THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

Predicting Critical Micelle Concentrations for Surfactants Using Graph Convolutional Neural Networks

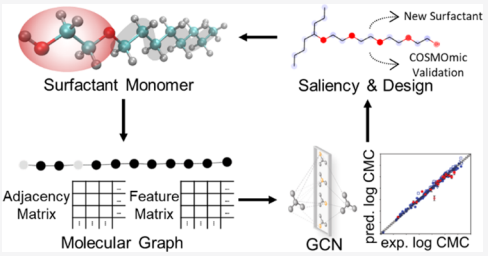
Shiyi Qin, Tianyi Jin, Reid C. Van Lehn,* and Victor M. Zavala*

Cite This: *J. Phys. Chem. B* 2021, 125, 10610−10620

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

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 study, we show that graph convolutional neural networks (GCNs) can predict CMCs directly from the surfactant molecular structure. In particular, 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 on a more inclusive data set than previously proposed methods and that it can generalize to anionic, cationic, zwitterionic, and nonionic surfactants using a single model. Molecular saliency maps revealed how atom types and surfactant molecular substructures contribute to CMCs and found this behavior 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 (Figure 1a−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.[3−5] Surfactants have also been utilized in 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 the hydrophilic surfactant head groups toward the aqueous phase and positioning the hydrophobic surfactant tail groups within the micelle core (Figure 1e,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-thermodynamic theory (MTT) models,[15−18] molecular dynamics (MD) simulations,[19−21] and descriptor-based quantitative structure−property relationship (QSPR) models[12,22−27] have been used to predict CMCs. These

Received: June 15, 2021
Revised: August 24, 2021
Published: September 9, 2021

https://doi.org/10.1021/acs.jpcb.1c05264
J. Phys. Chem. B 2021, 125, 10610−10620

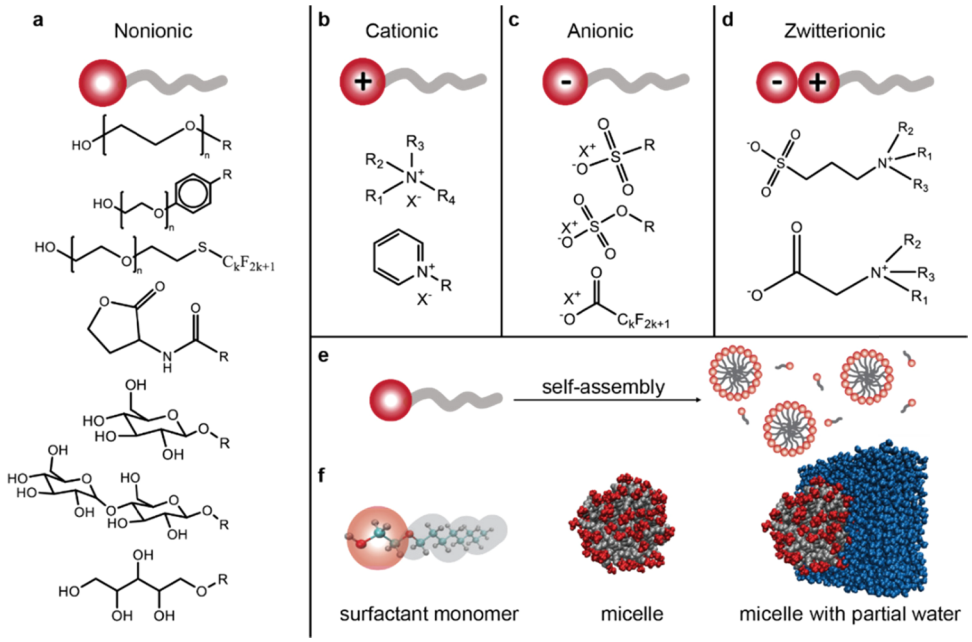


Figure 1. Overview of surfactant molecular structures and self-assembly process in micelles. (a–d) Sample structures of four classes of surfactants included in the experimental data set. Surfactants are categorized by the properties of their head groups as nonionic (a), cationic (b), anionic (c), or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

approaches have been shown to predict CMCs with relatively high accuracy, but they have a number of limitations. For instance, the MTT methods need numerous input parameters obtained either from experiments or empirical/MD models,[18] MD simulations usually require large system sizes, long simulation times, and assumptions regarding the number of surfactants within a micelle,[19−21] and QSPR models are often applicable to a single class of surfactants and may need density functional theory calculations to obtain quantum-chemical molecular descriptors.[28] 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 an input to convolutional neural networks (CNNs) to predict toxicity, activity, and solvation properties of different molecules.[29] 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.[30] Wu et al. employed graph neural networks (GNNs) trained on molecular graphs to predict various molecular properties.[31] 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 data sets such as Tox21[32] (for toxicity classification) and ESOL[33] (for water solubility regression).[31,34,35] 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).[31,34−36] 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),[37] 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 preserve topological information of the molecular structure through an adjacency matrix). GCNs extract features from molecular graph representations using convolutional operations that aggregate encoded information from molecular structures. We aim to use minimal input information of a surfactant monomer to predict the CMC. 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 molecular-level descriptors (constitutional, topological, geometrical, and quantum-chemical) of a surfactant monomer to predict CMCs.[12,22,23] One of the key advantages of GCNs is that they intrinsically capture the constitutional and topological information of a surfactant monomer without the need to calculate certain molecular descriptors explicitly (as in previous QSPR models). Atom features then propagate through graph convolutions; this type of convolution approximates physical interactions between atoms. Moreover, given that GCNs perform convolutions at an atomic (node) 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 data set that contains only 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 data set 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

https://doi.org/10.1021/acs.jpcb.1c05264
J. Phys. Chem. B 2021, 125, 10610−10620

predictive limitations of the proposed approach is the limited availability of experimental data. To address this issue, we created a synthetic data set that mimics the basic structural features of surfactants; this approach allowed us to construct a large and controlled data set to examine whether the GCN architecture can 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 and then validating GCN predictions by comparing them to CMC trends obtained from complementary molecular simulations.

Our results indicate that the proposed GCN generalizes well (as indicated by our cross-validation results and our predictions with both experimental and synthetic data sets). Our results also provide evidence that information encoded in the molecular structure of surfactants is sufficient to predict collective behavior. Moreover, our results aim to motivate further studies that focus on expanding databases via experiments (which is time-consuming) or via high-fidelity molecular simulation models.

METHODS

Preparation of the Experimental Surfactant CMC Data Set. We gathered experimental CMC data for 202 surfactants, including 122 nonionic surfactants, 35 cationic surfactants, 34 anionic surfactants, and 11 zwitterionic surfactants (Figure 1a–d), from multiple literature sources to form our data set.[1,12,21,38] All CMCs were measured at room temperature (between 20 and 25 °C) in water and converted to log CMC values (base 10). The data set 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 data set, respectively.

Since past approaches used for CMC predictions (e.g., QSPR models[12,22,23]) typically focused on a single class of surfactant, we first conducted baseline predictions on a subset of the original data set containing only nonionic surfactants to compare with the 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 data set containing all nonionic, anionic, cationic, and zwitterionic surfactants. This data set was partitioned into 160 training samples, 20 validation samples (ninefold CV), and 22 testing samples. The testing samples were selected using stratified sampling[39] to include surfactants that cover a wide range of the input CMCs and were held out during model training and validation. Supporting Information Table S1 lists all the surfactants studied and indicates the surfactants that were selected as test samples.

Molecular Graph Representation. Surfactant structures were converted to molecular graphs that were provided as an 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

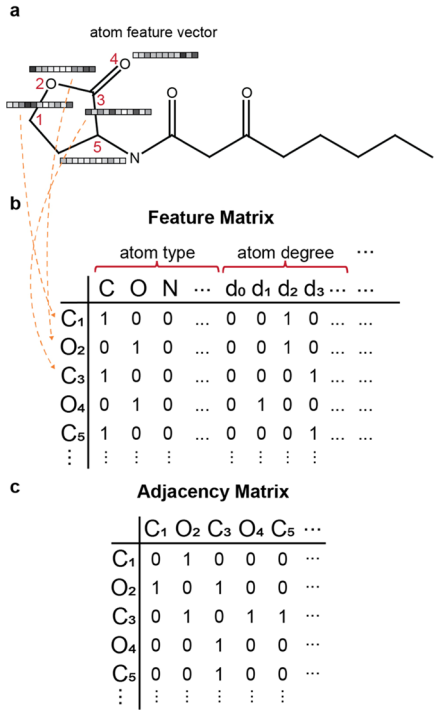


Figure 2. Data representation of an example surfactant. (a) Molecular graph of an example surfactant monomer. Atoms are represented as nodes and bonds are represented as undirected edges. Hydrogen atoms are implicit. The atom feature vectors are illustrated as colored bars next to each atom. (b) Atom 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) Adjacency matrix showing 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.

(number of connected edges to it), and charge in the form of a feature matrix (Figure 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 Supporting Information Table S2; as a comparison, a previous study[12] computed over 300 constitutional, topological, geometrical, and quantum-chemical descriptors to develop a QSPR model. In addition to the atom features, the molecular graph encodes topology using an adjacency matrix that captures atom connectivity (Figure 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).[12,23,24] For each cationic or anionic surfactant, the counterion was represented as a node disconnected from other nodes in the molecular graph.

GCN Architectures. The GCN proposed in this study 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

https://doi.org/10.1021/acs.jpcb.1c05264
J. Phys. Chem. B 2021, 125, 10610−10620

graph convolutions we used here are based on the original GCN implementation[37] where the hidden state of each node is updated using the information from its neighboring nodes. The node updating procedure is expressed by eq 1.

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

where $h_i^{(t)}$ represents the hidden state of node i at timestep t, b represents the bias, W represents the 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 was performed across all nodes in the graph to produce a fixed-size graph-level feature vector. This feature vector was then passed to fully connected layers. Finally, a linear transformation was 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[40] (version 0.4.3post2) together with the RDKit (version 2019.03.2).

GCN Hyperparameter Tuning. The major hyperparameters we varied are the number of graph convolutional layers (1−3), the number of fully connected hidden layers (1−3), and the number of hidden neurons (128, 256, and 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 the GCN architecture was selected as described in the text. Mean CV root-mean-squared-error (RMSE), median CV RMSE, and model complexity were all taken into consideration for the final architecture of the GCN.

Synthetic Data Set Generation. A larger synthetic data set was established to systematically study the predictive power of GCNs. To generate the synthetic molecules, a 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 the 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.

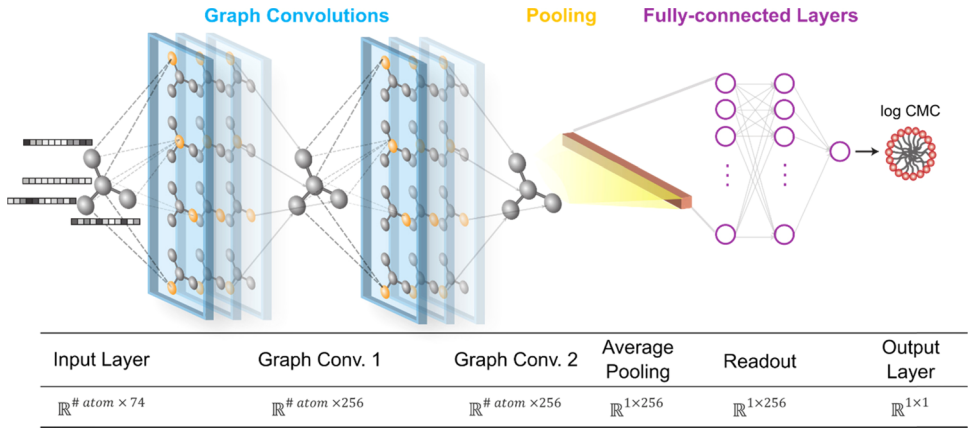
After the synthetic molecule structures were generated, three types of synthetic properties were calculated and used as prediction labels; here, we used three linear equations that capture constitutional, topological, and combined information of a synthetic molecule, respectively. The linear equations are dependent on molecular descriptors that have been used in QSPR models;[12,23,41,42] the constitutional descriptors are the number of C, number of O, and number of rings, while the topological descriptors are Balaban index[43] (a measure of average distance-based connectivity) and Bertz CT index[44] (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. Additional details on this procedure are included in the Supporting Information.

Saliency Map Generation. Saliency maps were created to gain insight into features of the molecular structure that best explains CMC values. Saliency maps are correlated with the sensitivity analysis that directly reveals how the change in the input would affect the output.[45] 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 using $x \odot \frac{\partial y}{\partial x}$ where $\odot$ denotes the element-wise multiplication operation;[45] this method was evaluated by Shrikumar et al.[46] on a genomic data set and has been proved to have correspondence with the first-order Taylor approximation of how the output would be influenced by the input change. To guide physical interpretation, multiplying the input by the gradient bit-wise shows the relative importance of an input feature if it is present. For example, the product of an input categorical feature for carbon and its corresponding gradient indicates how much contribution the presence (versus absence) of a carbon has to the log CMC, given that the input features that are associated with the atom types are one-hot-encoded features (binary features indicating the presence or absence of an atom type category). To generate a node-level gradient value and study how the atom type affects CMC predictions, we took the sum of the gradients that are related to atom types using eq 2 and normalized the value between −1 and 1 (the sign of a gradient value was kept during normalization); a similar interpretation of the total node importance was implemented in StellarGraph.[47]

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

CMC Calculations Using COSMOmic. The COnductor-like Screening Model for Realistic Solvation (COSMO-RS) model and its extension, COSMOmic, were used to compute the free energy of micellization to obtain CMCs from a molecular-scale simulation to compare with GCN predictions.[48,49] These calculations were also used to validate GCN predictions for designed surfactants for which no experimental data are available. The workflow behind the COSMOmic CMC calculation is illustrated in Supporting Information Figure S2a.[48] As input, COSMOmic requires structures of the surfactant monomer and micelles of interest (obtained from an atomistic MD 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 a constant temperature of 298.15 K using GROMACS 2016.[50] 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.[51,52] For simulations of micelles, 100 monomers were assembled and solvated in water using PACKMOL[53] and equilibrated for 10−40 ns. The simulation time was checked for each sample to confirm that the systems were equilibrated. Monomer and micelle

https://doi.org/10.1021/acs.jpcb.1c05264
J. Phys. Chem. B 2021, 125, 10610−10620



Input Layer	Graph Conv. 1	Graph Conv. 2	Average Pooling	Readout	Output Layer
$\mathbb{R}^{\#\ atom \times 74}$	$\mathbb{R}^{\#\ atom \times 256}$	$\mathbb{R}^{\#\ atom \times 256}$	$\mathbb{R}^{1\times 256}$	$\mathbb{R}^{1\times 256}$	$\mathbb{R}^{1\times 1}$

Figure 3. GCN architecture. The proposed GCN takes a molecular graph as an input, convolutes across each input twice in series (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 two fully connected neural network layers. The dimension for each layer is summarized in the table shown at the bottom.

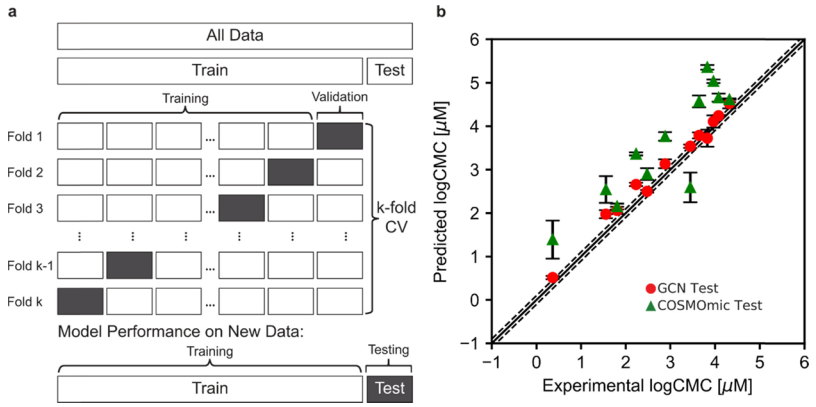


Figure 4. GCN model validation and testing for nonionic surfactant data set. (a) Training, validation, and testing procedure for hyperparameter tuning. We first divide the data set into training and testing sets. The training set is then split into training and validation sets for a k-fold CV procedure. After the architecture is determined, we train the GCN on the entire training data set and test the model on the held-out test set to evaluate the 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 MD simulations.

configurations were selected based on structural metrics as detailed in the Supporting Information. Monomer configurations were used as an input to Gaussian 16 to compute the screening charge densities (COSMO files). Geometry optimization in implicit water (conductor-like polarizable continuum model) was performed using density functional theory at the BVP86/TZVP/DGA1 level of theory. A single point calculation was then performed to generate ideal screening charges (at the infinite dielectric constant limit) on the molecular surface using the same level of theory.[54]

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.[55] The partition coefficient can be related to the free energy as a function of the radial distance from the micelle center, r, as given by eq 3.[48]

$$\Delta G(r) = -RT \ln K_m(r) \tag{3}$$

where $\Delta G(r)$ 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, it is related to the CMC as given by eq 4.

$$\Delta G_{mic} = RT \ln \text{CMC} \tag{4}$$

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

RESULTS AND DISCUSSION

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 (Figure 3). A graph convolution layer updates each atom by aggregating the features of itself and 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

https://doi.org/10.1021/acs.jpcb.1c05264
J. Phys. Chem. B 2021, 125, 10610−10620

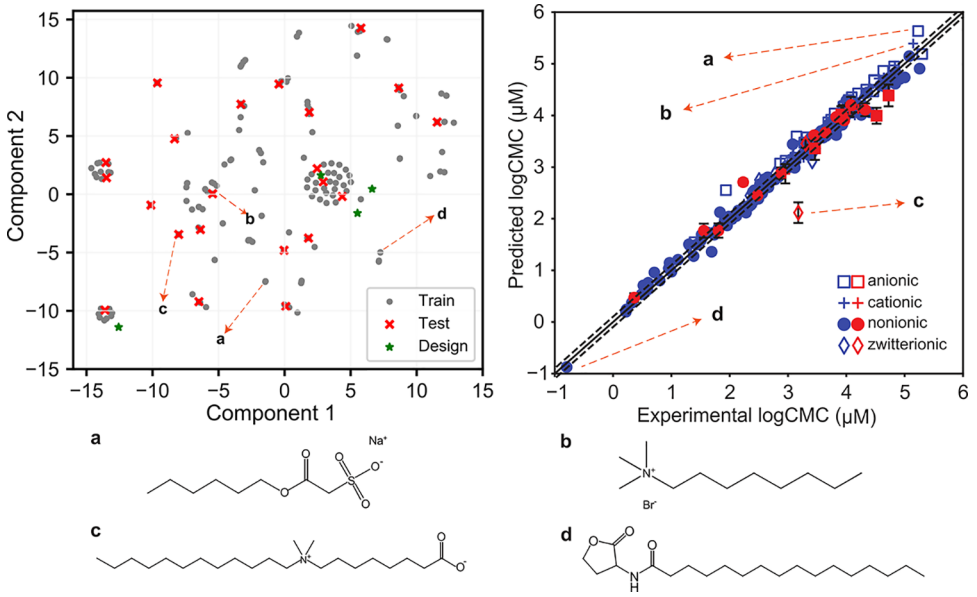


Figure 5. GCN predictions for all classes of surfactants. Left: low-dimensional distribution of surfactant fingerprints using t-SNE. The test samples (red crosses) are widespread, and most of the designed surfactants (green points) fall outside of the clusters of the existing data set. Right: parity 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.

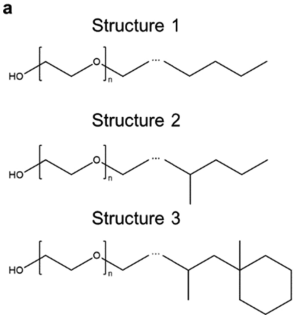
activation) of the updated features. The GCN model contains a total of 216,833 parameters, corresponding to operators of the graph convolutions as well as bias terms. Because the number of parameters is relatively large compared with the size of the data set, as measures to prevent overfitting, we used early stopping to terminate training when the validation performance starts to degrade and CV to estimate the predictive power of the model architecture on unseen data. However, the proposed GCN architecture is relatively simple compared with the preset GCN architecture (with 586,625 parameters) implemented in MoleculeNet[31] for a single regression task (Supporting Information Table S5). The GCN architecture was determined by hyperparameter tuning using the nonionic surfactant subset and by performing 11-fold CV (Figure 4a). The 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 (Supporting Information 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 Supporting Information Figure S1.

We tested the ability of the GCN model to generalize new data by training the model using all training samples and then testing on held-out test samples, as illustrated in Figure 4a. For the nonionic data set, 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 data set 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. 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 data set encompasses a wider variety of nonionic surfactants (other than sugar-based ones), such as fluorinated thiol ethoxylates and acyl-homoserine lactones.

Molecular Simulations for CMC Prediction as a Validation Method. We compared the predictive performance of molecular simulations on the same test data set using COSMOmic.[49] Atomistic MD simulations were first conducted to obtain input structures for each surfactant monomer and the corresponding spherical micelle in the test set (Supporting Information Figure S2). These structures were used as inputs 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.[56] The RMSE obtained from the COSMOmic calculations is 0.91 with a best-fit slope of 0.92, which is less accurate than that of 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

https://doi.org/10.1021/acs.jpcb.1c05264
J. Phys. Chem. B 2021, 125, 10610−10620

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 synthetic 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.

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 predicting CMC values for newly designed surfactants for which experimental data are not available.

GCN for CMC Prediction for all Surfactants. We trained the same GCN architecture on the full data set containing all four classes of surfactants and performed ninefold CV. Instead of tuning the hyperparameters, CV was used to compare the model performance to that of the previous data set 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 shown in Supporting Information Figure S1. We again tested the model performance on a test data set, which contains the same 12 nonionic test samples as well as 4 additional cationic, 4 anionic, and 2 zwitterionic samples. We verified the distribution of the test samples using t-distributed stochastic neighbor embedding (t-SNE),[57] a nonlinear dimension reduction technique to visualize high-dimensional data, on the molecular fingerprints[58] of the surfactants. Figure 5 illustrates that the test samples are widespread, indicating the inclusion of unlike surfactant structures and classes in the test data set, which cover a much more diverse spectrum of surfactants than the data sets used in previous QSPR models.

Figure 5 also 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) surfactants, 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 data set 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[12,22−26] have tried to predict CMCs for all classes of surfactants using a single model; as such, the proposed GCN exhibits a significant development in surfactant CMC prediction.

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 data set. 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 data set that encompasses 1820 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 Supporting Information). The methodology used for generating the synthetic molecules captures three types of surfactant-like structures (Figure 6a): (1) a head-tail linear structure, (2) a head-tail linear structure with single or double branching, and (3) a 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. The synthetic properties were constructed from molecular descriptors used in QSPR models[12,23,41,42] and were categorized as constitutional, topological, and combined labels calculated by a linear combination of corresponding descriptors (Supporting Information 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 the model performance instead of the RMSE. We used the proposed architecture to train the GCN with 10-fold CV, leading to 1638 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 (Figure 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-

https://doi.org/10.1021/acs.jpcb.1c05264
J. Phys. Chem. B 2021, 125, 10610−10620

specific). In particular, for the synthetic molecules with a 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.

Although the above synthetic data analysis confirms that GCN can extract structural information from a molecule, the extent still depends on the target property. The current models to create synthetic labels might be an oversimplification of the CMC's dependence on a molecular structure. However, the preliminary algorithm provided for generating the synthetic molecules can serve as a starting point for creating general surfactant structures with valid SMILES strings. For instance, the algorithm asks for inputs such as chain length and side group structure and thus has the potential to generate a large surfactant data set within seconds for high-throughput screening.

Molecular Saliency Maps. Compared with QSPR models, GCNs are in general more complex and less interpretable. Therefore, we computed molecular saliency maps to further understand the 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 the 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 exhibit 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 S3 showcases more saliency patterns for surfactant groups with similar structures.

Screening of New Surfactants. To further validate the generalizability of the trained GCN model, we designed new surfactants with unseen structures 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 data set 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); this structure is different from the alcohol ethoxylates in the training set, which are mostly linear. Our intuition that these modifications in the surfactant design would increase CMCs was confirmed by GCN predictions. To

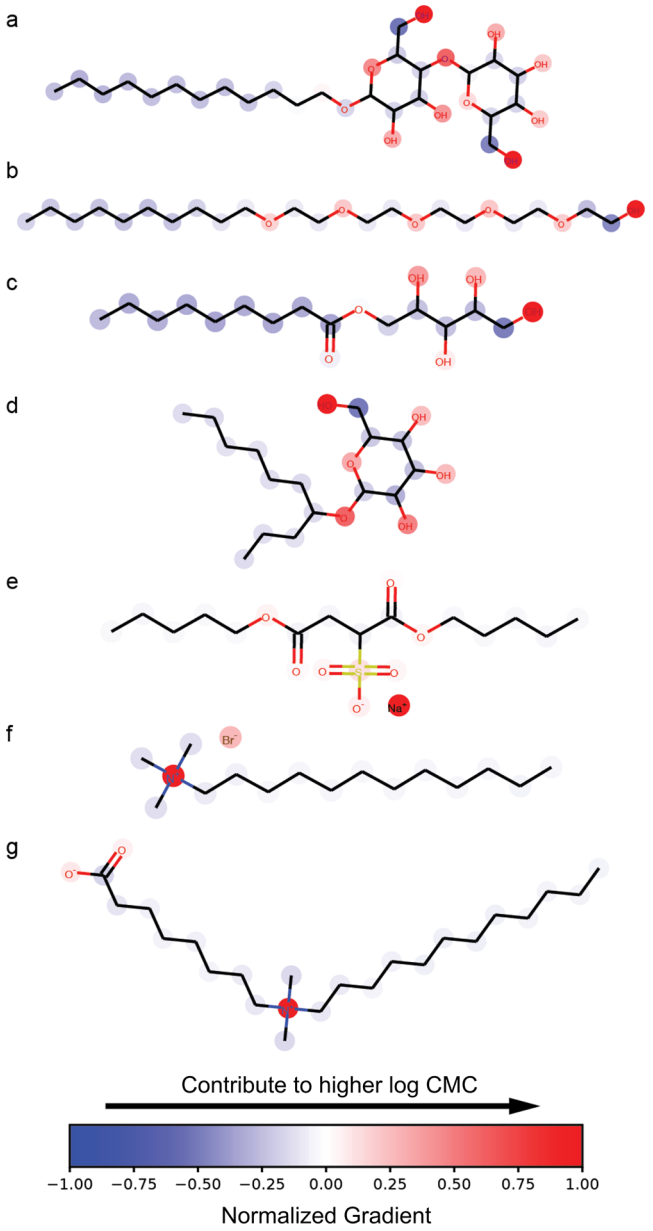


Figure 7. Molecular saliency maps. Selected examples from nonionic (a–d), cationic (e), anionic (f), and zwitterionic (g) surfactants. The gradient values are calculated for each node followed by normalization between −1 and 1 where the sign is kept. The node is then colored based on the normalized gradients.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

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 led 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 data set 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 that the removal of side chains dominates the behavior, leading to a higher

https://doi.org/10.1021/acs.jpcb.1c05264
J. Phys. Chem. B 2021, 125, 10610−10620

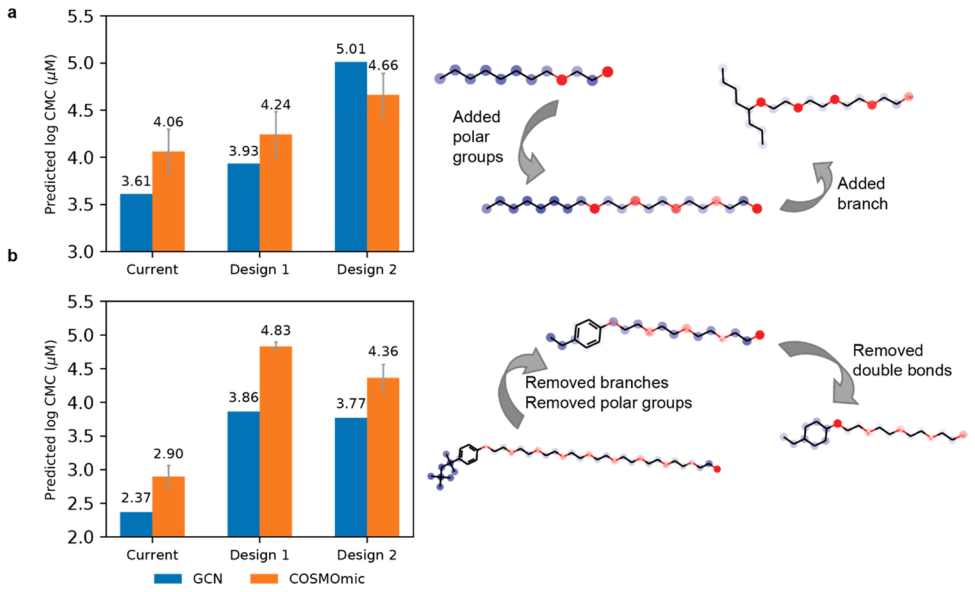


Figure 8. CMC predictions using GCN and COSMOmic calculations. Predicted log CMC values from the trained GCN model and from COSMOmic calculations for surfactants not included in the training/validation/testing data. (a) Surfactant design series 1 where we start with a simple alcohol ethoxylate structure. 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. None of the newly designed surfactants are included in the training data set. "Design 2" in both series also has unobserved structures, including the short and branched alcohol ethoxylate backbone in (a) and the cyclohexane group in (b).

CMC. For the second design, we broke the π bonds in the benzene ring and reduced it to a cyclohexane group, which is an unobserved structure in the training data; 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.

CONCLUSIONS

We developed a simple 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 can be generalized 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 and then predict new CMCs with the GCN. These CMCs were then validated by calculations using COSMOmic to confirm that the predicted CMCs are reasonable.

We focused on a basic GCN architecture because our goal is to study whether the CMC can be predicted directly from the structure of a single surfactant monomer by graph convolutions, which explicitly aggregate features from adjacent atoms and implicitly aggregate features from farther neighbors via the second convolution layer and the readout layers. We anticipate that the findings in this study may serve as a starting point and baseline for predictive models of CMC using GNN techniques. A few notable advantages of the GCN over the existing methods include its minimal input requirement, fast prediction speed, and good generalizability to surfactant classes. Compared to MD simulations, which can take hours or even days, the trained GCN only requires 0.01 s to make a prediction. Additionally, the proposed GCN has the flexibility to predict CMCs for multiple surfactant classes, whereas a QSPR model requires molecular descriptor calculations and recomputation when switching models for a different class/subclass of surfactants, which may lead to an overall higher computational cost. 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. In this study, however, CV, early stopping, the study of the synthetic data set, and the COSMOmic validation of CMCs for unobserved surfactant structures have revealed the ability of the GCN to extract surfactant information. There are a few limitations of the current scope of the study, which may be of interest for further investigation. For instance, 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, for example, by incorporating higher-order neighboring features to graph convolutions[59] to better capture the cooperative behavior. Future studies will also explore the use of GCN with graph-based inverse molecular design techniques[60] that introduce an encoder−decoder framework for automated surfactant design.

https://doi.org/10.1021/acs.jpcb.1c05264
J. Phys. Chem. B 2021, 125, 10610−10620

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.1c05264.

Algorithms to generate the synthetic data set, CMC calculations using COSMOmic, experimental and predicted CMC values, list of atom features, summary of hyperparameter tuning, equations to calculate the labels of the synthetic molecules, list of parameters from MoleculeNet implemented in DeepChem for the pretrained GCN, CV RMSEs, CMC estimations using COSMOmic, saliency patterns for surfactants with similar structures, and GCN model validations and testing for a nonionic surfactant data set (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Reid C. Van Lehn − *Department of Chemical and Biological Engineering, University of Wisconsin − Madison, Madison, Wisconsin 53706, United States;* orcid.org/0000-0003-4885-6599; Email: vanlehn@wisc.edu

Victor M. Zavala − *Department of Chemical and Biological Engineering, University of Wisconsin − Madison, Madison, Wisconsin 53706, United States;* orcid.org/0000-0002-5744-7378; Email: victor.zavala@wisc.edu

Authors

Shiyi Qin − *Department of Chemical and Biological Engineering, University of Wisconsin − Madison, Madison, Wisconsin 53706, United States*

Tianyi Jin − *Department of Chemical and Biological Engineering, University of Wisconsin − Madison, Madison, Wisconsin 53706, United States;* orcid.org/0000-0002-9974-2041

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.1c05264

Notes

The authors declare no competing financial interest.

■ 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*; 3rd Edition; John Wiley and Sons: 2005; pp. 1−380.

(4) Castro, M. J. L.; Ojeda, C.; Cirelli, A. F. Advances in surfactants for agrochemicals. *Environ. Chem. Lett.* **2014**, *12*, 85−95.

(5) Hill, K.; Rhode, O. Sugar-based surfactants for consumer products and technical applications. *Lipid/Fett* **1999**, *101*, 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. *Chem. Eng. Res. Des.* **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. *Green Chem.* **2016**, *18*, 14−19.

(8) Kumar, G. P.; Rajeshwarrao, P. Nonionic surfactant vesicular systems for effective drug delivery—an overview. *Acta Pharm. Sin. B* **2011**, *1*, 208−219.

(9) Lorenzetto, T.; Berton, G.; Fabris, F.; Scarso, A. Recent Designer Surfactants for Catalysis in Water. *Catal. Sci. Technol.* **2020**, *10*, 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*, No. e03861.

(12) Gaudin, T.; Rotureau, P.; Pezron, I.; Fayet, G. New QSPR Models to Predict the Critical Micelle Concentration of Sugar-Based Surfactants. *Ind. Eng. Chem. Res.* **2016**, *55*, 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. *J. Fluoresc.* **2018**, *28*, 465−476.

(14) Fluksman, A.; Benny, O. A robust method for critical micelle concentration determination using coumarin-6 as a fluorescent probe. *Anal. Methods* **2019**, *11*, 3810−3818.

(15) Reif, I.; Mulqueen, M.; Blankschtein, D. Molecular-Thermodynamic Prediction of Critical Micelle Concentrations of Commercial Surfactants. *Langmuir* **2001**, *17*, 5801−5812.

(16) Goldsipe, A.; Blankschtein, D. Molecular-Thermodynamic Theory of Micellization of pH-Sensitive Surfactants. *Langmuir* **2006**, *22*, 3547−3559.

(17) Ren, Z. H. A molecular-thermodynamic approach to predict the micellization of binary surfactant mixtures containing amino sulfonate amphoteric surfactant and nonionic surfactant. *AIChE J.* **2017**, *63*, 5076−5082.

(18) Sresht, V.; Lewandowski, E. P.; Blankschtein, D.; Jusufi, A. Combined Molecular Dynamics Simulation−Molecular-Thermodynamic Theory Framework for Predicting Surface Tensions. *Langmuir* **2017**, *33*, 8319−8329.

(19) Vishnyakov, A.; Lee, M. T.; Neimark, A. V. Prediction of the critical micelle concentration of nonionic surfactants by dissipative particle dynamics simulations. *J. Phys. Chem. Lett.* **2013**, *4*, 797−802.

(20) Santos, A. P.; Panagiotopoulos, A. Z. Determination of the critical micelle concentration in simulations of surfactant systems. *J. Chem. Phys.* **2016**, *144*, No. 044709.

(21) 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. B* **2020**, *124*, 3616−3628.

(22) 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.

(23) Roy, K.; Kabir, H. QSPR with extended topochemical atom (ETA) indices: Modeling of critical micelle concentration of non-ionic surfactants. *Chem. Eng. Sci.* **2012**, *73*, 86−98.

(24) 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. *J. Colloid Interface Sci.* **1997**, *187*, 113−120.

(25) Katritzky, A. R.; Pacureanu, L.; Dobchev, D.; Karelson, M. QSPR Study of Critical Micelle Concentration of Anionic Surfactants Using Computational Molecular Descriptors. *J. Chem. Inf. Model.* **2007**, *47*, 782−793.

(26) Katritzky, A. R.; Pacureanu, L. M.; Slavov, S. H.; Dobchev, D. A.; Karelson, M. QSPR Study of Critical Micelle Concentrations of Nonionic Surfactants. *Ind. Eng. Chem. Res.* **2008**, *47*, 9687−9695.

(27) 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. *Chem. Rev.* **2010**, *110*, 5714−5789.
(28) Puzyn, T.; Suzuki, N.; Haranczyk, M.; Rak, J. Calculation of quantum-mechanical descriptors for QSPR at the DFT level: Is it necessary? *J. Chem. Inf. Model.* **2008**, *48*, 1174−1180.
(29) 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, arXiv:1706.06689.
(30) Hirohara, M.; Saito, Y.; Koda, Y.; Sato, K.; Sakakibara, Y. Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC Bioinform.* **2018**, *19*, 526.
(31) 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. *Chem. Sci.* **2018**, *9*, 513−530.
(32) Tox21 Challenge. https://tripod.nih.gov/tox21/challenge/.
(33) Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000−1005.
(34) Meng, M.; Wei, Z.; Li, Z.; Jiang, M.; Bian, Y. Property Prediction of Molecules in Graph Convolutional Neural Network Expansion. In *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), 2019-10-01*; IEEE: 2019.
(35) Li, R.; Wang, S.; Zhu, F.; Huang, J. Adaptive Graph Convolutional Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*; 2018.
(36) Tokui, S.; Oono, K.; Hido, S.; Clayton, J. Chainer: a Next-Generation Open Source Framework for Deep Learning. In *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*; 2015; vol. *5*, pp. 1−6.
(37) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*; 2016.
(38) Mukerjee, P.; Mysels, K. J. *Critical micelle concentrations of aqueous surfactant systems*; National Standard reference data system: 1971.
(39) 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.
(40) 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, arXiv:1909.01315.
(41) 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*, 1462−1470.
(42) Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. Topological indices for structure-activity correlations. In *Topics in Current Chemistry*; Springer: Berlin Heidelberg, 1983; pp. 21−55.
(43) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(44) Bertz, S. H. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3599−3601.
(45) Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems*; 2018, December; pp. 9505−9515.
(46) Shrikumar, A.; Greenside, P.; Shcherbina, A.; Kundaje, A. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. 2017, arXiv:1605.01713.
(47) Data61, C. *Stellargraph machine learning library*; GitHub Repository: 2018.
(48) Jakobtorweihen, S.; Yordanova, D.; Smirnova, I. Predicting Critical Micelle Concentrations with Molecular Dynamics Simulations and COSMOmic. *Chem. Ing. Tech.* **2017**, *89*, 1288−1296.
(49) 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*, No. e0246187.
(50) 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.; et al. GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845−854.
(51) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **2008**, *29*, 1859−1865.
(52) 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. *J. Comput. Chem.* **2017**, *38*, 1879−1886.
(53) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **2009**, *30*, 2157−2164.
(54) Eckert, F. *COSMOtherm User Manual, Version C2.1, Release 01.10*; COSMOlogic GmbH: Leverkusen, Germany, 2010.
(55) Klamt, A.; Huniar, U.; Spycher, S.; Keldenich, J. COSMOmic: A mechanistic approach to the calculation of membrane-water partition coefficients and internal distributions within membranes and micelles. *J. Phys. Chem. B* **2008**, *112*, 12148−12157.
(56) Zana, R.; Weill, C. Effect of temperature on the aggregation behaviour of nonionic surfactants in aqueous solutions. *J. Phys. Lett.* **1985**, *46*, 953−960.
(57) Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2759−2605.
(58) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(59) Zhou, Z.; Li, X. Convolution on Graph: A High-Order and Adaptive Approach. 2017, arXiv:1706.09916.
(60) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Am. Assoc. Adv. Sci.* **2018**, *361*, 360−365.

https://doi.org/10.1021/acs.jpcb.1c05264
J. Phys. Chem. B 2021, 125, 10610−10620