Investigating the impact of sugar-based surfactants structure on surface tension at critical micelle concentration with structure-property relationships

Théophile Gaudin[a), b)], Patricia Rotureau[b)], Isabelle Pezron[a)], Guillaume Fayet[b),*]

a) Sorbonne Universités, Université de Technologie de Compiègne, EA 4297 TI-MR, rue du Dr Schweitzer, 60200 Compiègne, France

b) INERIS, Parc Technologique Alata, BP2, 60550 Verneuil-en-Halatte, France

*Corresponding author: guillaume.fayet@ineris.fr; tel: +33(0)344618126

Abstract

Hypothesis. Surface tension of aqueous solutions of surfactants at their critical micelle concentrations (γ_{CMC}), may be quantitatively linked to the surfactant structure using Quantitative Structure Property Relationships (QSPR), all other factors held equal (temperature, presence of additive or salts). Thus, QSPR models can allow improved understanding and quantification of structure-γ_{CMC} trends, direct γ_{CMC} predictions, and finally help to design renewable substitutes for petroleum-based surfactants.

Experiments and methods. A dataset of 70 γ_{CMC} of single surfactants at ambient temperature has been gathered from several research papers. Then, descriptors of the whole structure, of polar heads and of alkyl chains of the 70 surfactants were calculated and introduced in multilinear regressions to evidence the most predictive and physically meaningful structure property relationships.

Findings. The best model, based on quantum chemical descriptors, achieved a standard error of 2.4 mN/m on an external validation. Simpler models were also achieved based solely on the count of H atoms of the polar head but with prediction error of 2.9 mN/m. Among all identified factors affecting γ_{CMC} of sugar-based surfactants (polar head size, alkyl chain length and branching), polar head size was found to exhibit the only effect clearly taken into account by all the models.

Keywords: sugar-based surfactant; quantitative structure-property relationships; surface tension at critical micelle concentration; bio-based chemical

1 Introduction

In recent years, biorefinery [1] attracted a growing interest, for academic and industrial researchers, toward the access of chemical building blocks and specialties from renewable resources. Among the products than can be issued from biorefinery, surfactants are already quite well-implemented in the market, and notably sugar-based surfactants [2]. These compounds are amphiphilic molecules comprising at least one hydrophilic part, the polar head, and one hydrophobic part, the alkyl chain. They are present in many applications like detergents, cosmetics, inks, paints. They are also used to collect valuable materials such as minerals (froth flotation) and oil (Enhanced Oil Recovery), or to facilitate protein identification and analysis [3].

Sugar-based surfactants can be either directly obtained from biorefinery or indirectly, through renewable building blocks [4]. These surfactants are increasingly used as substitutes to conventional petroleum-based ones (notably ethylene-oxide nonionic surfactants) due to their renewability and similar performances in applications [2]. Their structures can be very diverse and sometimes complex, consisting of polar heads with many alcohol moieties in particular configurations, pyranose cycles etc. [5]. To evidence the best sugar-based surfactant for target applications, large experimental screening should be performed. However, since such intensive experimental syntheses and characterizations are time and cost expensive, estimations of the relevant properties of possible candidates using predictive methods could be of particular interest to select the most promising ones for detailed experiment campaigns [6].

The maximum effectiveness of surface tension reduction is a key indicator of surfactant performances, notably for applications requiring foaming and wetting abilities, like in detergents and cosmetics formulations [7, 8]. Indeed, the lowering of surface tensions originates from the ability of the surfactant to energetically favor the solvent/air interfaces. By the same way, energetically favorable interfaces also allow a better spreading of the liquid on hard surfaces.

The surface tension reduction reaches a maximum at the critical micelle concentration (CMC) and becomes almost constant at higher concentrations of surfactant in the aqueous solution. So, the surface tension at CMC (γ_{CMC}) is used to represent the effectiveness of the surfactant to lower the surface tension of the aqueous solution [7, 8].

The key structure-property trend identified by experimentalists is the increase of γ_{CMC} with polar head size and, a decrease, lower in comparison, with the alkyl chain length [7]. A similarly low influence of the alkyl chain branching has also been observed with a decreasing effect on γ_{CMC} [7]. Quantifying the relationship between these structural elements and γ_{CMC} could help evidencing the relative contributions of the different structural elements and combine them in a single equation. Quantitative Structure Property Relationship (QSPR) approach can be used to achieve such quantification. QSPR models are statistical relationships between molecular structure and properties, using molecular descriptors (e. g., the number of O atoms). The only QSPR models developed including non-petrochemically derived surfactants were dedicated to *CMC*, and no QSPR model applicable to the prediction of γ_{CMC} of sugar-based surfactants was identified in literature. Only three QSPR models, proposed by Wang et al. [9-11], were identified for the effectiveness in surface tension reduction of other classes of surfactants, such as anionic surfactants. Although good performances were obtained, with standard errors between 0.12 and 0.88 mN/m (for databases of 20 to 34 surfactants), none of them were externally validated. Thus, it is not possible to estimate their predictive power. In that context, we developed new validated QSPR models to predict the γ_{CMC} of sugar-based surfactants as done for CMC, in a previous work [13]. We developed QSPR models with different types of descriptors: including quantum chemical descriptors, in order to access physically meaningful models, or only simple topological and constitutional descriptors to favor easy-to-use models. Moreover, in QSPR approach, alkyl chain and polar heads can be characterized separately through fragment-based molecular descriptors (for example, the number of C atoms in the alkyl chains or the molecular weight of the polar head), as already done, e. g., for CMC [13-15]. Thus, considering the particular structure of surfactants, both integral descriptors of the whole surfactants and fragment-based descriptors were investigated. At last, the final models were associated

to previous models for the prediction of CMC to access the full tensiometric curves of sugar-based surfactants under the thermodynamic framework of Abbott's approach.

2 Computational details

2.1 Experimental dataset

The number and quality of the experimental data used to develop a QSPR model are of primary importance to allow good fitting and validation. All uncertainty and lack of homogeneity into these data propagates into the final model leading to increased prediction errors. In the present case, 70 γ_{CMC} for aqueous solutions of sugar-based surfactants without additives (like salts) have been extracted from literature [9-11]. Only γ_{CMC} measured between 293K and 303K (i.e. close to room temperature) were considered, since temperature can influence γ_{CMC} values [16]. The aqueous solutions of the highest purity were also targeted, since impurities can affect γ_{CMC} values [8].

Moreover, whenever possible, the Krafft temperatures of the molecules were checked to be lower than 298K, since surfactants exhibiting Krafft temperatures higher than 298K are not expected to form micelles at room temperature due to solubility issues [17]. In such case, no γ_{CMC} should be expected.

It has to be kept in mind that some variance can still exist for γ_{CMC} measured by different experimentalists, as shown in Table 1. Some large discrepancies might even be observed, like the deviation of 11.6 mN/m between the gathered data for 6-O-Dodecanoylsucrose. Such discrepancy, of yet unclear origin, illustrates the need for a cautious selection of the data then used to train the models, in particular on solubility and purity of surfactants. However, other examples well illustrate the fact that residual uncertainty of several mN/m remains, even after these checks.

<< TABLE 1 >>

The final dataset was constituted of 70 different sugar-based surfactants (Table 2), with various polar heads (cyclic, acyclic and even mixed), with linear, branched and/or unsaturated alkyl chains, and with various linkages (ether, thioether, ester, amide and methylamide). The distribution of γ_{CMC} data is represented in Figure 1, with data ranging between 24.1 mN/m and 44.1 mN/m.

<< TABLE 2 >>

<< FIGURE 1 >>

In the perspective of the external validation of models, the dataset was divided into two parts. The training set used for the development of the model was constituted of 47 surfactants (representing 2/3 of the dataset) whereas the validation set, used to evaluate the predictive power of the models was composed of 23 surfactants (1/3 of the dataset). To ensure that the surfactants of the validation set are at best in the applicability domain of the model, this partition was performed by a property-ranged approach. Surfactants were classified by increasing order of γ_{CMC} and the ones of the validation set were regularly selected (e.g. 2^{nd}, 5^{th}, 8^{th} etc.). The representativeness of the validation set in terms of the chemical diversity was also checked based on a Principal Component Analysis performed using the whole set of computed descriptors. As shown in Figure 2, the molecules of both the training and the validation sets revealed well-distributed in the global chemical space of the investigated surfactants.

<< FIGURE 2 >>

2.2 *Molecular descriptors*

The molecular structures of the 70 studied sugar-based surfactants of the dataset were optimized using the Density Functional Theory (DFT) at B3LYP/6-31+G(d,p) level after preliminary conformation analyses and as defined in recent works, on the impact of the conformations of sugar-based surfactants on molecular descriptors [12]. Frequency calculations were performed at the same level of theory to ensure that each conformation corresponds to a local minimum, i.e. presenting no imaginary frequency.

In the same way, the structures of the 31 hydrophilic (polar heads) and 20 hydrophobic (alkyl chains) fragments constituting the 70 molecules of the dataset were also optimized and checked by frequency calculations at B3LYP/6-31+G(d,p) level after, when necessary, specific conformation analyses. The separation between the polar head and the alkyl chain was set before the first heteroatom, as illustrated in Figure 3. Then, the fragments were hydrogen-saturated. The Gaussian09 [18] suite of programs was used for all these calculations.

<< FIGURE 3 >>

It has to be noticed that some sugar-based surfactants were enantiomeric [19], diastereomeric [20], or anomeric [21, 22] mixtures in aqueous solution. Enantiomeric mixtures comprise surfactants with D- and L-sugar alcohol polar heads. Diastereomeric mixtures originate from surfactants with ramified alkyl chains, with one chiral carbon at the ramification. Finally, anomeric mixtures consist in surfactants with polar heads containing a free anomeric alcohol. For these isomeric mixtures, isomers were considered as different conformations of the same compound. So, the geometries of all relevant isomers were optimized and the most stable one was finally retained.

Based on these quantum chemical structures (surfactants and fragments), more than 300 constitutional, topological, geometrical and quantum-chemical descriptors were computed using CODESSA software [23] for each surfactant and each fragment. Additional descriptors were also obtained directly from the quantum-chemical calculations like descriptors arising from conceptual DFT [24, 25] (electronegativity, hardness, softness and electrophilicity index), partial charges (notably on the polar head, calculated based on Mulliken [26] and Natural Populations Analyses [27] as implemented into Gaussian09 software). Finally, 953 descriptors were calculated for each surfactant: 326 related to their entire structure and 627 related to their fragments (polar head and alkyl chain).

2.3 *Development and validation of the models*

All QSPR models developed in this study consist in Multi Linear Regressions (MLR) with the general form of Eq. 1:

$$\gamma_{CMC} = a_0 + \sum_i a_i D_i \qquad (1)$$

where D_i is the descriptor i, and a_0 is the regression coefficient of D_i, and a_0 is the intercept.

To avoid building overfitted models, a descriptor selection was performed with the Best Multi-Linear Regression (BMLR) approach as implemented in CODESSA software [23]. This stepwise variable selection method has been deeply described in previous works [28, 29] and successfully used in

particular in a recent work for the CMC of sugar based surfactants [13]. Among the models proposed by the algorithm, the final model was chosen as the best compromise between correlation and number of descriptors to avoid any over-parameterization.

The goodness of fit of the model was measured by the determination coefficient (R^2), the mean absolute error (MAE) and the root mean square error (RMSE) between predicted and experimental values for the training set. Moreover, Student's t-test at a confidence level of 95% was performed to check the relevance of each descriptor into the regression.

Leave-one-out (LOO) and leave-many-out (LMO) cross-validations were used to assess the robustness of the model via the Q^2_{CV}, Q^2_{3CV}, Q^2_{7CV} and Q^2_{10CV} coefficients (for LOO, 3-fold, 7-fold and 10-fold cross-validations, respectively). Robust models are expected to present high Q^2 values, close to R^2 and one close to each other.

Moreover, to ensure that models did not issue from chance correlations, a Y-scrambling test [30] was realized. Random permutations of experimental property values within the training set were performed (500 iterations) and new models were refitted. To evaluate the impact of randomization, average (R^2_{YS}) and standard deviation (SD_{YS}) in the R^2 of the new models were calculated. Low R^2_{YS} are expected to avoid chance correlation. Rücker [30] proposed that R^2_{YS} should be superior to 2.3 SD_{YS} for a model to be considered as not issued from chance correlations.

Then, an external validation was performed by applying the model to the molecules of the validation set to evaluate its predictive power. The coefficient of determination R^2_{EXT}, the mean absolute error MAE_{EXT} and the root mean square error $RMSE_{EXT}$ were calculated. In addition, a series of external validation metrics were used: Q^2_{F1} [31], Q^2_{F2} [32], Q^2_{F3} [33], CCC [34], $\overline{r^2_m}$ and Δr^2_m [35]. To estimate the reliability of a QSPR model from these metrics, some criteria have been proposed like the threshold values proposed by Chirico et al. [34]: $R^2_{EXT} > 0.70$, $Q^2_{Fn} > 0.70$, CCC > 0.85, $\overline{r^2_m} > 0.65$, $\Delta r^2_m < 0.20$.

At last, the applicability domain (AD) [36, 37] of each model has been defined in terms of ranges of values of the calculated descriptors and the experimental property in the training set. All external

validation metrics presented above were calculated again considering only the molecules of the validation set within the applicability domain (R^2_{IN}, MAE_{IN}, $RMSE_{IN}$, $Q^2_{F1,IN}$, $Q^2_{F2,IN}$, $Q^2_{F3,IN}$, CCC_{IN}, $\overline{r^2_m}_{IN}$, $\Delta r^2_{m,IN}$) and represent the expected predictive power of the model, i.e. inside its applicability domain.

3 Results and discussion

Six new QSPR models were developed in this study depending on the type of descriptors used to build them. Three of them include integral descriptors, i.e. based on the whole surfactant molecule. One used all types of descriptors (i.e. including quantum-chemical descriptors); one was limited to topological and constitutional descriptors; the third one focused on constitutional descriptors to favor simpler models. The three other models were developed on the same scheme focusing on the fragment-based descriptors. The details and predictions issued from each model are available in Supporting Information (Tables S1-S7).

3.1 Performances of the developed QSPR models

Table 3 summarizes the performances of the different models developed in this study. The errors in prediction of the models range from 2.4 to 3.0 mN/m (in terms of $RMSE_{IN}$). The best model includes integral quantum-chemical descriptors (i/all), with $RMSE_{IN}$ = 2.4 mN/m and R^2_{IN} = 0.78.

Although this error is higher than expected experimental uncertainties using a single robust protocol (about 1 mN/m), this remains of the same order of magnitude than the observed variability among experimental γ_{CMC} values collected in the literature (e.g. with deviations within 2.5 mN/m and 2.9 mN/m for dodecanoyl-N-methylglucamine and N-dodecyl lactobionamide, respectively, as illustrated in Table 1). As the models were developed on these collected data, observed errors in prediction are completely relevant and satisfactory.

<< TABLE 3 >>

The only model fulfilling all validation criteria of Chirico is the i/all model (a shown in Table S2). So this model, based on the integral descriptor and including quantum chemical ones, is recommended to

access the most accurate γ_{CMC} predictions. But it requires preliminary quantum chemical calculations that involve some theoretical chemistry knowledge and facilities (Gaussian software). In the perspective of in silico molecular design, simpler models were also looked for, to allow the fast screening of large number of molecular structures. In that context, the two models based only on simple constitutional descriptors (i/c and f/c in Table 4) present the advantage to be based on the only knowledge of the surfactant elemental composition to access first estimation of γ_{CMC} even if they did not fulfill all validation criteria of Chirico. Even if the quantum chemistry based model is recommended for final accurate prediction of γ_{CMC} or to evidence the change of γ_{CMC} only related to the alkyl chain, these simpler models allow faster screening of large series of virtual sugar-based surfactants to evidence those that could present a target range of γ_{CMC}. For these reasons, these three models were further analyzed.

3.2 Model with all types of descriptors

From the 326 integral descriptors calculated for the whole surfactant molecule, a five-parameter model (eq. 2) was found as the best compromise between correlation and number of descriptors among the 15 equations sorted out by the BMLR method:

$$\gamma_{CMC} = 1.014\, n_O - 66.8\, V_{O,avg} - 444.0\, HACA_{1,TMSA} + 5.45\, q_{head} - 132.6\, n_{H,rel} + 234.58 \quad (2)$$

with n_O the number of O atoms, $V_{O,avg}$ the average valency of a O atom, $HACA_{1,TMSA}$ the relative hydrogen acceptor charged surface area, q_{head} the Mulliken partial charge of the polar head, and $n_{H,rel}$ the relative number of H atoms.

The model is characterized by good fitting (R^2 = 0.92, RMSE = 1.4 mN/m) and robustness ($Q^2_{LOO} = Q^2_{10CV} = Q^2_{7CV} = Q^2_{3CV}$ = 0.90). Moreover, the criterion of Rücker [30] for Y-scrambling validation is fulfilled ($R^2 - R^2_{YS} = 0.81 > 2.3SD_{YS} = 0.15$) ensuring against chance correlation.

As shown in Figure 4, a good predictive power is also obtained for the 18 molecules in AD out of the 23 molecules of the validation set (R^2_{IN} = 0.78, $RMSE_{IN}$ = 2.4 mN/m, $Q^2_{F1,IN}$ = 0.74, $Q^2_{F2,IN}$ = 0.73, $Q^2_{F3,IN}$ = 0.76, CCC_{IN} = 0.86, $\overline{r^2_m}_{IN}$ = 0.69, $\Delta r_m^2{}_{IN}$ = 0.10), all the criteria of Chirico et al. [34] being satisfied.

<< FIGURE 4 >>

When analyzing deeper the five molecules falling out of AD (presented in Table 4), four of them presented values of descriptors out of AD, but very close to the AD limits. Besides, for these surfactants, calculated γ_{CMC} revealed close to experiments. The last one, octyl glycol (Figure 5) presents a larger error (6.8 mN/m) and is also the only one to be significantly out of AD for several variables. This could be due to its particularly small polar head, with only one free alcohol moiety. So, the model may be less efficient for such surfactants and it is not surprising to find them out of its AD.

<< TABLE 4 >>

<< FIGURE 5 >>

According to the t-test values, the descriptor contributing the most to the prediction of γ_{CMC} in the model is the number of oxygens n_O, with predicted γ_{CMC} increasing with n_O. In sugar-based surfactants, which are polyhydroxylated, usually, the larger the surfactant, the larger is the number of oxygen atoms. So, n_O seems to account for the size of the polar head (as illustrated in Figure S1), in agreement with the main identified trend that γ_{CMC} increases with the size of the polar head. Besides, the same structural trend is involved in the main descriptors of the six models developed in this study (as shown in Figures S1-S5).

A secondary structural trend is identified in this model. The relative number of H atoms, $n_{H,rel}$, increases with the length of the alkyl chain for an identical polar head. Thus, this descriptor might contribute to account for the impact of the alkyl chain length on γ_{CMC}.

3.3 Models with constitutional descriptors

In the perspective of faster estimation of γ_{CMC}, without any quantum chemical calculation, two alternative models, based on constitutional descriptors, were also highlighted.

The first, two-parameter, model (in Eq. 3) was chosen among the 6 equations sorted out by the BMLR method when focusing on the only 36 constitutional integral descriptors:

$$\gamma_{CMC} = 1.026\ n_O + 1.838\ n_N + 22.84 \quad (3)$$

with n_O the number of O atoms and n_N the number of N atoms.

A fitting with experimental data was obtained at a level of R^2 of 0.78 and RMSE = 2.2 mN/m. This model proved to be robust in cross-validation with $Q^2_{CV} = Q^2_{10CV} = Q^2_{3CV} = 0.75$ and $Q^2_{7CV} = 0.77$. The Y-scrambling ensured that the model was not issued from chance correlation since low values of R^2 were found for the models obtained after randomization with $R^2_{YS} = 0.04$ and $SD_{YS} = 0.04$.

As noticed in previous section, the predictive performances of the model revealed lower than Eq. 2, and did not fulfill the criteria of Chirico: $R^2_{IN} = 0.67$, $Q^2_{F1,IN} = 0.60$, $Q^2_{F2,IN} = 0.60$, $Q^2_{F3,IN} = 0.64$, $CCC_{IN} = 0.76$, $\overline{r^2_m}_{IN} = 0.52$, $\Delta r_m^2{}_{IN} = 0.27$ (as illustrated in Figure 6).

<< FIGURE 6 >>

The last developed model was obtained focusing on the 76 constitutional fragment-based descriptors. The best compromise between correlation and number of descriptors was identified as the simple one-parameter model in Eq. 4.

$$\gamma_{CMC} = 0.568\ n_{H,h} + 21.86 \quad (4)$$

with $n_{H,h}$ the number of hydrogen atoms of the polar head.

The fitting performance on the training set was similar to Eq. 3, with $R^2 = 0.81$ and RMSE = 2.1 mN/m and a good robustness was characterized by cross-validation with $Q^2_{CV} = 0.79$, $Q^2_{10CV} = 0.79$, $Q^2_{7CV} = 0.80$ and $Q^2_{3CV} = 0.80$. The predictive power was lower than Eq. 2, with $RMSE_{IN} = 2.9$ mN/m (cf. Figure 7) and none of the Chirico criteria were fulfilled ($R^2_{IN} = 0.69$, $Q^2_{F1,IN} = 0.60$, $Q^2_{F2,IN} = 0.60$, $Q^2_{F3,IN} = 0.64$, $CCC_{IN} = 0.75$, $\overline{r^2_m}_{IN} = 0.51$, $\Delta r_m^2{}_{IN} = 0.28$).

<< FIGURE 7 >>

With its small polar head, octyl glycol was the only surfactant of the validation set out of the applicability domain of both models, with a slightly too small number of hydrogen and oxygen atoms in its polar

head (6 vs. AD range of [8;42] for $n_{H,h}$ and 2 vs. AD range of [3;22] for n_O), even if its error revealed small for both models (1.8 for Eq. 3 and 1.4 mN/m for Eq. 4).

In these two equations, the main descriptors, n_O and $n_{H,h}$ (in Eq. 3 and Eq. 4, respectively), quantify the increase of γ_{CMC} with polar head size (as shown in Figures S1 and S5, respectively). Moreover, it is interesting to note the presence of n_N in Eq. 4. With its positive regression coefficient (+1.838), this descriptor seems to account for the contribution of amide or amine moieties to increase the polar head size (which itself is observed to increase γ_{CMC}). Thus, it can be argued that both models only reflect the impact of polar head size on γ_{CMC}, which confirms polar head size as the dominant structural factor when compared to others such as alkyl chain length or branching.

To the end, although the statistical performances of Eq. 3 and 4 are lower than those of Eq. 2, they can be useful for pre-screening purposes, as only the raw formula of the polar head, e.g. $C_6H_{12}O_6$, is needed to apply them and obtain a first estimation of γ_{CMC}, with a standard error in prediction of 2.9-3.0 mN/m.

3.4 Application: prediction of tensiometry curves.

In our recent work [13], two QSPR models have been proposed to predict the critical micelle concentration of sugar-based surfactants. The first one (in Eq. 5) included quantum chemical descriptors of the whole structure of surfactant with a low error in prediction of $RMSE_{IN}$ = 0.32 (log).

$$log\ CMC = -1.83\ {}^{1}AIC - 3.70\ {}^{2}ACIC + 3.99 \cdot 10^{-2}\ \eta + 0.209\ T_e + 1.08 \quad (5)$$

with ${}^{1}AIC$ the Average Information Content of order 1, ${}^{2}ACIC$ the Average Complementary Information Content of order 2, η the hardness, and T_e the topographic electronic index calculated from all atomic pairs using Zefirov's partial charge model [38].

The second one (in Eq. 6) is a simpler fragment-based model including only constitutional descriptors with a slightly higher error in prediction with $RMSE_{IN}$ = 0.36 (log).

$$log\ CMC = -20.00\ n_{rel,S,h} - 2.65 \cdot 10^{-2}\ M_{w,c} - 63.78\ n_{rel,single,c} + 64.75 \quad (6)$$

with $n_{rel,S,h}$ the relative number of S atoms in the polar head, $M_{w,c}$ the molecular weight of the alkyl chain and $n_{rel,single,c}$ the relative number of single bonds in the alkyl chain.

Based on these models for CMC and the new ones for γ_{CMC}, it is possible to predict the surface tension of the aqueous solution as a function of surfactant concentration (i.e. tensiometry curves) by using the approximations of the web interactive model of Abbott [39], presented in his book [40].

In the first assumption, the surface tension is linked to the concentration curve under the consideration of the Langmuir-Szyszkowski isotherm by Eq. 7.

$$\gamma = \gamma_w - RT\,\Gamma_m \ln(1 + KC) \quad (7)$$

in which γ is the surface tension (in N/m), γ_w is the surface tension of water (0.0728 N/m at 298 K), R is the ideal gas constant (8.314 J/K/mol), T is the temperature (in K), Γ_m is the limiting surface concentration, K is the absorption constant (in L/mol), and C is the concentration (in mol/L).

In his second assumption, Abbott considers that the surfactant concentration resulting in a 20 mN/m surface tension decrease (C_{20}) is ten times lower than CMC. At last, the third assumption considered, based on Rosen's work [7], that, starting from C_{20}, the surface is saturated in surfactants and that surface tension becomes constant after CMC. Based on these three assumptions, a series of equations, summarized in Eqs. 8 and 9, were derived allowing a1ccess to Γ_m and K.

$$\Gamma_m = \frac{0.02 + \gamma_{CMC} - \gamma_w}{RT \ln(0.1)} \quad (8)$$

$$K = \frac{\exp\left(\frac{\gamma_w - \gamma_{CMC}}{RT\Gamma_m}\right) - 1}{CMC} \quad (9)$$

Thus, using Eqs. 8 and 9 in combination with Eq. 7, the entire tensiometry curve can be estimated for a given surfactant only from the knowledge of CMC and γ_{CMC}, for instance from QSPR models. In this study, this approach was tested for nine surfactants using either experimental values of γ_{CMC} and CMC

or predicted ones (from the quantum chemical models, Eqs. 2 and 5 or from simpler ones, Eqs. 4 and 6). Calculation parameters and results are provided in Supporting Information (Table S8).

The nine investigated surfactants were selected since they were included in the validation sets of both the models for CMC and for γ_{CMC}, and sufficiently detailed surface tension/concentration curves (i.e. with at least 8 or 9 experimental data points) were available for all of them. These surfactants are quite structurally diverse, including one or two cyclic and/or acyclic sugar units in the polar heads, short to long alkyl chains, and with different kinds of linkages, as illustrated in Supporting Information (Figure S6).

From a general point of view, relatively good agreement was found between experimental and predicted surface tensions on the tensiometric curves proposed in Supporting Information (Figures S7-S15). As shown in Table 5, low errors were obtained from Abbott's assumptions alone, i.e. from experimental values of CMC and γ_{CMC}, demonstrating the relevance of the approach for the studied sugar-based surfactants.

<< TABLE 5 >>

The Abbott's assumption failed only for one surfactant, S-hexyl-1-thio-D-lyxitol, as shown in Figure S14. For this case study, surface tension decrease more with concentration than expected. The hypothesis of reduction of 20 mN/m at 10% of CMC may be not valid.

When using predicted values of CMC and γ_{CMC}, the observed deviations are in line with the prediction errors on these two parameters. Low errors on both properties result in a good agreement between the calculated curve and the experimental data, as for nonyl-β-D-glucoside in Figure 8(a). Then, larger errors on CMC lead to a shifting of the breaking point (associated to CMC), as seen for N-decyl-N-methyl gluconamide in Figure 8(b). Errors on γ_{CMC} affect more particularly the surface tensions at concentrations higher than CMC by increasing or decreasing the surface tension at this constant level, as in the case of S-octyl 1-thio-D-lyxitol in Figure 8 (c).

<< FIGURE 8 >>

Overall, it can be seen that combining the Abbott's approximations with predicted γ_{CMC} and CMC from QSPR models represents a promising approach to estimate tensiometry curves of sugar-based surfactants from the only knowledge of their molecular structure.

4 Conclusion

In this study, QSPR models for the prediction of γ_{CMC} of sugar-based surfactants are proposed. The model presenting the highest predictive power included quantum chemical descriptors. Easier to use but less predictive models were obtained with only constitutional descriptors. Their detailed analyses confirmed that γ_{CMC} can be predicted from the sole molecular structures of sugar-based surfactants, for single surfactant solutions at ambient temperature, and enabled to better elucidate structural trends. Specifically, all models emphasize the primary role of polar head size for γ_{CMC} estimation with respect to other structural factors (especially in the simplest ones), possibly masked by the experimental uncertainty, and quantify its impact on γ_{CMC}. Associated with predictive models of CMC, these models even allow an estimation of the entire tensiometric curves of a surfactant from the only knowledge of its molecular structure using the Abbott's approximations. The developed models thus yield useful information that may contribute to anticipate foaming and wetting potentials of surfactant solutions at various concentrations. Previous QSPR models were proposed for γ_{CMC} of ethylene oxide derivatives [9, 11], alkyl sulfonates and alkyl sulfates [10], but none of them were applicable to sugar-based surfactants. So, these models fill a gap in the anticipation of the properties of these promising alternatives to petroleum-based surfactants. To improve the predictive capabilities of these new models, systematic series of measurements using homogeneous methods, carried out on surfactants solutions of highest possible purity and confirmed water solubility, would be beneficial, in particular to be able to better take into account secondary structural trends like alkyl chain length and branching. However, as demonstrated in this study, these models, complementary to previous predictive models of CMC [13, 41], represent powerful tools towards a faster design of bio-based sugar-based surfactants even before their synthesis.

Acknowledgements

This work was performed, in partnership with the SAS PIVERT, within the frame of the French Institute for the Energy Transition (Institut pour la Transition Energétique (ITE) P.I.V.E.R.T. (www.institut-pivert.com) selected as an Investment for the Future ("Investissements d'Avenir"). This work was supported, as part of the Investments for the Future, by the French Government under the reference ANR-001-01. Calculations were performed using HPC resources from GENCI-CCRT (Grant 2013-t2013086639).

Supporting Information

Experimental and predicted γ_{CMC} values from the developed QSPR models (Table S1). Details of the developed QSPR models (Tables S2-S7). Structures of the surfactants investigated using Abbott's approximations, associated parameters and predicted tensiometric curves (Figures S1-S10 and Table S8).

References

[1] S. Octave, D. Thomas, Biorefinery: Toward an industrial metabolism, Biochimie 91 (2009) 659-664.
[2] D. Alba, Marchés actuels des produits biosourcés et évolutions à horizons 2020 et 2030, ADEME, 2015.
[3] E.P. Carpenter, K. Beis, A.D. Cameron, S. Iwata, Overcoming the challenges of membrane protein crystallography, Curr. Opin. Struct. Biol. 18 (2008) 581-586.
[4] M. Kjellin, I. Johansson, Surfactants from Renewable Resources, 1 ed., John Wiley & Sons, Ltd, 2010.
[5] C.C. Ruiz, Sugar-Based Surfactants: Fundamentals and Applications, CRC Press, Taylor & Francis Group, 2009.
[6] G. Fayet, P. Rotureau, How to use QSPR-type approaches to predict properties in the context of Green Chemistry, Biofuels, Bioprod. Biorefin. 10 (2016) 738-752.
[7] M.J. Rosen, J.T. Kunjappu, Surfactants and Interfacial Phenomena, 4th ed., John Wiley & Sons, Inc., 2012.
[8] D. Myers, Surfactant Science and Technology, 3rd ed., Wiley-Interscience, 2006.
[9] Z.W. Wang, G.Z. Li, J.H. Mu, X.Y. Zhang, A.J. Lou, Quantitative Structure-Property Relationship on Prediction of Surface Tension of Nonionic Surfactants, Chin. Chem. Lett. 13 (2002) 363-366.
[10] Z.W. Wang, D.Y. Huang, G.z. Li, X.Y. Zhang, L.L. Liao, Effectiveness of Surface Tension Reduction by Anionic Surfactants—Quantitative Structure–Property Relationships, J. Disper. Sci. Technol. 24 (2003) 653-658.
[11] Z.W. Wang, J.L. Feng, H.J. Wang, Z.G. Cui, G.Z. Li, Effectiveness of Surface Tension Reduction by Nonionic Surfactants with Quantitative Structure-Property Relationship Approach, J. Disper. Sci. Technol. 26 (2005) 441-447.
[12] T. Gaudin, P. Rotureau, I. Pezron, G. Fayet, Conformations of n-alkyl-α/β-d-glucopyranoside surfactants: Impact on molecular properties, Comp. Theor. Chem. 1101 (2017) 20-29.
[13] T. Gaudin, P. Rotureau, I. Pezron, G. Fayet, New QSPR Models to Predict the Critical Micelle Concentration of Sugar-Based Surfactants, Ind. Eng. Chem. Res. 55 (2016) 11716-11726.
[14] P.D.T. Huibers, V.S. Lobanov, A.R. Katritzky, D.O. Shah, M. Karelson, Prediction of Critical Micelle Concentration Using a Quantitative Structure−Property Relationship Approach. 1. Nonionic Surfactants, Langmuir 12 (1996) 1462-1470.
[15] A.R. Katritzky, L.M. Pacureanu, S.H. Slavov, D.A. Dobchev, M. Karelson, QSPR Study of Critical Micelle Concentrations of Nonionic Surfactants, Ind. Eng. Chem. Res. 47 (2008) 9687-9695.
[16] M.E. Mahmood, D.A.F. Al-Koofee, Effect of Temperature Changes on Critical Micelle Concentration for Tween Series Surfactant, GJSFR-B: Chemistry 13 (2013) 1-7.
[17] Y. Moroi, Relationship between solubility and micellization of surfactants: The temperature range of micellization, in: K. Hummel, J. Schurz (Eds.), Dispersed Systems, Steinkopff, 1988, 55-61.
[18] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G.A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H.P. Hratchian, A.F. Izmaylov, J. Bloino, G. Zheng, J.L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J.A. Montgomery, J.E. Peralta, F. Ogliaro, M. Bearpark, J.J. Heyd, E. Brothers, K.N. Kudin, V.N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J.M. Millam, M. Klene, J.E. Knox, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, R.L. Martin, K. Morokuma, V.G. Zakrzewski, G.A. Voth, P. Salvador, J.J. Dannenberg, S. Dapprich, A.D. Daniels, Farkas, J.B. Foresman, J.V. Ortiz, J. Cioslowski, D.J. Fox, Gaussian 09, Revision B.01, Wallingford CT, 2009.
[19] M.P. Savelli, P. Van Roekeghem, O. Douillet, G. Cavé, P. Godé, G. Ronco, P. Villa, Effects of tail alkyl chain length (n), head group structure and junction (Z) on amphiphilic properties of 1-Z-R-d,l-xylitol compounds (R=CnH2n+1), Int. J. Pharm. 182 (1999) 221-236.
[20] S. Matsumura, K. Imai, S. Yoshikawa, K. Kawada, T. Uchibori, Surface Activities, Foam Suppression, Biodegradability and Antimicrobial Properties of s-Alkyl Glucopyranosides, J. Jpn. Oil. Chem. Soc. 40 (1991) 709-714.

[21] U.R.M. Kjellin, P.M. Claesson, E.N. Vulfson, Studies of N-Dodecyllactobionamide, Maltose 6'-O-Dodecanoate, and Octyl-β-glucoside with Surface Tension, Surface Force, and Wetting Techniques, Langmuir 17 (2001) 1941-1949.
[22] C. Boyère, G. Broze, C. Blecker, C. Jérôme, A. Debuigne, Monocatenary, branched, double-headed, and bolaform surface active carbohydrate esters via photochemical thiol-ene/-yne reactions, Carbohydr. Res. 380 (2013) 29-36.
[23] Codessa, www.semichem.com/codessa/.
[24] H. Chermette, Chemical reactivity indexes in density functional theory, J. Comput. Chem. 20 (1999) 129-154.
[25] P. Geerlings, F. De Proft, W. Langenaeker, Conceptual density functional theory, Chem. Rev. 103 (2003) 1793-1873.
[26] R.S. Mulliken, Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I, J. Chem. Phys. 23 (1955) 1833-1840.
[27] A.E. Reed, R.B. Weinstock, F. Weinhold, Natural population analysis, J. Chem. Phys. 83 (1985) 735-746.
[28] G. Fayet, P. Rotureau, L. Joubert, C. Adamo, QSPR modeling of thermal stability of nitroaromatic compounds: DFT vs. AM1 calculated descriptors, J. Mol. Model. 16 (2010) 805-812.
[29] G. Fayet, P. Rotureau, L. Joubert, C. Adamo, Development of a QSPR model for predicting thermal stabilities of nitroaromatic compounds taking into account their decomposition mechanisms, J. Mol. Model. 17 (2010) 2443-2453.
[30] C. Rücker, G. Rücker, M. Meringer, y-Randomization and Its Variants in QSPR/QSAR, J. Chem. Inf. Model. 47 (2007) 2345-2357.
[31] A. Tropsha, P. Gramatica, V.K. Gombar, The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, J. Comb. Sci. 22 (2003) 69-77.
[32] G. Schüürman, R.-U. Ebert, J. Chen, B. Wang, R. Kühne, External Validation, Prediction employing the predictive squared correlation coefficient - test set activity mean vs. training set activity mean, J. Chem. Inf. Model. 48 (2008) 2140-2145.
[33] V. Consonni, D. Ballabio, R. Todeschini, Comments on the Definition of the Q2 Parameter for QSAR Validation, J. Chem. Inf. Model. 49 (2009) 1669-1678.
[34] N. Chirico, P. Gramatica, Real External Predictivity of QSAR models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection, J. Chem. Inf. Model. 52 (2012) 2044-2058.
[35] K. Roy, I. Mitra, S. Kar, P.K. Ojha, R.N. Das, H. Kabir, Comparative Studies on Some Metrics for External Validation of QSPR Models, J. Chem. Inf. Model. 52 (2012) 396-408.
[36] J. Jaworska, N.-J. Nina, T. Aldenberg, QSAR applicability domain estimation by projection of the training set descriptor space: a review., Altern. Lab. Anim. 33 (2005) 445-459.
[37] A. Gissi, D. Galadeta, M. Floris, S. Olla, A. Carotti, E. Novellino, E. Benfenati, O. Nicolotti, An Alternative QSAR-Based Approach for Predicting the Bioconcentration Factor for Regulatory Purposes, Altex 31 (2014) 23-36.
[38] N.S. Zefirov, M.A. Kirpichenok, F.F. Izmailov, M.I. Trofimov, Scheme for the Calculation of the Electronegativities of Atoms in a Molecule in the Framework of Sanderson's Principle., Dokl. Akad. Nauk. SSSR 296 (1987) 883-887.
[39] S. Abbott, CMC and Γ. www.stevenabbott.co.uk/practical-surfactants/cmc.php. (Accessed 24/11/17 2016).
[40] S. Abbott, Surfactant Science: Principles and Practice, 2016.
[41] N. Anoune, M. Nouiri, Y. Berrah, J.-Y. Gauvrit, P. Lanteri, Critical micelle concentrations of different classes of surfactants: A quantitative structure property relationship study, J. Surfact. Deterg. 5 (2002) 45-53.
[42] S. Matsumura, K. Imai, S. Yoshikawa, K. Kawada, T. Uchibor, Surface activities, biodegradability and antimicrobial properties of n-alkyl glucosides, mannosides and galactosides, J. Am. Oil Chem. Soc. 67 (1990) 996-1001.
[43] K. Shinoda, T. Yamanaka, K. Kinoshita, Surface Chemical Properties in Aqueous Solutions of Non-ionic Surfactants Octyl Glycol Ether, α-Octyl Glyceryl Ether and Octyl Glucoside, J. Phys. Chem. 63 (1959) 648-650.

[44] K. Shinoda, T. Yamaguchi, R. Hori, The surface tension and the critical micelle concentration in aqueous solution of β-D-alkyl glucosides and their mixtures, Bull. Chem. Soc. Jpn. 34 (1961) 237-241.
[45] H.-K. Yuan, Q.-X. Li, Y.-L. Li, Study on Properties of N-Lauryl-N-Methylglucamide, China Surf. Det. Cosm. 40 (2010) 170-173.
[46] Y.-P. Zhu, M.J. Rosen, P.K. Vinson, S.W. Morrall, Surface Properties of N-Alkanoyl-N-methyl Glucamines and Related Materials, J. Surfact. Deterg. 2 (1999) 357-362.
[47] L. Syper, K.A. Wilk, A. Sokołowski, B. Burczyk, Synthesis and surface properties of N-alkylaldonamides, in: G.J.M. Koper, D. Bedeaux, C. Cavaco, W.F.C. Sager (Eds.), Trends in Colloid and Interface Science XII, Steinkopff 1998, 199-203.
[48] M. Ferrer, F. Comelles, F.J. Plou, M.A. Cruces, G. Fuentes, J.L. Parra, A. Ballesteros, Comparative Surface Activities of Di- and Trisaccharide Fatty Acid Esters, Langmuir 18 (2002) 667-673.
[49] I. Söderberg, C.J. Drummond, D. Neil Furlong, S. Godkin, B. Matthews, Non-ionic sugar-based surfactants: Self assembly and air/water interfacial activity, Colloids Surf., A 102 (1995) 91-97.
[50] N. Becerra, C. Toro, A.L. Zanocco, E. Lemp, G. Günther, Characterization of micelles formed by sucrose 6-O-monoesters, Colloids Surf., A 327 (2008) 134-139.
[51] J. Lalot, I. Stasik, G. Demailly, D. Beaupère, P. Godé, Synthesis and amphiphilic properties of S-alkylthiopentonolactones and their pentitol derivatives, J. Colloid Interface Sci. 273 (2004) 604-610.
[52] H. Minamikawa, M. Hato, Headgroup effects on phase behavior and interfacial properties of β-3,7-dimethyloctylglycoside/water systems, Chem. Phys. Lipids 134 (2005) 151-160.
[53] R. Aveyard, B.P. Binks, J. Chen, J. Esquena, P.D.I. Fletcher, Surface and Colloid Chemistry of Systems Containing Pure Sugar Surfactant, Langmuir 14 (1998) 4699-4709.
[54] J.A. Molina-Bolívar, J. Aguiar, J.M. Peula-García, C.C. Ruiz, Surface Activity, Micelle Formation, and Growth of n-Octyl-β-d-Thioglucopyranoside in Aqueous Solutions at Different Temperatures, J. Phys. Chem. B 108 (2004) 12813-12820.
[55] D. Plusquellec, C. Brenner-Hénaff, P. Léon-Ruaud, S. Duquenoy, M. Lefeuvre, H. Wróblewski, An Efficient Acylation of Free Glycosylamines for the Synthesis of N-Glycosyl Amino Acids and N-Glycosidic Surfactants for Membrane Studies, J. Carbohyd. Chem. 13 (1994) 737-751.
[56] Y. Zhu, M. Durand, V. Molinier, J.-M. Aubry, Isosorbide as a novel polar head derived from renewable resources. Application to the design of short-chain amphiphiles with hydrotropic properties, Green Chem. 10 (2008) 532-540.
[57] P. Boullanger, Y. Chevalier, Surface Active Properties and Micellar Aggregation of Alkyl 2-Amino-2-deoxy-β-d-glucopyranosides, Langmuir 12 (1996) 1771-1776.
[58] C.A. Ericsson, O. Söderman, V.M. Garamus, M. Bergström, S. Ulvenlund, Effects of Temperature, Salt, and Deuterium Oxide on the Self-Aggregation of Alkylglycosides in Dilute Solution. 1. n-Nonyl-β-d-glucoside, Langmuir 20 (2004) 1401-1408.
[59] T. Zhang, R.E. Marchant, Novel Polysaccharide Surfactants: The Effect of Hydrophobic and Hydrophilic Chain Length on Surface Active Properties, J. Colloid Interface Sci. 177 (1996) 419-426.
[60] J.A. Molina-Bolívar, J.M. Hierrezuelo, C. Carnero Ruiz, Self-assembly, hydration, and structures in N-decanoyl-N-methylglucamide aqueous solutions: Effect of salt addition and temperature, J. Colloid Interface Sci. 313 (2007) 656-664.
[61] M. Okawauchi, M. Hagio, Y. Ikawa, G. Sugihara, Y. Murata, M. Tanaka, A Light-Scattering Study of Temperature Effect on Micelle Formation of N-Alkanoyl-N-methylglucamines in Aqueous Solution, Bull. Chem. Soc. Jpn. 60 (1987) 2719-2725.
[62] B. Burczyk, K.A. Wilk, A. Sokołowski, L. Syper, Synthesis and Surface Properties of N-Alkyl-N-methylgluconamides and N-Alkyl-N-methyllactobionamides, J. Colloid Interface Sci. 240 (2001) 552-558.
[63] G. Milkereit, V.M. Garamus, K. Veermans, R. Willumeit, V. Vill, Structures of micelles formed by synthetic alkyl glycosides with unsaturated alkyl chains, J. Colloid Interface Sci. 284 (2005) 704-713.
[64] C.M. Persson, U.R.M. Kjellin, J.C. Eriksson, Surface Pressure Effect of Poly(ethylene oxide) and Sugar Headgroups in Liquid-Expanded Monolayers, Langmuir 19 (2003) 8152-8160.
[65] K. Wilk, L. Syper, B. Burczyk, I. Maliszewska, M. Jon, B. Domagalska, Preparation and properties of new lactose-based surfactants, J. Surfact. Deterg. 4 (2001) 155-161.

Table 1. Different γ_{CMC} values gathered in literature for the same surfactants.

surfactant	γ_{CMC} (mN/m)	reference
Octyl-β-D-Glucoside	30.5	[42]
	30.8	[21]
	31.0	[43]
	31.2	[44]
Dodecanoyl-N-methyl Glucamine	27.5	[45]
	30.0	[46]
N-Dodecyl Lactobionamide	35.1	[47]
	38.0	[21]
6-O-Dodecanoyl Sucrose	31.5	[48]
	37.4	[49]
	43.1	[50]

Table 2. Experimental γ_{CMC} data.

surfactant	γ_{CMC} (mN/m)	T (K)[a]	set[b]	reference
S-Octyl 5-Thio-D-Arabinonolactone	24.1	293	T	[51]
Octyl-D,L-Glycerol	24.5	298	V	[43]
1-Butylhexyl-β-D-Glucoside	25.1	298	T	[20]
S-Hexyl 5-Thio-D-Arabinonolactone	25.1	293	T	[51]
1-O-Nonanoyl-D,L-Xylitol	25.4	298	V	[19]
S-Octyl 5-Thio-D-Xylonolactone	25.7	293	T	[51]
1-O-Decanoyl-D,L-Xylitol	26.0	298	V	[19]
1-Propylheptyl-β-D-Glucoside	26.0	298	T	[20]
Dodecanoyl-N-Methylglyceramine	26.1	298	T	[46]
S-Hexyl 5-Thio-D-Xylonolactone	26.3	293	T	[51]
Octyl Glycol	26.7	298	V	[43]
1-Ethyloctyl-β-D-Glucoside	27.0	298	T	[20]
S-Hexyl 1-Thio-L-Ribitol	27.2	293	T	[51]
3,7-Dimethyloctyl-β-D-Glucoside	27.7	298	V	[52]
1-Methylnonyl-β-D-Glucoside	28.0	298	V	[20]
Decyl-β-D-Galactoside	28.0	298	T	[42]
Dodecanoyl-N-Methylxylamine	28.0	298	T	[46]
S-Hexyl 1-Thio-L-Xylitol	28.2	293	T	[51]
Dodecyl-α-D-Mannoside	28.4	298	T	[42]
6-O-[(Hexyloctyl)-3-Propylsulfide)ethanoyl]-D-Mannose	28.5	298	T	[22]
Decyl-α-D-Mannoside	28.5	298	V	[42]
Decyl-β-D-Glucoside	28.5	298	T	[53]
1-O-Octanoyl-D,L-Xylitol	28.6	298	V	[19]
Octyl-β-D-Thioglucoside	28.7	298	T	[54]
Octanoyl-β-D-Galactosylamine	29.3	298	T	[55]
S-Octyl 1-Thio-D-Lyxitol	29.3	293	V	[51]
Pentyl Isosorbide	29.8	298	T	[56]
1-O-Heptanoyl-D,L-Xylitol	30.0	298	T	[19]
1-O-Pentanoyl-D,L-Xylitol	30.0	298	T	[19]
S-Hexyl 1-Thio-D-Lyxitol	30.0	293	V	[51]
2-Amino-2-Deoxy-Nonyl-β-D-Glucoside	30.2	298	T	[57]
Nonyl-β-D-Glucoside	30.4	293	V	[58]
Octyl-α-D-Mannoside	30.5	298	T	[42]
Octyl-β-D-Galactoside	30.5	298	V	[42]
Octyl-β-D-Glucoside	30.5	298	T	[42]
Dodecyl-β-D-Galactoside	31.5	298	T	[42]
Octyl-D-Maltonamide	31.8	298	T	[59]
Butyl Isosorbide	32.1	298	V	[56]
Hexyl-D-Maltonamide	32.2	298	T	[59]
2-Amino-2-Deoxy-Octyl-β-D-Glucoside	32.5	298	V	[57]
Decanoyl-N-Methylglucamine	32.5	303	T	[60]
Nonanoyl-N-Methylglucamine	32.9	303	T	[61]
Decyl-D-Lactobionamide	33.0	298	T	[47]

1-O-Hexanoyl-D,L-Xylitol	33.2	298	V	[19]
Dodecyl-D-Maltonamide	33.6	298	T	[59]
3,7-Dimethyloctyl-β-D-Maltoside	33.7	298	T	[52]
Decyl-D-Maltonamide	33.9	298	V	[59]
N-Octadecyl-N-Methyl Lactobionamide	34.6	293	T	[62]
N-Oleyl-N-Methyl Gluconamide	34.8	293	T	[62]
N-Tetradecyl-N-Methyl Gluconamide	35.0	293	V	[62]
[N-(Oleoyl)-2 -Ethylamino]-β-D-Maltoside	35.1	298	T	[63]
N-Dodecyl-N-Methyl Gluconamide	35.2	293	T	[62]
Dodecyl-β-D-Maltoside	35.3	295	V	[64]
Decyl-β-D-Maltoside	35.6	298	T	[53]
N-Hexadecyl-N-Methyl Lactobionamide	35.6	293	T	[62]
N-Tetradecyl-N-Methyl Lactobionamide	35.7	293	V	[62]
N-Oleyl-N-Methyl Lactobionamide	35.8	293	T	[62]
N-Dodecyl-N-Methyl Lactobionamide	36.0	293	T	[62]
N-Decyl-N-Methyl Gluconamide	36.7	298	V	[62]
N-Tetradecanoyl-N-Methyl Lactitolamine	37.0	298	T	[65]
N-Dodecanoyl-N-Methyl Lactitolamine	37.2	298	T	[65]
6-O-Dodecanoylsucrose	37.4	293	V	[49]
Oleyl-β-D-Maltoside	37.8	298	T	[63]
N-Decyl-N-Methyl Lactobionamide	38.9	293	T	[62]
6'-O-Dodecanoylmaltose	39.0	295	V	[21]
3,7-Dimethyloctyl-β-D-Maltotrioside	39.6	298	T	[52]
N-Decanoyl-N-Methyl Lactitolamine	40.3	298	T	[65]
Oleyl-β-D-Maltotrioside	42.5	298	V	[63]
6-O-Dodecanoylstachyose	43.0	293	T	[49]
6-O-Dodecanoylraffinose	44.1	293	T	[49]

[a] measurement temperature; [b] T for training set, V for validation set.

Table 3. Summary of the performances of the new QSPR models for surface tension at critical micelle concentration of sugar-based surfactants

	model	n_{desc}	R^2	RMSE (mN/m)	R^2_{IN}	$RMSE_{IN}$ (mN/m)	n_{out}
i/all	integral/all types (Eq. 2)	5	0.92	1.4	0.78	2.4	5
i/ct	integral/constitutional and topological	1	0.75	2.4	0.65	3.0	2
i/c	integral/constitutional (Eq. 3)	2	0.78	2.2	0.67	2.9	1
f/all	fragments/all types	2	0.87	1.8	0.76	2.6	1
f/ct	fragments/constitutional and topological	3	0.84	1.9	0.70	2.6	2
f/c	fragments/constitutional (Eq. 4)	1	0.81	2.1	0.69	2.9	1

n_{desc}: number of descriptors; n_{out}: number of molecules of the validation set out of AD of the model

Table 4. Surfactants out of the AD of Eq. 2

surfactant	out of AD variable(s)	AD ranges	calculated γ_{CMC} (mN/m)	experimental γ_{CMC} (mN/m)
Octyl Glycol	n_O = 2 $n_{H,rel}$ = 0.65 $\gamma_{CMC,pred}$ = 19.9	n_O: [3 ; 22] $n_{H,rel}$: [0.52 ; 0.63] $\gamma_{CMC,pred}$: [24.1 ; 44.1]	19.9	26.7
Octyl-D,L-Glycerol	$n_{H,rel}$ = 0.64	$n_{H,rel}$: [0.52 ; 0.63]	23.6	24.5
S-Hexyl 1-thio-D-Lyxitol	$V_{O,avg}$ = 1.7528	$V_{O,avg}$: [1.7557 ; 1.9500]	29.2	30.0
S-Octyl 1-thio-D-Lyxitol	$V_{O,avg}$ = 1.7528	$V_{O,avg}$: [1.7557 ; 1.9500]	29.7	29.3
Dodecyl-β-D-Maltoside	q_{head} = 0.5719	q_{head}: [-0.6052 ; 0.5532]	37.1	35.3

Table 5. MAE for the surface tensions of surfactant solutions predicted by the Abbott's approach from experimental CMC and γ_{CMC} (exp) or using predicted ones from quantum chemical models (qc) or from constitutional ones (simple)

Surfactant	ref	n_{data}	exp	qc			simple		
			MAE (mN/m)	MAE (mN/m)	Δlog CMC[a] (mol/L)	$\Delta\gamma_{CMC}$[a] (mN/m)	MAE (mN/m)	Δlog CMC[a] (mol/L)	$\Delta\gamma_{CMC}$[a] (mN/m)
6-O-dodecanoylsucrose	[49]	9	0.6	1.3	0.16	1.3	1.1	0.17	3.1
nonyl-β-D-glucoside	[58]	8	0.4	1.2	0.16	0.4	3.1	0.27	1.7
N-decyl-N-methyl gluconamide	[62]	15	1.5	3.9	0.03	4.5	3.7	0.09	6.3
N-tetradecyl-N-methyl gluconamide	[62]	10	0.6	1.2	0.11	2.6	3.6	0.34	4.6
N-tetradecyl-N-methyl lactobionamide	[62]	10	0.9	5.2	0.46	0.8	1.8	0.15	0.3
1-O-hexanoyl-D,L-xylitol	[19]	8	0.7	3.4	0.31	4.4	3.5	0.30	4.5
1-O-nonanoyl-D,L-xylitol	[19]	10	0.9	2.2	0.01	2.9	5.2	0.30	3.3
S-hexyl 1-thio-D-lyxitol	[51]	9	5.6	2.7	0.46	0.8	3.0	0.48	1.3
S-octyl 1-thio-D-lyxitol	[51]	8	1.0	6.4	0.56	0.4	3.2	0.22	0.6

a) Δlog CMC and $\Delta\gamma_{CMC}$: absolute deviation in prediction for log CMC and γ_{CMC}, respectively.

Figure 1. Distribution of γ_{CMC} data within the dataset.

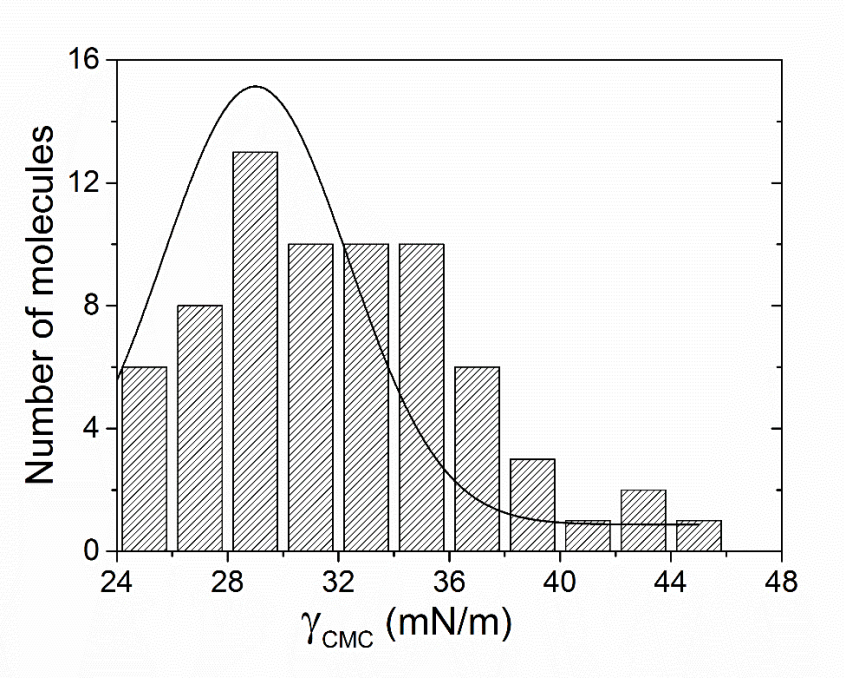


Figure 2. Repartition of molecules of training (circles) and validation (triangles) sets in the whole chemical space of the dataset as defined by Principal Component Analysis based on all calculated descriptors

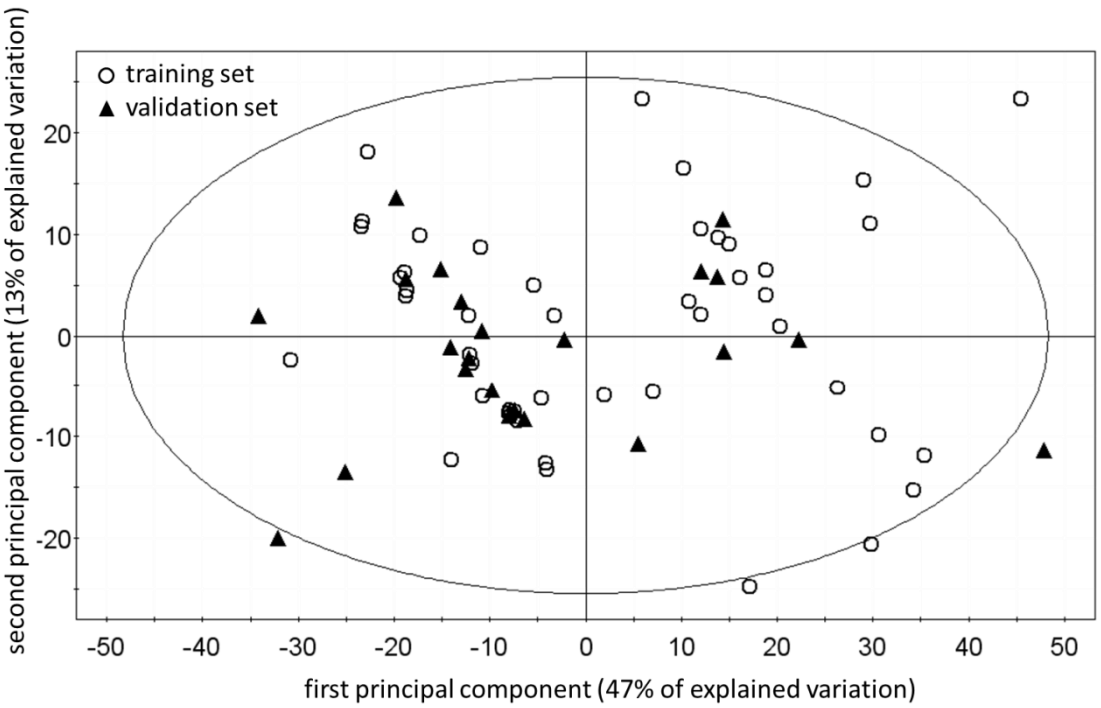


Figure 3. Definition of fragments for the polar head and the alkyl chain for Octyl-β-D-Glucoside.

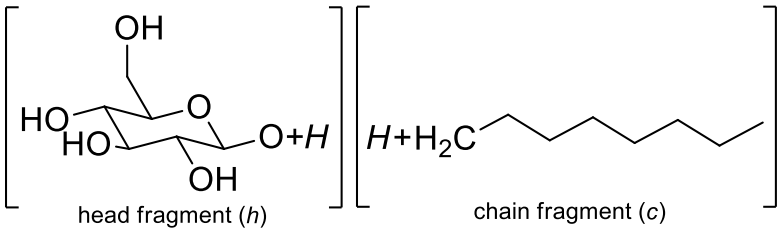


Figure 4. Experimental vs. calculated γ_{CMC} for the model based on integral descriptors of all types (Eq. 2)

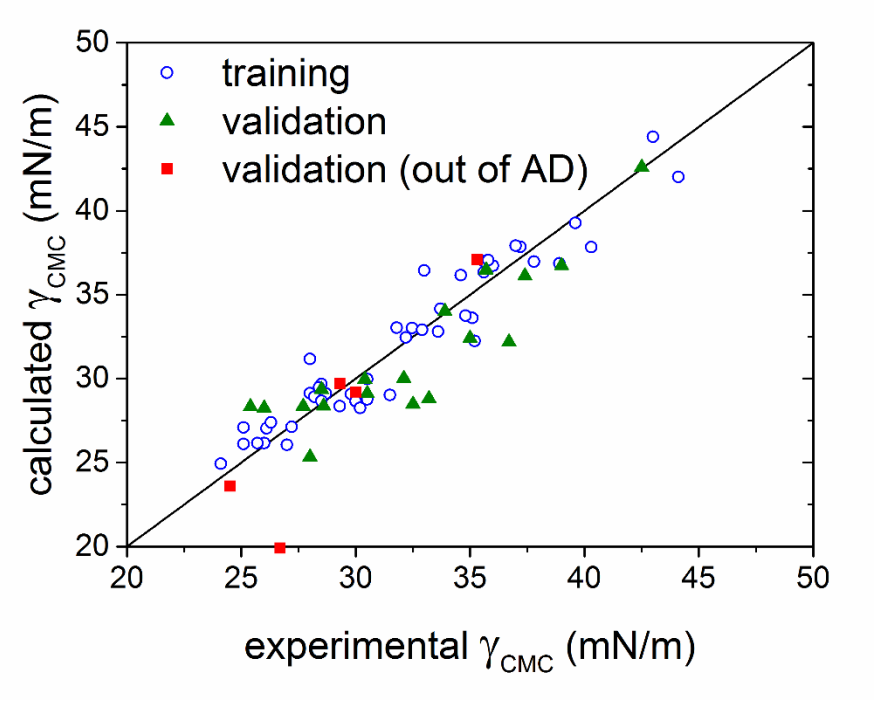
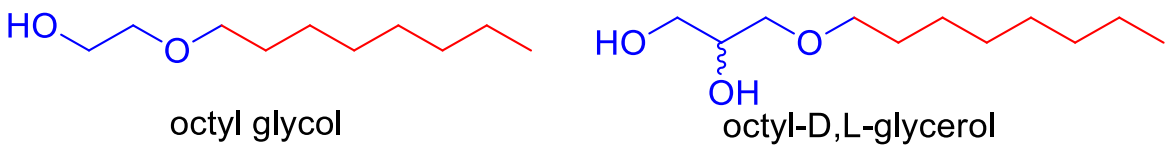


Figure 5. Octyl Glycol and Octyl-D,L-Glycerol



octyl glycol

octyl-D,L-glycerol

Figure 6. Experimental vs. calculated γ_{CMC} for the model based on constitutional integral descriptors (Eq. 3)

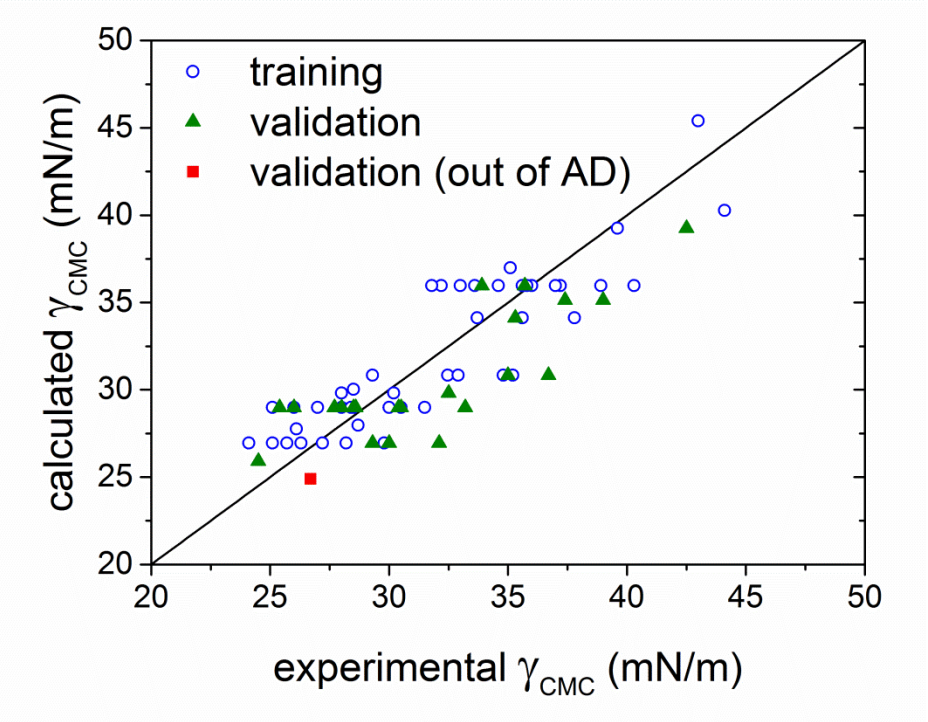


Figure 7. Experimental vs. calculated γ_{CMC} for the model based on constitutional fragment-based descriptors (Eq. 4)

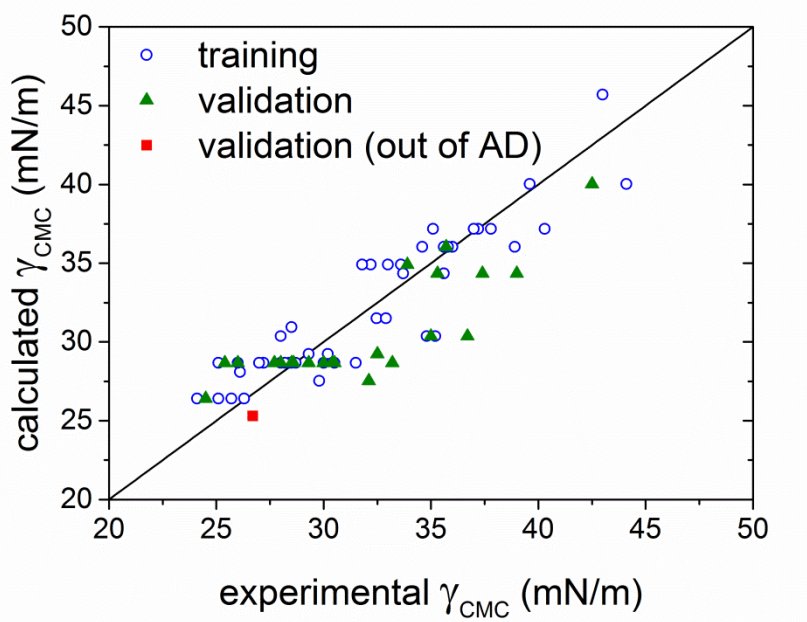


Figure 8. Predicted tensiometric curves vs. experimental tensiometric data for (a) Nonyl-β-D-Glucoside [58], (b) N-Decyl-N-methyl Gluconamide [62], (c) S-Octyl 1-thio-D-Lyxitol [51].