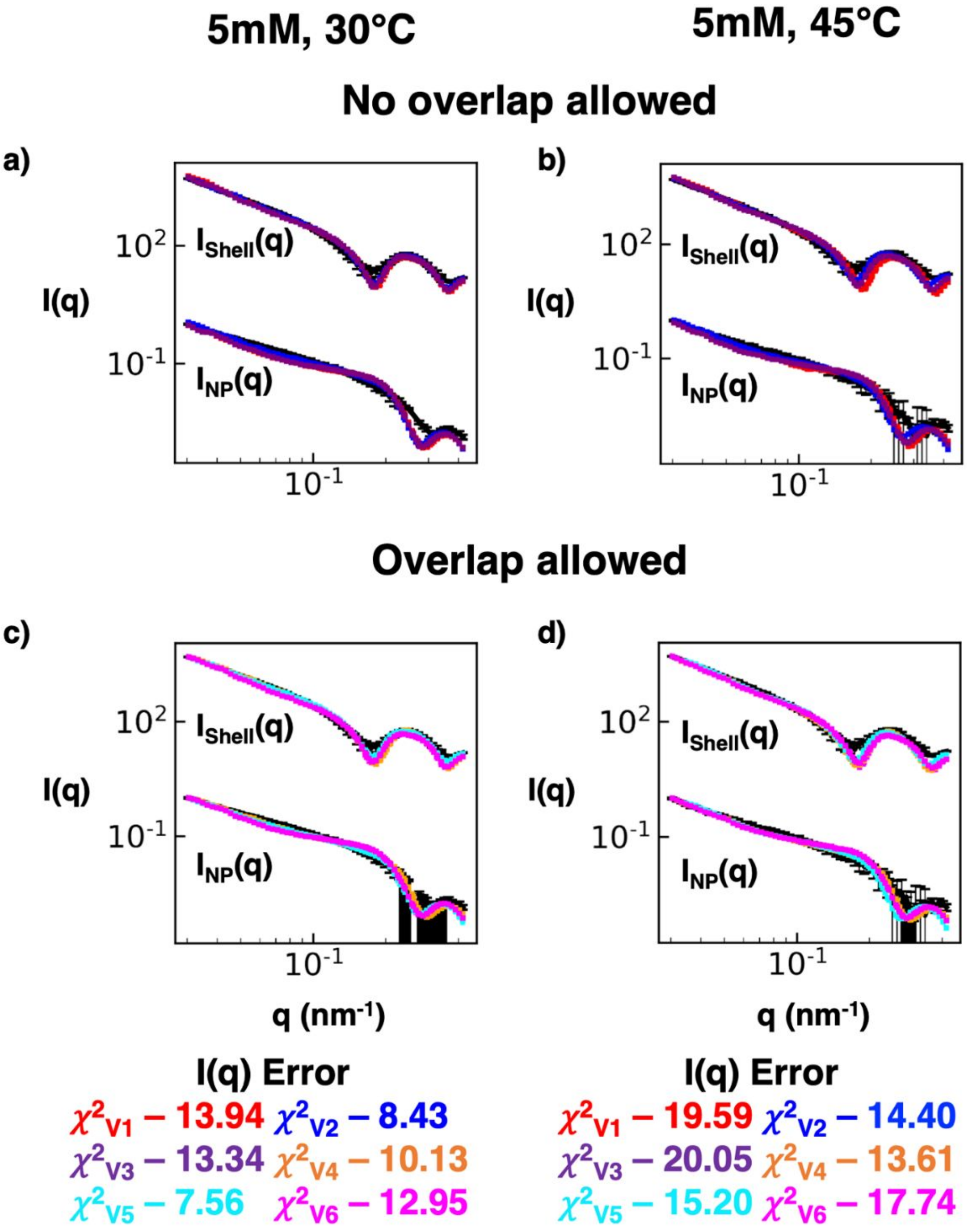


Figure S18*: Same as Figure S16 for* ***5mM salt concentration***

We compare the output from the various versions of ‘P(q) and S(q) CREASE’ to analytical model fits using the open-source commonly-used scattering fitting software package SASfit.[15, 16] The analytical modeling is performed by fitting a P(q) model and a S(q) model. The S(q) model used is the sticky hard sphere model. We consider two core-shell P(q) models; one assumes a disperse core size and constant shell thickness, and the other assumes a constant core size and disperse shell thickness. **ESI Figure S19** compares the ‘P(q) and S(q) CREASE’ to the analytical model fits at 5mM salt concentration and both solution temperatures considered in this study. Overall, we find that the ‘P(q) and S(q) CREASE’ results achieve substantially closer scattering matches to the target scattering profile than the analytical fits with the analytical fits having nearly twice the error.

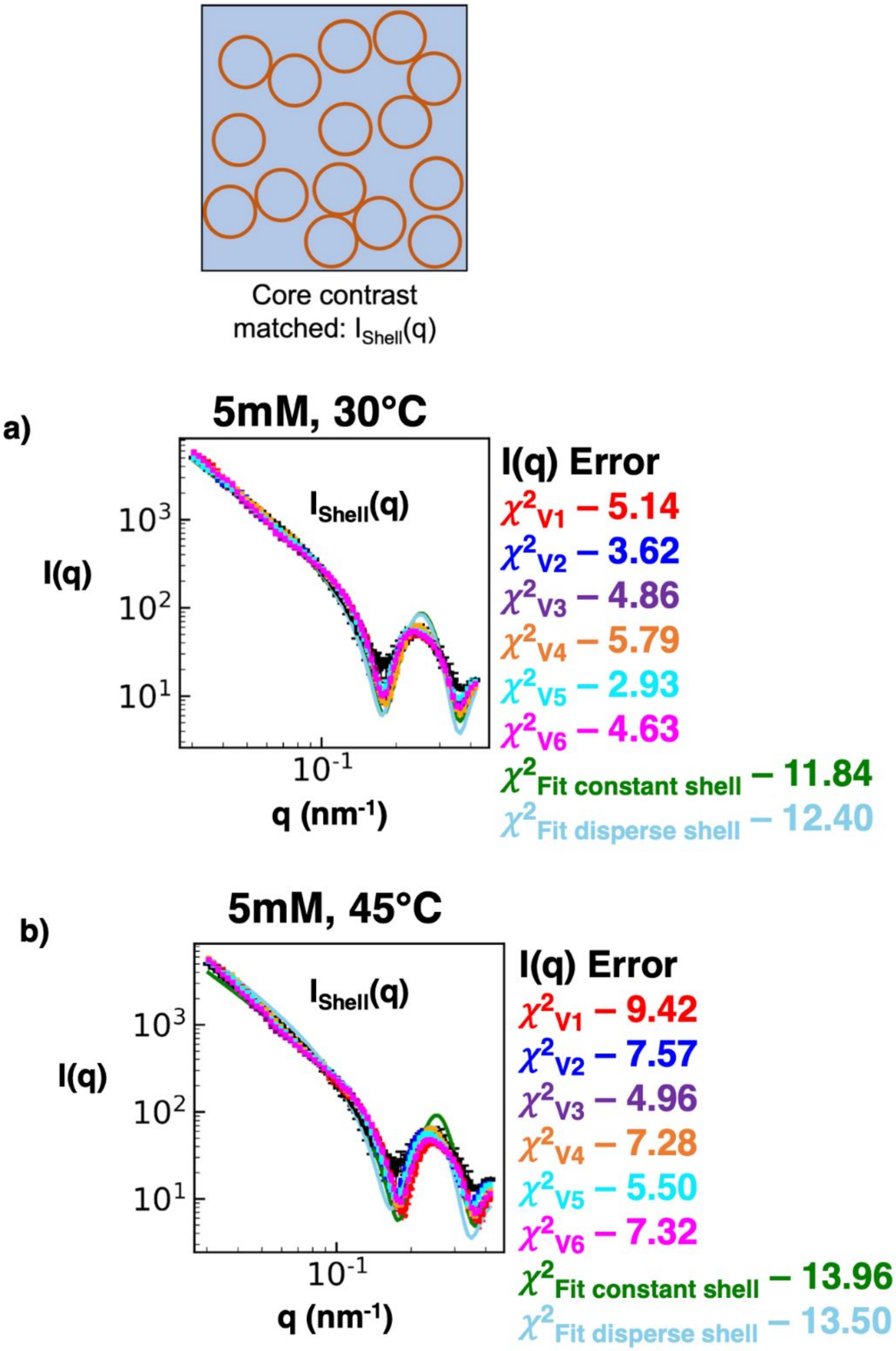


Figure S19*: ‘P(q) and S(q) CREASE’ applied to experimental small angle neutron scattering of surfactant adsorbed nanoparticles to simultaneously identify the nanoparticle and surfactant shell dimensions and the structural arrangement in solution for various solution. a) Scattering intensity, I(q), from the surfactant shell, $I_{shell}(q)$. The black line is the experimental contrast matched small*

angle neutron scattering profile, the colored symbols are the scattering profile results from various versions of 'P(q) and S(q) CREASE', and the colored lines are the analytical fits for the two analytical form factors (constant shell or disperse shell) considered. The red symbol is 'P(q) and S(q) CREASE' assuming a constant surfactant shell thickness; the blue symbol is 'P(q) and S(q) CREASE' assuming the surfactant shell thickness scales with the nanoparticle size; the purple symbol is 'P(q) and S(q) CREASE' assuming the surfactant shell has an average thickness and thickness dispersity independent of the nanoparticle; the orange symbol is 'P(q) and S(q) CREASE' assuming a constant surfactant shell thickness and allowing overlap between neighboring coated nanoparticles; the cyan symbol is 'P(q) and S(q) CREASE' assuming the surfactant shell thickness scales with the nanoparticle size and allowing overlap between neighboring coated nanoparticles; the magenta symbol is 'P(q) and S(q) CREASE' assuming the surfactant shell has an average thickness and thickness dispersity independent of the nanoparticle and allowing overlap between neighboring coated nanoparticles; the green line is the SASfit analytical fit assuming a sticky hard sphere S(q) model and a core-shell P(q) model with disperse core size and constant shell thickness; the light blue line is the SASfit analytical fit assuming a sticky hard sphere S(q) model and a core-shell P(q) model with constant core size and disperse shell thickness. a) and b) are the 30°C and 45°C solution temperature respectively. The χ^2 value is a quantitative measure of the scattering matches between the target (black) and 'P(q) and S(q) CREASE' variation or SASfit model with a lower value indicating a closer fit. Note that here the χ^2 value is only for the $I_{shell}(q)$. The error bars are the experimental SANS standard deviation and the standard deviation of the average of 3 independent runs of the 'P(q) and S(q) CREASE'. The SASfit model has no error prediction.

References:

(1) 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 Central Science* **2022**, *8* (7), 996-1007. DOI: 10.1021/acscentsci.2c00382.
(2) Hammouda, B. SANS from polymers—review of the recent literature. *Journal of Macromolecular Science®, Part C: Polymer Reviews* **2010**, *50* (1), 14-39.
(3) Schmidhuber, J. Deep learning in neural networks: An overview. *Neural networks* **2015**, *61*, 85-117.
(4) 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. *Journal of the American Chemical Society* **2019**, *141* (37), 14916-14930.
(5) 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* (2), 783-796.
(6) Wessels, M. G.; Jayaraman, A. Machine learning enhanced computational reverse engineering analysis for scattering experiments (crease) to determine structures in amphiphilic polymer solutions. *ACS Polymers Au* **2021**, *1* (3), 8581-8591.
(7) Wu, Z.; Jayaraman, A. Machine learning enhanced computational reverse-engineering analysis for scattering experiments (CREASE) for analyzing fibrillar structures in polymer solutions. *Macromolecules* **2022**, *55* (24), 11076-11091. DOI: 10.1021/acs.macromol.2c02165
(8) Lu, F.; Vo, T.; Zhang, Y.; Frenkel, A.; Yager, K. G.; Kumar, S.; Gang, O. Unusual packing of soft-shelled nanocubes. *Science advances* **2019**, *5* (5), eaaw2399.
(9) Yager, K. G.; Zhang, Y.; Lu, F.; Gang, O. Periodic lattices of arbitrary nano-objects: modeling and applications for self-assembled systems. *Journal of Applied Crystallography* **2014**, *47* (1), 118-129.
(10) Jeffries, C. M.; Ilavsky, J.; Martel, A.; Hinrichs, S.; Meyer, A.; Pedersen, J. S.; Sokolova, A. V.; Svergun, D. I. Small-angle X-ray and neutron scattering. *Nature Reviews Methods Primers* **2021**, *1* (1), 1-39.
(11) Kikhney, A. G.; Svergun, D. I. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS letters* **2015**, *589* (19), 2570-2577.
(12) Svergun, D. I.; Koch, M. H. Small-angle scattering studies of biological macromolecules in solution. *Reports on Progress in Physics* **2003**, *66* (10), 1735.
(13) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D. LAMMPS-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications* **2022**, *271*, 108171.
(14) 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* (11), 1925-1936.
(15) Breßler, I.; Kohlbrecher, J.; Thünemann, A. F. SASfit: a tool for small-angle scattering data analysis using a library of analytical expressions. *Journal of applied crystallography* **2015**, *48* (5), 1587-1598.

(16) Kohlbrecher, J.; Breßler, I. Updates in SASfit for fitting analytical expressions and numerical models to small-angle scattering patterns. *Journal of Applied Crystallography* **2022**, *55* (6).