Journal of Chemical&Engineering Data

Editors' Choice

pubs.acs.org/jced

Article

Machine Learning Prediction of Critical Micellar Concentration Using Electrostatic and Structural Properties as Descriptors

Published as part of Journal of Chemical & Engineering Data special issue "In Honor of Frederico W. Tavares".

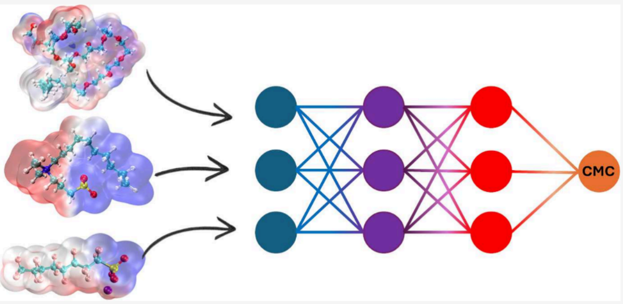
Gabriel D. Barbosa* and Alberto Striolo

Cite This: https://doi.org/10.1021/acs.jced.5c00388

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Understanding and predicting surfactants' critical micelle concentration (CMC) remains a key challenge for the rational design of functional amphiphiles. In this work, we develop a deep learning framework to predict CMCs using quantum chemically derived descriptors, focusing on electrostatic surface potential (ESP) and structural features. We employ a comprehensive temperature-dependent data set comprising over 1300 CMC values across diverse surfactant classes. Fourteen molecular descriptors are extracted via density functional theory (DFT) calculations and used as input, alongside temperature. A fully connected neural network trained on these features yields accurate predictions, achieving performance comparable to previously published graph-based models. To support model interpretability, we explicitly assessed ESP distributions for representative surfactants. SHapley Additive exPlanations (SHAP) and partial dependence analyses reveal that molecular volume, ESP variance, and solvation free energy are the dominant predictors, aligning with established thermodynamic theories. These results demonstrate that DFT-derived electrostatic and geometric descriptors can enable robust and interpretable CMC prediction, offering a physically grounded alternative to black-box models. The methodology and insights presented here may also inform the design of nanostructured soft materials, including surfactant-assisted platforms for hydrogen storage.



■ INTRODUCTION

Surfactants are amphiphilic compounds composed of hydrophilic and hydrophobic segments, enabling them to self-assemble into ordered structures (micelles) that can encapsulate either water or oil, depending on the surrounding medium. This self-assembling propensity governs their interfacial behavior and phase organization, with headgroup and tailgroup structures playing key roles in determining surfactant function, which is responsible for surfactant applications such as manufacturing[1,2] enhanced oil recovery,[3,4] drug delivery, and cosmetics.[5,6] Beyond these traditional roles, surfactants have emerged as key agents in designing and stabilizing soft and porous materials, including those used for hydrogen storage.[7−10]

Numerous empirical and semiempirical models have been developed to understand how molecular structure influences macroscopic properties and aggregation behavior, commonly characterized by the critical micelle concentration (CMC).[11] From a historical perspective, empirical relationships such as the Stauff−Klevens equation have long established a logarithmic decrease in CMC with increasing hydrophobic chain length, underscoring the central role of tail−water interactions in micellization.[12] As highlighted by Nagarajan and Ruckenstein,[13] the pioneering work of Tanford[14,15] provided a geometric and thermodynamic foundation for understanding micelle formation. Israelachvili and co-workers[16,17] introduced the concept of packing parameter to predict aggregate morphologies, based on Tartar and others'[15,18] considerations on the correlations between micelle size and surfactant tail dimensions. However, these models generally could not make quantitative predictions of the aggregation behavior directly from molecular structure and solution conditions. This limitation motivated the development of more rigorous statistical thermodynamic treatments, which began to incorporate chain conformation and interfacial features at the molecular level.[19−21] These foundational studies laid the groundwork for fully predictive molecular thermodynamic models, such as those developed in the seminal works by Puvvada and Blankschtein[22] and Nagarajan and Ruckenstein.[13]

Molecular thermodynamic theories have been updated to enhance predictive performance. For example, Nagarajan[23] explicitly incorporated surfactant tail conformation into the packing parameter formalism. Subsequent developments included numerical improvements, such as the application of

Received: June 17, 2025
Revised: July 15, 2025
Accepted: July 17, 2025

thermodynamic stability criteria, and extensions to model a broader range of amphiphilic molecules,[24] account for ion-specific dispersive effects,[25] describe micelle size distributions,[26] and predict CMC of inverse micelles.[27,28]

From an alternative approach, classical density functional theory (DFT), traditionally applied to inhomogeneous fluids,[29−31] was extended to describe interfacial and self-assembly phenomena through the inhomogeneous Statistical Associating Fluid Theory (iSAFT).[32] This theoretical framework has been applied to model the CMC of ethoxylated surfactants,[33] block copolymer micelles,[34] and the influence of short-chain alcohols on surfactant aggregation.[35] In addition to iSAFT, other second-order thermodynamic perturbation theories have been employed to investigate the interfacial behavior of surfactants,[36−40] although these efforts are for the most part restricted to nonionic surfactants.

As an alternative approach that leverages quantum chemical calculations, continuum solvation models such as COSMO-SAC and COSMO-RS[41−43] have been adapted to predict micellization phenomena by treating the micelle as a distinct phase. For instance, COSMO-RS has been used to estimate CMCs by solving thermodynamic equilibrium conditions.[44] More sophisticated approaches, such as COSMOmic,[45,46] account for micelle internal structure through layer-wise charge distributions obtained from molecular dynamics (MD) simulations. Further developments have overcome the need for explicit simulations by iteratively optimizing micelle structure using a self-consistent framework.[47] Despite their good performance, especially for nonionic surfactants, software availability and computational demands often limit COSMO-based methods.[48]

From a molecular perspective, several studies have employed simulations, in particular molecular dynamics (MD), to investigate the interfacial behavior of amphiphilic molecules.[49−51] Our group has contributed both atomistic and coarse-grained simulations.[52−54] Notably, Jorge[55] employed atomistic simulations to study the self-assembly of n-decyltrimethylammonium bromide, estimating the CMC based on the concentration of free surfactant molecules in solution. While this approach effectively captured aggregation behavior, it was later critiqued by Jusufi and Panagiotopoulos,[56] who argued that relying on the free monomer concentration to predict the CMC may, in some cases, be problematic due to the system-dependent nature of the proposed extrapolation.

Hybrid methodologies have also emerged. A seminal work by Sresht et al.[57] combined MD simulations with molecular thermodynamic theory to compute surface tension isotherms, using MD-derived parameters to inform the thermodynamic model. However, micellization was not incorporated into the underlying phase equilibrium framework. Inspired by this approach, Cárdenas et al.[58] introduced a strategy that accounts for micellization by using structural features at the water−surfactant interface. In parallel, Kanduc et al.[59] proposed a thermodynamically consistent framework that circumvents the time-scale limitations of classical MD by computing transfer free energies via enhanced sampling and alchemical path techniques. This approach enables the prediction of CMCs and adsorption isotherms directly from atomistic models without requiring the explicit observation of micelle formation, thereby providing a rigorous bridge between simulation and experiment. However, this promising approach has only been demonstrated for simple nonionic surfactants so far.

With the enhanced computational capabilities now available, graph neural networks (GNNs) have emerged as promising tools for predicting surfactant properties, including CMCs, offering advantages over traditional physics-based models when sufficient data are available. Recent works have applied standard machine learning models, such as artificial neural networks, random forest, and support vector machine, to predict CMCs of a small set of ionic surfactants in mixed solvent systems using basic structural and solvent descriptors.[60,61] For instance, Qin et al.[62] trained a GNN on experimental data to predict CMCs across multiple surfactant classes and provided interpretable saliency maps linked to molecular features. Building on this CMC database, Moriaty et al.[63] further advanced the previously published GNN by incorporating Gaussian processes for uncertainty quantification, thereby achieving a clearer assessment of the applicability domain. Brozos et al.[64] introduced a particularly valuable data set comprising nearly a thousand experimentally measured CMC values with explicit temperature dependence, spanning a wide range of surfactant classes. Leveraging this comprehensive data, their temperature-aware GNN model improved predictive performance compared to previous approaches.

Machine learning approaches could also be used for integrating physicochemical descriptors into predictive frameworks, which is particularly valuable for extracting mechanistic insights.[65−68] Previous studies have shown that descriptors derived from electrostatic surface potential (ESP) and calculated using density functional theory (DFT) correlate well with the condensed-phase behavior of various compounds.[65,69−73] In our prior work,[71] we found that ESP descriptors capture interfacial behavior trends in a small set of fluorinated and branched surfactants. However, a systematic investigation has not yet been conducted to assess the influence of ESP descriptors on our ability to predict surfactant aggregation, specifically, the CMC. To address this, in this work, the data set from Brozos et al.[64] is leveraged to investigate how ESP descriptors influence CMC, using deep neural network surrogate models. The trained model is analyzed to extract physical insights and identify the key molecular features influencing CMC behavior. Building on prior literature, the model presented here not only performs well with respect to experimental data, but it also helps identify the underlying molecular features driving micellization.

The remainder of this manuscript is organized as follows. First, we describe the methodology used to compute ESP descriptors based on DFT calculations. Next, we present the data set and the machine learning approach used to train the model and interpret CMC behavior. In the results section, we examine the distribution of selected descriptors used as input parameters for the new machine learning model and illustrate their variability using representative surfactants. We then evaluate the predictive performance of the surrogate model and explore the rationale behind its predictions. Finally, we conclude with a summary of key insights obtained from the model.

■ COMPUTATIONAL METHODS

Density Functional Theory Calculations. Considering the high conformational flexibility of the surfactants studied, we first performed an exhaustive conformer search. We adopted the metadynamics-driven Conformer−Rotamer Ensemble Sampling Tool (CREST) of Pracht et al.[74] Starting from an initial structure, metadynamics (MTD) simulations were run at the GFN2-xTB level,[75−78] using the root-mean-square deviation (RMSD) as the biased collective variable. Several bias-parameter pairs (different pushing strengths and widths) ensured broad conformer exploration. Each conformer was energy-minimized, and low-energy fragments were recombined using the built-in genetic crossing (GC) algorithm. Whenever a lower-energy minimum was found, the search was automatically restarted.

Initial 3D geometries were generated from SMILES strings using OpenBabel,[79] and then prefiltered using a genetic algorithm with the Universal Force Field.[80] The lowest-energy seed was refined at the GFN2-xTB level and passed to CREST's complete iMTD-GC workflow with the analytical linearized Poisson−Boltzmann (ALPB) implicit-water solvation.[81] For molecules containing fewer than 200 atoms, we used 500 ps MTD runs. Larger surfactants were treated with 100 ps runs to keep wall times manageable. After removing redundant structures (identified based on energy, RMSD, and rotational constants), the global minimum was selected, and its electrostatic potential (ESP) descriptors were extracted. These values were then used as inputs for neural-network CMC training.

The conformers selected via the algorithm just described were initially optimized using GFN2-xTB, followed by an optimization using the composite approach B97−3c.[82] To incorporate bulk-solvent effects, the geometries were reoptimized at the same B97−3c level using the SMD water continuum model.[83] Finally, single-point energies in aqueous solution were computed using the range-separated hybrid functional ωB97X-V with the def2-TZVP basis set and the SMD implicit solvation model (water), following the protocol of Mariano et al.[84] Calculations conducted at this level of theory have shown consistently strong performance across several molecular properties.[85] All DFT calculations were carried out with the software ORCA 6.0.[86] The computational workflow is summarized in Figure 1.

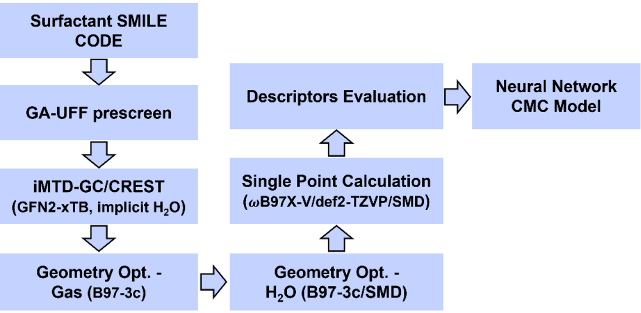

Surfactant SMILE CODE
GA-UFF prescreen
iMTD-GC/CREST (GFN2-xTB, implicit H_2O)
Geometry Opt. - Gas (B97-3c)
Geometry Opt. - H_2O (B97-3c/SMD)
Single Point Calculation (ωB97X-V/def2-TZVP/SMD)
Descriptors Evaluation
Neural Network CMC Model

Figure 1. Computational workflow used to generate descriptors for CMC prediction.

The Multiwfn package[87−89] was used to calculate general interaction property functions (GIPFs).[90] The following descriptors were computed on the van der Waals (vdW) surface and on the electron density isosurface set at 0.001 e/Bohr^3: molecular volume (V_m), surface area (SA), average electrostatic surface potential ($\bar{V}$), positive and negative average ESP values ($\bar{V}^+$ and $\bar{V}^-$), ESP extrema ($V_{\max}$ and $V_{\min}$), total ESP variance (σ^2_{tot}), molecular polarity index (MPI), ESP-defined polar surface area (regions with $|V| > 10$ kcal/mol), the average deviation of the ESP from its surface mean (Π), and $\nu\sigma^2_{tot}$, where ν is the degree of charge balance. A full description of these descriptors can be found elsewhere.[69] The solvation free energy and the HOMO−LUMO energy gap, computed at the ωB97X-V/def2-TZVP/SMD level, was also included as input parameters for the neural network model.

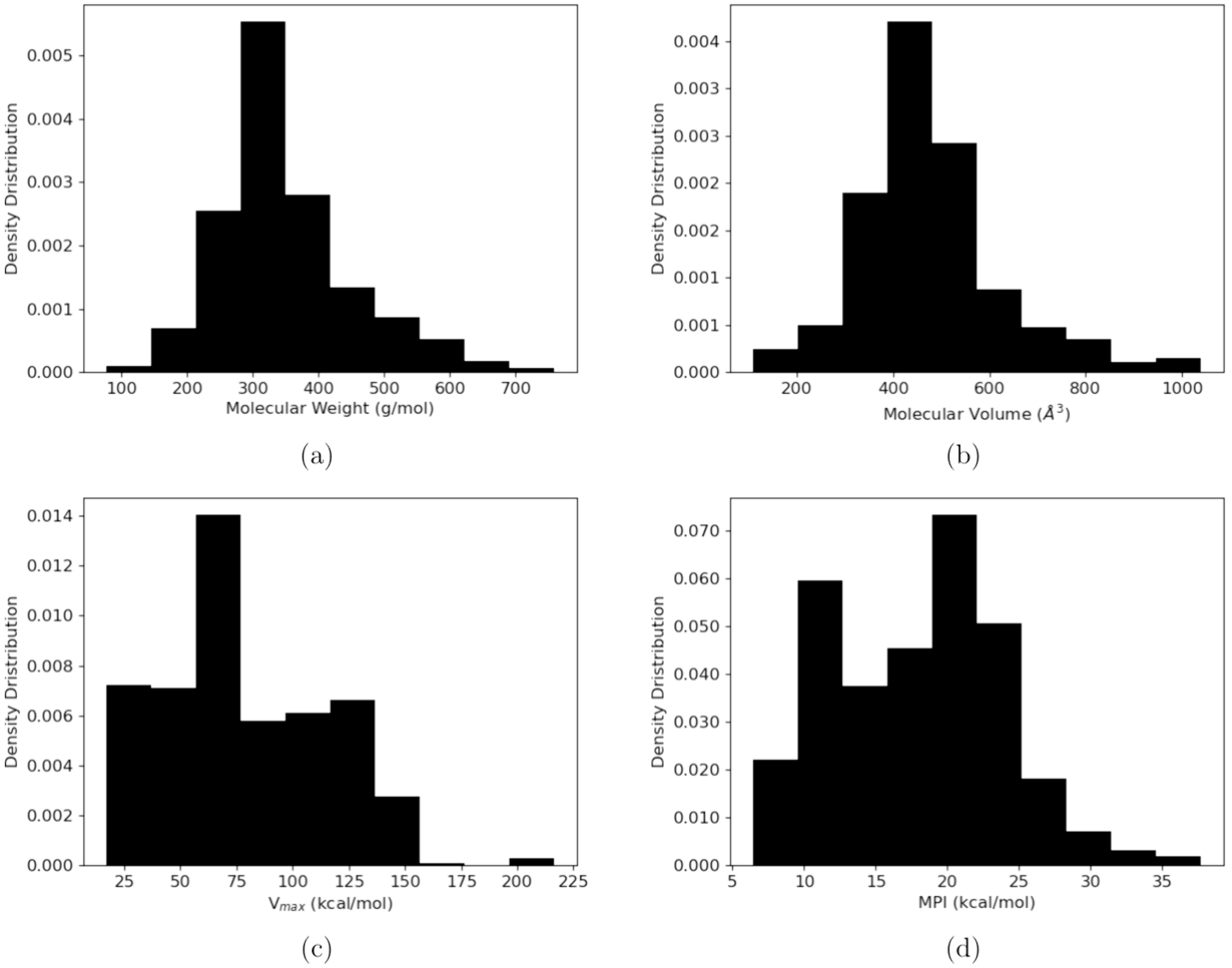
Descriptors derived from quantum chemical calculations, particularly those based on ESP, have been shown to correlate with a wide range of condensed-phase and interfacial properties.[65,69−73] Notably, ESP- and geometry-based descriptors have been successfully used to predict surface tension and speed of sound in ionic liquids through machine learning models trained on DFT-derived features.[65] This broader applicability motivates their use in surfactant modeling as physically meaningful features within data-driven frameworks.

The combination of GFN2-xTB, B97−3c, and ωB97X-V/def2-TZVP/SMD was selected to provide a consistent and computationally efficient framework for modeling both small and bulkier amphiphiles. Each level of theory employed has been extensively validated for a broad range of organic molecules: GFN2-xTB has demonstrated reliable performance in predicting molecular geometries, conformational energetics, and thermochemistry;[91−93] B97−3c has proven effective at producing accurate molecular structures at low computational cost,[82,94] making it well suited for intermediate refinement. For final single-point energy evaluations, the range-separated hybrid functional ωB97X-V, paired with the def2-TZVP basis set and the SMD solvation model, has shown strong predictive power for solvation thermodynamics and noncovalent interactions.[84,94] It is also worth pointing out that, while explicit solvent models are likely to provide more accurate predictions of the properties of individual solvated surfactants, implicit solvent models allow a computationally efficient sampling of the surfactants' properties, thus enabling a GNN model for CMC predictions.

Data Set and Model Training. We employed the experimental database used by Brozos et al.[64] This data set compiles temperature-dependent critical micelle concentration (CMC) measurements for a wide range of surfactants. It contains 1377 CMC values covering 492 unique surfactant structures, including 201 anionic, 171 nonionic, 90 cationic, and 30 zwitterionic species. Each entry includes the isomeric SMILES representation of the surfactant, the corresponding CMC value, and the temperature at which the measurement was performed, spanning a temperature range from 0 to 90 °C. For 227 surfactants, the CMC was measured at multiple temperatures, allowing for exploring temperature-CMC relationships.

As for data splitting, following Brozos et al.,[64] we treated each surfactant−temperature−CMC triplet as an independent data point during model training. Brozos et al. also demonstrated that including CMC measurements at different temperatures for the same surfactant in the training set slightly improves predictive performance when extrapolating to new conditions. In our case, approximately 60% of the data set (840 data points) was used for training, while the remaining 40% (537 data points) was reserved for validation and testing. To assess the robustness and generalization performance of the model, we performed k-fold cross-validation with values of k = 2 through k = 10, following best practices for deep learning regression workflows.[95] This approach allowed us to systematically assess how model stability and predictive performance vary across different data partitioning schemes, hence assessing the generalization capabilities of the new GNN model.

A fully connected neural network was constructed to predict the CMC from molecular descriptors and temperature. The input layer consists of 15 features: 14 ESP-derived or molecular descriptors, plus temperature. The model architecture includes three hidden layers with 32, 64, and 16 neurons, respectively. Each hidden layer uses a Rectified Linear Unit (ReLU) activation function, followed by layer normalization and dropout



(a)

(b)

(c)

(d)

Figure 2. Density distributions of selected molecular properties for the surfactants evaluated in this work: (a) molecular weight, (b) molecular volume (V_m), (c) maximal electrostatic surface potential ($V_{\max}$), and (d) molecular polarity index (MPI).

regularization (dropout rate = 0.0504). The final output layer yields a single scalar value corresponding to the standardized natural logarithm of the CMC. The hyperparameters, including hidden layer sizes, dropout rate, learning rate, and activation function, were optimized using a grid search implemented via Ray Tune,[96] with the test set strictly held out to ensure an unbiased evaluation of model performance.

Model training was performed using the mean squared error (MSE) loss function and the Adam optimizer with a learning rate of 1.59×10^{-4}. Training was conducted over 4000 epochs with a batch size of 32. Input features and target variables were all standardized. Early stopping was not applied; instead, the model was periodically evaluated on the validation set every 10 epochs to monitor overfitting and learning progression. All models were implemented using the PyTorch framework.

Model performance was evaluated using the Mean Absolute Error (MAE) and the coefficient of determination (R^2). To further interpret the influence of each input feature on the predicted CMC values, we employed SHAP (SHapley Additive exPlanations)[66,67] and partial dependence plots (PDPs).[68] A permutation-based SHAP explainer was used to estimate the contribution of each feature to the model output. Both SHAP and PDP analyses were conducted using the SHAP Python package.[66]

■ RESULTS AND DISCUSSION

Surfactant Properties and Electrostatic Surface Potential. In this section, we analyze the distributions of key molecular descriptors for the surfactants included in the data set. The density distributions of selected properties are shown in Figure 2. The molecular weight distribution (Figure 2, panel (a)) is approximately unimodal and skewed toward low values, with a peak around 350 g/mol and a long tail extending up to 700 g/mol. Most of the evaluated surfactants have molecular weights in the range from 250 to 450 g/mol. The lightest molecule in the data set is 1,2-propanediol (a nonionic surfactant),[97] while the heaviest is the ethoxylated surfactant $C_{12}E_{14}$.[98]

A similar trend is observed for the distribution of molecular volume V_m (Figure 2, panel (b)), which peaks near 450 Å^3 and spans from about 200 to 1000 Å^3. Values for maximum ESP ($V_{\max}$, Figure 2(c)) show a broader range, spanning from 20 to 200 kcal/mol, and a weak bimodal distribution suggesting distinct polarization behaviors between different surfactant classes. Values for molecular polarity index MPI (Figure 2(d)) are more evenly spread, with a distribution centered around 20 kcal/mol, which highlights the varying degree of hydrophilicity and charge localization present in the data set.

We further examined the molecular structures of selected surfactants chosen based on the descriptor distributions shown

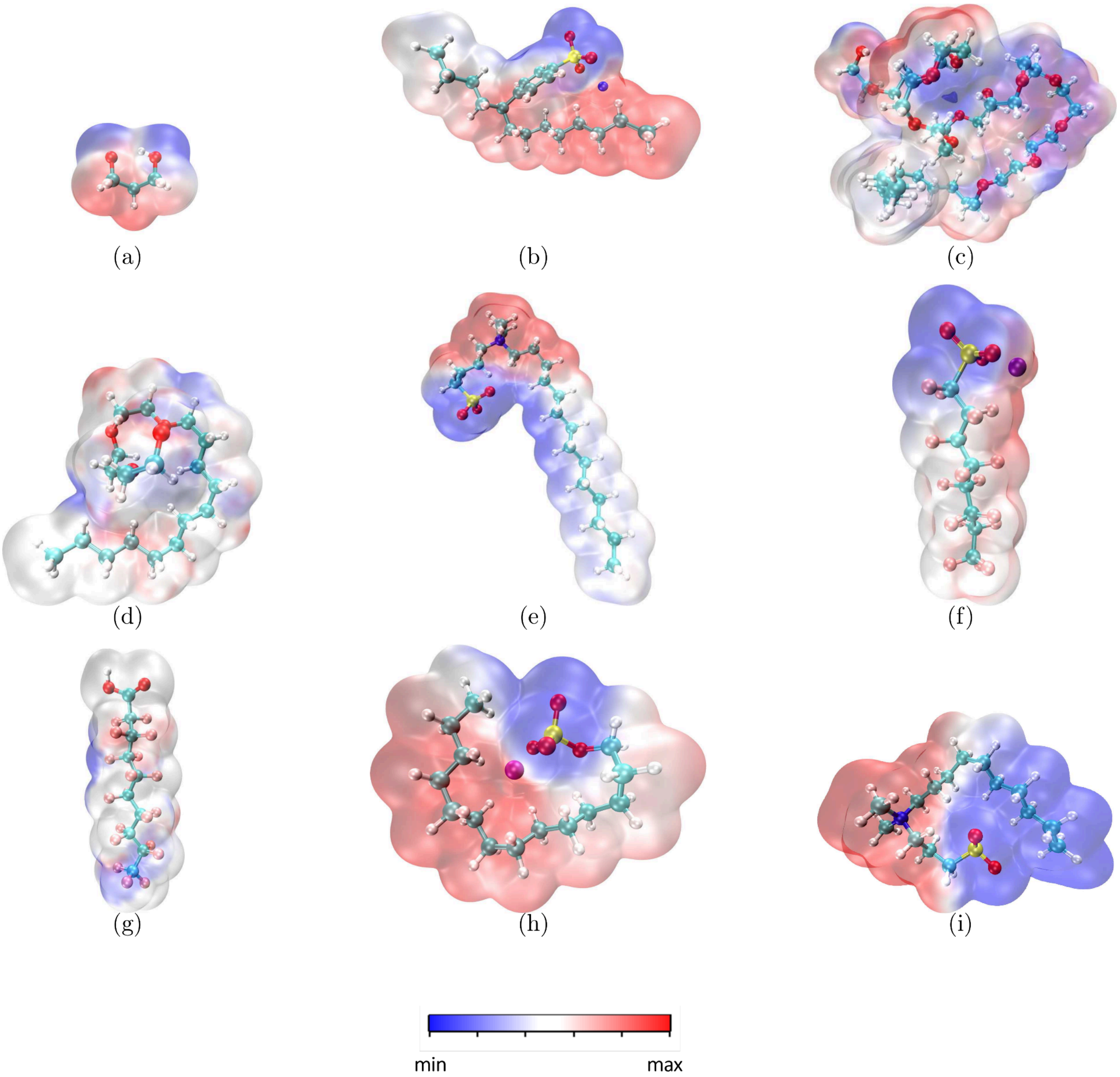


Figure 3. ESP mapped onto the molecular surfaces of selected surfactants (ρ = 0.001 e/Bohr^3, min = −0.03 au and max = 0.03 au). Surfactants were selected to represent the minimum, near-average, and maximum values of three key molecular descriptors: (top row) V_m, (middle row) maximum ESP (V_{max}), and (bottom row) MPI. Surfactant labels: (a) M107, (b) M385, (c) M13, (d) M166, (e) M46, (f) M403, (g) M471, (h) M223, and (i) M21. Color bar code: carbon - cyan; oxygen - red; hydrogen - white; sulfur - yellow; nitrogen - blue; lithium - purple.

Table 1. Summary of the Predicted GIPF Features for the Surfactants Illustrated in Figure 3[a]

Surfactant	V_m (Å^3)	V_{min} (kcal/mol)	V_{max} (kcal/mol)	SA (Å^2)	$\overline{V}$ (kcal/mol)	σ^2_{tot} (kcal/mol)^2	Π (kcal/mol)	MPI (kcal/mol)	Polar SA (%)	$\nu\sigma^2_{tot}$ (kcal/mol)^2
M107	107.8	−47.6	58.2	120.8	2.8	316.1	17.9	18.5	73.1	75.3
M385	476.8	−71.0	131.7	404.4	4.2	981.4	20.2	21.4	67.6	237.3
M13	1040.4	−70.1	44.8	745.0	0.0	303.6	12.7	12.7	47.8	48.6
M166	474.5	−45.2	16.8	391.7	1.9	170.6	7.2	7.8	28.7	16.2
M46	547.1	−100.1	78.1	490.6	−1.1	1419.3	27.1	27.3	56.4	329.8
M403	345.3	−47.9	216.6	308.5	3.2	1407.0	15.8	16.0	37.6	187.1
M471	353.7	−32.4	83.3	317.9	1.0	302.1	6.7	6.4	15.7	36.3
M223	410.7	−57.3	156.1	343.1	3.0	501.5	16.6	17.8	68.3	113.0
M21	428.3	−91.3	79.5	348.0	−2.5	1081.4	37.5	37.7	13.7	268.2

[a] Surfactant codes are used to match entries in the database of Brozos et al.[64] Molecular formulas, CAS numbers, and molecular weights for these compounds are listed in Table 2.

Table 2. Molecular Identifiers Corresponding to the Surfactants Listed in Table 1, Including Label, Name, Molecular Formula, CAS Registry Number, and Molecular Weight[a]

Label	Name	Molecular Formula	CAS number	Molecular Weight (g/mol)
M107	Propane-1,3-diol	$C_3H_8O_2$	504-63-2	76.09
M385	Sodium 4-(tridecan-5-yl) benzenesulfonate	$C_{19}H_{31}O_3SNa$	130462-56-5	362.5
M13	3,6,9,12,15,18,21,24,27,30,33,36,39-tridecaoxahenpentacontan-1-ol	$C_{38}H_{78}O_{14}$	24938-91-8	759
M166	2-(2-(2-(dodecyloxy)ethoxy)ethoxy)ethan-1-ol	$C_{18}H_{38}O_4$	3055-94-5	318.5
M46	4-(dimethyl(tetradecyl)ammonio)butane-1-sulfonate	$C_{20}H_{43}NO_3S$	22313-73-1	377.6
M403	Lithium 1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,8-heptadecafluorooctane-1-sulfonate	$C_8F_{17}LiO_3S$	29457-72-5	506.1
M471	2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,10-nonadecafluorodecanoic acid	$C_{10}HF_{19}O_2$	335-76-2	514.09
M223	Lithium tetradecyl sulfate	$C_{14}H_{29}LiO_4S$	52886-14-3	300.4
M21	3-(decyldimethylammonio)propane-1-sulfonate	$C_{15}H_{33}NO_3S$	15163-36-7	307.5

[a] A complete list of all evaluated surfactants is provided in the Supporting Information.

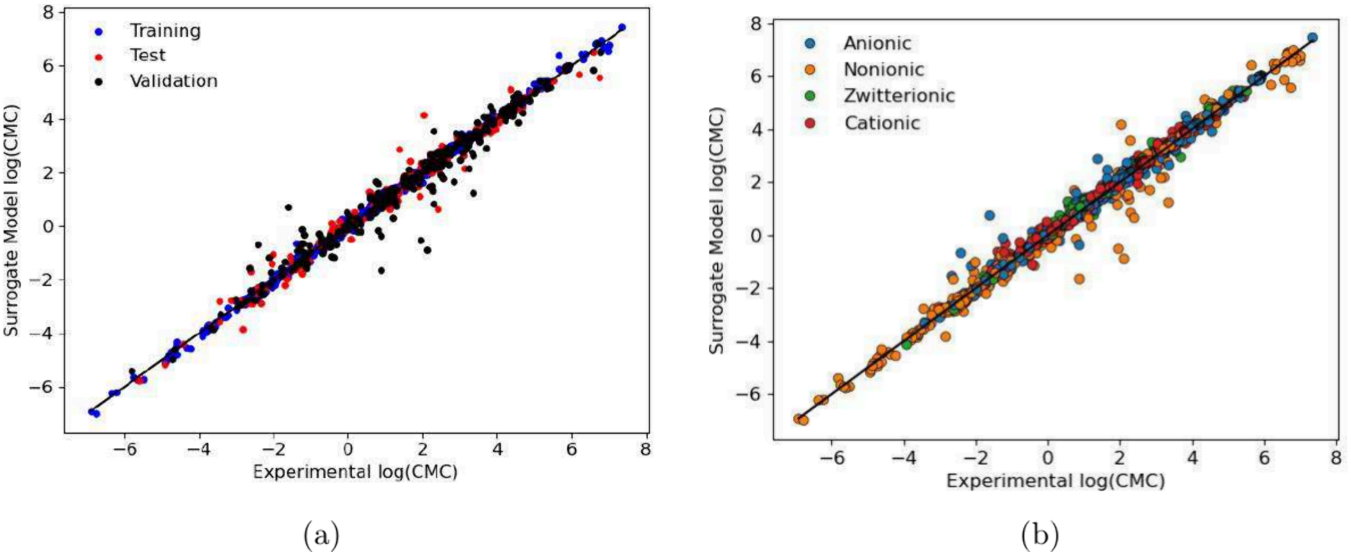


Figure 4. Parity plot for the neural network model trained to predict the critical micellar concentration of the evaluated surfactants: (a) training, test, and validation sets are presented separately; (b) model predictions across the different types of surfactants.

in Figure 2. Figure 3 provides a qualitative visualization of the ESP surfaces for surfactants representing minimum, average, and maximum values of V_m ((a), (b), and (c)), V_{max} ((d), (e), and (f)), and MPI ((g), (h), and (i)). The surfactant with the smallest molecular volume (M107, Figure 3(a)) is a small, nonionic glycol-like molecule. The surfactant with an average molecular volume (M385, Figure 3(b)) is a branched ionic species, whereas the surfactant with the largest volume (M13, Figure 3(c)) corresponds to a long-chain ethoxylated molecule characterized by an extended oxygenated backbone.

Considering V_{max}, the surfactant with the smallest V_{max} (M166, Figure 3(d)) is a nonionic ethoxylated species, where the lower positive potential is likely due to the presence of numerous oxygen atoms. The near-average V_{max} surfactant (M46, Figure 3(e)) is a zwitterionic molecule, characterized by a localized electron-poor region near the cationic nitrogen group. The surfactant with the highest V_{max} (M403, Figure 3(f)) is a fluorinated species that presents a pronounced positive region near the lithium cation.

The surfactant with the lowest MPI (Figure 3(g)) is a fluorinated molecule; consistent with previous studies, which attribute the low polarity of fluorinated surfactants to the low polarizability of fluorocarbon chains, leading to weaker intermolecular forces and distinct hydrophobic−oleophobic characteristics.[53,99−101] A common anionic surfactant represents the near-average MPI value (Figure 3(h)), while the surfactant with the highest MPI (Figure 3(i)) is a zwitterionic molecule. The coexistence of positively and negatively charged groups within the same backbone plays an important role in modulating the CMC behavior of zwitterionic surfactants.[102]

Table 1 summarizes the numerical values of the GIPF descriptors for the selected surfactants. The corresponding molecular formulas, CAS numbers, and molecular weights are provided in Table 2. These values provide a quantitative complement to the ESP surfaces shown in Figure 3. Molecular volume and surface area increase significantly from M107 to M13, ranging from approximately 108 to 1040 Å^3 and 121 to 745 Å^2, respectively. In terms of electrostatic features, M403 shows the highest V_{max}, followed by M223 and M385. Conversely, M46 exhibits the most negative V_{min}, indicating strong localized electron-rich regions.

Surfactants with broad ESP distributions, such as M403 and M46, show the highest ESP variances (σ^2_{tot}) and large $\nu\sigma^2_{tot}$ values, reflecting their pronounced electrostatic heterogeneity. Zwitterionic M21 shows the highest Π (a proxy for charge separation over ESP) value, consistent with its significant spatial charge separation, while fluorinated M471 exhibits the lowest Π (6.7), in line with its weak polarity and uniform surface potential. MPI and polar surface area show similar trends: M21 has the highest MPI and a large polar SA, while M471 exhibits the lowest values for both. Interestingly, the average ESP ($\overline{V}$) values are all relatively close to zero, within a few kcal/mol, in contrast to the much more pronounced extrema observed for V_{min} and V_{max}.

A complete list of the DFT-derived electrostatic and structural descriptors computed for all surfactants is available in the Supporting Information (CSV format).

CMC Surrogate Model. Based on the calculated molecular descriptors discussed above, we now explore how these features correlate with condensed-phase behavior. Specifically, we analyze the performance of the trained model in predicting CMC, aiming to gain a deeper understanding of the molecular features that contribute to the micellization process. The learning curve (Figure S1) of the trained model shows stable and smooth convergence, with both training and validation losses decreasing rapidly within the first few hundred epochs and then stabilizing; no indication of overfitting is observed.

The predictive accuracy of the surrogate model is illustrated in the parity plot shown in Figure 4(a). Predicted CMC values are plotted against experimental values for the training, validation, and test sets. The data points align closely along the diagonal, indicating strong agreement between predicted and reference values across all subsets. Notably, no significant overfitting or systematic bias is observed, and the model performs consistently on unseen data, demonstrating robust generalization.

To contextualize the performance of our surrogate model against previously reported approaches, we compared our results with the recent work by Brozos et al.,[64] who developed a graph neural network (GNN) ensemble to predict CMC values. Their model achieved an RMSE of 0.24 and an R^2 of 0.95 for a similarly sized test set (approximately 218 unseen data points). In comparison, our model, trained on physically motivated molecular descriptors, achieved an RMSE of 0.38, an MAE of 0.27, and an R^2 of 0.95. The comparable R^2 values suggest that the descriptors used here are sufficient to capture the dominant factors influencing micellization. Furthermore, the promising performance achieved, compared to that of other GNN-based models in the literature,[62,63,103] confirm the relevance of electrostatic and structural descriptors for CMC prediction. Consistently, k-fold cross-validation (see Figure S2) confirmed that the model performance, measured by MAE, remains stable across different data splits, indicating that the results are not sensitive to partitioning. Notably, the MAE from our specific train−test split lies within the variability observed across folds. It is worth highlighting that, the smooth convergence of the learning curves, stable cross-validation results, and strong test-set performance collectively indicate no signs of overfitting.

The model performance as assessed across different surfactant classes is illustrated in Figure 4(b). Qualitatively, the model exhibits consistent predictive accuracy for anionic, cationic, nonionic, and zwitterionic surfactants. Quantitatively, class-specific metrics computed over the entire data set are summarized in Table S1. The model performed exceptionally well for zwitterionic and cationic surfactants, achieving low MAEs and high R^2 values. The nonionic class showed slightly higher error, possibly due to the bulkier structures and greater conformational flexibility that are typical for surfactants in this class. While conformer search was applied to mitigate this variability, the broader range of structural motifs in nonionic surfactants likely increases the challenge of accurately predicting the GIPF features.

We analyzed the trained model using SHapley Additive exPlanations (SHAP) to gain deeper insight into the molecular features driving CMC predictions. Grounded in cooperative game theory, SHAP attributes to each input parameter a contribution value that reflects its impact on individual predictions.[104] Applied to the surrogate model developed here, the SHAP analysis identifies which molecular descriptors most strongly influence the predicted CMC values, offering interpretable insights into the physicochemical patterns captured by the surrogate model. In this context, SHAP values quantify how each input feature contributes to the predicted CMC: positive SHAP values indicate that a feature increases the predicted CMC relative to the average prediction, while negative SHAP values lead to negative contributions compared to the average prediction.

Figure 5 shows the SHAP summary plot, which ranks the descriptors by their contribution to the predictions. Each point represents a SHAP value for an individual prediction, with color indicating the magnitude of the corresponding feature: red for high values and blue for low values. Among the features, V_m, Π, and solvation free energy exhibit the most pronounced influence on the predicted CMC values. Overall, the model relies most heavily on global geometric and electrostatic descriptors, while features such as temperature and the HOMO−LUMO gap play a comparatively minor role. It is important to recognize that, even though a few input parameters appear to dominate the outcome, even the features with comparatively low SHAP values are important in determining the CMC, especially given the nonlinearity of the underlying GNN model.

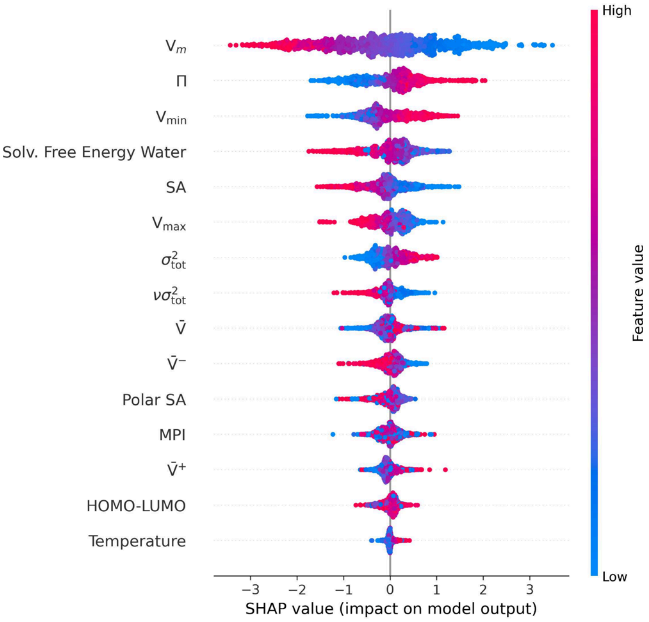
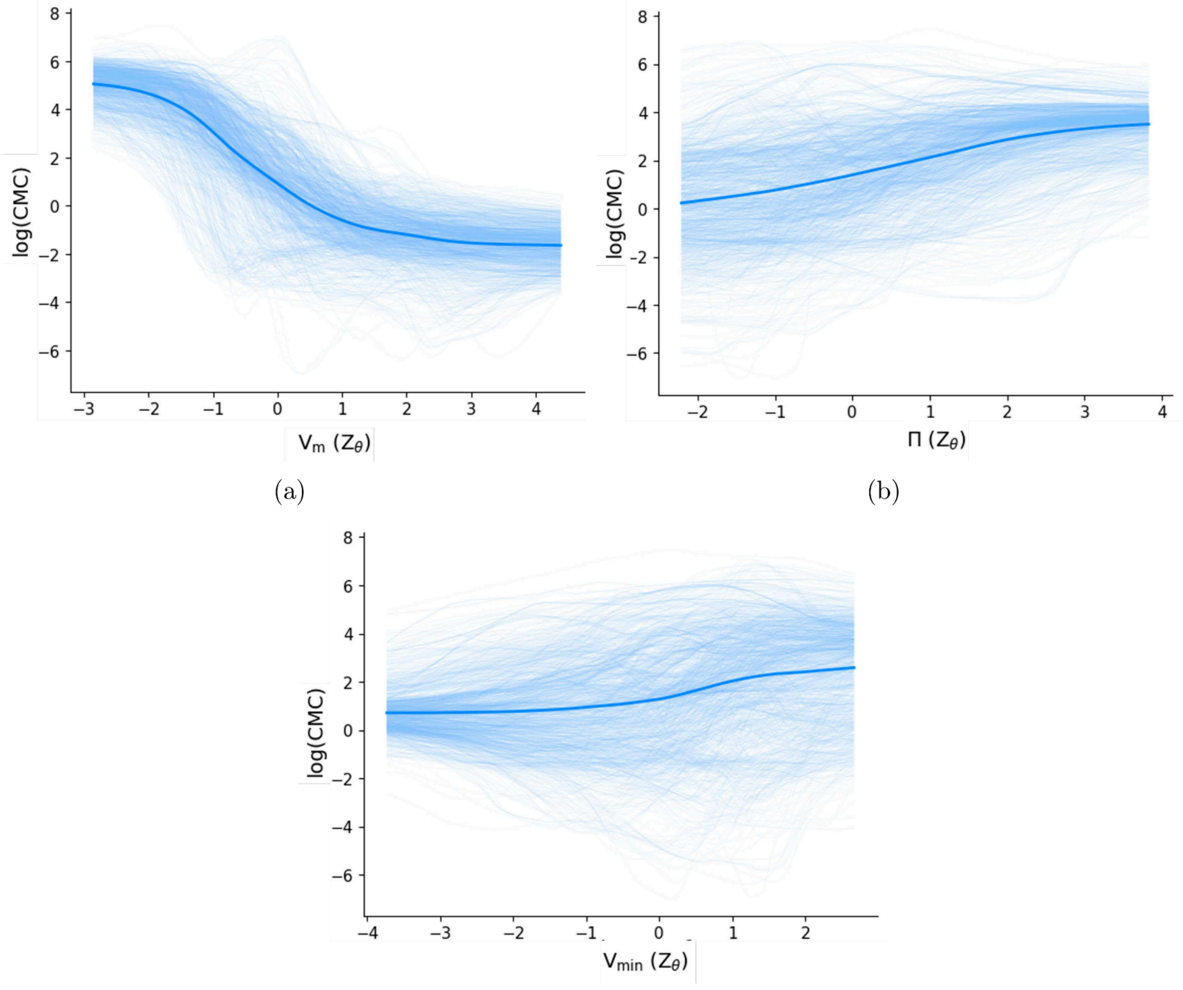


Figure 5. SHAP summary plot showing the impact of each descriptor on the predicted CMC. Features are ranked by their mean absolute SHAP value (top to bottom), with each point representing a single prediction. The color scale indicates the corresponding feature value.

Molecular volume emerges as a key descriptor. This result aligns with classical molecular thermodynamic theory, proposed by Nagarajan and Ruckenstein,[13] in which the volume of the hydrophobic tail plays a central role in determining micelle packing efficiency and aggregate geometry. Subsequent extensions of Nagarajan and Ruckenstein[13] framework have explicitly incorporated surfactant volume exclusion effects, embedding molecular volume terms into free energy formulations and predicting micelle structures.[25,105] The surrogate



(c)

Figure 6. Partial dependence plots (PDPs, dark blue lines) and individual conditional expectation (ICE) curves (light blue lines) showing the marginal effect of (a) V_m, (b) Π, and (c) V_{min}.

model prediction of higher CMC values for surfactants with lower V_m may reflect their diminished amphiphilic character, such as M107, a compact, nonionic glycol-like molecule. This trend is strongly supported by experimental data and molecular thermodynamic theory, which predicts that surfactants with shorter hydrophobic tails exhibit higher CMCs due to the reduced magnitude of the tail transfer free energy, for instance.[13]

The solvation free energy in water also emerges as a meaningful descriptor, with more negative values correlating with higher predicted CMCs. This trend is physically intuitive: highly polar or hydrophilic surfactants tend to be strongly stabilized in the aqueous phase, favoring the dispersed monomeric state over aggregation, yielding higher CMCs. In contrast, surfactants with less favorable solvation energies (i.e., less negative solvation free energy) are more prone to self-assemble, yielding lower CMCs.

The average electrostatic surface potential ($\overline{V}$) showed low SHAP value distribution and flat partial dependence behavior. This outcome is consistent with the physical interpretation of $\overline{\mathrm{V}}$ as a global descriptor that often averages out localized electrostatic variations and typically converges toward near-zero values. Similarly, diminished contributions were observed for $\overline{V}^+$ and $\overline{V}^-$, with SHAP values indicating no clear trend, either positive or negative, on the predicted CMC.

In contrast, the extrema of the electrostatic potential (V_{max} and V_{min}) captured more pronounced and spatially localized effects, both of which significantly impacted model predictions. This trend is consistent with recent simulations showing that the interfacial ESP rises with aggregation.[106] Likewise, the results of Bernardino and Farias de Moura[107] revealed that more negative electrostatic potential regions, especially those arising from close sulfate headgroup arrangements, stabilize ion−headgroup coordination patterns that mirror electrostatic bridging. These findings suggest that local electrostatic extrema may facilitate self-assembly by stabilizing interfacial interactions such as headgroup−headgroup association, counterion coordination, or localized hydrogen bonding.[108]

Among other ESP-based descriptors, Π, defined as the average deviation of the ESP from its surface mean, emerged as one of the most impactful descriptors in the SHAP analysis. As a proxy for internal charge separation or molecular polarity, Π reflects the extent of surface electrostatic contrast. Molecules with larger Π exhibit more pronounced intramolecular electrostatic gradients, enhancing interactions with bulk water and influencing the packing of surfactants in micellar structures. The input feature σ^2_{tot}, which decomposes into positive and negative components, captures the overall heterogeneity of the molecular surface. Higher variance can signal the coexistence of strongly electron-rich and electron-poor regions, favorable conditions for electrostatic complementarity during aggregation. The product $\nu\sigma^2_{\text{tot}}$, a scaled form of the total variance, has previously been proposed as a valuable metric for identifying molecules with strong tendencies to engage in electrostatic self-interactions.[69] In the surrogate model, moderate SHAP contributions from this descriptor support the idea that pronounced electrostatic contrast, even when delocalized, promotes intermolecular association, thereby reducing CMC. Overall, these results highlight the relevance of global charge separation and localized potential extremes in driving surfactant aggregation.

Finally, to explore the influence of individual descriptors on the surrogate GNN model predictions, we computed partial dependent plots (PDPs) and individual conditional expectation (ICE) curves.[68] In general, PDPs reflect the average effect of a feature across the data set, whereas ICE curves reveal instance-level responses and potential heterogeneity in feature influence. In the following, we focus on the three most impactful descriptors identified by SHAP; namely, V_m, Π, and V_{min}.

The PDP and ICE plots in Figure 6 are consistent with the SHAP-derived feature attributions. The input feature V_m shows the most pronounced and coherent effect on log(CMC), with a sigmoidal average trend indicating that increases in V_m strongly reduce the predicted CMC up to a plateau. The descriptors Π and V_{min} also align with the SHAP analysis in terms of their average directionality; however, their corresponding ICE curves are more widely distributed, suggesting that interactions with other features may modulate their effects on CMC or that their contributions are nonlinear and context-dependent.

Finally, it is worth stressing that the surrogate model relies on molecular descriptors extracted from the lowest-energy conformer of each surfactant, selected via CREST-based conformer sampling and DFT refinement. Although only a single representative structure was used per molecule, several key descriptors, such as V_m, Π, and V_{min}, are inherently sensitive to molecular conformation. This sensitivity is reflected in the individual conditional expectation (ICE) curves and partial dependence plots shown in Figure 6, where variation in predicted CMC values arises from changes in these conformation-dependent features. While an exhaustive conformer-averaging approach was computationally infeasible, our results suggest that the selected descriptors capture conformational effects relevant to micellization behavior.

CONCLUSIONS

In this work, we revisited the critical micelle concentration prediction problem using a data-driven approach grounded in quantum chemical descriptors. Leveraging a recent temperature-dependent experimental database containing 1377 data points, we evaluated whether electrostatic surface potential features, computed via density functional theory, capture key physicochemical trends underlying surfactant aggregation.

We first characterized the chemical diversity of the surfactants present in the data set by calculating and analyzing the distribution of descriptors such as molecular weight, volume, maximal ESP, and molecular polarity index. A neural network trained on these features shows very good performance in predicting experimental CMC values across surfactant classes. Interpretability analyses revealed that structural descriptors, particularly molecular volume, play a dominant role in micellization. Thermodynamic properties such as solvation free energy also proved influential and consistent with established molecular thermodynamics theories. The SHAP and PDP/ICE analyses collectively reinforce the conclusion that structural and electrostatic features, particularly those capturing spatial asymmetries and localized potential extremes, play central roles in directing surfactant self-assembly. The consistency between model interpretation tools and known physical principles supports the relevance of the chosen descriptors and suggests that data-driven frameworks rooted in molecular properties can yield accurate and mechanistically meaningful CMC predictions. These insights may directly inform the molecular design of surfactants with tailored aggregation behavior, by revealing how specific electrostatic and geometric features influence micellization. Such understanding enables more rational, structure-based development of functional amphiphiles for applications in formulations, drug delivery, and soft materials.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jced.5c00388.

CSV file with Molecular Modeling Sample Table and quantum descriptors (ZIP)

PDF file with learning curves, k-fold cross-validation results, and model performance by surfactant class (PDF)

AUTHOR INFORMATION

Corresponding Author

Gabriel D. Barbosa − *School of Sustainable Chemical, Biological and Materials Engineering, The University of Oklahoma, Norman, Oklahoma 73019, United States;* orcid.org/0000-0001-7220-9899; Email: gdbarbosa2@ou.edu

Author

Alberto Striolo − *School of Sustainable Chemical, Biological and Materials Engineering, The University of Oklahoma, Norman, Oklahoma 73019, United States;* orcid.org/0000-0001-6542-8065

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jced.5c00388

Author Contributions

Gabriel D. Barbosa: Methodology, simulations, data analysis, and writing—original draft. Alberto Striolo: Methodology evaluation, supervision, funding acquisition, and writing—review and editing.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge the generous allocations of computing resources from the OU Supercomputing Center for Education and Research (OSCER) at the University of Oklahoma. The authors also acknowledge the University of Oklahoma's Office of the Vice President for Research and Partnerships (OVPRP) and the Data Institute for Societal Challenges (DISC) for providing GPU resources. This work was supported, in part, by the US National Science Foundation under Grant No. 2317726 and by the School of Sustainable Chemical, Biological, and Materials Engineering at the University of Oklahoma. AS acknowledges financial support from the Asahi Glass Chair in

Chemical Engineering. We thank BASF Personal Care and Nutrition, particularly Dr. Sandip Bhattacharya and Dr. Christina Kohlmann, for providing the surfactant database used in this study.

REFERENCES

(1) Ateş, A.; Qiao, R.; Lattimer, B. Y. Fuel Resistance of Firefighting Surfactant Foam Formulations. *Fire* **2025**, *8*, 44.

(2) Balasubramaniyan, B.; Pradhan, S. K.; Rajesh, P.; Dash, S. A comparative analysis of the interactions between the anionic sunset yellow dye with cationic, nonionic, and anionic surfactants. *Res. Chem. Intermed.* **2025**, *51*, 2197−2222.

(3) Yang, Y.-H.; Kang, C.-Y.; Liu, T.-F.; Li, H.; Yu, H.-M.; Liu, Z.-Z.; Fan, H.-M. Synergistic stabilization of emulsions by microspheres and surfactants for enhanced oil recovery. *Petroleum Science* **2025**, *22*, 2535−2545.

(4) Li, Q.; Wang, X.; Tang, Y.; Ge, H.; Zhou, X.; Li, D.; Wang, H.; Zhang, N.; Zhang, Y.; Wang, W. Effect of Mixed-Charge Surfactants on Enhanced Oil Recovery in High-Temperature Shale Reservoirs. *Processes* **2025**, *13*, 1187.

(5) Nagaraj, K.; Anbazhagan, G. K.; Govindasamy, R. S.; Muthu, V.; Suraj, R.; Samritha, S.; Muddana, S.; Reddy, E. R.; Sivakumar, D.; Cd, H. S.; et al. Biomimetic surfactants for tunable interfacial properties in drug delivery, biomedical coatings and tissue engineering. *Int. J. Pharm.* **2025**, *677*, 125658.

(6) Luengo, G. S.; Aubrun, O.; Restagno, F. Foams In Cosmetics: New trends in Detergency, Friction, Coatings. *Curr. Opin. Colloid Interface Sci.* **2025**, *77*, 101906.

(7) Sanogo, B.; Dogra, P.; Kalita, K.; Zhang, X. Surfactant-Mediated Interfacial Hydrogen Evolution Reaction. *ACS Appl. Mater. Interfaces* **2025**, *17*, 19512−19525.

(8) Pratthana, C.; Aguey-Zinsou, K.-F. Surfactant induced synthesis of LiAlH4 and NaAlH4 nanoparticles for hydrogen storage. *Applied Sciences* **2022**, *12*, 4742.

(9) Salman, M. S.; Yang, Y.; Zubair, M.; Bedford, N. M.; Aguey-Zinsou, K.-F. Core−shell NaBH4@ Ni Nanoarchitectures: A Platform for Tunable Hydrogen Storage. *ChemSusChem* **2022**, *15*, No. e202200664.

(10) Ren, L.; Li, Y.; Zhang, N.; Li, Z.; Lin, X.; Zhu, W.; Lu, C.; Ding, W.; Zou, J. Nanostructuring of Mg-based hydrogen storage materials: recent advances for promoting key applications. *Nano-Micro Letters* **2023**, *15*, 93.

(11) Israelachvili, J. N. *Intermolecular and surface forces*; Academic Press, 2011.

(12) Klevens, H. Structure and aggregation in dilate solution of surface active agents. *Journal of the American Oil Chemists Society* **1953**, *30*, 74−80.

(13) Nagarajan, R.; Ruckenstein, E. Theory of surfactant self-assembly: a predictive molecular thermodynamic approach. *Langmuir* **1991**, *7*, 2934−2969.

(14) Tanford, C. Theory of micelle formation in aqueous solutions. *J. Phys. Chem.* **1974**, *78*, 2469−2479.

(15) Tanford, C. *The hydrophobic effect: formation of micelles and biological membranes*; J. Wiley & Sons, 1980; Vol. *233*.

(16) Israelachvili, J. N.; Mitchell, D. J.; Ninham, B. W. Theory of self-assembly of hydrocarbon amphiphiles into micelles and bilayers. *Journal of the Chemical Society, Faraday Transactions 2: Molecular and Chemical Physics* **1976**, *72*, 1525−1568.

(17) Israelachvili, J. N.; Mitchell, D. J.; Ninham, B. W. Theory of self-assembly of lipid bilayers and vesicles. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **1977**, *470*, 185−201.

(18) Tartar, H. A theory of the structure of the micelles of normal paraffin-chain salts in aqueous solution. *J. Phys. Chem.* **1955**, *59*, 1195−1199.

(19) Gruen, D. W. A statistical mechanical model of the lipid bilayer above its phase transition. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **1980**, *595*, 161−183.

(20) Dill, K. A.; Cantor, R. S. Statistical thermodynamics of short-chain molecule interphases. 1. theory. *Macromolecules* **1984**, *17*, 380−384.

(21) Shaul, B. A.; Gelbart, W. M. Theory of chain packing in amphiphilic aggregates. *Annu. Rev. Phys. Chem.* **1985**, *36*, 179−211.

(22) Puvvada, S.; Blankschtein, D. Molecular-thermodynamic approach used to predict micellization, phase behavior, and phase separation of micellar solutions. *Langmuir* **1990**, *6*, 894−895.

(23) Nagarajan, R. Molecular packing parameter and surfactant self-assembly: the neglected role of the surfactant tail. *Langmuir* **2002**, *18*, 31−38.

(24) Moreira, L. A.; Firoozabadi, A. Thermodynamic modeling of the duality of linear 1-alcohols as cosurfactants and cosolvents in self-assembly of surfactant molecules. *Langmuir* **2009**, *25*, 12101−12113.

(25) Lukanov, B.; Firoozabadi, A. Specific ion effects on the self-assembly of ionic surfactants: a molecular thermodynamic theory of micellization with dispersion forces. *Langmuir* **2014**, *30*, 6373−6383.

(26) Santos, M.; Tavares, F.; Biscaia Jr, E. Molecular thermodynamics of micellization: micelle size distributions and geometry transitions. *Brazilian Journal of Chemical Engineering* **2016**, *33*, 515−523.

(27) Moreira, L. A.; Firoozabadi, A. Molecular Thermodynamic Modeling of Droplet-Type Microemulsions. *Langmuir* **2012**, *28*, 1738−1752.

(28) Lukanov, B.; Firoozabadi, A. Molecular thermodynamic modeling of reverse micelles and water-in-oil microemulsions. *Langmuir* **2016**, *32*, 3100−3109.

(29) Sermoud, V.; Barbosa, G.; Soares, E.; Barreto, A.; Tavares, F. Exploring the multiple solutions of the classical density functional theory using metadynamics based method. *Adsorption* **2021**, *27*, 1023−1034.

(30) Klink, C.; Gross, J. A density functional theory for vapor−liquid interfaces of mixtures using the perturbed-chain polar statistical associating fluid theory equation of state. *Ind. Eng. Chem. Res.* **2014**, *53*, 6169−6178.

(31) Wu, J.; Jiang, T.; Jiang, D.-e.; Jin, Z.; Henderson, D. A classical density functional theory for interfacial layering of ionic liquids. *Soft Matter* **2011**, *7*, 11222−11231.

(32) Tripathi, S.; Chapman, W. G. Microstructure of inhomogeneous polyatomic mixtures from a density functional formalism for atomic mixtures. *J. Chem. Phys.* **2005**, *122*, DOI: 10.1063/1.1853371.

(33) Wang, L.; Haghmoradi, A.; Liu, J.; Xi, S.; Hirasaki, G. J.; Miller, C. A.; Chapman, W. G. Modeling micelle formation and interfacial properties with iSAFT classical density functional theory. *J. Chem. Phys.* **2017**, *146*, DOI: 10.1063/1.4978503.

(34) Xi, S.; Wang, L.; Liu, J.; Chapman, W. Thermodynamics, microstructures, and solubilization of block copolymer micelles by density functional theory. *Langmuir* **2019**, *35*, 5081−5092.

(35) Lu, J.; Gonzalez de Castilla, A.; Muller, S.; Xi, S.; Chapman, W. G. Dualistic Role of Alcohol in Micelle Formation and Structure from iSAFT Based Density Functional Theory and COSMOplex. *Industrial & engineering chemistry research* **2023**, *62*, 1968−1983.

(36) Li, X.-S.; Lu, J.-F.; Li, Y.-G.; Liu, J.-C. Studies on UNIQUAC and SAFT equations for nonionic surfactant solutions. *Fluid phase equilibria* **1998**, *153*, 215−229.

(37) Li, X.-S.; Lu, J.-F.; Li, Y.-G. Study on ionic surfactant solutions by SAFT equation incorporated with MSA. *Fluid phase equilibria* **2000**, *168*, 107−123.

(38) Rehner, P.; Bursik, B.; Gross, J. Surfactant modeling using classical density functional theory and a group contribution PC-SAFT approach. *Ind. Eng. Chem. Res.* **2021**, *60*, 7111−7123.

(39) Rother, M.; Sadowski, G. Hydrophobic interactions described using hetero-segmented PC-SAFT: 1. Alcohol/water mixtures. *Fluid Phase Equilib.* **2024**, *582*, 114102.

(40) Rother, M.; Sadowski, G. Hydrophobic interactions described using hetero-segmented PC-SAFT: 2. Surfactants and their aqueous solutions. *Fluid Phase Equilib.* **2025**, *593*, 114342.

(41) Klamt, A. Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. *J. Phys. Chem.* **1995**, *99*, 2224−2235.

(42) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. Refinement and parametrization of COSMO-RS. *J. Phys. Chem. A* **1998**, *102*, 5074−5085.

(43) Lin, S.-T.; Sandler, S. I. A priori phase equilibrium prediction from a segment contribution solvation model. *Industrial & engineering chemistry research* **2002**, *41*, 899−913.

(44) Turchi, M.; Karcz, A.; Andersson, M. First-principles prediction of critical micellar concentrations for ionic and nonionic surfactants. *J. Colloid Interface Sci.* **2022**, *606*, 618−627.

(45) Klamt, A.; Huniar, U.; Spycher, S.; Keldenich, J. COSMOmic: a mechanistic approach to the calculation of membrane- water partition coefficients and internal distributions within membranes and micelles. *J. Phys. Chem. B* **2008**, *112*, 12148−12157.

(46) Jakobtorweihen, S.; Yordanova, D.; Smirnova, I. Predicting critical micelle concentrations with molecular dynamics simulations and COSMOmic. *Chemie Ingenieur Technik* **2017**, *89*, 1288−1296.

(47) Klamt, A.; Koch, L.; Terzi, S.; Huniar, U.; Schwöbel, J.; Gaudin, T. COSMOplex: self-consistent simulation of self-organizing inhomogeneous systems based on COSMO-RS. *Phys. Chem. Chem. Phys.* **2019**, *21*, 9225.

(48) Herbert, J. M. Dielectric continuum methods for quantum chemistry. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2021**, *11*, No. e1519.

(49) Silva, G. M.; Pérez-Sanchéz, G.; Pantano, D. A.; Loehlé, S.; Coutinho, J. A. Using the SAFT-γ-Mie to Generate Coarse-Grained Force Fields Useable in Molecular Dynamics Simulations: Describing the Micellar Phases of Polyalkylglycols in Aqueous Solutions. *Ind. Eng. Chem. Res.* **2023**, *62*, 5658−5667.

(50) Pérez-Sánchez, G.; Costa, F. M.; Silva, G. M.; Piñeiro, M. M.; Coutinho, J. A. Coarse-grain molecular dynamics simulation framework to unravel the interactions of surfactants on silica surfaces for oil recovery. *Colloids Surf., A* **2023**, *670*, 131583.

(51) Wang, P.; Misra, R. P.; Zhang, C.; Blankschtein, D.; Wang, Y. Surfactant-Aided Stabilization of Individual Carbon Nanotubes in Water around the Critical Micelle Concentration. *Langmuir* **2024**, *40*, 159−169.

(52) Kobayashi, T.; Kotsi, K.; Dong, T.; McRobbie, I.; Moriarty, A.; Angeli, P.; Striolo, A. The solvation of Na+ ions by ethoxylate moieties enhances adsorption of sulfonate surfactants at the air-water interface. *J. Colloid Interface Sci.* **2025**, *682*, 924−933.

(53) Barbosa, G. D.; Tavares, F. W.; Striolo, A. Molecular Interactions of Perfluorinated and Branched Fluorine-Free Surfactants at Interfaces: Insights from a New Reliable Force Field. *J. Chem. Theory Comput.* **2024**, *20*, 7300−7314.

(54) Khedr, A.; Striolo, A. DPD parameters estimation for simultaneously simulating water−oil interfaces and aqueous nonionic surfactants. *J. Chem. Theory Comput.* **2018**, *14*, 6460−6471.

(55) Jorge, M. Molecular dynamics simulation of self-assembly of n-Decyltrimethylammonium Bromide micelles. *Langmuir* **2008**, *24*, 5714−5725.

(56) Jusufi, A.; Panagiotopoulos, A. Z. Explicit-and implicit-solvent simulations of micellization in surfactant solutions. *Langmuir* **2015**, *31*, 3283−3292.

(57) Sresht, V.; Lewandowski, E. P.; Blankschtein, D.; Jusufi, A. Combined molecular dynamics simulation−molecular-thermodynamic theory framework for predicting surface tensions. *Langmuir* **2017**, *33*, 8319−8329.

(58) Cárdenas, H.; Kamrul-Bahrin, M. A. H.; Seddon, D.; Othman, J.; Cabral, J. T.; Mejía, A.; Shahruddin, S.; Matar, O. K.; Müller, E. A. Determining interfacial tension and critical micelle concentrations of surfactants from atomistic molecular simulations. *J. Colloid Interface Sci.* **2024**, *674*, 1071−1082.

(59) Kanduc, M.; Stubenrauch, C.; Miller, R.; Schneck, E. Interface adsorption versus bulk micellization of surfactants: insights from molecular simulations. *J. Chem. Theory Comput.* **2024**, *20*, 1568−1578.

(60) Soria-López, A.; García-Martí, M.; Mejuto, J. C. Ionic surfactants critical micelle concentration modelling in water/organic solvent mixtures using random forest and support vector machine algorithms. *Tenside Surfactants Detergents* **2025**, *62*, 8−18.

(61) Soria-Lopez, A.; García-Martí, M.; Barreiro, E.; Mejuto, J. C. Ionic surfactants critical micelle concentration prediction in water/organic solvent mixtures by artificial neural network. *Tenside Surfactants Detergents* **2024**, *61*, 519−529.

(62) Qin, S.; Jin, T.; Van Lehn, R. C.; Zavala, V. M. Predicting critical micelle concentrations for surfactants using graph convolutional neural networks. *J. Phys. Chem. B* **2021**, *125*, 10610−10620.

(63) Moriarty, A.; Kobayashi, T.; Salvalaglio, M.; Angeli, P.; Striolo, A.; McRobbie, I. Analyzing the accuracy of critical micelle concentration predictions using deep learning. *J. Chem. Theory Comput.* **2023**, *19*, 7371−7386.

(64) Brozos, C.; Rittig, J. G.; Bhattacharya, S.; Akanny, E.; Kohlmann, C.; Mitsos, A. Predicting the temperature dependence of surfactant cmcs using graph neural networks. *J. Chem. Theory Comput.* **2024**, *20*, 5695−5707.

(65) Mohan, M.; Smith, M. D.; Demerdash, O. N.; Simmons, B. A.; Singh, S.; Kidder, M. K.; Smith, J. C. Quantum chemistry-driven machine learning approach for the prediction of the surface tension and speed of sound in ionic liquids. *ACS Sustainable Chem. Eng.* **2023**, *11*, 7809−7821.

(66) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* **2017**, *30*.

(67) Štrumbelj, E.; Kononenko, I. A general method for visualizing and explaining black-box regression models. Adaptive and Natural Computing Algorithms: 10th International Conference, ICANNGA 2011, Ljubljana, Slovenia, April 14−16, 2011, Proceedings. *Part II* **2011**, *10*, 21−30.

(68) Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning. **2009**. DOI: 10.1007/978-0-387-84858-7.

(69) Murray, J. S.; Brinck, T.; Lane, P.; Paulsen, K.; Politzer, P. Statistically-based interaction indices derived from molecular surface electrostatic potentials: a general interaction properties function (GIPF). *Journal of Molecular Structure: THEOCHEM* **1994**, *307*, 55−64.

(70) Liu, X.; Turner, C. H. Computational study of the electrostatic potential and charges of multivalent ionic liquid molecules. *J. Mol. Liq.* **2021**, *340*, 117190.

(71) Barbosa, G. D.; Turner, C. H. Investigating the molecular-level thermodynamics and adsorption properties of per- and poly-fluoroalkyl substances. *J. Mol. Liq.* **2023**, *389*, 122826.

(72) Byrd, E. F. C.; Rice, B. M. Improved Prediction of Heats of Formation of Energetic Materials Using Quantum Mechanical Calculations. *J. Phys. Chem. A* **2006**, *110*, 1005−1013.

(73) Rice, B. M.; Hare, J. J.; Byrd, E. F. C. Accurate Predictions of Crystal Densities Using Quantum Mechanical Molecular Volumes. *J. Phys. Chem. A* **2007**, *111*, 10874−10879.

(74) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169−7192.

(75) Grimme, S.; Bannwarth, C.; Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (Z= 1−86). *J. Chem. Theory Comput.* **2017**, *13*, 1989−2009.

(76) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652−1671.

(77) Pracht, P.; Caldeweyher, E.; Ehlert, S.; Grimme, S. A robust non-self-consistent tight-binding quantum chemistry method for large molecules. *ChemRxiv* **2019**, DOI: 10.26434/chemrxiv.8326202.v1.

(78) Spicher, S.; Grimme, S. Robust atomistic modeling of materials, organometallic, and biochemical systems. *Angew. Chem., Int. Ed.* **2020**, *59*, 15665−15673.

(79) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011**, *3*, 33.

(80) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard, W. A., III; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American chemical society* **1992**, *114*, 10024−10035.

(81) Ehlert, S.; Stahn, M.; Spicher, S.; Grimme, S. Robust and efficient implicit solvation model for fast semiempirical methods. *J. Chem. Theory Comput.* **2021**, *17*, 4250−4261.

(82) Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S. B97−3c: A revised low-cost variant of the B97-D density functional method. *J. Chem. Phys.* **2018**, *148*, 064104.

(83) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **2009**, *113*, 6378−6396.

(84) Mariano Colombari, F.; Marcos Nascimento, V.; Liu, Y. L.; de Morais Rocha, G. J.; Driemeier, C. Density functional theory with implicit solvents for accurate estimation of aqueous and organic solvation free energies of lignin fragments. *ACS Sustainable Chem. Eng.* **2022**, *10*, 10870−10878.

(85) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184−32215.

(86) Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA quantum chemistry program package. *J. Chem. Phys.* **2020**, *152*, DOI: 10.1063/5.0004608.

(87) Lu, T.; Chen, F. Multiwfn: A multifunctional wavefunction analyzer. *J. Comput. Chem.* **2012**, *33*, 580−592.

(88) Lu, T.; Chen, F. Quantitative analysis of molecular surface based on improved Marching Tetrahedra algorithm. *Journal of Molecular Graphics and Modelling* **2012**, *38*, 314−323.

(89) Lu, T.; Manzetti, S. Wavefunction and reactivity study of benzo [a] pyrene diol epoxide and its enantiomeric forms. *Structural chemistry* **2014**, *25*, 1521−1533.

(90) Bader, R. F. W.; Carroll, M. T.; Cheeseman, J. R.; Chang, C. Properties of atoms in molecules: atomic volumes. *J. Am. Chem. Soc.* **1987**, *109*, 7968−7979.

(91) Bursch, M.; Neugebauer, H.; Grimme, S. Structure Optimisation of Large Transition-Metal Complexes with Extended Tight-Binding Methods. *Angew. Chem., Int. Ed.* **2019**, *58*, 11078−11087.

(92) Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C.et al.; CREST—A program for the exploration of low-energy molecular chemical space. *J. Chem. Phys.* **2024**, *160*, DOI: 10.1063/5.0197592.

(93) Dohm, S.; Bursch, M.; Hansen, A.; Grimme, S. Semiautomated Transition State Localization for Organometallic Complexes with Semiempirical Quantum Chemical Methods. *J. Chem. Theory Comput.* **2020**, *16*, 2002−2012.

(94) Bursch, M.; Mewes, J.-M.; Hansen, A.; Grimme, S. Best-practice DFT protocols for basic molecular computational chemistry. *Angew. Chem.* **2022**, *134*, No. e202205735.

(95) White, A. D. Deep Learning for Molecules and Materials. *Living Journal of Computational Molecular Science* **2022**, *3*, 1499.

(96) Liaw, R.; Liang, E.; Nishihara, R.; Moritz, P.; Gonzalez, J. E.; Stoica, I. Tune: A Research Platform for Distributed Model Selection and Training. *arXiv* **2018**, 1807.05118.

(97) Kato, Y. Formation of a micelle-like structure in aqueous solution of glycols. *Chem. Pharm. Bull.* **1962**, *10*, 771−788.

(98) Berthod, A.; Tomer, S.; Dorsey, J. G. Polyoxyethylene alkyl ether nonionic surfactants: physicochemical properties and use for cholesterol determination in food. *Talanta* **2001**, *55*, 69−83.

(99) Zhou, R.; Jin, Y.; Shen, Y.; Zhao, P.; Zhou, Y. Synthesis and application of non-bioaccumulable fluorinated surfactants: a review. *Journal of Leather Science and Engineering* **2021**, *3*, 1−15.

(100) Buck, R. C.; Murphy, P. M.; Pabon, M. Chemistry, properties, and uses of commercial fluorinated surfactants. *Polyfluorinated chemicals and transformation products* **2012**, *17*, 1−24.

(101) Park, K.-H.; Berrier, C.; Lebaupain, F.; Pucci, B.; Popot, J.-L.; Ghazi, A.; Zito, F. Fluorinated and hemifluorinated surfactants as alternatives to detergents for membrane protein cell-free synthesis. *Biochem. J.* **2007**, *403*, 183−187.

(102) Mortara, L.; Cortez, M. P.; Lacerda, C. D.; Carretero, G. P.; Schreier, S.; Chaimovich, H.; Lima, F. S.; Cuccovia, I. M. Micellization of Zwitterionic Surfactant with Opposite Dipoles is Differently Affected by Anions. *Langmuir* **2025**, *41*, 9480−9487.

(103) Brozos, C.; Rittig, J. G.; Bhattacharya, S.; Akanny, E.; Kohlmann, C.; Mitsos, A. Graph neural networks for surfactant multi-property prediction. *Colloids Surf., A* **2024**, *694*, 134133.

(104) Shapley, L. S. Notes on the n-person game—ii: The value of an n-person game. *Lloyd S Shapley* **1951**, *7*.

(105) Nagarajan, R.; Ruckenstein, E. Molecular theory of microemulsions. *Langmuir* **2000**, *16*, 6400−6415.

(106) Hodala, A. J.; Wood, I. G.; Carbone, P. Understanding the influence of electrostatic interactions on observed pKa shifts in surfactant aggregates using classical simulations. *J. Mol. Liq.* **2025**, *427*, 127410.

(107) Bernardino, K.; Farias de Moura, A. Electrostatic potential and counterion partition between flat and spherical interfaces. *J. Chem. Phys.* **2019**, *150*, DOI: 10.1063/1.5078686.

(108) Pike, S. J.; Hutchinson, J. J.; Hunter, C. A. H-bond acceptor parameters for anions. *J. Am. Chem. Soc.* **2017**, *139*, 6700−6706.