Computers and Chemical Engineering 198 (2025) 109085

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Predicting the temperature-dependent CMC of surfactant mixtures with graph neural networks

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

[a] BASF Personal Care and Nutrition GmbH, Henkelstrasse 67, 40589 Duesseldorf, Germany
[b] RWTH Aachen University, Process Systems Engineering (AVT.SVT), 52074 Aachen, Germany
[c] Forschungszentrum Jülich GmbH, Institute of Energy and Climate Research IEK-10 – Energy Systems Engineering, 52425 Jülich, Germany
[d] JARA Center for Simulation and Data Science (CSD), 52056 Aachen, Germany

ARTICLE INFO

Keywords:
Surfactants
Mixtures
Critical micelle concentration
Graph neural network
Commercial surfactant

ABSTRACT

Surfactants are key ingredients in various industries such as personal and home care with the critical micelle concentration (CMC) being of major interest. Predictive models for CMC of pure surfactants have been developed based on recent ML methods, however, in practice surfactant mixtures are typically used due to performance, environmental, and cost reasons. Herein, we develop a graph neural network (GNN) framework for surfactant mixtures to predict the temperature-dependent CMC. We collect data for 108 surfactant binary mixtures, to which we add data for pure species from our previous work Brozos et al. (2024). We then develop and train GNNs and evaluate their accuracy across different prediction test scenarios for binary mixtures relevant to practical applications. The final GNN models demonstrate very high predictive performance when interpolating between different mixture compositions and for new binary mixtures with known species. Extrapolation to binary surfactant mixtures where either one or both surfactant species are not seen before, yields accurate results for the majority of surfactant systems. We further find superior accuracy of the GNN over a semi-empirical model based on activity coefficients, which has been widely used to date. We then explore if GNN models trained solely on binary mixture and pure species data can also accurately predict the CMCs of ternary mixtures. Finally, we experimentally measure the CMC of 4 commercial surfactants that contain up to four species and industrial relevant mixtures and find a very good agreement between measured and predicted CMC values.

1. Introduction

Surfactant mixtures commonly exhibit advantageous synergistic properties compared to single surfactants and are therefore preferred in many applications (Rosen and Kunjappu, 2012; Rosen and Hua, 1982; Huang and Ren, 2017; Kumar Shah et al., 2022; Geng et al., 2017; Moulik et al., 2021; Cheng et al., 2020), such as personal care products, detergents, pharmaceuticals and others (Shaban et al., 2020; Nitschke and Costa, 2007; Massarweh and Abushaikha, 2020; Tadros, 2005; Hunter and Fowler, 1998; Rodríguez Patino et al., 2008). Commercial formulations developed in the cosmetics and detergents industries contain almost exclusively surfactant mixtures (Myers, 2020; Grady, 2023; Kelleppan et al., 2023). This is due either to the existence of a homologous distribution in an industrial grade surfactant or the combination of surfactants driven by performance, cost, and sustainability aspects. Therefore, understanding and modeling surfactant mixtures is essential for the design of further tailored products.

The properties of surfactant mixtures are highly dependent on the interaction effects between the surfactant structures. Surfactants are amphiphilic molecules composed of a hydrophilic (head) and a hydrophobic (tail) part. An important property of a surfactant mixture is the critical micelle concentration (CMC), which is the minimum concentration of surfactant that causes the formation of surfactant micelles (Rosen and Kunjappu, 2012; Myers, 2020). Surfactant mixtures are generally described as complex systems due to the difference between the bulk and micelle concentration of the pure species, i.e., the mixture components (Grady, 2023; Clint, 1975). To elaborate further, when a surfactant species is part of a mixture, its concentration in the bulk will typically differ from its concentration in the micelle. A binary mixture exhibits synergism, if at any molar fraction the mixture CMC, denoted as CMC_{mix}, is lower than the CMC of either pure species, and antagonism if CMC_{mix} is higher than the CMC of either

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

https://doi.org/10.1016/j.compchemeng.2025.109085
Received 4 January 2025; Received in revised form 23 February 2025; Accepted 28 February 2025
Available online 10 March 2025

pure species (Rosen and Kunjappu, 2012; Alargova et al., 2001). A smaller CMC indicates that a lower surfactant concentration is needed to form micelles and is therefore desirable for cleansing applications as a smaller amount of surfactant is required.

The mixing behavior of surfactants is influenced by factors such as the surfactant molecular structure, temperature, presence of electrolytes and the pH (Rosen and Kunjappu, 2012; Grady, 2023). Herein, we only consider the influence of the surfactant molecular structure and temperature. That is, no electrolytes are present at any surfactant solution and possible pH variations are not here considered. Binary mixtures of two nonionic surfactants have been shown to form ideal micelles, i.e., the two pure species are ideally mixed on the micelle (Holland and Rubingh, 1983), which is due to the lack of electrostatic repulsive forces between the head groups (Huang and Ren, 2017; Myers, 2020; Zhang et al., 2004). Nonionic/anionic and nonionic/cationic surfactant systems tend to mix nonideally, i.e., nonideal mixing in the micelle, and behave synergistically (Rosen and Hua, 1982; Kumar Shah et al., 2022; Grady, 2023; Zhang et al., 2004; Ren et al., 2014; Haque et al., 1996; Hines et al., 1997b; Ko et al., 2004). However, not all mixtures between nonionic and ionic surfactants exhibit synergism, as was shown by Alargova et al. (2001). The formation of micelles in anionic/cationic mixtures benefits from the reduction of the repulsive forces between the head groups and thus a synergistic behavior is observed (Grady, 2023; Phaodee and Sabatini, 2021). Anionic/anionic mixtures have been shown to exhibit antagonistic micellar behavior (López-Fontán et al., 1999), due to the large repulsive forces between the head groups in surfactant micelles. Zwitterionic surfactants, surfactants with two oppositely charged head groups, have a different charge depending on the pH of the solution, and hence mixtures containing them can exhibit multiple behaviors based on the solution conditions. Attractive interactions between anionic and zwitterionic surfactants have been reported in the literature (Hines et al., 1998, 1997a; McLachlan and Marangoni, 2006); similarly, slight deviation from ideal mixed micelles between cationic and zwitterionic surfactants is found (McLachlan and Marangoni, 2006). For more details on possible interactions between surfactant classes, we refer to reviews in Refs. Kumar Shah et al. (2022), Grady (2023), Phaodee and Sabatini (2021), Abe (2004). Overall, binary surfactant mixtures are complex systems and their behavior is not fully understood yet.

The importance of the CMC_{mix} for many applications, as they were described in the first paragraph of this work, has motivated the development of mathematical models over the years. In the following, we provide a brief review of the major models developed for predicting CMC_{mix}. For systems where micelle formation is ideal, such as nonionic/nonionic combinations, John H. Clint (Clint, 1975) proposed the following equation to calculate the CMC_{mix} of a binary mixture based on the mole fractions, x_1 and x_2, and the CMCs of the two surfactants as pure species:

$$\frac{1}{CMC_{\text{mix}}} = \frac{x_1}{CMC_1} + \frac{x_2}{CMC_2} \tag{1}$$

To account for nonideal micelle formation, the activity coefficients (γ_1, γ_2) of surfactants 1 and 2 are introduced, leading to the following equation:

$$\frac{1}{CMC_{\text{mix}}} = \frac{x_1}{\gamma_1\, CMC_1} + \frac{x_2}{\gamma_2\, CMC_2}, \tag{2}$$

Note the inverse relationship on the activity coefficients in contrast to the usual multiplication of activity coefficients and mole fractions. One way to calculate the activity coefficients was proposed by Rubingh based on the regular solution theory (Holland and Rubingh, 1983; Rubingh, 1979). Rubingh's model introduces the interaction parameter β as a way to account for the nonideal mixed micelle formation, which for a binary mixture can be estimated from the experimentally obtained CMC values of the pure species and one CMC_{mix} value at any mixture composition. Hence, β signifies the deviation from ideal mixing and synergistic/antagonistic mixing (Grady, 2023). Knowing β, allows us to calculate the activity coefficients by Holland and Rubingh (1983):

$$\gamma_1 = \exp(\beta \cdot x_2^2) \tag{3}$$

$$\gamma_2 = \exp(\beta \cdot x_1^2) \tag{4}$$

Following Rubingh's theory, β should remain constant for all mixture compositions. Rubingh's theory has been extensively applied to various surfactant systems due its simplicity, however model inconsistencies regarding the interaction parameter β are reported, that is, β did not remain constant for all mixture compositions (Kumar Shah et al., 2022; Grady, 2023; McLachlan and Marangoni, 2006; Sonu and Saha, 2013). For more insights on the two empirical equations, namely Eqs. (1) and (2), we refer to the original works (Clint, 1975; Rubingh, 1979; Holland and Rubingh, 1983). The estimation of the CMC_{mix} can also be accomplished using numerous molecular-thermodynamic frameworks that have been proposed in the literature (Shiloach and Blankschtein, 1997, 1998; Iyer and Blankschtein, 2014; Srinivasan and Blankschtein, 2005). However, these models rely on multiple analytical approximations, are computationally intensive, have been applied on small number of systems and are not applicable on surfactants with a complex structure (Iyer et al., 2013).

Machine learning (ML) models, specifically graph neural networks (GNNs), have shown very promising results in predicting pure species CMC for a wide variety of surfactants (Brozos et al., 2024b; Qin et al., 2021; Moriarty et al., 2023; Brozos et al., 2024a). GNNs treat every molecule as a graph, with atoms corresponding to nodes and bonds to edges. They learn to extract necessary structural information in an end-to-end framework and map it to the molecular property of interest. Recently, GNNs were expanded to binary mixtures for mixture properties such as activity coefficients and viscosity with very promising results (Rittig et al., 2023a,b; Sanchez Medina et al., 2022, 2023; Qin et al., 2023; Bilodeau et al., 2023; Vermeire and Green, 2021). Yet, surfactant mixtures have not been investigated. Here, we extend GNNs to CMC predictions of surfactant mixtures at different temperatures.

We propose GNNs for CMC prediction of surfactant mixtures between surfactants of all surfactant classes, namely anionics, cationics, nonionics and zwitterionics. For this, we collect CMC data for 108 binary surfactant mixtures from the literature at multiple temperatures. The assembled data set consists of 599 data points. We enrich the assembled data set by concatenating pure species data from our previous work (Brozos et al., 2024b). Hence, the final data set contains 1,924 CMC values for both pure species and binary surfactant mixtures at various temperatures. To treat surfactant mixtures, we consider two GNN architectures: (i) a composition-based weighted linear summation of the so-called molecular fingerprints (Gilmer et al., 2017) of each pure species and (ii) a mixture graph accounting for inter-molecular interactions and hydrogen bonding. Since the performance of ML models can vary significantly depending on the test split considered (Yang et al., 2019; Zahrt et al., 2020), we implement different splits/test scenarios for binary mixtures that are relevant for practical application of the model, e.g., predicting the CMC_{mix} at different compositions or with new surfactant structures. In summary, we evaluate the accuracy of both GNN models for:

- 4 different test scenarios that are based on potential practical applications of the model for binary surfactant mixtures collected from the literature
- 6 ternary surfactant mixtures collected from the literature
- 4 commercial surfactants, of which 2 are composed of two species (binary mixtures), 1 consists of three species (ternary mixture) and 1 consists of four species (quaternary mixture) for which the CMC is experimentally determined as part of work
- mixtures of one commercial surfactant and sodium dodecyl sulfate (SDS) for which we experimentally determine the CMC as part of work.

In all these four subcases, the GNN models are not retrained; note in particular that the training was done solely on binary and pure species data and the GNNs need to extrapolate to ternary and quartenary mixtures, which can thus be considered as out-of-distribution predictions. In our recent work, we showed that GNNs significantly outperformed a descriptor-based baseline model for temperature-dependent CMCs of single species (Brozos et al., 2024b). Since a large fraction of the data used in our recent work (Brozos et al., 2024b), was also used in this work, we decided against a baseline model for surfactant mixtures.

We structure the article as follows: in Section 2, we provide a detailed description of the assembled data sets and the train–test splits. Section 3 refers to the architecture of the two GNN models and to the experimental procedures, while Section 4 offers an analysis and discussion of the model results. We summarize our findings in Section 5.

The final model predictions for the 6 ternary mixtures and for the 2 out of 4 test sets for binary mixtures are provided along with the source code, as open source in our GitHub repository. The model hyperparameters and architecture are presented in Section 3. The training data set and the trained models remain the property of BASF and could be made available upon request.

2. Data set

We provide a general description of the data curated for binary and ternary mixtures (Sections 2.1 and 2.3), along with an analysis of the splitting procedure used (Section 2.2). The newly generated experimental data for binary, ternary, and quaternary surfactant mixtures are presented in a later part of this work (cf. Section 3.4).

2.1. Data set overview

We collect 108 binary surfactant mixtures at various temperatures from available literature sources (Zhang et al., 2004; Hines et al., 1997b; Martín et al., 2010; Hierrezuelo et al., 2006; Prasad et al., 2006; Moulik et al., 1996; Treiner and Makayssi, 1992; ud Din et al., 2009; Maiti et al., 2010). The data set contains 515 experimental mixture points, i.e., data points where at least two surfactants coexist in the aqueous solution ($x_1 \neq 0$ and $x_2 \neq 0$), from a total of 68 pure species structures. A statistical overview of the 515 mixture points is provided in the Supporting Information (SI), Figure S1. We visualize the 108 binary surfactant mixtures as a mixture network, where each node represents one of the 68 surfactants, and each edge the existence of a binary mixture between two surfactants in Fig. 1. For visualization of the structural similarity of the surfactants, we apply the t-distributed neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008) on generated extended-connectivity fingerprints (ECFP_10) for each pure species (Rogers and Hahn, 2010). As shown in Fig. 1, the 68 surfactants are distinguished in clusters based on their classes. Some exceptions are also observed, with the most noticeable being the three nonionic surfactants inside the blue circle. All three of them are n-alkyl-n-methylglucamides, with different alkyl chain lengths (Mega-8,-9 and -10) (Ko et al., 2004). All three of them contain a nitrogen atom, as well as a branched alcohol compared to the other nonionic surfactants. As is evident in Fig. 1, from the number of graph edges, some surfactants are well studied and are present in multiple mixtures, while others are present only in one mixture.

Based on the preceding discussion, one would expect the assembled data set to have a total size of 583, which will include the 515 mixture points and the 68 pure species data points. However, it is worth noting that some of the 68 pure species are present in mixtures at more than one temperature. For instance, SDS (a widely researched and used surfactant) is a component in mixtures at 4 different temperatures. Consequently, the herein assembled data set contains 4 distinct data points for pure SDS, i.e., the CMC at 4 different temperatures. The same holds true for other surfactants, resulting in a final data set size of 599 data points. Furthermore, to enrich molecular diversity and temperature-dependency information, we concatenate 1,377 CMC data points for pure species at various temperatures from our previous work to the binary mixture data set (Brozos et al., 2024b). In the combined data set, duplicate entries between the CMC values of the 68 newly collected surfactants as pure species and the 1,377 old data points arise, which are averaged. An example is given in the SI, Table S1. Overall, the final assembled data set consists of 1,924 data points. The minimum experimental temperature of our data is 0 °C and the maximum is 90 °C.

2.2. Data splits

We split our data set in different ways into training and test set to evaluate the model for different practical settings. As the choice of the test set can impact how the model performance is interpreted, leading to over-confident results (Zahrt et al., 2020), using different test scenarios is also beneficial to assess the robustness of the model. So we evaluate the performance of the model under different test scenarios. We implement 4 types of data set splitting: (i) composition interpolation (comp-inter), (ii) mixture compositions extrapolation (mix-comp-extra), (iii) mixture surfactant extrapolation (mix-surf-extra), (iv) mixture extrapolation (mix-extra) as described in the following paragraph. An overview of the 4 splits can be found in the SI, Table S2.

The *comp-inter* test set contains mixture points of previously seen binary mixtures but at different compositions. To select the mixture points, we identify all binary surfactant mixtures with at least two different mixture compositions. Out of the 108 binary mixtures present in our data set, 96 of them fulfill this criteria. For each of them, a mixture point is randomly selected, removed from the training set, and assigned to the test set. By utilizing the comp-inter test set, we can assess model's performance in predicting new mixture points of a binary mixture for which measured other mixture points are readily available.

The *mix-comp-extra* test set refers to binary surfactant mixtures, where the surfactants were seen during training in other surfactant mixture combinations. In other words, the training set includes the two surfactant structures of a binary mixture, either solely as pure species or as components of other mixtures. However, their combination (mixture) remains completely unseen. Two subsets, each containing 10 binary mixtures are randomly selected, consisting of 46 and 58 mixture points, respectively. Similarly to the comp-inter test set, the mix-comp-extra test set only contains mixture points. We can thereby evaluate whether the model is able to predict binary surfactant combinations for which experimental mixture data is not available yet.

The *mix-surf-extra* test set expands the extrapolation character of the mix-comp-extra split by completely excluding one of the two surfactants of a binary mixture from the training data set. That is, one surfactant is not included in the training set, either as pure species or as a component of other mixtures, whereas the other surfactant is included, either as pure species or in other mixtures. This test scenario reflects the isolation/synthesis of a new surfactant structure, for which no previous measurements are available. To construct the mix-surf-extra test set, we select 4 representative surfactants, namely: n-decanoyl-n-methylglucamide (Mega-10), n-dodecyl-β-D-maltoside (β-$C_{12}G_2$), cetylpyridinium chloride (CPC) and SDS. Removing them from the training set simultaneously would result in a huge information loss, accounting to about 30 percent of the mixture points, and thus diminish model performance. Therefore, we decided to remove each of them (and corresponding mixtures) separately. Accordingly, we constructed 4 sub-test sets. The model performance on the mix-surf-extra test set is reported on the combination of the 4 sub-test sets (cf. Section 4). For example, for Mega-10 there are 9 binary mixtures (45 mixture points) in the assembled data set at 30 °C (Martín et al., 2010; Hierrezuelo et al., 2006). The sub-test set will consist both the 45 mixture points and the Mega-10 as pure species at 30 °C. The data points for pure

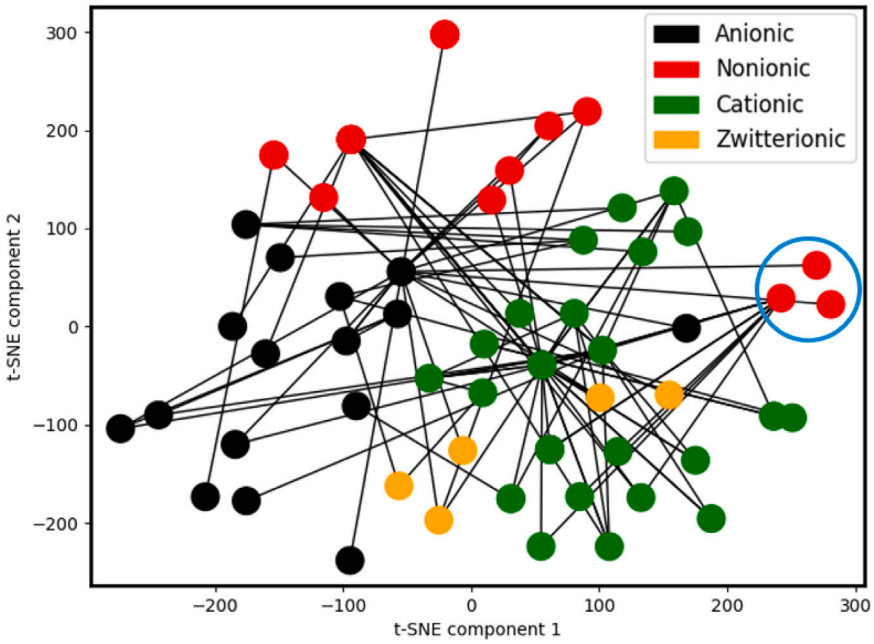


Fig. 1. Mixture network of the curated data set. Each node represents a surfactant structure, and each edge the existence of a binary mixture between two surfactants. Correspondingly, the surfactants are also categorized based on their class. Each node is plotted on a 2D map obtained by applying t-SNE (van der Maaten and Hinton, 2008) on generated molecular fingerprints (ECFP_10) (Rogers and Hahn, 2010). The surfactants are well distinguished in clusters based on their classes. An exception are the three n-alkyl-n-methylglucamides surfactants (Mega-8,-9 and -10) enclosed in the blue circle.

Mega-10 at temperatures different than 30 °C (Prasad et al., 2006) are excluded from both the training and the test set, since we already demonstrated GNN's ability to predict the temperature-dependent CMC of pure surfactants in our previous work (Brozos et al., 2024b) and we herein focus on surfactant mixtures. Similarly, the other sub-test sets contains 3 binary mixtures (14 mixture points) at 25 °C for β-$C_{12}G_2$ (Zhang et al., 2004; Hines et al., 1997b), 6 binary mixtures (19 mixture points) at 25 and 30 °C for CPC (Moulik et al., 1996; Treiner and Makayssi, 1992; ud Din et al., 2009; Maiti et al., 2010), and 15 binary mixtures (82 mixture points) at 22, 25, 30, 35 and 40 °C for SDS (Haque et al., 1996; Hines et al., 1997b, 1998; Maiti et al., 2010; López-Fontán et al., 2000). We can thereby evaluate whether the model is able to predict binary mixture combinations of surfactants for which one of them has not been seen before during model training.

The mix-extra test set considers a scenario where none of the surfactants in a binary mixture are encountered during training, either as pure species or as components of other binary mixtures. In this scenario, the model has to predict CMCs of new/unseen surfactant structures, as well as binary mixtures of them. First, we identify 5 binary mixtures composed by 6 surfactants (2 anionic and 4 nonionic) that fulfill the criteria described above (Huang and Ren, 2017; Ren et al., 2014; Haque et al., 1996). To enhance structural complexity and variety of the mix-extra test set, we further assign 2 binary mixtures composed by 3 cationic surfactants (Treiner and Makayssi, 1992). However, one of the 3 cationic surfactants (cetalkonium chloride) is present in a mixture with CPC (Treiner and Makayssi, 1992), which as described above, is widely present in the data set. We discard this mixture only for the mix-extra split. Thus, the training set contains 100 binary mixtures (and the data points of pure species) and the test set contains 7 binary mixtures. Hence, we can evaluate whether the model is able to predict binary mixtures of which both surfactants have never seen by the model before.

2.3. Ternary mixtures from literature sources

We further assemble a small external data set containing 6 ternary mixtures from literature sources (Moulik et al., 2021; Huang and Ren, 2019). The ternary data set contains 16 mixtures points at 25 and 30 °C, composed from 8 pure species structures. All 8 surfactant structures are present in the collected data set described in Section 2.1. Here, we investigate whether the model can effectively generalize to ternary mixtures containing species for which CMC data are readily available as pure species and in binary mixtures. Note that the ternary mixtures are not used for training. Additionally, we underscore the lack of CMC data for ternary surfactant mixtures in literature, likely due to the combinatorial increase of required measurements and resources incurred to do such measurements.

3. Methods

We begin this section by describing the composition-based weighted linear summation GNN architecture (Section 3.1). We then introduce the mixture graph framework; the second GNN architecture (Section 3.2) and describe the hyperparameters and model implementation (Section 3.3). Lastly, we provide a description of the experimental material and procedures (Section 3.4).

3.1. Weighted sum GNNs for surfactant mixtures

We develop a composition-based weighted linear GNN model for predicting the temperature-dependent CMC of surfactant mixtures. The concept of GNN models has been described in detail in several literature sources, see Refs. Yang et al. (2019), Zhou et al. (2020), Wu et al. (2021), Reiser et al. (2022), Khemani et al. (2024), Hamilton et al. (2017), Schweidtmann et al. (2020), Rittig et al. (2023c). We build on our GNN model for predicting the temperature-dependent CMC of pure species from our previous work (Brozos et al., 2024b). Here, we extend the architecture for mixtures. Each surfactant molecule in the mixture is represented as a graph $G_i = (V, E)$, where V are the vertices (atoms), E are the edges (bonds) between nodes, $i \in \mathcal{N}$ is a component of the surfactant mixture and $\mathcal{N}$ is the total number of mixture components. The GNN first encodes structural information from pure species through graph convolutions and a pooling step into a molecular fingerprint (FP), which is denoted as FP_i. To treat mixtures, we propose the weighted

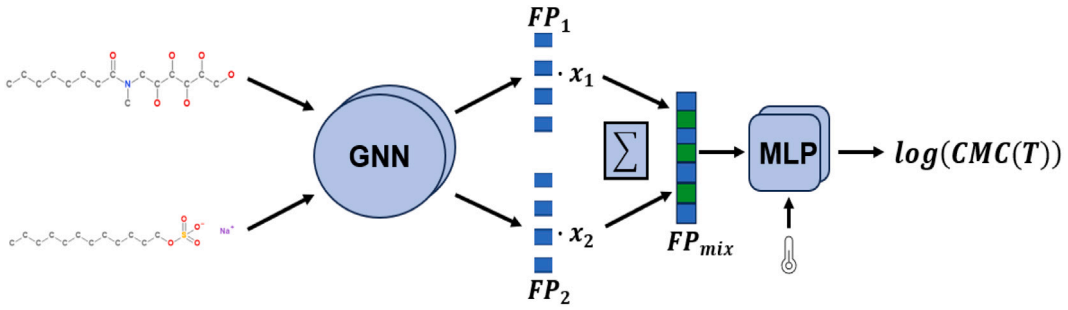


Fig. 2. Schematic representation of the WS-GNN architecture for a binary mixture.

linear summation of the learned FPs, i.e., the mole fraction x_i is used as weight. Hence, the mixture fingerprint FP_{mix} is calculated through Eq. (5). This model architecture enables also considering pure species, i.e., $x_2 = 0$ in the case of binary mixtures. Furthermore, since weighted summation is a permutation invariant operation, the architecture ensures permutation invariance with respect to the order of representation of the surfactants within a mixture: if the order of the surfactants is changed, FP_{mix} remains the same. We denote this architecture as **WS-GNN** (weighted sum GNN) and a schematic representation for an exemplary binary mixture is given in Fig. 2. The choice of WS-GNN is motivated by recent successes in predicting mixture properties with similar architectures, that is, a weighted summation of individual molecular fingerprints (Leenhouts et al., 2025; Zhang et al., 2024). Further promising alternatives, e.g., using the attention for combining molecular into mixture representations as recently proposed by Zhang et al. (2024), could also be explored in future work. Note that Eq. (5) does not explicitly encode the inverse relationship between CMC_{mix} and CMC_{i} described in Eqs. (1), (2). The reason for this is to allow greater flexibility during training, and since Eqs. (1), (2) are only semi-empirical, they may not always hold true, as many authors have found (Huang and Ren, 2017; Zhang et al., 2004; Misselyn-Bauduin et al., 2000).

$$FP_{\text{mix}} = \sum_{i \in N} x_i \cdot FP_i \tag{5}$$

The FP_{mix} is mapped to the temperature-dependent CMC through a standard multi-layer perceptron (MLP). The non-linearity introduced in the MLP allows the non-linear mixing behavior of surfactants to be captured. In other words, the nonideal mixing behavior of surfactants can be captured by a GNN model through the non-linearity of the activation functions in the network. The temperature is concatenated to the first hidden layer, similar to our previous work (Brozos et al., 2024b), where more detailed information regarding the MLP architecture and temperature dependency can be found.

3.2. GNNs with mixture graph

Alternatively, to capture molecular interactions in mixtures, more advanced geometries that consider hydrogen bonding information have recently been proposed (Sanchez Medina et al., 2023; Qin et al., 2023). Here, the construction of a new graph is proposed, where nodes represent the pure species of each mixture and edges inter- and intramolecular interactions (Sanchez Medina et al., 2023; Qin et al., 2023). The pure species fingerprints are used as node feature vectors and number of hydrogen acceptors and donors as edge features (Sanchez Medina et al., 2023; Qin et al., 2023). Afterwards, the mixture graph is passed into a GNN layer to account for intermolecular interactions (Sanchez Medina et al., 2023; Qin et al., 2023). We note that Van der Waals forces are not considered in this architecture. The architecture which considered hydrogen bonding information, outperformed alternative architectures in predicting activity coefficients of binary and ternary mixtures in the original paper of Qin et al. (2023).

We adapt the proposed mixture graph architectures by first multiplying the composition x_i of surfactant i by the corresponding FP_i. The composition-adjusted FP_i are utilized as node feature vectors of the mixture graph, as proposed in previous works (Sanchez Medina et al., 2023; Qin et al., 2023). For edge features, we also consider the number of hydrogen bond acceptors and donors, but we calculate them through Lipinski's rule of five (Lipinski et al., 2001) compared to the previous works, as we observed a slightly higher model accuracy. We did not apply a weighting factor to the edges. The mixture graph is then passed into a graph convolutional layer to account for surfactant interactions (Sanchez Medina et al., 2023; Qin et al., 2023). We use the GINE-operator (Xu et al., 2018; Hu et al., 2020) as a graph convolutional layer. To extract the FP_{mix}, a summation pooling layer is added, as was also proposed in our recent work (J.G. and Mitsos, 2024). Analogously to the WS-GNN, the pooling layer ensures permutation invariance for the surfactant order within the mixture. We denote this architecture as **MG-GNN** (mixture graph GNN) and illustrate it for an exemplary binary mixture in Fig. 3.

3.3. Implementation, hyperparameter tuning and ensemble learning

All models are implemented in PyTorch Geometric (PyG) (Fey and Lenssen, 2019). We perform hyperparameter tuning based on the WS-GNN architecture (cf. Section 3.1). Specifically, we train the GNN model on 20 different, randomly selected validation sets, similar to our previous works (Brozos et al., 2024b,a; Rittig et al., 2023a). The size of the validation set is kept constant at 385 points, which represents 20% of the whole data set size. We use the root mean square error (RMSE) on the comp-inter split for hyperparameter tuning. The hyperparameters investigated during hyperparameter tuning are provided in the SI, Table S3.

To improve the predictive capabilities of ML models, ensemble learning is a widely employed. The models trained on different seeded validation sets may produce noisy results. By averaging out the predictions of all them, robust and generalized predictions are obtained (Breiman, 1996; Dietterich, 2000; Ganaie et al., 2022). We average the predictions of the 20 different trained models for each split type mentioned in Section 2.2, and report only the prediction accuracy of the ensemble of GNNs (Brozos et al., 2024b; Rittig et al., 2023a). We further apply ensemble learning when combining the two developed GNN models, described in Sections 3.1 and 3.2. That is, the predictions from the two proposed GNNs architectures (WS-GNN and MG-GNN) are averaged.

3.4. Materials and surface tension measurements

As industrial grade surfactants are usually composed from more than one species, we aim to identify such examples and validate the accuracy of our model. The commercial surfactant Dehyton® AB 30 (D-AB30) is provided by BASF and was recently used in a commercially formulation study (Cao et al., 2021). A gas chromatography (GS) analysis is also provided from the supplier. The main two species (≥ 98%)

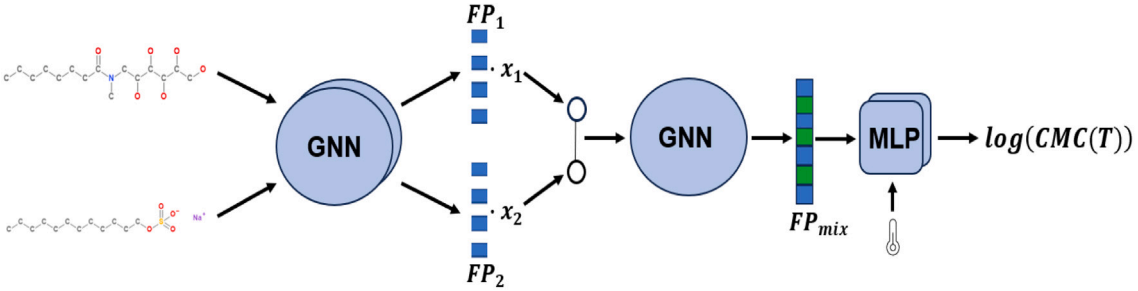


Fig. 3. Schematic representation of the MG-GNN architecture for binary mixtures.

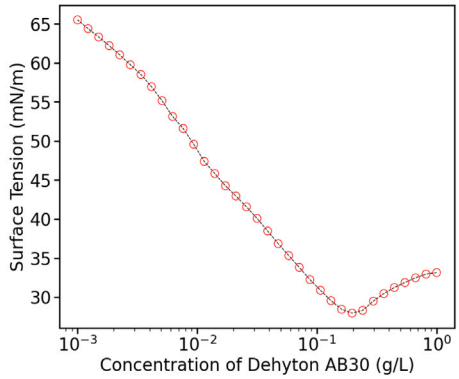


Fig. 4. Surface tension measurement of D-AB30 at 23 °C.

are lauryl and myristyl betaine. Therefore, the product can be well categorized as a binary surfactant mixture between two zwitterionic surfactants. We note that both pure species exist in our collected data set (cf. Section 2.1) but no mixture between two zwitterionic surfactants is available. Furthermore, highly purified unary SDS (≥ 99%) was bought from Sigma Aldrich and used as received. Mixtures between SDS and D-AB30 at different mole fractions were prepared and measured. Since D-AB30 is a binary mixture, the prepared mixtures are considered ternary ones. Three more commercial surfactants, namely Sulfopon® 1214 G (≥ 96%), Texapon® V 95 G (≥ 96%) and Texapon® K 30 UP (≥ 98%) were provided by BASF, together with a GC analysis. The Sulfopon® 1214 G (S-1214G) is composed from two species, namely SDS and sodium tetradecyl sulfate (STS) while the Texapon® V 95 G (T-V95G) consists three species, namely SDS, STS and sodium hexadecyl sulfate. We note, that no mixture data between these species exist in our collected data sets (cf. Sections 2.1 and 2.3). The Texapon® K 30 UP (T-K30UP) consists four species, namely SDS, STS, sodium hexadecyl sulfate and sodium octadecyl sulfate.

Surface tension measurements were performed at 23 °C using the Force Tensiometer – K100 (Krüss, Germany) and the Wilhelmy plate method with a standardized platinum-iridium plate. The surfactant solutions were freshly prepared in deionized water and adjusted to pH 4.9 prior to the measurements. The surface tension was plotted against the logarithm of the surfactant concentration to determine the CMC. 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. To ensure reproducibility, the measurements were repeated three times. The surface tension measurement of D-AB30 is illustrated in Fig. 4.

4. Results and discussion

We begin this section by discussing the results of the two GNN models on the 4 different test scenarios for binary mixtures (Section 4.1). Afterwards, we analyze the model performance on different surfactant class combinations (Section 4.2) and the applicability of the GNN model to ternary mixtures (Section 4.3). We then compare model predictions with experimental values for commercial surfactants and mixtures of them (Section 4.4). We conclude this section by comparing the GNN models with a semi-empirical model (Section 4.5).

4.1. Predictive performance on binary surfactant mixtures

In Table 1, an overview of the prediction accuracy of the two developed GNN architectures is presented for the 4 different test sets (cf. Section 2.2). We report 4 different error metrics for each test set. We scale the CMC (μM) values using a (based on 10) logarithmic scale and hence all error metrics refer to the log(CMC), which are scaled in μM throughout the paper. Note that in Table 3 the CMC is also reported in mM, as previous literature merely reports experimentally measured CMC values in mM. From Table 1, we observe that the best performing architecture varies for each split, i.e., no clear best-performing GNN model can be identified. Therefore, we consider averaging their predictions (cf. Section 3.3) and denote the model as **combined**. Another motivation is that the combined framework leverages learning from both architectures and offers a more flexible and robust framework. The combined results are reported in Table 1 and parity plots on the 4 test scenarios are illustrated in Fig. 5.

In the comp-inter split, the combined model outperforms the best performing MG-GNN model, as indicated by the slight decrease in RMSE from 0.251 to 0.249 (cf. Table 1). In the mix-comp-extra split, it yields an RMSE of 0.313, which is very close to the lowest RMSE (0.302) achieved by the WS-GNN model. In both test scenarios, the RMSE is similar to that reported from previous accurate GNN models applied to pure species (Brozos et al., 2024b; Qin et al., 2021; Moriarty et al., 2023; Brozos et al., 2024a). In the parity plots of the two test sets (Figs. 5a,b), the majority of the points lie very close to the diagonal, as can also be seen from the high R^2 values, 0.93 and 0.88 respectively. Therefore, the model can be used to provide highly accurate predictions for these two test scenarios.

In the mix-surf-extra split, the combined model yields an RMSE of 0.551, which is also close to the RMSE (0.507) of the WS-GNN model. Exclusion of the pure species of a binary mixture from the training set, drastically reduces the model performance, as a significantly higher error is exhibited in the mix-surf-extra test set. However, most of the predictions in Fig. 5c also lie close to the diagonal. Here, the noticeable outliers significantly increase the RMSE. Furthermore, the combined model exhibits an RMSE of 0.344 on the mix-extra test, very similar to the lowest RMSE (0.333) achieved by the MG-GNN model. The predictions lie closer to the diagonal as in the mix-surf-extra test set, which is illustrated from the higher R^2 value, 0.69 compared to 0.57. The results of extrapolation to new surfactant structures and binary mixtures indicate that model performance depends on the new, unseen structures introduced to the model. In essence, a high fraction of the predictions are very accurate (indicated from the parity plots), but outliers with high deviation are more common than in the comp-inter

Table 1
Summary of the prediction accuracy of the ensemble GNNs models, i.e., WS-GNN and MG-GNN architectures, as well as of the combination of the two GNN architectures on the 4 test scenarios. MAE = mean absolute error, MAPE = mean absolute percentage error (unit %).

Model		comp-inter	mix-comp-extra	mix-surf-extra	mix-extra
WS-GNN		0.264	**0.302**	**0.507**	0.441
MG-GNN	RMSE	0.251	0.334	0.665	**0.333**
Combined		**0.249**	0.313	0.551	0.344
WS-GNN		0.183	**0.182**	**0.382**	0.366
MG-GNN	MAE	0.178	0.215	0.486	**0.215**
Combined		**0.171**	0.196	0.406	0.274
WS-GNN		6.831	**7.352**	**15.916**	14.758
MG-GNN	MAPE	6.344	8.452	20.336	**8.2**
Combined		**6.292**	7.827	17.334	10.789
WS-GNN		0.93	**0.89**	**0.62**	0.54
MG-GNN	R^2	0.93	0.86	0.43	0.64
Combined		**0.93**	0.88	0.57	**0.69**

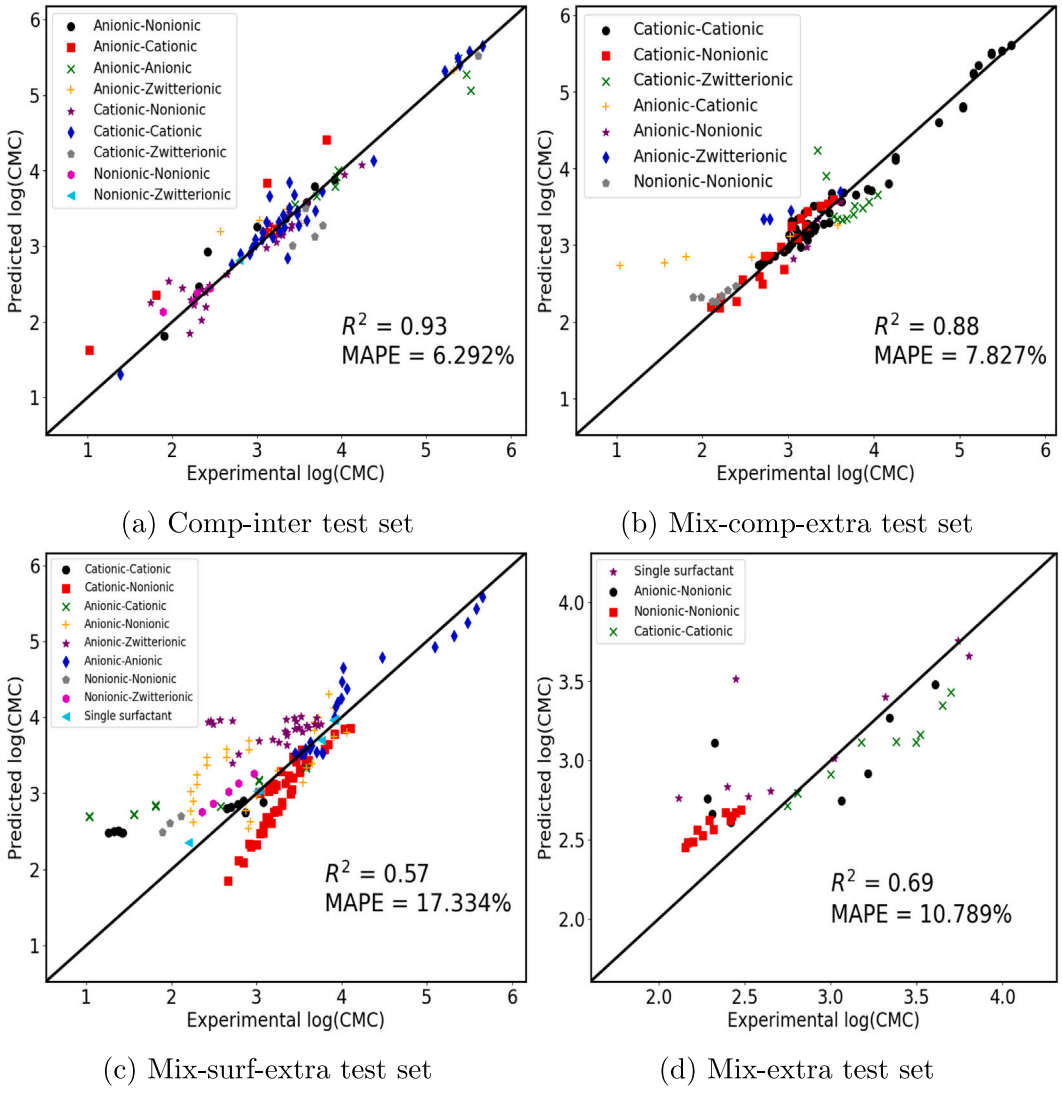


Fig. 5. Parity plots on the 4 test sets. The predictions are made by the combined GNN model. The data points are highlighted with different colors and markers based on the classes of the two mixture species. The logarithm is applied to CMC in μM (base 10).

and mix-comp-extra splits. Therefore, the model should be used with increased awareness in such extrapolation scenarios.

Excluding the comp-inter split, the combined model did not yield the lowest RMSE compared to the WS-GNN and MG-GNN models. We explain this performance by the larger error exhibited from one out of the two models in each of the three splits. To elaborate, in the comp-inter split the relative difference between the two RMSEs is about 5.18%, while in the other three splits it is between 10.6% and 32.43%. Notably, in these three splits, the RMSE of the combined model is significantly closer to the lowest RMSE, rather than the highest. Thus, our findings suggest that a GNN framework that combines multiple architectures can improve performance in test sets, where one architecture may underperform.

In addition, we investigate the mixtures with the highest deviations for the mix-comp-extra and mix-surf-extra test sets, as they are the two test sets with the highest number of major outliers. In the case of the mix-comp-extra test set, the mixture with the highest error is between two ionic surfactants, namely an anionic/cationic mixture (Maiti et al., 2010), symbolized by a yellow cross in Fig. 5b. We observe that the CMC_{mix} at x_{SDS} equal to 0.33, 0.5 and 0.67 reduces by up to an order of magnitude compared to the CMC_{mix} a x_{SDS} equal to 0.14, 0.8 and 0.85. We do not observe a similar reduction to the CMC_{mix} for any of the other mixtures present in our data set, and therefore the model cannot accurately capture it, and high model deviations are observed. In the mix-surf-extra test set, a similarly high prediction error is observed for the same anionic/cationic mixture (Maiti et al., 2010). Furthermore, as illustrated by the black dots in Fig. 5c, a cationic/cationic mixture also appears as a major outlier too (ud Din et al., 2009). After examining of similar mixtures, namely between a gemini cationic and a conventional cationic surfactant, in the training data set, we observe that this mixture displays a high synergistic behavior, namely $\beta \leq -5.99$, while all other similar mixtures show a $\beta \geq -2.2$ (Sonu and Saha, 2013; Rodríguez et al., 2008; Alargova et al., 2001). We hypothesize that the lack of cationic (gemini)/cationic mixtures with such high synergistic behavior from the training set is the underlying reason for these large outliers. Finally, an anionic/zwitterionic mixture also appears as a major outlier in the mix-surf-extra test set (Hines et al., 1997b). We note that since all anionic/zwitterionic mixtures consist of SDS as the anionic surfactant (Hines et al., 1997a; Bakshi et al., 1993), no anionic/zwitterionic mixtures were present in the training data for the mix-surf-extra sub-test set (cf. Section 2.2), which may have led to the high model error observed.

4.2. Predictive performance per surfactant mixtures classes

In this subsection, we provide an analysis of the model results per surfactant class combination. We categorize the binary surfactant mixtures based on the classes of the two species and we identify descriptive examples in each test set split, where model predictions are in agreement with experimental measurements, as well as examples where model refinement/further extension is required. The predictions are made with the combined GNN model. All the comparisons refer again to the logarithmic scale. For the interested reader, selective examples with absolute CMC values are illustrated in the SI, Figure S2.

4.2.1. Mixtures with nonionic surfactants

Nonionic surfactants typically show ideal micelle formation when mixed with other nonionic surfactants and synergistic behavior when mixed with ionic surfactants (cf. Section 1). A nonionic/nonionic mixture between octylphenol polyoxyethylene ether (OP-10) and OP-4 at 25 °C is presented in Figs. 6a,b (Huang and Ren, 2017) in two different test scenarios. In the case of mix-comp-extra test set, the combined GNN model accurately predicts the trend and the CMC_{mix} in all mole fractions. Higher errors are observed for the mix-extra test set. The model significantly overestimates the CMC of pure OP-4, thus leading to overestimated CMC_{mix} predictions in lower OP-10 mole fractions. Fig. 6c refers to the mixture of dimethylene-1,2-bis(dodecyldimethylammonium bromide) (12-2-12), a dimeric cationic surfactant, and $C_{12}E_8$ (Alargova et al., 2001) and Fig. 6d represents the mixture of cetyltrimethylammonium bromide (CTAB), a cationic surfactant, and Mega-10 (Hierrezuelo et al., 2006). Hence, both cases belong to cationic/ nonionic combinations. In both cases, highly accurate predictions are demonstrated throughout all mole fractions. To summarize, in all 4 examples the trend is well captured except for mixture (b) at composition $x_1 = 0$.

4.2.2. Ionic - ionic mixtures

Mixtures between ionic surfactants can exist in three different combinations, namely anionic/cationic (opposite charge of the head group), anionic/anionic and cationic/cationic (similar head group charge). Here, we investigate all three of them. In Fig. 7 one exemplary mixture for each combination is presented. Fig. 7a refers to a mixture between SDS and CPC at 25 °C (Maiti et al., 2010), while Fig. 7b to a mixture between sodium hexyl sulfate (SHS) and SDS at 35 °C (López-Fontán et al., 2000). Both mixtures are extracted from the mix-surf-extra test set. Fig. 7c refers to a mixture between tetradecyltrimethylammonium chloride (TTAC) and benzyldimethyltetradecylammonium chloride at 25 °C present in the mix-extra test set (Treiner and Makayssi, 1992). The mixture containing two anionic species shows antagonistic behavior, which is accurately captured by the model.

However, the model's sensitivity to mole fraction variances could be further improved. It is important to note that due to the removal of SDS from the training set, only three mixtures between anionic surfactants were present in the training data. Furthermore, in the anionic/cationic mixture, the strong synergism at mole fraction 0.5 is not fully captured, the combined GNN model accurately predicts the existence and point of synergism as well as the order of magnitude of CMC_{mix} at other mixture compositions. The synergistic behavior between two cationic surfactants is accurately captured by the combined GNN model, as evident in Fig. 7c. The predicted CMC_{mix} values at all mole fractions match the measured ones.

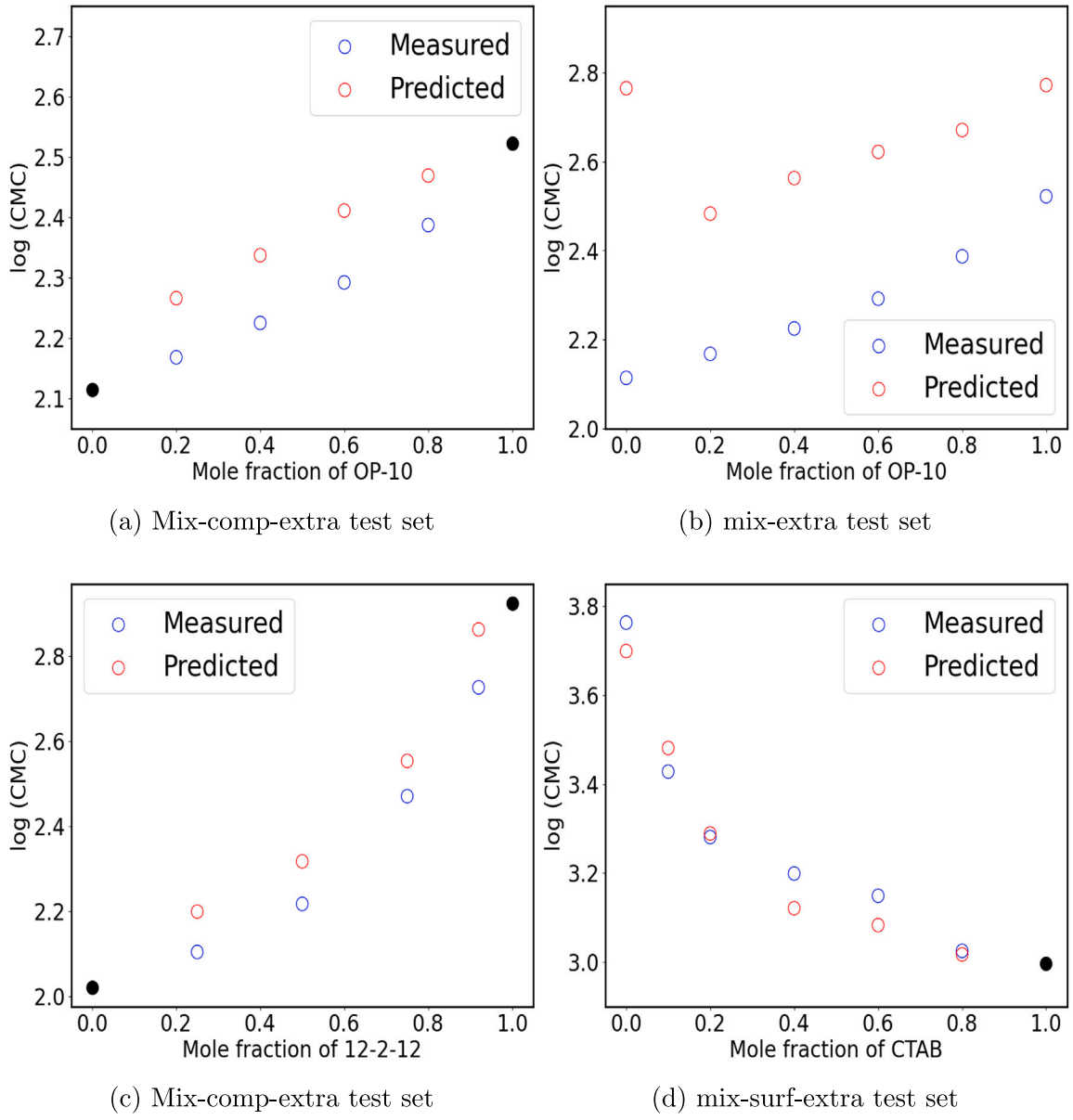
4.2.3. Mixtures with zwitterionics

Binary mixtures, where at least one surfactant is zwitterionic, can exhibit nonideal mixed micelles and are highly pH sensitive, as the properties of zwitterionic surfactants can vary significantly at different pHs (Rosen and Kunjappu, 2012). However, research on them has remained limited and hence only minor examples are present on our test sets. An anionic/zwitterionic mixture between SDS and lauramine oxide (LDAO) at 25 °C is shown in Fig. 8a (Bakshi et al., 1993), while a cationic/zwitterionic mixture between dodecyltrimethylammonium bromide (DTAB) and sulfobeteine-12 (S-12) is presented in Fig. 8b (McLachlan and Marangoni, 2006).

Both mixtures are present on the mix-comp-extra test set. A further example, namely a mixture between S-12 and $C_{12}G_2$ at 25 °C is provided in Fig. 8c (Hines et al., 1997b). For the first mixture, the model fails to capture the synergism of the system and significant deviations from the measured CMC_{mix} are also observed. Similar behavior is observed at the mixture between DTAB and S-12, but the model exhibits accurate sensitivity regarding the mole fraction variations. For the mixture of S-12 and $C_{12}G_2$ good agreement between predicted CMCs and measured are observed. All over the combined model is able to capture the trend and provide good predictions.

4.3. CMC predictions on ternary mixtures from literature sources

We adapt both GNN architectures to predict ternary mixtures described in Section 2.3, while training them exclusively on binary mixtures and pure species data. Here, we are interested in whether GNN models trained on CMC values of binary mixtures and pure species can accurately scale up to ternary mixtures with no further information. We use the trained WS-GNN and MG-GNN models from the comp-inter split to perform predictions on the 6 ternary mixtures described in Section 2.3. The results are given in Table 2. The WS-GNN model performs fairly good with and RMSE of 0.165 while the MG-GNN model exhibits a very high RMSE of 1.824. Since the MG-GNN model exhibits such a high error, combining the architectures would not provide any benefits and therefore we do not report any combined results. Hence, the simple WS-GNN is able to perform above average predictions on surfactant ternary mixtures even trained exclusively on binary and pure species data. We hypothesize that the weighted summation on the WS-GNN model, ensures that the mixture fingerprint always remains in the



(a) Mix-comp-extra test set

(b) mix-extra test set

(c) Mix-comp-extra test set

(d) mix-surf-extra test set

Fig. 6. GNN predictions versus experimental data on different surfactant mixtures classes are present in three test scenarios. Panel (a) and (b) refer to a mixture between nonionic OP-10 and OP-4 (Huang and Ren, 2017). Panels (c) and (d) refer to mixtures between cationic and nonionic surfactants, namely 12-2-12/$C_{12}E_8$ and CTAB/Mega-10 (Alargova et al., 2001; Hierrezuelo et al., 2006). Black filled circles represent data points present in the training set.

same order of magnitude. On the other side, in the MG-GNN model, a third node is introduced in the mixture graph. After the summation pooling step, the order of magnitude of the mixture fingerprint does not remain similar to that of binary mixtures. Here, with no data on ternary mixtures during training the MG-GNN model completely fails to give any reasonable predictions. Replacing the summation with a mean pooling step led only to slight performance improvements. Thus, only the WS-GNN model could be used when dealing with ternary mixtures. We provide a parity plot with the predictions from the WS-GNN model in the SI, Figure S3.

4.4. Experimental validation and commercial surfactants

The experimentally measured mixtures comprise, binary, ternary and quaternary mixtures. The combined GNN model is utilized for binary mixtures, while the WS-GNN is used for ternary and quaternary mixtures, in accordance with the discussion above. The results are summarized in Table 3. In the case of commercial surfactants that are composed of two species, the combined GNN model demonstrates high accuracy, especially for the D-AB30 with an absolute error (AE) of 0.022. To better understand why the performance of the model on S-1214G, i.e., AE of 0.262, does not match the one of D-AB30, we investigate their impurities. In the case of D-AB30 only two further species exist in small concentrations. In contrast, there exist more than five further species in the S-1214G.

Hence, we hypothesize the non treatment of impurities from our model as a reason for this deviation. For the ternary mixtures, we observe a MAE of 0.178 which is slightly higher than the one found in

Table 2
Error metrics of WS-GNN and MG-GNN models on the ternary mixtures data set. MAPE is in unit of %.

Model	RMSE	MAE	MAPE
WS-GNN	**0.165**	**0.13**	**5.4**
MG-GNN	1.824	1.687	74.805

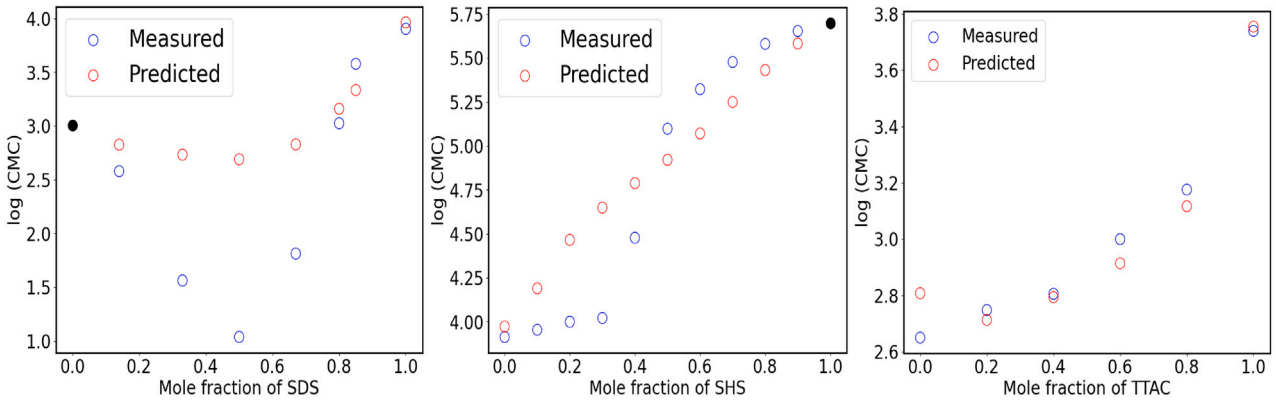


Fig. 7. GNN predictions versus experimental data on three different surfactant mixture combinations present on the mix-surf-extra and mix-extra test sets. Panel (a) refers to a mixture between SDS and CPC (Maiti et al., 2010), panel (b) to a mixtures between SHS and SDS (López-Fontán et al., 2000) and panel (c) between TTAC and benzyldimethyltetradecylammonium chloride (Treiner and Makayssi, 1992). Black filled circles represent data points present in the training set.

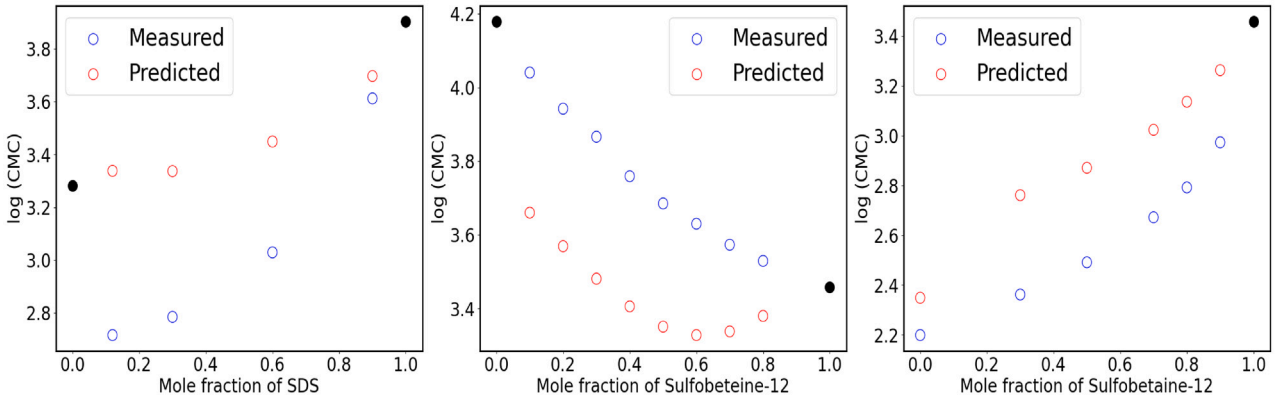


Fig. 8. GNN predictions versus experimental data on mixtures with one zwitterionic species. Panel (a) refers to a mixture between SDS and LDAO (Bakshi et al., 1993), panel (b) S-12 and DTAB (McLachlan and Marangoni, 2006) and panel (c) between S-12 and $C_{12}G_2$ (Hines et al., 1997b). Black filled circles represent data points present in the training set.

Table 3
Model predictions for *log*(CMC) versus the experimental data, measured as part of this work. On parentheses the absolute CMC values in mM are provided.

Sample	Predicted *log*(CMC)			Measured *log*(CMC)
	WS-GNN	MG-GNN	Combined	
vv D-AB30	2.9 (0.79)	2.83 (0.68)	2.86 (0.73)	2.84 (0.7)
S-1214G	3.7 (5.02)	3.59 (3.92)	3.65 (4.47)	3.39 (3.53)
SDS (0.4)/D-AB30 (0.6)	2.86 (0.72)	–	–	2.75 (0.56)
SDS (0.6)/D-AB30 (0.4)	3.18 (1.51)	–	–	2.91 (0.8)
SDS (0.2)/D-AB30 (0.8)	2.65 (0.45)	–	–	2.55 (0.35)
T-V95G	3.68 (4.79)	–	–	3.90 (3.59)
T-K30UP	3.44 (2.74)	–	–	3.3 (1.98)

Section 4.3, namely 0.13. We note that the mixture between SDS and D-AB30 at a 0.4–0.6 composition, exhibits the highest error and when is not taken into account, the MAE reduces to 0.147. For the T-K30UP the model predictions matches the experimental value, although no quaternary mixtures were considered during training. The high purity of T-K30UP positively contributes on the model predictions. Overall, our developed GNN models can be effectively applied on commercial surfactants that are composed up to four species and help guide research and development. Accounting for the impurities is critical for further model refinement and should be addressed in a future work.

4.5. Comparison to semi-empirical model

We further compare the combined GNN model to the semi-empirical model described by Eq. (2), as we also did in a recent work of ours (Nevolianis et al., 2024). We calculate the activity coefficients with HANNA, a hard-constraint neural network (HANNA) that ensures thermodynamic consistency that was trained on the Dortmund Data Bank, one of the largest collections of experimental activity coefficients, and was recently proposed by Specht et al. (2024). To calculate the CMCs of the two mixture species, we consider our recently developed GNN model for temperature-dependent CMCs of pure species (Brozos

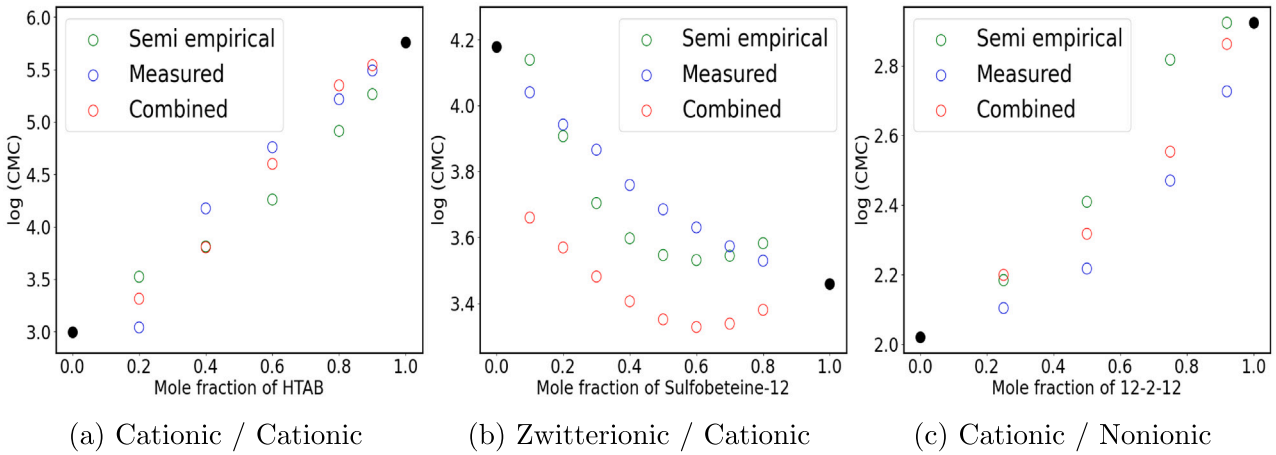


Fig. 9. Comparison between hybrid model predictions, GNN predictions and experimental data versus on binary surfactant mixtures from the mix-comp-extra test set. Panel (a) refers to a mixture between n-hexyltrimethylammonium bromide (HTAB) and CTAB (López-Fontán et al., 1999), panel (b) S-12 and DTAB (McLachlan and Marangoni, 2006) and panel (c) between 12-2-12 and $C_{12}E_8$ (Alargova et al., 2001). Black filled circles represent data points present in the training set.

Table 4
Comparison between the prediction accuracy of the combined GNN model versus the semi-empirical model on 2 test scenarios. MAE = mean absolute error, MAPE = mean absolute percentage error (unit %).

Model		mix-comp-extra	mix-extra
Combined	RMSE	**0.313**	**0.344**
Semi-empirical		0.568	0.783
Combined	MAE	**0.196**	**0.274**
Semi-empirical		0.415	0.621
Combined	MAPE	**7.827**	**10.789**
Semi-empirical		15.204	21.995

et al., 2024b). The results of the semi-empirical model on the logarithmic scale are given in Table 4 and are compared with the combined model described in Section 4.1. Furthermore, the predictions of both models, namely semi-empirical and combined, for three exemplary binary mixtures are graphically represented alongside the corresponding experimental measurements in Fig. 9.

The results show that the combined model outperforms the semi-empirical model in both of the test scenarios, with a reduced RMSE of about half of the semi-empirical model. Investigating the three mixtures of Fig. 9, we observe that in all cases except for mixture (b), the combined model outperforms the semi-empirical one. Yet, the semi-empirical model accurately captures the variations of the log(CMC) with the mole fraction, and might be used only for a first estimate of the CMC_{mix} given the significant higher accuracy of the combined GNN model trained on the mixture CMC data directly.

5. Conclusion

We present a GNN-based framework for the prediction of temperature-dependent CMCs of surfactant mixtures. We collect data from literature sources for 108 binary mixtures, to which we concatenate data for pure species from our previous work. We develop and test two GNN architectures. In the first architecture, a weighted summation on the molecular fingerprints of the two mixture components is applied. In the second, a more complex mixture graph to capture inter-molecular and intramolecular interactions is proposed. We implement different test splits to uncover model capabilities and limitations for different practical applications.

Both GNN models exhibit very high performance in the interpolation between mixture compositions scenario, showing that GNNs are capable of accurately predicting CMC_{mix} values at new mixture compositions. Extrapolation to new mixtures from known surfactants is also handled very well from both GNN models. Further extrapolation to mixtures where either one or both pure species of the mixture are previously unseen by the model, decreased model accuracy as expected. However, the model predictions remained accurate with some notable exceptions. Our findings indicate that no clear best-performing GNN architecture can be identified and hence we decide to combined them for more robust results. The final combined model leads to highly accurate CMC predictions for the first two test cases, while for the other two cases reasonable predictions with slightly higher errors are observed. Further analysis is conducted on the predictive performance on different class combinations.

We further investigate if a GNN model trained on binary mixture and pure species data can be applied to ternary mixtures, without further model adjustment. We collect a small external data set of 6 ternary mixtures and test both GNN models. We find that the WS-GNN model can predict the CMC_{mix} of ternary mixtures with high accuracy, in the sense of low RMSE (cf. Table 2), hence demonstrating potential for further development and applications, such as consideration of the unreacted surface-active raw material. Experimental validation on 4 commercial surfactants that contain up to 4 species and ternary mixtures is conducted and very good agreement between predictions and measurements is observed. Therefore, the model can accurately screen commercial surfactants and accelerate industrial research and development in a sustainable way.

Since the CMC of surfactants, mixtures, and pure species is influenced by the pH of the solution, accounting for this factor in future work would be very interesting. This is particularly pressing in the case of zwitterionic surfactants. Furthermore, understanding the impact of different surface-active species present in the final product would allow us to either incorporate such information during model training or refine a high-fidelity data set. Extending the models to ethoxylated alkylsulfates would be highly industrial relevant, however dealing with the wide product distribution of ethoxylation remains challenging. Lastly, the lack of data availability for surfactant mixtures of higher order limits the development and/or testing of ML models for CMC prediction and should be addressed in future works. We however anticipate that the model can predict CMC_{mix} values for mixtures with any number of surfactants, i.e., beyond quaternary systems presented in this work. The pre-requisite is that the structures of all species and the surface-active impurities are known.

CRediT authorship contribution statement

Christoforos Brozos: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Jan G. Rittig:** Writing – review

& editing, Software, Methodology, Formal analysis, Conceptualization. **Elie Akanny:** Writing – review & editing, Supervision, 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, Elie Akanny, Sandip Bhattacharya and Christina Kohlmann are employees of BASF Personal Care and Nutrition GmbH which uses surfactants commercially.

Acknowledgments

The BASF authors (C. Brozos, E. Akanny, S. Bhattacharya, 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).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compchemeng.2025.109085.

Data availability

Data will be made available on request.

References

Abe, M., 2004. Mixed Surfactant Systems, second ed. Marcel Dekker, New York 2005.
Alargova, R., Kochijashky, I., Sierra, M., Kwetkat, K., Zana, R., 2001. Mixed micellization of Dimeric (Gemini) surfactants and conventional surfactants: II. CMC and micelle aggregation numbers for various mixtures. J. Colloid Interface Sci. 235, 119–129.
Bakshi, M.S., Crisantino, R., De Lisi, R., Milioto, S., 1993. Volume and heat capacity of sodium dodecyl sulfate-dodecyldimethylamine oxide mixed micelles. J. Phys. Chem. 97, 6914–6919.
Bilodeau, C., Kazakov, A., Mukhopadhyay, S., Emerson, J., Kalantar, T., Muzny, C., Jensen, K., 2023. Machine learning for predicting the viscosity of binary liquid mixtures. Chem. Eng. J. 464, 142454.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.
Brozos, C., Rittig, J.G., Bhattacharya, S., Akanny, E., Kohlmann, C., Mitsos, A., 2024a. Graph neural networks for surfactant multi-property prediction. Colloids Surfaces A: Physicochem. Eng. Asp. 694, 134133.
Brozos, C., Rittig, J.G., Bhattacharya, S., Akanny, E., Kohlmann, C., Mitsos, A., 2024b. Predicting the temperature dependence of surfactant CMCs using graph neural networks. J. Chem. Theory Comput. 20, 5695–5707.
Cao, L., Russo, D., Felton, K., Salley, D., Sharma, A., Keenan, G., Mauer, W., Gao, H., Cronin, L., Lapkin, A.A., 2021. Optimization of formulations using robotic experiments driven by machine learning DoE. Cell Rep. Phys. Sci. 2, 100295.
Cheng, K.C., Khoo, Z.S., Lo, N.W., Tan, W.J., Chemmangattuvalappil, N.G., 2020. Design and performance optimisation of detergent product containing binary mixture of anionic-nonionic surfactants. Heliyon 6, e03861.
Clint, J.H., 1975. Micellization of mixed nonionic surface active agents. J. Chem. Soc. Faraday Trans. 1 (71), 1327–1334.
Dietterich, T.G., 2000. Ensemble methods in machine learning. In: Proceedings of the First International Workshop on Multiple Classifier Systems. Berlin, Heidelberg. pp. 1–15.
Fey, M., Lenssen, J.E., 2019. Fast graph representation learning with PyTorch geometric. ArXiv abs/1903.02428. (Accessed 1 June 2024).
Ganaie, M., Hu, M., Malik, A., Tanveer, M., Suganthan, P., 2022. Ensemble deep learning: A review. Eng. Appl. Artif. Intell. 115, 105151.
Geng, T., Zhang, C., Jiang, Y., Ju, H., Wang, Y., 2017. Synergistic effect of binary mixtures contained newly cationic surfactant: Interaction, aggregation behaviors and application properties. J. Mol. Liq. 232, 36–44.
Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E., 2017. Neural message passing for quantum chemistry. Int. Conf. Mach. Learn.
Grady, B.P., 2023. Surfactant mixtures: A short review. J. Surfactants Deterg. 26, 237–250.
Hamilton, W.L., Ying, Z., Leskovec, J., 2017. Inductive representation learning on large graphs. Neural Inf. Process. Syst..
Haque, M.E., Das, A.R., Rakshit, A.K., Moulik, S.P., 1996. Properties of mixed micelles of binary surfactant combinations. Langmuir 12, 4084–4089.
Hierrezuelo, J., Aguiar, J., Carnero Ruiz, C., 2006. Interactions in binary mixed systems involving a sugar-based surfactant and different n-alkyltrimethylammonium bromides. J. Colloid Interface Sci. 294, 449–457.
Hines, J., Fragneto, G., Thomas, R., Garrett, P., Rennie, G., Rennie, A., 1997a. Neutron reflection from mixtures of sodium dodecyl sulfate and dodecyl betaine adsorbed at the hydrophobic solid/aqueous interface. J. Colloid Interface Sci. 189, 259–267.
Hines, J.D., Thomas, R.K., Garrett, P.R., Rennie, G.K., Penfold, J., 1997b. Investigation of mixing in binary surfactant solutions by surface tension and neutron reflection: Anionic/nonionic and zwitterionic/nonionic mixtures. J. Phys. Chem. B 101, 9215–9223.
Hines, J.D., Thomas, R.K., Garrett, P.R., Rennie, G.K., Penfold, J., 1998. Investigation of mixing in binary surfactant solutions by surface tension and neutron reflection: Strongly interacting anionic/zwitterionic mixtures. J. Phys. Chem. B 102, 8834–8846.
Holland, P.M., Rubingh, D.N., 1983. Nonideal multicomponent mixed micelle model. J. Phys. Chem. 87, 1984–1990.
Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., Leskovec, J., 2020. Strategies for pre-training graph neural networks. In: International Conference on Learning Representations.
Huang, J., Ren, Z.H., 2017. Synergistic interaction between nonionic octylphenol polyoxyethylene ethers and effect of hydrophilic chain. J. Surfactants Deterg. 20, 1197–1203.
Huang, J., Ren, Z.H., 2019. Micellization and interactions for ternary mixtures of amino sulfonate surfactant and nonionic octylphenol polyoxyethylene ethers in aqueous solution: 1 blending with nonionic surfactants with smaller numbers of hydrophilic unit. J. Mol. Liq. 278, 53–60.
Hunter, J.E., Fowler, Jr., J.F., 1998. Safety to human skin of cocamidopropyl betaine: A mild surfactant for personal-care products. J. Surfactants Deterg. 1, 235–239.
Iyer, J., Blankschtein, D., 2014. Molecular-thermodynamic framework to predict the micellization behavior of mixtures of fluorocarbon-based and hydrocarbon-based surfactants. J. Phys. Chem. B 118, 2377–2388, PMID: 24512047.
Iyer, J., Mendenhall, J.D., Blankschtein, D., 2013. Computer simulation–molecular-thermodynamic framework to predict the micellization behavior of mixtures of surfactants: Application to binary surfactant mixtures. J. Phys. Chem. B 117, 6430–6442, PMID: 23634888.
J.G., Rittig, Mitsos, A., 2024. Thermodynamics-consistent graph neural networks. https://arxiv.org/abs/2407.18372.
Kelleppan, V.T., Butler, C.S., Williams, A.P., Vidallon, M.L.P., Giles, L.W., King, J.P., Sokolova, A.V., de Campo, L., Pearson, G.R., Tabor, R.F., Tuck, K.L., 2023. Components of cocamidopropyl betaine: Surface activity and self-assembly of pure alkyl amidopropyl betaines. Colloids Surfaces A: Physicochem. Eng. Asp. 656, 130435.
Khemani, B., Patil, S., Kotecha, K., Tanwar, S., 2024. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. J. Big Data 11, 18.
Ko, J.-S., Oh, S.-W., Kim, Y.-S., Nakashima, N., Nagadome, S., Sugihara, G., 2004. Adsorption and micelle formation of mixed surfactant systems in water. IV. Three combinations of SDS with MEGA-8, -9 and -10. J. Oleo Sci. 53, 109–126.
Kumar Shah, S., Chakraborty, G., Bhattarai, A., De, R., 2022. Synergistic and antagonistic effects in micellization of mixed surfactants. J. Mol. Liq. 368, 120678.
Leenhouts, R.J., Larsson, T., Verhelst, S., Vermeire, F.H., 2025. Property prediction of fuel mixtures using pooled graph neural networks. Fuel 381, 133218.
Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J., 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25.1. Adv. Drug Deliv. Rev. 46, 3–26, Special issue dedicated to Dr. Eric Tomlinson, Advanced Drug Delivery Reviews, A Selection of the Most Highly Cited Articles, 1991-1998.
López-Fontán, L.J., Suárez, J.M., Mosquera, V., Sarmiento, F., 1999. Mixed micelles of n-alkyltrimethylammonium bromides: influence of alkyl chain length. Phys. Chem. Chem. Phys. 1, 3583–3587.
López-Fontán, J.L., Suárez, M.J., Mosquera, V., Sarmiento, F., 2000. Micellar behavior of n-alkyl sulfates in binary mixed systems. J. Colloid Interface Sci. 223, 185–189.
Maiti, K., Bhattacharya, S., Moulik, S., Panda, A., 2010. Physicochemistry of the binary interacting mixtures of cetylpyridinium chloride (CPC) and sodium dodecylsulfate (SDS) with special reference to the catanionic ion-pair (coacervate) behavior. Colloids Surfaces A: Physicochem. Eng. Asp. 355, 88–98.

Martín, V.I., Rodríguez, A., Graciani, M.d.M., Robina, I., Moyá, M.L., 2010. Study of the micellization and micellar growth in pure alkanediyl-α-ω-bis(dodecyldimethylammonium) bromide and MEGA10 surfactant solutions and their mixtures. Influence of the spacer on the enthalpy change accompanying sphere-to-rod transitions. J. Phys. Chem. B 114, 7817–7829.

Massarweh, O., Abushaikha, A.S., 2020. The use of surfactants in enhanced oil recovery: A review of recent advances. Energy Rep. 6, 3150–3178.

McLachlan, A.A., Marangoni, D.G., 2006. Interactions between zwitterionic and conventional anionic and cationic surfactants. J. Colloid Interface Sci. 295, 243–248.

Misselyn-Bauduin, A.M., Thibaut, A., Grandjean, J., Broze, G., Jérôme, R., 2000. Mixed micelles of anionic-nonionic and anionic-zwitterionic surfactants analyzed by pulsed field gradient NMR. Langmuir 16, 4430–4435.

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

Moulik, S.P., Haque, M.E., Jana, P.K., Das, A.R., 1996. Micellar properties of cationic surfactants in pure and mixed states. J. Phys. Chem. 100, 701–708.

Moulik, S.P., Rakshit, A.K., Naskar, B., 2021. 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, 535–549.

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

Nevolianis, T., Rittig, J.G., Mitsos, A., Leonhard, K., 2024. Multi-fidelity graph neural networks for predicting toluene/water partition coefficients. ChemRxiv.

Nitschke, M., Costa, S., 2007. Biosurfactants in food industry. Trends Food Sci. Technol. 18, 252–259.

Phaodee, P., Sabatini, D.A., 2021. Anionic and cationic surfactant synergism: Minimizing precipitation, microemulsion formation, and enhanced solubilization and surface modification. J. Surfactants Deterg. 24, 551–562.

Prasad, M., Chakraborty, I., Rakshit, A.K., Moulik, S.P., 2006. Critical evaluation of micellization behavior of nonionic surfactant MEGA 10 in comparison with ionic surfactant tetradecyltriphenylphosphonium bromide studied by microcalorimetric method in aqueous medium. J. Phys. Chem. B 110, 9815–9821.

Qin, S., Jiang, S., Li, J., Balaprakash, P., Van Lehn, R.C., Zavala, V.M., 2023. Capturing molecular interactions in graph neural networks: a case study in multi-component phase equilibrium. Digit. Discov. 2, 138–151.

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

Reiser, P., Neubert, M., Eberhard, A., Torresi, L., Zhou, C., Shao, C., Metni, H., van Hoesel, C., Schopmans, H., Sommer, T., Friederich, P., 2022. Graph neural networks for materials science and chemistry. Commun. Mater. 3, 93.

Ren, Z.H., Luo, Y., Zheng, Y.C., Shi, D.P., Mei, P., Li, F.X., 2014. Interaction behavior between an amino sulfonate surfactant and octylphenol polyoxyethylene Ether (10) in aqueous solution. J. Solut. Chem. 43, 853–869.

Rittig, J.G., Ben Hicham, K., Schweidtmann, A.M., Dahmen, M., Mitsos, A., 2023a. Graph neural networks for temperature-dependent activity coefficient prediction of solutes in ionic liquids. Comput. Chem. Eng. 171, 108153.

Rittig, J.G., Felton, K.C., Lapkin, A.A., Mitsos, A., 2023b. Gibbs–Duhem-informed neural networks for binary activity coefficient prediction. Digit. Discov. 2, 1752–1767.

Rittig, J.G., Gao, Q., Dahmen, M., Mitsos, A., Schweidtmann, A.M., 2023c. Machine learning and hybrid modelling for reaction engineering: Theory and applications. R. Soc. Chem..

Rodríguez, A., Graciani, M.d.M., Moreno-Vargas, A.J., Moyá, M.L., 2008. Mixtures of monomeric and dimeric surfactants: Hydrophobic chain length and spacer group length effects on non ideality. J. Phys. Chem. B 112, 11942–11949.

Rodríguez Patino, J.M., Carrera Sánchez, C., Rodríguez Niño, M.R., 2008. Implications of interfacial characteristics of food foaming agents in foam formulations. Adv. Colloid Interface Sci. 140, 95–113.

Rogers, D., Hahn, M., 2010. Extended-connectivity fingerprints. J. Chem. Inf. Model. 50, 742–754.

Rosen, M.J., Hua, X.Y., 1982. Synergism in binary mixtures of surfactants: II. Some experimental data. J. Am. Oil Chem. Soc. 59, 582–585.

Rosen, M., Kunjappu, J., 2012. Surfactants and Interfacial Phenomena: Rosen/Surfactants 4E. John Wiley & Sons, Ltd.

Rubingh, D.N., 1979. Solution Chemistry of Surfactants, vol. 2, Plenum Press, New York, pp. 337–354.

Sanchez Medina, E.I., Linke, S., Stoll, M., Sundmacher, K., 2022. Graph neural networks for the prediction of infinite dilution activity coefficients. Digit. Discov. 1, 216–225.

Sanchez Medina, E.I., Linke, S., Stoll, M., Sundmacher, K., 2023. Gibbs–Helmholtz graph neural network: capturing the temperature dependency of activity coefficients at infinite dilution. Digit. Discov. 2, 781–798.

Schweidtmann, A.M., Rittig, J.G., König, A., Grohe, M., Mitsos, A., Dahmen, M., 2020. Graph neural networks for prediction of fuel ignition quality. Energy & Fuels 34, 11395–11407.

Shaban, S.M., Kang, J., Kim, D.-H., 2020. Surfactants: Recent advances and their applications. Compos. Commun. 22, 100537.

Shiloach, A., Blankschtein, D., 1997. Prediction of critical micelle concentrations and synergism of binary surfactant mixtures containing zwitterionic surfactants. Langmuir 13, 3968–3981.

Shiloach, A., Blankschtein, D., 1998. Predicting micellar solution properties of binary surfactant mixtures. Langmuir 14, 1618–1636.

Sonu, A.K., Saha, S.K., 2013. Study on mixed micelles of cationic gemini surfactants having hydroxyl groups in the spacers with conventional cationic surfactants: Effects of spacer group and hydrocarbon tail length. Ind. Eng. Chem. Res. 52, 5895–5905.

Specht, T., Nagda, M.K., Fellenz, S., Mandt, S., Hasse, H., Jirasek, F., 2024. HANNA: Hard-constraint neural network for consistent activity coefficient prediction. ArXiv abs/2407.18011.

Srinivasan, V., Blankschtein, D., 2005. Prediction of conformational characteristics and micellar solution properties of fluorocarbon surfactants. Langmuir 21, 1647–1660.

Tadros, T., 2005. Applied Surfactants. John Wiley & Sons, Ltd, pp. 399–432, Chapter 12.

Treiner, C., Makayssi, A., 1992. Structural micellar transition for dilute solutions of long chain binary cationic surfactant systems: a conductance investigation. Langmuir 8, 794–800.

ud Din, K., Sheikh, M.S., Dar, A.A., 2009. Interaction of a cationic gemini surfactant with conventional surfactants in the mixed micelle and monolayer formation in aqueous medium. J. Colloid Interface Sci. 333, 605–612.

van der Maaten, L., Hinton, G.E., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Vermeire, F.H., Green, W.H., 2021. Transfer learning for solvation free energies: From quantum chemistry to experiments. Chem. Eng. J. 418, 129307.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S., 2021. A comprehensive survey on graph neural networks. IEEE Trans. Neural Netw. Learn. Syst. 32, 4–24.

Xu, K., Hu, W., Leskovec, J., Jegelka, S., 2018. How powerful are graph neural networks? ArXiv abs/1810.00826. (Accessed 10 May 2024).

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., Barzilay, R., 2019. Analyzing learned molecular representations for property prediction. J. Chem. Inf. Model. 59, 3370–3388.

Zahrt, A.F., Henle, J.J., Denmark, S.E., 2020. Cautionary guidelines for machine learning studies with combinatorial datasets. ACS Comb. Sci. 22, 586–591.

Zhang, H., Lai, T., Chen, J., Manthiram, A., Rondinelli, J.M., Chen, W., 2024. Learning molecular mixture property using chemistry-aware graph neural network. PRX Energy 3, 023006.

Zhang, R., Zhang, L., Somasundaran, P., 2004. Study of mixtures of n-dodecyl-β-d-maltoside with anionic, cationic, and nonionic surfactant in aqueous solutions using surface tension and fluorescence techniques. J. Colloid Interface Sci. 278, 453–460.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M., 2020. Graph neural networks: A review of methods and applications. AI Open 1, 57–81.