Colloids and Surfaces A: Physicochemical and Engineering Aspects 694 (2024) 134133

Contents lists available at ScienceDirect

Colloids and Surfaces A: Physicochemical and Engineering Aspects

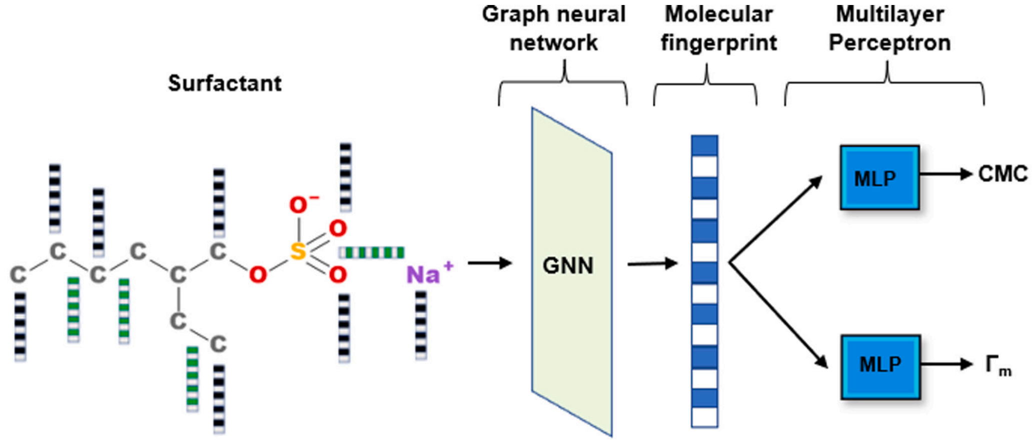
journal homepage: www.elsevier.com/locate/colsurfa

Graph neural networks for surfactant multi-property prediction

Christoforos Brozos [a,b], Jan G. Rittig [b], Sandip Bhattacharya [a], Elie Akanny [a], Christina Kohlmann [a], Alexander Mitsos [d,b,c,*]

[a] BASF Personal Care and Nutrition GmbH, Henkelstrasse 67, 40589 Duesseldorf, Germany
[b] RWTH Aachen University, Process Systems Engineering (AVT.SVT), Aachen 52074, Germany
[c] Forschungszentrum Jülich GmbH, Institute of Energy and Climate Research IEK-10 – Energy Systems Engineering, Jülich, Germany
[d] JARA Center for Simulation and Data Science (CSD), Aachen, Germany

GRAPHICAL ABSTRACT



ARTICLE INFO

Dataset link: https://github.com/brozosc/GNNs-for-surfactant-multi-property-prediction

Keywords:
Surfactants
Critical micelle concentration
Surface excess concentration
Graph neural network
Multi-task learning

ABSTRACT

Surfactants are of high importance in different industrial sectors such as cosmetics, detergents, oil recovery and drug delivery systems. Therefore, many quantitative structure–property relationship (QSPR) models have been developed for surfactants. Each predictive model typically focuses on one surfactant class, mostly nonionics. Graph Neural Networks (GNNs) have exhibited a great predictive performance for property prediction of ionic liquids, polymers and drugs in general. Specifically for surfactants, GNNs can successfully predict critical micelle concentration (CMC), a key surfactant property associated with micellization. A key factor in the predictive ability of QSPR and GNN models is the data available for training. Based on extensive literature search, we create the largest available CMC database with 429 molecules and the first large data collection for surface excess concentration (Γ_m), another surfactant property associated with foaming, with 164 molecules. Then, we develop GNN models to predict the CMC and Γ_m and we explore different learning approaches, i.e., single- and multi-task learning, as well as different training strategies, namely ensemble and transfer learning. We find that a multi-task GNN with ensemble learning trained on all Γ_m and CMC data performs best. Thus, our results show that the simultaneous use of data from highly correlated properties can improve the predictability of surfactant properties for which only a small amount of experimental data is available. Finally, we test the ability of our CMC model to generalize on industrial grade pure component surfactants. The GNN yields highly accurate predictions for CMC, showing great potential for future industrial applications.

https://doi.org/10.1016/j.colsurfa.2024.134133
Received 2 January 2024; Received in revised form 26 April 2024; Accepted 28 April 2024
Available online 3 May 2024

1. Introduction

Surfactants are highly relevant molecules used in a wide range of everyday products, such as food, cosmetics, detergents and drugs [1–6]. Surface-active agents (surfactants) are amphiphilic molecules with a hydrophobic chain and a polar hydrophilic head. The surfactant molecule orients itself at the interface between two phases with the hydrophobic portion oriented towards the hydrophobic phase (e.g., air/oil) and the hydrophilic portion oriented towards the hydrophilic phase. Due to their structure, surfactants are surface/interface active and they are able to lower the surface/interfacial tension [2,7]. Owing to their properties, surfactants are widely used in a variety of applications such as detergents, dispersion stabilizers, foaming agents, lubricants and as pharmaceuticals among many others [2,8–10]. Surfactants are classified based on their hydrophilic head group, into ionics and nonionics, with the former further classified into anionics, cationics and zwitterionics. Important surfactant properties to various applications are the critical micelle concentration (CMC), the surface tension (γ), the surface excess concentration (Γ_m), the Cloud Point (CP) and the Krafft Temperature (KT) [11,12]. These properties are used to identify new surfactants with desired performance.

Amphiphilic molecules such as surfactants form micelles, i.e., aggregates. Micelles ultimately dictate the surface/interface activity and strongly impacts the solubilization and detergency, or cleaning ability of a surfactant solution [13]. The minimum surfactant concentration at which such micelles are formed in a solution is called *critical micelle concentration* (CMC) [14,15]. The CMC is an important value in a wide range of applications such as shampoos [16], bio-materials design for drug delivery systems [17], polymeric micelles [18,19] and oil recovery [20]. In addition, some studies have reported correlation between CMC and surfactant toxicity [21] and between CMC and foam stability [22]. The CMC is influenced by multiple factors, like temperature, solvent, pH, chemical structure, pressure conditions and size of the tail and head groups [14,23,24]. Determination of CMC is time-consuming and expensive, and several methods can be used like tensiometry [25], refractive index [15], calorimetry [26], viscosity and conductivity measurements [27]. In most methods, a break-point in the measured property (e.g., surface tension or conductivity) vs. surfactant concentration curve is observed, and the CMC is defined to be at that point [15,28,29].

Since surfactants prefer to exist and adsorb at the interfaces, we can define their adsorption effectiveness as the surface excess concentration (Γ_m) [14]. Γ_m is an important surfactant property, as it is a measure of surfactant concentration at air/water and oil/water interfaces. Additionally, surface excess concentration has been shown to influence foaming, emulsification and the kinetics of surfactant-induced pore wetting [14,30]. Like CMC, Γ_m is influenced by surfactant structure and temperature [14,31]. The implicit calculation of the surface excess concentration is possible, as is the CMC, from a surface tension measurement plot using the Gibbs adsorption equation [14,32].

Due to the tedious and expensive nature of experiments, the prediction of surfactant properties without experiments has been a focus of research for many years, mainly, through the use of quantitative structure–property relationship (QSPR) models. An overview of QSPR models in surfactants was given by Hu et al. [33], with most of the QSPR studies aiming to correlate molecular descriptors with the CMC [23,34–37]. The developed models showed good predictive performance [23,34–37]. Nevertheless, all of them share a common limitation: they are applicable only on a single surfactant class (nonionics, cations etc.). Besides the CMC, similar modeling techniques have been applied to other important surfactant properties like the cloud point of nonionic surfactants, the hydrophile–lipophile balance (HLB) or the minimum surface tension at CMC [33,38–40]. Very recently, the first QSPR model for predicting Γ_m was developed [41]. The QSPR model was trained on data generated from the Szyszkowski equation using experimentally measured SFT-log(c) profiles and not on the experimental data directly [41]. However, the authors stress that their QSPR model has limitations in the surfactant categories included, e.g., the model is not applicable to fluorinated surfactants [41].

In the recent years, graph neural networks (GNNs) have been intensively researched in the field of molecular property prediction, with numerous GNN models existing in the literature, even regarding the same target property [42,43]. In more detail, GNNs have been applied to a variety of chemical applications such as ferromagnetic materials [44], the biodegradability of molecules [45] and the activity coefficients [46–49]. GNNs are a deep learning technique, where each molecule is represented as a graph, with atoms corresponding to nodes and bonds to edges. In contrast to classical QSPR methods, where molecular descriptors are typically selected manually and therefore require domain knowledge, GNNs can extract, in an automated way, all the necessary structural-related information which are later used in the regression task for property prediction. For surfactants, Qin et al. [50] used GNNs to predict the CMC of multiple surfactant classes. They showed that GNNs can be efficiently used as an alternative to classical descriptor-based QSPR in surfactants, with very promising results, using a database of 200 molecules, which is relatively small for training a machine learning model.

Herein, we create the largest CMC and Γ_m data sets available and we use them to develop GNN models for their prediction. First, we extend the publicly available CMC data set of Qin et al. [50] to 429 molecules through an extensive literature search. Then, we construct a second data set of 99 molecules with duplicate values with the aim of investigating possible benefits of transfer learning in CMC prediction. For the Γ_m, no publicly available database was found during our research. Therefore, we construct one with 164 surfactant molecules varying from multiple surfactant class types. Compared to the work of Seddon et al. [41], we include a wider range of surfactant categories, e.g., fluorinated components, and we consider the impact of counterions on Γ_m [14]. Compared to successfully developed QSPR CMC prediction models [23,34–37], we here include CMC data for all surfactant classes. Note that the collected data sets include only measurements that can be found in various sources of publicly available literature. Please further note that the data set was collected with resources from BASF and therefore remains the company's property. We provide the part of the data set that we use for model evaluation, i.e., the test set, publicly available at https://github.com/brozosc/GNNs-for-surfactant-multi-property-prediction. This allows future work to use this test set for model evaluation and comparison. The other part of the data set used for model training could be made available upon request.

Furthermore, we establish a GNN model for the prediction of surface excess concentration (Γ_m) and a GNN model for the prediction of CMC, both trained on the above mentioned new databases. In contrast to previous work, the GNN models developed here explicitly capture edge features and a broader surfactant domain. Additionally, we investigate multi-task learning to overcome data limitations and ensemble learning to enhance the predictive performance. Then, we experimentally measure 3 industrial grade surfactants, previously unseen by the GNN model. Finally, we predict their CMC with our GNN model, which was trained exclusively on literature data with mainly purified surfactants, and demonstrate the model's ability to generalize to unpurified industrial surfactants.

We construct the rest of this work as following: Firstly, we analyze our databases, data sampling procedure and the industrial surfactants

* Corresponding author at: RWTH Aachen University, Process Systems Engineering (AVT.SVT), Aachen 52074, Germany.
E-mail address: amitsos@alum.mit.edu (A. Mitsos).

Table 1
CMC values of dodecylpyridinium bromide reported in literature at 25 °C.

Method	Value (mM)	Source
Tensiometry	11.5	[51]
Conductometry	11.3	[51]
Conductometry	10	[52]
Light scattering	11.6	[53]

used (Section 2). Thereafter, we give an overview of how GNN models work, the methods we applied and a brief overview of the hyperparameter selection (Section 3). We then present our results, compare them with previous works and discuss limitations and possible solutions (Section 4). Lastly, we summarize our work and suggest possible future improvements (Section 5).

2. Data sets

We now analyze the existing databases and describe our methodology for data collection (Section 2.1). Following, we discuss the estimation of CMC and how we handled duplicated values for transfer learning (Section 2.2). Finally, we analyze the collected data sets (Section 2.3) and we present three industrial grade surfactants for model testing (Section 2.4).

2.1. Existing database analysis and data collection

We started our work by building on the publicly available database of 202 substances from Qin et al. [50] for CMC prediction. For the surface excess concentration Γ_m, we had to exclusively rely on tables in books and publications, such as [14,25], because no constructed data set was found in the literature. At first, literature data (CMC and Γ_m) was extracted from multiple sources [14,15,34] for all the molecules at temperatures between 20–28 °C. Note that since temperature massively impacts both properties, we only focus on the temperature range defined above. We also traced back to the individual articles referenced in the sources mentioned above [14,15,34] and extracted additional CMC and Γ_m data. This procedure resulted in an extended data set of 429 distinct substances for CMC. In addition, we simultaneously collected 164 different Γ_m values from multiple sources.

2.2. CMC data collection procedure and duplicate values

During data collection, we often found multiple CMC values for the same surfactant, differing from source to source, due to factors such as purity levels, measuring method of choice and mathematical evaluation of experimental data [28,29]. An example is given in Table 1, where for the same surfactant 4 different values have been reported. The CMC variations are discussed in previous works and remain an issue in surfactant science [15,28,29].

To handle duplicate values, we decided for a ranking according to the measurement method. Here, we prefer CMC values obtained via tensiometry because one of our targets is to evaluate our model on industrial grade surfactants using CMC values measured through tensiometry. If tensiometry data was not available, we favored data from refractometry measurements since it was found to be reliable by Mukerjee and Mysels [15]. If data only from other methods was available, i.e., neither tensiometry nor refractometry, we also included it into our main data set. All remaining values for a surfactant, i.e., duplicates, are not included in the main data set but rather collected in a separate data set, which we utilize for a transfer learning approach (cf. Section 3.5). We note that for most surfactants where duplicate values exist, the values tend to be very similar to each other and in some cases even equal.

Table 2
Number of surfactants per class for each database collected in this work. DV = Duplicate Values.

	CMC	Γ_m	DV-CMC
Nonionics	220	86	19
Anionics	130	44	44
Cationics	55	13	27
Zwitterionics	24	21	9
Total substances	429	164	99

2.3. Data sets analysis

The described sampling process in Section 2.2 led us to construct the three data sets shown in Table 2, together with a detailed surfactant class distribution. We observe that in the two main databases, nonionic surfactants is the dominant class followed by anionics. This class distribution matches with consumption data of surfactants in 2000 [14,31], where anionic and nonionic surfactants are the most used in industrialized areas. In other words, the research focus is matching the industrial output. Afterwards, a statistical overview of the target properties, i.e. CMC and Γ_m, is presented in Fig. 1. We note that Γ_m shows a natural normal distribution without applying the logarithm. Both data sets have similar mean, median, 5th and 95th percentile values although no comparison between them should be made, as CMC values are scaled. The smallest and biggest values in both data sets are similar too. Finally, a correlation plot between $\log$ CMC and Γ_m is given in Fig. 2, containing surfactants for which both CMC and Γ_m values are collected.

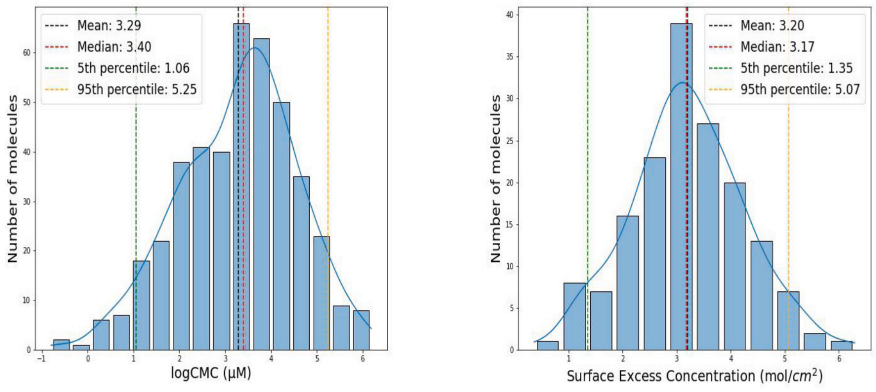
2.4. Industrial surfactants

With an estimated market size of around $40 billion in 2020 [54], surfactants are also heavily researched in the industry. As Myers [31] points out, the majority of academic interest in surfactants, focuses generally on highly purified compounds while the industry is either using complex mixtures to obtain the desired performance or unpurified compounds due to economical reasons. For surface properties like CMC, many authors have noticed the effect of impurities on CMC through the years [15,29,55]. In this study, we examine to what extent GNN models trained exclusively on literature data can generalize for unpurified industrial surfactants.

Three industrial-grade pure-component surfactants were used, provided by BASF, as obtained from production site without further purification. The main species in these surfactants are (S1) Texapon 842 UP (Sodium Caprylyl Sulfate), (S2) Texapon EHS (Sodium 2-Ethylhexyl Sulfate) and (S3) Texapon K 12 G (Sodium Dodecyl Sulfate). We exclude those three molecules from the training set. According to manufacturing process, we expect the presence of unreacted raw material, alcohols in this case, and reaction by-products.

3. Methods

In this section, we first present the fundamentals of a GNN model (Section 3.1), the general training settings of current works (Section 3.2) and the hyperparameter selection (Section 3.3). Afterwards, we refer to the learning techniques applied on this paper (Sections 3.4–3.6). The CMC of the three industrial surfactants was determined by plotting the surface tension as a function of the logarithm of the surfactant concentration. From this plot, two linear regions were determined, which correspond to the linear concentration-dependent and the linear concentration-independent region, respectively. The CMC value is then obtained from the intersection of the straight lines. Finally, for the surface tension measurement a Force Tensiometer – K100 (Krüss, Germany), at 23 °C was used.



(a) CMC

(b) Γ_m

Fig. 1. Statistical overview of CMC and Γ_m databases, assembled from literature in this work. Normal distribution of the data set would ensure an even train-validation-test split without artifact.

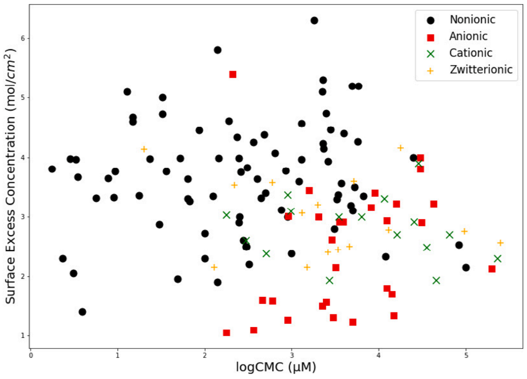


Fig. 2. Correlation plot between log CMC and Γ_m for all surfactant classes. In the plot, 141 surfactants are presented, for which both CMC and Γ_m data was collected from the literature.

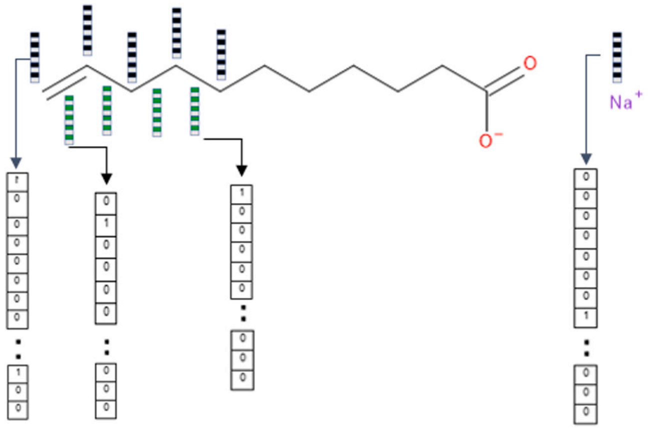


Fig. 3. A surfactant molecule represented as underictedgraph. In every node (atom) a feature vector (black-white color) of size 30 and in every edge (bond) a feature vector (green-white) of size 12 is assigned. The feature vectors encode chemical information about each individual atom and edge respectively. As can be seen in the atom features vector, different atoms have different entries, which distinguishes them from the other atoms. Similarly, bonds are distinguished through their feature vector too.

3.1. Graph neural networks

In GNN models, every molecule is treated as an undirected graph, where atoms correspond to nodes and bonds to edges. A feature vector, containing chemical information, is assigned to each atom and each edge. Our node and edge features of choice are shown in Tables 3 and 4 respectively, motivated from our previous works [47,56] and past literature [42]. Please note that hydrogen atoms are not considered as individual nodes but are implicitly represented in the node feature vector. A surfactant example is given in Fig. 3. The molecular graph then passes through graph convolutions, where neighbor information, i.e., neighboring node and edge features, is aggregated for each node in the graph accordingly. The network depth L, i.e., the number of graph convolutional layers, defines the neighborhood pool from which structural information will be aggregated. We herein use edge-conditioned graph convolutional layers [57] and a gated recurrent unit (GRU) [58], similar to the message passing framework by Gilmer et al. [59] and our previous works [45,47,56]. In contrast to Qin et al. [50] who used graph convolutional layers only considering node features, we thus explicitly include bond type information in learning the molecular structure which potentially facilitates distinguishing molecules with similar heavy atoms but different bonds, e.g., alkanes versus alkenes. After the last graph convolutional layer, the final updated atom feature vectors are pooled into a final molecular fingerprint vector $\mathbf{h}_{FP}$ [59] through a permutation invariant function, i.e., summation of all node vectors. The $\mathbf{h}_{FP}$ contains all the necessary structure-related information of a specific molecule required for molecular property prediction, thereby replacing the selected descriptors in classical QSPR techniques mentioned in Section 1.

Our model is implemented in the Pytorch Geometric (PyG) framework [60]. For the attributed molecular graph generation, we use the SMILES string [61] of each molecule and RDKit (*version 2022.3.5*), an open-source toolkit for cheminformatics.

3.2. General training settings

We use the high-quality data set CMC to define the hyperparameters of our GNN models. For the target property CMC, the log CMC is calculated and then standardized to a zero mean and a standard deviation of one. The train and test sets are separated randomly in a 85%–15% ratio respectively. The training set contains 191 nonionic, 106 anionic, 46 cationic and 21 zwiterrionic surfactant molecules. Thus, each surfactant class is represented in the training set. For the hyperparameter selection an internal validation set is used, which is a subset of the training set with 20 substances each. The loss function is the mean squared error (MSE) and the optimizer is Adam [62]. In general, we use the same general training settings as in our previous works [47,56] and the interested reader can refer to them for more information. For every modeling approach, which will be introduced in Sections 3.4–3.6, as

Table 3
Atom features used in the molecular graph representation. All features are implemented as one-hot-encoding.

Feature	Description	Dimension
Atom type	Atom type (C, N, O, S, F, Cl, Br, Na, I, B, K, H, Li)	13
is aromatic	If the atom is part of an aromatic system	1
hybridization	sp, sp^2, sp^3, sp^3d, or sp^3d^2	5
# bonds	Number of bonds the atom is involved in	6
# Hs	Number of bonded hydrogen atoms	5
Total		**30**

Table 4
Edge features used in the molecular graph representation. All features are implemented as one-hot-encoding.

Feature	Description	Dimension
Bond type	single, double, triple, or aromatic	4
is in a ring	whether the bond is part of a ring	1
conjugated	whether the bond is conjugated	1
stereo	none, any, E/Z, or cis/trans	6
Total		**12**

well as in industrial surfactants application, 40 models were trained on 40 individual training subsets and the results are averaged and reported in Section 4.

3.3. Hyperparameter selection

With the hyperparameter selection procedure, the aim is to find the suitable hyperparameters of our GNN model. For a robust hyperparameter selection, we test each model in 40 different internal validation sets. We perform a grid search for the following hyperparameters of the GNN model, varying them within the respective ranges: Graph convolutional type $\in \{NNConv, GINEConv\}$, number of graph convolutional layers $\in \{1, 2, 3\}$, usage of GRU $\in \{True, False\}$, the batch size $\in \{4, 8, 16\}$, the initial learning rate $\in \{0.005, 0.01, 0.05\}$, dimensions of molecular fingerprint and of MLP $\in \{64, 128, 256\}$. In other words, the hidden layers of the graph convolution part and the hidden layers of the MLP are always the same size. The optimum combination is a GNN architecture with an initial learning rate of 0.005, a hidden state size of 64, a batch size of 16, total graph convolutional layers of 1, the NNConv graph convolutional type and the usage of GRU for the message passing scheme. Our edge feature network, similar to our previous work [56], consists three layers with the following number of neurons: #1 12, #2: 64, and #3: 4096. The architecture exhibits 306,561 learnable parameters in total.

3.4. Single- and multitask learning

A surfactant molecule usually has multiple target properties, as we discussed above in Section 1. The classical learning approach, *single-task learning,* is to train individual models for every property of interest. In single-task learning, model parameters are directly optimized based exclusively on a single target property only, and not transferred to another property prediction task. In that sense, all available QSPR methods for surfactants (discussed in Section 1) are single-task learning. On the other hand, in *multi-task learning* multiple target properties are simultaneously predicted [63,64]. The simultaneous prediction has been shown to improve the modeling accuracy of GNNs in molecular property prediction [44,56,65]. Normally during multi-task learning, the graph convolutional layers are shared and individual MLPs are constructed for each target property. The benefits of this approach, are mainly models' ability to generalize, learn faster, reduce overfitting [66,67] and data efficiency [67].

In the present work, we investigate the prediction of CMC and of Γ_m with both single- and multi-task learning. Specifically, we develop a single-task learning model for each property individually and multi-task learning models for simultaneous prediction of the properties. Since both of the properties come from the same measurement procedure and therefore are correlated, we expect an improved model accuracy during multi-task learning. Note that if only a CMC or Γ_m experimental value was available for a surfactant molecule during multi-task learning, the loss function for that surfactant molecule was calculated only for the property for which a measurement was available. In the multi-task learning loss function, the individual prediction errors of the CMC and Γ_m are weighted equally, i.e., no weighting factor is applied.

3.5. Transfer learning

Another technique for improving machine learning models is transfer learning [68,69]. During transfer learning, a model is usually pretrained on a data set, for example a synthetic one, and then the model parameters are used to initialize the training on a new unseen data set. This technique is very useful when only small data sets are available. In the field of GNNs, researchers investigated the benefits and limitations of transfer learning [70–72].

As we described in Section 2.2, we collect duplicate values to apply transfer learning to single-task CMC prediction, with the scope of utilizing bigger portions of experimental data from the literature. We use the data set DV-CMC (Table 2) to pre-train the model, i.e., learn the graph convolutions and MLP parameters, and afterwards we initiate our single-task CMC model with them. All the initialized parameters are optimized based on the CMC data set (Table 2).

3.6. Ensemble learning

Training and using single models can lead to under- or/and over-predictions. A well-known technique to mitigate this phenomenon in machine learning is ensemble learning [73,74]. In ensemble learning, multiple models are trained on different subsets of training data set and their final predictions are averaged, resulting in more robust and generalized predictions [73–75].

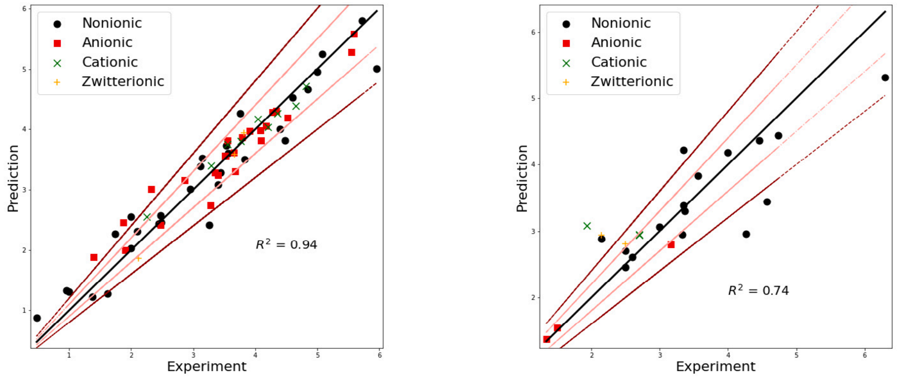
We use ensemble learning both for our single- and multi-task models mentioned in Section 3.4, by training 40 different models in 40 different subsets of our training data set, in each case. Afterwards, we use the 40 different models to perform predictions in our test set, which are averaged to obtain the final scores.

4. Results & discussion

In this section, we firstly summarize the predictive performance of our models (Section 4.1). Afterwards we compare our findings with previous similar work (Section 4.2) and finally we conclude with investigation of model applicability on the selected industrial surfactants (Section 4.3). An overview of the performance of the developed models is reported in Table 5. For every task we report the root mean squared error (RMSE), the mean absolute error (MAE) and the variances on the validation and test sets.

4.1. Predictive performance

The single-task GNN model for CMC exhibits an average RMSE of 0.27 on validation set and 0.33 on test set, while the variance is bigger in the test set than in the validation set. For Γ_m the average RMSE in test set is lower than the one in the validation set, with the former equal to 0.85 and the later equal to 1.02. Using the logarithm of Γ_m did not improve the performance. Our model exhibits great performance in predicting the log CMC, but fails to exhibit similar performance in Γ_m prediction. The reason for the model's under-performance may be the small size of the data set used (140 molecules) for the training and the ambiguous measuring procedures.



(a) CMC test set: Predicted versus experimental value of log(CMC) in μM.

(b) Γ_m test set Predicted versus experimental value of Γ_m in mol/cm^2.

Fig. 4. Multi-task GNN ensemble models for (a) CMC and (b) Γ_m for all surfactant classes. The light red dashed lines represent the 10% error and the dark red dashed lines the 20% error.

Table 5
Summary of model accuracy for different predictive tasks over 40 different runs. In each case the standard deviation is also given, except ensemble learning. In the above table we use the following abbreviations: STL = single-task learning, MTL = multi-task learning, TL = transfer learning, EL = ensemble learning, MAE = mean absolute error, RMSE = root mean squared error.

	CMC		Γ_m	
	RMSE	MAE	RMSE	MAE
STL (val)	0.27 ± 0.027	0.2 ± 0.026	1.02 ± 0.169	0.85 ± 0.15
STL (test)	0.33 ± 0.033	0.25 ± 0.026	0.8 ± 0.143	0.57 ± 0.146
STL & TL (val)	0.27 ± 0.034	0.2 ± 0.033		
STL & TL (test)	0.33 ± 0.042	0.26 ± 0.034		
MTL (val)	0.26 ± 0.075	0.2 ± 0.053	0.3 ± 0.121	0.26 ± 0.116
MTL (test)	0.36 ± 0.041	0.27 ± 0.031	0.59 ± 0.051	0.43 ± 0.044
STL & EL (test)	**0.28**	**0.21**	**0.76**	**0.53**
MTL & EL (test)	**0.31**	**0.23**	**0.56**	**0.4**

In multi-task learning, the GNN model for CMC prediction exhibits an average RMSE of 0.26 on validation set and 0.36 on test set. For Γ_m prediction, the average RMSE in test set is again lower than the one in the validation set, with the former equal to 0.59 and the later equal to 0.43. In the CMC task, the model perform identical with the one in single-task learning, both for validation and test sets, while in the Γ_m task the multi-task model exhibits significantly better performance on the validation set, with the RMSE reducing by 60%, and improved performance on the test set, with the RMSE reducing by 20%. Therefore, we conclude that the data limitations of the Γ_m database as single target property can be overcomed by applying multi-task learning. On the other hand, the CMC model did not benefit from the additional data and showed identical results with a slightly higher variance but an overall similar accuracy compared to the single-task learning.

The transfer learning approach, i.e., pre-training the model using the 99 collected duplicate values described in Section 2.2, is applied only on the single-task CMC model. The RMSE on the validation set remains the same as in single-task learning, equal to 0.27 and on the test set equal to 0.33. Interestingly, the standard deviation increases in both sets. The increase may be due to the broader range of target values for the same property, which leads the model to deviate more from the true value. We observed that transfer learning slightly reduced the final model training time, i.e., the model reached its optimum sooner. Besides the slight reduction of final model training time, using duplicate values for transfer learning led to similar performance.

Ensemble learning, i.e., averaging the predictions of the 40 trained models, slightly reduces the RMSE on test set for single-task learning to 0.28 and for multitask learning to 0.31 in the CMC case. A similar RMSE reduction is observed on test set for Γ_m accordingly, to 0.76 for single-task learning and to 0.56 for multitask learning. We use the ensembled results in multitask CMC and Γ_m learning to draw the parity plots, shown in Fig. 4 on the independent test sets. The parity plots (measured vs predicted values) for CMC and Γ_m show a high determination coefficient for the former, R^2_{CMC} = 0.94, and moderate one for the latter $R^2_{\Gamma_m}$ = 0.74. We demonstrate that the GNN approach in the present work, can predict CMC and Γ_m across all surfactant classes.

In addition, we present the three components with the highest absolute CMC error in Fig. 5. For molecule one, the combination of high CMC value and lack of similar molecules, i.e. small alkyl chain with high number of ethylene oxides in the training set, may be the reasons why the model fails to perform well. For molecules two and three, we suspect the measurement may have an impact on the result since identical molecules can be found in the training set.

Similarly, the four components with the highest absolute Γ_m error are illustrated in Fig. 6. Molecule four is similar as in the CMC case, which supports the measurement impacted hypothesis from before. On the other side molecules one and three have complex structure, where at a similar chemistry is lacking in the training set. Therefore, we can assume that the model fails to capture the property–structure relationship in this case. The same reasoning can be applied to molecule two, where we also lack similar molecules in the training set.

4.2. Comparison with previous works

We next compare our results to the work of Qin et al. [50] and their GNN model for CMC prediction, which used a subset of our CMC training data but is also applicable for a wide choice of surfactant classes. Qin et al. [50] report a test RMSE of 0.30 which is similar to ours of 0.28. Note that our test set is almost three times the size of the one used by Qit et al. [50], so that we cover a higher variance of molecules. Besides the slight RMSE improvement in test set, our model reduces also the average RMSE on the validation set from 0.39 to 0.27. As can be seen in Fig. 7, no major outliers were observed in the 40 models. Finally, most of the models exhibited a test RMSE in the range of 0.26–0.29 while Qin et al. [50] reported a broader test RMSE range of 0.28–0.45.

Overall, the comparison shows that the general performance of our model on the CMC is similarly high with previous ones, although a direct quantitative comparison is not possible due to the different data sets used. As there is no model for predicting surface excess concentration directly from experimental data (cf. Section 1), we are unable to compare our results for Γ_m with other works.

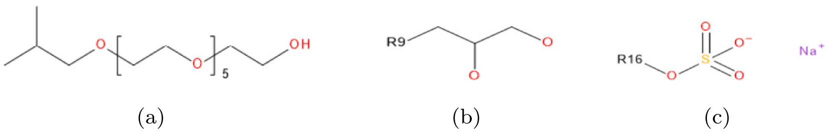


Fig. 5. Outliers in multi-task learning for CMC prediction. Two of them, (a) and (b) belong to nonionic class and the third, (c) to anionic class.

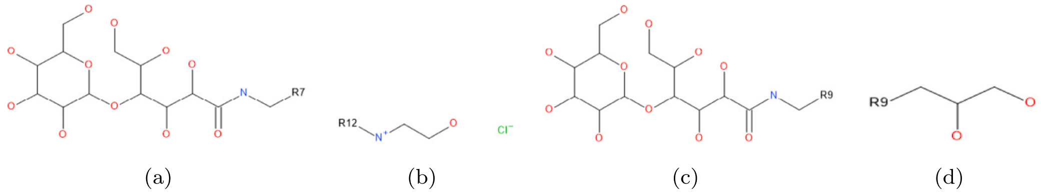


Fig. 6. Outliers in multi-task learning for Γ_m prediction. Three of them, (a), (c) and (d) belong to nonionic class and the second, (b) to cationic class.

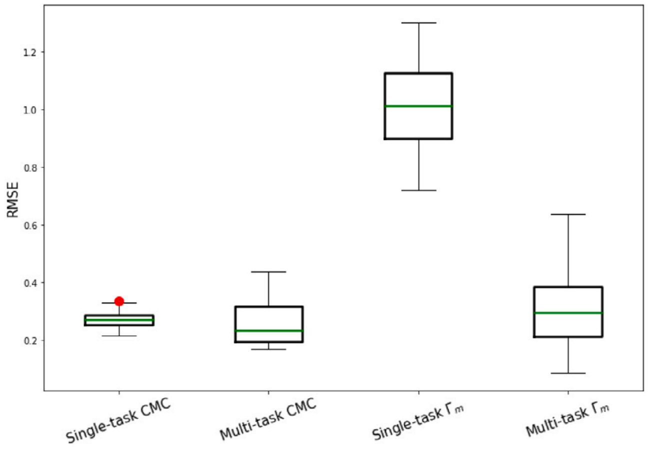


Fig. 7. Distribution plot of RMSE on internal validation set for all the learning tasks. The boxplots are the results of 40 runs in different internal validation sets. The red points represent the outliers.

4.3. Industrial surfactants

We then apply our developed GNN model to predict on the three pure component industrial surfactants described in Section 2.4. According to the above discussed results, the best learning approach for CMC is the combination of single-task with ensemble learning. We use the 40 trained models from Section 4.1 to perform ensembled predictions on the three surfactants. The predicted log CMC values, as well the experimental measured ones, are given in Table 6 for comparison. For all of the three, the predicted values are very close to the measured ones. Overall, the data indicates that the developed GNN model trained on literature data can accurately predict the CMCs for all three single-molecule industrial unpurified surfactants.

We note that similar molecules are used in the training set and none of the three exhibits high structural complexity. Specifically, the training set contains straight-chain, long alkyl sulfates with a different number of carbons than S1 and S3, but an overall very similar structure. The training set also includes short-chain alkyl alcohols (different head group) and branched-chain, long alkyl sulfates that are structurally similar to S2 [14,76]. On the other hand, the impurities effects on CMC are not learnt from the model and their implementation could potentially be future area of research. Future work could also focus on testing the applicability of GNN model in research of bio-surfactants, with many of them naturally exhibit high structural complexity.

Table 6
Comparison of predicted values from our single-task GNN with ensemble versus experimentally calculated values for three selected industrial grade surfactants. For the predicted values, the standard deviation over 40 individual runs is also given. The above values, are the logarithmic ones.

	Predicted CMC (μM)	Measured CMC (μM)
S1	4.87 ± 0.11	4.88
S2	4.95 ± 0.17	4.98
S3	3.91 ± 0.08	3.86

5. Conclusions and future work

We apply GNNs to pure component surfactants to predict CMC and Γ_m. Based on extensive literature scanning, we generate a database for CMC with double the size of existing. We also construct a data set for Γ_m. As GNNs have been successfully used for CMC prediction [50], we extend the GNN architecture to simultaneously predict the surface excess concentration Γ_m in a multi-task learning, thereby utilizing correlations between these two properties. In contrast to previous works, we herein implement a GNN architecture where edge features are explicitly captured. Furthermore, to the best our of knowledge, we develop the first openly available ML model to predict Γ_m from the surfactant structure. Furthermore, we collect additional CMC values from the literature and investigate if transfer learning can increase the model accuracy.

All GNN models exhibit high-accuracy CMC predictions, on a comparable level to a recently developed GNN model by Qin et al. [50] but for an extended spectrum of surfactants. For Γ_m on the other hand, the single-task GNNs fail to capture the property–structure relationship; here, we find that multi-task GNNs effectively utilizes the CMC data to substantially enhance prediction accuracy for Γ_m. Therefore, we find multi-task learning to be an effective learning technique to overcome data availability problems in the field of surfactant property prediction. In all cases, ensemble learning increases the prediction accuracy. For transfer learning, however, we observe no improvements in the model accuracy. Finally, we test the best GNN model for CMC on three unpurified industrial surfactants and find highly accurate predictions matching our laboratory measurements, thereby indicating strong potential for further industrial applications.

Furthermore, our GNN models are subject to certain limitations. As is typically the case for ML models, the applicability of our GNN models is limited to surfactants with similar structure as the ones contained in the training data set. Stereochemistry is also not taken into account in this work, for example in the case of n-dodecyl-D-maltoside we only use the CMC of the α isomer [77]. Another limitation, shortly mentioned above, is not including information regarding the purity of each compound. We only considered highly purified compounds reported in literature and future research should focus in incorporating surface active impurities to the GNN model.

Future work could extend the relative small database for surface excess concentration Γ_m, thus yielding higher performance predictive models. For Γ_m noise is often encountered in reported values due to it is implicit calculation through various approaches of the Gibbs adsorption equation [14]. This noise prohibits the GNN model to better capture the structure–property relationship. Finally, prediction of further surfactant properties based on the structure would be highly interesting for surfactant formulators.

CRediT authorship contribution statement

Christoforos Brozos: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Data curation, Conceptualization. **Jan G. Rittig:** Writing – review & editing, Software, Methodology, Formal analysis, Conceptualization. **Sandip Bhattacharya:** Writing – review & editing, Methodology, Conceptualization. **Elie Akanny:** Writing – review & editing, Methodology, Conceptualization. **Christina Kohlmann:** Writing – review & editing, Supervision, Funding acquisition. **Alexander Mitsos:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jan Rittig and Alexander Mitsos have no competing interests. Christoforos Brozos, Sandip Bhattacharya, Elia Akanny and Christina Kohlmann are employees of BASF Personal care which uses surfactants commercially. We have submitted a patent disclosure partially based on the methods presented herein.

Data availability

All python scripts and the test data used in this work are available as open-source at https://github.com/brozosc/GNNs-for-surfactant-multi-property-prediction. We do not directly provide the training set used in this work, as it remains the property of BASF, but it could be made available upon request.

Acknowledgments

The BASF authors (C. Brozos, S. Bhattacharya, E. K. Akanny and C. Kohlmann) were funded by the BASF Personal Care and Nutrition GmbH. J. G. Rittig and A. Mitsos acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 466417970 – within the Priority Programme “SPP 2331: Machine Learning in Chemical Engineering”. Additionally, J. G. Rittig acknowledges the support of the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE).

References

[1] I.M.M. Vieira, B.L.P. Santos, D.S. Ruzene, D.P. Silva, An overview of current research and developments in biosurfactants, J. Ind. Eng. Chem. 100 (2021) 1–18, http://dx.doi.org/10.1016/j.jiec.2021.05.017, URL https://www.sciencedirect.com/science/article/pii/S1226086X21002914.

[2] S.M. Shaban, J. Kang, D.-H. Kim, Surfactants: Recent advances and their applications, Compos. Commun. 22 (2020) 100537, http://dx.doi.org/10.1016/j.coco.2020.100537, URL https://www.sciencedirect.com/science/article/pii/S2452213920302655.

[3] M. Nitschke, S. Costa, Biosurfactants in food industry, Trends Food Sci. Technol. 18 (5) (2007) 252–259, http://dx.doi.org/10.1016/j.tifs.2007.01.002, URL https://www.sciencedirect.com/science/article/pii/S0924224407000362.

[4] T. Tadros, Surfactants in Personal Care and Cosmetics, John Wiley & Sons, Ltd, 2005, pp. 399–432, http://dx.doi.org/10.1002/3527604812.ch12, Ch. 12. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/3527604812.ch12. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/3527604812.ch12.

[5] S.A. Adu, P.J. Naughton, R. Marchant, I.M. Banat, Microbial biosurfactants in cosmetic and personal skincare pharmaceutical formulations, Pharmaceutics 12 (2020).

[6] A. Szűts, P. Szabó-Révész, Sucrose esters as natural surfactants in drug delivery systems—a mini-review, Int. J. Pharm. 433 (1) (2012) 1–9, http://dx.doi.org/10.1016/j.ijpharm.2012.04.076, URL https://www.sciencedirect.com/science/article/pii/S037851731200422X.

[7] F. Zhang, S. Li, Q. Zhang, J. Liu, S. Zeng, M. Liu, D. Sun, Adsorption of different types of surfactants on graphene oxide, J. Mol. Liq. 276 (2019) 338–346, http://dx.doi.org/10.1016/j.molliq.2018.12.009, URL https://www.sciencedirect.com/science/article/pii/S0167732218334196.

[8] K. Knop, R. Hoogenboom, D. Fischer, U. Schubert, Poly(ethylene glycol) in drug delivery: Pros and cons as well as potential alternatives, Angew. Chem. Int. Ed. 49 (36) (2010) 6288–6308, http://dx.doi.org/10.1002/anie.200902672, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.200902672. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.200902672.

[9] L. Schramm, E. Stasiuk, G. Marangoni, Surfactants and their applications, Annu. Rep. Prog. Chem. Sect. C 99 (2003) 3–48, http://dx.doi.org/10.1039/B208499F.

[10] F. Gallou, N. Isley, A. Ganic, U. Onken, M. Parmentier, Surfactant technology applied toward an active pharmaceutical ingredient: more than a simple green chemistry advance, Green Chem. 18 (2015) http://dx.doi.org/10.1039/C5GC02371H.

[11] Z. Yang, C.G. Brouillette, Chapter thirteen - a guide to differential scanning calorimetry of membrane and soluble proteins in detergents, in: A.L. Feig (Ed.), Calorimetry, in: Methods in Enzymology, vol. 567, Academic Press, 2016, pp. 319–358, http://dx.doi.org/10.1016/bs.mie.2015.08.014, URL https://www.sciencedirect.com/science/article/pii/S0076687915004668.

[12] A. Al-Sabagh, N. Nasser, M. Migahed, N. Kandil, Effect of chemical structure on the cloud point of some new non-ionic surfactants based on bisphenol in relation to their surface active properties, Egypt. J. Petrol. 20 (2) (2011) 59–66, http://dx.doi.org/10.1016/j.ejpe.2011.06.006, URL https://www.sciencedirect.com/science/article/pii/S1110062111000079.

[13] A. Patist, S. Oh, R. Leung, D. Shah, Kinetics of micellization: Its significance to technological processes, Colloids Surf. A 176 (2001) 3–16, http://dx.doi.org/10.1016/S0927-7757(00)00610-5.

[14] M. Rosen, J. Kunjappu, Surfactants and Interfacial Phenomena: Rosen/Surfactants 4E, John Wiley & Sons, Ltd, 2012, http://dx.doi.org/10.1002/9781118228920.

[15] P. Mukerjee, K.J. Mysels, Critical Micelle Concentrations of Aqueous Surfactant Systems, Tech. Rep., National Standard reference data system, 1971.

[16] C.J. Thompson, N. Ainger, P. Starck, O.O. Mykhaylyk, A.J. Ryan, Shampoo science: A review of the physiochemical processes behind the function of a shampoo, Macromol. Chem. Phys. 224 (3) (2023) 2200420, http://dx.doi.org/10.1002/macp.202200420, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/macp.202200420. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/macp.202200420.

[17] H. Su, F. Wang, W. Ran, W. Zhang, W. Dai, H. Wang, C.F. Anderson, Z. Wang, C. Zheng, P. Zhang, Y. Li, H. Cui, The role of critical micellization concentration in efficacy and toxicity of supramolecular polymers, Proc. Natl. Acad. Sci. USA 117 (2020) 4518–4526.

[18] M. Ghezzi, S. Pescina, C. Padula, P. Santi, E. Del Favero, L. Cantù, S. Nicoli, Polymeric micelles in drug delivery: An insight of the techniques for their characterization and assessment in biorelevant conditions, J. Control. Release 332 (2021) 312–336, http://dx.doi.org/10.1016/j.jconrel.2021.02.031, URL https://www.sciencedirect.com/science/article/pii/S0168365921001061.

[19] S. Perumal, R. Atchudan, W. Lee, A review of polymeric micelles and their applications, Polymers 14 (2022).

[20] A. Kumar, A. Mandal, Critical investigation of zwitterionic surfactant for enhanced oil recovery from both sandstone and carbonate reservoirs: Adsorption, wettability alteration and imbibition studies, Chem. Eng. Sci. 209 (2019) 115222, http://dx.doi.org/10.1016/j.ces.2019.115222, URL https://www.sciencedirect.com/science/article/pii/S0009250919307146.

[21] D.R. Perinelli, M. Cespi, L. Casettari, D. Vllasaliu, M. Cangiotti, M.F. Ottaviani, G. Giorgioni, G. Bonacucina, G.F. Palmieri, Correlation among chemical structure, surface properties and cytotoxicity of n-acyl alanine and serine surfactants, Eur. J. Pharmaceut. Biopharmaceut. 109 (2016) 93–102, http://dx.doi.org/10.1016/j.ejpb.2016.09.015, URL https://www.sciencedirect.com/science/article/pii/S0939641116306154.

[22] T. Majeed, T.I. Sölling, M.S. Kamal, Foamstability: The interplay between salt-, surfactant- and critical micelle concentration, J. Pet. Sci. Eng. 187 (2020) 106871, http://dx.doi.org/10.1016/j.petrol.2019.106871, URL https://www.sciencedirect.com/science/article/pii/S0920410519312872.

[23] A.R. Katritzky, L.M. Pacureanu, S.H. Slavov, D.A. Dobchev, M. Karelson, Qspr study of critical micelle concentrations of nonionic surfactants, Ind. Eng. Chem. Res. 47 (23) (2008) 9687–9695, http://dx.doi.org/10.1021/ie800954k.

[24] S. Thiruvengadam, M. Murphy, J.S. Tan, K. Miller, A generalized theoretical model for the relationship between critical micelle concentrations, pressure, and temperature for surfactants, J. Surfactants Deterg. 23 (2) (2020) 273–303, http://dx.doi.org/10.1002/jsde.12360, arXiv:https://aocs.onlinelibrary.wiley.com/doi/pdf/10.1002/jsde.12360. URL https://aocs.onlinelibrary.wiley.com/doi/abs/10.1002/jsde.12360.

[25] M. Dahanayake, A.W. Cohen, M.J. Rosen, Relationship of structure to properties of surfactants. 13. surface and thermodynamic properties of some oxyethylenated sulfates and sulfonates, J. Phys. Chem. 90 (11) (1986) 2413–2418, http://dx.doi.org/10.1021/j100402a032.

[26] O. Ortona, V. Vitagliano, L. Paduano, L. Costantino, Microcalorimetric study of some short-chain nonionic surfactants, J. Colloid Interface Sci. 203 (2) (1998) 477–484, http://dx.doi.org/10.1006/jcis.1998.5519, URL https://www.sciencedirect.com/science/article/pii/S0021979798955199.

[27] D.J. Jobe, V.C. Reinsborough, Micellar properties of sodium alkyl sulfoacetates and sodium dialkyl sulfosuccinates in water, Can. J. Chem. 62 (1984) 280–284.

[28] D.R. Perinelli, M. Cespi, N. Lorusso, G.F. Palmieri, G. Bonacucina, P. Blasi, Surfactant self-assembling and critical micelle concentration: One approach fits all? Langmuir : ACS J. Surf. Colloids 36 (2020) 5745–5753.

[29] S.P. Moulik, A.K. Rakshit, B. Naskar, Evaluation of non-ambiguous critical micelle concentration of surfactants in relation to solution behaviors of pure and mixed surfactant systems: A physicochemical documentary and analysis, J. Surfactants Deterg. 24 (4) (2021) 535–549, http://dx.doi.org/10.1002/jsde.12503, arXiv:https://aocs.onlinelibrary.wiley.com/doi/pdf/10.1002/jsde.12503. URL https://aocs.onlinelibrary.wiley.com/doi/abs/10.1002/jsde.12503.

[30] Z. Wang, Y. Chen, F. Zhang, S. Lin, Significance of surface excess concentration in the kinetics of surfactant-induced pore wetting in membrane distillation, Desalination 450 (2019) 46–53, http://dx.doi.org/10.1016/j.desal.2018.10.024, URL https://www.sciencedirect.com/science/article/pii/S0011916418316436.

[31] D. Myers, Surfactant Science and Technology, fourth ed., John Wiley & Sons, Ltd, 2020.

[32] M.J. Rosen, A.W. Cohen, M. Dahanayake, X.Y. Hua, Relationship of structure to properties in surfactants. 10. surface and thermodynamic properties of 2-dodecyloxypoly(ethenoxyethanol)s, c12h25(oc2h4)xoh, in aqueous solution, J. Phys. Chem. 86 (4) (1982) 541–545, http://dx.doi.org/10.1021/j100393a025.

[33] J. Hu, X. Zhang, Z. Wang, A review on progress in qspr studies for surfactants, Int. J. Mol. Sci. 11 (2010) 1020–1047, http://dx.doi.org/10.3390/ijms11031020.

[34] T. Gaudin, P. Rotureau, I. Pezron, G. Fayet, New qspr models to predict the critical micelle concentration of sugar-based surfactants, Ind. Eng. Chem. Res. 55 (45) (2016) 11716–11726, http://dx.doi.org/10.1021/acs.iecr.6b02890.

[35] M. Mattei, G.M. Kontogeorgis, R. Gani, Modeling of the critical micelle concentration (cmc) of nonionic surfactants with an extended group-contribution method, Ind. Eng. Chem. Res. 52 (34) (2013) 12236–12246, http://dx.doi.org/10.1021/ie4016232.

[36] K. Roy, H. Kabir, Qspr with extended topochemical atom (eta) indices: Exploring effects of hydrophobicity, branching and electronic parameters on logcmc values of anionic surfactants, Chem. Eng. Sci. 87 (2013) 141–151, http://dx.doi.org/10.1016/j.ces.2012.10.002, URL https://www.sciencedirect.com/science/article/pii/S0009250912005970.

[37] X. Li, G. Zhang, J. Dong, X. Zhou, X. Yan, M. Luo, Estimation of critical micelle concentration of anionic surfactants with qspr approach, J. Mol. Struct.: THEOCHEM 710 (1) (2004) 119–126, http://dx.doi.org/10.1016/j.theochem.2004.08.039, URL https://www.sciencedirect.com/science/article/pii/S0166128004006323.

[38] J. Wu, F. Yan, Q. Jia, Q. Wang, Qspr for predicting the hydrophile-lipophile balance (hlb) of non-ionic surfactants, Colloids Surf. A 611 (2021) 125812, http://dx.doi.org/10.1016/j.colsurfa.2020.125812, URL https://www.sciencedirect.com/science/article/pii/S0927775720314059.

[39] Y. Shi, F. Yan, Q. Jia, Q. Wang, Norm descriptors for predicting the hydrophile-lipophile balance (hlb) and critical micelle concentration (cmc) of anionic surfactants, Colloids Surf. A 583 (2019) 123967, http://dx.doi.org/10.1016/j.colsurfa.2019.123967, URL https://www.sciencedirect.com/science/article/pii/S0927775719309574.

[40] Z.-W.W. Mei-Ling Chen, H.-J. Duan, Qspr for hlb values of nonionic surfactants using two simple descriptors, J. Dispers. Sci. Technol. 30 (10) (2009) 1481–1485, http://dx.doi.org/10.1080/01932690903123338, arXiv:https://doi.org/10.1080/01932690903123338.

[41] D. Seddon, E.A. Müller, J.T. Cabral, Machine learning hybrid approach for the prediction of surface tension profiles of hydrocarbon surfactants in aqueous solution, J. Colloid Interface Sci. 625 (2022) 328–339, http://dx.doi.org/10.1016/j.jcis.2022.06.034, URL https://www.sciencedirect.com/science/article/pii/S0021979722010013.

[42] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, R. Barzilay, Analyzing learned molecular representations for property prediction, J. Chem. Inf. Model. 59 (8) (2019) 3370–3388, http://dx.doi.org/10.1021/acs.jcim.9b00237, pMID: 31361484. arXiv:https://doi.org/10.1021/acs.jcim.9b00237.

[43] O. Wieder, S. Kohlbacher, M. Kuenemann, A. Garon, P. Ducrot, T. Seidel, T. Langer, A compact review of molecular property prediction with graph neural networks, Drug Discov. Today: Technol. 37 (2020) 1–12, http://dx.doi.org/10.1016/j.ddtec.2020.11.009, URL https://www.sciencedirect.com/science/article/pii/S1740674920300305.

[44] M.L. Pasini, P. Zhang, S.T. Reeve, J.Y. Choi, Multi-task graph neural networks for simultaneous prediction of global and atomic properties in ferromagnetic systems*, Mach. Learn.: Sci. Technol. 3 (2) (2022) 025007, http://dx.doi.org/10.1088/2632-2153/ac6a51.

[45] J.G. Rittig, Q. Gao, M. Dahmen, A. Mitsos, A.M. Schweidtmann, Graph neural networks for the prediction of molecular structure–property relationships, 2022, arXiv:2208.04852.

[46] E.I. Sanchez Medina, S. Linke, M. Stoll, K. Sundmacher, Graph neural networks for the prediction of infinite dilution activity coefficients, Digit. Discov. 1 (2022) 216–225, http://dx.doi.org/10.1039/D1DD00037C.

[47] J.G. Rittig, K. Ben Hicham, A.M. Schweidtmann, M. Dahmen, A. Mitsos, Graph neural networks for temperature-dependent activity coefficient prediction of solutes in ionic liquids, Comput. Chem. Eng. 171 (2023) 108153, http://dx.doi.org/10.1016/j.compchemeng.2023.108153.

[48] K.C. Felton, H. Ben-Safar, A.A. Lapkin, DeepGamma: A deep learning model for activity coefficient prediction, 2021.

[49] J.G. Rittig, K.C. Felton, A.A. Lapkin, A. Mitsos, Gibbs-Duhem-informed neural networks for binary activity coefficient prediction, Digit. Discov. 2 (2023) 1752–1767, http://dx.doi.org/10.1039/D3DD00103B.

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

[51] M.J. Rosen, M. Dahanayake, A.W. Cohen, Relationship of structure to properties in surfactants. 11. surface and thermodynamic properties of n-dodecyl-pyridinium bromide and chloride, Colloids Surf. 5 (2) (1982) 159–172, http://dx.doi.org/10.1016/0166-6622(82)80071-1, URL https://www.sciencedirect.com/science/article/pii/0166662282800711.

[52] J. Škerjanc, K. Kogej, J. Cerar, Equilibrium and transport properties of alkylpyridinium bromides, Langmuir 15 (15) (1999) 5023–5028, http://dx.doi.org/10.1021/la981710+, arXiv:https://doi.org/10.1021/la981710+.

[53] W. Ford, R. Ottewill, H. Parreira, Light-scattering studies on dodecylpyridinium halides, J. Colloid Interface Sci. 21 (5) (1966) 522–533, http://dx.doi.org/10.1016/0095-8522(66)90050-X, URL https://www.sciencedirect.com/science/article/pii/009585226690050X.

[54] The global surfactants market is projected to grow from 41.22 billion in 2021 to 57.81 billion by 2028 at a cagr of 4.9 read more at:- https://www.fortunebusinessinsights.com/surfactants-market-102385. (2021). URL https://www.fortunebusinessinsights.com/surfactants-market-102385.

[55] A. Patist, S.S. Bhagwat, K.W. Penfield, P. Aikens, D.O. Shah, On the measurement of critical micelle concentrations of pure and technical-grade nonionic surfactants, J. Surfactants Deterg. 3 (1) (2000) 53–58, http://dx.doi.org/10.1007/s11743-000-0113-4.

[56] A.M. Schweidtmann, J.G. Rittig, A. König, M. Grohe, A. Mitsos, M. Dahmen, Graph neural networks for prediction of fuel ignition quality, Energy Fuels 34 (9) (2020) 11395–11407, http://dx.doi.org/10.1021/acs.energyfuels.0c01533, arXiv:https://doi.org/10.1021/acs.energyfuels.0c01533.

[57] M. Simonovsky, N. Komodakis, Dynamic edge-conditioned filters in convolutional neural networks on graphs, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 29–38, http://dx.doi.org/10.1109/CVPR.2017.11.

[58] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder–decoder for statistical machine translation, 2014, arXiv:1406.1078.

[59] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural message passing for quantum chemistry, 2017, arXiv:1704.01212.

[60] M. Fey, J.E. Lenssen, Fast graph representation learning with pytorch geometric, 2019, arXiv abs/1903.02428.

[61] D. Weininger, Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1) (1988) 31–36, http://dx.doi.org/10.1021/ci00057a005.

[62] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, CoRR abs/1412.6980.

[63] R. Caruana, Multitask learning, Mach. Learn. 28 (1) (1997) 41–75, http://dx.doi.org/10.1023/A:1007379606734.

[64] Y. Zhang, Q. Yang, A survey on multi-task learning, IEEE Trans. Knowl. Data Eng. 34 (2017) 5586–5609.

[65] F. Capela, V. Nouchi, R.V. Deursen, I.V. Tetko, G. Godin, Multitask learning on graph neural networks applied to molecular property predictions, 2019, arXiv:1910.13124.

[66] S. Ruder, An overview of multi-task learning in deep neural networks, 2017, arXiv:1706.05098.

[67] M. Crawshaw, Multi-task learning with deep neural networks: A survey, 2020, arXiv abs/2009.09796.

[68] D. Hendrycks, K. Lee, M. Mazeika, Using pre-training can improve model robustness and uncertainty, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 2712–2721, URL https://proceedings.mlr.press/v97/hendrycks19a.html.

[69] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proc. IEEE PP (2020) 1–34, http://dx.doi.org/10.1109/JPROC.2020.3004555.

[70] C. Grambow, Y.-P. Li, W.H. Green, Accurate thermochemistry with small data sets: A bond additivity correction and transfer learning approach, J. Phys. Chem. A 123 (27) (2019) 5826–5835, http://dx.doi.org/10.1021/acs.jpca.9b04195.

[71] X. Han, Z. Huang, B. An, J. Bai, Adaptive transfer learning on graph neural networks, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, KDD '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 565–574, http://dx.doi.org/10.1145/3447548.3467450.

[72] N. Kooverjee, S. James, T. van Zyl, Investigating transfer learning in graph neural networks, Electronics 11 (8) (2022) http://dx.doi.org/10.3390/electronics11081202, URL https://www.mdpi.com/2079-9292/11/8/1202.

[73] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140, http://dx.doi.org/10.1023/A:1018054314350.

[74] T.G. Dietterich, Ensemble methods in machine learning, in: Proceedings of the First International Workshop on Multiple Classifier Systems, MCS '00, Springer-Verlag, Berlin, Heidelberg, 2000, pp. 1–15.

[75] M. Ganaie, M. Hu, A. Malik, M. Tanveer, P. Suganthan, Ensemble deep learning: A review, Eng. Appl. Artif. Intell. 115 (2022) 105151, http://dx.doi.org/10.1016/j.engappai.2022.105151, URL https://www.sciencedirect.com/science/article/pii/S095219762200269X.

[76] Y. Kato, Formation of a micelle-like structure in aqueous solution of glycols, Chem. Pharm. Bull. 10 (1962) 771–788.

[77] C. Pagliano, S. Barera, F. Chimirri, G. Saracco, J. Barber, Comparison of the α and β isomeric forms of the detergent n-dodecyl-d-maltoside for solubilizing photosynthetic complexes from pea thylakoid membranes, Biochim. Biophys. Acta (BBA) - Bioenerg. 1817 (8) (2012) 1506–1515, http://dx.doi.org/10.1016/j.bbabio.2011.11.001, photosynthesis Research for Sustainability: From Natural to Artificial. URL https://www.sciencedirect.com/science/article/pii/S000527281100260X.