scientific reports

Check for updates

OPEN

Structure-based modeling of critical micelle concentration (CMC) of anionic surfactants in brine using intelligent methods

Danial Abooali[1]✉ & Reza Soleimani[2]✉

Critical micelle concentration (CMC) is one of the main physico-chemical properties of surface-active agents, also known as surfactants, with diverse theoretical and industrial applications. It is influenced by basic parameters such as temperature, pH, salinity, and the chemical structure of surfactants. Most studies have only estimated CMC at fixed conditions based on the surfactant's chemical parameters. In the present study, we aimed to develop a set of novel and applicable models for estimating CMC of well-known anionic surfactants by considering both the molecular properties of surfactants and basic affecting factors such as salinity, pH, and temperature as modeling parameters. We employed the quantitative-structural property relationship technique to employ the molecular parameters of surfactant ions. We collected 488 CMC values from literature for 111 sodium-based anionic surfactants, including sulfate types, sulfonate, benzene sulfonate, sulfosuccinate, and polyoxyethylene sulfate. We computed 1410 optimized molecular descriptors for each surfactant using Dragon software to be utilized in the modelling processes. The enhanced replacement method was used for selecting the most effective descriptors for the CMC. A multivariate linear model and two non-linear models are the outputs of the present study. The non-linear models were produced using two robust machine learning approaches, stochastic gradient boosting (SGB) trees and genetic programming (GP). Statistical assessment showed highly applicable and acceptable accuracy of the newly developed models ($R_{SGB}^2 = 0.999395$ and $R_{GP}^2 = 0.954946$). The ultimate results showed the superiority and greater ability of the SGB method for making confident predictions.

Abbreviations

AE	Absolute error
ANFIS	Adaptive neuro-fuzzy inference system
ANNs	Artificial neural networks
BEHp2	Highest eigenvalue no. 2 of Burden matrix/weighted by atomic polarizabilities
CEOR	Chemical enhanced oil recovery
CIC2	Complementary information content (neighborhood symmetry of 2-order)
CMC	Critical micelle concentration
D	Dipole moment
EEig12x	Eigenvalue 12 from edge adj. matrix weighted by edge degrees
E_{HOMO}	Energy of the highest occupied molecular orbital
E_{LUMO}	Energy of the lowest unoccupied molecular orbital
EOR	Enhanced oil recovery
ERM	Enhanced replacement method
E_t	Total energy of molecule
f-I_{BAL}	Balaban distance connectivity index
FSR	Forward stepwise regression
G3s	3Rd component symmetry directional WHIM index/weighted by atomic electro-topological states

[1]Young Researchers and Elite Club, Central Tehran Branch, Islamic Azad University, Tehran, Iran. [2]Department of Chemical Engineering, Faculty of Chemical Engineering, Tarbiat Modares University, P.O. Box 14115-143, Tehran, Iran. ✉email: danial.abooali@gmail.com; soleimanire@gmail.com

GA-MLR	Genetic algorithm multivariate linear regression
GB	Gradient boosting
GFA	Genetic function approximation
GP	Genetic programming
HGP	Hydrophobic group position
ICA	Imperialist competitive algorithm
KH0	Kier and Hall molecular connectivity index of zero-th order
KH1	Kier and Hall molecular connectivity index of first-order
KS3	Kier shape index of third-order
Lop	Lopping centric index
MAE	Mean absolute error
MM2	Molecular mechanics
n	Number of samples in the dataset
N_T	Total atom number
P	Pressure
PSO	Particle swarm optimization
Q^2_{boot}	LOO cross-validation squared correlation coefficient of bootstrapping
Q^2_{ext}	Squared correlation coefficient of external-validation
Q^2_{LNO}	Leave-N-out cross-validation squared correlation coefficient
Q^2_{LOO}	Leave-one-out cross-validation squared correlation coefficient
Q^2_{yi}	Y-randomization LOO cross-validation squared correlation coefficient
$Q_{C\text{-}max}$	Maximum net atomic charges on carbon atom
QSPR	Quantitative-structural property relationship
R^2	Squared correlation coefficient
R^2_{boot}	Squared correlation coefficient of bootstrapping test
R^2_{ext}	Squared correlation coefficient of external-validation test
R^2_{yi}	Squared correlation coefficient of y-randomization test
RA^{-1}	Reciprocal of randic index
RM	Replacement method
RMSD	Root-mean-square deviation
RMSECV	Root-mean-square error of cross-validation
RSD	Residual standard deviation
RNC	Relative number of carbon atoms
S_{eq}	NaCl equivalent salinity
SGB	Stochastic gradient boosting
SVM	Support vector machine
T	Temperature
TDIP	Total dipole moment
WHIM	Weighted holistic invariant molecular
WI	Wiener number
$y_i^{cal.}$	Predicted dependent variable
$y_i^{exp.}$	Experimental dependent variable
$\overline{y}^{exp.}$	Average of experimental dependent variable
ΔH_f	Molar heat of formation
Π	Octanol/water partition coefficient

The industrial applications of surfactant solutions demonstrate the growing importance of these systems in everyday life[1]. Surfactants are utilized in various industries, including enhanced oil recovery (EOR)[2], cleaners and detergents[3,4], emulsifiers and dispersing agents[5], foods[6], coatings[7], and many other chemical, petroleum, and pharmaceutical processes[1].

Surfactants are amphiphilic compounds consisting of hydrophilic (polar head) and hydrophobic (nonpolar tail) parts. Due to this unique structure, surfactants tend to accumulate at the surface of solutions such as water or brine. Once the surface is saturated with surfactant molecules, the remaining particles accumulate in the bulk and form micelles[8].

Among different types of surfactants, anionic surfactants are known for their high foaming properties, and some industries such as chemical EOR (CEOR), detergents, and cleaners, often use them in specific applications. In the present study, we investigated several anionic surfactants to better understand their behavior and properties.

Critical micelle concentration (CMC) is an important property of surfactants that has been investigated in many theoretical and experimental studies. The CMC is defined as the maximum concentration of a surfactant at which micelles do not form or the concentration at which micelles begin to form[8,9].

In concentrations larger than CMC, the solution is considered micellar and exhibits different behavior from a dilute solution (e.g., a solution with concentration less than the CMC). From an industrial and economic point of view, operating surfactant systems at the CMC often results in specific efficiencies. In addition, several theoretical and thermodynamic studies have been carried out to estimate various properties of surfactant systems based on the same properties at the CMC. A good example in this area is the estimation of the surface tension of a surfactant solution from the surface excess concentration at the CMC[8,9]. The CMC is a straightforward way to assess the behavior of surfactant solutes on surfaces and colloids, making it a valuable tool for evaluating their potential industrial and pharmaceutical applications[10,11]. In certain situations, it is desirable for surfactants to have a low CMC, such as when they are used to dissolve hydrophobic drugs in micellar cores with minimal surfactant quantities[10,12]. Additionally, in applications like foaming, wetting, and hard surface cleaning, where a low product surface tension is often desired, micelles act as surfactant reservoirs above the CMC, allowing for product dilution without significant changes in surface tension. On the other hand, in cases like membrane protein extraction, a high CMC is preferred since the extraction efficiency typically plateaus at around four times the CMC of the surfactant due to self-association[10,13].

Due to the numerous applications of CMC, knowledge about the values of this specific property is essential under different conditions. Experimental measurements are a reliable way to access to accurate values. However, conducting experiments in laboratories is not always simple, especially at high temperatures and pressures. In some cases, experimental measurements are expensive and/or time-consuming and may involve uncertainties about impurities, possible decompositions, etc. The application of estimation methods and mathematical models may be effective in this area. Empirical modeling, as a famous method, and different mathematical-statistical algorithms are available for developing computational correlations. Well-known tools such as genetic programing (GP), artificial neural networks (ANNs), particle swarm optimization (PSO), adaptive neuro-fuzzy inference system (ANFIS), support vector machines (SVMs), stochastic gradient boosting (SGB) trees, etc., are applied.

In order to estimate the properties of chemical compounds, molecular based approaches such as group-contribution and quantitative structure–property relationship (QSPR) are preferred[14]. In the group-contribution method, properties of chemical compounds are estimated by analyzing different parts of their molecular structures, such as functional groups, singular and multiple bonds, etc. This is an interesting method that can sometimes achieve high accuracy. However, there are some disadvantages, such as its limited applicability to certain isomers as well as chemical compounds with novel structure.

QSPR is another estimation approach in which the considered property (objective function) is estimated from a number of chemical parameters of the components called ”molecular descriptors”[15]. The molecular descriptors relate solely to the molecular structures of components and are calculated by applying certain mathematical rules. One of the important advantages of a QSPR model is the ability to estimate the properties of newly designed chemical compounds only solely from their molecular descriptors. In this study, the QSPR technique was applied to produce novel models for CMC as functions of molecular descriptors.

There are several mathematical models for estimating the CMC of anionic surfactants. In 1953, Klevens[16] proposed a relationship between the CMC and the number of carbon atoms in the surfactant tail (N) as follows:

$$\log(\mathrm{CMC}) = \mathrm{A} - \mathrm{BN} \tag{1}$$

A and B are constants for homologue series of surfactants under fixed condition. This model is simple, but it is valid for fixed conditions and structurally simple surfactants.

In the main studies of CMC modelling, the QSPR approach has been used. Huibers et al.[17] developed a multi-variable linear model based on QSPR from a data set of 119 anionic surfactants at 40 °C. The model is as follows:

$$\begin{aligned} \log_{10}(\mathrm{CMC}) = {} & (1.89 \pm 0.11) - (0.314 \pm 0.01)\,\mathrm{t\text{-}sum\text{-}KH0} - (0.034 \pm 0.003)\mathrm{TDIP} \\ & - (1.45 \pm 0.18)\,\mathrm{h\text{-}sum\text{-}RNC} \end{aligned} \tag{2}$$

In this equation, the descriptor “t-sum-KH0”, which is the zeroth-order Kier and Hall molecular connectivity index, is considered as a variable for the hydrophobic part (tail) of the surfactant. This parameter is related to the molecular volume and surface area. “TDIP” represents the total dipole moment of the surfactant and is a descriptor for the entire molecule.“h-sum-RNC” is the relative number of carbon atoms in the hydrophilic moiety (head) and reflects the diversity of head group structures[18].

Huibers et al.[17] also developed a multi-variable linear correlation for the types of sulfate and sulfonates using 66 data points at 40 °C:

$$\begin{aligned} \log_{10}(\mathrm{CMC}) = {} & (2.42 \pm 0.07) - (0.537 \pm 0.009)\,\mathrm{KH1} - (0.019 \pm 0.002)\,\mathrm{KS3} \\ & + (0.096 \pm 0.005)\,\mathrm{HGP} \end{aligned} \tag{3}$$

KH1 is the first-order Kier and Hall molecular connectivity index, which is a parameter that correlates with molecular volume and surface area. KS3 is the of third-order Kier shape index that is related to molecular shape. HGP determines the carbon number attached to the hydrophilic moiety and is located on the longest chain of the surfactant’s molecule[17,18].

Another linear model was produced by Jalali-Heravi and Konouz[19] using 31 anionic surfactants (27 alkyl sulfates and 4 alkane sulfonates) at 40 °C. The correlation was presented as follows:

$$\begin{aligned} \log_{10}(\mathrm{CMC}) = {} & -(3.1373 \pm 0.4374) - (9.7401 \pm 1.3165) \times 10^{-4} \times \mathrm{WI} \\ & + (11.0284 \pm 2.2709)\mathrm{RA}^{-1} + (6.704 \pm 0.6150)\,\mathrm{D} \end{aligned} \tag{4}$$

In this equation, WI, which is the Wiener number, a topological descriptor that measures molecule compactness. RA^{-1} is the reciprocal of Randic index, a criterion for quantifying molecular branching and D is the molecular dipole moment.

In 2002, Wang et al.[20] proposed a QSPR linear model for 40 anionic surfactants. This model involved a number of quantum mechanical descriptors:

$$\log_{10}(\mathrm{CMC}) = 0.546 - 0.269\mathrm{KH0} - 0.0037\ \Delta\mathrm{H_f} + 0.000224\ \mathrm{E_t} + 0.382\ \mathrm{E_{HOMO}} + 0.493\ \mathrm{E_{LUMO}} - 0.0134\ \mathrm{D} \tag{5}$$

In this equation, KH0, E_t, ΔH_f, E_{HOMO} and E_{LUMO} represent the Kier and Hall molecular connectivity index of zeroth order, total energy of the molecule, molar heat of formation, energy of the highest occupied molecular orbital, and energy of the lowest unoccupied molecular orbital, respectively.

The model of Robert et al.[21] was another correlation produced in 2002 which was generated by adopting the octanol/water partition coefficient for 16 anionic surfactants, including primary alcohol sulfate and primary alcohol ester sulfate at 50 °C. They applied two variables in their correlation: Π_h, which is the octanol/water partition coefficient of the hydrophobic moiety and is defined as the octanol/water partition coefficient of the whole molecule minus the octanol/water partition coefficient of the negatively charged fragment SO_3^- or OSO_3^- [18], and L,which is the length of hydrophobic moiety as a C–C single bond unit. The following model is their suggested correlation:

$$\log_{10}(\mathrm{CMC}) = 1.5(\pm 0.3) - 0.39\ (\pm 0.05)\ \Pi_H - 0.08\ (\pm 0.02)\ \mathrm{L} \tag{6}$$

A multi-variate linear model was presented by Li et al.[22] in 2004. They optimized the hydrophobic–hydrophilic structures of 98 anionic surfactants, including sodium alkyl sulfates, sodium alkyl sulfonates, sodium alkyl benzene sulfonates, and potassium alkyl carboxylates, and calculated quantum chemical data to develop their correlation:

$$\log_{10}(CMC) = (1.89 \pm 0.0671) - (0.0697 \pm 0.00151)\ \mathrm{N_T} - (0.0323 \pm 0.0015)\mathrm{D} + (0.381 \pm 0.0305)\ \mathrm{Q_{C\text{ - }max}} \tag{7}$$

In this equation, N_T represents the total number of atoms, and $Q_{C\text{-}max}$ represents the maximum net atomic charges on the carbon atom.

Li et al.[23] also developed a linear model in 2006 for 36 sodium alkyl benzene sulfonates using the same method as their previous work:

$$\log_{10}(\mathrm{CMC}) = -0.213 - 0.261\{\mathrm{KH0}\} + 0.598\{\mathrm{f} - \mathrm{I_{BAL}}\} - 0.0191\{\mathrm{D}\} \tag{8}$$

$f–I_{BAL}$ is the Balaban distance connectivity index of the hydrophobic segment, which stands for molecular size and compactness.

Katritzky et al.[18,24] recommended using topological, solvation, and charge-related molecular descriptors for developing models, due to the significant driving force of the intermolecular interactions between anionic surfactants and water. However, different categories of descriptors have been used in modeling, and acceptable results have been presented.

A general investigation shows that almost all suggested mathematical correlations for estimating CMC have been constructed based on chemical descriptors in constant conditions of temperature (T), mostly in aqueous solutions without salinity. However, CMC is a physico-chemical quantity of surfactants that is highly influenced by some basic parameters. Along with the chemical structure of a surfactant, the salinity of solution, temperature (T), pressure (P), and pH are the most effective parameters on CMC, as shown in previous studies[25–29].

The impact of temperature on the CMC of surfactants in water is intricate and follows a non-linear trend. Initially, the CMC decreases with temperature until it reaches a minimum, after which it starts to increase with a further increase in temperature. This is due to the fact that higher temperatures lead to reduced hydration of the hydrophilic part of the surfactant molecule, which facilitates the formation of micelles. However, at the same time, the increase in temperature also interferes with the structured water molecules surrounding the hydrophobic part of the surfactant molecule, which impedes micelle formation. Thus, the balance between the favorable and unfavorable effects of temperature on micellization determines whether the CMC increases or decreases over a certain temperature range[30]. Generally, the addition of salt to anionic surfactant solutions results in a reduction of surface tension, with the effect becoming more significant at higher salt concentrations. This phenomenon is attributed to the electrostatic interactions that facilitate the migration of surfactant monomers towards the interface[31].

The amin objective of this study was to generate novel and accurate models that incorporates both the effective parameters on CMC, including chemical descriptors and physical variables, for several widely-used common anionic surfactants. In this study, the QSPR method was coupled with two robust machine-learning approaches,-SGB and GP. New predictive methods were developed with applicability and confidence for estimating CMC. of the inclusion of physical properties such as T, pH and salinity along with the chemical descriptorsfor estimating of CMC is a novel and innovative approach. Additionally, the use of SGB and GP methods to develop CMC models is a new technique.

Materials and methods

Data set.

The total dataset includes 488 sets (i.e. observations) of experimental data adopted from the literature[11,19,25,32–42]. Each set (observation) contains basic parameters, including the salinity of the solution (in the the form of NaCl equivalent salinity), temperature (T), pH, and CMC at atmospheric pressure. The collected data involve 111 widely-used sodium-based anionic surfactants, including sodium alkyl sulfates, sodium alkane sulfonates, sodium alkyl benzene-sulfonates, sodium di-alkyl sulfosuccinate, and sodium alkyl (X) oxy-ethylene sulfates (X represents mono, di, tri or tetra).

It should be noted that NaCl equivalent salinity (S_{eq}) is defined as the salinity of brine in which all dissolved salts (cations and anions) have been replaced with a certain amount of sodium chloride so that the brine resistivity keeps the same[43,44]. It is a usual and simple method for representing salinity where a common criterion (the amount of NaCl) is applied instead of a diverse variety of salts. Additionally, the pH of solutions collected in the dataset is attributed to the dissolved salts (i.e. effects of cations and anions of the salts) without the effects of surfactant ions, and there are no acid or base additives in the collected data. The ranges of all variables have been shown in Table 1.

To generate the data-based models, the entire dataset was first randomly divided into two subsets. According to the literature[45–49], 90% of the data was considered as training data, and the remaining data points were utilized as test data. The training dataset was used to develop the CMC model, while the test data was used to test the estimation ability of the newly developed model.

Molecular descriptors generation. Molecular descriptors of a compound are numerical chemical specifications calculated from the chemical structure of the component. They are computed using certain mathematical rules that are available in specialized software[50,51]. Firstly, the chemical structure of the compound should be accurately drawn in an appropriate software. In the present study, the structures of surfactant ions (anions) were drawn in ChemBio3D Ultra, which is a module of the ChemBioOffice software[52]. Then, the drawn structures were optimized by minimizing the energy level using molecular mechanics (MM2). The optimized structures were saved as SDF files[53] and fed to the Dragon software for calculating the descriptors. The online version of Dragon software is freely available[54]. Dragon software calculates different categories of descriptors, including (1) 0D-constitutional descriptors (atom and group counts), (2) 1D-functional groups and atom-centered fragments, (3) topological, autocorrelations, connectivity indices, information indices, and eigenvalue-based indices, (4) weighted holistic invariant molecular (WHIM) and geometry, topology, and atom-weights assembly (GETAWAY) descriptors, and so on. For more information about molecular descriptors, please refer to the literature[55].

In the next step, descriptors with the same value for all compounds in the dataset, i.e.,non-informative descriptors, were excluded. Finally, a set of 1410 optimized descriptors were considered for each compound in the modeling process.

Selection of the most informative descriptors as surfactants variables. In the QSPR approach, after computing the descriptors, a small subset of the most effective descriptors should be selected as model chemical (e.g., structural) parameters along with other (basic) variables. In other words, a small number of descriptors should be chosen from the large pool. There are different methods for subset variable selection, such as genetic algorithm-based multivariate linear regression (GA-MLR)[15], genetic function approximation (GFA)[51], forward stepwise regression (FSR), replacement method (RM)[56,57], enhanced replacement method (ERM)[56,58], and so on.

In this study, the ERM was used to select the best subset. A detailed explanation of the ERM procedure can be found elsewhere[56,58,59]. In the ERM method, the user determines the number of descriptors that the algorithm should find, and ERM will find them in the form of a multivariate linear regression. The main challenge is to determine a simple regression with a minimum number of descriptors that provides appropriate accuracy. To select the best descriptors in this study, we first attempted to find two descriptors using the training dataset. The ERM algorithm developed the best linear regression with two descriptors. Then, the number of descriptors was increased one by one to enhance the accuracy of the multivariate regression. For each regression, the correlation coefficient (R^2) and residual standard deviation (RSD) were calculated using the following formulas:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{exp.} - y_i^{cal.}\right)^2}{\sum_{i=1}^{n}\left(y_i^{exp.} - \overline{y}^{exp.}\right)^2} \tag{9}$$

$$RSD = \sqrt{\frac{\sum_{i=1}^{n}(y_i^{exp.} - y_i^{cal.})^2}{n - d - 1}} \tag{10}$$

In the equations, $y_i^{exp.}$,$y_i^{cal.}$, and $\overline{y}^{exp.}$ represent the experimental, estimated, and average of experimental values of objective function ($\log_{10}$ CMC), respectively. n is the number of samples in the dataset (training dataset), and d is the number of descriptors in the linear regression. A lower value of RSD and a higher value of R^2 are desired. The results of the descriptor selection step have been shown in Fig. 1. It can be inferred from Fig. 1 that increasing the number of descriptors beyond five had no positive effect on the estimation capability of the linear

Parameters		Range
Temperature	T (K)	273.15–363.15
NaCl equivalent salinity	S_{eq} (ppm)	0–70,131.36
pH	pH	6.146–11.133
Critical micelle concentration	Log_{10} (CMC)	−1.39794 to 2.99564

Table 1. The ranges of basic variables in the present study.

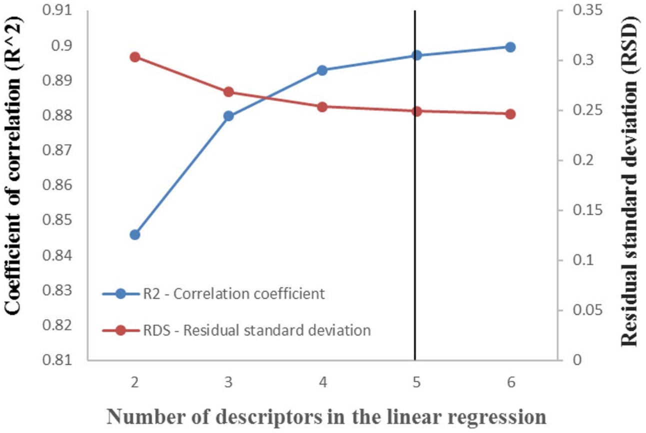


Figure 1. The effect of number of molecular descriptors on the prediction capability in descriptors selection step.

regression. Therefore, a subset of five molecular descriptors was considered, and the determined descriptors are presented in Table 2.

Developing and validation of linear multi-variable model for CMC.

The determined descriptors along with T, S_{eq} and pH were utilized to generate a multivariate linear regression model for CMC. To evaluate the predictive performance of the model, several common statistical criteria were emplyed. The root-mean-square deviation (RMSD), mean absolute error (MAE), and R^2 which are widely used parameters, were utilized in this study.

$$\mathrm{RMSD} = \sqrt{\left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(y_i^{exp.} - y_i^{cal.}\right)^2} \quad (11)$$

$$\mathrm{MAE} = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\left|y_i^{exp.} - y_i^{cal.}\right| \quad (12)$$

$y_i^{exp.}$,$y_i^{cal.}$, and n represent the experimental, estimated and number of samples of the dependent variable in the dataset, respectively. Lower values of RMSD and MAE, which indicate proximity to zero, are more desirable. The R^2 value should be close to unity. In addition to the common statistical criteria, several specific statistical techniques are used in the QSPR modeling approach to validate any QSPR linear model. The main QSPR validation methods include leave-one-out (LOO) cross-validation, leave-N-out (LNO) cross-validation, bootstrapping, y-randomization, and external validation. Although the explanation of these specific techniques has been proposed in some studies[60],a brief review is presented here.

In LOO cross-validation, each sample in the training dataset is excluded once, and a new multivariate linear regression is generated without that sample. Using the new regression, the dependent variable of the excluded sample is estimated. The values of the correlation coefficient (Q^2) and root mean square error of cross-validation (RMSECV) are then computed using the following equations:

$$\mathrm{RMSECV} = \sqrt{\left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(y_i^{exp.} - y_i^{cal.}\right)^2} \quad (13)$$

Molecular descriptor	Descriptor type	Definition
Lop	Topological descriptors	Lopping centric index
CIC2	Information indices	Complementary information content (neighborhood symmetry of 2-order)
EEig12x	Edge adjacency indices	Eigenvalue no. 12 from edge adj. matrix weighted by edge degrees
BEHp2	Burden eigenvalue descriptors	Highest eigenvalue no. 2 of Burden matrix/weighted by atomic polarizabilities
G3s	WHIM descriptors	3rd component symmetry directional WHIM index/weighted by atomic electrotopological states

Table 2. The selected molecular descriptors as chemical variables.

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{exp.} - y_i^{cal.}\right)^2}{\sum_{i=1}^{n}\left(y_i^{exp.} - \overline{y}^{exp.}\right)^2} \tag{14}$$

where $y_i^{exp.}$,$y_i^{cal.}$, $\overline{y}^{exp.}$, and n represent the experimental, estimated, average of experimental values, and the number of samples in the training dataset, respectively.

LNO cross-validation is similar to LOO, with the only difference being that in LNO cross-validation, a group of samples is excluded instead of just one. The values of RMSECV and Q^2 are recalculated for LNO cross-validation. In LOO cross-validation, repeating the test does not affect RMSECV and Q^2. However, in LNO ross-validations, RMSECV and Q^2 can vary due to the repetition of the test. In this study, the LNO cross-validation test was repeated three times and the results were reported. In developing a QSPR linear model, the minimum acceptable values for statistical variables are $Q^2 > 0.5$ and $R^2 > 0.6$. A difference between Q^2 and R^2 that exceeds 0.2–0.3 indicates overfitting in the QSPR linear modeling process[60].

In the bootstrapping technique, the entire dataset is randomly divided into training and test datasets multiple times. For each split, a respective multivariate linear regression is generated, and LOO cross-validation is performed. The values of R^2 and Q^2 are then calculated and their averages are reported (i.e. R^2_{boot} and Q^2_{boot}). In bootstrapping, a data point may be excluded once, multiple times, or never. In the present study, bootstrapping was performed 5000 times.

The y-randomization method is used to assess the possibility of chance correlation between the dependent and independent variables of a QSPR linear model. In the y-randomization test, the original matrix of independent variables values is fixed, and the vector of dependent variable is randomized. A regression is then constructed between the randomized variables. If there is no chance correlation, the resulting multivariate regression should be of poor quality. Y-randomization is performed multiple times, and the values of R^2 and LOO correlation coefficient (Q^2) are calculated for each regression (i.e. R^2_{yi} and Q^2_{yi}). The results of y-randomization are usually presented graphically as R^2_i versus Q^2_i. When $Q^2_{yi} < 0.2$ and $R^2_{yi} < 0.2$, there is no chance correlation risk[14,60]. In the present study, y-randomization was performed 1000 times.

External validation is another method in which the main dataset is randomly split into structurally similar sets of training data and an external validation set (i.e., a test set). In the present study, at first, 10% of the entire dataset was randomly selected as the external validation set (i.e., the test set) and was used to evaluate the estimation applicability.

After developing and evaluating the multi-variable linear model, the SGB and GP algorithms were applied to generate nonlinear models for CMC using the independent variables (i.e. the determined descriptors, T, and S_{eq}). Nonlinear models often provide more accuracy and estimation power.

Stochastic gradient boosting (SGB).

In the current inquiry, the stochastic gradient boosting (SGB) tree framework was implemented over collected data to model CMC.

Stochastic Gradient Boosting is an improvement on the classic Gradient Boosting method, created by Friedman[61]. By incorporating Breiman's bagging approach[62], it boosts accuracy and efficiency by randomly sampling the training data[63,64]. This results in better prediction performance[65], and the technique has been proven effective in many industries and applications[66–76].

In more general terms, Gradient Boosting (GB) is an effective algorithm that transforms weak hypotheses into strong ones by combining a series of ensemble learners made up of simple base or weak learners[77,78]. A weak learner is defined as one whose performance is only slightly better than random chance, and in the case of GB, decision trees (such as regression trees) are commonly used as weak learners. To avoid overfitting, the construction of trees is often constrained by limiting the number of levels or choosing the best split points based on minimizing a loss function.

The overall goal of the algorithm is to minimize the loss of the model by adding weak learners using a gradient descent-like procedure. At each iteration, a new weak learner is added that focuses on the cases that the previous weak learner did not predict correctly, thus reducing the loss. The output of each generated tree is then added to the output of the sequence of trees to gradually improve the final output of the model.

Stochastic GB is a variation of GB where a subsample of the total training set is randomly selected for each iteration, and the base learner is fit on that subsample without replacement[61,64]. This reduces the risk of overfitting and allows for self-validation of the model internally by using out-of-bag error estimates. Additionally, the algorithm becomes faster since regression trees are generated on smaller datasets at each iteration. The review of the literature has shown the high ability of this new branch of decision tree algorithm in chemical engineering areas[79,80].

When developing the SGB model, the error values sharply decreased with an increasing number of trees until the error rate stabilized (see Fig. 2). The SGB algorithm selected a solution with 2736 number of trees, which was the solution that returned the minimum error in the form of RMSD for the test data set ($RMSD_{test} = 0.05203$).

To achieve the most generalizable model, determining the learning rate was crucial. The learning rate is the specific weight at which consecutive simple trees are added to the prediction equation, and it is considered the most important parameter. To identify the optimal value, a sensitivity analysis was performed, which demonstrated the effects of learning rate on the performance of the SGB model for predicting CMC, as illustrated in Fig. 3. The optimized parameter was determined to be 0.09. Using the SGB tree, the importance degrees of all the model parameters were also determined.

Genetic programming (GP).

Genetic programing (GP) is an algorithm used in the present study to develop the CMC model. GP is a well-known machine learning approaches for optimization and modeling studies which was introduced in the 1990s by John Koza[81]. The GP procedure is inspired by biological generation

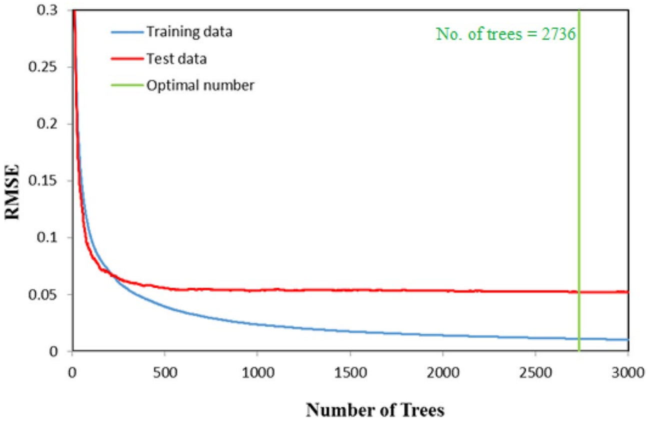


Figure 2. The graph of RMSD over the successive boosting steps for the training and test samples using SGB method.

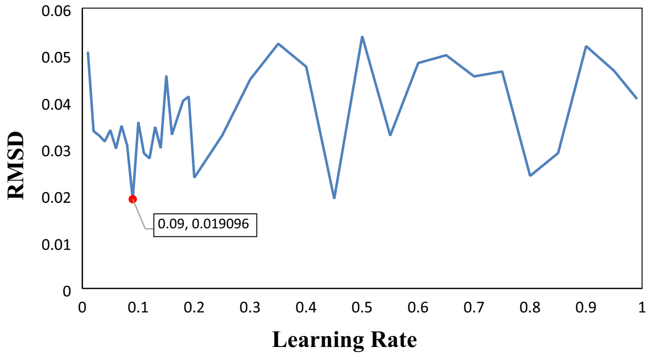


Figure 3. The effects of learning rate on the performance of the SGB model for predicting CMC.

phenomenon in which computer programs evolve evolutionarily in a machine learning algorithm to perform tasks.

In the GP process, a population of mathematical functions is first randomly generated from pre-determined user-defined mathematical operators. Then, some of these functions are randomly chosen to be arranged in the form of one or several "genes". A Gene is represented as a chromosome-like syntactic tree structure that operates on input data, i.e., the training dataset(as shown in Fig. 4)[82,83].

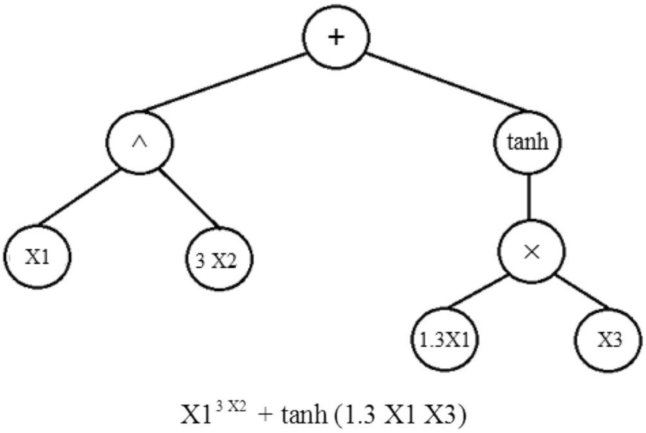


Figure 4. Schematic of a simple GP gene including the operators: +, ^, ×, tanh.

After the primary genes are determined from the first population (known as parents), the overall primary GP model is developed by a weighted summation of the genes with a bias term. However, the primary model does not provide the desired accuracy, and a modification process is required. In the next step, the tree structures of primary genes are modified by crossing over the best performing trees and cutting some sections of trees to be exchanged between themselves. This modification mainly results in a new population (next generation or children) due to changes in the mathematical functions[84].

The generation is iterated several times in a regular process until the last population is generated, which includes the most-optimized functions with a specific arrangement of genes to solve the problem[85]. In the modeling applications of GP, regression between the objective function and independent variables is also known as "multi-gene symbolic regression". It is an effective technique that includes one or more genes (individual usual GP trees) providing simple and fast processing to perform tasks[83,86].

In this study, the number of populations and number of generations were set as 180 each, and the mathematical operators + , −, × , /, and exp (exponential) were employed. GP was run over the input data, and the output model with acceptable accuracy was obtained.

Results and discussion

Multi-variable linear correlation of CMC.

The multi-variable linear model for CMC of anionic surfactants in brine is presented below:

$$\begin{aligned}\mathrm{Log}_{10}(CMC) = {} & 31.817705(\pm 1.59767) + 0.002290(\pm 0.00066) \times \mathrm{T} - 0.083577(\pm 0.02999) \times \mathrm{pH} \\ & - 0.000023(\pm 0.000002) \times \mathrm{S}_{\mathrm{eq}} - 0.498878(\pm 0.03992) \times \{\mathrm{CIC2}\} - 0.465377(\pm 0.03149) \times \{\mathrm{EEig12x}\} \\ & - 0.445544(\pm 0.05699) \times \{\mathrm{Lop}\} - 7.805830(\pm 0.44219) \times \{\mathrm{BEHp2}\} - 2.840368(\pm 0.36536) \times \{\mathrm{G3s}\}\end{aligned} \tag{15}$$

The variables of the new developed model have been presented in Tables 1 and 2. The determined descriptors (shown in Table 2) are "CIC2"[87], "EEig12x"[88], "Lop"[88,89], "BEHp2"[90], and "G3s"[91].

CIC2 is a complementary information content of 2nd order neighborhood symmetry from the category of information indices descriptors. It is a measure of the degree of diversity of elements in the structure[87].

The Lop descriptor is a lopping centric index categorized in topological descriptors, which are usually obtained from a hydrogen-depleted molecular graph. A molecular graph is a labeled graph whose vertices correspond to the atoms of the compound labeled with the kinds of atoms, and the edges correspond to chemical bonds labeled with the types of bonds[89].

Lop is an index defined as the mean information content derived from the pruning partition of a graph[88].

EEig12x is one of the edge adjacency indices descriptors, which stands for the 12th eigenvalue of the edge adjacency matrix weighted by edge degrees. The edge adjacency matrix derived from a molecular graph encodes the connectivity between graph edges[88].

BEHp2 belongs to the Burden eigenvalue category from 2D topological descriptors. It is a measure of molecule/ion polarizability defined as the 2nd highest eigenvalue of the Burden matrix, which is weighted by atomic polarizabilities[90,92].

G3s is a WHIM descriptor and is defined as the 3rd component symmetry directional WHIM index weighted by atomic electro-topological states. WHIM specifications are used to calculate 3D molecular information based on molecular size, shape, symmetry, diversity of atoms, etc.[91].

The statistical parameters of the multivariate linear correlation, including QSPR specific validation parameters, are presented in Tables 3 and 4. The values of R^2, RMSD, and MAE show medium accuracy of the linear model. The validity of the linear model was checked by LOO cross-validation, LNO cross-validation, bootstrapping, y-randomization, and external validation techniques. The LNO cross-validation parameters are shown in Table 4, and the bootstrapping test was performed 5000 times. The low difference between the values of Q^2_{LOO},

$n_{total} = 488$	$n_{train} = 440$	$n_{test} = 48$
$R^2_{total} = 0.9059$	$R^2_{train} = 0.9061$	$R^2_{test} = 0.9047$
$RMSD_{total} = 0.2382$	$RMSD_{train} = 0.2367$	$RMSD_{test} = 0.2514$
$MAE_{total} = 0.1734$	$MAE_{train} = 0.1711$	$MAE_{test} = 0.1947$
$Q^2_{LOO} = 0.8988$	$RMSECV_{LOO} = 0.2456$	$Q^2_{boot} = 0.8990$
$R^2_{boot} = 0.9064$	$Q^2_{ext} = 0.9048$	$R^2_{ext} = 0.9047$

Table 3. Statistical parameters of multivariate linear model for CMC of anionic surfactants in brine. The subscripts "total", "train" and "test" are attributed to total dataset, training dataset and test dataset, respectively.

1st	2nd	3th	Average
$Q^2_{L-25\%-O} = 0.8940$	$Q^2_{L-25\%-O} = 0.8943$	$Q^2_{L-25\%-O} = 0.8991$	$Q^2_{L-25\%-O} = 0.8958$
$RMSECV_{LNO} = 0.2514$	$RMSECV_{LNO} = 0.2511$	$RMSECV_{LNO} = 0.2453$	$RMSECV_{LNO} = 0.2493$

Table 4. Statistical parameters of LNO cross-validation for linear model of CMC.

Q^2_{LNO}, $Q^2_{boot,}$ Q^2_{ext}, R^2_{boot}, and R^2_{ext} indicates that the linear model has been developed without occurring overfitting. The y-randomization test was repeated 1000 times, and the results are shown in Fig. 5. According to this test, the values of Q^2_{yi} and R^2_{yi} (i.e., y-randomization data points) are of poor quality compared to the linear model correlation coefficient (R^2) and Q^2_{LOO} (indicated as a red point in Fig. 5), which verifies that there is no risk of chance correlation in the multi-variable linear model of CMC.

The estimated CMC by Eq. (15) versus experimental data is presented in Fig. 6. Based on Tables 3 and 4 and Fig. 6, the linear model has acceptable accuracy. However, the prediction ability is not excellent enough. The results of non-linear models are proposed in the next section.

Non-linear models of CMC.

The SGB and GP programs were run over the input data to produce new models for the CMC of anionic surfactants in a brine solution. The execution of the SGB algorithm in this study follows the explanations in Friedman[61,64]. The new GP model is a mathematical relation as follows:

$$\begin{aligned}\mathrm{Log_{10}(CMC)} =&0.0006095\ S_{eq} - 13.76\{CIC2\}+0.0003308\{EEig12x\} - 6.882\{BEHp2\}+0.001219\{EEig12x\}^2\\ &-1.096\left(\exp(-\{EEig12x\}(\{EEig12x\}+\{G3s\})) + \exp\left(-2\{EEig12x\}^2\right)\right)+13.56\exp(-\exp(\{G3s\} - \{CIC2\}))\\ &-(\{BEHp2\}+\{G3s\})\left(0.0001654\ S_{eq} + 31.68\left(\frac{\{CIC2\}+\{EEig12x\}}{1.407\ T + 2.815\ S_{eq}}\right)\right)0.01423\{EEig12x\}\left(\frac{pH + \{G3s\}}{\{G3s\}}\right)\\ &-3.961(\{CIC2\}+\{EEig12x\})\left(\frac{S_{eq} - 9.438}{9.007\ (T + S_{eq})}\right)+0.8278(\{CIC2\} - \exp(-\{CIC2\}))\left(\{CIC2\}\left(1 - \frac{1}{pH}\right)+9.447\right)\\ &-0.5913(2.908\{Lop\} - \{CIC2\})\exp(\{BEHp2\}-5.068) - +0.4997\exp(2\{EEig12x\} - \{CIC2\})+26.1\end{aligned} \tag{16}$$

Table 5 shows the statistical parameters of the presented models. The values of R^2, RMSD, and MAE represent the acceptable applicability of SGB and GP models and the high accuracy and superiority of the SGB method. Figures 7 and 8 show the estimated CMC versus the experimental values for the GP and SGB models, respectively. The calculated data by the SGB model has been scattered well on the 45 degree line (y = x), verifying excellent accuracy.

Figure 9presents the curves of cumulative frequency versus absolute errors of the objective function (Log $_{10}$ (CMC)) for the SGB and GP models, as well as the linear correlation. The maximum absolute error of the SGB model in this figure is 0.18. Moreover, the absolute errors of 82.2% of all datasets are less than 0.01, and the absolute errors of 99.2% of the data are below 0.1 for the new SGB model. Figure 10 shows absolute errors over the total dataset for the linear (top plot), GP (middle plot), and SGB (bottom plot) models. As observed in Figs. 9 and 10, the estimation accuracy has been enhanced from the linear model to the SGB model, and the accuracy of the SGB method is the highest.

The relative importance of independent variables, including descriptors (Lop, CIC2, EEig12x, BEHp2, and G3s), T, pH, and S_{eq}, has been determined by the SGB algorithm in the calibration of the SGB model, and the results have been depicted in Fig. 11. A higher value of a variable indicates stronger relative importance on the

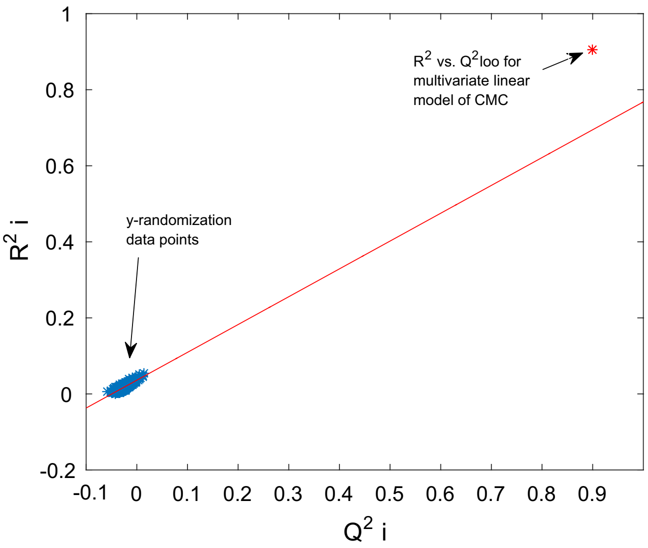


Figure 5. The result of y-randomization test for multi-variable linear model of CMC.

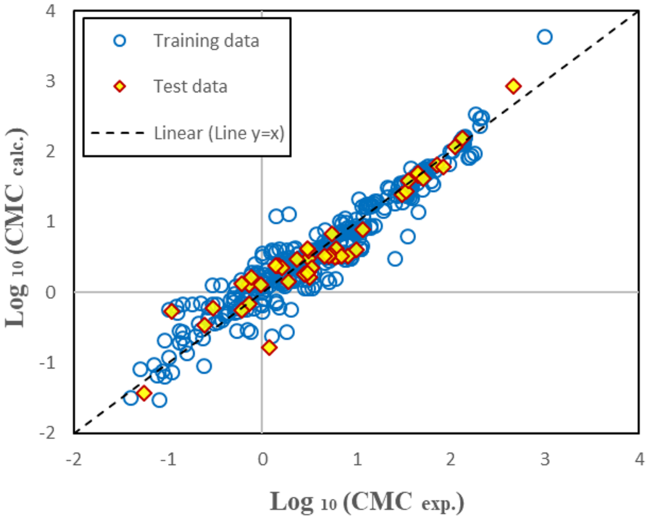


Figure 6. The estimated CMC versus experimental data for multivariate linear model over training and test datasets.

Statistical parameters	SGB model			GP model		
	All	Train	Test	All	Train	Test
R^2	0.999395	0.999808	0.991658	0.954946	0.953866	0.963834
RMSD	0.019096	0.010993	0.052034	0.164829	0.165879	0.154869
MAE	0.008387	0.005457	0.036536	0.111650	0.111228	0.115514

Table 5. Statistical parameters of non-linear models for CMC of anionic surfactants in brine.

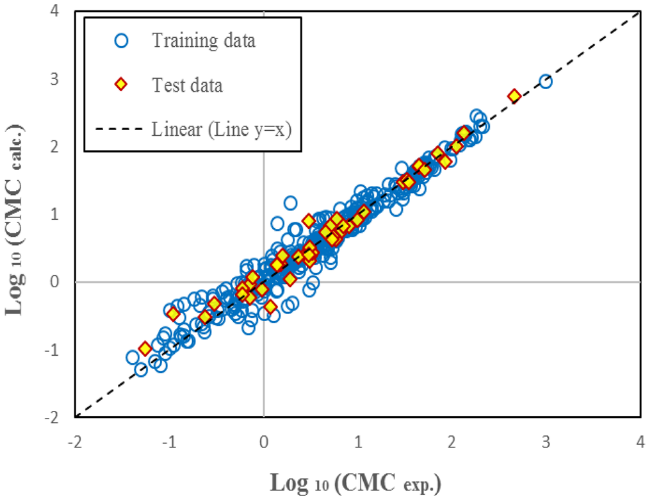


Figure 7. The estimated CMC versus experimental values for GP model over training and test datasets.

response. As shown, the descriptor Lop is the more effective factor among the input variables in the development of the SGB model.

The application of the proposed models has been shown in Table 6 for the estimation of the CMC of sodium dodecyl sulfate as a sample in the dataset.

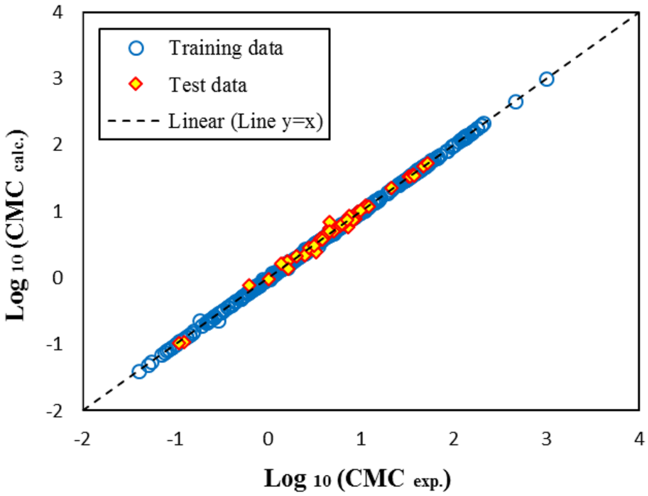


Figure 8. The estimated CMC versus experimental values for SGB model over training and test datasets.

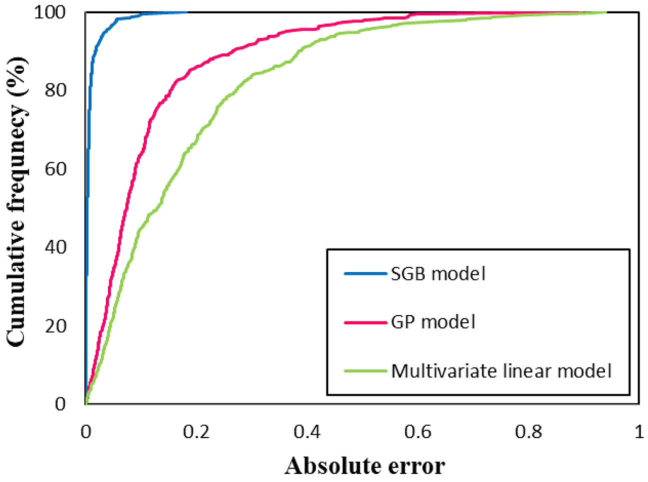


Figure 9. Cumulative frequency of the new developed models.

The generation of new models with high accuracy for the CMC of surfactant solutions containing different types of salts based on the QSPR approach and the application of GP and SGB for producing non-linear models are novelties of the present study. Using a wide range of salinities and temperatures, as well as various types of anionic surfactants in the modelling procedure, has increased the estimation applicability and prediction performance of the newly developed models.

Conclusion

The estimation of CMC is one of the most important interests of the academic and industrial communities dealing with surfactants. The present study was conducted to obtain novel methods for the estimation of the CMC of well-known, highly-used anionic surfactants as functions of both physical parameters (T, pH and salinity) and chemical factors (Lop, CIC2, EEig12x, BEHp2, and G3s) and to avoid the expensive and time-consuming laboratory measurements. CMC estimation at different temperatures and salinities is considered novel and innovative. The QSPR molecular approach, along with the ensemble learning framework of stochastic gradient boosting (SGB) and genetic programming (GP) procedures, was used to produce models for CMC in brine. The implemented algorithms are reliable and applicable for predicting CMC. However, the output of SGB is more accurate in terms of statistical parameters. This inquiry also encourages the scientific and engineer communities

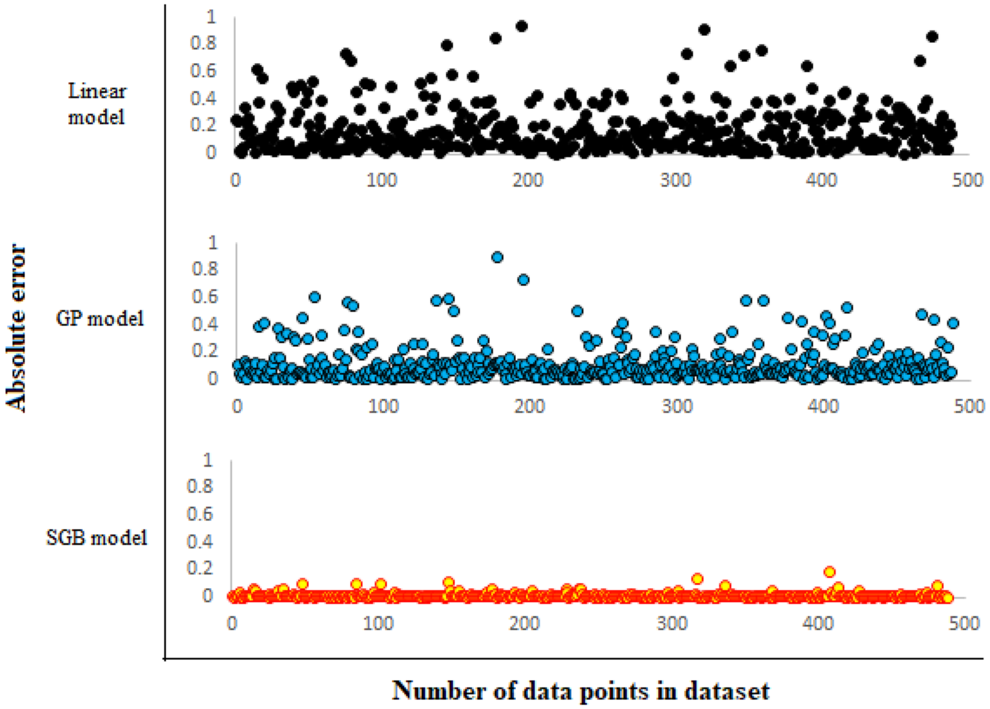


Figure 10. Absolute errors of data points over all dataset for linear model (top), GP model (middle) and SGB model (down). It is observed that the estimation accuracy has been increased from top to down.

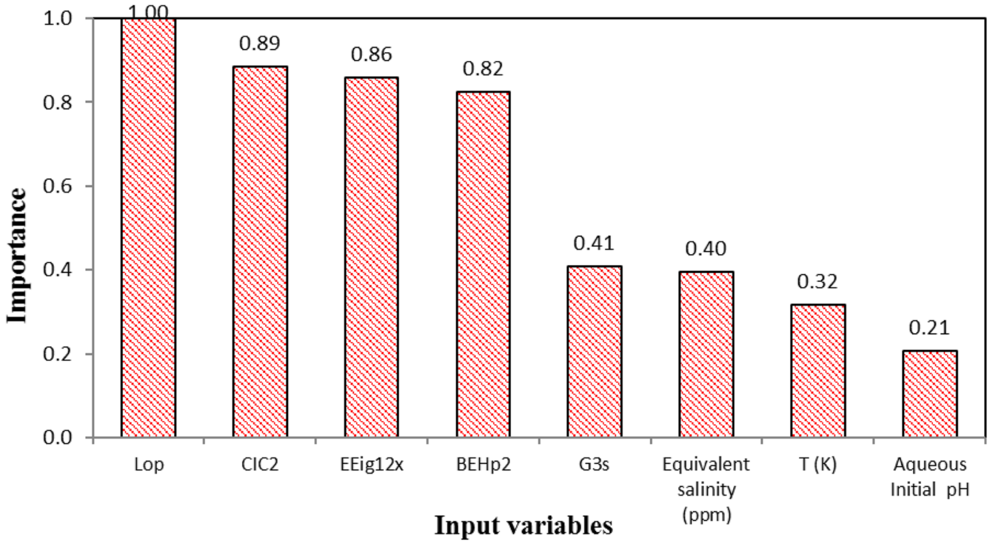


Figure 11. Relative importance of independent variables on the CMC based on the SGB algorithm.

Surfactant	Structure	Physical variables	Descriptors of anionic part (without Na^+ ion)	CMC

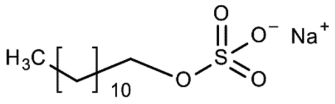
sodium dodecyl sulfate		T = 298.15 K
		pH = 7
		S_{eq} = 309.75 ppm
			Lop = 2.911
			CIC2 = 2.838
			EEig12x = 0
			BEHp2 = 3.574
			G3s = 0.275
				log $(CMC)_{exp.}$ = 0.805
				log $(CMC)_{calc.}$ _linear model = 0.516 (AE[a] = 0.289)
				log $(CMC)_{calc.}$ _GP model = 0.729 (AE = 0.076)
				log $(CMC)_{calc.}$ _SGB model = 0.804 (AE = 0.001)

Table 6. Application of the new QSPR models of the present study for estimation of CMC of a sample component. [a]The abbreviation AE indicates absolute error calculated as : AE $=|y^{exp.} - y^{calc.}|$.

to further investigate the use of the novel branch of soft computing frameworks. Developing such models for CMC provides new applications in the simulation and control of surfactant systems, as well as prediction of CMC for newly designed anionic surfactants.

Data availability

All the literature datasets analyzed in this study are available at a reasonable request from the corresponding authors.

Received: 14 March 2023; Accepted: 10 August 2023
Published online: 17 August 2023

References

1. Schramm, L. L., Stasiuk, E. N. & Marangoni, D. G. 2 Surfactants and their applications. *Ann. Rep. Sect. C (Phys. Chem.)* **99**, 3–48 (2003).
2. Massarweh, O. & Abushaikha, A. S. The use of surfactants in enhanced oil recovery: A review of recent advances. *Energy Rep.* **6**, 3150–3178 (2020).
3. Suárez, L., Díez, M. A., García, R. & Riera, F. A. Membrane technology for the recovery of detergent compounds: A review. *J. Ind. Eng. Chem.* **18**, 1859–1873 (2012).
4. Falbe, J. *Surfactants in Consumer Products: Theory, Technology and Application*. (Springer Science & Business Media, 2012).
5. Hellgren, A.-C., Weissenborn, P. & Holmberg, K. Surfactants in water-borne paints. *Prog. Org. Coat.* **35**, 79–87 (1999).
6. Kralova, I. & Sjöblom, J. Surfactants used in food industry: A review. *J. Dispers. Sci. Technol.* **30**, 1363–1383 (2009).
7. Adams, J. W. Organosilicone Surfactants: Properties, Chemistry, and Applications. *Surface Phenomena and Additives in Water-Based Coatings and Printing Technology*, 73–82 (1991).
8. Myers, D. *Surfactant science and technology*. (John Wiley & Sons, 2005).
9. Rosen, M. J. *Surfactants and Interfacial Phenomena*. (Wiley, 2004).
10. Gaudin, T. *et al.* Impact of the chemical structure on amphiphilic properties of sugar-based surfactants: A literature overview. *Adv. Coll. Interface. Sci.* **270**, 87–100 (2019).
11. Mukerjee, P. & Mysels, K. J. *Critical Micelle Concentrations of Aqueous Surfactant Systems*. (National Standard reference data system, 1971).
12. Rangel-Yagui, C. O., Pessoa, A. Jr. & Tavares, L. C. Micellar solubilization of drugs. *J. Pharm. Pharm. Sci* **8**, 147–163 (2005).
13. Arachea, B. T. *et al.* Detergent selection for enhanced extraction of membrane proteins. *Protein Expr. Purif.* **86**, 12–20 (2012).
14. Abooali, D. & Sobati, M. A. Novel method for prediction of normal boiling point and enthalpy of vaporization at normal boiling point of pure refrigerants: A QSPR approach. *Int. J. Refrigerat.* **40**, 282–293 (2014).
15. Gharagheizi, F. & Sattari, M. Prediction of triple-point temperature of pure components using their chemical structures. *Ind. Eng. Chem. Res.* **49**, 929–932 (2009).
16. Klevens, H. Structure and aggregation in dilate solution of surface active agents. *J. Am. Oil. Chem. Soc.* **30**, 74–80 (1953).
17. Huibers, P. D., Lobanov, V. S., Katritzky, A., Shah, D. & Karelson, M. Prediction of critical micelle concentration using a quantitative structure–property relationship approach. *J. Colloid Interface Sci.* **187**, 113–120 (1997).
18. Hu, J., Zhang, X. & Wang, Z. A review on progress in QSPR studies for surfactants. *Int. J. Mol. Sci.* **11**, 1020–1047 (2010).
19. Jalali-Heravi, M. & Konouz, E. Prediction of critical micelle concentration of some anionic surfactants using multiple regression techniques: A quantitative structure-activity relationship study. *J. Surfact. Deterg.* **3**, 47–52 (2000).
20. Wang, Z.-W., Li, G.-Z., Zhang, X. & Li, L. Prediction on critical micelle concentration of anionic surfactants in aqueous solution: quantitative structure-property relationship approach. *Acta Chimica Sinica-Chinese Edition* **60**, 1548–1552 (2002).
21. Roberts, D. W. Application of octanol/water partition coefficients in surfactant science: A quantitative structure—property relationship for micellization of anionic surfactants. *Langmuir* **18**, 345–352 (2002).
22. Li, X. *et al.* Estimation of critical micelle concentration of anionic surfactants with QSPR approach. *J. Mol. Struct. (Thoechem)* **710**, 119–126 (2004).
23. Xuefeng, L. *et al.* Correlation of critical micelle concentration of sodium alkyl benzenesulfonates with molecular descriptors. *Wuhan Univ. J. Nat. Sci.* **11**, 409–414 (2006).
24. Katritzky, A. R., Pacureanu, L., Dobchev, D. & Karelson, M. QSPR study of critical micelle concentration of anionic surfactants using computational molecular descriptors. *J. Chem. Inf. Model.* **47**, 782–793 (2007).
25. Chauhan, S. & Sharma, K. Effect of temperature and additives on the critical micelle concentration and thermodynamics of micelle formation of sodium dodecyl benzene sulfonate and dodecyltrimethylammonium bromide in aqueous solution: A conductometric study. *J. Chem. Thermodyn.* **71**, 205–211 (2014).
26. Hara, K., Kuwabara, H., Kajimoto, O. & Bhattacharyya, K. Effect of pressure on the critical micelle concentration of neutral surfactant using fluorescence probe method. *J. Photochem. Photobiol., A* **124**, 159–162 (1999).
27. Rahman, A. & Brown, C. Effect of pH on the critical micelle concentration of sodium dodecyl sulphate. *J. Appl. Polym. Sci.* **28**, 1331–1334 (1983).
28. Ren, Z. H. Mechanism of the salt effect on micellization of an aminosulfonate amphoteric surfactant. *Ind. Eng. Chem. Res.* **54**, 9683–9688 (2015).
29. Akhlaghi, N. & Riahi, S. Salinity effect on the surfactant critical micelle concentration through surface tension measurement. *Iran. J. Oil Gas Sci. Technol.* **8**, 50–63 (2019).
30. Rosen, M. J. & Kunjappu, J. T. *Surfactants and Interfacial Phenomena* (Wiley, 2012).
31. Rafique, A. S. *et al.* Micellar structure and transformations in sodium alkylbenzenesulfonate (NaLAS) aqueous solutions: Effects of concentration, temperature, and salt. *Soft Matter* **16**, 7835–7844 (2020).
32. Davis, A., Morton, S., Counce, R., DePaoli, D. & Hu, M.-C. Ionic strength effects on hexadecane contact angles on a gold-coated glass surface in ionic surfactant solutions. *Colloids Surf., A* **221**, 69–80 (2003).
33. Fletcher, P. D., Savory, L. D., Woods, F., Clarke, A. & Howe, A. M. Model study of enhanced oil recovery by flooding with aqueous surfactant solution and comparison with theory. *Langmuir* **31**, 3076–3085 (2015).
34. Fu, J. *et al.* A new technique for determining critical micelle concentrations of surfactants and oil dispersants via UV absorbance of pyrene. *Colloids Surf., A* **484**, 1–8 (2015).
35. Moradi, P., Najafi, M. & Khani, V. Adsorption and micellar phase properties of anionic surfactant in the presence of electrolyte and oil at different temperatures. *Fluid Phase Equilib.* **337**, 370–378 (2013).
36. Mulqueen, M. & Blankschtein, D. Theoretical and experimental investigation of the equilibrium oil–water interfacial tensions of solutions containing surfactant mixtures. *Langmuir* **18**, 365–376 (2002).
37. Nahringbauer, I. The interaction between polymer and surfactant as revealed by interfacial tension. *Trends Colloid Interface Sc. V* **1**, 200–205 (1991).

38. Puig, J., Mares, M., Miller, W. & Franses, E. Mechanism of ultralow interfacial tensions in dilute surfactant—oil—brine systems. *Colloids Surf.* **16**, 139–152 (1985).
39. Rosen, M. J., Wang, H., Shen, P. & Zhu, Y. Ultralow interfacial tension for enhanced oil recovery at very low surfactant concentrations. *Langmuir* **21**, 3749–3756 (2005).
40. Serrano-Saldaña, E. & Domı́nguez-Ortiz, A., Pérez-Aguilar, H., Kornhauser-Strauss, I. & Rojas-González, F.,. Wettability of solid/brine/n-dodecane systems: Experimental study of the effects of ionic strength and surfactant concentration. *Colloids Surfaces A Physicochem. Eng. Aspects* **241**, 343–349 (2004).
41. Zdziennicka, A., Szymczyk, K., Krawczyk, J. & Jańczuk, B. Critical micelle concentration of some surfactants and thermodynamic parameters of their micellization. *Fluid Phase Equilib.* **322**, 126–134 (2012).
42. Zhou, J. & Dupeyrat, M. Alcohol effect on interfacial tension in oil—water—sodium dodecyl sulphate systems. *J. Colloid Interface Sci.* **134**, 320–335 (1990).
43. Bassiouni, Z. *Theory, measurement, and interpretation of well logs* Vol. 4 (Society of Petroleum Engineers, 1994).
44. Limited, S. *Schlumberger log interpretation charts*. (Schlumberger, 1984).
45. Abooali, D., Soleimani, R. & Gholamreza-Ravi, S. Characterization of physico-chemical properties of biodiesel components using smart data mining approaches. *Fuel* **266**, 117075 (2020).
46. Chatterjee, S. *et al.* Particle swarm optimization trained neural network for structural failure prediction of multistoried RC buildings. *Neural Comput. Appl.* **28**, 2005–2016 (2017).
47. Gupta, A. K., Singh, S. K., Reddy, S. & Hariharan, G. Prediction of flow stress in dynamic strain aging regime of austenitic stainless steel 316 using artificial neural network. *Mater. Des.* **35**, 589–595 (2012).
48. Gyurova, L. A. & Friedrich, K. Artificial neural networks for predicting sliding friction and wear properties of polyphenylene sulfide composites. *Tribol. Int.* **44**, 603–609 (2011).
49. Soleimani, R., Abooali, D. & Shoushtari, N. A. Characterizing CO2 capture with aqueous solutions of LysK and the mixture of MAPA+ DEEA using soft computing methods. *Energy* **164**, 664–675 (2018).
50. Sobati, M. A. & Abooali, D. Molecular based models for estimation of critical properties of pure refrigerants: Quantitative structure property relationship (QSPR) approach. *Thermochim. Acta* **602**, 53–62 (2015).
51. Khajeh, A. & Modarress, H. QSPR prediction of surface tension of refrigerants from their molecular structures. *Int. J. Refrigerat.* **35**, 150–159 (2012).
52. C.B.O. Cambridgesoft, http://www.cambridgesoft.com/. (2015).
53. Dalby, A. *et al.* Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **32**, 244–255 (1992).
54. VCCLAB, Virtual Computational Chemistry Laboratory, http://www.vcclab.org. (2005).
55. Todeschini, R. & Consonni, V. *Molecular descriptors for chemoinformatics, volume 41 (2 volume set)*. Vol. 41 (John Wiley & Sons, 2009).
56. Mercader, A. G., Duchowicz, P. R., Fernández, F. M. & Castro, E. A. Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories. *Chemom. Intell. Lab. Syst.* **92**, 138–144 (2008).
57. Morales, A. H. *et al.* Application of the replacement method as a novel variable selection strategy in QSAR. 1. Carcinogenic potential. *Chemomet. Intell. Lab. Syst.* **81**, 180–187 (2006).
58. Mercader, A. G., Duchowicz, P. R., Fernández, F. M. & Castro, E. A. Advances in the replacement and enhanced replacement method in QSAR and QSPR theories. *J. Chem. Inf. Model.* **51**, 1575–1581 (2011).
59. Sobati, M. A., Abooali, D., Maghbooli, B. & Najafi, H. A new structure-based model for estimation of true critical volume of multi-component mixtures. *Chemom. Intell. Lab. Syst.* **155**, 109–119 (2016).
60. Kiralj, R. & Ferreira, M. Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. *J. Braz. Chem. Soc.* **20**, 770–787 (2009).
61. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
62. Breiman, L. Arcing the edge. (Technical Report 486, Statistics Department, University of California at Berkeley, 1997).
63. Kriegler, B. & Berk, R. Small area estimation of the homeless in Los Angeles: An application of cost-sensitive stochastic gradient boosting. *Ann. Appl. Stat.* **1**, 1234–1255 (2010).
64. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **1**, 1189–1232 (2001).
65. Kuhn, M. & Johnson, K. *Applied Predictive Modeling*. Vol. 810 (Springer, 2013).
66. Saeedi Dehaghani, A. H. & Soleimani, R. Prediction of CO_2-oil minimum miscibility pressure using soft computing methods. *Chem. Eng. Technol.* **43**, 1361–1371 (2020).
67. Abooali, D., Soleimani, R. & Gholamreza-Ravi, S. Characterization of physico-chemical properties of biodiesel components using smart data mining approaches. *Fuel* **266**, 117075 (2020).
68. Abooali, D., Soleimani, R. & Rezaei-Yazdi, A. Modeling CO_2 absorption in aqueous solutions of DEA, MDEA, and DEA+ MDEA based on intelligent methods. *Sep. Sci. Technol.* **55**, 697–707 (2020).
69. Soleimani, R., Abooali, D. & Shoushtari, N. A. Characterizing CO_2 capture with aqueous solutions of LysK and the mixture of MAPA+ DEEA using soft computing methods. *Energy* **164**, 664–675 (2018).
70. Hashemkhani, M. *et al.* Prediction of the binary surface tension of mixtures containing ionic liquids using Support Vector Machine algorithms. *J. Mol. Liq.* **211**, 534–552 (2015).
71. Soleimani, R. *et al.* Evolving an accurate decision tree-based model for predicting carbon dioxide solubility in polymers. *Chem. Eng. Technol.* **43**, 514–522 (2020).
72. Dehaghani, A. H. S. & Soleimani, R. Estimation of interfacial tension for geological CO_2 storage. *Chem. Eng. Technol.* **42**, 680–689 (2019).
73. Soleimani, R., Dehaghani, A. H. S. & Bahadori, A. A new decision tree based algorithm for prediction of hydrogen sulfide solubility in various ionic liquids. *J. Mol. Liq.* **242**, 701–713 (2017).
74. Brillante, L. *et al.* Investigating the use of gradient boosting machine, random forest and their ensemble to predict skin flavonoid content from berry physical–mechanical characteristics in wine grapes. *Comput. Electron. Agric.* **117**, 186–193 (2015).
75. Godinho, S., Guiomar, N. & Gil, A. Using a stochastic gradient boosting algorithm to analyse the effectiveness of Landsat 8 data for montado land cover mapping: Application in southern Portugal. *Int. J. Appl. Earth Obs. Geoinf.* **49**, 151–162 (2016).
76. Zhou, J., Li, X. & Mitri, H. S. Comparative performance of six supervised learning methods for the development of models of hard rock pillar stability prediction. *Nat. Hazards* **79**, 291–316 (2015).
77. Kearns, M. Thoughts on hypothesis boosting. *Unpublished manuscript* **45**, 105 (1988).
78. Mason, L., Baxter, J., Bartlett, P. L. & Frean, M. R. in *Advances in neural information processing systems*. 512–518.
79. Soleimani, R., Dehaghani, A. H. S. & Bahadori, A. A new decision tree based algorithm for prediction of hydrogen sulfide solubility in various ionic liquids. *J. Mol. Liq.* **242**, 701–713 (2017).
80. Soleimani, R., Mahmood, T. & Bahadori, A. Assessment of compressor power and condenser duty per refrigeration duty in three-stage propane refrigerant systems using a new ensemble learning tool. *Chemeca 2016: Chemical Engineering-Regeneration, Recovery and Reinvention*, 23 (2016).
81. Koza, J. R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. (Bradford, 1992).
82. Abooali, D. & Khamehchi, E. New predictive method for estimation of natural gas hydrate formation temperature using genetic programming. *Neural Computing and Applications*, 1–10.

83. Searson, D. P., Leahy, D. E. & Willis, M. J. in *Proceedings of the International multiconference of engineers and computer scientists.* 77–80 (Citeseer).
84. Abooali, D. & Khamehchi, E. Toward predictive models for estimation of bubble-point pressure and formation volume factor of crude oil using an intelligent approach. *Braz. J. Chem. Eng.* **33**, 1083–1090 (2016).
85. Abooali, D. & Khamehchi, E. Estimation of dynamic viscosity of natural gas based on genetic programming methodology. *J. Nat. Gas Sci. Eng.* **21**, 1025–1031 (2014).
86. Searson, D. GPTIPS: Genetic programming & symbolic regression for MATLAB. *User Guide* **2010** (2009).
87. Gharagheizi, F. & Alamdari, R. F. Prediction of flash point temperature of pure components using a quantitative structure–property relationship model. *Mol. Inf.* **27**, 679–683 (2008).
88. Todeschini, R. & Consonni, V. *Handbook of Molecular Descriptors*. Vol. 11 (Wiley, 2008).
89. Gold, V., Loening, K., McNaught, A. & Shemi, P. *IUPAC compendium of chemical terminology* (Blackwell Science, 1997).
90. Burden, F. R. Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.* **29**, 225–227. https://doi.org/10.1021/ci00063a011 (1989).
91. Todeschini, R. & Gramatica, P. SD-modelling and prediction by WHIM descriptors. Part 5. Theory development and chemical meaning of WHIM descriptors. *Mol. Inf.* **16**, 113–119 (1997).
92. Burden, F. R. A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix. *Mol. Inf.* **16**, 309–314 (1997).

Author contributions

D.A.: Conceptualization, Methodology, Software, Validation, Writing - original draft, Resources, Visualization, Data Curation, Investigation, Formal analysis, Supervision. R.S.: Project administration, Conceptualization, Validation, Visualization, Data Curation, Software, Writing - original draft, Writing - Review & Editing, Methodology, Formal analysis.

Competing interests

The authors certify that they have NO conflict over any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Additional information

Correspondence and requests for materials should be addressed to D.A. or R.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023