A model for the simulation of the $\mathrm{C_nE_m}$ nonionic surfactant family derived from recent experimental results

Michael A. Johnston,[†] Andrew Ian Duff,[‡] Richard L. Anderson,[‡] and William C. Swope[*,¶]

[†]*IBM Research Ireland, Dublin, Ireland*

[‡]*STFC Hartree Centre, SciTech Daresbury, Warrington, Cheshire WA4 4AD, UK*

[¶]*IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA*

E-mail: wswope30@gmail.com

Abstract

Using a comprehensive set of recently published experimental results for training and validation, we have developed computational models appropriate for simulations of aqueous solutions of poly(ethylene oxide) alkyl ethers, an important class of micelle-forming nonionic surfactants, usually denoted C_nE_m. These models are suitable for use in simulations that employ a moderate amount of coarse graining and especially for dissipative particle dynamics (DPD), which we adopt in this work.

The experimental data used for training and validation were reported earlier and produced in our laboratory using dynamic light scattering (DLS) measurements performed on twelve members of the C_nE_m compound family yielding micelle size distribution functions and mass weighted mean aggregation numbers at each of several surfactant concentrations. The range of compounds and quality of the experimental results were designed to support the development of computational models. An essential feature of this work is that all simulation results were analysed in a way that is consistent with the experimental data. Proper account is taken of the fact that a broad distribution of micelle sizes exists, so mass weighted averages (rather than number weighted averages) over this distribution are required for the proper comparison of simulation and experimental results.

The resulting DPD force field reproduces several important trends seen in the experimental critical micelle concentrations and mass averaged mean aggregation numbers with respect to surfactant characteristics and concentration. We feel it can be used to investigate a number of open questions regarding micelle sizes and shapes and their dependence on surfactant concentration for this important class of nonionic surfactants.

1 Introduction

Surfactants are amphiphilic molecules that play important roles in science and industry with uses that vary widely and include key components for soap, detergent and toothpaste formulations, emulsifiers and stabilizers for food products, and even aircraft deicing agents. In

cellular biology, surfactants are used to break up cell membranes, and in the pharmaceutical industry they are used to stabilize drug formulations and facilitate drug delivery.

Surfactant molecules have solvophobic and solvophilic components that allow them to form aggregates in solution that are stabilized by solvophobic interactions: the aggregates are structured with the solvophobic components of the molecules clustered together to avoid solvent and the solvophilic components organized to be in contact with solvent. The smallest of these structures are the spherical micelles, which can form even under dilute conditions. However, under various other thermodynamic conditions (e.g., solvent composition, higher surfactant concentration, temperature, salt content, pH) larger, more complex, structures can form such as micellar rods, long linear or branched wormlike micelles, vesicles, and micellar aggregates. At higher concentrations some surfactants can even form interesting phases with large scale lamellar or hexagonal structures.

As one of the simplest self-assembling structures, spherical micelles have received a lot of scientific scrutiny. Experimental techniques for studying micelles include static light scattering (SLS), measurement of diffusion constant by dynamic light scattering (DLS) and nuclear magnetic resonance (NMR) methods, surface tension (ST) measurements, careful volumetric measurements, measurements of sedimentation rates, and fluorescence spectroscopy performed on dyes included with the solvent-surfactant mixture. Studies usually attempt to determine the structural features (size and shape) of micelles as well as the lowest concentration in solvent at which they begin to form, and the thermodynamic and molecular attributes that affect these micellar characteristics. Since micelles are dynamic structures, the kinetics of micelle equilibration and the determination of time scales for structural reorganization are also of keen interest.

The two most basic micellar properties are the lowest surfactant concentration at which micelles first begin to form, known as the critical micelle concentration (CMC), and the average number of surfactant molecules in the micelles, known as the mean aggregation number (N_{agg}). In spite of extensive experimental work performed over approximately 60

years, there is often disagreement among research groups using, perhaps, different experimental techniques about even these basic properties. Many reasons for these discrepancies exist and have been discussed extensively in our previous[1] work. These reasons include the sensitivity of many of the results to compound and solvent purity and thermodynamic conditions, such as temperature, surfactant concentration, and the presence of salt. Also, different experimental techniques measure slightly different properties and with differing degrees of accuracy. For example, micelle sizes can be measured by SLS, producing a radius of gyration, or by DLS, producing a hydrodynamic radius. These size metrics have very different meanings and must be compared with caution. Attempts to produce the quantities of interest, such as mean aggregation numbers, from micelle size measurements involve numerous assumptions, such as the amount of water loading in the micelles, and so aggregation numbers from different research groups are rarely in close agreement. Many experimental results are interpreted assuming that all the micelles in the sample are spherical, although some work[2–5] suggests this is not universally true. DLS measurements have shown that for many compounds, there is actually a very broad distribution of micelle sizes, yet the interpretation of most experiments assumes this distribution is narrow. This matters because some experimental techniques report an average size that is number averaged over the micelle size distribution, whereas others report a size that is mass weighted or z-averaged over this distribution. The different weightings are inherent in the experimental method and associated data analysis. The average micelle sizes reported from different methods might be numerically similar if the size distribution is sufficiently narrow. However, we have seen that the width of the distribution depends on the chemical nature of the surfactant and can vary widely even within a closely related family of surfactants. All these issues can make it very difficult to compare results from different experimental techniques on a given compound and, especially, over a range of compounds.

The state of the experimental literature, coupled with the scientifically rich phenomenology and industrial significance of surfactants suggests that computer simulations could be

valuable in contributing to a deeper understanding of micellar behavior. Computer simulations can provide molecular level detail that is usually lacking in experiments on large ensembles of molecules. Two barriers exist in this effort. First, simulations of micelles and micellar phenomena are very computationally demanding. Since micelles form at low surfactant concentrations (typically near 1% surfactant, by weight), simulations that capture micelle formation need to be very large, and most of the material in the simulation must be solvent. If one wants to observe micelle formation in a simulation with the size distribution one might see in a real system, one must have enough surfactant material to be able to form several micelles. Furthermore, micelles form slowly on a computer simulation time scale, so not only do the simulations require large numbers of molecules, they must run for a very long time in order to reach an equilibrium with respect to the distribution of micelle sizes and shapes. Although detailed all atom simulations[6–14] have been made with some success, they are usually considered too expensive for a thorough investigation of micellar behavior. Currently, these performance challenges are met through the use of coarse grained models for the surfactant, such as united atom or beaded string models, whereby a single particle in the simulation actually represents several atoms of a surfactant molecule, or even multiple molecules of the solvent. These approaches also use reduced complexity in the interaction potentials (force fields), which tend to be softer and smoother than their all-atom counterparts, thereby permitting the use of larger time step sizes in the simulation. A common such approach is dissipative particle dynamics[15–18](DPD), which is the method employed here.

The second barrier to the use of computer simulations to study micellar phenomena is in finding validated and tested force fields that are sufficiently accurate for the materials, conditions, and phenomena of interest. Although some efforts have attempted to address this in the context of coarse grained simulations, they are often developed and validated to replicate known behavior for a single compound or a couple of compounds, or to capture the behavior for one compound over a range of concentrations or temperatures. These models can be very useful for studying many aspects of micellar phenomena. However, since they

were developed to model a small number of compounds, the parameters are often not useful for making predictions about compounds or behavior outside of their training set. The ability of a force field description to provide sufficiently accurate descriptions for a range of molecules (usually at the expense of any particular one of them) is known as transferability. Generally, this feature has to be designed into the parameter set through training with a range of molecules and an approach that balances the accuracy over the set of molecules.

Some efforts[19–21] have attempted to address the need for general purpose coarse grained force fields, but a generally accepted set of parameters for most molecules of interest is lacking. Such a set of parameters would depend on the nature of the molecule, the degree of coarse graining, and on the number of types of sites (number of bead types) used to represent the molecule. Furthermore, there are a number of functional forms available to describe interactions between coarse grained sites, each with its own set of parameters. At best, any such description might be limited to a small set of molecules, a specific range of thermodynamic conditions such as concentration, density, temperature and solvent, and it may provide an accurate description for only a limited set of properties and phenomena.

In spite of all these caveats and potential limitations one would hope that a force field developed to accurately reproduce experimental observables for a range of related physical properties and for a suitably small set of closely related compounds should be useful for the prediction and study of these and similar properties for other similar compounds. Systematic force field development, therefore, requires 1) a choice of what type(s) of compounds for which one wishes to have accurate models, 2) a choice of what observables one wishes to be able to model with accuracy, 3) sufficiently accurate experimental data on a range of compounds that span the desired types of compounds that can be used for training and validation of the force field model, and 4) a procedure for systematically improving the force field parameters so that computational results are brought into best agreement with experimental ones.

In this work we will describe the development of a set of parameters we feel are appropriate for the computational study of micellar phenomena for a simple family of nonionic

surfactants, the poly(ethylene oxide) alkyl ethers, with formulas $H(CH_2)_n(OCH_2CH_2)_mOH$, usually denoted C_nE_m. Although we optimized parameters to reproduce experimentally observed micellar behavior for a small family of compounds, we hope they are useful for the study of micellar behavior for closely related molecules with similar chemical functionality. Notably, we used experimental data generated in our lab especially for this effort, and we adopted Force Balance[22] for use in systematically optimizing the parameters. Force Balance has been used extensively[23,24] to generate and improve all atom force fields. Finally, we compare our final force field with a similar one produced previously[25] in our laboratory that was designed with different experimental observables in mind.

The structure of the paper is as follows: Section 2, Methods, discusses the approaches and procedures used in this work. Section 2.1, Experimental Data, discusses the choice of molecules, the properties selected for study, and the experimental data set used for force field development and testing. Section 2.2, Force Field, discusses the force field functional form used in simulations of these compounds and how the parameters in these functional forms were optimized to reproduce the experimental results. Note that some of the parameters were optimized to reproduce experimental densities for related materials, and the others were optimized to reproduce experimental CMC and N_{agg} data. Section 2.3 describes the Choice of Data for Training and Validation of the force field. Section 2.4, Computed Observables, discusses how the CMC and N_{agg} are computed from simulations. Section 2.5, Objective Function, discusses the construction of the function that was optimized with respect to force field parameters in order to align the computed and experimental observables. Section 2.6, Simulation Properties, discusses the simulation protocols, including how the system sizes were determined and how equilibration was established. Section 3, presents the Results. Section 4, Discussion, compares the quality of this force field with that of a different one, recently developed in our lab, and it examines the extent to which the new force field captures important experimental trends observed in the data. Finally, Section 6 provides a short Summary and Conclusion. An Appendix describes how the derivatives of observables

with respect to the force field parameters were computed through the computation of cross correlation functions.

2 Methods

2.1 Experimental Data

The choices of compounds and observables, as well as the development of a data set of experimental results were the subject of a previous[1] paper. These are for the simple nonionic surfactant family of poly(ethylene oxide) alkyl ethers, denoted C_nE_m (Table 1). These are fairly simple short diblock polymers with essentially two types of functional groups: a hydrophobic region consisting of the aliphatic block of the chain and a hydrophilic region that is a hydroxyl terminated polyether block. These constitute a good first choice to develop parameterization methodology. Because of their chemical simplicity, one might hope a very simple force field could be adequate for capturing much of the relevant behavior. We also hoped that the parameters we developed would be useful in the context of other surfactant compounds, many of which have structural components in common with the C_nE_m family. The micellar observables that were available from experiments and could also be computed from simulations included CMC and N_{agg}.

An extensive body of experimental literature exists for the C_nE_m compounds. However, there is often not consensus on even basic properties. Table 1 gives the range of CMC values seen in the literature. (Unless otherwise noted, these and all other properties are measured at 25 °C, our temperature of interest for this study. We note that both CMC and N_{agg} are temperature sensitive to varying degrees for compounds in this family.) After a careful review of the literature, target values for the CMC were selected to be used in the force field parameterization effort. For $C_{12}E_6$ and $C_{12}E_8$, the literature values for the CMCs fell into two distinct ranges so it was not possible to select a single target value. For $C_{12}E_6$ the CMC values were near to either 0.072 mM or 0.082 mM; and for $C_{12}E_8$ they were near to either

0.084 mM or 0.109 mM. The table shows only the larger value for each. (See the previous work[1] and extensive discussion in its Supporting Information about how the choices were made for these target CMC values.)

Table 1 shows other physical properties for the compounds selected. The compounds have a critical concentration (c_c) and temperature (T_c) from Schubert[26] (for LCST behavior) above which they phase separate into solvent-rich and surfactant-rich phases. There is also a value for the cloud temperature (T_{cloud}), the temperature of phase separation at a surfactant concentration of 10 g/L, from a compilation[27] of experimental results by Berthod. Note that $T_c \approx T_{cloud}$ and the phase separation behavior is very sensitive to the length of the hydrophilic block. This information is not used in the parameterization effort, but is included here since it is recognized that these surfactant compounds can exhibit unusual aggregation even 20 °C below the cloud temperature, suggesting that the $C_{12}E_5$ data should be used with a great deal of caution in parameterization since we are interested in micellar behavior rather than phase separation. Finally, the densities for pure surfactant are from a fit[27] to experimental density data at 25 °C, $\rho_{n,m} = (14n + 44m + 18)/(18.3n + 39.13m)$ (g/mL). These densities were used along with the density of water (0.997040 g/mL at 25 °C) to estimate aggregation numbers from hydrodynamic diameters, and also to convert between concentrations reported in mole fraction, mass fraction, and molarity (see Supporting Information for details).

An extensive review of the literature for N_{agg} values for these compounds, however, did not yield a coherent data set for force field development purposes. One needs values produced from a single experimental method, of consistent and known accuracy, and over the range of compounds to be used in the force field development. What exists, however, has been performed using about a dozen different experimental techniques, over an approximately 60 year time period, by different research groups, and with different assumptions employed in the data analysis. Coverage over the set of compounds has been inconsistent, with some compounds relatively unstudied and with others having scores of measurements performed. The values found and an extensive discussion about their unsuitability are in the Supporting

Table 1: Properties of molecules used in this study

Compound	Literature CMC, mM	Target CMC mM	Target CMC wt%	c_c mM	T_c C	T_{cloud} C	ρ g/mL
C6E3[26,28–31]	68-105	100.	2.35	624	46.0	45	1.030
C6E4[26,31–35]	72-106	106.	2.96	592	66.1		1.044
C6E5[29–33,35,36]	75-115	113.	3.65			75	1.054
C8E4[26,30,37,38]	6.5-11.7	8.0	0.246	230	40.8	40	1.010
C8E5[26,38–40]	6.0-11.0	9.0	0.316	270	61.7	60.4	1.023
C8E6[28,29,39,41]	7.6-10.8	9.9	0.392	310	74.4	74	1.034
C10E5[33,34,39]	0.68-1.00	0.86	0.0327			44	0.998
C10E6[28,34,39,40,42–44]	0.46-0.95	0.90	0.0381				1.010
C10E8[39,40,42,43,45–48]	0.28-1.15	1.00	0.0512			85	1.028
C12E5[26,33,34,39,49–53]	0.035-0.071	0.064	0.00261	37	32.0	31.7	0.978
C12E6[26,28,34,39,40,51,53]	0.060-0.087	0.082*	0.00371	55	51.3	52	0.990
C12E8[39,40,45–47,52–55]	0.056-0.109	0.109*	0.00589			78	1.010

Information of our previous[1] paper.

Table 2: Size results for C_6E_m and C_8E_m from DLS

Compound	Conc, mM	Conc, wt%	Conc/CMC	n_{cut}	$\langle D_H \rangle_M$, nm	$\langle N_{agg} \rangle_M$
C6E3	128.	3.01	1.28	8	4.54(0.02)	81(1)
	171.	4.01	1.71		4.65(0.03)	100(2)
C6E4	270.	7.51	2.55	10	3.999(0.004)	46.3(0.1)
	360.	10.01	3.40		4.02(0.01)	47.2(0.6)
C6E5	266.	8.56	2.35	8	3.747(0.003)	31.8(0.1)
	354.	11.4	3.14		3.718(0.008)	30.5(0.3)
C8E4	18.2	0.558	2.27	9	5.56(0.08)	110(4)
	24.2	0.744	3.03		5.94(0.04)	153(2)
C8E5	33.0	1.16	3.67	8	5.01(0.03)	65(1)
	55.0	1.93	6.11		5.10(0.08)	77(3)
C8E6	29.7	1.17	3.00	10	4.484(0.005)	44.8(0.4)
	49.5	1.96	5.00		4.64(0.02)	54(1)

Therefore, new DLS measurements were performed for each of twelve C_nE_m compounds that produced mass weighted mean hydrodynamic diameters ($\langle D_H \rangle_M$) as well as mass weighted micelle size distributions as functions of the hydrodynamic diameters. (See Tables 2 and 3.) Experiments were performed for each compound over a wide range of concentrations, from below the CMC of the compound up to as much as 70 times the CMC. The tables give data for each compound at some of the concentrations where we believe micelles were present

Table 3: Size results for $C_{10}E_m$ and $C_{12}E_m$ from DLS

Compound	Conc, mM	Conc, wt%	Conc/CMC	n_{cut}	$\langle D_H \rangle_M$, nm	$\langle N_{agg} \rangle_M$
C10E5	4.29	0.163	4.99	11	6.62(0.1)	158(5)
	8.59	0.326	9.99		6.61(0.04)	198(7)
	17.2	0.652	20.0		7.28(0.08)	280(5)
C10E6	4.50	0.191	5.00	10	5.61(0.01)	74.4(0.4)
	9.00	0.381	10.0		5.52(0.01)	75.4(0.3)
	18.0	0.763	20.0		5.55(0.05)	90(3)
C10E8	4.19	0.215	4.19	8	5.52(0.05)	61(1)
	5.24	0.268	5.24		5.80(0.01)	66.8(0.4)
	10.47	0.536	10.47		5.21(0.03)	57(2)
C12E5	1.92	0.0784	30.0	11	6.35	143
	2.56	0.105	40.0		6.38	143
	3.20	0.131	50.0		5.95	101
C12E6	1.78	0.0803	21.7	9	7.08(0.14)	186(11)
	2.66	0.120	32.4		6.40(0.05)	142(6)
	3.55	0.161	43.3		7.59(0.2)	257(20)
	6.22	0.281	75.9		6.53(0.2)	155(22)
	8.88	0.401	108.3		6.92(0.03)	158(3)
C12E8	2.18	0.118	20.0	8	6.20(0.01)	82.3(0.3)
	3.27	0.177	30.0		6.15(0.02)	80.9(0.4)
	4.36	0.236	40.0		6.14(0.01)	80.5(0.4)
	5.45	0.295	50.0		6.19(0.01)	81.3(0.3)

and that are appropriate for the parameterization effort. Although surfactant molecules aggregate to form micelles, these coexist with smaller clusters of surfactant molecules that we refer to as submicellar aggregates. Submicellar aggregates could not be detected by our DLS instrumentation because their scattering cross section is too small compared to that of any actual micelles that were present. The columns in the tables labeled n_{cut} give an estimate of the largest aggregate size that is not visible to our instrument and can serve as an operational definition in the analysis of computer simulations for the cutoff between aggregates that are too small to be considered micelles ($N \leq n_{cut}$) and those that are large enough ($N > n_{cut}$). Note that these cutoff values are all in the range of 8 to 11. Our earlier work[56] used values for this cutoff based on the size of aggregate clusters corresponding to the first minima in the cluster size distribution functions that were observed in simulations. Those were 13, 10 and 5 for C_6E_4, C_8E_4 and $C_{12}E_6$, respectively. Our experimental work suggests 10, 9 and 9, for these compounds, respectively, values which are actually quite similar to what was used earlier.

The tables also give the mass weighted mean aggregation number which is obtained using the following equation that was derived and discussed[1] earlier.

$$\langle N_{agg} \rangle_M = \frac{\pi}{6} \langle D_H^3 \rangle_M \left(\frac{1}{\rho_S} + \frac{m n_W}{\rho_W} \right)^{-1} \quad (1)$$

Note that in this equation the mass weighted average over the particle size distribution of D_H^3 is used rather than the cube of the average over this distribution of D_H itself. $\langle N_{agg} \rangle_M$ cannot be computed directly from $\langle D_H \rangle_M$. The equation also uses the number density of water, ρ_W, and of pure surfactant, ρ_S, as well as the number of water molecules, n_W, assumed to be bound to each of the m hydrophilic repeat units on the C_nE_m surfactant molecule. Throughout this work we have consistently used $n_W = 4$.

Unless otherwise stated, we will assume in the following that mean aggregation number, N_{agg}, refers to the *mass weighted* mean aggregation number, sometimes denoted $\langle N_{agg} \rangle_M$, as

produced from the analysis of the DLS measurements.

2.2 Force Field

Coarse graining scheme

This work uses the DPD method for simulation. There are many features relevant[16] to this method which will not be reviewed here, but the essential elements for the current discussion are the coarse graining of the molecular structure into a set of interacting beads, use of a small number of bead types, and the simple functional form for the interactions between beads. The level of coarse graining used here involves two to three heavy (non Hydrogen) atoms per bead and is illustrated in Figure 1. A solvent bead (type W) represents some number (here, two) of water molecules, a hydrophobic tail bead (type T) represents an ethylene repeat unit, and a hydrophilic head bead (type H) represents an ethylene oxide repeat unit. Some workers use greater or lesser degrees of coarse graining. We feel our approach captures most of the benefits of coarse graining without removing too much chemical detail. We also chose to use only three bead types in order to explore whether a model this simple is capable of describing the micellar behavior we have seen experimentally. Other approaches[25] might introduce, for example, additional terminal bead types for the hydrophobic end, such as a CH_3– or CH_3CH_2–, and/or a terminal bead type for the hydrophilic end, such as a hydroxyl –OH, methanol –CH_2OH, or diethoxy –OCH_2CH_2OH bead type. The LogP force field[25] against which we compare our results, uses a slightly different coarse graining strategy as well as additional bead types to describe the terminal groups.

Force field functional form

Although other functional forms can be used, it is common in DPD simulations to use an interaction energy between all pairs of beads that produces a pairwise additive conservative

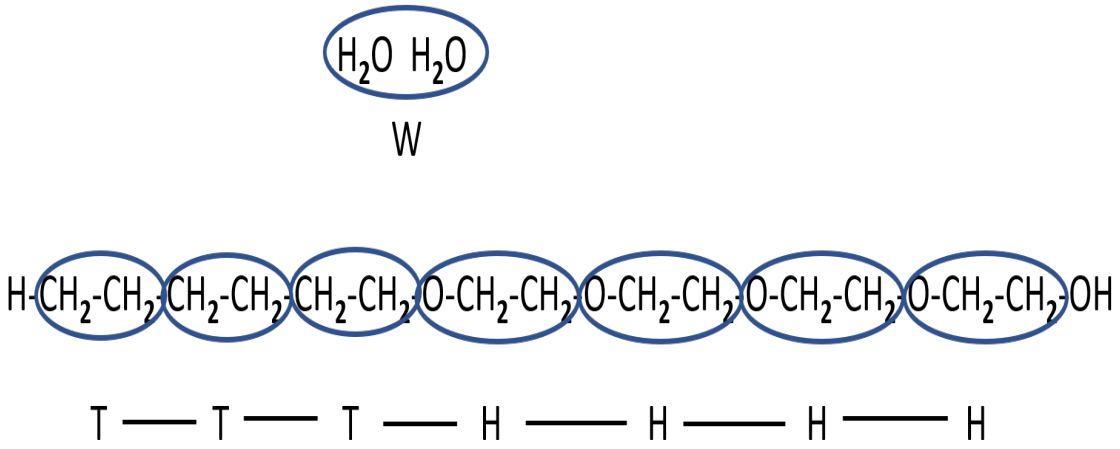

Figure 1: Coarse graining of two water molecules into a W-type bead and C_6E_4 into H_4T_3, a representation of the molecule with 3 hydrophobic T-type beads and 4 hydrophilic H-type beads.

force with a simple short ranged functional form.

$$\mathbf{F}_{i,j}(\mathbf{r}_{ij}) = \begin{cases} A_{IJ}\left(1 - r_{ij}/R_{c,IJ}\right)\hat{\mathbf{r}}_{ij} & r_{ij} < R_{c,IJ} \\ 0 & r_{ij} \geq R_{c,IJ} \end{cases} \tag{2}$$

where $\mathbf{F}_{i,j}$ is the contribution to the vector force on a bead i due to its interaction with a bead j, which depends on $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$, the vector displacement that points from bead j to bead i. We use lower case (i, j) to designate bead indices and upper case (I, J) to indicate their bead type. Hence, in this work I and J can each be any of W, H or T. $r_{ij} = |\mathbf{r}_{ij}|$ is the scalar distance between the beads, and $\hat{\mathbf{r}}_{ij} = \mathbf{r}_{ij}/r_{ij}$ is a unit vector that points from bead j to bead i. The force decreases linearly with the distance between the beads from a maximum of A_{IJ} at $r_{ij} = 0$ to zero for $r_{ij} \geq R_{c,IJ}$. The force depends parametrically on an amplitude, A_{IJ}, and a range or cutoff distance, $R_{c,IJ}$, that both depend on the types of beads involved. This functional form for the force corresponds to the following contribution

to the energy from the interaction between particles i and j:

$$U_{ij}(r_{ij}) = \begin{cases} \frac{1}{2} A_{IJ} R_{c,IJ} \left(1 - r_{ij}/R_{c,IJ}\right)^2 & r_{ij} < R_{c,IJ} \\ 0 & r_{ij} \geq R_{c,IJ} \end{cases} \tag{3}$$

This energy is repulsive for all interactions between beads. In DPD simulations, an external pressure is applied to establish and maintain a desired mean density.

Beads within a molecule also interact through an intramolecular potential that includes bond stretching $U_b(r_{ij})$ and angle bending $U_a(\theta_{ijk})$ contributions with the following functional forms:

$$\begin{aligned} U_b(r_{ij}) &= \frac{1}{2}\kappa_b (r_{ij} - r_0)^2 \\ U_a(\theta_{ijk}) &= \frac{1}{2}\kappa_a (\theta_{ijk} - \theta_0)^2 \end{aligned} \tag{4}$$

where r_0 and θ_0 are the equilibrium separation between bonded beads and the equilibrium angle between sets of three beads that share two bonds, respectively. The bead model representations used here for the C_nE_m molecules use $\kappa_b = 150$ $(k_B T/R_{\rm DPD}^2)$, for all bonds and $\kappa_a = 5$ $(k_B T/(\text{radian})^2)$, for all angles. $R_{\rm DPD}$ is the DPD unit of length and $k_B T$ is the DPD unit of energy. The equilibrium angles are all $\theta_0 = 180^\circ$. Values for the equilibrium bond lengths, r_0, were determined by considering all atom representations of the molecules in low energy, all trans, conformations. If T beads are placed at the centers of mass of successive -CH_2CH_2- groups and H beads are placed at the centers of mass of successive -OCH_2CH_2- groups, the distance between neighboring T beads is 2.50Å, between neighboring H beads is 3.84Å, and between a T bead and an H bead to which it is bonded is 3.17Å. If we use $R_{\rm DPD} = 5.641$Å (see below) to convert these distances to DPD distance units, they are $0.443 R_{\rm DPD}$, $0.68 R_{\rm DPD}$ and $0.562 R_{\rm DPD}$, respectively.

The contribution to the total potential energy from these conservative interactions is a

sum of such terms, as follows:

$$U_{\text{total}} = \sum_{\text{sites},i<j}^{N} U_{ij} + \sum_{\text{molecules},i} (U_{b,i} + U_{a,i}) \tag{5}$$

where the first sum is over all pairs of beads, and the second sum is over all molecules, with $U_{b,i}$ the total of all bond energies in molecule i, and $U_{a,i}$ the total of all angle energies in molecule i. (Note that all pairs of beads within each molecule *also* are included in the first sum and, therefore, interact through the repulsive DPD potential used for the intermolecular interactions.) For the parameter optimization procedure, we also require the derivative of the total potential energy, U_{total}, with respect to each of the parameters being optimized, $\partial U_{\text{total}}/\partial A_{IJ}$ and $\partial U_{\text{total}}/\partial R_{c,IJ}$ which are easily obtained by differentiating Eqn. 3. For the three bead types used in this work, there are six values of A_{IJ} and six values of $R_{c,IJ}$, corresponding to $IJ = WW, WH, WT, HH, HT, TT$. And there are six corresponding derivatives of the total energy with respect to the force amplitude parameters and six for the range parameters. In this work, the six ranges $R_{c,IJ}$ were held fixed at $R_{c,IJ} = R_{c,WW} = R_{\text{DPD}}$ and were not optimized, so only the six derivatives with respect to the interaction strengths A_{IJ} were computed.

Parameters determined using compound densities

In the early work[16] of Groot and Warren, simulation protocols and parameters were established that were appropriate for DPD simulations of water under ambient conditions. These used DPD beads interacting with $A_{WW} = 25$ (k_BT/R_{DPD}), and used the range of this interaction as the DPD unit of length $(R_{c,WW}/R_{\text{DPD}} = 1)$. Under thermodynamic conditions where this material behaves as a dense liquid like water, the resulting density was $\rho = 3$ (DPD beads)$/R_{\text{DPD}}^3$, under a control pressure in NpT simulations of $P \approx 23.7\ k_BT/R_{\text{DPD}}^3$. These parameters are often used[25] in DPD simulations to represent water and we will follow this practice here, thereby fixing A_{WW} and $R_{c,WW}$. However, one has some flexibility in

deciding how many water molecules such a DPD bead represents in these simulations. It has been standard in our work and that of others[56–58] to let a DPD bead represent *two* water molecules. With this assignment and the actual density of water we can establish the DPD unit of length in our simulations. Using a mass density of 1000 kg/m^3 and a molar mass of 18.02528 g/mole for water, one obtains a number density of $\rho_w = 0.03343$ (water molecules)/Å^3. If this number density for water molecules corresponds to a bead density of $\rho = 3$ beads/$R_{\rm DPD}^3$ and we assign two water molecules per bead, it establishes the DPD length scale as $R_{\rm DPD} = 5.641$Å.

With the DPD length scale so determined, we can use it, along with experimental densities for alkanes, to establish the TT bead interaction parameters, and with experimental densities for polyethylene glycol to establish the HH bead interaction parameters. For the TT interactions we use the experimental mass density of dodecane, $C_{12}H_{26}$, 0.7495 g/cm^3, which implies a number density of $\rho_{C_{12}H_{26}} = 0.002650$ molecules/Å^3 $= 0.4757$ molecules/$R_{\rm DPD}^3$. Using the coarse graining strategy that lets a T bead represent an ethyl unit, dodecane molecules are represented by chains of six T beads, resulting in a desired bead density of 2.854 (DPD T beads)/$R_{\rm DPD}^3$. DPD simulations were performed under the control ForceBalance[22], which iteratively adjusted A_{TT} to obtain this target density. These simulations used $R_{c,TT}/R_{\rm DPD} = 1$ with 32000 molecules of six T beads each and the NpT ensemble with $p = 23.8\ k_BT/R_{\rm DPD}^3$. Each of these simulations was for one million time steps, using a time step size of 0.04 DPD time units. The density equilibrated very early, usually during the first 100000 steps of the simulation, but the average density was computed in each case using data from the last 750000 steps. After a few iterations, ForceBalance converged to $A_{TT} = 37.325$ to yield the target T bead density of 2.854 (DPD T beads)/$R_{\rm DPD}^3$.

For the HH interactions, we use the mass density of 1.128 g/cm^3 for PEG400, a well characterized and commercially available narrow dispersity polymer mixture of molecules with an average molar mass of 400 g/mole that is a liquid under ambient conditions. This implies a number density of $\rho_{\rm PEG400} = 0.001698$ molecules/Å^3 $= 0.3048$ molecules/$R_{\rm DPD}^3$. Molecules

with an average molar mass of 400 g/mole have an average of 9.035 ethyl units, so we would like a value for A_{HH} that gives an average bead density of 2.754 (DPD H beads)/$R_{\rm DPD}^3$. As above for the T bead interactions, DPD simulations were performed under the control ForceBalance, which iteratively adjusted A_{HH} to obtain this target density. These simulations used $R_{c,HH}/R_{\rm DPD} = 1$ with 21333 linear molecules of nine H beads each and the NpT ensemble with $p = 23.8\ k_BT/R_{\rm DPD}^3$. Densities were averaged in each case using the last 750000 steps of one million step simulations, with a time step size of 0.04 DPD time units. After a few iterations, ForceBalance converged to $A_{HH} = 34.193$ to yield the target H bead density of 2.754 (DPD H beads)/$R_{\rm DPD}^3$.

Parameters determined using CMC and N_{agg} data

The remaining three parameters, A_{WT}, A_{WH}, A_{HT}, were iteratively optimized to best reproduce experimental CMC and N_{agg} values. The column labeled Start in Table 4 summarizes the values for A_{WW} and $R_{c,WW} = R_{\rm DPD}$ established from convention, and those for A_{HH} and A_{TT} established along with the use of $R_{c,HH} = R_{c,TT} = R_{\rm DPD}$ to reproduce experimental densities for alkane and polyethylene glycol. Also shown in the column labeled Previous are the values used in our earlier[56] work, where one can see that the older values for A_{HH} and A_{TT} are significantly smaller. We started our iterative optimization of A_{WH}, A_{WT}, and A_{HT} with values that were somewhat increased compared with what was used earlier in order keep them somewhat balanced. The ranges $R_{c,IJ}$ were not optimized.

Table 4: DPD interbead interaction parameters

Bead Type Pair	Previous		Start		Final	
I–J Pair	A_{IJ}	$R_{c,IJ}$	A_{IJ}	$R_{c,IJ}$	A_{IJ}	$R_{c,IJ}$
W–W	25	1.00	25	1.00	25	1.00
H–H	25	1.00	34.193	1.00	34.193	1.00
T–T	25	1.00	37.325	1.00	37.325	1.00
W–H	25	1.00	32.10	1.00	29.49	1.00
W–T	45	1.00	52.15	1.00	51.547	1.00
H–T	30	1.00	37.15	1.00	40.45	1.00

The LogP force field

For comparison, we also performed simulations using a different recently derived[25] DPD model, also developed in our lab for nonionic surfactants. This model uses a slightly different coarse graining strategy (see Figure 2) as well as two additional bead types: one, to describe the terminal methyl unit on the hydrophobic region, and another, to describe the terminal hydroxyl group on the hydrophilic region. The parameters for this model were optimized to reproduce experimental water-octanol partition coefficient data, so we refer to it here as the *LogP* parameter set. The optimization procedure that produced this model adjusted both interaction strengths and the ranges, although the WW interactions were the same as ours. These parameters are included in Table 5 along with ours for comparison in the six places where such a comparison is meaningful. One can see that there are interesting similarities and differences.

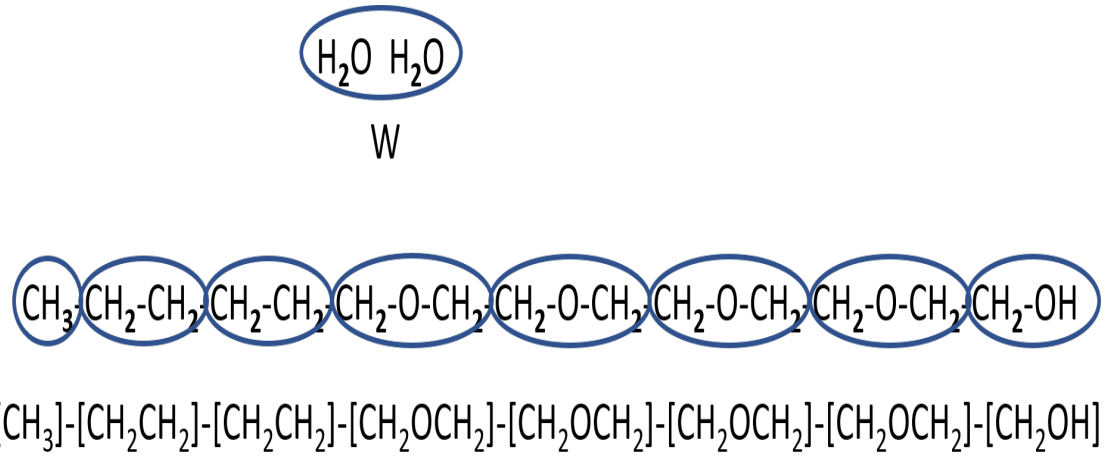

Figure 2: Coarse graining of two water molecules into a W-type bead and C_6E_4 into eight beads of four different types as given by an alternative[25] (LogP) coarse graining scheme, with a methyl and two ethyl beads used for the hydrophobic region and 4 dimethoxy ether beads and a methanol bead used for the hydrophilic region.

Table 5: DPD interbead interaction parameters

Bead Type Pair	LogP		This Work	
I–J Pair	A_{IJ}	$R_{c,IJ}$	A_{IJ}	$R_{c,IJ}$
W–W	25.0	1.0000	25.000	1.00
W–[CH_2OH]	14.5	0.9900		
W–[CH_2CH_2]	45.0	1.0370	51.547	1.00
W–[CH_3]	45.0	0.9775		
W–[CH_2OCH_2]	24.0	1.0580	29.49	1.00
[CH_2OH]–[CH_2OH]	14.0	0.9800		
[CH_2OH]–[CH_2CH_2]	26.0	1.0270		
[CH_2OH]–[CH_3]	26.0	0.9675		
[CH_2OH]–[CH_2OCH_2]	25.0	1.0480		
[CH_2CH_2]–[CH_2CH_2]	22.0	1.0740	37.325	1.00
[CH_2CH_2]–[CH_3]	23.0	1.0145		
[CH_2CH_2]–[CH_2OCH_2]	28.5	1.0950	40.45	1.00
[CH_3]–[CH_3]	24.0	0.9550		
[CH_3]–[CH_2OCH_2]	28.5	1.0355		
[CH_2OCH_2]–[CH_2OCH_2]	25.5	1.1160	34.193	1.00

2.3 Choice of Data for Training and Validation

The entire experimental data set[1] consists of CMC values for twelve compounds (Table 1) as well as mean aggregation numbers (Tables 2 and 3) and particle size distribution functions for them at each of several concentrations at which spherical micelles were thought to form. From this we selected training and validation data sets. The goal was to have enough training data to represent important trends and enough validation data to test for transferability of the resulting model, all while keeping the effort computationally tractable. Data for one of the compounds, $C_{12}E_5$, was omitted because the temperature of the measurements was close to the cloud temperature, and it appeared in the experiments that large aggregates of micelles, rather than isolated micelles, might have been forming. (However, this would be a very interesting case for simulation with a validated force field.) For the C_6E_3, C_8E_4, $C_{10}E_5$ and $C_{12}E_6$ compounds (C_nE_m with n=2m), we noticed there was an unusually strong dependence of micelle size on surfactant concentration, and it also appeared that there were longer tails in the particle size distribution functions. Our criteria (see below) for establishing

the appropriate size of simulations for modeling these compounds suggested extremely large simulations were required. In order to keep the effort tractable we included only the smaller two of these four, C_6E_3 and C_8E_4, omitting $C_{10}E_5$ and $C_{12}E_6$. The compound with the smaller suggested simulation size of these was C_6E_3, and it was included at two concentrations in the training set. C_8E_4 was also included, but only as a validation set compound and at only one concentration. (Compounds in the validation set are only simulated once, using the final optimized force field, whereas those in the training set have to be simulated dozens of times as the force field is iteratively optimized.)

In the C6 series, data from C_6E_5 was used along with that from C_6E_3 as training set data, and data from C_6E_4 (between them) was used for validation. In the C8 series, data from C_8E_5 was used for training, and from C_8E_6 data at one concentration was used for training and a different one for validation. In the C10 and C12 series, the simulation sizes needed were quite large. Data from $C_{10}E_6$ at one concentration was used for training, and a different one for validation. Data from $C_{10}E_8$ and $C_{12}E_8$ were used (two concentrations each) only for validation due to tractability concerns.

The training set, therefore, included representatives of two C6 compounds, two C8 compounds, and one C10 compound, with the C12 compounds excluded from the training set due to tractability concerns, but with $C_{12}E_8$ included in the validation set. We hoped that the training molecules and concentrations selected could adequately serve to capture specific observed trends seen in the target CMC and N_{agg} values as a function of surfactant concentration, and of the hydrophobic and hydrophilic chain lengths. The two surfactant concentrations chosen for each compound were both low enough to expect spherical micelles to be the predominant aggregate, but different enough to capture the general size dependence with respect to concentration seen in the data. The success of the final force field at reproducing several trends is discussed in the Discussion section.

The resulting CMC data consisted of five values for training (C_6E_3, C_6E_5, C_8E_5, C_8E_6, $C_{10}E_6$) and six values for validation (C_6E_4, C_8E_4, C_8E_6, $C_{10}E_6$, $C_{10}E_8$, $C_{12}E_8$). For the

micelle size data, two concentrations were simulated for all compounds (except only one for C_8E_4) resulting in eight training values (two concentrations each for C_6E_3, C_6E_5, C_8E_5, and one each for C_8E_6 and $C_{10}E_6$), and nine validation values (two each for C_6E_4, $C_{10}E_8$, and $C_{12}E_8$, and one each for C_8E_4, C_8E_6, and $C_{10}E_6$). We hoped that the five CMC values and eight $\langle N_{agg} \rangle_M$ values would be sufficient to determine optimal values for the three force field parameters, A_{WT}, A_{WH}, A_{HT}, without overfitting.

2.4 Computed Observables

In this work we develop parameters that can be used in simulations of micelles and are capable of reproducing experimental CMC and N_{agg} values. In this section we explain how these are computed. During each simulation at regular intervals particle coordinates are saved. Similar to previous[56] work, the particle coordinates are analysed by the UMMAP[59] software package that partitions the surfactant molecules into clusters. Two molecules are considered to be in the same cluster if any of the hydrophobic sites of one of them are within a DPD distance unit of any of the hydrophobic sites of the other. The result of such an analysis on a set of coordinates, say from a particular point in time during a simulation, is the number of clusters, and various attributes for each, such as the number and identity of constituent surfactant molecules and various shape attributes. Some of these clusters may consist of only a single molecule. We will represent by $M_{N,i}$ the number of clusters with N surfactant molecules seen in frame i, captured at time $t_i = i\Delta t$ during the simulation over total time T, which has had analyses performed at regular intervals of Δt. Clusters having more constituent molecules than some critical number, n_{cut}, are considered to be micelles. If they are that size or smaller, they are considered to be submicellar-sized clusters, and their constituent molecules are treated as free (i.e., unbound) molecules. The n_{cut} parameter of this analysis depends on the compound, but is approximately 10. (See Tables 2 and 3 and associated discussion.)

The CMC is computed using the time average number of *free* surfactant molecules, given

by the following expressions:

$$CMC = \frac{\langle N_{\mathrm{F}} \rangle}{\langle V \rangle} \tag{6}$$

$$\langle N_{\mathrm{F}} \rangle = (T/\Delta t)^{-1} \sum_{i=1}^{T/\Delta t} \left[\sum_{N \leq n_{cut}} N M_{N,i} \right] \tag{7}$$

The sum is over cluster sizes that are submicellar. All time averages are performed using simulation results after equilibration has been achieved. Since simulations are performed in the NpT ensemble, $\langle V \rangle$ is the mean volume of the simulation cell. Similarly, the time average of the number of surfactant molecules that are in micelles, and the time average of the number of micellar sized clusters are given by the following expressions:

$$\langle N_{\mathrm{Mic}} \rangle = (T/\Delta t)^{-1} \sum_{i=1}^{T/\Delta t} \left[\sum_{N > n_{cut}} N M_{N,i} \right] \tag{8}$$

$$\langle M_{\mathrm{Mic}} \rangle = (T/\Delta t)^{-1} \sum_{i=1}^{T/\Delta t} \left[\sum_{N > n_{cut}} M_{N,i} \right] \tag{9}$$

Note that at all times during the simulation, $N_{\mathrm{F}} + N_{\mathrm{Mic}} = N$, where N is the total number of surfactant molecules. Since N is constant during a simulation, only one (e.g., N_{Mic}) needs to be measured. The *number averaged* mean aggregation number would be given from these by the following expression:

$$\langle N_{agg} \rangle_N = \frac{\langle N_{\mathrm{Mic}} \rangle}{\langle M_{\mathrm{Mic}} \rangle} \tag{10}$$

Of greater interest in this work, however, is the computation of the *mass averaged* mean aggregation number. For this, we note that the fraction, f_N, of the total mass of micellar surfactant molecules that are in clusters with exactly N molecules (where $N > n_{cut}$) is given

by the following expression:

$$f_N = \frac{(T/\Delta t)^{-1}\sum_{i=1}^{T/\Delta t} N M_{N,i}}{(T/\Delta t)^{-1}\sum_{i=1}^{T/\Delta t}\sum_{L>n_{cut}} L M_{L,i}} \tag{11}$$

The numerator is the time average of the number of molecules in clusters of size N, and the denominator is the time average of the number of molecules in *all* micellar clusters, which is equal to $\langle N_{\rm Mic}\rangle$, defined above. The distribution function, f_N, is normalized to unity when summed over all $N > n_{cut}$. The computation of the mass averaged mean aggregation number uses this as the weighting function to yield the following:

$$\langle N_{agg}\rangle_M = \sum_{N>n_{cut}} N\, f_N \tag{12}$$

$$= \sum_{N>n_{cut}} \frac{(T/\Delta t)^{-1}\sum_{i=1}^{T/\Delta t} N^2 M_{N,i}}{(T/\Delta t)^{-1}\sum_{i=1}^{T/\Delta t}\sum_{L>n_{cut}} L M_{L,i}} \tag{13}$$

$$= \langle N_{\rm Mic}\rangle^{-1}(T\Delta t)^{-1}\sum_{i=1}^{T/\Delta t}\left[\sum_{N>n_{cut}} N^2 M_{N,i}\right] \tag{14}$$

In these expressions, we see that the time averages of three quantities are of interest:

$$x_{1,i} = \sum_{N>n_{cut}} M_{N,i} \tag{15}$$

$$x_{2,i} = \sum_{N>n_{cut}} N M_{N,i} \tag{16}$$

$$x_{3,i} = \sum_{N>n_{cut}} N^2 M_{N,i} \tag{17}$$

The time average of x_2 is used to obtain $\langle N_{\rm Mic}\rangle$ (Eqn. 8) and $\langle N_{\rm F}\rangle = N - \langle N_{\rm Mic}\rangle$, which is then used to compute the CMC (Eqns. 6 and 7). The time average of x_1 is used to obtain the average number of micelles, $\langle M_{\rm Mic}\rangle$ (Eqn. 9), and along with $\langle N_{\rm Mic}\rangle$ to obtain the number averaged mean aggregation number, $\langle N_{agg}\rangle_N$ (Eqn. 10). The time average of x_3 is used along with $\langle N_{\rm Mic}\rangle$ to obtain the mass averaged mean aggregation number, $\langle N_{agg}\rangle_M$ (Eqn. 14). From these three time series one also obtains estimates of the variances, time correlation

functions, and resulting statistical uncertainties for the CMC and $\langle N_{agg} \rangle_M$ quantities. The statistical uncertainties are computed as described previously[56] and account for the temporal correlation in the simulation data. Through the cross correlation functions of the x_i time series with the time series that describes the derivatives of the energy with respect to the force field parameters (see Appendix A), one also obtains the derivatives of the CMC and $\langle N_{agg} \rangle_M$ with respect to the force field parameters.

Some of the quantities being computed are ratios of two ensemble averages. Derivatives of these quantities with respect to force field parameters are required for the force field optimization procedure. The following expression gives, for example, the derivative of the mass weighted aggregation number with respect to some force field parameter, α:

$$\frac{\partial \langle N_{agg} \rangle_M}{\partial \alpha} = \frac{\partial \left(\langle X_3 \rangle / \langle X_2 \rangle \right)}{\partial \alpha} \tag{18}$$

$$= \frac{1}{\langle X_2 \rangle} \frac{\partial \langle X_3 \rangle}{\partial \alpha} - \frac{\langle X_3 \rangle}{\langle X_2 \rangle^2} \frac{\partial \langle X_2 \rangle}{\partial \alpha} \tag{19}$$

$$= \langle N_{agg} \rangle_M \left[\frac{1}{\langle X_3 \rangle} \frac{\partial \langle X_3 \rangle}{\partial \alpha} - \frac{1}{\langle X_2 \rangle} \frac{\partial \langle X_2 \rangle}{\partial \alpha} \right] \tag{20}$$

$$\tag{21}$$

The procedure for the computation of derivatives of ensemble averages with respect to the force field parameters is described in Appendix A.

2.5 Objective Function

Systematic force field optimization involves defining a positive definite objective function of the force field parameters which describes the deviation of observables computed in simula-

tions that use the model from target (experimental) values.

$$F(\alpha,\beta,\dots) = w_N \sum_{ij} \left[\frac{N_{ij}^{(sim)}(\alpha,\beta,\dots) - N_{ij}^{(expt)}}{\sigma_{N,ij}} \right]^2 + w_C \sum_{i} \left[\log\left(\frac{C_i^{(sim)}(\alpha,\beta,\dots)}{C_i^{(expt)}} \right) \right]^2 \quad (22)$$

Here we represent by $(\alpha,\beta,\dots)$, the force field parameters that we seek to vary in order to minimize the objective function, F. For notational convenience we use N_{ij} to represent mass weighted mean aggregation numbers, ($\langle N_{agg} \rangle_M$), for compound i measured at concentration j, allowing for the fact that the experimental aggregation numbers are not constant, but tend to grow with concentration for this class of compounds. The superscripts *expt* and *sim* indicate whether the value is obtained from experiment, making it, therefore, part of the training data, or, rather, from a simulation, in which case its value depends on the force field parameters being considered for optimization. The variable $\sigma_{N,ij}$ is an estimate of the experimental uncertainty in the aggregation numbers, which depends on the compound and concentration. This value serves to amplify the contribution to the objective function when the difference between simulation and experiment (training) is larger than this experimental uncertainty. Similarly, we use $C_i^{(expt)}$ and $C_i^{(sim)}$ to represent the experimental values and simulation results for the CMC of compound i, and indicate through arguments on $C_i^{(sim)}$ that the simulation result depends on the choice of the force field parameters. Where there were multiple concentrations simulated for a compound, as there were for most, the CMCs measured at each concentration were averaged to produce a single value for that compound for use in the objective function.

Values for $\sigma_{N,ij}$ were the uncertainties in experimental $\langle N_{agg} \rangle_M$ values from Tables 2 and 3 unless they were less than 5, in which case 5 was used. Given all the assumptions and approximations in the interpretation of the experimental results, we felt that for the purposes of the fitting exercise that uncertainties in the aggregation numbers should be at least this

large.

The use of the log function in the treatment of the CMC contribution to the objective function allows for the fact that the experimental CMC values and their uncertainties for this family of compounds have a very wide dynamic range and supports the goal that we are aiming to minimize relative (i.e., percent) errors in the CMC. (Experimental CMC values for these compounds range over three orders of magnitude and are typically reported with only one significant digit.) Use of the logarithm allows an equal contribution to the objective function to be made when the CMC from the simulation is off by a factor of two, say, from the corresponding experimental value regardless of whether the experimental value itself is nearer to 100 mM or 0.1 mM. On the other hand, experimental aggregation numbers have a much smaller dynamic range (30-200) and uncertainties are fairly constant (of order unity), and so we aim to minimize their absolute errors.

The w_N and w_C parameters are used to weight the relative importance of fitting each type of observable. Since the objective function cannot be optimized to zero, the values of these weights affect the final results. We established a heuristic rule for setting these weights that reflects the degree of experimental uncertainty typical in the observables. Namely, for each term in the summations, if the deviation between the simulation result and the experimental value is comparable to the experimental uncertainty, a value of 10 is contributed to the objective function. Specifically, for the contributions related to aggregation numbers, we assumed $\sigma_N \simeq 5$, a value that roughly captures typical variation in the experimental values among the concentrations for a given compound. Consequently, $w_N = 10$. For the contributions related to the CMC values, we observed *relative* variations in experimental CMCs from the literature (see the Supporting Information in the previous[1] paper) to be approximately 25% of their value for many compounds in this study. If the CMC value from simulation differs by this much from its corresponding experimental value, $C_i^{(sim)} = C_i^{(expt)} + 0.25 C_i^{(expt)}$, this results in a contribution of $w_C[\log(1.25)]^2$ to the objective function. Equating this to 10 determines $w_C = 1065$. That is, with these choices of $w_N = 10$ and

$w_C = 1065$, when the aggregation number from simulation is different from the experimental target value by 5, one gets a comparable contribution to the objective function as when the CMC from simulation differs from the experimental target value by 25%.

The experimental values used in the objective function are listed in Table 1 for the CMC and in Tables 2 and 3 for the $\langle N_{agg} \rangle_M$. Tables 6 and 7 also include these values and indicate which data were used as part of the training data (T) and which were used for validation (V) to determine how well the resulting model performed for compounds and/or concentrations that were not in the training set.

For the parameter optimization itself we made extensive use of an existing force field optimization framework, ForceBalance[22] which we have adapted for use in our micellar based calculations. Our main changes have been to: 1) implement the objective function described above; 2) incorporate analytic derivatives of this objective function with respect to force field parameters, and 3) interface the code with DL MESO[60], the simulation engine used in this work for the DPD simulations. The latter was facilitated by integrating ForceBalance with our locally developed workflow software that manages job submission and control of the highly compute intensive DPD simulations, parses results, invokes the clustering analyses, computes observables (CMCs and aggregation numbers) and their statistical uncertainties and their derivatives with respect to the force field parameters. These simulations need to be performed at multiple concentrations for each compound of the training set. The quantities needed to evaluate the objective function and its gradients are then passed to ForceBalance, which computes them and returns a new set of (hopefully) improved force field parameters. This process is repeated with adequate convergence after 12-15 iterations.

To perform the local minimization we used the quasi-Newton optimizer implemented in ForceBalance with the Gauss-Newton approximation, which constructs the Hessian using first order derivatives of the observables. The BFGS algorithm was also tried but was found to be less effective for our application. A trust radius setting of 1.0 was also used in ForceBalance.

2.6 Simulation Properties

System Sizes

Experimental results[1] suggest there are rather broad distributions of micelle sizes for these nonionic surfactants. Obviously, the size of micellar aggregates seen in a simulation, however, is bounded by the amount of surfactant present in the simulation. Since we are hoping to develop force field parameters for a surfactant model that can reproduce mass weighted aggregation numbers averaged over these broad distributions, it is important that the simulations be large enough to produce in sufficient numbers some of the larger micelles observed in the experimental distributions. In particular, simulations may have to be large enough to be able to form micelles potentially several times larger than suggested by the mass weighted mean aggregation number itself. Similarly, to reproduce the experimental CMC accurately we require the simulations to be large enough to also have a reasonable amount of free surfactant for precise estimates of the free surfactant concentration.

We used the CMC values from the literature, and the mean aggregation numbers and micelle size distributions from the DLS measurements for guidance to establish reasonable sizes for the simulations. For each compound and concentration simulated, we established two different system sizes, a *basic* size used for most of the simulations and a larger *refinement* size used for just the final stages of the parameter optimization. For the basic size, we required simulations to be large enough for each compound and concentration to be able to produce at least (1) five free surfactant molecules based on the target CMC; (2) five micelles with an average size equal to that of the experimental $\langle N_{agg} \rangle_M$; and (3) two micelles with an average size of N_{99}, corresponding to the size for which 99% of the total micellar mass is from micelles of this size or smaller, as indicated by the mass weighted particle size distribution. The refinement sizes were roughly twice as large as the basic sizes and determined as above, but with requirements for *ten* free surfactant molecules and *ten* micelles with an average size equal to the experimental aggregation number.

The last criterion addresses those compounds with a very slowly decaying particle size distribution. Typically, N_{99} is four to six times the value of the mass weighted aggregation number, but the factor can be as large as nine for some compounds. (See Supporting Information for tabulated values of N_{99}.) Depending on the compound and concentration, different of these three criteria establish the minimum reasonable system size, but usually it is requirement (2), to have at least a minimum number of micelles with an average size of $\langle N_{agg} \rangle_M$.

These criteria produced mathematical expressions and the simulation system sizes that are reported in the Supporting Information. Size characteristics of the systems actually simulated, N_{used}, are in a table in Supporting Information. (Sizes are also included as recommendations for compounds and concentrations that were not included in this study.) In order to shorten the parameter optimization process, the strategy was to optimize the force field parameters first using simulations with the smaller basic sized systems, then use the resulting parameter values as starting values in a second optimization that used simulations with the refinement sized systems. Simulations on the validation compounds also used the basic sizes.

We should note here that the surfactant concentrations used in the simulations do not exactly match the ones used in the experiments that yielded the target $\langle N_{agg} \rangle_M$ values (Tables 2 and 3). When constructing systems for simulation one must convert from experimental concentrations expressed in units of molarity to mole fraction, or mass fraction as used in system setup for the simulations. These conversions require assumptions about ideal mixing and require the experimental mass density of the pure surfactants. At early stages of the project we used experimental mass densities for some of the compounds, but could not obtain them for every compound in the entire set. Therefore, we switched to using the densities from the Berthod[27] fit for *all* such concentration conversions. This change caused apparent minor differences (usually less than 6%) between the concentrations used in the experiments to obtain the target data and those of the simulation, once the latter were converted back

to molarity. This difference just reflects experimental uncertainty in the measured mass densities for these compounds. We do not expect these discrepancies to affect our training or validation results. Supporting Information presents a more detailed discussion and a table with the actual concentrations used.

Simulation Parameters

Simulations were performed using the DL_MESO software package in the NpT ensemble using the standard DPD thermostat and a Langevin barostat[61,62], with a control temperature, $T = 1$, and pressure $p = 23.8$, in DPD units. With the potential in use, these conditions correspond roughly to dilute aqueous solutions at 25° C and 1 atm. The time step size for all simulations was 0.04 DPD time units, except for simulations that used the LogP force field, where a time step size of 0.02 was used. (The smaller time step size is more in keeping with what was used in the development of the LogP force field. Since they represent fewer atomic sites, the terminal beads in this model have a smaller mass than the rest, resulting in higher frequency motion. This, in turn, necessitated a smaller time step size to produce comparably accurate simulation trajectories.) The equations of motion were integrated using the Velocity Verlet algorithm[63]. Other conditions are consistent with and described in earlier work[56].

Simulations were run from system setup for a minimum of 120,000 DPD time units, corresponding to 3 million time steps ($dt = 0.04$) for all simulations in both developing and using our force field parameters, and 6 million time steps ($dt = 0.02$) for all simulations using the LogP force field, yielding the same amount of simulated time. Coordinates were saved to disk every 500 time steps. The determination of how much of the beginning of each simulation represented equilibration and what part was used for computing time averages (production) was determined on a case by case basis. Each of the saved coordinates of a simulation was analyzed and the temporal behavior of nine observables was monitored until there was no observed systematic drift in any of their behaviors. These observables were relevant to the

study and included the number of free surfactant molecules (in submicellar-sized clusters) and the number of bound surfactant molecules (in micellar-sized clusters), the maximum micelle size, the per frame number-averaged and weight-averaged cluster size, the sum of the squares of the sizes of all micelle-sized clusters, and the sum of the cubes of the sizes of all micelle-sized clusters. (These last metrics are relevant to detecting equilibration with respect to the numbers of the larger clusters in the size distribution.) For each simulation, the equilibration continued until all observables met the criteria for absence of systematic drift. Therefore, the amount of time spent during equilibration was different for each compound and at each of its concentrations, and for each iteration of the force field optimization. However, the median equilibration length was 79260 DPD time units, and the median production time was 40160 DPD time units. Our earlier work[56] describes this approach in greater detail.

Figure 3 shows a flow chart that gives an overview of the force field parameter optimization process.

3 Results

Representative results of the force field optimization process are shown in Figure 4. Two different optimizations were performed starting with different choices for the three parameters (A_{WH}, A_{WT}, and A_{HT}) being optimized. We observed that within the three dimensional space of the force field parameters being optimized there was a distinct and roughly planar region where micelles formed. On one side of this plane the parameters were such that the surfactant molecules were so soluble that no micelles formed, and on the other side they were so insoluble that the surfactant material phase separated without forming micelles. Within the planar region of parameter space where micelles were produced, different places on the plane produced micelles with different size characteristics. Away from this planar region, the objective function gradients were large and the optimization algorithm readily moved the search to the planar region. However, once on the planar region, the gradients were much

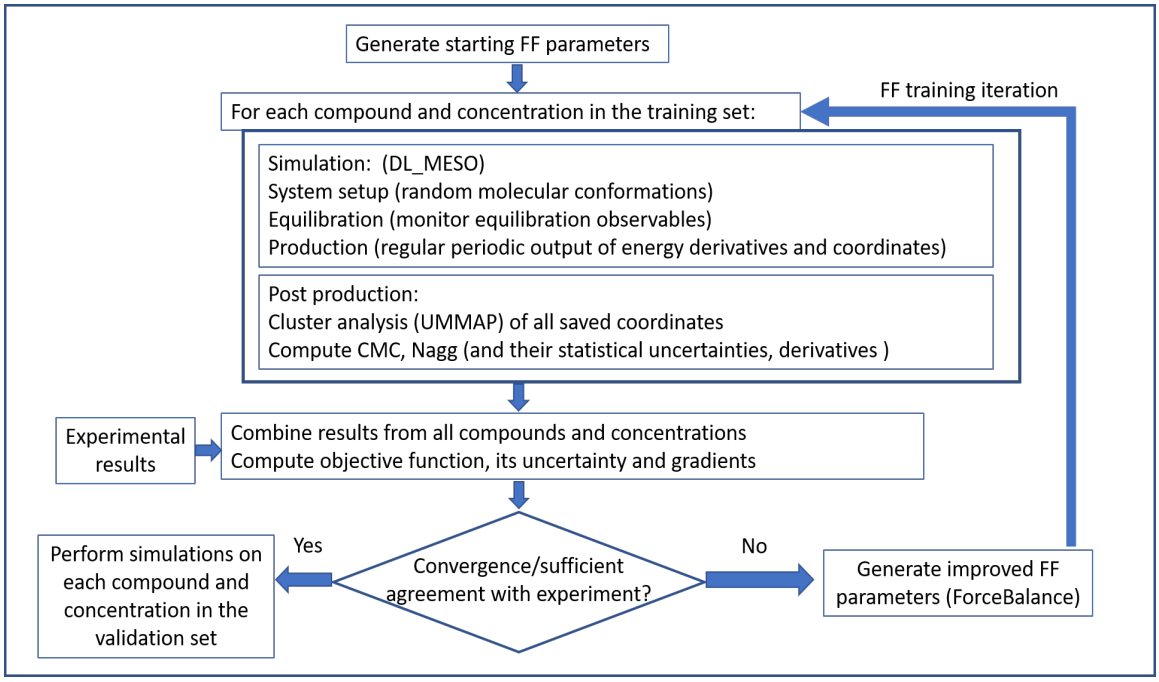

Figure 3: Flow Chart showing iterative optimization of the force field parameters using ForceBalance. Each iteration involves simulations of each molecule and at each of its concentrations in the training set. After analysis, each simulation produces CMC and N_{agg} values, as well as derivatives of these with respect to the force field parameters being optimized. These are used to obtain the objective function and its gradients with respect to the force field parameters.

smaller and the search became more difficult.

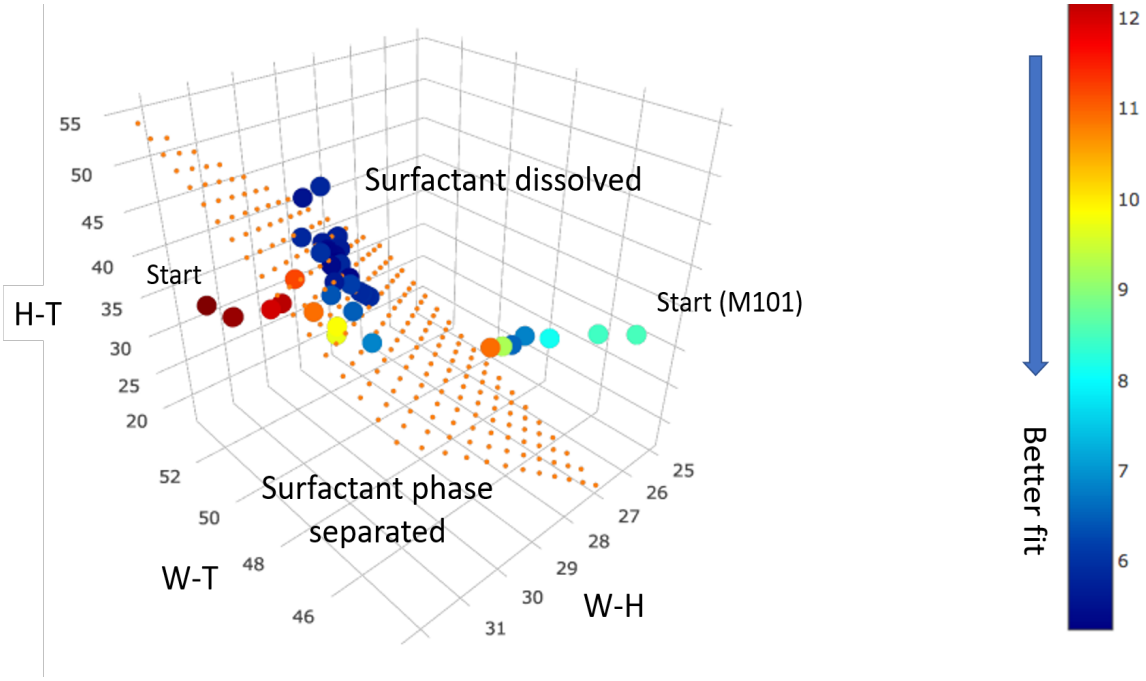

Figure 4: Paths in 3D parameter space for optimizations from two different starting points. Points are color coded based on the value of the objective function. The region of parameter space with relatively low objective function values corresponds to micelle-forming models with roughly correct CMC and $\langle N_{agg} \rangle_M$ trends and falls close to the plane depicted by the orange dots.

The resulting parameters are shown in Table 4 under the column labeled Final. The CMC and $\langle N_{agg} \rangle_M$ results for the training (T) and validation (V) simulations produced using these parameters (This FF) and using the LogP force field parameters (LogP FF) are tabulated in Tables 6 and 7. Also included in these tables is the Target data from experiment and the compound concentrations in the experiments that produced these. The CMC values in Table 6 are averages from simulations of each compound at the concentrations simulated as listed in Table 7.

4 Discussion

The results in Tables 6 and 7 are shown graphically in Figure 5 where simulations results are plotted against experimental results. The CMC results show four clusters of data, repre-

Table 6: Results of CMC obtained with optimized force field parameters

Compound	Model	T/V	Target CMC, wt%	This FF CMC, wt%	LogP FF CMC, wt%
C6E3	H3T3	T	2.35	1.304(0.005)	0.892(0.004)
C6E4	H4T3	V	2.96	1.83(0.02)	1.13(0.02)
C6E5	H5T3	T	3.65	2.26(0.01)	1.38(0.01)
C8E4	H4T4	V	0.246	0.412(0.003)	0.121(0.003)
C8E5	H5T4	T	0.316	0.421(0.007)	0.152(0.003)
C8E6	H6T4	T	0.392	0.51(0.01)	0.179(0.004)
C10E6	H6T5	T	0.038	0.11(0.01)	0.034(0.01)
C10E8	H8T5	V	0.051	0.138(0.001)	0.035(0.001)
C12E8	H8T6	V	0.0059	0.029(0.001)	0.009(0.001)

Table 7: Results of N_{agg} obtained with optimized force field parameters

Compound	T/V	Expt Conc, mM	Expt Conc, wt%	Target $\langle N_{agg} \rangle_M$	This FF $\langle N_{agg} \rangle_M$	LogP FF $\langle N_{agg} \rangle_M$
C6E3	T	128	3.01	81(1)	74(2)	26.7(0.1)
	T	171	4.01	100(2)	97(2)	28.0(0.1)
C6E4	V	270	7.51	46.3(0.1)	55(1)	27.6(0.5)
	V	360	10.01	47.2(0.6)	66(1)	30.2(0.4)
C6E5	T	266	8.56	31.8(0.1)	40.4(0.2)	25.8(0.1)
	T	354	11.4	30.5(0.3)	45.5(0.3)	28.0(0.1)
C8E4	V	24.2	0.744	153(2)	63(3)	37.1(0.4)
C8E5	T	33.0	1.16	65(1)	60(1)	36.2(0.5)
	T	55.0	1.93	77(3)	66.4(0.7)	36.5(0.9)
C8E6	V	29.7	1.17	44.8(0.4)	49(1)	30(1)
	T	49.5	1.96	54(1)	58.1(0.6)	31.9(0.7)
C10E6	V	4.50	0.191	74.4(0.4)	71.5(0.7)	26.81(0.02)
	T	9.00	0.381	75.4(0.3)	64.8(0.2)	23.6(0.8)
C10E8	V	5.24	0.268	66.8(0.4)	51.4(0.1)	24.3(0.3)
	V	10.47	0.536	57(2)	51.3(0.2)	26.6(0.2)
C12E8	V	3.27	0.177	80.9(0.4)	60.8(0.3)	24.2(0.1)
	V	5.45	0.295	81.3(0.3)	60(1)	24.13(0.02)

senting $C_{12}E_8$ on the left (smallest CMC), to $C_{10}E_m$, C_8E_m, and C_6E_m on the right (largest CMC). One can see that the training (black squares) and validation (red triangles) cases do comparably well at tracking the experimental data, but the slope is a bit too small, with the force field producing too low a CMC for the C_6E_m, and too high for the $C_{12}E_m$. We note that the LogP (blue circles) force field also performs very well at tracking the experimental CMC.

As can be seen from Figure 5 the aggregation number results were much more difficult to fit. It is clear that the training set data does a bit better at tracking the experimental results than the validation set data. However, both do much better than the LogP force field, which shows almost no sensitivity with respect to compound or concentration, with all sizes falling in a narrow range near 30.

One data point on Figure 5 that deserves discussion is the $\langle N_{agg} \rangle_M$ outlier for C_8E_4 with the experimental value of 153. Our optimized force field gives a value of 63 ± 3 for this. Our earlier work[1] included an extensive literature survey of experimental results from other workers and found aggregation numbers for this compound in the unusually wide range of 23 to 147, from a variety of different types of experiments. However, there were a couple of very credible values reported near 80. Our own DLS experiments, performed over a wide range of concentrations, showed the sizes for this compound to be a very sensitive function of concentration, possibly accounting for the wide disparity in the literature results. For example, we saw monotonic growth in aggregate size from 1X CMC (0.246 % wt; $\langle N_{agg} \rangle_M = 81$) to 20X CMC (4.92 % wt; $\langle N_{agg} \rangle_M = 380$), with no *plateau* in the size at concentrations just above the CMC, as one might expect, and as we saw in the behavior of most of the other compounds of that study. We did not believe the very large clusters to be spherical micelles, but, perhaps, worm-like micelles or aggregates of spherical micelles. For most of the compounds in that study aggregation numbers from experiments with concentrations in the range of 2X to 3X CMC exhibited a plateau, so we chose to interpret the aggregation numbers for this compound at concentrations of 2.27X CMC (0.558 % wt; $\langle N_{agg} \rangle_M = 110$)

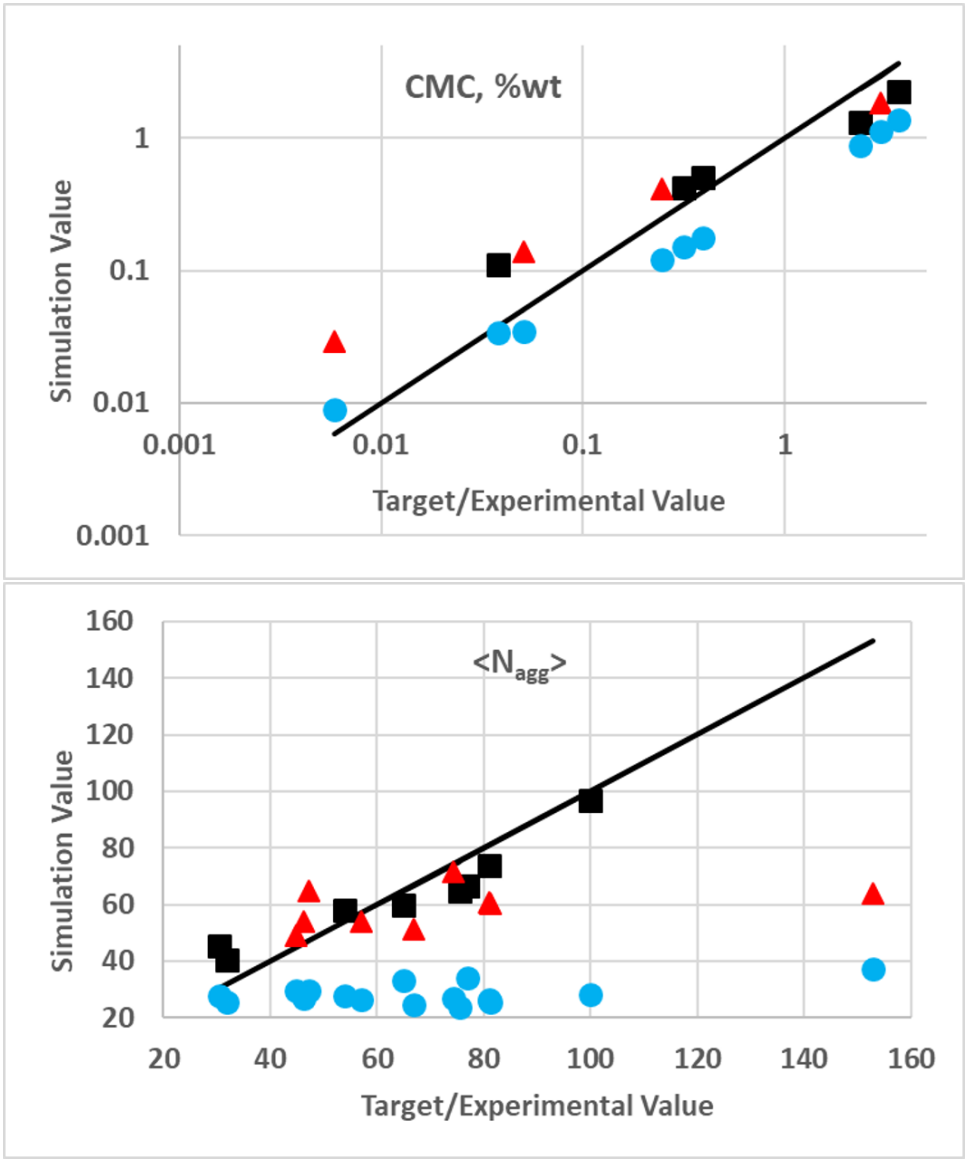

Figure 5: Comparison of simulation and experimental values for CMC and $\langle N_{agg} \rangle_M$ obtained for the training set (black squares), the validation set (red triangles), and from an alternatively optimized (LogP) force field (blue circles). The black lines are for reference and have unit slope and go through the origin in each case. Uncertainty estimates would produce error bars that are comparable to the sizes of the symbols. The outlier in the $\langle N_{agg} \rangle_M$ graph has prompted a reexamination of the experimental result.

and 3.03X CMC (0.744 % wt; $\langle N_{agg} \rangle_M = 153$) as the sizes representative of spherical micelles, but recognized at the time that the lack of a plateau was suspicious. With this in mind, the outlier on the graph could shift significantly to the left, perhaps as far as to correspond with an experimental value of $\langle N_{agg} \rangle_M = 80$, but it will still fall somewhat below the diagonal line.

We actually see this as a very positive result that illustrates the power of this force field and simulations in general. The experimental aggregation number for C_8E_4 was not used in training the model, so the value of $\langle N_{agg} \rangle_M = 63 \pm 3$ is actually a *prediction* that forced a more careful look at the experimental data, and is causing us to adjust our interpretation of the experimental results.

We note that some of the results we report for the LogP force field are in disagreement with what was originally[25] reported for that force field. In Table 6 of the original publication, there are CMC and $\langle N_{agg} \rangle_M$ values for several C_nE_m compounds, four of which are also in our set. The originally reported CMC values for C_6E_4, C_8E_4, $C_{10}E_6$, and $C_{12}E_8$ are, respectively, 2.1 ± 0.7, 0.2 ± 0.1, 0.03 ± 0.03, and 0.006 ± 0.005. These compare favorably with our measurements using the LogP field, which are 1.13 ± 0.02, 0.121 ± 0.003, 0.034 ± 0.01 and 0.009 ± 0.001. In each case the differences are comparable to or less than the uncertainty estimates of the original paper. The original $\langle N_{agg} \rangle_M$ results for C_6E_4 and C_8E_4, 33 ± 7 and 40 ± 5, respectively, also agree pretty well with ours, which fall in the range of $28 - 30$ for the former and near 37 for the later. However, for $C_{10}E_6$ the original work reports an aggregation number of 57 ± 7, whereas we get values in the range of 24 to 27. And for $C_{12}E_8$ the original work reports 64 ± 8, whereas we now get values near 24. Both studies used the UMMAP analysis software and were careful to report mass averaged aggregation numbers. We feel our newer values might be more reliable reflections of the LogP force field. Results in the original paper were produced using simulations at a concentration of 5 (% wt), which is 131X CMC for $C_{10}E_6$ and 850X CMC for $C_{12}E_8$. Our simulations were at much lower relative concentrations corresponding to 5X and 10X CMC for $C_{10}E_6$, and to 31X and 53X

CMC for $C_{12}E_8$. At the higher concentrations we were concerned there might be objects forming other than spherical micelles, such as worm-like micelles or aggregates of micelles, resulting in an artificially large value for the aggregation number. We, therefore, preferred to simulate at lower concentrations, but still significantly above the CMC. Our simulations were also performed using much larger system sizes and longer simulation times. The original simulations all used 325000 DPD beads, whereas our newer ones used 3779130 and 1265623 for the two concentrations of $C_{10}E_6$, and 5055463 and 5055468 for the two concentrations of $C_{12}E_8$. (See Supporting Information.) The equilibration and production simulation times in the original work, 10000 and 20000 DPD time units, respectively, were also much shorter than in our newer simulations. Our simulations were four times longer in overall length, and much of this effort went to equilibration, which we found to be slow for these compounds. Our simulations of $C_{10}E_6$ had equilibration/production times of 111000/9000 DPD time units at the lower concentration and 27140/92860 at the higher concentration. Our simulations of $C_{12}E_8$ had times of 104540/15460 DPD time units at the lower and 111830/8170 at the higher concentration.

Reproducing trends

Finally, we feel that a good model should be able to reproduce several notable *trends* seen in the experimental data set for the C_nE_m family of compounds as a function of surfactant concentration and as a function of the lengths of the hydrophobic (n) and hydrophilic (m) chains. These trends were discussed in the earlier[1] experimental study. Reproducing these trends is important if a model is to be useful for any kind of prediction of micellar properties of new compounds. Even with our limited data set, consisting of only nine compounds simulated at 17 concentrations we can check to see if these trends are observed. The trends seen in the CMC data are 1) that the CMC decreases by an order of magnitude with each addition of two carbon atoms to the hydrophobic block while keeping the length of the hydrophilic block fixed; 2) increasing the length of the hydrophilic block while keeping the

hydrophobic block length fixed leads to a slight increase in the CMC, and 3) the amount of the increase is smaller with increasing hydrophobic length.

With respect to the CMC trends, we note, first, that the force field models do not reproduce the absolute CMC values, often being off by a factor of 2-3. This is true for our optimized force field as well as for the LogP force field (Table 6 and Figure 5). However, the dynamic range of these results is three orders of magnitude, so errors of this size are possibly acceptable if the right trends are observed. Notably, both models get the first trend very well, seeing reduction in the CMC with lengthening of the hydrophobic chain, but the reduction is less than an order of magnitude with the addition of two carbon atoms that is seen in the experimental data. (This is apparent since the slopes in the CMC are less than unity in Figure 5.) The optimized force field shows reductions by factors of 4-5 instead of 10; the LogP force field does a bit better showing reductions by factors of about 6-8. The second CMC trend is also exhibited by both force fields, showing slight increases in the CMC with increases in the length of the hydrophilic chain, while holding the hydrophobic chain length fixed. With respect to the third CMC trend, for both force fields, these CMC increases become less significant with increasing hydrophobic chain length. However, the support for this is weak because the number of comparisons is small and the sizes of the differences are often close to the statistical uncertainties in the CMC values themselves.

The trends seen in the experimental $\langle N_{agg} \rangle_M$ values are 1) that they increase with increasing hydrophobic block length while keeping the length of the hydrophilic block fixed; 2) that they decrease with increasing hydrophilic block length while keeping the hydrophobic block fixed; and 3) there is a tendency for the aggregation numbers to increase for each compound as the surfactant concentration is increased. (This third trend is seen in all of the experimental data, except for minor violations in the cases of C_6E_5 and $C_{10}E_8$.)

With respect to the aggregation number trends, we see that the optimized force field shows the correct first trend, showing increases in aggregation number with increasing hydrophobic chain length while holding the hydrophilic chain length fixed. For the LogP force

field, this trend is seen in only two out of the four cases, passing for the smaller C_6E_m and C_8E_m compounds, but not for the larger $C_{10}E_m$ and $C_{12}E_m$. The optimized force field also shows the desired behavior with respect to the second trend, showing decreases in aggregation number with increasing hydrophilic chain length for a given hydrophobic chain length. The LogP potential misses this trend. The last trend, showing increases in aggregation number for each compound when the surfactant concentration is increased, is seen in nearly all of the results from the optimized force field (except for $C_{10}E_6$), but only in some of the results from the LogP force field.

For some of these compounds we have also performed preliminary investigations using the new force field of the *rate* of aggregation number growth with respect to surfactant concentration. These simulations used a wider range of concentrations than are in the training and validation sets. We have found, fully consistent with our experimental results, that there are much higher rates of growth for C_6E_3 and C_8E_4 (C_nE_m with n=2m) than for the other C6 and C8 compounds.

Reproducing size distribution functions

Two representative micelle size distribution functions from DLS measurements and simulations are shown in Figure 6 for C_6E_3 (experiment concentration 128 mM, or 3.01 wt%) and for $C_{10}E_8$ (experiment concentration 10.47 mM, or 0.536 wt%). Although mass weighted aggregation numbers, which are averages over these distribution functions, are used in the force field training and validation, the distribution functions themselves are not. Therefore, it is possible for the model to reproduce experimental aggregation numbers but not the underlying micelle size distribution. We see that for C_6E_3, there is actually very good qualitative agreement between the simulated and experimental distribution functions, producing mean aggregation numbers of 74 and 81, respectively. However, we see that for the larger $C_{10}E_8$, the agreement is not as good, even though the aggregation numbers are in quite good agreement, at 54.1 and 57. Generally, the agreement in distribution functions is good for the

smaller molecules and gets progressively worse with increasing molecule size, even though the mean aggregation numbers remain in good agreement. One can see for the larger molecule that the simulated distribution function is narrower and centered near its mean, whereas the experimental one is peaked at smaller sizes and achieves nearly the same mean by virtue of a longer more slowly decaying tail, indicating the presence of smaller numbers of very large micelles that are heavily weighted in the mass average aggregation number reported by the DLS measurement. This illustrates that getting good fits to aggregation numbers does not guarantee that the size distribution itself is accurate.

Caveat

A possible challenge with the approach used here for fitting CMC and aggregation number data deserves discussion. We are using the concentration of submicellar clusters as a surrogate for the CMC, as is commonly done, and we are using the properties of the larger clusters as if they were representative micelles. One should recall, however, that micelles would only be expected to form at all if the simulation is being performed at concentrations sufficiently above the CMC *of the force field model*. Normally, this is not a problem, but while one is doing force field optimization, it may be that at times one is using parameters for which the concentration being simulated is below the CMC of the model. In this case, the submicellar concentration will be below the CMC and is not a good surrogate for it. Also, any larger aggregates, if they form at all, are probably not good representative micelles in terms of their shape, stability or, in particular, aggregation number. In this case, the aggregates will likely be too small. Even if one is simulating at a concentration equal to the CMC, it is unlikely that any micelles will form. As a rule of thumb, we feel one should try to be simulating at concentrations of at least twice the CMC of the model. At this concentration, approximately half of the surfactant will be in micellar aggregates and half in free or submicellar aggregates. Of course, one usually does not know the CMC of a model being used *during* a force field optimization effort, but it should be verified at the end of the optimization exercise that

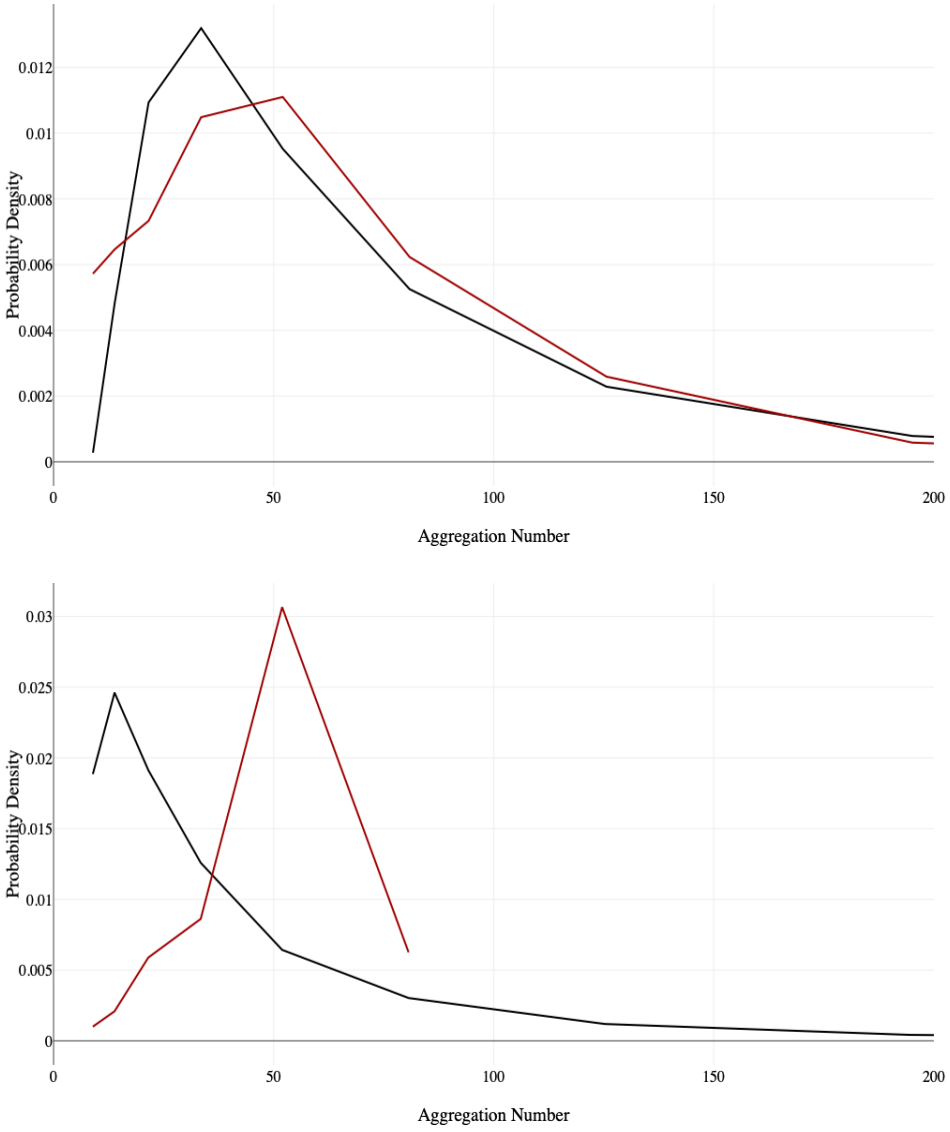

Figure 6: Comparison of micelle size distributions seen in simulation and experiment for C_6E_3 (top figure experiment concentration 128 mM, or 3.01 wt%) and for $C_{10}E_8$ (bottom figure, experiment concentration 10.47 mM, or 0.536 wt%). The figures show distributions from experiment (black) and from simulation (red). The curves represent mass weighted probability densities; $P(n)\,dn$ is the fraction of the total mass of micellar material that is from micelles with sizes between n and $n + dn$. Mass weighted aggregation numbers, which are averages over these distribution functions, are used in the training and validation.

all the simulations for all of the compounds in the test and training sets were performed at sufficiently high concentrations.

Usually, one selects concentrations for the training and validation simulations that are in the safe range of at least twice the experimental CMC, and hopes that after the optimization exercise is complete, the model CMC is close to the experimental CMC. One should be concerned, however, if the CMC of the final force field model ends up being higher than the experimental value. This was the case for six of the nine compounds of our study. The lowest concentrations selected for C_8E_4, C_8E_5, C_8E_6, $C_{10}E_6$, $C_{10}E_8$, and $C_{12}E_8$ were at least three times their respective experimental CMC. However, since the CMCs of the final models ended up being greater than the experimental values, these concentration ended up being, respectively, only factors of 1.9, 2.8, 2.4, 1.8, 1.9, and 6.0 times the CMC of the force field model. Fortunately, even the three lowest of these relative concentrations were high enough that we believe fully formed micelles were were produced in the simulations. This is somewhat corroborated in the cases of $C_{10}E_6$ and $C_{10}E_8$ where simulations were also performed for each at higher concentrations, and consistent aggregation number results were observed. In fact, we performed this kind of test for all of the surfactants of this study with simulations at higher concentrations (e.g., at concentrations of at least twice the CMC of the model) and checked for consistency of the aggregation numbers with those from the training and validation sets.

5 Summary and Conclusions

Using experimental results on micellar properties developed especially for this purpose, we have optimized force field parameters in the context of a coarse grained model of the C_nE_m surfactant molecules and a DPD functional form. These are extreme simplifying assumptions and part of the exercise was to determine if such a simple model could produce satisfactory results. We found that the resulting model captures important trends seen in the experi-

mental data in terms of the behavior of the CMC and aggregation numbers with respect to molecular attributes (lengths of the hydrophobic and hydrophilic components) and surfactant concentration. We compared the optimized force field with the LogP force field produced earlier in our laboratory and found that the new one yielded better micelle size properties. As such, it should be useful as a starting point for the study of micelles where size and shape characteristics are important. We also hope that these parameters are transferable, i.e., useful for the study of other surfactant compounds having chemically similar functional groups: alkane chain components and poly-ethyleneoxide components.

Our force field optimization approach used analytic gradients of the CMC and mean aggregation number simulation observables with respect to the force field parameters being optimized. These were produced during the simulation using analytic expressions, and time averaged, but are, nonetheless, subject to the usual uncertainties due to finite time sampling. The objective function we sought to optimize incorporated both data with experimental uncertainty, and simulation data with statistical uncertainty.

The objective function that was optimized to train the force field balanced the reproduction of experimental CMC data against that for the aggregation number data. We compared our optimized force field with the LogP force field, which was trained to reproduce experimental water-octanol partition coefficient data. We showed that that procedure does a very good job at reproducing experimental CMC values, but is not good at reproducing aggregation number values or trends.

Our study made extensive use of ForceBalance, which performed pretty well. However, we observed several situations where the optimization algorithm got stuck and needed to be restarted. This was often due to the algorithm taking too large of a step in force field parameter space that resulted in a large increase in the objective function. This could happen after several smaller successful steps. Then, sensing that the step was too large caused the algorithm to backtrack and explore its local parameter space with very small steps. Also, it is not surprising that statistical uncertainty in the objective function and the components of

its gradient created challenges to the optimization process, especially near convergence when the gradients and/or successive changes to the objective function were numerically small. We encountered several cases where the optimization algorithm suggested a move in parameter space that appeared to have improved the objective function, but, in fact, produced an increase in the objective function that had been masked by statistical uncertainty. Such situations are difficult for an automated optimization algorithm to detect and to recover from. This might be addressed by a procedure that enforces stronger requirements on the equilibration and statistical sampling in each simulation as convergence is approached in the parameter optimization. We feel that more experience with ForceBalance could help one better control or prevent these situations. Our typical solution was to manually restart the parameter search from the best previous set of values. We are also exploring the use of different optimization algorithms.

Some compounds ($C_{10}E_5$, $C_{12}E_5$, $C_{12}E_6$) included in the experimental study[1] were not included in the parameterization effort because they exhibited interesting or unusual behavior, or because a thorough investigation would require unusually large simulations. We feel it would be fruitful to apply the optimized force field to these compounds. For example, $C_{12}E_5$ at room temperature happens to be close to a phase transition. It would be very interesting to see if a force field trained on molecules and conditions far from a phase transition could, nonetheless, replicate some of the observed behavior for this compound. Similarly, the C_nE_m compounds with $n = 2m$ show pronounced aggregate size growth with respect to increasing concentration and the others do not. Also, for a few of the $C_{12}E_m$ molecules, various experiments are in disagreement about the value of the CMC, and simulations might suggest an explanation for this. Many of these compounds have micellar properties that show interesting temperature-dependent behavior, and it would interesting to see if this model can capture that even though it was not trained to.

Appendix: Derivatives of computed observables with respect to force field parameters

To perform the force field parameter optimization, we need derivatives of the ensemble averages of specific observables with respect to the force field parameters being optimized. These observables include the average number of micellar clusters within a specific size range, or the average number of surfactant molecules that are in clusters of submicellar size. In practice these derivatives can be obtained by finite difference calculations, or by the use the following analytical expressions, where the derivatives are expressed as ensemble averages of cross correlation functions. Consider the canonical ensemble average of an observable, $X(r)$, that is a function of particle coordinates, r, for which we have the following:

$$\langle X \rangle = \frac{\int dr\, X(r)\; e^{-\beta U(r)}}{\int dr\, e^{-\beta U(r)}} \tag{23}$$

where $\beta = 1/k_B T$ is the inverse temperature and $U(r)$ is the potential energy. In this expression, only the potential energy depends on the force field parameters. We use α to represent a generic force field parameter, note that $U = U(r;\alpha)$, and differentiate to obtain the following:

$$\begin{aligned}
\frac{\partial \langle X \rangle}{\partial \alpha} &= \frac{\int dr\, X(r)\; e^{-\beta U(r)} \left(-\beta \frac{\partial U(r)}{\partial \alpha}\right)}{\int dr\, e^{-\beta U(r)}} - \frac{\left(\int dr\, X(r)\; e^{-\beta U(r)}\right)\left(\int dr\, e^{-\beta U(r)} \left(-\beta \frac{\partial U(r)}{\partial \alpha}\right)\right)}{\left(\int dr\, e^{-\beta U(r)}\right)\left(\int dr\, e^{-\beta U(r)}\right)} \\
&= \left\langle -\beta X \frac{\partial U}{\partial \alpha} \right\rangle - \langle X \rangle \left\langle -\beta \frac{\partial U}{\partial \alpha} \right\rangle \\
&= -\beta \left\langle (X - \langle X \rangle) \left(\frac{\partial U}{\partial \alpha} - \left\langle \frac{\partial U}{\partial \alpha} \right\rangle \right) \right\rangle
\end{aligned} \tag{24}$$

This expression reduces the computation of the required derivatives to the evaluation of a fluctuation cross correlation function between the observable of interest and potential energy derivatives. It is trivial to extend the above derivation in the canonical ensemble to the isobaric-isothermal ensemble, which leads to equivalent expressions.

In application, we approximate the ensemble averages in these expressions as time averages over a sufficiently long production (i.e., post equilibration) phase of a simulation:

$$\langle X \rangle \approx \overline{X} = (T/\Delta t)^{-1} \sum_{i=1}^{T/\Delta t} X(r(t_i)) \tag{25}$$

$$\left\langle \frac{\partial U}{\partial \alpha} \right\rangle \approx \overline{\frac{\partial U}{\partial \alpha}} = (T/\Delta t)^{-1} \sum_{i=1}^{T/\Delta t} \frac{\partial U(r(t_i);\alpha)}{\partial \alpha} \tag{26}$$

$$\frac{\partial \langle X \rangle}{\partial \alpha} \approx (-\beta)(T/\Delta t)^{-1} \sum_{i=1}^{T/\Delta t} \left[X(r(t_i)) - \overline{X}\right] \left[\frac{\partial U(r(t_i);\alpha)}{\partial \alpha} - \overline{\frac{\partial U}{\partial \alpha}}\right] \tag{27}$$

Here we have assumed that particle coordinates, $r(t)$, and values for $\partial U(r(t))/\partial\alpha$ have been saved at times $t_i = i\Delta t$, at regular intervals over the length T of the production phase of the simulation. The cluster analysis is applied to each of the $T/\Delta t$ sets of coordinates to give the times series $X(r(t_i))$, where X represents one of the observables needed to compute the CMC and $\langle N_{agg} \rangle_M$ for the compound and composition being considered.

Use of this analytic expression has several advantages over, say, a finite difference approach. First, during a single simulation, derivatives with respect to *all* force field parameters of interest may be evaluated. Second, it is easy to implement in a standard MD software package and imposes almost no computation or storage penalty since the information needed to compute the energy derivatives is already available where the forces are evaluated. Third, if the energy derivatives, $\partial U/\partial\alpha$, are written out at the same time as and along with the coordinates, these gradient expressions may be evaluated even for complex properties that require post processing by an external software package.

We note, however, that the use of this approach necessitates longer simulations than are required to converge the observables themselves in order for the statistical uncertainty on their gradients to be small enough for the optimization algorithm. We have found that early in the optimization process this is not a problem, since the gradients are large compared to their uncertainties. But as one approaches a minimum in the objective function and the gradients get smaller in magnitude, longer simulations may be required to obtain the

required precision in the gradients.

Acknowledgment

The authors wish to thank Dr. David Bray and Dr. James McDonagh for many useful and stimulating scientific discussions and suggestions regarding this work. We thank Dr. Michael Seaton for continued assistance with the use of DL_MESO and his incorporation of modifications in the program to support computation and reporting of derivatives of the potential energy with respect to force field parameters. And we also thank Prof. Lee-Ping Wang for assistance with the use of ForceBalance.

Supporting Information

Supporting information consists of extra discussion of (1) Images of representative micellar aggregates, (2) Equations relating solute concentrations, mole fraction, bead fraction and weight fraction, (3) System sizes used in simulations.

References

(1) Swope, W. C.; Johnston, M. A.; Duff, A. I.; McDonagh, J. L.; Anderson, R. L.; Alva, G.; Tek, A. T.; Maschino, A. P. *J. Phys. Chem. B* **2019**, *123*, 1696–1707.

(2) Tanford, C.; Reynolds, J. A. *Biochim. Biophys. Acta, Rev. Biomembr.* **1976**, *457*, 133–170.

(3) Tanford, C.; Nozaki, Y.; Rohde, M. F. *J. Phys. Chem.* **1977**, *81*, 1555.

(4) Brown, W.; Pu, Z.; Rymden, R. *J. Phys. Chem.* **1988**, *92*, 6086–6094.

(5) Herrington, T. M.; Sahi, S. *J. Colloid Interface Sci.* **1988**, *121*, 107–120.

(6) Shelley, J. C.; Shelley, M. Y. *Curr. Opin. Colloid Interface Sci.* **2000**, *5*, 101–110.

(7) Tieleman, D. P.; van der Spoel, D.; Berendsen, H. J. C. *J. Phys. Chem. B* **2000**, *104*, 6380–6388.

(8) Bruce, C. D.; Berkowitz, M. L.; Perera, L.; Forbes, M. D. E. *J. Phys. Chem. B* **2002**, *106*, 3788–3793.

(9) Garde, S.; Yang, L.; Dordick, J. S.; Paulaitis, M. E. *Mol. Phys.* **2002**, *100*, 2299–2306.

(10) Boek, E. S.; Padding, J. T.; den Otter, W. K.; Briels, W. J. *J. Phys. Chem. B* **2005**, *109*, 19851–19858.

(11) Bandyopadhyay, S.; Tarek, M.; Lynch, M. L.; Klein, M. L. *Langmuir* **2000**, *16*, 942–946.

(12) MacKerell, A. D. *J. Phys. Chem.* **1995**, *99*, 1846–1855.

(13) Bogusz, S.; Venable, R. M.; Pastor, R. W. *J. Phys. Chem. B* **2000**, *104*, 5462–5470.

(14) Piotrovskaya, E. M.; Vanin, A. A.; Smirnova, N. A. *Mol. Phys.* **2006**, *104*, 3645–3651.

(15) Hoogerbrugge, P. J.; Koelman, J. M. V. A. *Europhys. Lett.* **1992**, *19*, 155–160.

(16) Groot, R.; Warren, P. *J. Chem. Phys.* **1997**, *107*.

(17) Moeendarbary, E.; Ng, T.; Zangeneh, M. *Int. J. Appl. Mechanics* **2009**, *01*, 737–763.

(18) Sanders, S. A.; Panagiotopoulos, A. Z. *J. Chem. Phys.* **2010**, *132*, 114902.

(19) Marrink, S. J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750–760.

(20) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.

(21) Shinoda, W.; DeVane, R.; Klein, M. L. *Soft Matter* **2008**, *4*, 2454–2462.

(22) Wang, L.-P.; Martinez, T. J.; Pande, V. S. *J. Phys. Chem. Lett.* **2014**, *5*, 1885–1891.

(23) Wang, L.-P.; Head-Gordon, T.; Ponder, J. W.; Ren, P.; Chodera, J. D.; Eastman, P. K.; Martinez, T. J.; S, P. V. *J. Phys. Chem. B* **2013**, *117*, 9956–9972.

(24) Wang, L.-P.; McKiernan, K. A.; Gomes, J.; Beauchamp, K. A.; Head-Gordon, T.; Rice, J. E.; Swope, W. C.; Martinez, T. J.; Pande, V. S. *J. Phys. Chem. B* **2017**, *121*, 4023–4039.

(25) Anderson, R. L.; Bray, D. J.; Ferrante, A. S.; Noro, M. G.; Stott, I. P.; Warren, P. B. *J. Chem. Phys.* **2017**, *147*, 094503.

(26) Schubert, K.-V.; Strey, R.; Kahlweit, M. *J. Colloid Interface Sci.* **1991**, *141*, 21–29.

(27) Berthod, A.; Tomer, S.; Dorsey, J. G. *Talanta* **2001**, *55*, 69–83.

(28) Corkill, J. M.; Goodman, J. F.; Harrold, S. P. *Trans. Faraday Soc.* **1963**, 202–207.

(29) Borthakur, A.; Zana, R. *J. Phys. Chem.* **1987**, *91*, 5957–5960.

(30) Frindi, M.; Michels, B.; Zana, R. *J. Phys. Chem.* **1992**, *96*, 8137–8141.

(31) Ambrosone, L.; Costantino, L.; D'Errico, G.; Vitagliano, V. *J. Colloid Interface Sci.* **1997**, *190*, 286–293.

(32) Mulley, B. A.; Metcalf, A. D. *J. of Colloid Sci.* **1962**, *17*, 523–530.

(33) Donbrow, M.; Jan, Z. A. *J. Pharm. Pharmacol.* **1963**, *15*, 825–830.

(34) Carless, J. E.; Challis, R. A.; Mulley, B. A. *J. Colloid Sci.* **1964**, *19*, 201–212.

(35) Mukerjee, P.; Mysels, K. J. *Critical Micelle Concentrations of Aqueous Surfactant Systems*; National Standard Reference Data System: Washington, D.C., 1971; Vol. NSRDS-NBS 36.

(36) Wieczorek, S. A. *J. Chem. Thermodyn.* **2000**, *32*, 529–537.

(37) Stubenrauch, C.; Nydén, M.; Findenegg, G. H.; Lindman, B. *J. Phys. Chem.* **1996**, *100*, 17028–17033.

(38) Kato, S.; Harada, S.; Sahara, H. *J. Phys. Chem.* **1995**, *99*, 12570–12575.

(39) Van Ede, J.; Nijmeijer, J. R. J.; Welling-Wester, S.; Örvell, C.; Welling, G. W. *J. Chromatogr. A* **1989**, *476*, 319–327.

(40) Casey, J. R.; Reithmeier, R. A. F. *Biochemistry* **1993**, *32*, 1172–1179.

(41) Corkill, J. M.; Goodman, J. F.; Ottewill, R. H. *Trans. Faraday Soc.* **1961**, *57*, 1627–1636.

(42) Funasaki, N.; Hada, S.; Neya, S. *J. Phys. Chem.* **1988**, *92*, 7112–7116.

(43) Funasaki, N.; Shim, H.-S.; Hada, S. *J. Phys. Chem.* **1992**, *96*, 1998–2006.

(44) Wieczorek, S. A. *J. Chem. Thermodyn.* **1999**, *31*, 661–673.

(45) Ueno, M.; Takasawa, Y.; Miyashige, H.; Tabata, Y.; Meguro, K. *Colloid and Polym. Sci.* **1981**, *259*, 761–766.

(46) Meguro, K.; Takasawa, Y.; Kawahashi, N.; Tabata, Y.; Ueno, M. *J. Colloid Interface Sci.* **1981**, *83*, 50–56.

(47) Takasawa, Y.; Ueno, M.; Sawamura, T.; Meguro, K. *J. Colloid Interface Sci.* **1981**, *84*, 196.

(48) Ueno, M.; Kimoto, Y.; Ikeda, Y.; Momose, H.; Zana, R. *J. Colloid Interface Sci.* **1987**, *117*, 179–186.

(49) Lange, H. Proceedings of the International Congress on Surface Activity. 1960; p 279.

(50) Lange, H. *Kolloid Z. Z. Polym.* **1965**, *201*, 131–136.

(51) Nishikido, N.; Moroi, Y.; Matuura, R. *Bull. Chem. Soc. Jpn.* **1975**, *48*, 1387–1390.

(52) Rosen, M. J.; Cohen, A. W.; Dahanayake, M.; Hua, X. Y. *J. Phys. Chem.* **1982**, *86*, 541–545.

(53) Chen, L.-J.; Lin, S.-Y.; Huang, C.-C.; Chen, E.-M. *Colloids Surf. A* **1998**, *135*, 175–181.

(54) Becher, P. *J. Phys. Chem.* **1959**, *63*, 1675–1676.

(55) Becher, P. *J. Colloid Sci.* **1961**, *16*, 49–56.

(56) Johnston, M. A.; Swope, W. C.; Jordan, K. E.; Warren, P. B.; Noro, M. G.; Bray, D. J.; Anderson, R. L. *J. Phys. Chem. B* **2016**, *120*, 6337–6351.

(57) Groot, R.; Rabone, K. *Biophysical Journal* **2001**, *81*, 725–736.

(58) McDonagh, J. L.; Shkurti, A.; Bray, D. J.; Anderson, R. L.; Pyzer-Knapp, E. O. *J. Chem. Inf. Model.* **2019**, *59*, 4278–4288.

(59) Bray, D. J.; Del Regno, A.; Anderson, R. L. *Molecular Simulation* **2020**, *46*, 308–322.

(60) Seaton, M. A.; Anderson, R. L.; Metz, S.; Smith, W. *Mol. Simul.* **2013**, *39*, 796–821.

(61) Jakobsen, A. F. *J. Chem. Phys.* **2005**, *122*, 124901.

(62) Jakobsen, A. F. *J. Chem. Phys.* **2006**, *125*, 029901.

(63) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637–649.