Computers and Chemical Engineering 204 (2026) 109425

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Predicting the equivalent alkane carbon number of oils using graph neural networks and quantum mechanical descriptors ☆

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

[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] BASF SE, Carl-Bosch-Strasse 38, 67056 Ludwigshafen, Germany
[d] Forschungszentrum Jülich GmbH, Institute of Energy and Climate Research IEK-10 – Energy Systems Engineering, 52425 Jülich, Germany
[e] JARA Center for Simulation and Data Science (CSD), 52056 Aachen, Germany

ARTICLE INFO

Keywords:
GNN
Oil
Hydrophobicity
EACN
QSPR

ABSTRACT

Oils are key components in personal care products, typically used as emollients to moisturize the skin and prevent dryness. The hydrophobicity of oils is crucial for designing tailor-made surfactant/oil/water (SOW) systems and can be quantified by the equivalent alkane carbon number (EACN). While various quantitative structure–property relationship (QSPR) models have been proposed, they require manual selection of informative descriptors. Recently, the largest publicly available data set with 121 EACN values was curated [Delforce et al. (2022), ACS Omega]. Herein, we apply graph neural networks (GNNs) for EACN prediction and propose a descriptor-augmented graph neural network (DA-GNN), that utilizes semi-empirical quantum mechanical (QM) descriptors. We first observe that the simple GNN model achieves highly accurate predictions. Afterwards, we find that incorporating QM descriptors enhances the accuracy of standard GNNs, resulting in improved EACN predictions by the DA-GNN model. However, the accuracy remains slightly lower than that of manually derived QSPR models, highlighting the need for sufficient data when training GNNs. We also experimentally measure the EACN of two cosmetic-type oils and analyze the predictive performance of both GNN models. The DA-GNN model accurately predicts their EACN values, making it practical for future applications in designing SOW systems.

1. Introduction

Oils are essential components of cosmetic formulations, as they can act as viscosity regulators, emollients, perfumes and solubilizers among others (Chuberre et al., 2019; Bhatnagar and Khurana, 2024; Akanny and Kohlmann, 2004). To design effective surfactant/oil/water (SOW) systems, knowledge of the hydrophobicity of the oil is necessary (Delforce et al., 2022; Lukowicz et al., 2018). The hydrophobicity of an oil can be quantified using the dimensionless equivalent alkane carbon number (EACN) (Cash et al., 1977; Lukowicz et al., 2015). The EACN corresponds to the number of carbon atoms in the linear alkane that exhibit lipophilicity equivalent to that of the given oil (Cash et al., 1977). For alkanes, the EACN is thus equal to the number of carbon atoms of the molecule. For non-alkane oils, in contrast, the EACN dependence on the molecular structure is not as straightforward. To experimentally determine the EACN, "fish diagrams" of the SOW system are constructed by varying the surfactant concentration and temperature (Queste et al., 2007). The "fish diagram" allows for the identification of the temperature T^* at the intersection of Winsor phases I to IV, which is then compared to the T^* values of n-alkanes to determine the EACN (Delforce et al., 2022). The EACN of the non-alkane corresponds to the carbon number derived from an interpolation function between the T^* values of adjacent n-alkanes. Notably, the learned interpolating function can also be extrapolated to negative values. The determination of EACN using "fish diagrams" consequently requires tedious experimental efforts.

The importance of the EACN and its lengthy experimental determination have motivated the development of predictive models based on structural molecular descriptors in the last years (Bouton et al., 2010; Lukowicz et al., 2018). Recently, Delforce et al. (2022) curated the largest available data set reported in the literature, which contains

☆ This article is part of a Special issue entitled: 'Machine learning in chemical engineering' published in Computers and Chemical Engineering.

* 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.2025.109425
Received 10 January 2025; Received in revised form 7 June 2025; Accepted 26 September 2025
Available online 30 September 2025

121 EACN values for various chemical species, such as n-alkanes, esters and ethers. They used the data set to develop a quantitative structure–property relationship (QSPR) model, referred to as NN-6N, for predicting the EACN of oils with semi-empirical quantum-mechanical (QM) descriptors, specifically σ-moments (Delforce et al., 2022). The choice of descriptors was motivated by their physical meaning and a selection step was performed. The selected descriptors were used to train a multi-layer perceptron (MLP) for EACN estimation (Delforce et al., 2022). Furthermore, the authors also proposed a graph machine (GM) approach (Goulon et al., 2006), that uses the 2D-structure of the oils as an input and extracts topological information regarding the molecule, i.e., topological descriptors, which are then used to train a MLP (Delforce et al., 2022). Their results showed that both QSPR models are capable of predicting the EACN of oils with high accuracy, with the GM outperforming the NN-6N model (Delforce et al., 2022). Notably, achieving this high level of predictive quality required manual adjustment of the QSPR models. Specifically, the authors added the number of carbon atoms (n) as an additional descriptor in the NN-6N model motivated by the definition of the EACN for n-alkanes. Additionally, special treatment was required for hexyl octanoate in the GM model, so the predictive quality of the QSPR models is affected by manual model adjustments.

In the last years, graph neural networks (GNNs), a deep learning technique, have been extensively applied in the field of molecular property prediction with very promising results (Zhou et al., 2020; Gilmer et al., 2017; Yang et al., 2019; Schweidtmann et al., 2020; Rittig et al., 2023a). For components of cosmetic formulations, such as surfactants, GNNs have recently been applied to predict the critical micelle concentration (CMC) and surface excess concentrations of both single species (Brozos et al., 2024b,c; Qin et al., 2021) and surfactant mixtures (Brozos et al., 2024a). Compared to classical QSPR methods, GNNs do not rely on pre-selected molecular descriptors; instead, they extract structure–property information within an end-to-end learning framework, thereby enhancing model flexibility and expressiveness (Gilmer et al., 2017; Li et al., 2024). Consequently, GNNs can learn additional structural information that may have been previously overlooked during the descriptor selection process. Herein, we investigate the application of GNNs for predicting the EACN of oil molecules.

We further consider augmentation of GNNs with semi-empirical QM descriptors. Recently, GNNs were coupled with QM descriptors, which add extra structural information, and yielded better results on data sets from computational chemistry, experimental physical chemistry, biophysics and experimental physiology, with up to 2000 entries (molecules) compared to pure GNNs (Li et al., 2024). To that extend, Biswas et al. (2023) showed that a coupled GNN with Abraham parameters outperformed the single GNN model in predicting critical properties and acentric factors of fluids. Furthermore, hybrid representations, i.e., GNNs coupled with molecular descriptors, outperformed baseline GNN models in subsequent studies at predicting regio-selectivity and activation energies, especially when only limited experimental data were available (Guan et al., 2021; Stuyver and Coley, 2022). However, it is noted that careful selection of the QM descriptors should take place, as QM descriptors uncorrelated to the target property can introduce unwanted noise to the model (Li et al., 2024), i.e., in that case the model needs to learn the irrelevance of these descriptors. In addition, calculating QM descriptors is computationally expensive and time-consuming since the descriptors must be calculated first for the molecules of interest, significantly increasing training and prediction costs. Moreover, the descriptor values can vary across different software versions. Given that the existing databases of EACNs consist of only a few hundred molecules, we also explore the augmentation of GNNs with semi-empirical QM descriptors.

We first develop a standard GNN model using the publicly available dataset collected by Delforce et al. (2022), which contains 121 molecules. We then augment the standard GNN model with semi-empirical QM descriptors, specifically the σ-moments, as calculated by Delforce et al. (2022), to assess whether they can enhance predictive accuracy of the standard GNN. Next, we compare the performance of the standard GNN model with that of the descriptor-augmented model. Our findings indicate that the descriptor-augmented GNN model outperforms the standard GNN model. Afterwards, the descriptor-augmented GNN model is compared with two state-of-the art models developed by Delforce et al. (2022) and the exact model predictions are analyzed.

We further measure the EACN values of two commercial emollients (oils). To apply the descriptor-augmented GNN model to these oils, we would need to perform computationally intensive calculations of their semi-empirical QM descriptors using COSMO-RS, similar to those performed by Delforce et al. (2022). To circumvent this, we explore replacing the existing semi-empirical QM descriptors with newly ones obtained from the commercial software COSMOquick (Loschen and Klamt, 2012). That is, we retrain the descriptor-augmented GNN model using the newly calculated descriptors and make EACN predictions for the two commercial oils. We then compare these predictions with the experimental values as well as with the predictions from the standard GNN model.

We structure this work as follows: In Section 2, we provide a description of the data sets used, the developed GNN models, and the calculations of the semi-empirical QM descriptors. In Section 3, we analyze and discuss the predictive quality of the GNNs and compare it to state-of-the-art QSPR models. Finally, we summarize our findings in Section 4.

2. Methods

We first describe the data set used for model development (Section 2.1) and the complementary data set (Section 2.2). We then shortly describe the architectures of the GNN model and of the descriptor augmented one (Section 2.3). We then outline the hyperparameters and the model selection methodology (Section 2.4). We conclude this section by describing the newly calculated semi-empirical QM descriptors (Section 2.5).

2.1. Data set and data scaling

The data set used in this study was recently compiled by Delforce et al. (2022) and contains experimental EACN values for 121 chemical species. We adhere to the same train-test split proposed by the authors (Delforce et al., 2022), which consists of a training set of 111 compounds and a test of 10 compounds (Delforce et al., 2022). For a comprehensive analysis of the data set, we refer reader to the original work of Delforce et al. (2022). As commonly done in ML, we scale the EACN values using the decimal logarithmic scale to obtain a normalized distribution. Since oils can have negative EACN values, with our data set containing values as low as −4, we initially shift all EACN values upward by 4.0001 before applying the logarithmic transformation. We use the logarithmic EACN values to train the model, but report all error metrics in absolute EACN values. Note that Delforce et al. (2022) did not report any data scaling prior to model training.

2.2. Complementary data set

We experimentally measure the EACN values of two more oils. Specifically, our complementary data set includes two single-species cosmetic-type oils (emollients): Cetiol® OE (purity ≥ 96%) and Eutanol® G 16 (purity ≥ 97%). In deviation from using the fish diagrams, we determine the EACN of the 2 oils from the phase inversion temperature (PIT) value of the $C_{10}EO_6$/Oil/2×10^{-2} M NaCl aq. system according to the CAPICO method developed by Förster et al. (1994). Here, we use an oil:water ratio of 1:1 and 10% total emulsifier content and a calibration based on a series of n-alkanes.

Table 1
Atom features used in the molecular graph representation. All features are implemented as one-hot-encoding.

Feature	Description	Dimension
Atom type	Atom type (C, Cl, N, O)	4
Is in a ring	If the atom is part of a ring	1
Is aromatic	If the atom is part of an aromatic system	1
Hybridization	sp, sp^2, sp^3	3
Chirality	Unspecified, clockwise, counter clockwise	3
# bonds	Number of bonds the atom is involved in	5
# Hs	Number of bonded hydrogen atoms	4
Total		**21**

Table 2
Edge features used in the molecular graph representation. All features are implemented as one-hot-encoding.

Feature	Description	Dimension
Bond type	Single, double, triple or aromatic	4
Is in a ring	Is the bond part of a ring ?	1
Conjugated	Is the bond conjugated ?	1
Stereo	None or E/Z	3
Total		**9**

2.3. Baseline and descriptor-augmented GNN models

We use the collected data to develop GNNs for predicting the EACN of oils. In the GNNs, each oil molecule is represented as a molecular graph, where each atom corresponds to a node and each bond corresponds to an edge. A feature vector containing chemical information is appended to both each node and each edge. The features used in this work are detailed in Tables 1 and 2, and are based on our previous works (Brozos et al., 2024b,c,a; Rittig et al., 2023a,b). The molecular graphs of the oils then enter the GNN, where a graph convolutional layer is employed to each node's feature vector, also referred to as hidden state, which gets updated with structural information pertaining to its neighborhood (Gilmer et al., 2017; Schweidtmann et al., 2020). Subsequently, the updated hidden states are encoded into a vector known as the molecular fingerprint (FP), which represents a learned molecular representation (Gilmer et al., 2017; Schweidtmann et al., 2020). Then the FP is fed into a MLP to predict the property of interest, i.e., the EACN value. We denote this model as **B-GNN** (baseline GNN).

We also explore the option to augment the GNN model with descriptors. In this study, the descriptors from the work of Delforce et al. (2022), specifically the calculated σ-moments (M_0^X, M_2^X and M_3^X) and the number of carbon atoms (n) are used. The descriptors are first normalized to a range between 0 and 10 to align with the magnitude of the FP. We select these descriptors because they have been shown to highly correlate with the EACN, making it particularly interesting to examine whether they can enhance the accuracy of a baseline GNN model. We note that these 4 descriptors correspond to the inputs of the NN-6N model developed by Delforce et al. (2022), allowing for a direct comparison. These descriptors are then concatenated to the FP. The descriptor-augmented FP is then fed into a MLP to predict the EACN. We refer this model as **DA-GNN** (descriptor-augmented GNN).

2.4. Hyperparameter tuning and model selection

The GNN models are implemented in Python using PyTorch Geometric (PyG) (Fey and Lenssen, 2019). We conduct hyperparameter tuning separately for both the B-GNN and DA-GNN models. That is, we train each model on 30 different, randomly selected validation sets. The size of the validation set is kept constant at 25 data points, which represents 21% of the entire data set size and approximately 23% of the training data set size. Given the small sizes of both the training and validation sets, noisy results may occur, i.e. the performance metrics may fluctuate significantly. To mitigate their impact, we select only the 10 models out of the 30 that exhibit the lowest validation error. We utilize the average root mean square error (RMSE) of these 10 models to determine the optimal hyperparameter combination for each GNN model. The selected hyperparameters are presented in Table 3. To further improve the predictive capability of the GNN models, we perform ensemble training (Dietterich, 2000) and average the predictions of the 10 GNN models to obtain a final EACN prediction.

Table 3
Hyperparameters of the GNN models investigated through a grid search. The hyperparameter dimensions refers to the size of the molecular fingerprint and the size of the MLP.

Hyperparameter	Range	DA-GNN	B-GNN	CQ-DA-GNN
Initial learning rate	(0.005, 0.01, 0.05)	0.005	0.01	0.005
Dimensions	(16, 32, 64)	32	32	64
Batch size	(4, 8)	8	8	4
Number of MLP layers	3	3	3	3
Activation function	ReLU	ReLU	ReLU	ReLU
Maximum epochs	300	300	300	300
Early stopping patience	40	40	40	40
Learning rate decay	0.8	0.8	0.8	0.8
Patience	3	3	3	3
Optimizer	Adam	Adam	Adam	Adam

2.5. COSMOquick calculations and descriptor selection

As COSMO-RS calculations for the molecules in the complementary data set can be computationally expensive, we explore the possibility of replacing them with semi-empirical QM descriptors obtained from the commercial software COSMOquick (Loschen and Klamt, 2012), which utilizes fragments from a database of 40,000 molecules generated at the BP86/def-TZVP level (Hornig and Klamt, 2005; Pattanaik et al., 2023). The advantage of using COSMOquick is that requires substantially less time compared to COSMO-RS. Initially, we replace the M_0^X, M_2^X and M_3^X descriptors calculated by Delforce et al. (2022) using COSMO-RS with those calculated using COSMOquick. We observe a deviation between the two sets of calculated descriptors, particularly in the case of M_3^X. Notably, the predictive performance of the DA-GNN model significantly decreases as a result. Therefore, we decide to calculate further semi-empirical QM descriptors using the commercial software COSMOquick and perform a descriptor selection step.

To select the most suitable semi-empirical QM descriptors out of the newly calculated descriptor set and reduce redundant information, we perform several pre-processing steps similar to those in previous works (Comesana et al., 2022; Seddon et al., 2022). For the selection, we only consider the 111 oil molecules in the training set (cf. Section 2.1). First, we identify and remove descriptors with low variance, such as those that have the same value in more than 95% of the data entries. Next, we eliminate descriptors with a Spearman coefficient lower than ±0.2 with respect to the EACN (Concise, 2008). In the final step, we calculate the Spearman coefficients for pairs of the remaining descriptors and remove one descriptor from each pair with a threshold of 0.6. Interestingly, 2 out of the 4 remaining semi-empirical QM descriptors are M_0^X (surface area of the molecule) and M_3^X (asymmetry of the σ-profile), similar to the findings of Delforce et al. (2022). The other 2 descriptors are the chemical potential in the gas phase (μ_i^{gas}) and the 2nd σ-moment that describes H-bonding donor capabilities ($M_{don_2}^{hb}$). A comparison of the descriptors utilized in this work with those used by Delforce et al. (2022) can be found in Table 4.

3. Results

We begin this section by analyzing the prediction accuracy of the B-GNN and DA-GNN models and comparing their performance (Section 3.1). Afterwards, we compare our findings with previous works (Section 3.2). We conclude this section by investigating the model performance on the complementary test set (Section 3.3).

Table 4
A comparison between the semi-empirical QM descriptors used in this work and those employed in the work of Delforce et al. (2022).

Descriptors in Delforce et al. (2022)	Descriptors in this work
M_0^X	M_0^X
M_2^X	M_3^X
M_3^X	$M_{don_2}^{hb}$
n	μ_i^{gas}

Table 5
Comparison between experimental EACN values and predictions from four models. The best model prediction for each oil molecule is highlighted in bold. For the two here developed models the standard deviation is also given.

Molecule	Exp.	Model predictions			
		B-GNN	DA-GNN	GM-5N	NN-6N
Hemisqualene	14.8	14.43 ± 0.77	14.17 ± 0.63	**14.7**	15
Isododecane	11.7	9.54 ± 1.06	10.34 ± 0.39	**11.9**	13.6
Dioctylether	10.3	10.96 ± 0.72	9.89 ± 0.58	10.8	**10.5**
Octyl octanoate	8.1	6.42 ± 0.43	6.58 ± 0.18	**8.7**	8.8
Isopropyl myristate	7.3	7.13 ± 0.71	**7.35** ± 0.35	7.6	6.5
Caryophyllene	6	4.03 ± 1.25	4.87 ± 0.67	6.6	**5.8**
Limonene	1.8	0.7 ± 0.73	0.49 ± 0.55	**2.5**	**2.5**
Linalyl acetate	−0.9	−1.61 ± 0.85	−0.64 ± 0.64	−0.7	**−0.8**
Rose oxide	−1.7	−0.13 ± 0.85	−1.22 ± 0.51	−2.5	**−1.8**
β-Ionone	−1.9	−0.92 ± 0.5	**−2.03** ± 0.38	−2.5	−2.3
RMSE		1.31	0.9	**0.51**	0.74
$\mathbf{R}^2$		0.956	0.987	**0.992**	0.986

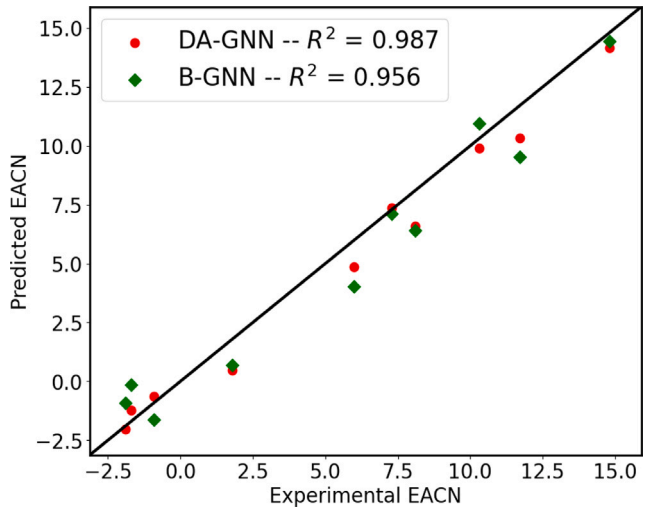

Fig. 1. Parity plot on the ensemble of the two GNN models on the test set.

3.1. Predictive accuracy of the B-GNN model and comparison with the DA-GNN model

We first evaluate the performance of the B-GNN model on the test set described in Section 2.1. The B-GNN model achieves an RMSE of 1.31 and an R^2 of 0.956. The predictions made by the B-GNN model on the test set are illustrated in a parity plot in Fig. 1, showing the rather good fit. Considering the small number of training examples and the prediction accuracy of absolute experimental EACN values, we can conclude that the B-GNN model effectively extracts the necessary structural information to perform highly accurate predictions.

Afterwards, we assess the predictive performance of DA-GNN model on the same test set. Here, we find that the DA-GNN model exhibits an RMSE of 0.9 and an R^2 of 0.987, i.e., the B-GNN model with semi-empirical QM descriptors substantially decreases the RMSE by approximately 35%. Fig. 1 also shows this improvement for most but not all oils. Therefore, we can conclude that the additional structural information provided by the descriptors enhances the accuracy of predictions, positively impacting the overall model performance. These results are in par with the observations made recently by Li et al. (2024), which indicate that inclusion of QM descriptors benefits data sets with up to 2000 data points.

3.2. Comparison with previous works

Next, we compare our work with that of Delforce et al. (2022) and their two developed models, which achieved state-of-the-art performance on the exact same data set. From our work, we examine the performance only of the DA-GNN model. As described in Section 1, the authors proposed two models; a GM that learns from the 2D structures of the oil, denoted by them as GM-5N, and a descriptor-based neural network, referred to as NN-6N (Delforce et al., 2022). According to the original work, the GM-5N model exhibited an RMSE of 0.51 and an R^2 of 0.992, while the NN-6N had an RMSE of 0.74 and an R^2 of 0.986 on the test set (Delforce et al., 2022). As the comparison in Table 5 shows, the DA-GNN model exhibits an RMSE of 0.9 and an R^2 of 0.987. Therefore, the RMSE of the DA-GNN is slightly higher than that of the NN-6N and significantly higher than that of the GM-5N model. Furthermore, we observe that the R^2 values of all three models are almost identical.

Furthermore, in Table 5, we present the exact model predictions of the DA-GNN model, alongside the predictions from the NN-6N and GM-5N models for each of the 10 oil molecules in the test set (Delforce et al., 2022). Notably, the DA-GNN model exhibits the lowest deviation from the experimental values for 2 out of the 10 molecules, despite having the highest RMSE among the three models. For these 2 molecules, the DA-GNN predicted EACN values almost perfectly match the experimental ones. Furthermore, the highest deviation is observed for octyl octanoate. When compared only with the NN-6N model, the number of predictions with the lowest deviation increases to 3, and interestingly the predictions from the two models differ significantly. Thus, the learned representation, i.e., FP, influences the predictions of the DA-GNN model, and they are not solely derived from the structural information provided by the molecular descriptors. However, in these low data regimes the QSPR models outperform both GNN model variations in terms of RMSE, as in previous studies (Sun et al., 2020). We hypothesize two reasons for this result: (1) GNNs have more parameters and thus require more data to train effectively, and (2) QSPR models leverage molecular descriptors that incorporate domain-specific priors, enabling better generalization in data-scarce regimes. Overall, all models discussed so far accurately predict the EACN values of oil molecules, with the three models using descriptors outperforming the baseline GNN model. For practical applications, we expect all four models to be useful, as their predictions can provide valuable guidance in formulation design. In terms of computational efficiency, we expect all four models to perform rapid inference on new compounds, usually within milliseconds; while, for training, the GNN models typically require more computational time for training. Here, we trained the GNN models on a local machine with a CPU in a couple of hours.

3.3. Model retraining and performance on the complementary data set

We further experimentally measure and curate a complementary data set following the procedure described in Section 2.2. To perform predictions on the complementary data set using the DA-GNN, computationally expensive calculations of their semi-empirical QM descriptors with COSMO-RS are necessary. To avoid this step, we calculate new semi-empirical QM descriptors for both the entire data set and the complementary data set with the commercial software COSMO-quick (Loschen and Klamt, 2012), a process described thoroughly in Section 2.5. Afterwards we utilize the newly calculated semi-empirical

Table 6
Comparison of experimental EACN values with predictions from the two GNN models developed in this study on the complementary test set. In each case the standard deviation is also given.

Commercial oil	Experimental EACN value	Model predictions	
		CQ-DA-GNN	B-GNN
Cetiol® OE	14	11.19 ± 0.73	10.96 ± 0.72
Eutanol® G 16	−0.76	−0.97 ± 1.83	10.96 ± 2.28
RMSE		**1.99**	**8.56**

QM descriptors to retrain the descriptor-augmented GNN model, which we now refer to as **CQ-DA-GNN** (COSMOquick descriptor-augmented GNN). We conduct a hyperparameter tuning step following the methodology outlined in Section 2.4 and report the optimal hyperparameters in Table 3. The CQ-DA-GNN model exhibits an RMSE of 0.96 and R^2 of 0.977 on the test set of 10 species, discussed in the previous section. Both metrics are identical to those exhibited by the DA-GNN model. We note that in CQ-DA-GNN model, the number of carbon atoms is not used as an additional descriptor.

The predictions of the CQ-DA-GNN and B-GNN models on the complementary test set are presented in Table 6. Notably, the B-GNN model significantly overpredicts the EACN value of Eutanol® G 16, resulting in a very high RMSE value of 8.56. In contrast, the CQ-DA-GNN model exhibits an RMSE of 1.99. Furthermore, the predictions of the CQ-DA-GNN are closer to the experimental value in both cases. This observation aligns with the findings in Section 3.1, where the descriptor-augmented GNN model outperformed the baseline GNN. Notably, the B-GNN model predicts the same EACN value for both oils. This is due to an averaging of the ten predicted EACN values rather than the model treating both oils as the same, which is also evident from the different standard deviations presented in Table 6. Since the structures of the two oils are very similar, i.e., dicaprylyl ether vs. 2-Hexyl-1-decanol, the B-GNN model predicts EACN values that are close to each other. However, the difference between the two experimental values signifies the difficulty in estimating the EACN of non-n-alkanes. It should also be noted that the EACN values of the complementary test set were not determined with the "fish diagram", as the majority of the EACN values on the training set. Therefore, some model deviations are expected due to the different experimental method used.

4. Conclusion

We investigate the applicability of GNNs for predicting the EACN values of cosmetic oils. A recently collected, publicly available data set containing EACN values for 121 oil molecules was used for model training and evaluation. We find that a standard GNN model can predict the EACN values of oil molecules with a high accuracy ($R^2 \geq 0.94$). Due to the small size of the data set, we explore a descriptor-augmented GNN for EACN prediction, following the work of Li et al. (2024). As descriptors, we initially use σ-moments and the number of carbon atoms. Our results show that the descriptor-augmented GNN outperforms the simple GNN model by 35% in terms of the RMSE, indicating that the additional structural information encapsulated in the descriptors enhances the predictive capability of the GNN model.

Afterwards, we compare the descriptor-augmented GNN with two previously developed QSPR models, one of which utilizes the same semi-empirical QM descriptors as inputs. We find that while the descriptor-augmented GNN does not match their RMSE, it exhibits a similar R^2. Thus, although the GNN models can extract structural information from the molecules to predict the EACN, their predictions still fall short compared to state-of-the-art QSPR models in these low-data regimes. This highlights the potential of GNN and the need for additional high-fidelity data. In essence, the current state-of-the-art models developed by Delforce et al. (2022) cannot be yet outperformed and hence they are more suitable for current usage.

Furthermore, we experimentally determine the EACN values for a complementary test set containing 2 cosmetic-type oils. We compute new semi-empirical QM descriptors using the commercial software COSMOquick (Loschen and Klamt, 2012) and perform a descriptor selection step. We then retrain the descriptor-augmented GNN with the newly calculated semi-empirical QM descriptors and perform predictions on the complementary test set. We observe very good agreement between the predicted and measured EACN values for the descriptor-augmented GNN, indicating the usefulness in industrial applications. Future work should investigate potential differences in EACN values resulting from different experimental methods. Systematic measurement of additional reliable EACN values for species not present in the existing database can enhance the applicability domain of the predictive models. Finally, the impact of impurities on the EACN should be also systematically addressed.

CRediT authorship contribution statement

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

Declaration of competing interest

C. Brozos, S. Bhattacharya and E. Akanny were funded by the BASF Personal Care and Nutrition GmbH which has commercial interest in the sector.

Acknowledgments

C. Brozos, S. Bhattacharya and E. Akanny were funded by the BASF Personal Care and Nutrition GmbH. P. Eiden was funded by BASF SE. J. G. Rittig and A. Mitsos acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Germany – 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).

Data availability

Data will be made available on request.

References

Akanny, E., Kohlmann, C., 2004. Predicting tactile sensory attributes of personal care emulsions based on instrumental characterizations: A review. Int. J. Cosmet. Sci..

Bhatnagar, S., Khurana, S., 2024. Oils and Fats As Raw Materials for Industry. John Wiley & Sons, Ltd, pp. 145–167, Chapter 6.

Biswas, S., Chung, Y., Ramirez, J., Wu, H., Green, W.H., 2023. Predicting critical properties and acentric factors of fluids using multitask machine learning. J. Chem. Inf. Model. 63, 4574–4588.

Bouton, F., Durand, M., Nardello-Rataj, V., Borosy, A.P., Quellet, C., Aubry, J.-M., 2010. A QSPR model for the prediction of the fish-tail temperature of CiE4/Water/Polar hydrocarbon oil systems. Langmuir 26, 7962–7970.

Brozos, C., Rittig, J.G., Akanny, E., Bhattacharya, S., Kohlmann, C., Mitsos, A., 2024a. Predicting the temperature-dependent CMC of surfactant mixtures with graph neural networks. https://arxiv.org/abs/2411.02224, (Accessed 8 November 2024).

Brozos, C., Rittig, J.G., Bhattacharya, S., Akanny, E., Kohlmann, C., Mitsos, A., 2024b. 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., 2024c. Predicting the temperature dependence of surfactant CMCs using graph neural networks. J. Chem. Theory Comput. 20, 5695–5707.

Cash, L., Cayias, J., Fournier, G., Macallister, D., Schares, T., Schechter, R., Wade, W., 1977. The application of low interfacial tension scaling rules to binary hydrocarbon mixtures. J. Colloid Interface Sci. 59, 39–44.

Chuberre, B., Araviiskaia, E., Bieber, T., Barbaud, A., 2019. Mineral oils and waxes in cosmetics: an overview mainly based on the current European regulations and the safety profile of these compounds. J. Eur. Acad. Dermatol. Venereol. 33, 5–14.

Comesana, A.E., Huntington, T.T., Scown, C.D., Niemeyer, K.E., Rapp, V.H., 2022. A systematic method for selecting molecular descriptors as features when training models for predicting physiochemical properties. Fuel 321, 123836.

2008. The Concise Encyclopedia of Statistics. Springer New York, New York, NY, pp. 502–505.

Delforce, L., Duprat, F., Ploix, J.-L., Ontiveros, J.F., Goussard, V., Nardello-Rataj, V., Aubry, J.-M., 2022. Fast prediction of the equivalent alkane carbon number using graph machines and neural networks. ACS Omega 7, 38869–38881.

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).

Förster, T., Rybinski, W.V., Tesmann, H., Wadle, A., 1994. Calculation of optimum emulsifier mixtures for phase inversion emulsification*. Int. J. Cosmet. Sci. 16, 84–92.

Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E., 2017. Neural message passing for quantum chemistry. In: International Conference on Machine Learning.

Goulon, A., Duprat, A., Dreyfus, G., 2006. Graph Machines and their Applications To Computer-Aided Drug Design: A New Approach To Learning from Structured Data. Unconventional Computation, Berlin, Heidelberg, pp. 1–19.

Guan, Y., Coley, C.W., Wu, H., Ranasinghe, D., Heid, E., Struble, T.J., Pattanaik, L., Green, W.H., Jensen, K.F., 2021. Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. Chem. Sci. 12, 2198–2208.

Hornig, M., Klamt, A., 2005. COSMOfrag: A novel tool for high-throughput ADME property prediction and similarity screening based on quantum chemistry. J. Chem. Inf. Model. 45, 1169–1177.

Li, S.-C., Wu, H., Menon, A., Spiekermann, K.A., Li, Y.-P., Green, W.H., 2024. When do quantum mechanical descriptors help graph neural networks to predict chemical properties? J. Am. Chem. Soc. 146, 23103–23120.

Loschen, C., Klamt, A., 2012. COSMOquick: A novel interface for fast σ -profile composition and its application to COSMO-RS solvent screening using multiple reference solvents. Ind. Eng. Chem. Res. 51, 14303–14308.

Lukowicz, T., Benazzouz, A., Nardello-Rataj, V., Aubry, J.-M., 2015. Rationalization and prediction of the equivalent alkane carbon number (EACN) of polar hydrocarbon oils with COSMO-RS σ -moments. Langmuir 31, 11220–11226, PMID: 26397810.

Lukowicz, T., Illous, E., Nardello-Rataj, V., Aubry, J.-M., 2018. Prediction of the equivalent alkane carbon number (EACN) of aprotic polar oils with COSMO-RS sigma-moments. Colloids Surfaces A: Physicochem. Eng. Asp. 536, 53–59, Special issue on Formula VIII.

Pattanaik, L., Menon, A., Settels, V., Spiekermann, K.A., Tan, Z., Vermeire, F.H., Sandfort, F., Eiden, P., Green, W.H., 2023. ConfSolv: Prediction of solute conformer-free energies across a range of solvents. J. Phys. Chem. B 127, 10151–10170.

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.

Queste, S., Salager, J., Strey, R., Aubry, J., 2007. The EACN scale for oil classification revisited thanks to fish diagrams. J. Colloid Interface Sci. 312, 98–107, In Memory of Professor Hironobu Kunieda.

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., Gao, Q., Dahmen, M., Mitsos, A., Schweidtmann, A.M., 2023b. Machine Learning and Hybrid Modelling for Reaction Engineering: Theory and Applications. R. Soc. Chem..

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.

Seddon, D., Müller, E.A., Cabral, J.T., 2022. Machine learning hybrid approach for the prediction of surface tension profiles of hydrocarbon surfactants in aqueous solution. J. Colloid Interface Sci. 625, 328–339.

Stuyver, T., Coley, C.W., 2022. Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability. J. Chem. Phys. 156, 084104.

Sun, X., Krakauer, N.J., Politowicz, A., Chen, W.-T., Li, Q., Li, Z., Shao, X., Sunaryo, A., Shen, M., Wang, J., Morgan, D., 2020. Assessing graph-based deep learning models for predicting flash point. Mol. Inform. 39, 1900101.

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.

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.