Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

Comparative Analysis of PharmHGT, GCN, and GAT Models for Predicting LogCMC in Surfactants.

Gabriela C. Theis Marchan[a], Teslim Olayiwola[a], Jose Romagnoli[a]*

[a] Louisiana State University, Department of Chemical Engineering, Baton Rouge, Louisiana 70803, United States
* Corresponding Author: jose@lsu.edu.

ABSTRACT

Predicting the critical micelle concentration (CMC) of surfactants is essential for optimizing their applications in various industries, including pharmaceuticals, detergents, and emulsions. In this study, we investigate the performance of graph-based machine learning models, specifically Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), and a graph-transformer model, PharmHGT, for predicting CMC values. We aim to determine the most effective model for capturing the structural and physicochemical properties of surfactants. Our results provide insights into the relative strengths of each approach, highlighting the potential advantages of transformer-based architectures like PharmHGT in handling molecular graph representations compared to traditional graph neural networks. This comparative study serves as a step towards enhancing the accuracy of CMC predictions, contributing to the efficient design of surfactants for targeted applications.

Keywords: Graph Neural Networks, Critical Micelle Concentration, Surfactants, Machine Learning, Molecular Property Prediction.

INTRODUCTION

Surfactants are versatile chemical compounds utilized across various industries, including cosmetics,(*1*) pharmaceuticals, detergents,(*2*) and enhanced oil recovery.(*3*) Their unique amphiphilic structure, containing both hydrophilic and hydrophobic components, enables them to interact with immiscible phases by reducing interfacial tension.(*4*) A key property influencing surfactant behavior is the critical micelle concentration (CMC), defined as the minimum concentration at which micelles begin to form spontaneously.

Traditional methods for CMC determination, such as surface tension measurements(*5, 6*) and fluorescence spectroscopy,(*7*) while accurate, are time-consuming and require specialized equipment. Computational approaches, including quantitative structure-property relationship (QSPR) models(*8*) and molecular dynamics simulations,(*9*) offer alternatives but face limitations in accuracy or computational efficiency. Recent advances in graph-based neural networks present promising opportunities for molecular property prediction, offering potentially faster and more accurate predictions while capturing complex structure-property relationships.

This study compares three graph-based deep learning architectures—PharmHGT, GCN, and GAT—for predicting LogCMC values. We aim to evaluate their relative strengths in capturing molecular features and their potential for accelerating surfactant design and optimization.

METHODOLOGY

Model Architectures

PharmHGT Model

Our work adapts the Pharmacophoric-constrained heterogeneous graph transformer (PharmHGT) architecture, originally developed by Jiang et al.(*10*) for molecular property prediction, to specifically address the unique challenges of surfactant modeling and LogCMC prediction. While maintaining the core heterogeneous graph architecture of PharmHGT, we introduce several key modifications to handle surfactant-specific characteristics such as amphiphilic structure, head group chemistry, and tail group interactions.

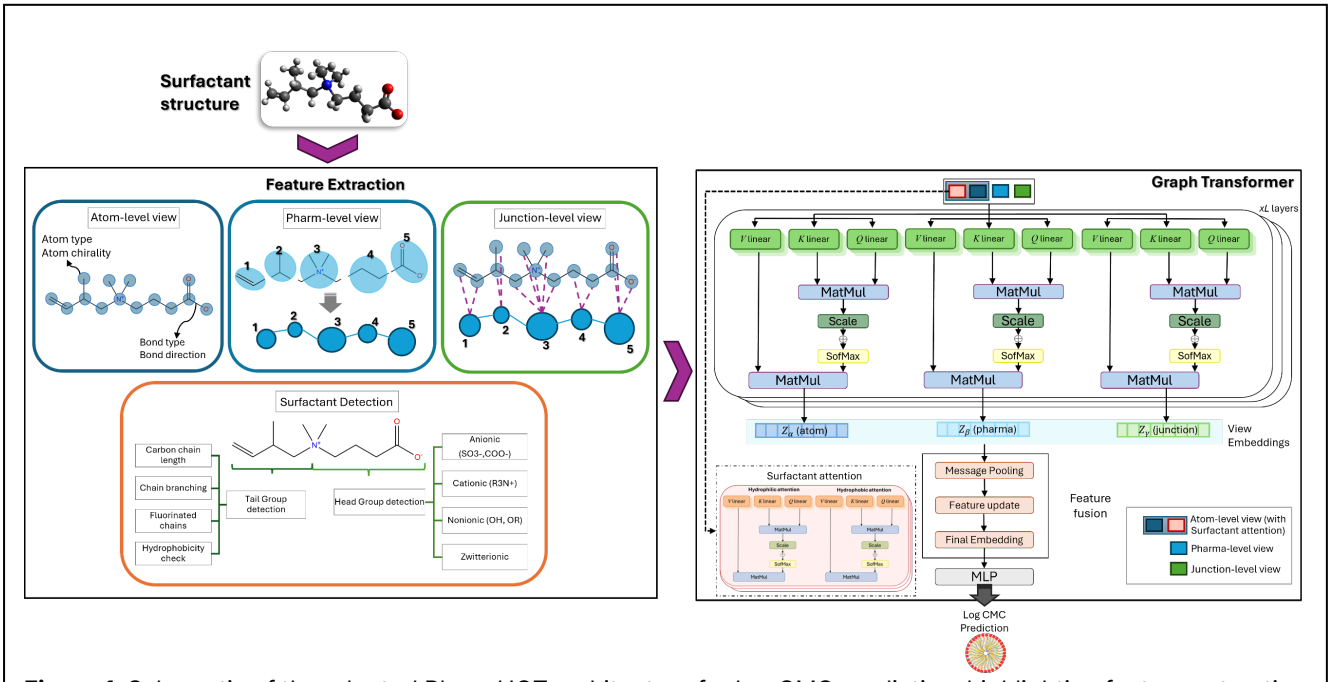

Figure 1: Schematic of the adapted PharmHGT architecture for LogCMC prediction, highlighting feature extraction (atom-level, pharmacophore-level, and junction-level), the surfactant detection module, attention mechanisms, and hierarchical feature fusion.

While the original PharmHGT framework was designed for capturing pharmacophoric structures through heterogeneous graph representation, we found it initially struggled to recognize the unique molecular structures of surfactants due to their unique amphiphilic nature and the interactions between hydrophilic and hydrophobic regions. However, the framework's underlying architecture for heterogeneous graph learning provided a valuable foundation. Our adaptation enhances this framework by incorporating surfactant-specific features and attention mechanisms, enabling it to effectively capture the distinct functional regions that significantly influence surfactant properties.

The adapted PharmHGT represents surfactants as heterogeneous graphs $G = (V, E)$, where V is the set of nodes representing molecular components (e.g., atoms or functional groups) and E is the set of edges representing chemical bonds or interactions. Node types are mapped using $\phi: V \to O$, and edge types are mapped using $\psi: E \to P$. This representation integrates three complementary views of the molecular structure:

The atom-level view (G_α), which captures atomic features such as atomic number, chirality, and bonding patterns. Additionally, it encodes surfactant-specific information, such as whether an atom belongs to a hydrophilic head group or hydrophobic tail group.

The pharmacophore-level view (G_β), which focuses on functional groups and their interactions, especially tail-head relationships. Pharmacophoric features are derived using MACCS keys and chemical feature detection.

The junction-level view (G_γ), which combines atom-level and pharmacophoric information to provide a comprehensive molecular representation.

These three complementary views are illustrated in Figure 1 in the feature extraction section, which provides an overview of the adapted PharmHGT model and its surfactant-specific enhancements.

The model employs a multi-head attention mechanism to dynamically learn the importance of different molecular features across atom-level, pharmacophore-level, and junction-level views. The attention mechanism is mathematically expressed as:

$$Attention(Q, K, V) = \sum_{p \in P} \Omega_p \cdot \sigma\left(\frac{Q_p K_p^T}{\sqrt{d_k}}\right) \cdot V_p \qquad (1)$$

where Q, K, and V are the query, key, and value matrices; Ω_p is a learnable weight matrix specific to view p; and σ represents the SoftMax function. This mechanism prioritizes interactions unique to surfactants, such as those between hydrophilic head groups and hydrophobic tail regions, achieved through dedicated attention modules. Feature propagation through the graph is governed by a multi-view message-passing framework (MVMP), expressed as Eq. 2:

$$M^{(t)}(X_v) = \sum_{u \in N(v)} Attention\left(H^{(t-1)}(X_v)W_Q, H^{(t-1)}(X_u)W_K, H^{(t-1)}(X_u)W_V\right) \qquad (2)$$

where $M^{(t)}(X_v)$ is the updated feature of node v at

step t; $N(v)$ represents the neighbors of node v; $H^{(t-1)}(X_v)$ is the feature of v from the previous step; and W_Q,W_K,W_V are learnable projection matrices. The MVMP framework integrates attention mechanisms to dynamically aggregate and prioritize molecular interactions, effectively capturing the diverse structural and functional relationships unique to surfactants.

After integrating surfactant-specific features into the atom-level embeddings, the model combines all structural views (atom-level, pharmacophore-level, and junction-level) through a hierarchical Readout Attention mechanism:

$$Z_{\gamma\beta} = ReadOutAttention(Z_\gamma, Z_\beta) \quad (3)$$

where Z_γ and Z_β are the feature embeddings from the junction-level and pharmacophore-level views, respectively. The resulting embedding $Z_{\gamma\beta}$ is further integrated with atom-level features using Eq. 4:

$$Z = ReadOutAttention(Z_\alpha, Z_{\gamma\beta}) \quad (4)$$

where Z_α is the atom-level embedding. Finally, the LogCMC prediction is obtained through a multi-layer perceptron (MLP), expressed as Eq. 5:

$$\hat{y} = MLP(Z) \quad (5)$$

GCN Model

Our GCN architecture consists of two graph convolutional layers followed by a three-layer fully connected regression network as we defined in our previous work.(*11*) The model processes molecular graphs by iteratively updating node features based on their local neighborhood information.(*11, 12*) Each node's feature vector encodes atomic properties and surfactant-specific characteristics. The first convolutional layer maps input features to a hidden representation, while the second layer refines these representations. A global mean pooling operation then aggregates node-level features into a graph-level representation, which is processed through the regression network for final prediction. This architecture effectively captures local molecular structure and chemical bonding patterns while maintaining computational efficiency.

GAT Model

The GAT architecture enhances the basic graph neural network framework by incorporating attention mechanisms that dynamically weight the importance of different atomic interactions.(*13*) Our implementation uses a dual-layer architecture with four attention heads in the first layer and a single consolidating head in the second layer. This multi-head attention allows the model to capture different aspects of molecular structure simultaneously. Each attention head computes importance weights for neighboring atoms, enabling the model to focus on the most relevant structural features for LogCMC prediction. The model combines local atomic features with global molecular context through a series of attention-weighted transformations, followed by a global pooling operation and regression layers.

Data and Implementation

Our study utilized two carefully curated datasets of surfactants compiled from literature sources.(*14-19*) Data1 consists of 285 non-ionic surfactants, providing a focused dataset for evaluating model performance on a single surfactant class. Unlike Data1, Data2 encompasses 365 surfactants with a broader chemical diversity: 285 non-ionic, 35 cationic, 34 anionic, and 11 zwitterionic surfactants. This diverse dataset enables the assessment of model generalization across different surfactant types.

For molecular representation, we used RDKit(*20*) to generate comprehensive feature sets for each surfactant. Atomic features included atomic number, formal charge, hybridization state, aromaticity, and surfactant-specific properties such as head/tail group identification. Bond features captured bond type, conjugation, and ring membership. For the PharmHGT model, additional pharmacophoric features were extracted using MACCS keys(*21*) and chemical feature detection.

In creating the regression models, we adopted the PyTorch(*22*) for neural network components and the Deep Graph Library (DGL)(*23*) for efficient graph operations. For model training and evaluation, we systematically divided the available experimental datasets (containing the chemical structure and their corresponding logCMC values) into training (70%), testing (20%), and validation (10%) sets, maintaining the consistent distribution of surfactant types across splits. Hyperparameter optimization was conducted using Optuna, running 100 trials per model. Table S1 in the digital supplementary material presents the optimal hyperparameters for each model-dataset combination. All models were trained using early stopping with patience ranging from 20-40 epochs (monitoring validation loss without improvement) to prevent overfitting while ensuring convergence. Data splitting was performed using stratified random sampling to maintain a consistent distribution of surfactant types across training, testing, and validation sets. (see Table S2 in the digital supplementary material)

RESULTS AND DISCUSSION

We evaluated the performance of PharmHGT, GCN, and GAT models on two distinct datasets (Data1 and Data2) using root mean square error (RMSE) and mean absolute error (MAE) as primary metrics. The comparative analysis reveals notable differences in predictive capabilities across model architectures and dataset compositions.

For Data1, which consists exclusively of non-ionic

surfactants (n=285), all three models demonstrated strong predictive capabilities, albeit with varying degrees of accuracy. As shown in Table 1, the PharmHGT model achieved the highest performance, with a test MSE of 0.090 and MAE of 0.207, followed by the GCN model (MSE: 0.238, MAE: 0.340) and the GAT model (MSE: 0.314, MAE: 0.369). Notably, the training metrics in Table 1 indicate strong learning capacity across all models, with PharmHGT showing the lowest training MSE (0.017) and highest training R^2 (0.990).

The superior performance of PharmHGT can be better understood by examining the structural complexity of Data1. Despite focusing solely on non-ionic surfactants, Data1 exhibits considerable structural diversity at both atomic and pharmacophoric levels. Quantitative diversity analysis reveals five unique atom types (C: 72.55%, O: 25.64%, N: 0.92%, F: 0.59%, S: 0.31%) with significant variance in key atom features (mean-variance: 0.036). At the pharmacophoric level, molecules contain a wide range of functional fragments (1-84 per molecule, average: 9.14±10.24) with varied fragment sizes (1-22 atoms, average: 3.38±3.68). Feature similarity analysis identifies multiple distinct structural clusters, indicating that the dataset encompasses diverse molecular scaffolds despite focusing on non-ionic surfactants. This structural complexity particularly benefits PharmHGT's heterogeneous graph representation, which better captures the varying importance of different structural features compared to the more uniform approaches of GCN and GAT.

Table 1: Model Performance on Data1 (Non-ionic Surfactants)

Model	MSE		R^2		MAE	
	Train	Test	Train	Test	Train	Test
GAT	0.075	0.314	0.955	0.845	0.158	0.369
GCN	0.042	0.238	0.974	0.883	0.091	0.340
PharmHGT	0.017	0.090	0.990	0.943	0.098	0.207

Figure 2 presents scatter plots comparing predicted versus experimental values for all three models on Data1 in the top row. The PharmHGT model exhibits the tightest clustering around the diagonal line (R^2 = 0.943), indicating superior prediction accuracy across the full range of CMC values. The GCN model shows good correlation (R^2 = 0.883), while the GAT model demonstrates slightly lower but still substantial predictive power (R^2 = 0.845). Both models show increased scatter at extreme CMC values compared to PharmHGT. The outlier non-ionic surfactants in Figure 2 feature: (1) extreme ethoxylation patterns (very short or very long chains), (2) highly branched hydrophobic tails, and (3) atypical oxygen atom distributions throughout the molecule rather than concentrated in head groups. PharmHGT's superior handling of these challenging structures (reducing prediction error by up to 31% compared to GAT) demonstrates the advantage of heterogeneous graph representation and transformer-based attention mechanisms for capturing complex structure-property relationships in surfactants with

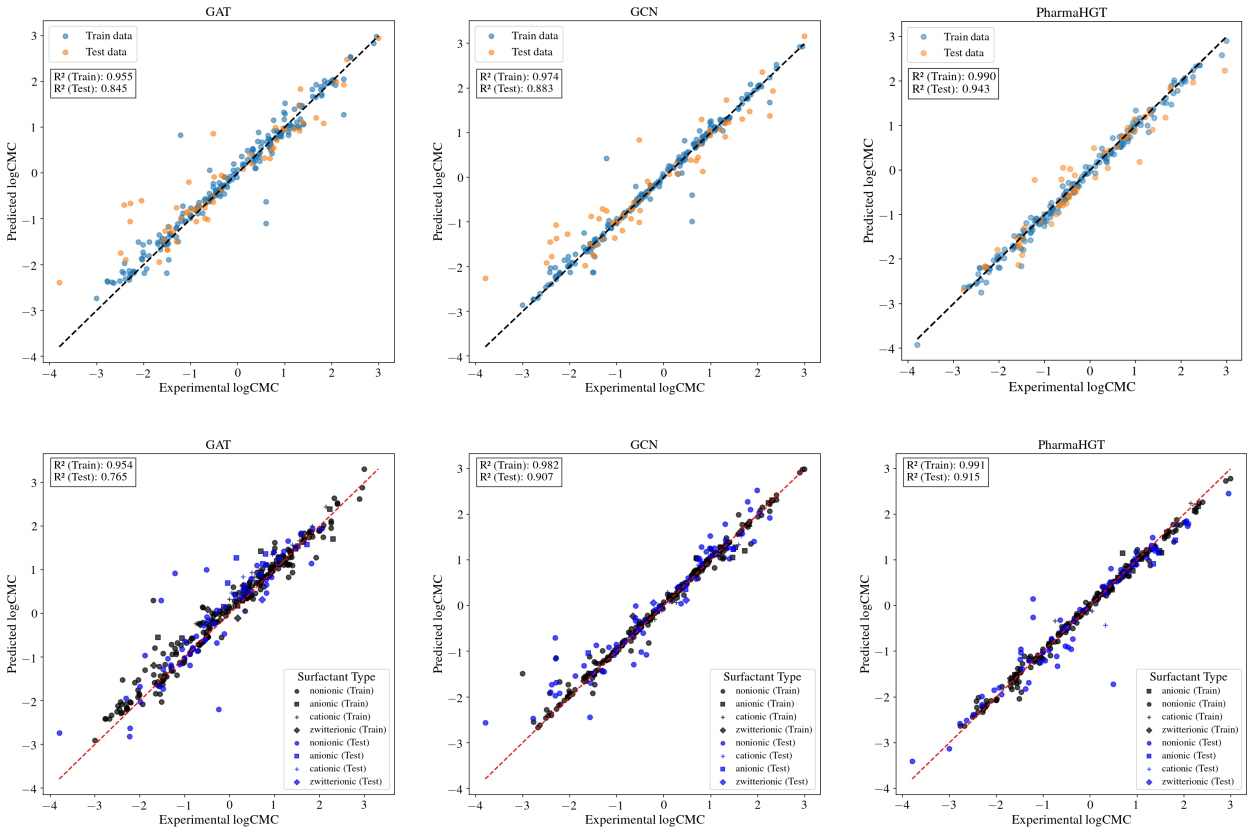

Figure 2: Scatter Plots of Predicted Vs. Experimental LogCMC Values for the top row representing Data Set 1 and the bottom row representing Data Set 2.

unconventional features.

The more diverse Data2 dataset (n=365), containing a mixture of non-ionic, cationic, anionic, and zwitterionic surfactants, presented a more challenging prediction task. As detailed in Table 2, the performance differences between models became more pronounced. The PharmHGT model maintained robust performance (MSE: 0.171, MAE: 0.250), demonstrating effective handling of diverse surfactant types. The GCN model showed comparable performance (MSE: 0.182, MAE: 0.307), while the GAT model exhibited higher prediction error (MSE: 0.318, MAE: 0.369). The training metrics in Table 2 reveal particularly strong learning capacity for both PharmHGT (training MSE: 0.014, R^2: 0.991) and GCN (training MSE: 0.027, R^2: 0.982), suggesting effective feature extraction from the more complex dataset.

Figure 2 illustrates the prediction performance on Data2 through scatter plots in the bottom row. The PharmHGT model maintains a strong correlation (R^2 = 0.915) despite the increased chemical diversity. The GCN model shows similar performance (R^2 = 0.907), while the GAT model (R^2 = 0.765) shows more substantial scatter across surfactant types. All models display some deviation from the ideal prediction line for surfactants with extreme LogCMC values, though this effect is least pronounced in the PharmHGT predictions.

Table 2: Model Performance on Data2 (Mixed surfactants)

Model	MSE		R^2		MAE	
	Train	Test	Train	Test	Train	Test
GAT	0.078	0.318	0.954	0.765	0.190	0.369
GCN	0.027	0.182	0.982	0.907	0.083	0.307
Pharm HGT	0.014	0.171	0.991	0.915	0.080	0.250

Tables 1-2 report results from a single stratified random split (70/20/10) with fixed random seeds to ensure reproducible partitioning. To evaluate the impact of model configuration on performance, we first assessed all models with identical default parameters on the same data split, then performed systematic hyperparameter optimization using Optuna (100 trials per model). Notably, PharmHGT with default parameters (R^2 = 0.925 on Data1) outperformed both GCN and GAT even before optimization (R^2 = 0.862 and 0.814 respectively), demonstrating

that its superior performance is not merely an artifact of hyperparameter tuning. While performance may vary slightly with different splits, the consistent ranking of models across both default and optimized configurations confirms the robustness of our findings with respect to data partitioning and parameter selection.

Analysis of Model Strengths

Based on the two dataset types studied in this work, it was discovered that PharmHGT performs better than GCN and GAT and the GAT model showing the least accuracy. The PharmHGT model's superior performance can be attributed to its sophisticated architecture that combines transformer-based attention mechanisms with heterogeneous graph representation. Its consistent performance across both datasets suggests effective capture of both local molecular features and global structural patterns. The GCN model showed robust performance, particularly on Data2, where it achieved comparable results to PharmHGT. This suggests that its message-passing mechanism effectively captures essential molecular features for LogCMC prediction. Lastly, the GAT model's performance revealed limitations in handling diverse surfactant types, with accuracy decreasing notably for the mixed surfactant dataset. This suggests that simple attention mechanisms may not fully capture the complex interactions present in diverse surfactant systems.

CONCLUSION

In this study, we conducted a comprehensive comparison of three graph-based deep learning architectures (PharmHGT, GCN, and GAT) for predicting LogCMC values of surfactants. Our investigation using two distinct datasets, one focused on non-ionic surfactants and another encompassing diverse surfactant types, revealed several important findings about the capabilities and limitations of each architecture.

The PharmHGT model demonstrated superior performance across both datasets, achieving the highest prediction accuracy (R^2 = 0.943 for non-ionic surfactants and R^2 = 0.915 for mixed surfactant types) and the most consistent generalization behavior. The model's ability to maintain high accuracy when transitioning from non-ionic to mixed surfactant types (with only minimal performance degradation) highlights its robustness in handling diverse molecular structures. This effectiveness stems from its sophisticated architecture that combines transformer-based attention mechanisms with heterogeneous graph representation.

The GCN model showed surprisingly strong performance, particularly for the mixed surfactant dataset where it achieved results comparable to PharmHGT (R^2 = 0.907). This suggests that well-designed message-passing mechanisms can effectively capture the essential molecular features governing surfactant behavior. The GAT model, while performing adequately for non-ionic surfactants (R^2 = 0.845), showed limitations in handling the increased complexity of mixed surfactant types (R^2 = 0.765), indicating that simple attention mechanisms may not fully capture the complex interactions present in diverse surfactant systems.

Future work could explore several promising directions. First, the models could be extended to predict additional surfactant properties beyond LogCMC, such as surface tension or aggregation numbers. Second, the architecture could be modified to incorporate explicit handling of environmental factors such as temperature, pH, and ionic strength, which significantly influence surfactant behavior. Finally, the strong performance of these models suggests potential applications in inverse design problems, where the goal is to generate novel surfactant structures with desired properties.

DIGITAL SUPPLEMENTARY MATERIAL

The digital supplementary material includes Table S1 detailing optimal hyperparameters for all models (batch size, learning rate, hidden units, epochs, and model-specific parameters); Table S2 showing Hyperparameter Ranges for GAT, GCN, and Graph Transformers. This information supports findings regarding model optimization, reproducibility, and result robustness. The supplementary material is available at https://psecommunity.org/LAPSE:2025.0036v1.

ACKNOWLEDGEMENTS

This work was supported by the U.S Department of Energy, Office of Science, under the Office of Basic Energy Science Separation Science program under Award No. DE-SC0022304. The authors gratefully acknowledge the computer time allotted by the high-performance computing center (HPC) at LSU and the Louisiana Network initiative.

REFERENCES

1. B. Bhattacharya, T. K. Ghosh, N. Das, Application of bio-surfactants in cosmetics and pharmaceutical industry. Sch. Acad. J. Pharm 6, 320-329 (2017).
2. C. Ceresa, L. Fracchia, E. Fedeli, C. Porta, I. M. Banat, Recent advances in biomedical, therapeutic and pharmaceutical applications of microbial surfactants. Pharmaceutics 13, 466 (2021).
3. C. Negin, S. Ali, Q. Xie, Most common surfactants employed in chemical enhanced oil recovery. Petroleum 3, 197-211 (2017).
4. M. J. Rosen, J. T. Kunjappu, Surfactants and interfacial phenomena. (John Wiley & Sons, 2012).

5. M. Bielawska, A. Chodzińska, B. Jańczuk, A. Zdziennicka, Determination of CTAB CMC in mixed water+ short-chain alcohol solvent by surface tension, conductivity, density and viscosity measurements. Colloids and Surfaces A: Physicochemical and Engineering Aspects 424, 81-88 (2013).
6. S. F. Burlatsky et al., Surface tension model for surfactant solutions at the critical micelle concentration. Journal of colloid and interface science 393, 151-160 (2013).
7. M. M. Mabrouk, N. A. Hamed, F. R. Mansour, Spectroscopic methods for determination of critical micelle concentrations of surfactants; a comprehensive review. Applied Spectroscopy Reviews 58, 206-234 (2023).
8. A. R. Katritzky, L. Pacureanu, D. Dobchev, M. Karelson, QSPR study of critical micelle concentration of anionic surfactants using computational molecular descriptors. Journal of chemical information and modeling 47, 782-793 (2007).
9. A. P. Santos, A. Z. Panagiotopoulos, Determination of the critical micelle concentration in simulations of surfactant systems. The Journal of chemical physics 144, (2016).
10. Y. Jiang, S. Jin, X. Jin, Pharmacophoric-constrained heterogeneous graph transformer model for molecular property prediction. Communications. Chemistry 6, (2023).
11. P. Naghshnejad, G. Theis Marchan, T. Olayiwola, R. Kumar, J. Romagnoli, Graph-Based Modeling and Molecular Dynamics for Ion Activity Coefficient Prediction in Polymeric Ion-Exchange Membranes. Industrial & Engineering Chemistry Research, (2024).
12. D. Deng et al., XGraphBoost: extracting graph neural network-based features for a better prediction of molecular properties. Journal of chemical information and modeling 61, 2697-2705 (2021).
13. P. Velickovic et al., Graph attention networks. stat 1050, 10-48550 (2017).
14. M. Nnadili et al., Surfactant-Specific AI-Driven Molecular Design: Integrating Generative Models, Predictive Modeling, and Reinforcement Learning for Tailored Surfactant Synthesis. Industrial & Engineering Chemistry Research 63, 6313-6324 (2024).
15. S. Qin, T. Jin, R. C. Van Lehn, V. M. Zavala, Predicting critical micelle concentrations for surfactants using graph convolutional neural networks. The Journal of Physical Chemistry B 125, 10610-10620 (2021).
16. R. A. Saunders, J. A. Platts, Correlation and prediction of critical micelle concentration using polar surface area and LFER methods. Journal of physical organic chemistry 17, 431-438 (2004).
17. S. Lee, J. Lee, H. Yu, J. Lim, Synthesis of environment friendly nonionic surfactants from sugar base and characterization of interfacial properties for detergent application. Journal of Industrial and Engineering Chemistry 38, 157-166 (2016).
18. L. Chaveriat, I. Gosselin, C. Machut, P. Martin, Synthesis, surface tension properties and antibacterial activities of amphiphilic D-galactopyranose derivatives. European Journal of Medicinal Chemistry 62, 177-186 (2013).
19. C. Yan, G. Li, The rise of machine learning in polymer discovery. Advanced Intelligent Systems 5, 2200243 (2023).
20. G. Landrum, RDKit: Open-source cheminformatics. 2006. Google Scholar, (2006).
21. J. L. Durant, B. A. Leland, D. R. Henry, J. G. Nourse, Reoptimization of MDL keys for use in drug discovery. Journal of chemical information and computer sciences 42, 1273-1280 (2002).
22. A. Paszke et al., Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32, (2019).
23. M. Wang et al., Deep graph library: A graph-centric, highly-performant package for graph neural networks. arXiv preprint arXiv:1909.01315, (2019).