Predicting Critical Micelle Concentrations for Surfactants using Graph Convolutional Neural Networks

Shiyi Qin,[¶] Tianyi Jin,[¶] Reid C. Van Lehn,[¶] and Victor M. Zavala[¶]

[¶]Department of Chemical and Biological Engineering

University of Wisconsin – Madison, 1415 Engineering Drive, Madison, WI 53706, USA

Abstract

Surfactants are amphiphilic molecules that are widely used in consumer products, industrial processes, and biological applications. A critical property of a surfactant is the critical micelle concentration (CMC), which is the concentration at which surfactant molecules undergo cooperative self-assembly in solution. Notably, the primary method to obtain CMCs experimentally—tensiometry—is laborious and expensive. In this work, we show that graph convolutional neural networks (GCNs) can predict CMCs directly from the surfactant molecular structure. Specifically, we developed a GCN architecture that encodes the surfactant structure in the form of a molecular graph and trained it using experimental CMC data. We found that the GCN can predict CMCs with higher accuracy than previously proposed methods and that it can generalize to anionic, cationic, zwitterionic, and nonionic surfactants. Molecular saliency maps revealed how atom types and surfactant molecular substructures contribute to CMCs and found this to be in agreement with physical rules that correlate constitutional and topological information to CMCs. Following such rules, we proposed a small set of new surfactants for which experimental CMCs are not available; for these molecules, CMCs predicted with our GCN exhibited similar trends to those obtained from molecular simulations. These results provide evidence that GCNs can enable high-throughput screening of surfactants with desired self-assembly characteristics.

Introduction

Surface-active agents (surfactants) are amphiphiles that consist of a lyophilic head and a lyophobic tail. Depending on the charge carried by the polar head group, surfactants can be categorized as nonionic, cationic, anionic, or zwitterionic (Fig. 1, a-d).[1] Given their ability to reduce surface tension and increase the solubility of insoluble or sparingly soluble substances,[2] surfactants are widely used for wetting, foaming, cleaning, emulsification, solubilization, lubrication, and flotation in industrial applications such as pharmaceuticals, personal care, detergents, coatings, food, and agriculture.[3,4,5] Surfactants have also been utilized for green chemistry, bioengineering, and other chemically relevant research fields; for example, surfactants have been shown to enhance oil recovery,[6] reduce environmental footprints in pharmaceuticals,[7] improve drug delivery effectiveness,[8] and enable catalysis in aqueous media.[9]

When dissolved in water, surfactant monomers will undergo a cooperative aggregation process, called self-assembly, to form spherical micelles or related aggregate structures.[10] Self-assembly is thermodynamically favorable because the micelle structure minimizes the water-exposed hydrophobic surface area by orienting hydrophilic surfactant head groups towards the aqueous phase and positioning hydrophobic surfactant tail groups within the micelle core (Fig. 1, e-f).[1] The formation of micelles in a solution can induce significant changes in key solution properties including the electrical conductivity, surface tension, light scattering, and reactivity.[1, 10] Consequently, predicting conditions under which surfactants self-assemble is important for surfactant selection and design.[11] A critical parameter that characterizes surfactant self-assembly behavior is the critical micelle concentration (CMC), which is the minimum surfactant concentration at which self-assembly occurs.[1, 10] CMCs are strongly influenced by the molecular structure of the surfactant (such as the tail length and the head area); for instance, it is typically

observed that the shorter the hydrophobic tail and the larger the hydrophilic head, the higher the CMC.[3, 10] However, this type of qualitative analysis cannot easily translate into quantitative predictions of CMCs, which limits the screening and rational design of surfactants. Moreover, the primary method to obtain CMCs experimentally—tensiometry—is laborious and expensive.[12-14]

As an alternative to experiments, computational methods such as molecular dynamics (MD) simulations[15-17] and descriptor-based quantitative structure-property relationship (QSPR) models[12, 18-23] have been used to predict CMCs. These approaches have been shown to predict CMCs with relatively high accuracy, but they have a number of limitations; for instance, MD simulations usually require large system sizes, long simulation times, and assumptions regarding the number of surfactants within a micelle,[15-17] whereas QSPR models are often applicable to a single class of surfactants and may need density functional theory calculations to obtain quantum-chemical molecular descriptors.[24] Recent advances in machine learning methods for molecular property prediction can help overcome some of these obstacles. Goh et al. used 2D molecule "images" as input to convolutional neural networks (CNNs) to predict toxicity, activity, and solvation properties of different molecules.[25] Hirohara et al. used one-hot-encoded simplified molecular-input line-entry system (SMILES) strings combined with molecular descriptors as CNN inputs to predict functional substructures.[26] Wu et al. employed graph neural networks (GNNs) trained on molecular graphs to predict various molecular properties.[27] GNNs have similarly been shown to outperform other machine learning methods, including logistic regression, support vector machine, kernel ridge regression, and random forests in different benchmark datasets such as Tox21[28] (for toxicity classification) and ESOL[29] (for water solubility regression).[27, 30, 31] Most studies in this area, however, have focused on predicting common molecular properties for which a large number of data samples are available (such as solubility and toxicity).[27, 30-32] To the best of our knowledge,

GNNs have not been used for CMC prediction. The CMC is also distinct from these related properties because it describes the *cooperative* behavior of a collection of molecules in solution, rather than the property of a single molecule.

In this study, we show that graph convolutional neural networks (GCNs),[33] a basic architecture in the family of GNNs, can predict CMC values directly from the molecular graph of a surfactant monomer. Molecular graphs are intuitive and flexible data representations that encode information on component atoms and atom connectivity (*e.g.*, they encode topological information of the molecular structure). GCNs extract features from molecular graph representations by using convolutional operations that aggregate encoded information from molecular structures. We hypothesize that GCNs can capture important structural information that can enable CMC predictions; this hypothesis is motivated by the observation that QSPR models use topological and constitutional descriptors of a surfactant to predict CMCs.[12, 18, 19] Moreover, given that GCNs perform convolutions at an atomic level, they provide flexibility to handle surfactants of different sizes without the need for artificial data manipulations (*e.g.*, CNN models require zero-padding to handle molecules of different sizes).

We present a GCN architecture that was first trained and tuned using a dataset that only contains nonionic surfactants. This architecture is used to confirm the ability of the model to extract hidden molecular features that enable CMC predictions. We then trained the architecture using an expanded dataset that contains nonionic, anionic, cationic, and zwitterionic surfactants. We show that the GCN model achieves a higher prediction accuracy on a broader spectrum of surfactants than previous QSPR models reported in the literature. However, one of the obstacles in understanding the predictive limitations of the proposed approach is the lack of available experimental data. To address this issue, we created a synthetic dataset that mimics the basic

structural features of surfactants; this approach allowed us to construct a large and controlled dataset to examine whether the GCN architecture has the ability to capture the intrinsic topological and constitutional information of a molecule. We also used gradient information to generate molecular saliency maps and with this gain understanding of how a surfactant structure influences its CMC. Finally, we illustrate the potential of our approach to enable molecular design and screening by deriving new surfactant structures from the existing ones, then validating GCN predictions by comparing them to CMC trends obtained from complementary molecular simulations.

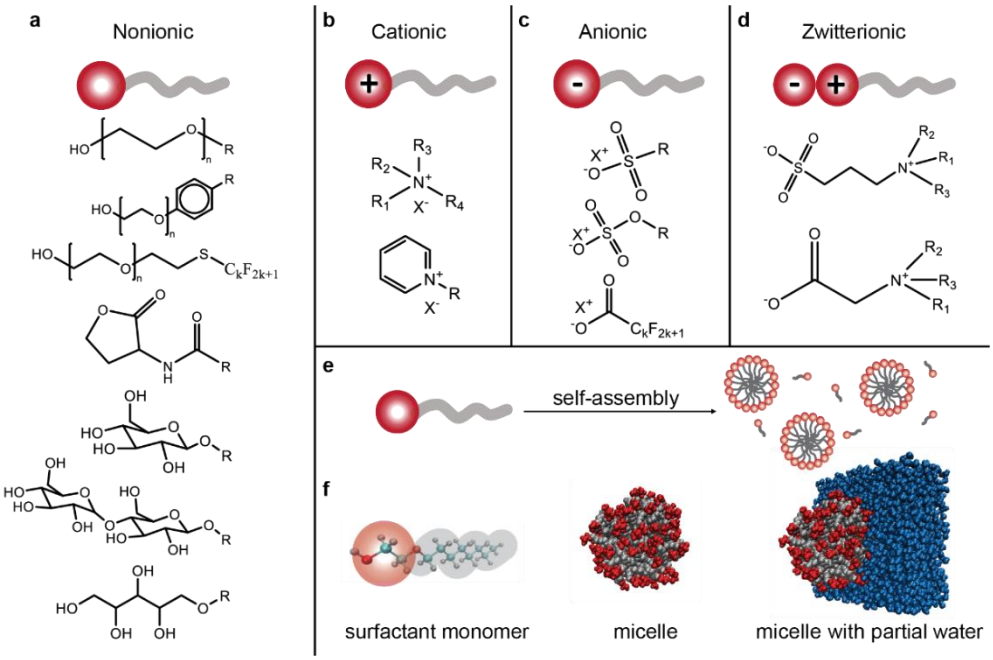


Figure 1. Overview of surfactant molecular structures and self-assembly process in micelles. (a)-(d) Sample structures of four classes of surfactants included in the experimental dataset. Surfactants are categorized by the properties of their head groups as nonionic (a), cationic (b), anionic (c), or zwitterionic (d). Additional structures not shown here are listed in SI Table S1. (e) Surfactant monomers aggregate into spherical micelles in water with hydrophilic head groups facing towards the solvent and hydrophobic tail groups sequestered inside the micelle core. (f) Snapshots of a surfactant micelle from a representative molecular dynamics simulation with water shown in blue.

Results and Discussion

Preparation of the experimental surfactant CMC dataset

We gathered experimental CMC data for 202 surfactants, including 122 nonionic surfactants, 35 cationic surfactants, 34 anionic surfactants, and 11 zwitterionic surfactants (Fig. 1, a-d), from multiple literature sources to form our dataset.[1,12,17,34] All CMCs were measured at room temperature (between 20-25 °C) in water and converted to log CMC values (base 10). The dataset was split into training (~90%) and testing (~10%) subsets, and we performed *k*-fold cross-validation (CV) for hyperparameter tuning. In *k*-fold CV, the training subset was randomly divided into *k* groups. The model was then trained *k* times with a different group held out each time as the validation set and the remaining *k-1* groups used as a training set. The value of *k* was determined such that the training subset and the validation subset contained approximately 80% and 10% samples of the original dataset, respectively.

Since past approaches used for CMC predictions (*e.g.*, QSPR models[12, 18, 19]) typically focus on a single class of surfactant, we first conducted baseline predictions on a subset of the original dataset containing only nonionic surfactants to compare to past results. This subset was partitioned into 100 training samples, 10 validation samples (11-fold CV), and 12 testing samples. To analyze the generalizability of the GCN model to multiple classes of surfactants, we used the full dataset containing all nonionic, anionic, cationic, and zwitterionic surfactants. This dataset was partitioned into 160 training samples, 20 validation samples (9-fold CV), and 22 testing samples. The testing samples were selected using stratified sampling[35] to include surfactants that cover a wide domain of the input CMCs and were held out during model training and validation. SI Table S1 lists all surfactants studied and indicates the surfactants that were selected as test

samples. All data and scripts needed to reproduce the results can be found here: https://github.com/zavalab/ML/tree/master/CMC_GCN.

Molecular graph representation

Surfactant structures were converted to molecular graphs and these were provided as input to the GCN. In this data representation, atoms were represented as nodes and bonds as edges, as illustrated in Figure 2a. Hydrogen atoms were treated implicitly. Each node encoded atomic information such as the atom type, degree (number of connected edges to it), and charge in the form of a feature matrix (Fig. 2b); for instance, the atom type was one-hot encoded into 43 categorical features based on the predefined list of chemical elements. Edge features (*e.g.*, bond type) were not explicitly included but were captured by atom features such as hybridization and aromaticity. This representation resulted in 74 features per atom, with a full list summarized in SI Table S2; as a comparison, a previous study[12] computed over 300 constitutional, topological, geometrical, and quantum-chemical descriptors to develop a QSPR model. Besides atom features, the molecular graph encodes topology through an adjacency matrix that captures atom connectivity (Fig. 2c). This data representation thus differs from that used in QSPR models, in which topological information is only indirectly captured via molecular-level descriptors (*e.g.*, topological indices).[12, 19, 20] For each cationic or anionic surfactant, the counterion was represented as a node disconnected from other nodes in the molecular graph.

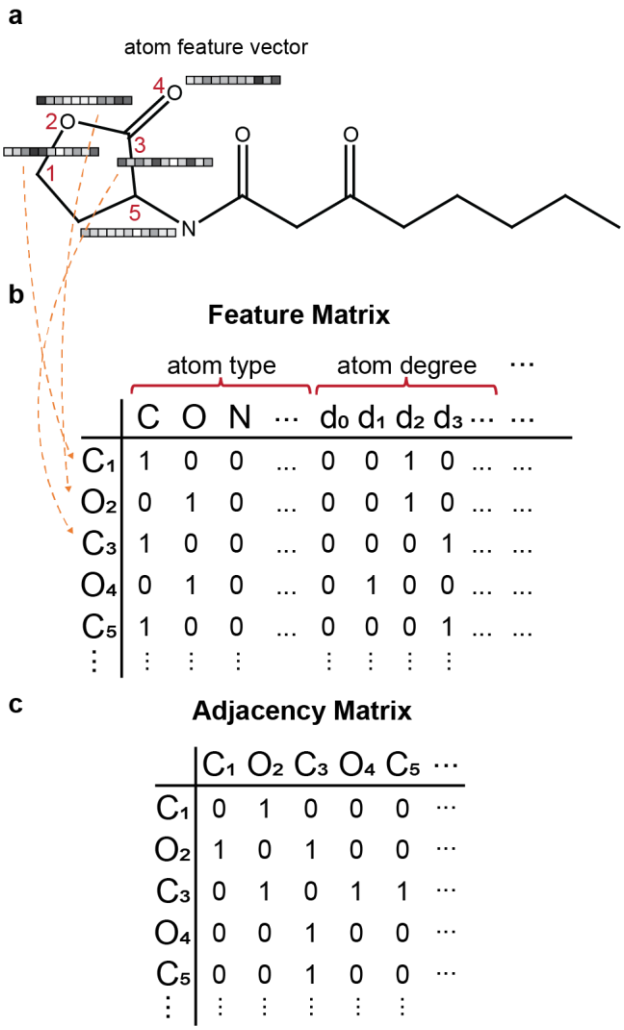


Figure 2. Data representation of an example surfactant. (a) The molecular graph of an example surfactant monomer. Atoms are represented as nodes and bonds are represented as undirected edges. Hydrogen atoms are implicit. The atom feature vectors are illustrated as colored bars next to each atom. (b) Atom features are encoded as fixed-length atom feature vectors. The presence or absence of each feature is labeled as "1" or "0", respectively, resulting in a feature matrix for each molecule with dimensions given by the number of atoms and the number of features per atom. (c) The adjacency matrix shows the connectivity between atoms; a value of one is assigned to the matrix entry (i, j) if there is a bond that connects atom i and atom j.

GCN for CMC predictions of nonionic surfactants

The proposed GCN architecture consists of two graph convolutional layers, one average pooling layer, two fully-connected hidden layers, and one final output layer (Fig. 3). A graph convolution layer updates each atom by aggregating the features of itself and of its neighbors and maps the

updated features into a hidden layer with 256 hidden features. The hidden features are generated from nonlinear transformations (linear mapping with trainable parameters followed by ReLU activation) of the updated features. The GCN model contains a total of 216,833 parameters, corresponding to operators of the graph convolutions as well as bias terms. Because the number of parameters is relatively large compared with the size of the dataset, as measures to prevent overfitting, we used early-stopping to terminate training when validation performance starts to degrade and CV to estimate the predictive power of the model architecture on unseen data. The GCN architecture was determined by hyperparameter tuning using the nonionic surfactant subset and by performing 11-fold CV (Fig. 4a). The root-mean-squared-errors (RMSEs) between the experimental and predicted log CMC values (obtained with 11-fold CV) have a mean value of 0.32. Although we were able to achieve a lower average CV RMSE of 0.30 when we increased both the number of convolutional layers and the number of fully-connected hidden layers, the median and the standard error did not improve (SI Table S3). Therefore, we decided to select a simpler architecture for less computational time and potentially better model interpretability. For this GCN architecture, RMSEs for 9 out of the 11 models trained during CV fall between 0.15 and 0.34, with only one major outlier at 0.90 and one minor outlier at 0.47. RMSEs are summarized in SI Figure S1.

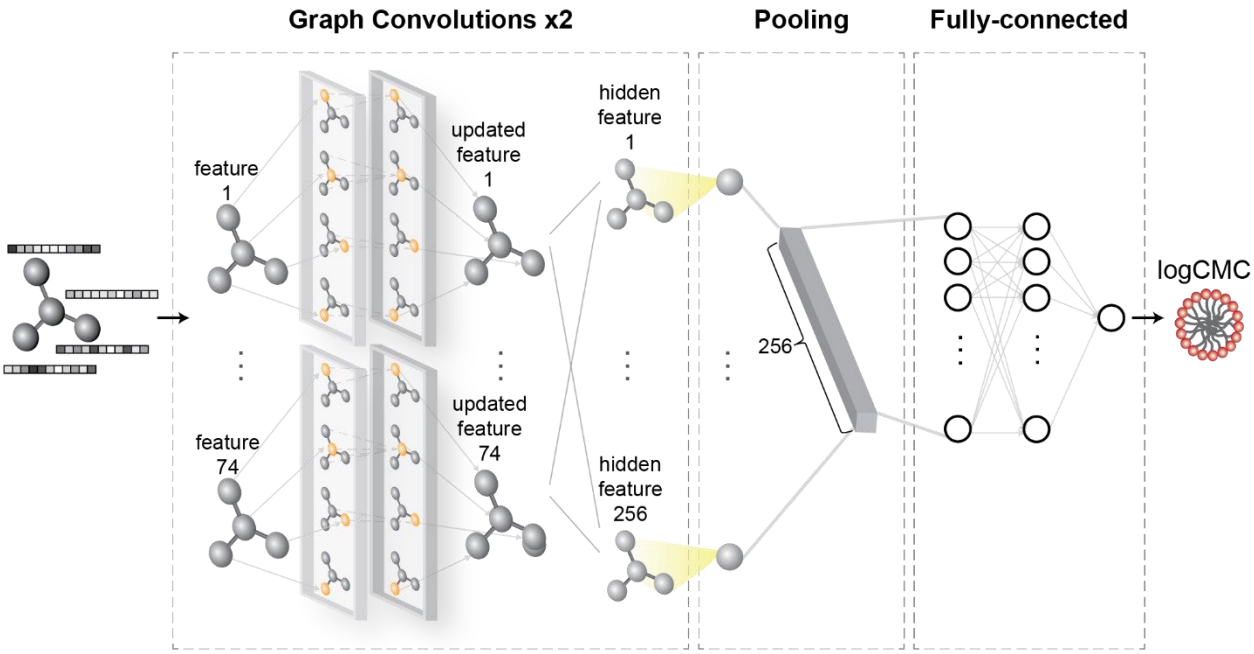


Figure 3. GCN architecture. The proposed GCN takes a molecular graph as input, convolutes across each input twice (by updating the atom features and mapping the updated features into hidden features), averages the atom-level hidden features into molecule-level hidden features, and calculates the final prediction of the log CMC value from fully-connected neural network layers.

We tested the ability of the GCN model to generalize to new data by training the model using all training samples then testing on held-out test samples, as illustrated in Figure 4a. For the nonionic dataset, 12 test samples were selected to include various nonionic surfactant structures, covering samples with structures listed in Figure 1a as well as other surfactant classes such as glucamine and lactobioamide. Each test sample prediction was calculated as the average of the prediction results from three training runs with different parameter initializations. The test dataset has an RMSE of 0.23 ($R^2 = 0.96$) and a best-fit slope of 0.95. Figure 4b shows a parity plot between the predicted and experimental log CMC values for the training and testing sets. The RMSE of the test data lies in the middle range of the CV RMSEs, indicating that the model is not overfitted. Our model performs better than a previous QSPR model developed to predict the CMC values of 108 sugar-based nonionic surfactants, for which the best test RMSE reported was 0.32

($R^2 = 0.93$).[12] Furthermore, our dataset encompassed a wider variety of nonionic surfactants (other than sugar-based ones), such as fluorinated thiol ethoxylates and acyl-homoserine lactones.

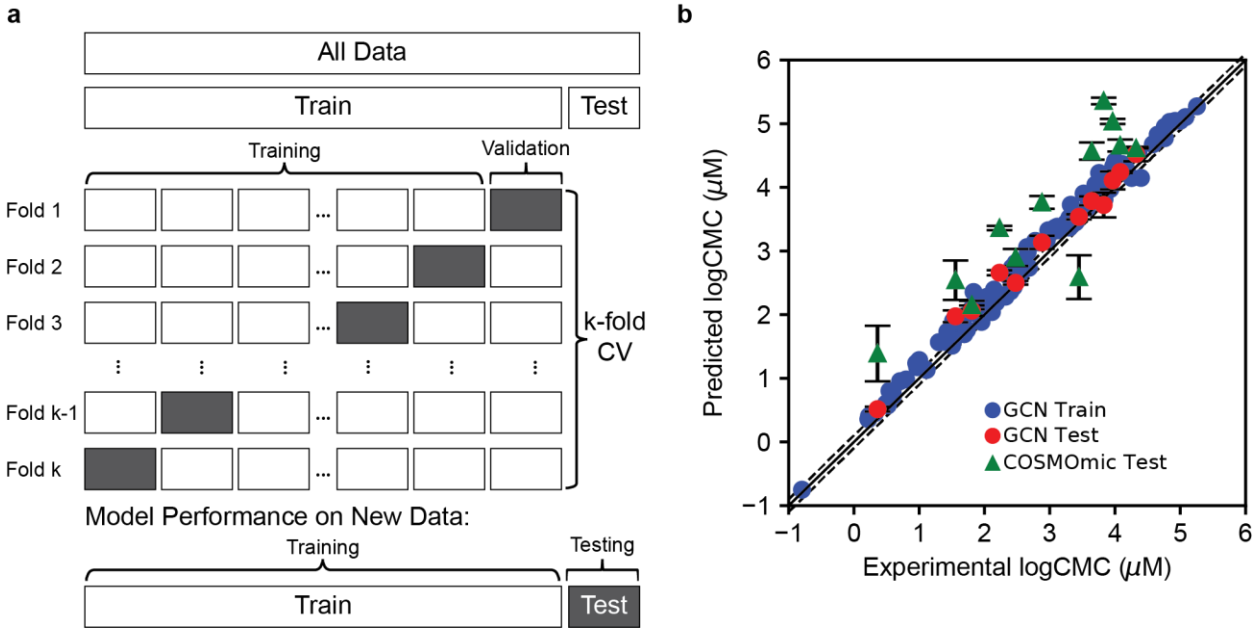


Figure 4. GCN model validation and testing for nonionic surfactant dataset. (a) Training, validation, and testing procedure for hyperparameter tuning. We first divide the dataset into training and testing sets. The training set is then split into training and validation sets for a *k*-fold cross-validation (CV) procedure. After the architecture is determined, we train the GCN on the entire training dataset and test the model on the held-out test set to evaluate model performance on the new data. (b) Parity plot between the predicted and experimental log CMCs for nonionic surfactants. The best-fit slope is 0.95 ($R^2 = 0.96$) for the GCN test set and 0.92 ($R^2 = 0.39$) for the COSMOmic test set. The dashed lines show a 10% error range of log CMCs. The error bars of the GCN test data are standard errors computed from three training trials with different parameter initializations. The error bars shown for the COSMOmic test data are the standard errors computed from three different monomer configurations obtained from molecular dynamics simulations.

Molecular simulations for CMC predictions as a validation method

We compared the predictive performance of molecular simulations on the same test dataset using the COnductor-like Screening Model for Realistic Solvation (COSMO-RS) model and its extension, COSMOmic, which can compute aggregation free energies.[36] Atomistic MD simulations were first conducted to obtain input structures for each surfactant monomer and corresponding spherical micelle in the test set (SI Fig. S2). These structures were used as input for COSMOmic calculations to compute the free energy of micellization in order to obtain the CMC.

This simulation protocol is faster than experiments and can be applied to any surfactant molecule (thus providing the flexibility to study the effects of structure on CMC values). However, the protocol requires an aggregation number (*e.g.*, the number of surfactants within the micelle), which we assumed to be 100 for all surfactants modeled in this study because this value is typical of nonionic surfactants.[37] The RMSE obtained from the COSMOmic calculations is 0.91 with a best-fit slope of 0.92, which is less accurate than the GCN predictions. When predicting large log CMC values (log CMC > 4), this method deviates more from experimental values, which in part could be due to variations in the aggregation number. While COSMOmic tends to overestimate log CMC values, in general, it predicts the correct trend. Therefore, this method can be used as an additional source of information to validate trends in predicted CMC values for newly designed surfactants for which experimental data are not available.

GCN for CMC predictions for all surfactants

We trained the same GCN architecture on the full dataset containing all four classes of surfactants and performed 9-fold CV. Instead of tuning hyperparameters, CV was used to compare the model performance to that of the previous dataset with only nonionic surfactants. The resulting CV RMSE on all types of surfactants has a mean value of 0.39 with no significant outliers. The majority of the CV RMSEs lie in the range of 0.28-0.45, as summarized in SI Figure S1. We again tested model performance on a test dataset, which contains the same 12 nonionic test samples as well as 4 additional cationic, 4 anionic, and 2 zwitterionic samples. Figure 5 shows a parity plot between the experimental and predicted log CMC values for the training and testing sets. We found that the average CV RMSE is 0.30 with a best-fit slope of 0.91. The RMSE is higher than that of the model trained on nonionic surfactants (as expected). Cationic surfactants have the lowest test

RMSE (0.07), followed by nonionic (0.18) and anionic (0.32), and the model performs worst for zwitterionic surfactants (0.76). The most significant outlier is the zwitterionic surfactant shown in Figure 5, which may be due to the presence of long alkyl groups (with a backbone of 22 atoms). Another potential reason for the high test RMSE obtained in zwitterionic surfactants may be the small number of test samples. The parity plot also suggested a slightly lower accuracy for surfactants with relatively large log CMC values (> 4.5), as also observed in the COSMOmic calculations. Despite the major outlier found for a zwitterionic surfactant, the overall predictability of the GCN model still outperforms that of a prior QSPR model[12] developed for only sugar-based nonionic surfactants. The differences in the molecular structures found in our dataset further highlight the wide variety of surfactants that the GCN model can capture. To the best of our knowledge, none of the previously reported QSPR models[12, 18-22] have tried to predict CMCs for all classes of surfactants using a single model; as such, the proposed GCN provides a significant development in surfactant CMC prediction.

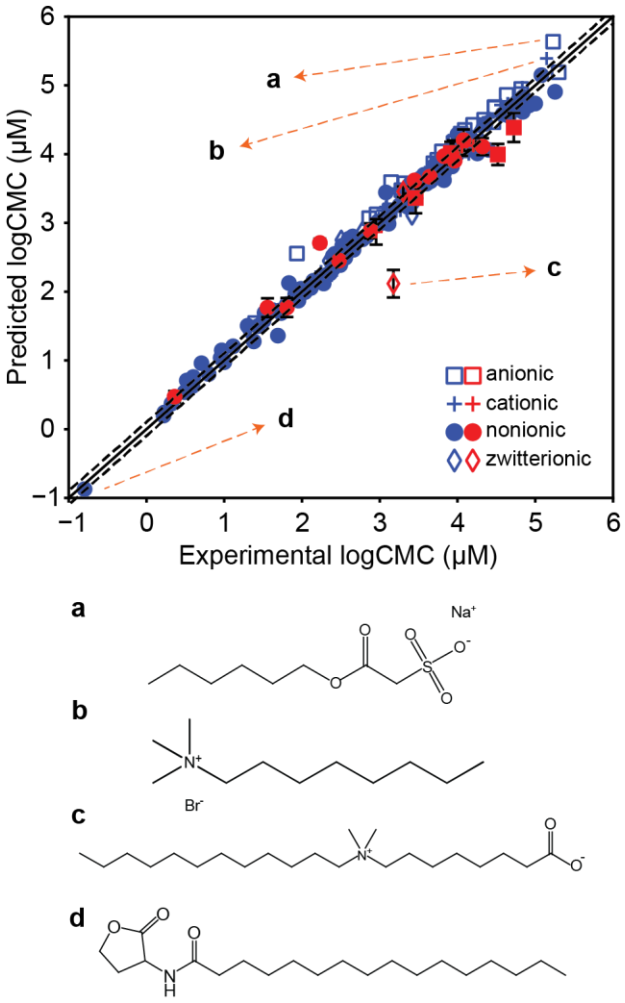


Figure 5. GCN predictability in all classes of surfactants. Parity plot between the predicted and experimental log CMC values (training data in blue and test data in red). The best-fit slope of the test data is 0.91 ($R^2 = 0.92$), and the test RMSE is 0.30. Molecular structures are shown for the selected extreme points. Structure (a) is an anionic surfactant (minor outlier) with a high log CMC value. Structure (b) is a cationic surfactant (minor outlier) with a high log CMC value. Structure (c) is a zwitterionic surfactant (major outlier). Structure (d) is a nonionic surfactant with a low log CMC value.

Systematic analysis using synthetic molecular structures

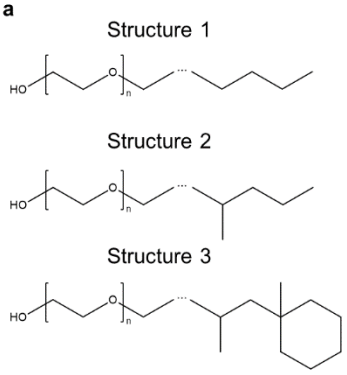
Although the proposed GCN shows promising results, the model might suffer from overfitting given the limited size of the experimental dataset. Therefore, to further validate the assumption that the GCN can capture structural information of surfactants when trained on more data samples, we studied the model performance on a synthetic dataset that encompasses 1,820 human-generated molecules. With control over the length of alkyl backbones as well as the quantity and location of

functional groups such as alkyl branches and rings, we developed three types of synthetic molecules and assigned three types of synthetic labels to each of the synthetic molecules based on its atom constitution and structure (details are provided in the Supplemental Information). The methodology used for generating the synthetic molecules captures three types of surfactant-like structures (Fig. 6a): (1) head-tail linear structure, (2) head-tail linear structure with single- or double-branching, and (3) head-tail linear structure (with and without branches) combined with a cyclohexane group. "Head" represents a surfactant head group that is constituted by linearly connected ethoxy groups. "Tail" represents a surfactant tail group that is constituted by a linear alkyl chain. "Branch" represents a randomly positioned side chain which is either a methyl or ethyl group. After the synthetic molecule structures were generated, three types of synthetic properties were calculated and used as prediction labels; here, we used three linear equations that capture constitutional, topological, and combined information of a synthetic molecule, respectively. The linear equations are dependent on molecular descriptors that have been used in QSPR models[12, 19, 38, 39]; the constitutional descriptors are number of C, number of O, and number of rings while the topological descriptors are Balaban index[40] (a measure of average distance-based connectivity) and Bertz CT index[41] (a measure of molecular complexity). Each descriptor was rescaled to obtain values between 0 and 1 and random weights were assigned to construct the linear equations, as summarized in SI Table S4.

Because the synthetic labels were computed from different equations, the magnitude of the labels may vary from subset to subset. As such, we used the slope of the parity plot between the synthetic labels and the predicted values to compare model performance instead of the RMSE. We used the proposed architecture to train the GCN with 10-fold CV, leading to 1,638 training samples and 182 validation samples in each CV fold. Overall, we found that the GCN predictions were

most accurate if the prediction label was a function of both constitutional and topological descriptors when all the synthetic molecules were used for training (Fig. 6b). These results indicate that a GCN architecture can effectively capture the topology of a molecule, regardless of the molecule size and structure (unlike QSPR models, which are usually structure-specific). In particular, for the synthetic molecules with linear structure, the trained GCN can make near-perfect predictions if the labels are dependent on topological or combined descriptors. The model also showed significant improvement for the synthetic molecules with rings when topological information plays a role in the molecular property of interest. In the case of CMC, we can infer that GCN serves as a more effective approach to make predictions than QSPR models given its ability to extract the same type of descriptors (constitutional and topological) that would be recognized well by a QSPR model without the need of explicit descriptor calculations.

a



b

Best-fit slope of the parity plot from 10-fold CV

Data \ Label	Constitutional	Topological	Combined
Structure 1	1.013±0.007	1.001±0.005	1.000±0.004
Structure 2	1.003±0.004	0.993±0.004	0.993±0.004
Structure 3	0.982±0.003	1.000±0.004	0.995±0.005
All synthetic molecules	**0.998±0.001**	**0.999±0.002**	**1.000±0.001**

Figure 6. GCN prediction performance on synthetic data. (a) Example structures of three types of synthetic molecules. (b) The GCN architecture was trained on all the synthetic molecules for each of the three synthetical labels. For each CV fold, the slope of the best-fit line of the validation data was recorded, and the averaged CV slope was then calculated.

Molecular saliency maps

We computed molecular saliency maps to further understand information that the GCN identifies in molecular structures to make predictions. The gradients of input atom features were first calculated and summed for each node, followed by normalization between -1 and 1. Figure 7 shows saliency maps computed for example surfactants that represent each of the four classes of surfactants. Atoms (nodes in the graph representation) are colored based on their normalized gradients, with red indicating more positive contributions and blue indicating more negative contributions to the CMC. The saliency maps confirm that polar atoms (such as O and N) contribute to higher CMC values whereas nonpolar atoms (such as C) contribute to lower CMC values, in agreement with qualitative expectations. From the saliency maps, we also confirm that topological information is being exploited by the GCN. For example, the branched tail nodes in sample d exhibits lighter blue colors compared with the unbranched tail nodes in samples a, b, and c. These patterns match the physical intuition that a surfactant tends to have a lower CMC value if it has a long and unbranched tail group or a small head group area.[10]

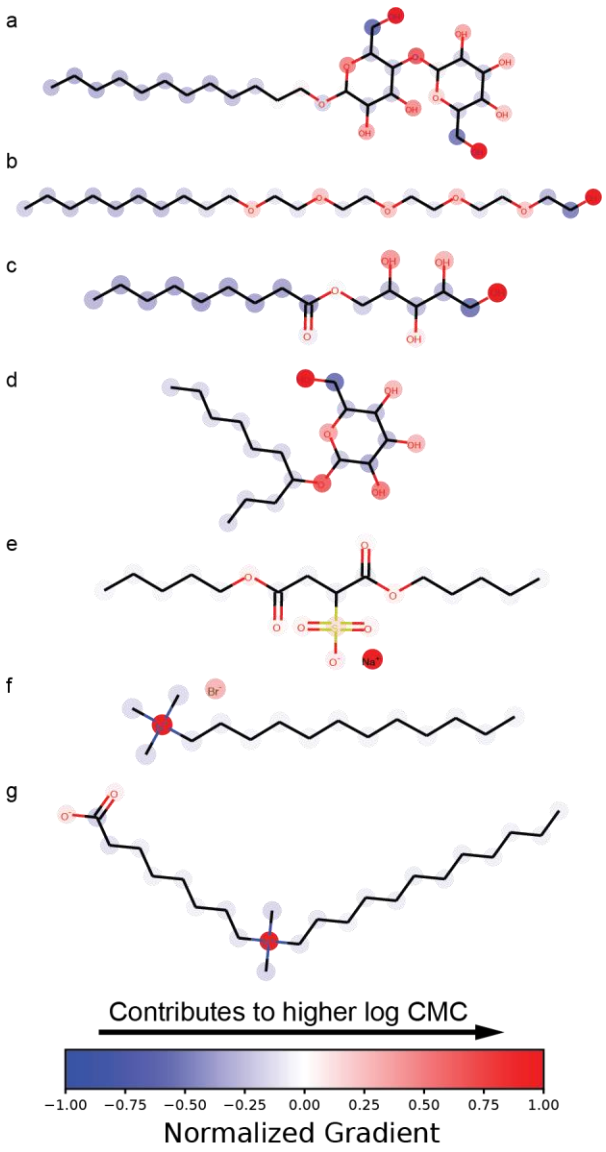


Figure 7. Molecular saliency maps. Selected examples from nonionic (a-d), cationic (e), anionic (f), and zwitterionic (g) surfactants. The gradient values are calculated for each node, followed by normalization between -1 and 1 where the sign is kept. The node is then colored based on the normalized gradients. The higher the value (darker red), the more a node contributes to a higher log CMC, and vice versa.

Screening of new surfactants

To further validate the generalizability of the trained GCN model, we designed new surfactants based on features found in the surfactants studied. Two series of surfactant designs are shown in Figure 8. In series 1, we started with a known structure of alcohol ethoxylate. By adding three ethoxy

groups to the polar head, we created a new surfactant that is not in the existing dataset and which is expected to have a higher CMC due to the addition of polar groups. We then tried to further increase the CMC by converting the linear alkyl chain into a branched one, as suggested by the saliency map analysis (Figure 7). Our intuition that these modifications in the surfactant design would increase CMCs was confirmed by GCN predictions. To further validate this result, we calculated CMCs using COSMOmic because this framework predicts similar trends as GCNs and experiments (Figure 4). As expected, the COSMOmic calculations lead to similar variations in the CMC, with slightly larger values predicted as also observed in Figure 4. For the second series, we selected a more complex surfactant structure from our dataset as the baseline design. The first design was obtained by removing side chains from the surfactant tail and by reducing the length of the polar head chain. These modifications simultaneously will tend to increase and decrease the CMC; as such, it is difficult to predict from intuition alone whether the new structure would have a higher or lower CMC. The GCN prediction shows the removal of side chains dominates the behavior, leading to a higher CMC. For the second design, we broke the π bonds in the benzene ring and reduced it to a cyclohexane group; because benzene has a higher polarity than cyclohexane, we expected that this design would have a lower CMC, as also shown by the GCN. For both designs, COSMOmic calculations again led to identical trends. These results validate that the GCN provides predictions that are physically intuitive. Overall, the proposed GCN architecture demonstrates the potential to be used as a tool that can help accelerate surfactant screening and design.

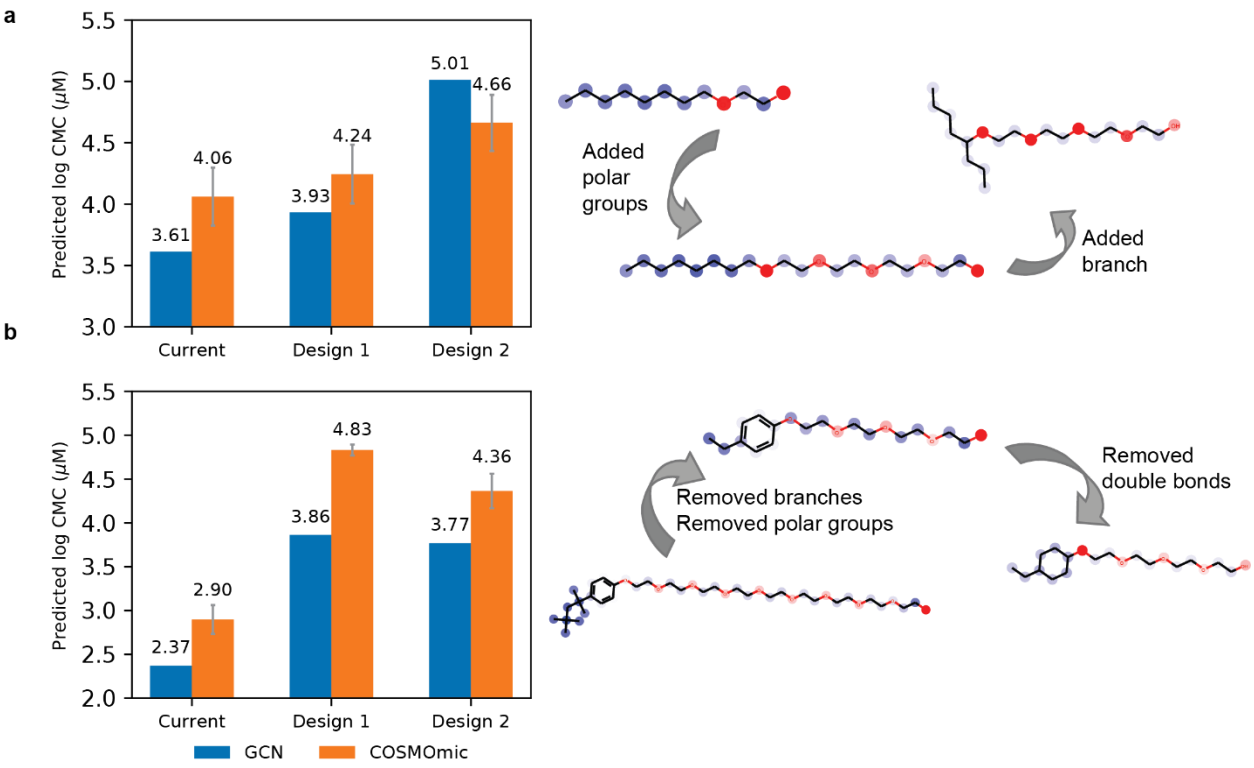


Figure 8. CMC predictions using GCN and COSMOmic calculations. Predicted log CMC values for new surfactants from trained GCN model and COSMOmic calculations. (a) Surfactant design series 1 where we start with a simple alcohol ethoxylate structure. Design 1 has additional ethoxy groups in the polar head and Design 2 further converts the linear chain into a branched one. (b) Surfactant design series 2 where we start with a complex alcohol phenol ethoxylate structure. Design 1 removes the branches from the nonpolar tail, and Design 2 reduces the benzene ring to a cyclohexane group.

Conclusions

We developed a GCN architecture to predict CMCs of surfactants directly from their molecular structure. We have found that the GCN predicts surfactant CMCs more accurately than previously developed QSPR models and generalizes to nonionic, cationic, anionic, and zwitterionic surfactants. Saliency analysis reveals that the GCN has the ability to capture important atomic types and molecular substructures that influence the CMC (such as polarity and head/tail lengths) even though the corresponding descriptors are not explicitly taken into account. Using the GCN, we demonstrated the ability to utilize the saliency map analysis to guide the design of new

surfactants for which experimental data are not available, then predict new CMCs with the GCN. These CMCs were then validated by calculations using COSMOmic to confirm that the predicted CMCs are reasonable.

A few notable advantages of the GCN over existing methods include its minimal input requirement, fast prediction speed, and good generalizability to surfactant types. Compared to MD simulations which can take hours or even days, the GCN only requires 0.01 seconds to make a prediction. This increased computational efficiency allows for surfactant screening and, when used in combination with product design models, can potentially enable the design of novel surfactants. Given the limited amount of experimental CMC data, high-throughput screening may still require additional training, although the study of the synthetic dataset has revealed the ability of the GCN to extract surfactant information. Moreover, the bond features are only implicitly captured in the atom feature vectors in our current GCN architecture, and each graph convolution only propagates the neighboring atom features. Therefore, we anticipate that alternative architectures of GNNs may be able to achieve higher prediction accuracies (*e.g.*, by incorporating higher-order neighboring features to graph convolutions).[42] Future work will also explore the use of GCN with graph-based inverse molecular design techniques[43] that introduce an encoder-decoder framework for automated surfactant design.

Methods

GCN architectures

The GCN proposed in this work is comprised of three major components: graph convolution, average pooling, and readout layers. Convolutional layers serve as a feature extraction step that incorporates both constitutional and topological information of a molecular graph. The graph

convolutions we used here are based on the original GCN implementation[33] where the hidden state of each node is updated using the information from its neighboring nodes. The node updating procedure is summarized in Equation 1.

$$h_i^{(t)} = ReLU\left[b^{(t)} + W^{(t)^T} \sum_{j \in \{\mathcal{N}(i) \cup i\}} \frac{1}{c_{ij}} h_j^{(t-1)}\right] \quad (1)$$

where $h_i^{(t)}$ represents the hidden state of node i at timestep t, b represents bias, W represents weight matrix, $\mathcal{N}(i)$ represents the set of neighboring nodes of node i, and $c_{ij} = \sqrt{d_i d_j}$ is a normalization term which denotes the square root of the product of node i's degree d_i and node j's degree d_j. The initial $h_i^{(0)}$ state of a node is the atom feature vector x described earlier in the text. After graph convolutional operations, average pooling is performed across all nodes in a graph to produce a fixed-size graph-level feature vector. This feature vector is then passed to fully connected layers. Finally, a linear transformation is performed to predict the log CMC. The model was constructed using Pytorch (version 1.2.0), and the molecular graphs and atom features were generated using the *Deep Graph Library*[44] (version 0.4.3post2) together with *RDKit* (version 2019.03.2).

GCN hyperparameter tuning

The major hyperparameters we varied are the number of graph convolutional layers (1 to 3), the number of fully connected hidden layers (1 to 3), and the number of hidden neurons (128, 256, 512). The model was trained with a mean-squared-error loss function, the Adam optimizer, a learning rate of 0.005, and a batch size of 5. The maximum epoch was set to 200 and, when early-stopping was enabled, the training process was terminated if the model performance on the

validation set did not improve for 20 epochs to help avoid overfitting. CV was also conducted to prevent overfitting and select the GCN architecture as described in the text. Mean CV RMSE, median CV RMSE, and model complexity were all taken into consideration for the final architecture of the GCN.

Synthetic dataset generation

A molecule backbone was first created by incorporating two components: a head part and a tail part, each comprised of repeated units (ethoxy groups for head and carbons for tail) to resemble a simple surfactant structure. We varied the backbone length and corresponding head-tail ratio. To add variety to the synthetic data, we introduced branches including one or two methyl or ethyl groups to the linear backbones; cyclohexane rings were also included at random positions in the linear backbones for more data complexity. The above structural design was translated into SMILES strings for which the feasibility and duplicity were checked. Additional details on this procedure are included in the Supplemental Information.

Saliency map generation

Saliency maps were created to gain insight into features of the molecular structure that best explain CMC values. To obtain a saliency map, gradients of the input atom features $\frac{\partial y}{\partial x}$ for each node were first calculated using backpropagation. Here y represents the predicted log CMC and x represents the atom feature vector. Element-wise multiplication was then performed between the input and the gradient through $x \odot \frac{\partial y}{\partial x}$ where $\odot$ denotes the element-wise multiplication operation.[45] To generate a node-level gradient value and study how atom type affects CMC predictions, we took

the sum of the gradients which are related to atom types using Equation 2 and normalized the value between -1 and 1 (the sign of a gradient value was kept during normalization).

$$Saliency = \sum_{x \in atom\ type} x \odot \frac{\partial y}{\partial x} \quad (2)$$

CMC calculations using COSMOmic

COSMOmic was used to compute the free energy of micellization to obtain CMCs from a molecular-scale simulation.[46] The workflow behind the COSMOmic CMC calculation is summarized in SI Figure S2a.[46] As input, COSMOmic requires structures of the surfactant monomer and micelle of interest (obtained from an atomistic molecular dynamics simulation) and screening charge densities for each of the different types of molecules in the system (obtained from quantum chemistry calculations). MD simulations were performed at a constant pressure of 1 bar and constant temperature of 298.15 K using Gromacs 2016.[47] Surfactants were modeled using the CHARMM36 force field with the TIP3P water model. Molecular structures and force field parameters were generated using the CHARMM-GUI Input Generator.[48, 49] For simulations of micelles, 100 monomers were assembled and solvated in water using PACKMOL[50] and equilibrated for 10-40ns. The simulation time was checked for each sample to confirm that the systems were equilibrated. Monomer and micelle configurations were selected based on structural metrics as detailed in the Supplemental Information. Monomer configurations were used as input to Gaussian 16 to compute the screening charge densities (COSMO files). Geometry optimization in implicit water (Conductor-like Polarizable Continuum Model, CPCM) was performed using density functional theory at the BVP86/TZVP/DGA1 level of theory. A single point calculation

was then performed to generate the ideal screening charges (at the infinite dielectric constant limit) on the molecular surface using the same level of theory.[51]

Given the input structure of a micelle and screening charge densities for a surfactant monomer, COSMOmic (implemented in *COSMOtherm*, version 19.0.05) divides the micelle into a series of concentric spherical shells and computes the water-micelle partition coefficient (K_m) of the surfactant in each shell using COSMO-RS calculations.[52] The partition coefficient can be related to the free energy as a function of the radial distance from the micelle center, *r*, by Equation 4.[46]

$$\Delta G(r) = -RT \ln K_m(r) \tag{4}$$

$\Delta G(r)$ is the free energy for moving a molecule from a position in bulk water to the specific value of r. The lowest value of $\Delta G(r)$ is defined as the free energy of micellization (ΔG_{mic}) and for nonionic surfactants is related to the CMC by Equation 5.

$$\Delta G_{mic} = RT \ln \mathrm{CMC} \tag{5}$$

In this expression, the CMC is expressed in mole fraction units by dividing concentrations by the molarity of water.

Acknowledgments

We acknowledge partial support from the US National Science Foundation through the University of Wisconsin Materials Research Science and Engineering Center (DMR-1720415).

References

1. Rosen, M. J.; Kunjappu, J. T., *Surfactants and Interfacial Phenomena: Fourth Edition*. John Wiley and Sons: 2012.
2. Torchilin, V. P., Structure and design of polymeric surfactant-based drug delivery systems. Elsevier: 2001; Vol. 73, pp 137-172.
3. Myers, D., *Surfactant Science and Technology: Third Edition*. John Wiley and Sons: 2005; p 1-380.
4. Castro, M. J. L.; Ojeda, C.; Cirelli, A. F., Advances in surfactants for agrochemicals.
5. Hill, K.; Rhode, O., Sugar-based surfactants for consumer products and technical applications. *Lipid / Fett* **1999,** *101* (1), 25-33.
6. Barati, A.; Najafi, A.; Daryasafar, A.; Nadali, P.; Moslehi, H., Adsorption of a new nonionic surfactant on carbonate minerals in enhanced oil recovery: Experimental and modeling study. *Chemical Engineering Research and Design* **2016,** *105*, 55-63.
7. Gallou, F.; Isley, N. A.; Ganic, A.; Onken, U.; Parmentier, M., Surfactant technology applied toward an active pharmaceutical ingredient: more than a simple green chemistry advance. Royal Society of Chemistry: 2015; Vol. 18, pp 14-19.
8. Kumar, G. P.; Rajeshwarrao, P., Nonionic surfactant vesicular systems for effective drug delivery—an overview. *Acta Pharmaceutica Sinica B* **2011,** *1* (4), 208-219.
9. Lorenzetto, T.; Berton, G.; Fabris, F.; Scarso, A., Recent Designer Surfactants for Catalysis in Water. *Catalysis Science & Technology* **2020,** *10* (14), 4492-4502.
10. Israelachvili, J., *Intermolecular and Surface Forces*. Elsevier Inc.: 2011.
11. Cheng, K. C.; Khoo, Z. S.; Lo, N. W.; Tan, W. J.; Chemmangattuvalappil, N. G., Design and performance optimisation of detergent product containing binary mixture of anionic-nonionic surfactants. *Heliyon* **2020,** *6* (5), e03861-e03861.
12. Gaudin, T.; Rotureau, P.; Pezron, I.; Fayet, G., New QSPR Models to Predict the Critical Micelle Concentration of Sugar-Based Surfactants. *Industrial & Engineering Chemistry Research* **2016,** *55* (45), 11716-11726.
13. 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. *Journal of Fluorescence* **2018,** *28* (1), 465-476.
14. Fluksman, A.; Benny, O., A robust method for critical micelle concentration determination using coumarin-6 as a fluorescent probe. *Analytical Methods* **2019,** *11* (30), 3810-3818.
15. Vishnyakov, A.; Lee, M. T.; Neimark, A. V., Prediction of the critical micelle concentration of nonionic surfactants by dissipative particle dynamics simulations. *Journal of Physical Chemistry Letters* **2013,** *4* (5), 797-802.
16. Santos, A. P.; Panagiotopoulos, A. Z., Determination of the critical micelle concentration in simulations of surfactant systems. *Journal of Chemical Physics* **2016,** *144* (4), 044709-044709.
17. Gahan, C. G.; Patel, S. J.; Boursier, M. E.; Nyffeler, K. E.; Jennings, J.; Abbott, N. L.; Blackwell, H. E.; Van Lehn, R. C.; Lynn, D. M., Bacterial Quorum Sensing Signals Self-Assemble in Aqueous Media to Form Micelles and Vesicles: An Integrated Experimental and Molecular Dynamics Study. *J. Phys. Chem* **2020,** *2020*, 3628-3628.
18. Li, X.; Zhang, G.; Dong, J.; Zhou, X.; Yan, X.; Luo, M., Estimation of critical micelle concentration of anionic surfactants with QSPR approach. *Journal of Molecular Structure: THEOCHEM* **2004,** *710* (1-3), 119-126.

19. Roy, K.; Kabir, H., QSPR with extended topochemical atom (ETA) indices: Modeling of critical micelle concentration of non-ionic surfactants. *Chemical Engineering Science* **2012**.
20. Huibers, P. D. T.; Lobanov, V. S.; Katritzky, A. R.; Shah, D. O.; Karelson, M., Prediction of critical micelle concentration using a quantitative structure-property relationship approach. 2. Anionic surfactants. *Journal of Colloid and Interface Science* **1997,** *187* (1), 113-120.
21. Katritzky, A. R.; Pacureanu, L.; Dobchev, D.; Karelson, M., QSPR Study of Critical Micelle Concentration of Anionic Surfactants Using Computational Molecular Descriptors. *Journal of Chemical Information and Modeling* **2007,** *47* (3), 782-793.
22. Katritzky, A. R.; Pacureanu, L. M.; Slavov, S. H.; Dobchev, D. A.; Karelson, M., QSPR Study of Critical Micelle Concentrations of Nonionic Surfactants. *Industrial & Engineering Chemistry Research* **2008,** *47* (23), 9687-9695.
23. Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A., Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chemical Reviews* **2010,** *110* (10), 5714-5789.
24. Puzyn, T.; Suzuki, N.; Haranczyk, M.; Rak, J., Calculation of quantum-mechanical descriptors for QSPR at the DFT level: Is it necessary? *Journal of Chemical Information and Modeling* **2008,** *48* (6), 1174-1180.
25. Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N. O.; Baker, N., Chemception: A Deep Neural Network with Minimal Chemistry Knowledge Matches the Performance of Expert-developed QSAR/QSPR Models. **2017**.
26. Hirohara, M.; Saito, Y.; Koda, Y.; Sato, K.; Sakakibara, Y., Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC Bioinformatics* **2018,** *19*.
27. Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V., MoleculeNet: A Benchmark for Molecular Machine Learning. *Chemical Science* **2017,** *9* (2), 513-530.
28. Tox21 Challenge. https://tripod.nih.gov/tox21/challenge/.
29. Delaney, J. S., ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *Journal of Chemical Information and Computer Sciences* **2004,** *44* (3), 1000-1005.
30. Meng, M.; Wei, Z.; Li, Z.; Jiang, M.; Bian, Y. In *Property Prediction of Molecules in Graph Convolutional Neural Network Expansion*, 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), 2019-10-01; IEEE: 2019.
31. Li, R.; Wang, S.; Zhu, F.; Huang, J., Adaptive Graph Convolutional Neural Networks. **2018**.
32. Tokui, S.; Oono, K.; Hido, S.; Clayton, J. *Chainer: a Next-Generation Open Source Framework for Deep Learning*.
33. Kipf, T. N.; Welling, M., Semi-Supervised Classification with Graph Convolutional Networks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings* **2016**.
34. Mukerjee, P.; Mysels, K. J. *Critical micelle concentrations of aqueous surfactant systems*; National Standard reference data system: 1971.
35. Neyman, J., On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection. In *Breakthroughs in Statistics: Methodology and Distribution*, Kotz, S.; Johnson, N. L., Eds. Springer New York: New York, NY, 1992; pp 123-150.

36. Jin, T.; Patel, S. J.; Van Lehn, R. C., Molecular simulations of lipid membrane partitioning and translocation by bacterial quorum sensing modulators. *PLOS ONE* **2021,** *16* (2), e0246187-e0246187.
37. Zana, R.; Weill, C., Effect of temperature on the aggregation behaviour of nonionic surfactants in aqueous solutions. *Journal de Physique Lettres* **1985,** *46* (20), 953-960.
38. Huibers, P. D. T.; Lobanov, V. S.; Katritzky, A. R.; Shah, D. O.; Karelson, M., Prediction of critical micelle concentration using a quantitative structure-property relationship approach. 1. Nonionic surfactants. *Langmuir* **1996,** *12* (6), 1462-1470.
39. Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O., Topological indices for structure-activity correlations. In *Topics in Current Chemistry*, Springer Berlin Heidelberg: pp 21-55.
40. Balaban, A. T., Highly discriminating distance-based topological index. *Chemical Physics Letters* **1982,** *89* (5), 399-404.
41. Bertz, S. H., The First General Index of Molecular Complexity. *Journal of the American Chemical Society* **1981,** *103* (12), 3599-3601.
42. Zhou, Z.; Li, X. *Convolution on Graph: A High-Order and Adaptive Approach*.
43. Sanchez-Lengeling, B.; Aspuru-Guzik, A., Inverse molecular design using machine learning:Generative models for matter engineering. American Association for the Advancement of Science: 2018; Vol. 361, pp 360-365.
44. Wang, M.; Zheng, D.; Ye, Z.; Gan, Q.; Li, M.; Song, X.; Zhou, J.; Ma, C.; Yu, L.; Gai, Y.; Xiao, T.; He, T.; Karypis, G.; Li, J.; Zhang, Z., Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph

Neural Networks. **2020**.
45. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B., Sanity Checks for Saliency Maps. *Advances in Neural Information Processing Systems* **2018,** *2018-December*, 9505-9515.
46. Jakobtorweihen, S.; Yordanova, D.; Smirnova, I., Predicting Critical Micelle Concentrations with Molecular Dynamics Simulations and COSMOmic. *Chemie Ingenieur Technik* **2017,** *89* (10), 1288-1296.
47. Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; Van Der Spoel, D.; Hess, B.; Lindahl, E., GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013,** *29* (7), 845-854.
48. Jo, S.; Kim, T.; Iyer, V. G.; Im, W., CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* **2008,** *29* (11), 1859-1865.
49. Kim, S.; Lee, J.; Jo, S.; Brooks, C. L.; Lee, H. S.; Im, W., CHARMM-GUI ligand reader and modeler for CHARMM force field generation of small molecules. *Journal of Computational Chemistry* **2017,** *38* (21), 1879-1886.
50. Martinez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M., PACKMOL: A package for building initial configurations for molecular dynamics simulations. *Journal of Computational Chemistry* **2009,** *30* (13), 2157-2164.
51. Eckert, F., COSMOtherm User Manual, Version C2.1, Release 01.10. 2009.
52. 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. *Journal of Physical Chemistry B* **2008,** *112* (38), 12148-12157.