colloids and interfaces

MDPI

Article

A GNN-Based QSPR Model for Surfactant Properties

Seokgyun Ham *, Xin Wang, Hongwei Zhang, Brian Lattimer and Rui Qiao *

Department of Mechanical Engineering, Virginia Tech, Blacksburg, VA 24061, USA; xinwang@vt.edu (X.W.); hongwei@vt.edu (H.Z.); lattimer@vt.edu (B.L.)

* Correspondence: seokgyunh@vt.edu (S.H.); ruiqiao@vt.edu (R.Q.); Tel.: +1-540-231-7199 (R.Q.)

Abstract: Surfactants are among the most versatile molecules in the chemical industry because they can self-assemble in bulk solutions and at interfaces. Predicting the properties of surfactant solutions, such as their critical micelle concentration (CMC), limiting surface tension (γ_{cmc}), and maximal packing density (Γ_{max}) at water–air interfaces, is essential to their rational design. However, the relationship between surfactant structure and these properties is complex and difficult to predict theoretically. Here, we develop a graph neural network (GNN)-based quantitative structure–property relationship (QSPR) model to predict the CMC, γ_{cmc}, and Γ_{max}. Ninety-two surfactant data points, encompassing all types of surfactants—anionic, cationic, zwitterionic, and nonionic—are fed into the model, covering a temperature range of [20–30 °C], which contributes to its generalization across all surfactant types. We show that our models have high accuracy (R^2 = 0.87 on average in tests) in predicting the three parameters across all types of surfactants. The effectiveness of the QSPR model in capturing the variation of CMC, γ_{cmc}, and Γ_{max} with molecular design parameters are carefully assessed. The curated dataset, developed model, and critical assessment of the developed model will contribute to the development of improved surfactants QSPR models and facilitate their rational design for diverse applications.

Keywords: molecular property prediction; surfactant; graph neural networks (GNN); quantitative structure–property relationship (QSPR)

check for updates

Citation: Ham, S.; Wang, X.; Zhang, H.; Lattimer, B.; Qiao, R. A GNN-Based QSPR Model for Surfactant Properties. *Colloids Interfaces* **2024**, *8*, 63. https://doi.org/10.3390/colloids8060063

Academic Editor: Eugene V. Aksenenko

Received: 26 September 2024
Revised: 7 November 2024
Accepted: 12 November 2024
Published: 19 November 2024

1. Introduction

Surfactants are amphiphilic molecules featuring hydrophilic (water-attracting) and hydrophobic (water-repelling) motifs simultaneously. Their amphiphilic nature facilitates the interaction between immiscible phases such as oil and water, thereby reducing their interfacial tension [1]. The latter underpins their extensive usage in numerous consumer and industrial applications. For example, surfactants are essential in detergent formulation, where they play a critical role in allowing aqueous solution to penetrate deeply into soiled fabrics and emulsify, eventually removing hydrophobic contaminants [2]. In the petroleum industry, surfactants facilitate the release of oil droplets trapped on oil-wet surfaces, thereby enhancing oil recovery. Similarly, surfactant usage is integral in pharmaceutical, food, and agriculture industries [3–6].

Most of the functionalities of surfactants originate from their ability to self-assemble in bulk solutions and at fluid–fluid interfaces. The self-assembly of surfactants and their impact on fluid–fluid interfaces can be characterized by many properties; the most important include Γ_{max}, CMC, and γ_{cmc}. The maximum surface excess concentration, Γ_{max}, measures the maximum density of surfactant molecules adsorbed at an interface. The critical micelle concentration, CMC, denotes the concentration at which surfactant molecules in solution begin to self-assemble and transition from monomeric to aggregated forms. The surface tension at CMC, γ_{cmc}, represents the surface tension of a solution when micelles first begin to form, marking the point at which the addition of more surfactant no longer notably lowers the surface tension. These three parameters are crucial metrics in the practical application of surfactants. For example, a lower CMC indicates a more efficient surfactant, as it requires a lower surfactant concentration to achieve micellization, which is critical in applications ranging from detergency [7] to drug delivery systems [8]. Likewise, a lower γ_{cmc} corresponds to a more potent surfactant.

Designing surfactants with tailored $\mathit{\Gamma}_{max}$, CMC, and γ_{cmc} is a central problem in their applications. Such design, when practiced by trial-and-error, can be time-consuming and costly due to the vast chemical space of surfactant molecules. Quantitative structure–property relationship (QSPR) models that predict these properties from surfactant molecules' chemical structure can significantly accelerate their design. QSPR models for $\mathit{\Gamma}_{max}$, CMC, and γ_{cmc} of surfactants can be developed using theoretical, simulation, and data-driven approaches. Models built by theoretical approach often fall short in accuracy, as they over-simplify the complex surfactant–surfactant and surfactant–fluid interactions at play [9]. For example, among polyethylene oxide surfactants (PEO), $\mathit{\Gamma}_{max}$ is a function of the number of ethylene oxide units only. The pure simulation approach powered by molecular dynamics (MD) modeling, while potentially more accurate, is computationally expensive and time-consuming. This is because $\mathit{\Gamma}_{max}$, CMC, and γ_{cmc}, all related to surfactant self-assembly, are governed by intricate collective, emergent behaviors of surfactant molecules that can only be accurately captured through long simulations of large MD systems. The experimental approach also has drawbacks to accurately measuring surfactant properties. The most common way of calculating $\mathit{\Gamma}_{max}$ is to utilize the Gibbs adsorption equation $\Gamma \sim \mathrm{d}\gamma/\mathrm{d}\ln(\mathrm{c})$, but this method requires a very large number of measurements and long equilibration time to be accurate, especially in low concentration ranges [10]. Similarly, inconsistent CMC values because of different determination techniques has been reported [11]. Given these limitations, there is a pressing need for a data-driven approach that can leverage datasets curated from experimental and simulation studies of surfactants [12,13]. QSPR models developed using this approach could potentially offer a cost-effective and reliable approach to designing surfactants for specific applications.

The data-driven approach for QSPR model development, however, is not without its challenges. One significant hurdle is a sparsity of available data [12]. In molecular design, the combinatorial space of possible surfactant structures is virtually infinite. One can narrow down this space by leveraging chemical knowledge. For example, a surfactant molecule can be divided into head- and tail groups. The former can be nonionic, cationic, anionic, or zwitterionic, while the latter can feature hydrocarbon, fluorocarbon, and silicone tails of different lengths. However, even with chemical knowledge, the potential molecular designs remain extraordinarily high-dimensional (e.g., numerous nonionic surfactant head groups are possible). Building a dataset for $\mathit{\Gamma}_{max}$, CMC, and γ_{cmc} that comprehensively covers the high-dimensional chemical space of surfactants is difficult, if not impossible. The sparsity of data, coupled with the complexity and variability of surfactant behavior, complicates the development of robust data-driven models. As a result, there is an ongoing need for innovative approaches that can efficiently explore and model this high-dimensional space to achieve accurate predictions of $\mathit{\Gamma}_{max}$, CMC, and γ_{cmc}.

Many methods have been used to develop data-driven QSPR models with data scarcity. One classical method involves building a property function using molecular descriptors through molecular vector embedding. The molecular descriptors are selected either by domain knowledge [14–20], multi-linear regression [21–23], Pearson coefficient [24], support vector machine, [25–27] or random forests [28–30]. While they can show good performance in predicting surfactant properties, the descriptors from molecular descriptor libraries are not readily interpretable. For example, Seddon et al. reported good performance in predicting $\mathit{\Gamma}_{max}$ and CMC, but the descriptors revealed as important input features, e.g., VE3Sign_RG (the logarithmic coefficient sum of the last eigenvector from the reciprocal squared geometrical matrix) are hard to interpret intuitively. One can mitigate the data sparsity issue by narrowing the data scope to certain types of surfactants. For example, work reported by Kania et al. [28] investigated only three types of functional groups in nonionic surfactants. Similarly, work by Wu et al. [14] is limited to only nonionic surfactants, and work by Creton et al. [31] focuses only on perfluoroalkyl substance. While this method can construct good models for certain surfactant types, its relatively narrow chemical space limits its utility in exploring and searching for optimal surfactants for target applications.

With the rise of machine learning (in particular, deep learning), the graph neural network (GNN) has become a popular data-driven method for developing QSPR models. In GNN, a molecule with atoms and bonds can be treated as a graph with nodes and edges. Since GNN takes the constitutional atomic features directly from the molecule structure information, it enables neural network models to learn the complicated relationship without additional embedding vectors. Qin et al. [32] studied the CMC prediction of surfactants covering all main categories using the graph convolutional network (GCN) algorithm. They also built saliency maps to study the positive or negative contribution of atoms on CMC. Moriarty et al. [33] also used the GCN architecture and Gaussian process to predict CMC with uncertainty quantification. Moreover, Brozos et al. studied the predictive models of Γ_{max} and CMC [34] and investigated the temperature dependence of CMC using a similar architecture [35].

The above recent studies suggest that GNN and its derivatives can be a powerful method for developing QSPR models for surfactants. Nevertheless, these models have not yet been utilized to predict the full set of Γ_{max}, CMC, and γ_{cmc} when the dataset has a modest size, contains inevitable noise, yet covers a chemical space spanned by a wide spectrum of surfactants. Therefore, the question of whether an effective QSPR model for these properties can be developed under such practical constraints remains open. Furthermore, chemical knowledge of surfactant properties is not directly encoded in data-driven QSPR models (e.g., how γ_{cmc} varies with the tail length), and overfitting remains a major concern even with advanced ML algorithms. Therefore, the question of whether these models can capture variations in (Γ_{max}, CMC, and γ_{cmc}) while surfactant structure features remain open. Indeed, chemists have acquired such knowledge through decades of experimental, theoretical, and computational studies. It is worth examining to what extent the predictions of a data-driven QSPR model align with such knowledge.

In this work, we develop GNN-based QSPR models to predict Γ_{max}, CMC, and γ_{cmc} using a dataset of 92 surfactants with a wide variety of head- and tail groups. The rest of the manuscript is organized as follows. Section 2 discusses the curation and noise/uncertainty of the dataset, the architecture of the QSPR model, and the selection of the model's hyperparameters. Section 3 analyzes the performance of the QSPR model and assesses its capability to capture the variation of Γ_{max}, CMC, and γ_{cmc} as a function of surfactant design parameters, including head- and tail types, and tail length.

2. Materials and Methods

2.1. Dataset Curation

To build a dataset of Γ_{max}, γ_{cmc}, and CMC for surfactant solutions, we collected the surface tension isotherms of 92 surfactants measured experimentally at water-air interfaces (see Table S1 in the Supplementary Materials for a list of the surfactants and their origins). There are 78 surfactants with hydrocarbon tails, 6 with fluorocarbon tails, and 8 with silicone tails. The number of anionic, cationic, zwitterionic, nonionic surfactants is 23, 5, 8, and 56, respectively. The sodium ion is the counterion for anionic surfactants, while the bromide ion is the counterion for cationic surfactants. Therefore, our dataset spans a vast chemical space despite its small size. A sample surface tension isotherm is shown in Figure 1a for the aqueous solution of sodium dodecyl sulfate (SDS), a canonical anionic surfactant. To extract Γ_{max}, γ_{cmc}, and CMC of the surfactant solution, we fit the surface tension isotherms to the Szyszkowski equation at surfactant concentration below CMC and γ_{cmc} at higher concentrations:

$$\gamma = \begin{cases} \gamma_0 - RT\Gamma_{max}\ln(1+K_L c)\ if\ c < CMC,\ nonionic\ surfactant \\ \gamma_0 - 2RT\Gamma_{max}\ln(1+K_L c)\ if\ c < CMC,\ ionic\ surfactant \\ r_{cmc}\ if\ c > CMC \end{cases} \tag{1}$$

where γ_0 is the surface tension in the absence of surfactants (72 mN/m for the air-water interfaces at 25 °C), R is the ideal gas constant, T is the absolute temperature, K_L is the Langmuir constant, and c is the surfactant concentration. We only gather isotherms at temperatures between 20 and 30 °C, and no additional salts are presented in the surfactant solution. The maximum surfactant density at interfaces $\mathit{\Gamma}_{max}$ reflects the slope of the surface tension curve before reaching the CMC, and its value is calculated based on the nonlinear least square optimization function scipy.optimize.curve fit in the scipy library [36], implemented in Python (version 3.12.1). For γ_{cmc} and CMC values, we take either reported values from the literature or the intersection of the two guidelines before and after CMC (see Figure 1a) suggested by authors of the surface tension isotherm. Values of $\mathit{\Gamma}_{max}$, γ_{cmc}, and CMC thus obtained are listed in Table S1 of the Supplementary Materials. Figure 1b–d show the histograms of $\mathit{\Gamma}_{max}$, γ_{cmc}, and logCMC of the surfactants curated. The histogram of $\mathit{\Gamma}_{max}$ is skewed toward left while that of γ_{cmc} is skewed toward right. The logCMC histogram is balanced. The ranges of $\mathit{\Gamma}_{max}$, γ_{cmc}, and logCMC are [1.6, 11] × 10^{-6} mol/m^2, [14.5, 45.5] mN/m, and [−4.8, −0.9] (mol/L) in logarithmic scale, respectively. Our dataset thus covers a rather broad range of $\mathit{\Gamma}_{max}$, γ_{cmc}, and CMC, which is consistent with the wide variety of surfactants it includes.

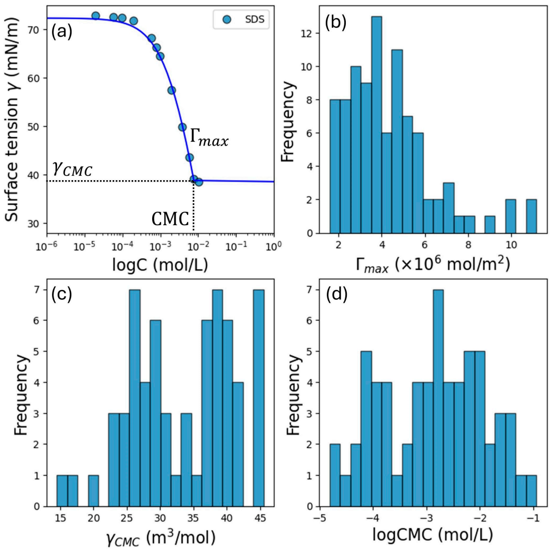

Figure 1. (**a**) The surface tension isotherm of a representative surfactant sodium dodecyl sulfate (SDS). The dots are experiment data from the literature [37], and the blue line is the fit to the Szyszkowski equation. (**b**–**d**) The histograms of $\mathit{\Gamma}_{max}$ (**b**), γ_{cmc} (**c**), and CMC (**d**) of the 92 surfactants in the curated dataset.

For many surfactants in our dataset, complete information for their properties is not available: often the surface tension isotherm was provided only at concentrations considerably below CMC and CMC was not reported. For example, the surface tension isotherm of sodium octyl carboxylate (C_8OONa) reported by Neys et al. [38] shows a monotonic decrease in surface tension, showing that micelles do not form at the surfactant concentration studied. On the other hand, the data of sodium octyl sulfate from Feinerman et al. [39] show its full profile that contains CMC information. In total, only 64 of the surfactants have known CMC values and the surface tension at CMC γ_{cmc}. The rest do not have reported CMC values while $\mathit{\Gamma}_{max}$ can still be determined with the reported surface tension isotherm.

It is also found that the $\mathit{\Gamma}_{max}$ extracted from surface tension isotherms is sometimes sensitive to the surface tension data near CMC. As illustrated in Figure 2, for sodium octyl sulfate (C_8SO_4Na), $\mathit{\Gamma}_{max}$ is fitted to be 5.02×10^{-6} mol/m^2 when the seven data points [(logC [M], γ [mN/m]) = (0.002, 71.388), (0.006, 70.214), (0.020, 63.732), (0.038, 55.975), (0.055, 50.566), (0.089, 43.931)] and CMC data (0.114, 40.563) are used in fitting. However, the fitted Γ_{max} decreases to 4.01 $\times 10^{-6}$ mol/m^2 when an additional point near CMC [(logC [M], γ [mN/m]) = (0.128, 40.614)] (colored as a green dot in Figure 2) is used in fitting. Such sensitivity, along with the fact that CMC is not always known with great accuracy [11], implies that some noise and uncertainty inevitably exist in our dataset.

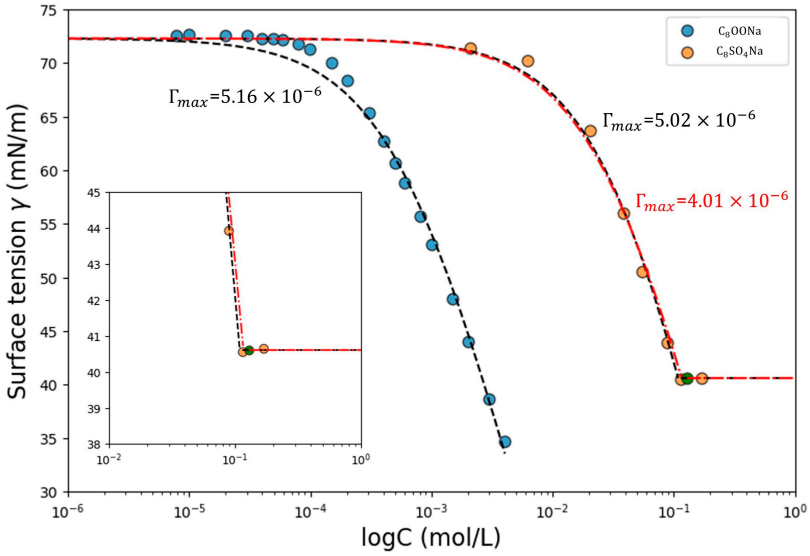

Figure 2. Inherent sensitivity of Γ_{max} (mol/m^2) when it is obtained by fitting the surface tension isotherm via the Szyszkowski equation. The inset is a zoom-in view of the surface tension profile near the CMC of surfactant C_8SO_4Na.

2.2. Architecture of Machine Learning Model

Our ML model seeks to predict Γ_{max}, γ_{cmc}, and CMC of a surfactant using its chemical structure (specifically, its SMILES string) as an input. Figure 3 shows the overview of the architecture of our model. The model consists of three modules: (1) an initial feature encoding module that converts the SMILES string of a surfactant into an adjacency matrix, an atom feature matrix, and a bond feature matrix, (2) a molecular embedding module that learns the molecular embedding vector via message passing networks, and (3) a fully connected neural network that has three end nodes corresponding to Γ_{max}, γ_{cmc}, and logCMC of the surfactant. Below we outline only the first two modules because the third module is straightforward.

Table 1. Atom features considered for surfactant molecules [1].

Atom Features	Descriptions	Size
Atom type	Type of atom (e.g., C, N, O), by atomic number	100
Number of bonds	Number of bonds the atom is involved in	6
Formal charge	Integer electric charge assigned to atom	5
Chirality	Unspecified, tetrahedral CW/CCw, or others	4
Hybridization	sp, sp2, sp3, sp3d, or sp3d2	5
Number of hydrogens	Number of bonded hydrogen atoms	5
Aromaticity	Whether this atom is part of an aromatic system	1
Atomic mass	Mass of the atom, divided by 100	1
Partial charge [2]	Non-integer electric charge assigned by the OPLS force field	1
LJ potential σ [2]	The distance parameter of the Lennard-Jones potential assigned by the OPLS force field	1
LJ potential ϵ [2]	The potential well depth parameter of the Lennard-Jones potential assigned by the OPLS force field	1

[1] All features are one-hot encoded except the atomic mass and MD-computed features. [2] MD-computable features.

Table 2. Bond features considered for surfactant molecules [1].

Bond Features	Descriptions	Size
Bond type	Single, double triple, or aromatic	4
Conjugated	Whether the bond is conjugated	1
In-ring	Whether the bond is part of a ring	1
Stereo	None, any, E/Z, or cis/trans	6

[1] All features are one-hot encoded.

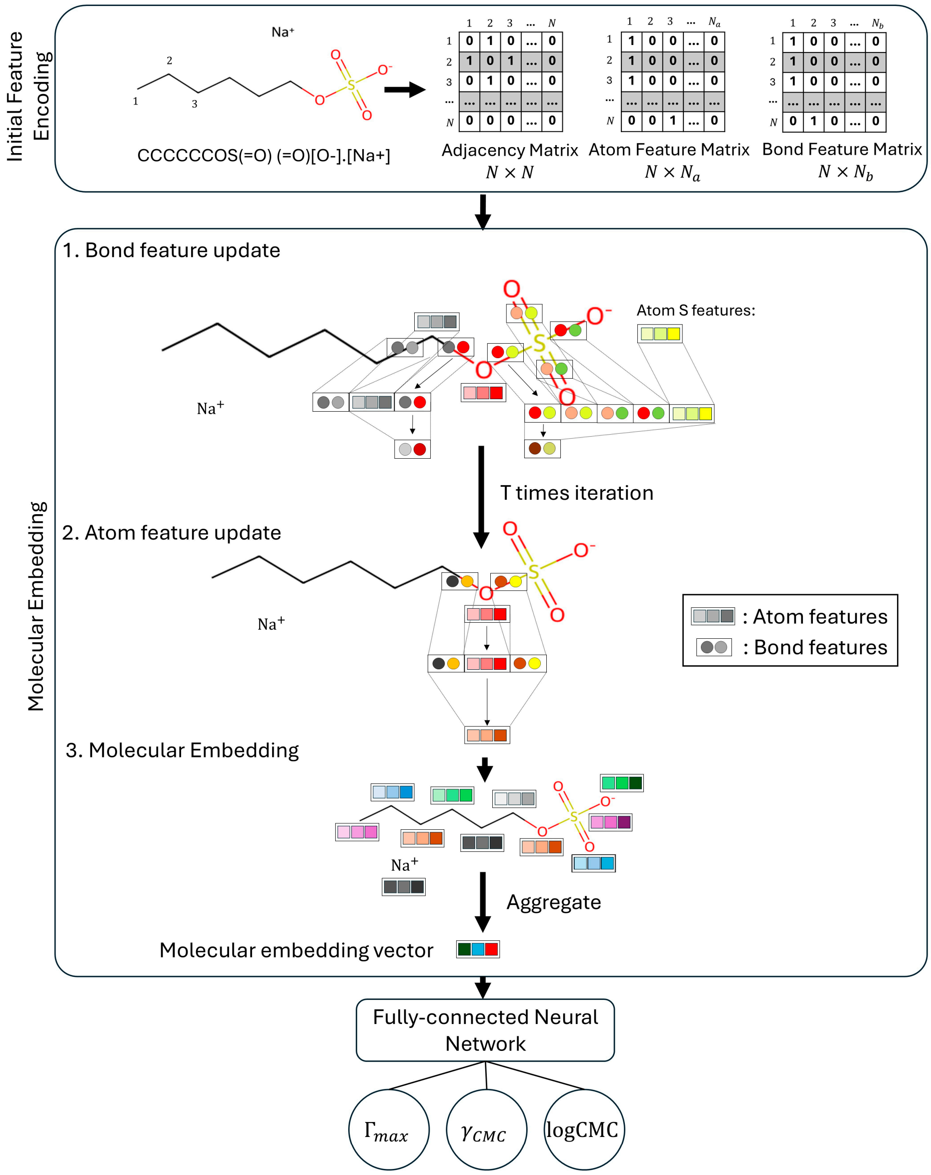

Figure 3. An overview of the architecture of ML model for predicting surfactant properties. In the initial feature encoding module, a SMILES string is used to generate a corresponding adjacency matrix, an atom feature matrix, and a bond feature matrix (see Tables 1 and 2 for the atom and bond features included in this work; N_a (N_b) is the number of atom (bond) features). In the molecular embedding module, firstly the bond feature is updated by sharing the information from the bond itself, neighbor atoms, and one of the neighbor atom's bonds. Once the bond feature is updated over T cycles, the atom feature is updated using the updated bond feature. Once the atom features are updated, they are aggregated to a molecular embedding vector by averaging the atom feature vectors of each atom. The molecular embedding vector is fed into the fully connected neural network with three node outputs targeting surfactant properties Γ_{max}, γ_{cmc}, and logCMC.

In the initial feature encoding module, a SMILES string representing the surfactant structure is used to generate an adjacency matrix, an atom feature matrix, and a bond feature matrix. The adjacency matrix is a square matrix encoding the connectivity between atoms (1 as connected or 0 not connected), and its size is the product of the number of atoms in the surfactant molecule (N). The atom feature matrix stores atomic features. We selected the features listed in Table 1, where the i-th row and j-th column of the matrix corresponds to the j-th feature of i-th atom. The bond feature matrix contains the features of chemical bonds between atoms, and Table 2 lists the features we selected. In the molecular embedding module, we chose the directed message-passing neural networks (D-MPNN) algorithm developed by McGill and colleagues [40,41] as the message-passing network. This is because D-MPNN has excellent capability in molecular property prediction. For example, its effectiveness has been demonstrated with [21] public datasets including QM9, ESOL, and FreeSolv, as well as datasets collected by other authors. [42–44] In D-MPNN, the molecular embedding module starts by introducing the directed bond feature e^d_{vw}. This vector is devised to avoid messages passed along any path of the form $\mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_\mathrm{n}$ ($\mathrm{v}_\mathrm{i} = \mathrm{v}_{\mathrm{i+2}}$ for any i) [39] and is given by:

$$e^d_{vw} = cat(x_v, e_{vw}) \tag{2}$$

where x_v and e_{vw} are the atom feature vector of atom v and the bond feature vector of the bond between atoms v and w. These vectors are a subset of the atom feature matrix and bond feature matrix obtained from the previous feature encoding module, respectively. It is noted that x_w is excluded from the concatenation because it is directional. The directed edge feature vectors are used to initialize the edge hidden feature h^0_{vw} with a learned matrix W_i and activation function τ, followed by the message-passing process between neighbor edge hidden features.

$$h^0_{vw} = \tau\left(W_i e^d_{vw}\right) \tag{3}$$

$$h^{t+1}_{vw} = \tau\left(h^0_{vw} + W_j \sum_{k \in (N(v) \backslash w)} h^t_{kw}\right) \tag{4}$$

Note that during one cycle of this message-passing layer to update h^t_{vw}, the information from the bond itself, the neighbor atoms, and the other bonds of one of the neighbor atoms are shared. This update is iterated across all bonds as one cycle of the message-passing layer. Once this process iterates T times of message-passing layers to obtain final hidden bonds feature vectors h^T_{ij} where T is the number of messages passing layers and a hyperparameter, they are used to build the hidden atom features h_v via

$$q = cat\left(x_v, \sum_{k \in (N(v))} h^T_{kw}\right) \tag{5}$$

$$h_v = \tau(W_k q) \tag{6}$$

The hidden atom features h_v are aggregated into the molecular embedding vector h_m by taking the average of the hidden atom features. x_w is an optional vector of additional molecular descriptors that can concentrate with the molecular embedding vector. Here we add two geometric descriptors that can be computed using MD simulations (Gromacs 2020.3, access date: 1 November 2023–13 November 2024), and they are described in the next section.

$$h'_m = \frac{1}{n} \sum_{v \in V} h_v \tag{7}$$

$$h_m = cat\left(h'_m, x_m\right) \tag{8}$$

where n is the size of h_v. The molecular embedding vector is fed into the fully connected layers which have three nodes at the end targeting the surfactant properties Γ_{max}, γ_{cmc}, and logCMC.

2.3. Additional Features and Descriptors by MD Simulations

Directly predicting surfactant properties such as maximum surface density or CMC through MD simulations is challenging, as it requires long simulation times. However, MD simulations can still provide valuable insights. To explore their potential utility, we investigated the use of computationally inexpensive MD simulations as molecular features and descriptors to enhance the predictive power of our property models. Therefore, in addition to the default list of the features, we introduce additional atom features and descriptors that are computable using MD simulations (see Figure 4). The partial charge and the Lennard-Jones parameters σ and ϵ of each atom in a surfactant molecule are essential atom-level parameters that govern the strength of electrostatic and van der Waals interactions between atoms. Therefore, they are introduced and later examined whether they help improve model performance. The MD-computable features in Table 1 are obtained by the OPLS force fields using the LigPargen web server (https://zarbi.chem.yale.edu/ligpargen/, accessed on 13 November 2024) [45] and are concatenated in the initial atom features x_v. In Figure 4a, the MD-computable features of a carbon, an oxygen, and a sulfur atom of SDS are displayed. It is noted that the hydrogen atoms connected to the carbon atoms are weakly positively charged, so the tail remains overall neutral as intended.

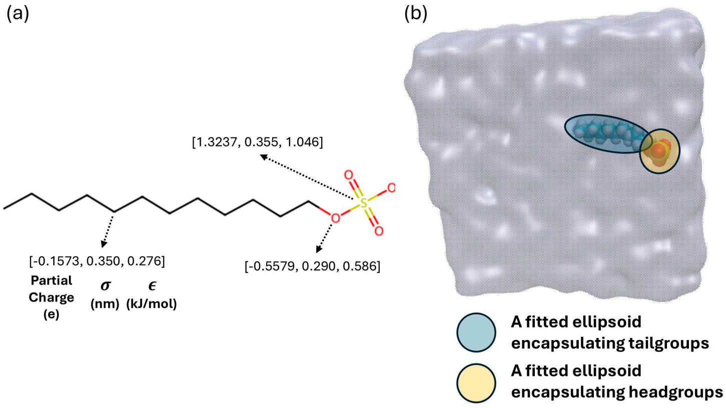

Figure 4. (**a**) Three MD-computable atom features: partial charge and the Lennard-Jones parameters σ and ϵ. (**b**) A snapshot of a sample MD system used to calculate two geometric descriptors of a surfactant molecule derived from the fitted ellipsoids encapsulating its tailgroup and headgroup.

It is known the geometrical properties of surfactant molecules play a key role in their self-assembly and adsorption at water-air interfaces [1]. Although measuring these geometrical properties is difficult in experiments, they can be computed through MD simulations. Therefore, it is worth computing some of these geometrical properties and investigating whether incorporating them into the ML model can improve its performance. Here we build a $5 \times 5 \times 5$ nm^3 size MD system where a single surfactant molecule is immersed in water (see Figure 4b). The system is run for 15 ns in the NVT ensemble, with the last 10 ns taken as the production run. The subgroup atoms (e.g., headgroup or tailgroup) of the surfactant molecule are selected and fit to an ellipsoid by equating the moment of inertia of the selected atoms to that of an ellipsoid (see Supplementary Materials for the full detail of the MD setting and ellipsoid fitting). We then use the surface area of the fitted ellipsoid encapsulating headgroup and tailgroup as descriptors (see Table 3). These descriptors are concatenated in the molecular embedding vector as x_m.

Table 3. The list of additional molecular descriptors.

Additional Molecular Descriptors	Descriptions	Size
Surface area of tailgroups	Surface area of a fitted ellipsoid encapsulating tailgroups	1
Surface area of headgroups	Surface area of a fitted ellipsoid encapsulating headgroups	1

2.4. Hyperparameter Setting

We use random data splitting based on 0.8/0.1/0.1 for the train/validation/test. Table 4 shows the list of hyperparameters tried and their final selection obtained using the grid-search algorithm. The criteria to determine the best set of hyperparameters is to minimize the average of the three surfactant properties' root mean square error (RMSE). Each RMSE value is averaged from five ensembles initialized with different random weights, which has been demonstrated to help increase model accuracy [41,46].

Table 4. Hyperparameters of the ML model, with the selected values highlighted in bold.

Hyperparameters	Values
Number of message-passing layers	2, **3**, 4
The size of hidden bond feature h_{vw}	200, **300**, 400
Number of fully connected layers	2, **3**, 4
Dropout rate	**0**, 0.1, 0.2

3. Results

3.1. Model Evaluation

We vary the choice of the initial atom features x_v as well as the molecular descriptors x_m to investigate how different sets of inputs affect the model performance. Table 5 lists the combination of the atom features and the additional molecular descriptors used as inputs. Specifically, we tried four sets of inputs by including or excluding the MD-computable atom features (partial charge, σ, and ϵ) and MD-computed molecular descriptors (the surface area of surfactant tailgroup and headgroup). The default set in Table 5 refers to the first eight atom features listed in Table 1, which do not require MD simulations to obtain.

Table 5. The list of atomic features and additional molecular descriptor sets (the default set refers to the first eight atom features in Table 1).

Set	Atom Feature Set	Additional Molecular Descriptors
1	Default	None
2	Default + partial charge + σ+ϵ	None
3	Default	Surface areas of tailgroup and headgroup
4	Default + partial charge + σ+ϵ	Surface areas of tailgroup and headgroup

Figure 5 shows the RMSE values of the target surfactant properties with different input sets. For Γ_{max}, Set 3 shows the smallest RMSE (1.04) in the test set. Set 3 also shows the smallest test RMSE (2.46) for γ_{cmc}. Set 1 shows the smallest test RMSE (0.28) for logCMC. It appears that the concatenation of information regarding the surface area of a surfactant's headgroup and tailgroup into the molecular embedding vector helps the model better capture its properties. For example, the RMSE is reduced by 7% in γ_{cmc} with the surface area information, and the RMSE is reduced by 3% when predicting Γ_{max}. However, its effect does not always improve the model's performance. When predicting logCMC, the RMSE increases by 10%. Overall, introducing MD computed features into our model only marginally affects its performance. We envision that other MD-computed features or other ways of incorporating MD-computed features into the model may improve its performance.

A systematic investigation of such practice is beyond the scope of this work. Instead, below we select Set 1 in Table 5 as the baseline input set to further analyze the model performance.

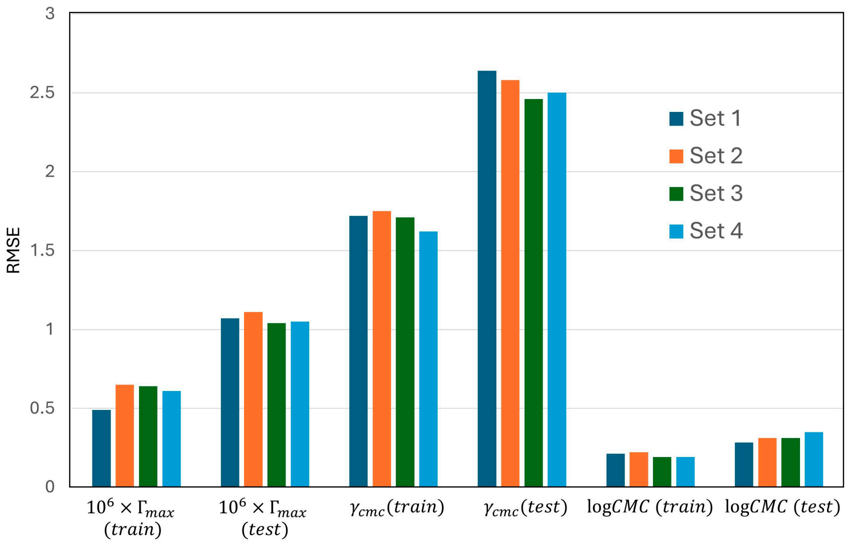

Figure 5. RMSE of predicted Γ_{max}, γ_{cmc}, and logCMC in the training and test sets with four different inputs listed in Table 5.

Figure 6 shows the parity plot of the three target properties of all 92 surfactants in our dataset. The values are the averaged values of the five ensembles predicted based on Set 1 input in Table 5. The R^2 value in the train set for Γ_{max}, γ_{cmc}, and logCMC is 0.94, 0.95, and 0.95, respectively, and that in the test set is 0.82, 0.90, and 0.90, respectively. Though all three surfactant properties are predicted reasonably well by the model in the test set as well as in the training set, it appears that the model is overfitting with respect to Γ_{max}. This is likely due to the inherent uncertainty/noise of Γ_{max} as illustrated in Figure 2. A similar phenomenon has also been reported by Seddon et al. [28], where the prediction of Γ_{max} is weaker than that of other properties logKL and logCMC. Nevertheless, the parity plots show that our model performs well in predicting the three surfactant properties considered here.

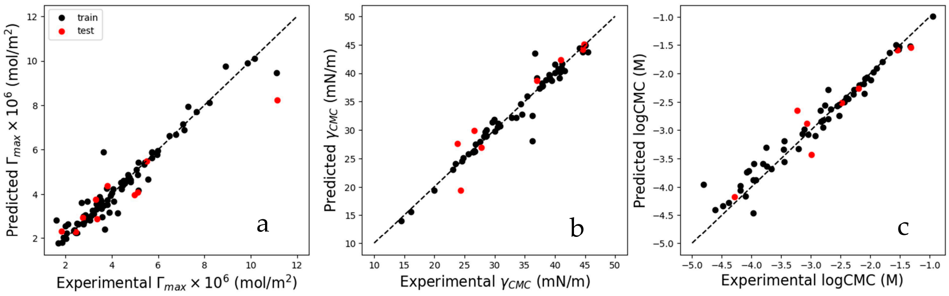

Figure 6. The parity plot of surfactant properties (**a**) Γ_{max}, (**b**) γ_{cmc}, and (**c**) logCMC.

3.2. Surfactant Parameter Space Exploration

The ultimate goal of a QSPR model is to predict surfactant properties over a wide chemical space. The good performance of our model shown in the previous section for the surfactants in our dataset is encouraging. However, the dataset does not cover the high-dimensional chemical space of surfactants well (e.g., many hydrocarbon surfactants with nonionic headgroups are included, but far fewer fluorocarbon surfactants with similar headgroups exist in the dataset), and overfitting is possible even with advanced ML models. Therefore, it is important to examine the behavior of the model with a wide variety of chemical structures. Because rigorous validation is difficult due to a lack of experiment data, we shall compare the semi-quantitative trend and scaling law predicted by the model with those expected from chemical knowledge. Because our model is trained using very sparse data over the high-dimensional space of surfactant structure and the data likely contain some noise, capturing those trends and scaling law, even though possible for experienced chemists, is not guaranteed a priori for the data-driven model we develop here.

Below, we generate the landscape of surfactant properties as a function of several molecular design parameters based on our models. Specifically, we vary the surfactant structure to investigate the variation of Γ_{max}, γ_{cmc}, and logCMC. While numerous groups of surfactants vary in structure, we select the nonionic head (polyethylene oxide, PEO), and anionic head types (sulfate, sulfonate, and carboxylate) as the representative surfactants, because these functional groups are commonly found in surfactants. Figure 7 shows the schematic of the surfactant head- and tail types used in this study. For PEO, we vary the number of carbons in its tail N_C and the number of ethylene oxide (EO) units N_{EO} in its headgroup. With the PEO head, we limit the choice of tail types only to hydrocarbon. For anionic surfactants, we vary not only the surfactant head types (sulfonate, sulfate, and carboxylate) but also the tail types to hydrocarbon (C_xH_y) and fluorocarbon (C_xF_y) tails. The chain length N_C of the hydrocarbon and fluorocarbon tails is also varied.

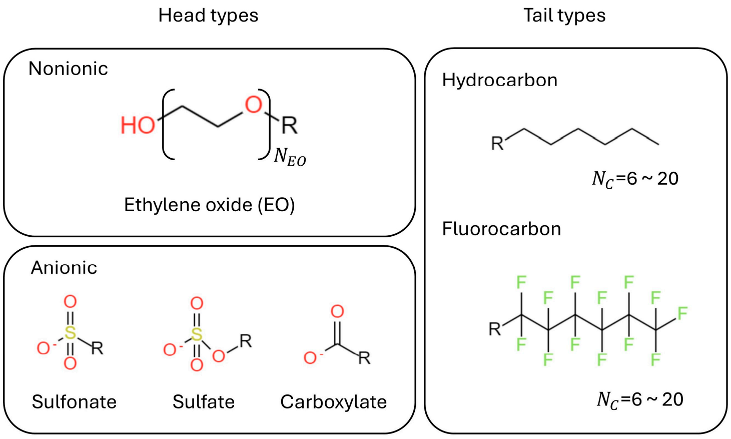

Figure 7. Illustrations of surfactant head- and tail types used in the surfactant structural exploration. N_C is the number of carbons in a surfactant's hydrophobic tail.

3.2.1. Polyethylene Oxide (PEO) Surfactants

Figure 8a depicts the variation of Γ_{max} as a function of N_C and N_{EO} for $6 \leq N_C \leq 20$ and $4 \leq N_{EO} \leq 16$ predicted by our model, and the experimental data available in our dataset are marked by color-coded crosses and triangles (all symbols are color-coded based on N_{EO}). Despite the minimal experimental data in the parameter space examined here, the predicted dependence of Γ_{max} on N_C and N_{EO} is smooth, suggesting that our model is likely not overfitted.

The predicted dependence of Γ_{max} on N_C and N_{EO} is largely consistent with the trend observed for PEO surfactants and that expected for nonionic hydrocarbon surfactants. Specifically, Γ_{max} is predicted to increase with decreasing N_{EO}, which agrees with that observed for PEO surfactants [1] and also aligns with the expectation that the packing density of nonionic surfactants at fluid interfaces typically increases as the size of their hydrophilic head decreases [1]. The model predicts that Γ_{max} generally increases with the tail length N_C although an oscillatory behavior is observed at $N_C \gtrsim 12$ when N_{EO} is less than about 10. The former is consistent with the observation that Γ_{max} of PEO surfactants generally increases with the length of their hydrophobic group, which can be attributed to the stronger lateral interactions and thus denser packing of their tails [1]. The reliability of the more complicated dependence of Γ_{max} at small N_{EO} and large N_C predicted by the model cannot be ascertained yet. However, at small N_C, the favorable PEO-water and PEO-PEO interactions and hydrophobic interactions between hydrocarbon tails may be insufficient to enforce a dense packing of long hydrocarbon tails due to the entropic penalty associated with the latter. Therefore, at small N_{EO}, the dependence of Γ_{max} on N_C likely would not follow the monotonic trends found at large N_{EO}, and a more complicated relation, as suggested by our model, may emerge.

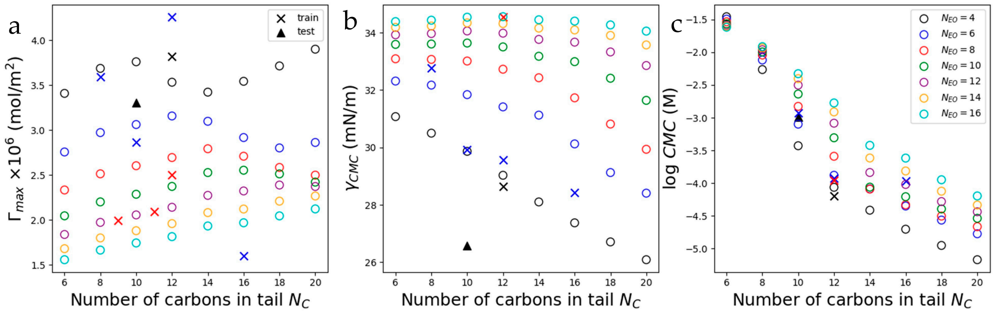

Figure 8. The dependence of (**a**) Γ_{max}, (**b**) γ_{cmc}, and (**c**) logCMC on the number of carbons in the tail (N_C) and the number of ethylene oxide unit (N_{EO}) of PEO surfactants. Circles are model predictions and crosses and triangles are experimental data; all symbols are color-coded based on N_{EO} as marked in the legend.

At a more quantitative level, the model seems robust enough to overcome likely noise in the training dataset and can capture an important scaling law. At (N_C, N_{EO}) = (12, 6), the predicted Γ_{max} is 3.16 µmol/m^2, which is 26% lower than the experimental data; a larger discrepancy of 75% is observed at (N_C, N_{EO}) = (16,6). In light of the uncertainties of experimentally derived Γ_{max} highlighted in Section 2 and Figure 2, it is likely that the reported Γ_{max} has substantial noise and the model has avoided overfitting these data. The minimum area per PEO surfactant molecule, given by Amin = $1/\Gamma_{max}$, follows the well-known scaling law $A_{min} \sim \sqrt{N_{EO}}$ [47]. Figure 9 shows the relationship between A_{min} and $\sqrt{N_{EO}}$. While the model prediction satisfies the scaling law well, the experimental data do not align with the scaling law. The latter suggests that the experimental data in Figure 8a likely contain significant uncertainty. Despite such uncertainty, the model can capture the scaling law of Γ_{max} established in the literature. This is likely because the experimental data within our dataset but outside of the (N_C, N_{EO}) range considered in Figure 8 work as regularizers to prevent the overfit to the noisy Γ_{max}. Moreover, the nature of multi-task prediction utilized in our model framework helps to generate robust high-level representations [48].

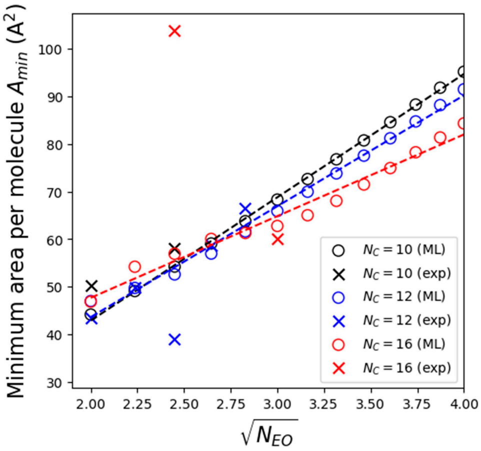

Figure 9. The variation of the minimum area per molecule A_{min} relative to the square root of EO units (N_{EO}) in PEO surfactants. The model predictions are denoted using circles and the experiment data is denoted using a cross.

Figure 8b shows the predicted variation of γ_{cmc} for $6 \leq N_C \leq 20$ and $4 \leq N_{EO} \leq 16$, and the experimental data available in our dataset are marked by color-coded crosses and triangles. The consistency between the model prediction and available experimental data is greatly improved compared to that for Γ_{max}. The largest difference in γ_{cmc} is 11% and occurs at (N_C, N_{EO}) = (10, 4), where the model prediction and experimental data are 29.86 and 26.57 mN/m, respectively. In terms of qualitative trend, γ_{cmc} generally decreases as N_C increases and N_{EO} decreases. These trends are difficult to establish using the experimental data in the (N_C, N_{EO}) space examined here alone. Nevertheless, the reduction of surface tension by surfactants generally increases with their tail length [49] and their packing density at water-air interfaces [1]. Figure 8a shows that, overall, Γ_{max} decreases (increases) with an increase of N_{EO} (N_C). Therefore, the trends shown in Figure 8b seem reasonable and warrant further experimental validations.

Figure 8c shows the predicted variation of CMC with N_C and N_{EO}. The agreement between the predictions and available experimental data is comparable to that for γ_{cmc}. Overall, logCMC decreases as N_C increases for a given N_{EO}, as has been observed for a wide variety of surfactants [50]. The predicted logCMC decreases as N_{EO} decreases for a given N_C, and the reduction of logCMC with N_{EO} is much weaker than that per methylene unit in the hydrophobic chain. Both trends are consistent with the literature [1], although the validity of the more subtle details of logCMC variation shown in Figure 8c, e.g., the insensitivity of logCMC relative to N_{EO} at $N_C \leq 10$, are difficult to ascertain.

3.2.2. Anionic Surfactants

In this section, we investigate the variation of Γ_{max}, γ_{cmc}, and logCMC of anionic surfactants with different types of head- and tail groups, and the length of tail groups. We consider three anionic head groups (sulfate, sulfonate, and carboxylate) and two tail group types (hydrocarbon and fluorocarbon). The number of carbon atoms N_C of hydrocarbon and fluorocarbon tails is varied from 6 to 20.

Γ_{max} of anionic surfactants. Comparison of the data in Figure 10a,d shows that Γ_{max} is more affected by the tail type than the head type, with Γ_{max} of fluorocarbon surfactants always higher than that of hydrocarbon surfactants for all three head types. This result aligns well with the fact that the cross-section area of the fluorocarbon chain (about 27 A^2) is much higher than that of the hydrocarbon chain (about 20 A^2) [51] For both fluorocarbon and hydrocarbon surfactants, Γ_{max} increases with the tail length for $N_C \leq 12$; at larger N_C, the effects of tail length on Γ_{max} becomes marginal. These trends for the hydrocarbon surfactants can be loosely inferred from the limited, and somewhat noisy experimental data in our dataset (see the crosses and triangles in Figure 10a), and agree well with the experimental observations that, for ionic surfactants, their Γ_{max} increases with N_C for short tails but varies little once the N_C exceeds 10 [1].

The model predicts a complicated dependence of Γ_{max} on the head type. For hydrocarbon surfactants with the same N_C, Γ_{max} follows the order of carboxylate > sulfonate > sulfate at $N_C \geq 10$; at smaller N_C, the order changes to sulfonate > carboxylate ≈ sulfate. For fluorocarbon surfactants, Γ_{max} consistently follows the order of carboxylate > sulfate > sulfonate. These trends do not always align with the trend that may be estimated from the limited experimental data in Figure 10a. Inferring these trends from data for other surfactants is not straightforward given that Γ_{max} is governed by the complex interplay between tail–tail, head–head, and head–water interactions, which depends on the structure of head and tail in a manner that defies simple extrapolation in the high-dimensional chemical space. While this makes validating the model prediction difficult, it also highlights the potential of ML-based models in suggesting new experimental studies to explore complex structure–property relations.

γ_{cmc} of anionic surfactants. Comparison of the data in Figure 10b and e shows that for surfactants with the same anionic headgroup, the predicted γ_{cmc} is lower for fluorocarbon tailgroups than for hydrocarbon surfactants. While experimental data supporting this trend is unavailable in the parameter space we examine here, this trend aligns well with the

fundamental studies of surface tension reduction by surfactants. Specifically, as reviewed by Czajka, Hazell, and Eastoe [52], surface tension is governed by the additive contributions of dispersion and polar interactions. For surfactants with the same headgroups but different tail groups, their difference in γ_{cmc} is dominated by the difference in the dispersion interactions. Because of the bulkier size of fluorocarbon tails and the lower polarizability of fluorine atoms than hydrogen atoms, the dispersion contribution of fluorocarbon tails is smaller than that of hydrocarbon tails [52]. Consequently, γ_{cmc} of fluorocarbon anionic surfactants is lower than their hydrocarbon counterparts.

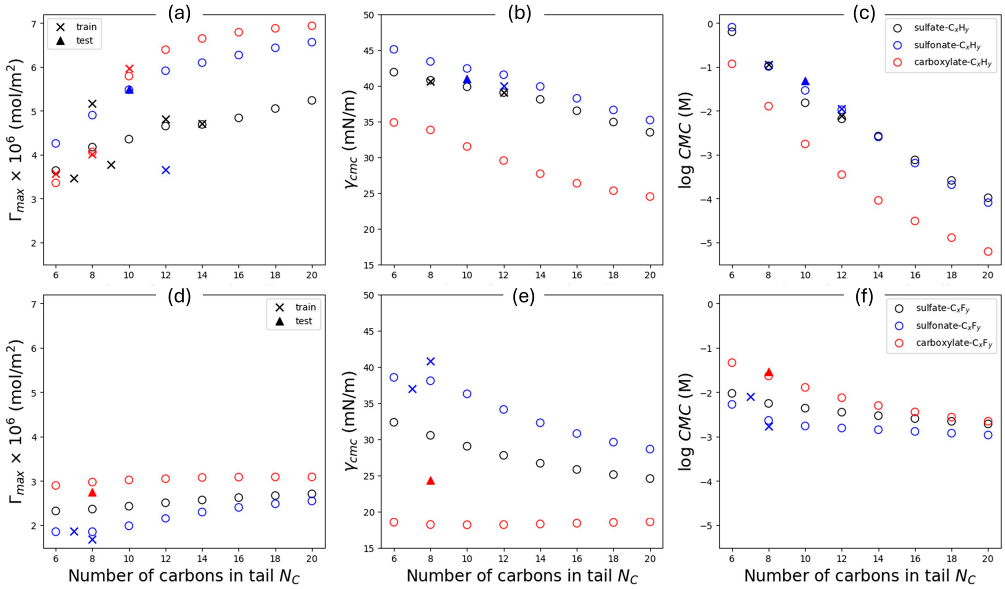

Figure 10. The variation of Γ_{max}, γ_{cmc} and logCMC for anionic surfactants with hydrocarbon (**a**–**c**) and fluorocarbon tails (**d**–**f**). C_xH_y and C_xF_y denote hydrocarbon and fluorocarbon tails, respectively. Circles and squares are model predictions; crosses and triangles are experimental data. Symbols are color-coded based on the surfactants' anionic head (black: sulfate; blue: sulfonate; red: carboxylate).

Our model predicts three main trends for the dependence of γ_{cmc} on the tail length of sur factants considered. First, γ_{cmc} of all hydrocarbon surfactants decreases with the tail length N_C, which does not contradict the two data points in our dataset. Second, for fluorocarbon surfactants, the decrease of γ_{cmc} with N_C is observed for sulfate and sulfonate headgroups, in contradiction to the data for sulfonate surfactants (cf. the two blue crosses in Figure 10f), which is likely due to experimental noise. Finally, for fluorocarbon surfactants with carboxylate headgroups, their γ_{cmc} is independent of tail length N_C. The first two predictions are consistent with the general trend that γ_{cmc} decreases with increasing tail length. The fact that this trend is recovered robustly in the presence of some noise in the training data is encouraging. The third prediction is likely valid. The limiting surface tension of fluorocarbon surfactants is usually in the range of 15–25 mN/m [52]. For fluorocarbon carboxylate surfactants, a γ_{cmc} of about 18 mN/m is already achieved at N_C = 6, making the reduction of γ_{cmc} with increasing N_C unlikely.

Our model predicts that the γ_{cmc} of surfactants with the same tailgroup follows the trend of sulfonate > sulfate > carboxylate, i.e., carboxylate headgroups are most effective in lowering surface tension, followed by sulfate and then sulfonate headgroups. Such a trend is difficult to establish with the experimental data shown in Figure 10b,e. However, in a related study, Liu and colleagues [53]. studied the surface activities of sodium alkyl polyethylene oxide carboxylate (AE_3C), sodium alkyl polyethylene oxide sulfate (AE_3S), and sodium alkyl polyethylene oxide sulfonate (AE_3SO). The terminal motifs of these surfactants' headgroups are the same as those studied here, and their tailgroups have 12–14 carbon atoms. They found that the γ_{cmc} of surfactants with sulfonate group is highest, followed by those with sulfate group, and then by those with carboxylate. This study thus lends indirect support to our model prediction, even though a direct prediction of our prediction is not yet available.

logCMC of anionic surfactants. Comparison of the data in Figure 10c,f shows that logCMC of hydrocarbon surfactants decreases sharply and almost linearly with the tail length N_C. For fluorocarbon surfactants, the decrease of logCMC with N_C is much slower. The former prediction is consistent with the general trend of logCMC of anionic hydrocarbon surfactants [1] and the data for $C_nH2_{n+1}OSO_3Na$ (n = 8 to 14) reported by Srinivasan and Blankschtein [54]. The latter prediction, however, is inconsistent with some reported data in the literature. For example, the data presented by Kancharla and colleagues showed that the logCMC of $C_{n-1}F_{2n-1}COONa$ decreases by 1.7 as n increases from 5 to 9 [55], which is even more steep compared to that for $C_nH_{2n+1}OSO_3Na$ [54] The models by Creton [31] also support the steep slope of the graph, showing that the logCMC of $C_{n-1}F_{2n-1}COONa$ decreases from −1 to −3 as n increases from 6 to 10, and logCMC of $C_{n-1}F_{2n-1}SO_3Na$ decreases from −2 to −3 as n increases from 6 to 9. This inconsistent trend observed in our model compared to the references arises likely due to the limited number of fluorinated surfactant data.

For hydrocarbon surfactants with the same tail length, our model predicts that their logCMC follows the order of sulfate ≈ sulfonate > carboxylate. While there is no logCMC data in our dataset to support these trends, they seem unreliable based on the data reported by Sadeghi and Shahabi: they found that the logCMC of $C_{12}H_{25}SO_3Na$ is higher by about 0.18 than that of $C_{12}H_{25}SO_4Na$ [56]. For fluorocarbon surfactants with the same tail length, our model predicts that their logCMC follows the order of carboxylate > sulfate > sulfonate. This trend seems to be supported by a pair of data points in our dataset (cf. the red triangle and blue cross in Figure 10f), but its robustness is unclear at present.

4. Conclusions

In summary, we develop a GNN-based QSPR model to predict surfactant properties Γ_{max}, γ_{cmc}, and CMC.Ninty-two surface tension isotherms are curated and used to extract the three properties. Atom and bond features of surfactant molecules are incorporated as model inputs; MD-computed atom features and molecular descriptors are also introduced to test whether these physics-inspired features help improve the model. The full feature vectors in the model are updated through the D-MPNN algorithm and aggregated into the molecular embedding vector, which is connected to fully connected neural network layers to predict the three surfactant properties.

Our model predicts the three surfactant properties reasonably well: the average R^2 is 0.87 and the RMSE is 1.07, 2.64, and 0.28 for Γ_{max}, γ_{cmc}, and CMC, respectively. The model performance is only marginally improved by introducing MD-computed features of surfactant molecules. We also systematically examine the variation of predicted properties for surfactants with different types and sizes of head- and tail groups. With a few exceptions, within the chemical space explored, the trends (and, sometimes, the scaling law) of the three properties predicted by the model often agree well with those established theoretically and experimentally in the literature, despite that little experimental data regarding the explored chemical space is available in the training dataset and that those trends are not encoded directly into the model.

The fact that our QSPR model was trained on a small and noisy dataset in the vast chemical space of surfactants not only describes the dataset reasonably well but also captures the variation of surfactant properties in the chemical space is encouraging. The quality of the model can be improved by expanding the training dataset. In this regard, analysis of the trends predicted by the current model can help researchers gather new data strategically. Indeed, it would be helpful to experimentally explore the chemical space in which the current model predicts unusual trends or trends contradictory to expectations based on conventional 'chemical' wisdom or data from related surfactants. Such experiments help explore potentially new surfactant behaviors to expand our chemical knowledge, and their data will improve the reliability of the QSPR model, both of which will benefit the rational design of surfactants for diverse applications.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/colloids8060063/s1, Table S1: Surfactant property data collection from literature [57–72]. Figure S1: Surface tension isotherms of surfactants.

Author Contributions: Author Contributions: Conceptualization, S.H., B.L. and R.Q.; methodology, S.H. and R.Q.; software, S.H.; validation, S.H., X.W. and R.Q.; formal analysis, S.H., H.Z. and R.Q.; investigation, S.H., X.W., H.Z., B.L. and R.Q.; resources, S.H.; data curation, S.H.; writing—original draft preparation, S.H. and R.Q.; writing—review and editing, X.W., H.Z. and B.L.; visualization, S.H.; supervision, R.Q.; project administration, R.Q.; funding acquisition, B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Strategic Environmental Research and Development Program under grant number WP20-3048.

Data Availability Statement: All the data presented in the study are included in the Supplementary Materials, further inquiries can be directed to the corresponding author.

Acknowledgments: We thank Virginia Tech Advanced Research Center (ARC) for providing computational resources.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rosen M, J. *Surfactants and Interfacial Phenomena*, 4th ed.; John Wiley & Sons: Hoboken, NJ, USA, 2012; Chapter 1, 3, 7.
2. Bajpai, D.; Tyagi, V. Laundry Detergents: An Overview. *J. Oleo Sci.* **2007**, *56*, 327–340. [CrossRef]
3. Ceresa, C.; Fracchia, L.; Fedeli, E.; Porta, C.; Banat, I.M. Recent advances in biomedical, therapeutic and pharmaceutical ap-plications of microbial surfactants. *Pharmaceutics* **2021**, *13*, 466. [CrossRef] [PubMed]
4. Kralova, I.; Sj¨oblom, J. Surfactants used in food industry: A review. *J. Dispers. Sci. Technol.* **2009**, *30*, 1363–1383. [CrossRef]
5. Appah, S.; Jia, W.; Ou, M.; Wang, P.; Asante, E.A. Analysis of potential impaction and phytotoxicity of surfactant-plant surface interaction in pesticide application. *Crop. Prot.* **2020**, *127*. [CrossRef]
6. Ramsey, R.; Stephenson, G.; Hall, J. A review of the effects of humidity, humectants, and sur factant composition on the ab-sorption and efficacy of highly water-soluble herbicides. *Pestic. Biochem. Physiol.* **2005**, *82*, 162–175. [CrossRef]
7. Goddard, E.D. *Polymer*/surfactant interaction—Its relevance to detergent systems. *J. Am. Oil Chem. Soc.* **1994**, *71*, 1–16. [CrossRef]
8. Torchilin, V.P. Structure and design of *polymer*ic surfactant-based drug delivery systems. *J. Control. Release* **2001**, *73*, 137–172. [CrossRef]
9. Mochizuki, K. The packing parameter of bare surfactant does not necessarily indicate mor phological changes. *J. Colloid Interface Sci.* **2023**, *631*, 17–21. [CrossRef]
10. Yunfei, H.; Yazhuo, S.; Honglai, L.; Dominique, L.; Anniina, S. Surfactant Adsorption onto Interfaces: Measuring the Surface Excess in Time. *Langmuir* **2012**, *28*, 3146–3151. [CrossRef]
11. Scholz, N.; Behnke, T.; Resch-Genger, U. Determination of the Critical Micelle Concentration of Neutral and Ionic Surfactants with Fluorometry, Conductometry, and Surface Tension—A Method Comparison. *J. Fluoresc.* **2018**, *28*, 465–476. [CrossRef]
12. Dou, B.; Zhu, Z.; Merkurjev, E.; Ke, L.; Chen, L.; Jiang, J.; Zhu, Y.; Liu, J.; Zhang, B.; Wei, G.-W. Machine Learning Methods for Small Data Challenges in Molecular Science. *Chem. Rev.* **2023**, *123*, 8736–8780. [CrossRef] [PubMed]
13. Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; van Hoe sel, C.; Schopmans, H.; Sommer, T. others Graph neural networks for materials science and chemistry. *Commun. Mater.* **2022**, *3*, 93. [CrossRef] [PubMed]
14. Wu, J.; Yan, F.; Jia, Q.; Wang, Q. QSPR for predicting the hydrophile-lipophile balance (HLB) of non-ionic surfactants. *Colloids Surf. A Physicochem. Eng. Asp.* **2020**, *611*, 125812. [CrossRef]
15. Camarda, K.V.; Bonnell, B.W.; Maranas, C.D.; Nagarajan, R. Design of surfactant solutions with optimal macroscopic proper-ties. *Comp. Chem. Eng.* **1999**, *23*, S467–S470. [CrossRef]
16. Huibers, P.D.; Shah, D.O.; Katritzky, A.R. Predicting Surfactant Cloud Point from Molecular Structure. *J. Colloid Interface Sci.* **1997**, *193*, 132–136. [CrossRef]
17. Zoeller, N.J.; Blankschtein, D. Development of User-Friendly Computer Programs To Predict Solution Properties of Single and Mixed Surfactant Systems. *Ind. Eng. Chem. Res.* **1995**, *34*, 4150–4160. [CrossRef]
18. Yuan, S.; Cai, Z.; Xu, G.; Jiang, Y. Quantitative Structure–Property Relationships of Surfactants: Critical Micelle Concentration of Anionic Surfactants. *J. Dispers. Sci. Technol.* **2002**, *23*, 465–472. [CrossRef]
19. Yao, H.L.; Shi, Y.C.; Yuan, S.L.; Li, G.Z. Quantitative Structure–property Relationship on Prediction of Cloud Point of Sur-factants. *J. Dispers. Sci. Technol.* **2009**, *30*, 1223–1230. [CrossRef]

20. Li, X.; Zhang, G.; Dong, J.; Zhou, X.; Yan, X.; Luo, M. Estimation of critical micelle concentration of anionic surfactants with QSPR approach. *J. Mol. Struct. THEOCHEM* **2004**, *710*, 119–126. [CrossRef]
21. Li, Y.; Xu, G.; Luan, Y.; Yuan, S.; Xin, X. Property Prediction on Surfactant by Quantitative Structure-Property Relationship: Krafft Point and Cloud Point. *J. Dispers. Sci. Technol.* **2005**, *26*, 799–808. [CrossRef]
22. Ghasemi, J.; Abdolmaleki, A.; Asadpour, S.; Shiri, F. Prediction of Solubility of Nonionic Solutes in Anionic Micelle (SDS) Using a QSPR Model. *QSAR Comb. Sci.* **2008**, *27*, 338–346. [CrossRef]
23. Wang, Z.W.; Huang, D.Y.; Gong, S.P.; Li, G.Z. Prediction on Critical Micelle Concentration of Nonionic Surfactants in Aqueous Solution: Quantitative Structure–property Relationship Approach. *Chin. J. Chem.* **2003**, *21*, 1573–1579. [CrossRef]
24. Mavaddat, M.; Riahi, S. A molecular structure based model for predicting optimal salinity of anionic surfactants. *Fluid Phase Equilibria* **2016**, *409*, 354–360. [CrossRef]
25. Ren, Y.; Zhao, B.; Chang, Q.; Yao, X. QSPR modeling of nonionic surfactant cloud points: An update. *J. Colloid Interface Sci.* **2011**, *358*, 202–207. [CrossRef]
26. Anoune, N.; Nouiri, M.; Berrah, Y.; Gauvrit, J.; Lanteri, P. Critical micelle concentrations of different classes of surfactants: A quantitative structure property relationship study. *J. Surfactants Deterg.* **2002**, *5*, 45–53. [CrossRef]
27. Ren, Y.; Liu, H.; Yao, X.; Liu, M.; Hu, Z.; Fan, B. The accurate QSPR models for the prediction of nonionic surfactant cloud point. *J. Colloid Interface Sci.* **2006**, *302*, 669–672. [CrossRef]
28. Kania, D.; Yunus, R.; Omar, R.; Rashid, S.A.; Jan, B.M.; Aulia, A. Adsorption of non-ionic surfactants on organoclays in drilling fluid investigated by molecular descriptors and Monte Carlo random walk simulations. *Appl. Surf. Sci.* **2020**, *538*, 148154. [CrossRef]
29. Welling, S.H.; Clemmensen, L.K.; Buckley, S.T.; Hovgaard, L.; Brockhoff, P.B.; Refs gaard, H.H. In silico modelling of per-meation enhancement potency in Caco-2 monolayers based on molecular descriptors and random forest. *Eur. J. Pharm. Biopharm.* **2015**, *94*, 152–159. [CrossRef]
30. Seddon, D.; Müller, E.A.; Cabral, J.T. Machine learning hybrid approach for the prediction of surface tension profiles of hydrocarbon surfactants in aqueous solution. *J. Colloid Interface Sci.* **2022**, *625*, 328–339. [CrossRef]
31. Creton, B.; Barraud, E.; Nieto-Draghi, C. Prediction of critical micelle concentration for per- and polyfluoroalkyl substances. *SAR QSAR Environ. Res.* **2024**, *35*, 309–324. [CrossRef]
32. 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. [CrossRef] [PubMed]
33. Moriarty, A.; Kobayashi, T.; Salvalaglio, M.; Angeli, P.; Striolo, A.; McRobbie, I. Analyz ing the accuracy of critical micelle con-centration predictions using deep learning. *J. Chem. Theory Comp.* **2023**, *19*, 7371–7386. [CrossRef] [PubMed]
34. Brozos, C.; Rittig, J.G.; Bhattacharya, S.; Akanny, E.; Kohlmann, C.; Mitsos, A. Graph neural networks for surfactant mul-ti-property prediction. *Colloids Surf. A Physico. chem. Eng. Asp.* **2024**, *694*, 134133. [CrossRef]
35. Brozos, C.; Rittig, J.G.; Bhattacharya, S.; Akanny, E.; Kohlmann, C.; Mitsos, A. Predicting the Temperature Dependence of Sur-factant CMCs Using Graph Neural Networks. *J. Chem. Theory Comp.* **2024**, *20*, 5695–5707. [CrossRef]
36. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J. others SciPy 1.0: Fundamental algo rithms for scientific computing in Python. *Nat. methods* **2020**, *17*, 261–272. [CrossRef]
37. Mańko, D.; Zdziennicka, A.; Jańczuk, B. Adsorption and Aggregation Activity of Sodium Dodecyl Sulfate and Rhamnolipid Mixture. *J. Surfactants Deterg.* **2016**, *20*, 411–423. [CrossRef]
38. Neys, B.; Joos, P. Equilibrium surface tensions and surface potentials of some fatty acids. *Colloids Surf. A Physicochem. Eng. Asp.* **1998**, *143*, 467–475. [CrossRef]
39. Fainerman, V.; Miller, R.; Möhwald, H. General relationships of the adsorption behavior of surfactants at the water/air interface. *J. Phys. Chem. B* **2002**, *106*, 809–819. [CrossRef]
40. Heid, E.; Greenman, K.P.; Chung, Y.; Li, S.-C.; Graff, D.E.; Vermeire, F.H.; Wu, H.; Green, W.H.; McGill, C.J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2023**, *64*, 9–17. [CrossRef]
41. Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M. others Ana-lyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388. [CrossRef]
42. McGill, C.; Forsuelo, M.; Guan, Y.; Green, W.H. Predicting Infrared Spectra with Message Passing Neural Networks. *J. Chem. Inf. Model.* **2021**, *61*, 2594–2609. [CrossRef] [PubMed]
43. Buterez, D.; Janet, J.P.; Kiddle, S.J.; Oglic, D.; Li'o, P. Transfer learning with graph neu ral networks for improved molecular property prediction in the multi-fidelity setting. *Nat. Commun.* **2024**, *15*, 1517. [CrossRef] [PubMed]
44. Fu, L.; Shi, S.; Yi, J.; Wang, N.; He, Y.; Wu, Z.; Peng, J.; Deng, Y.; Wang, W.; Wu, C. others ADMETlab 3.0: An updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support. *Nucleic Acids Res.* **2024**, gkae236.
45. Dodda, L.S.; de Vaca, I.C.; Tirado-Rives, J.; Jorgensen, W.L. LigParGen web server: An automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res.* **2017**, *45*, W331–W336. [CrossRef] [PubMed]
46. Wang, A.Y.T.; Murdock, R.J.; Kauwe, S.K.; Oliynyk, A.O.; Gurlo, A.; Brgoch, J.; Persson, K.A.; Sparks, T.D. Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices. *Chem. Mater.* **2020**, *32*, 4954–4965. [CrossRef]
47. Sedev, R. Limiting Area per Molecule of Nonionic Surfactants at the Water/Air Interface. *Langmuir* **2000**, *17*, 562–564. [CrossRef]

48. Cang, Z.; Wei, G.W. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comp. Biol.* **2017**, *13*, e1005690. [CrossRef]
49. Smit, B.; Schlijper, A.G.; Rupert, L.A.M.; Van Os, N.M. Effects of chain length of surfactants on the interfacial tension: Molecular dynamics simulations and experiments. *J. Phys. Chem.* **1990**, *94*, 6933–6935. [CrossRef]
50. Holmberg, K. *Handbook of Applied Surface and Colloid Chemistry*, 1st ed.; John Wiley & Sons: West Sussex, UK, 2001; Chapter 19.
51. Fameau, A.-L.; Cousin, F.; Saint-Jalmes, A. Morphological Transition in Fatty Acid Self-Assemblies: A Process Driven by the Interplay between the Chain-Melting and Surface-Melting Process of the Hydrogen Bonds. *Langmuir* **2017**, *33*, 12943–12951. [CrossRef]
52. Czajka, A.; Hazell, G.; Eastoe, J. Surfactants at the Design Limit. *Langmuir* **2015**, *31*, 8205–8217. [CrossRef]
53. Liu, X.; Zhao, Y.; Li, Q.; Jiao, T.; Niu, J. Adsorption behavior, spreading and thermal stability of anionic-nonionic surfactants with different ionic headgroup. *J. Mol. Liq.* **2016**, *219*, 1100–1106. [CrossRef]
54. Chander MP, C.; Kartick, J.; Gangadhar, P. Vijayachari, Ethno medicine and healthcare practices among Nicobarese of Car Nicobar-an indigenous tribe of Andaman and Nicobar Islands, *J. Ethnopharmacol.* **2014**, *158*, 18–24. [CrossRef] [PubMed]
55. Kancharla, S.; Jahan, R.; Bedrov, D.; Tsianou, M.; Alexandridis, P. Role of chain length and electrolyte on the micellization of anionic fluorinated surfactants in water. *Colloids Surf. A Physicochem. Eng. Asp.* **2021**, *628*, 127313. [CrossRef]
56. Sadeghi, R.; Shahabi, S. A comparison study between sodium dodecyl sulfate and sodium dodecyl sulfonate with respect to the thermodynamic properties, micellization, and interaction with poly(ethylene glycol) in aqueous solutions. *J. Chem. Thermodyn.* **2011**, *43*, 1361–1370. [CrossRef]
57. Corkill, J.M.; Goodman, J.F.; Harrold, S.P. Thermodynamics of micellization of non-ionic detergents. *Trans. Faraday Soc.* **1964**, *60*, 202–207. [CrossRef]
58. Carless, J.; Challis, R.; Mulley, B. Nonionic surface-active agents. Part V. The effect of the alkyl and the polyglycol chain length on the critical micelle concentration of some monoalkyl polyethers. *J. Colloid Sci.* **1964**, *19*, 201–212. [CrossRef]
59. Rosen, M.J.; Cohen, A.W.; Dahanayake, M.; Hua, X.Y. Relationship of structure to 15 properties in surfactants. 10. Surface and thermodynamic properties of 2-dodecyloxypoly (ethenoxyethanol) s, $C_{12}H_{25}(OC_2H_4)_xOH$, in aqueous solution. *J. Phys. Chem.* **1982**, *86*, 541–545. [CrossRef]
60. Zdziennicka, A.; Szymczyk, K.; Krawczyk, J.; Ja´nczuk, B. Activity and thermodynamic parameters of some surfactants adsorption at the water–air interface. *Fluid Ph. Equilib.* **2012**, *318*, 25–33. [CrossRef]
61. Dahanayake, M.; Cohen, A.W.; Rosen, M.J. Relationship of structure to properties of *surfactant*s. 13. Surface and thermodynamic properties of some oxyethylenated sulfates and sulfonates. *J. Phys. Chem.* **1986**, *90*, 2413–2418. [CrossRef]
62. Rosen, M.J.; Kwan, C.-C. Relationship of structure to properties in *surfactant*s. 8. Synthesis and properties of sodium 3-alkyltetrahydropyranyl 4-sulfates. *J. Phys. Chem.* **1979**, *83*, 2727–2732. [CrossRef]
63. Mędrzycka, K.; Zwierzykowski, W. Adsorption of Alkyltrimethylammonium Bromides at the Various Interfaces. *J. Colloid Interface Sci.* **2000**, *230*, 67–72. [CrossRef] [PubMed]
64. Nguyen, C.V.; Nguyen, T.V.; Phan, C.M. Adsorption of alkyltrimethylammonium bromide surfactants at the air/water inter-face. *Int. J. Heat Mass Transf.* **2017**, *106*, 1035–1040. [CrossRef]
65. Zhu, Y.P.; Rosen, M.J.; Vinson, P.K.; Morrall, S.W. Surface properties of N-alkanoyl-N-methyl glucamines and related materi-als. *J. Surfactants Deterg.* **1999**, *2*, 357–362. [CrossRef]
66. Shinoda, K.; Yamanaka, T.; Kinoshita, K. Surface Chemical Properties in Aqueous Solutions of Non-ionic Surfactants Octyl Glycol Ether, α-Octyl Glyceryl Ether and Octyl Glucoside. *J. Phys. Chem.* **1959**, *63*, 648–650. [CrossRef]
67. Zhao, F.; Rosen, M.J. Relationship of structure to properties of surfactants. 12. Synthesis and surface properties of long-chain 2-pyridinium alkanoates. *J. Phys. Chem.* **1984**, *88*, 6041–6043. [CrossRef]
68. Chevalier, Y.; Storet, Y.; Pourchet, S.; Le Perchec, P. Tensioactive properties of zwitterionic carboxybetaine amphiphiles. *Langmuir* **1991**, *7*, 848–853. [CrossRef]
69. Kumpulainen, A.J.; Persson, C.M.; Eriksson, J.C. Headgroup and Hydrocarbon Tail Effects on the Surface Tension of Sugar-Based Surfactant Solutions. *Langmuir* **2004**, *20*, 10935–10942. [CrossRef]
70. Varga, I.; Mészáros, R.; Stubenrauch, C.; Gilányi, T. Adsorption of sugar *surfactant*s at the air/water interface. *J. Colloid Interface Sci.* **2012**, *379*, 78–83. [CrossRef]
71. Gentle, T.E.; Snow, S.A. Absorption of Small Silicone Polyether Surfactants at the Air/Water Surface. *Langmuir* **1995**, *11*, 2905–2910. [CrossRef]
72. Shinoda, K.; Hato, M.; Hayashi, T. Physicochemical properties of aqueous solutions of fluorinated surfactants. *J. Phys. Chem.* **1972**, *76*, 909–914. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.