New QSPR models to predict the critical micelle concentration of sugar-based surfactants

Théophile Gaudin, Patricia Rotureau, Isabelle Pezron, Guillaume Fayet

New QSPR Models to predict the Critical Micelle Concentration of Sugar-Based Surfactants

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

a) Sorbonne Universités, Université de Technologie de Compiègne, EA 4297 TIMR, 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

Sugar-based surfactants represent a fruitful field of research in the context of sustainable chemistry since they can be obtained from renewable resources. In this work, new Quantitative Structure Property Relationships (QSPR) models for the critical micelle concentration (CMC) dedicated to sugar-based surfactants are proposed in order to reduce testing in a screening perspective. An important literature compilation allowed the constitution of a dataset of 83 sugar-based surfactants for which accurate CMC values were found. Then, a series of QSPR models were developed based on molecular descriptors of the whole molecule, and of the hydrophobic and hydrophilic fragments taken separately. Different models were considered by including quantum-chemical descriptors with hope to access physically based models, and by using only simple constitutional descriptors to favor fast and easy prediction. The best QSPR model was obtained including quantum-chemical descriptors of the whole molecular structure with a root mean square error (RMSE) of 0.32 (log) evaluated on a validation set of 27 molecules. A simpler model with good performances was also found (with a RMSE of 0.36 (log) on the validation set), including only constitutional-based fragment descriptors, that can be easily computed from the 2-dimension structure of the hydrophilic and hydrophobic fragments.

1 Introduction

Surfactants are amphiphilic molecules, constituted by at least one hydrophilic part (the polar head), and at least one hydrophobic part (the alkyl chain)[1]. This particular configuration favors their self-aggregation into micelles or membranes, and their adsorption at interfaces[1]. Amphiphilic molecules are widely found in nature[2], for example as constituents of cell membranes. They are key components in industrial and customer applications[3], such as agrochemicals, paints, inks, hard surface cleaning, or cosmetics[4]. Surfactants help to collect valuable materials, such as metals through froth flotation[5] or oil by Enhanced Oil Recovery (EOR) process[6]. They are also used to solubilize and crystallize membrane proteins[7] enabling their identification and analysis in biological and medical research[8].

Sugar-based surfactants are an important subfamily of surfactants[9], characterized by their polar head constituted by carbohydrates such as glucose[10], maltose[11] or sucrose[12], and their derivatives. For this reason, sugar-based surfactants can be obtained from renewable resources such as starch[13], and are often biocompatible and easily biodegradable[14]. So, they are commonly considered as environmentally-friendly alternatives to conventional petroleum-based non-ionic surfactants[15], particularly regarding soft detergents or personal care products, cosmetics and pharmaceutical formulations[16].

Surfactants can bear electric charge(s) in the polar head. Cationic surfactants bear positive charge(s) whereas anionic surfactant bear negative charge(s). Non charged surfactants are called non-ionic surfactants. Charge has tremendous impact on surfactant properties, notably increasing critical micelle concentration (CMC) by orders of magnitude due to repulsion between charged surfactant polar heads[1]. Since most commercially available sugar-based surfactants are non-ionic[9], we focused on this class of sugar-based surfactants.

CMC is a fundamental property of surfactants representing the onset of micelle formation in solution[1]. Micelles result from the self-association of surfactant molecules into aggregates in which surfactant polar heads are oriented towards the aqueous phase, whereas alkyl chains are positioned inside the

micelle core[17] to minimize the contact between hydrophobic moieties and water molecules. Crucial changes of solution properties occur above the CMC due to the presence of micelles[1], such as a near constant surface tension, and a rapid increase of light scattering intensity. For applications such as wetting and foaming, which target a low surface tension, increasing surfactant concentration above the CMC will not induce a further decrease of surface tension, but micelles will act as reservoirs of additional surfactants. More importantly, the solubility of organic water-insoluble materials such as lipophilic drugs, dyes or agrochemicals will drastically increase above the CMC due to their solubilisation in hydrophobic micellar cores, which almost behave as a liquid hydrocarbon phase. For such reasons, the knowledge of CMC values serves as a valuable guide for evaluating and comparing the potential of surfactants in various applications, such as detergency and cosmetic formulation where wetting and foaming ability are optimized[3], or drug formulation for which hydrophobic drug solubilization[18] may be an issue. In addition, it has to be noticed that this property is also of interest for regulatory purposes. Indeed, to comply with REACH regulation[19], lipophilicity of substances has to be characterized by the octanol-water partition coefficient, and for surfactants, this partition coefficient has to be measured below CMC[20]. At last, recent works highlighted potential trends between the CMC and the cytotoxicity of sugar-based surfactants[21].

Experimentalists observed, in the case of conventional non-ionic surfactants, that the CMC is primarily influenced by the alkyl chain, and in particular by its size[1, 3, 22]. It is also known to increase notably with branching or unsaturation of the alkyl chain, and slightly with the size of the polar head[1].

CMC can be measured by a variety of experimental techniques[23]. The most commonly used[3] is tensiometry which consists in measuring the surface tension of aqueous solutions of surfactants at several concentrations, and identifying a break in the curve of surface tension versus concentration (in log)[24]. In order to reduce time and cost of experimental screening, the disposal of computational methods to access early estimations of CMC based on the only knowledge of the surfactant molecular structure would be of great interest[25-27].

CMC was estimated for some systems by molecular simulation, e. g. molecular dynamics[28] or coarse-graining[29], with some success. For example, the CMC of madecassic acid was predicted as 96 μM for an experimental value of 62 μM[28]. These approaches, in principle, can be applied to any systems, including sugar-based surfactants. However, surfactant/solvent systems are complex and therefore such direct simulation for these systems remain challenging[26] through significant computation times and careful parameterizations for each calculation.

To predict directly CMC from molecular characteristics, some empirical equations have been also proposed by experimentalists[1, 30, 31] by relating log CMC and the number of C atoms in the alkyl chain (reflecting the size of the hydrophobic part) for homologous series of surfactants, including some sugar-based ones, but each of these can only be used within homologous series of compounds for which it was developed. Moreover, such models are limited to surfactants that bear linear/saturated alkyl chains.

More recently, Quantitative Structure-Property Relationships (QSPR) were also developed to predict the CMC of surfactants[26, 32] . This method already demonstrated its predictive potential for a wide range of properties[27, 33, 34]. QSPR models are mathematical relationships between the molecular structure, characterized by molecular descriptors, and the target property. Various data mining methods are used to derive such models[35] like Multi-Linear Regressions (MLR)[36] or Artificial Neural Networks[37].

Among existing QSPR models for surfactant's CMC, the ones of Huibers et al.[31] and Katritzky et al.[38] are worth mentioning. Indeed, they are based on original fragment-based descriptors calculated from the polar head and the alkyl chain instead of the whole surfactant. Fragment-based descriptors have been already used for various kinds of properties, like boiling point[39] or melting point[40] notably due to their simplicity of interpretation in some properties and versatility[41]. This approach is especially meaningful for surfactants regarding their particular structure, in which polar head and alkyl chain fragments have different impact on amphiphilic properties.

Although many QSPR models have been proposed for CMC, even for non-ionic surfactants[38], none of them are specifically dedicated to the CMC of sugar-based surfactants. Besides, only a few authors even considered sugar-based surfactants[31, 38, 42-47], and only three of their QSPR models were tested on a validation set[43, 45, 46]. Khayamian et al.[43] developed a neural network for the CMCs of non-ionic surfactants (including sugar-based surfactants) validated with an average relative error of 1.5% but on only 5 molecules. Roy et al.[45] developed a MLR model also for CMCs of non-ionic surfactants, for which a RMSE of prediction of 0.40 (log) can be calculated for the 28 molecules used for its validation (with only 3 sugar-based surfactants). To our knowledge, the best model including sugar-based surfactants is the group contribution approach proposed by Mattei et al.[46], developed on the basis of 150 non-ionic surfactants including 40 sugar-based ones. Indeed, they obtained a standard deviation of 0.17 (log) evaluated on a validation set of 30 non-ionic surfactants (of which 10 were sugar-based ones). To further assess its predictive power for sugar-based surfactants specifically, we tested it on the dataset of 83 sugar-based surfactants used in the present study (cf. Computational Details). A RMSE of only 0.93 (log) was found, as shown in Supporting Information, revealing that more accurate QSPR models could be searched towards reliable predictions for the particular family of sugar-based surfactants.

In that context, we developed new QSPR models specifically dedicated to the CMC of sugar-based surfactants. The models are based on different types of descriptors, including quantum chemical descriptors, in order to access physically meaningful models, or only based on simple topological and constitutional descriptors to favor easy-to-use models. Moreover, considering the particular structure of surfactants, both integral descriptors of the whole surfactants and fragment-based descriptors computed separately for the polar head and the alkyl chain of the surfactant were investigated.

2 Computational details

2.1 Experimental dataset

The performances of QSPR models are critically dependent on the number and quality of the data employed for its development. For this reason, an important amount of published data on the

properties of sugar-based surfactants was gathered[48] to constitute the largest database on sugar-based surfactants, to the best of our knowledge, with more than 2500 data for more than 600 sugar-based surfactants.

Then, a detailed analysis of the CMC data was performed, to extract the most reliable CMCs in order to constitute the dataset used for the robust development and validation of QSPR models. In particular, all selected CMCs were measured at a temperature as close as possible to room temperature (i.e. between 20°C and 25°C), since temperature can strongly influence CMC values[49]. Moreover, whenever possible, the Krafft temperatures of the surfactants were checked to be lower than 25°C. Indeed, surfactants exhibiting Krafft temperatures higher than 25°C are not expected to form micelles at room temperature due to solubility issues[50]. The aqueous solutions of the highest purity were also targeted, since impurities can affect CMC values[3].

Even after such selection, it has to be kept in mind that some variabilities can still exist for CMCs measured by different experimentalists, as shown in table 1. One example was given by Syper et al.[51] in which, for dodecyl-D-lactobionamide, the measured CMC of 1.3 mM was compared with a previously published one of 3.4 mM, that corresponds to a log difference of 0.4.

Finally, the selected dataset was constituted of 83 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). Our dataset covers a wide range of CMC values, from 0.0033 mM to 180 mM. In this study, CMCs were analyzed in decimal logarithm (in M), as performed by other workers[31, 38, 42-47], since a linear dependency of log CMC with the length of the alkyl chain was evidenced by experimentalists[1]. Finally, the distribution of log CMC data is close to normality, with data ranging between -5.5 and -0.7 and a maximum around -2.5 (see Figure 1).

To allow the evaluation of the predictive power of the models, the dataset was divided into two parts. A training set of 56 surfactants (representing 2/3 of the dataset) was used for the development of the model and a validation set of 27 surfactants (1/3 of the dataset) was used. 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 log CMC and the ones of the validation set were regularly selected (e.g. 2nd, 5th, 8th etc.) to represent 1/3 of the dataset. The similarity of the chemical diversity in both sets was checked based on a Principal Component Analysis computed from the whole set of calculated descriptors (cf. §2.2), as shown in Figure 2, and the surfactants of both the training and the validation sets revealed well-distributed in the global chemical space of the investigated surfactants.

2.2 *Molecular descriptors*

The molecular structures of the 83 studied sugar-based surfactants of the dataset were optimized from Density Functional Theory (DFT) at B3LYP/6-31+G(d,p) level after preliminary conformation analyses to identify the best (most stable) conformation to calculate descriptors. Frequency calculations were performed at the same level of theory to ensure each conformation corresponds to a local minimum in the potential energy surface.

Moreover, the structures of the 34 hydrophilic (polar heads) and 20 hydrophobic (alkyl chains) fragments constituting the 83 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[52] suite of programs was used for all these calculations.

It has to be noted that 26 out of the 83 sugar-based surfactants of our dataset are in the form of enantiomeric[53], diastereomeric[54], or anomeric[55, 56] 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. In all such isomeric mixtures, the isomers were considered as different conformations of the same compound. The geometries of all relevant isomers were optimized and the most stable one was finally retained.

Based on these quantum chemical structures, more than 300 constitutional, topological, geometrical and quantum-chemical descriptors were computed using CODESSA software[57] for each surfactant and each fragment. Additional descriptors were also obtained directly from the quantum-chemical calculations. Descriptors arising from conceptual DFT[58, 59] (electronegativity, hardness, softness and electrophilicity index) were calculated from the energies of the Highest Occupied Molecular Orbital (E_{HOMO}) and the Lowest Unoccupied Molecular Orbital (E_{LUMO}). Moreover, the partial charge of the polar head and of the first hydrocarbon fragment of the alkyl chain (CH_2 or CH here) were also calculated based on Mulliken[60] and Natural Populations Analyses[61] (as implemented into Gaussian09 software), to take into account the possibility of electron withdrawing from polar head to alkyl chain in surfactants as proposed by Huibers[62].

2.3 *Model development and validation*

All QSPR models were built from the training set based on Multi Linear Regressions (MLR) in the general form of eq. 1:

$$\log CMC = a_0 + \sum_i a_i D_i \qquad (1)$$

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

To avoid building overfitted models, which have no predictive power[63], the Best Multi-Linear Regression (BMLR) approach was used as implemented in CODESSA program[57]. This variable selection method, which has been described in detail and successfully used in previous works[64, 65], generates a list of models with an increasing numbers of descriptors, from which the final model is chosen by the user. Here, the final model was chosen as the best compromise between correlation and number of descriptors, again to avoid against any over-parameterization.

To evaluate the performances of each model, robustness and ability of the model to predict properties for molecules that were not used in the parameterisation were tested[66]. The goodness of fit 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 measure the robustness of the model, i.e. the dependence of the fitting of the model to any molecule(s) of the training set 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. To ensure that models did not issue from chance correlations, a Y-scrambling test[67] was realized on the training set. Random permutations of experimental property values 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[67] 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. The difference between the maximal randomized R^2 and the actual R^2 of the new models, $\Delta R^2_{YS,max}$, was also checked as proposed by Nicolotti et al.[68].

Then, the model was applied for 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, series of validation metrics were used: Q^2_{F1}[69], Q^2_{F2}[70], Q^2_{F3}[71], CCC[72, 73], $\overline{r^2_m}$ [74] and Δr^2_m[74] (see Supporting Information, Table S3, for detailed formulas). Based on these validation metrics, the thresholds values proposed by Chirico et al.[75] (presented in Table 3) were used to estimate the reliability of a QSPR model.

It has to be kept in mind that a QSPR model is only expected to provide good predictions for molecules which are similar to those used to develop the model. So, the applicability domain[76,77] (AD) of each model has been defined considering the range of values of the calculated descriptors and the experimental property in the training set. At last, all validation metrics presented above were

calculated again considering only the molecules within the applicability domain. This is accounted for by the IN subscript (R^2_{IN}, MAE_{IN}, $RMSE_{IN}$, $Q^2_{F1,IN,}$ $Q^2_{F2,IN,}$ $Q^2_{F3,IN,}$ CCC_{IN}, $\overline{r^2_m}_{IN}$, $\Delta r^2_{m,IN}$).

3 Results

Six new QSPR models were developed in this study. Three of them include integral descriptors based on all types of descriptors, limited to topological and constitutional descriptors, or only on constitutional descriptors to favor simple models. The three others are based on fragment descriptors on the same scheme, to take into account the specificities of surfactants. All details on the developed models are available in Supporting Information, including equations, performances, ADs, and descriptor definitions.

3.1 Models based on integral descriptors

3.1.1 Model with all descriptors

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

$$\log CMC = -1.8\,{}^1AIC - 3.7\,{}^2ACIC + 4.0\cdot 10^{-2}\eta + 0.21T_e + 1.1 \qquad (2)$$

with ^{1}AIC the Average Information Content of order 1 (t-test = -4.4), ^{2}ACIC the Average Complementary Information Content of order 2 (t-test = -16.6), η the hardness[59], (t-test = 9.6), and T_e the topographic electronic index calculated from all atomic pairs using Zefirov's partial charge model[78] (t-test = 9.0).

The model is characterized by a high goodness of fit (R^2 = 0.93, RMSE = 0.31 (log)) and robustness ($Q^2_{CV} = Q^2_{10CV} = Q^2_{3CV}$ = 0.91, Q^2_{7CV} = 0.90). The criterion of Rücker[67] for Y-scrambling validation is fulfilled: $R^2 - R^2_{YS} = 0.85 > 2.3SD_{YS} = 0.12$. A high $\Delta R^2_{YS,max}$ of 0.69 also confirms that the model is not issued from chance correlation. The model also reveals a very good predictivity, satisfying Chirico's[75] criteria (R^2_{IN} = 0.91, $RMSE_{IN}$ = 0.32 (log), $Q^2_{F1,IN}$ = 0.90, $Q^2_{F2,IN}$ = 0.90, $Q^2_{F3,IN}$ = 0.92,

CCC_{IN} = 0.94, $\overline{r^2_m}_{IN}$ = 0.76, $\Delta r_m^2{}_{IN}$ = 0.10) in its applicability domain (with 26 out of 27 molecules of the validation set in AD), as shown in Figure 4. The only surfactant out of AD is hexyl-D-maltonamide, because its ${}^1AIC$ value is slightly too high. With a large polar head (2 sugar residues) and a small alkyl chain (only 6 C atoms), its log CMC was still well predicted, as illustrated in Figure 4, probably because its ${}^1AIC$ value is not far from the AD range (2.92 vs. AD range of [1.78;2.85]).

3.1.2 Topological/Constitutional descriptors

When limiting the analysis to the 74 constitutional and topological descriptors, a four-parameter model (eq. 3) was also found as the best compromise between correlation and number of descriptors.

$$\log CMC = -3.7\,{}^0AIC - 3.5\,{}^2ACIC + 1.6 n_O - 0.7 n_S + 8.5 \qquad (3)$$

with ^{0}AIC the Average Information Content of order 0 (t-test = -6.4), ^{2}ACIC the Average Complementary Information Content of order 2 (t-test = -16.7), n_O the number of O atoms (t-test = 6.8), and n_S the number of S atoms (t-test = -4.1).

A good correlation was found (R^2 = 0.87, RMSE = 0.41 (log)), and the model appears as robust (Q^2_{CV} = 0.84, $Q^2_{10CV} = Q^2_{7CV}$ = 0.83, Q^2_{3CV} = 0.84) as the quantum-chemical based model (eq. 2). The Y-scrambling demonstrates that it does not originate from chance correlation according to Rücker's criteria: $R^2 - R^2_{YS} = 0.79 > 2.3SD_{YS} = 0.11$.

Although slightly lower than for eq. 2, the predictivity of this model revealed nevertheless also good (R^2_{IN} = 0.89, $RMSE_{IN}$ = 0.37 (log), $Q^2_{F1,IN} = Q^2_{F2,IN}$ = 0.87, $Q^2_{F3,IN}$ = 0.89, CCC_{IN} = 0.93, $\overline{r^2_m}_{IN}$ = 0.73, $\Delta r_m^2{}_{IN}$ = 0.12), on 24 out of 27 molecules of the validation set that belong to AD. All molecules out of AD (S-Hexyl-5-Thio-D-Arabinonolactone, S-Hexyl-5-Thio-D-Xylonolactone and Hexyl-D-Maltonamide) have slightly too high ${}^0AIC$ values (1.71, 1.71 and 1.68, respectively, with respect to an AD range of [1.17;1.67]). Nevertheless, good predictions were obtained for all three out-of-AD surfactants, with errors between 0.2 and 0.5 (log) since they keep close from the AD. Again, all Chirico criteria were fulfilled.

3.1.3 Model with constitutional descriptors

The four-parameter model in eq. 4 was chosen among the six equations sorted out by the BMLR method when focusing on the only 36 constitutional descriptors.

$$\log CMC = -62 n_{rel,C} - 7.1 \cdot 10^{-2} n_H - 0.8 n_S + 0.6 n_{rings} + 17.4 \quad (4)$$

with $n_{rel,C}$ the relative number of C atoms (t-test = -8.9), n_H the number of H atoms (t-test = -8.3), n_S the number of S atoms (t-test = -4.2), and n_{rings} the number of rings (t-test = 5.2).

R^2 = 0.82 and RMSE = 0.47 (log) were obtained for this very simple model. It also proved robust (Q^2_{CV} = Q^2_{10CV} = Q^2_{7CV} = 0.79, Q^2_{3CV} = 0.78) and the Y-scrambling ensured that the model was not issued from chance correlation with low values of R^2 for the models obtained after randomization according to the criteria of Rücker with $R^2 - R^2_{YS} = 0.75 > 2.3\ SD_{YS} = 0.12$.

The predictive power of the model has proved to be good in terms of most of the validation metrics ($R^2_{EXT} = R^2_{IN} = 0.78$, $Q^2_{F1} = 0.76$, $Q^2_{F2} = 0.76$, $Q^2_{F3} = 0.78$, and CCC = 0.85), and all molecules of the validation set are in its applicability domain. However, Roy's metrics stand below Chirico's criteria ($\overline{r^2_m} = 0.55 < 0.65$, $\Delta r_m^2 = 0.23 > 0.20$).

Although the predictive performances of this model revealed better than Mattei's model[46] ($RMSE_{EXT}$ of 0.92 (log)) , eq. 4 keeps higher than more complex models (eqs. 2 and 3) with $RMSE_{EXT} = RMSE_{IN} = 0.52$ (log).

3.2 *Fragment descriptors*

3.2.1 Models with all types of descriptors

A total of 627 descriptors were obtained when combining the descriptors of the polar and hydrophobic fragments. Based on this large set of descriptors, a first three-parameter model (eq. 5) was obtained, selected as the best compromise between correlation and number of descriptors among the 10 equations sorted out by the BMLR method.

$$\log CMC = 5.7 \cdot 10^{-2} \eta_h - 2.10 \cdot 10^{-2} TMSA_c + 7.4 \cdot 10^{-3\ 2} IC_h - 8.2 \quad (5)$$

with η_h the hardness of the polar head (t-test = 9.3), $TMSA_c$ the total molecular surface area of the alkyl chain (t-test = -23.7) and ^2IC_h is the Information content (order 2) of the polar head (t-test = 8.6).

An excellent fitting of the training set data was obtained, (R^2 = 0.93, RMSE = 0.30 (log)) and the model appears to be robust (Q^2_{CV} = 0.92, Q^2_{10CV} = Q^2_{7CV} = 0.91, Q^2_{3CV} = 0.90) and not issued from chance correlation as evidenced by Y-scrambling (R^2 - R^2_{YS} = 0.85 > 2.3SD_{YS} = 0.08).

Although being slightly lower than the QSPR model for integral-based descriptors of all types ($RMSE_{IN}$ = 0.32 (log) for this model, eq. 2), eq. 5 also has a high predictive ability (R^2_{IN} = 0.88, $RMSE_{IN}$ = 0.36 (log), $Q^2_{F1,IN}$ = 0.88, $Q^2_{F2,IN}$ = 0.88, $Q^2_{F3,IN}$ = 0.90, CCC_{IN} = 0.93, $\overline{r^2_m}_{IN}$ = 0.79, $\Delta r_m^2{}_{IN}$ = 0.11), in agreement with Chirico's[75] criteria. Only one surfactant, Hexyl-D-Maltonamide, was found to be out of AD as defined by the experimental log CMC (calculated log CMC of -0.66 vs. AD range of [-5.48;-0.74]).

3.2.2 Topological/Constitutional descriptors

A simpler model was again reached using only the 150 topological and constitutional fragment descriptors. A two-parameter model (eq. 6) was derived from the BMLR method.

$$\log CMC = -19 n_{rel,S,h} - 2.6 \cdot 10^{-2\ 2} CIC_c - 0.2 \quad (6)$$

with $n_{rel,S,h}$ the relative number of S atoms in the polar head (t-test = -6.2) and $^2CIC_c$ the Complementary Information Content of order 2 of the alkyl chain (t-test = -18.4).

Good correlation (R^2 = 0.87, RMSE = 0.41 (log)) and robustness (Q^2_{LOO} = 0.84, Q^2_{10CV} = Q^2_{7CV} = 0.85, Q^2_{3CV} = 0.86) characterize this model. The Y-scrambling test ensures that the model was not obtained by chance correlation: R^2 - R^2_{YS} = 0.83 > 2.3SD_{YS} = 0.08.

This model exhibits a satisfying predictivity (R^2_{IN} = 0.89, $RMSE_{IN}$ = 0.36 (log), Q^2_{F1} = Q^2_{F2} = 0.88, Q^2_{F3} = 0.90, CCC = 0.94, $\langle r_m^2 \rangle$ = 0.81, Δr_m^2 = 0.10) in its applicability domain, with all molecules of the validation set included. Once again, it fulfils Chirico's criteria. The predictivity of the

fragment-based model including constitutional and topological descriptors is identical to the corresponding integral-based model ($RMSE_{IN}$ = 0.36 (log)).

3.2.3 Constitutional descriptors

At last, focusing on the 72 fragment constitutional descriptors, a three-parameter model (eq. 7) was found.

$$\log CMC = -20 n_{rel,S,h} - 2.7 \cdot 10^{-2} M_{w,c} - 6 \cdot 10^{1} n_{rel,single,c} + 64.8 \quad (7)$$

with $n_{rel,S,h}$ the relative number of S atoms in the polar head (t-test = -6.5), $M_{w,c}$ the molecular weight of the alkyl chain (t-test = -16.1) and $n_{rel,single,c}$ the relative number of single bonds in the alkyl chain (t-test = -4.1).

This model is well-fitted with the training set surfactants (R^2 = 0.86, RMSE = 0.41 (log)), and presents a good robustness ($Q^2_{CV} = Q^2_{10CV} = Q^2_{3CV}$ = 0.84, Q^2_{7CV} = 0.85). The Y-scrambling confirms that it is not issued from chance correlation: $R^2 - R^2_{YS} = 0.81 > 2.3SD_{YS} = 0.10$.

Eq. 7 demonstrates similar predictive performances than other fragment-based models, as shown in Figure 5 (R^2_{EXT} = 0.88, $RMSE_{EXT}$ = 0.36 (log), $Q^2_{F1} = Q^2_{F2}$ = 0.88, Q^2_{F3} = 0.90, CCC = 0.94, $\langle r_m^2 \rangle$ = 0.79, Δr_m^2 = 0.11) and its applicability domain includes all the molecules of the validation set.

Therefore, this model is particularly appealing since it is based on very simple descriptors and remains reliable enough for good-quality and fast estimation of CMC of sugar-based surfactants in the perspective of molecular screening or discovery, notably for formulation specialists.

4 Discussion

4.1 Molecular descriptors

Different categories of descriptors are identified among those included into the new QSPR models of this study that relate to observed experimental trends of CMC. Some descriptors are related to the size

of the alkyl chain, to the size of the polar head, to the presence of a sulfur linkage and to the saturation of the alkyl chain.

The size of the alkyl chain is encoded by three descriptors ($M_{w,c}$, $TMSA_c$, $^2CIC_c$). The molecular weight of the alkyl chain $M_{w,c}$ and the total molecular surface area of the alkyl chain ($TMSA_c$) are proportional to alkyl chain size. At last, the topological descriptor Complementary Information Content of order 2 in the alkyl chain, $^2CIC_c$, is also mainly influenced by the alkyl chain length since it increases with the number of atoms. These descriptors are the most significant ones in all the models regarding the t-test. The size of the alkyl chain relates to the hydrophobic effect[79], which drives the micellisation/adsorption behavior of conventional surfactants. Hydrophobic effect is defined as the entropy loss induced by the water structuration around alkyl chains. When increasing surfactant concentration, the alkyl chains tend to avoid contact with water either by adsorbing at water interfaces or by forming micelles. The micelle formation onset, represented by the CMC, is likely to happen at lower concentrations for larger alkyl chains because of a higher hydrophobic effect. For their relation to alkyl chain size, and their high statistical significance, these four descriptors are meaningful regarding CMC property.

Three of the new models also contain descriptors related to the size of the polar head (n_O, n_{rings} and ^2IC_h). Eq. 3 contains the number of O atoms, n_O, and eq. 4 contains the number of rings, n_{rings}. As the sugar-based surfactants considered in this study are constituted by multiple alcohol functional groups, and rings are only on their polar heads, these descriptors relate to the size of the polar head. Eq. 6 contains the second-order information content for the polar head, ^2IC_h, which increases with the total number of atoms in the polar head. So, ^2IC_h also increases with the size of the polar head. Polar head size was experimentally identified as impacting CMC[1, 3]. Thus, presence of such descriptors in QSPR equations for CMC is not surprising.

Four out of the six developed QSPR models include the number or relative number of S atoms. In all these models, the associated regression coefficient is negative, which means that the presence of S may decrease the CMC. Among the surfactants of the dataset, the S atom is always a linking atom

between the polar head and the alkyl chain and appears in 17 surfactants out of 83. So, it can be thought that such a S linkage represents a particular feature, in agreement with experimental literature suggesting that S linkage increases the hydrophobic character of sugar-based surfactants[80].

To further study the specific character of S-linked surfactants, we analyzed partial charges on the linker and along the chain, as in a previous study[81], for a series of surfactants by only changing their linkages (-O-, -(C=O)-O-, -(C=O)-NH-, -O-(C=O)-, -NH-(C=O)-, -(C=O)-N(Me)-, -N(Me)-(C=O)-) on octyl-β-D-thioglucoside, a S-linked surfactant (structures of the modeled surfactants are provided in Supporting Information, Figure S1). All linkages in the dataset were represented in the 8 modeled structures. As shown in Figure 6, octyl-β-D-thioglucoside exhibits a particular feature. Indeed, its -S- linker presents a positive partial charge of +0.2 whereas all other linkers present a negative one (between -0.2 to -0.6). Some of the tested surfactants (with –O-, -(C=O)-O-, -(C=O)-NH-, -(C=O)-NMe- linkages) have +0.2 to +0.3 in partial charge for the first CH_2 and near 0 for the other CH_2 of the alkyl chain, whereas for -O-(C=O)-, -NH-(C=O)-, -N(Me)-(C=O)- and -S- linkages, very low partial charge is already observed on the first CH_2 (-0.1 to 0.0). This may make the first CH_2 more hydrophobic and therefore lower CMCs might be anticipated for such compounds, which is confirmed at least for -S- linked surfactant octyl-β-D-thioglucoside (CMC = 8,5 mM vs. 20 mM for octyl-β-D-glucoside, see Table 2).

Moreover, the plot of experimental log CMC versus $^2CIC_c$ (the main descriptor from the best QSPR model containing S related descriptors according to t-test, eq. 6) confirms that S-linked surfactants constitute a particular subfamily of sugar-based surfactants (see Figure 7). This pushed us to consider developing a new local model for the 66 non-S linked surfactants only based on $^2CIC_c$, but no improvement was obtained from it (R^2_{EXT} = 0.92), since R^2_{EXT} = 0.92 was also calculated for non-S linked surfactants of the global model (eq. 6).

At last, it is also worth mentioning that the relative number of single bonds in the alkyl chain, $n_{rel,single,c}$, appears in one of our models (eq. 7) with a negative regression coefficient. This implies that unsaturated alkyl chains should have higher CMCs. This is supported by general considerations about

surfactants and originates from a less hydrophobic character of unsaturated alkyl chains[1, 3]. Therefore, this descriptor is also physically meaningful.

4.2 Performance of the developed QSPR models

From a general point of view, the new QSPR models developed for the CMC of sugar-based surfactants revealed good performances with both the descriptors of the whole molecule and the ones using fragment-based approach, as shown in Table 4. In particular, they provide an improvement with respect to the best developed model identified in literature, i. e. the model of Mattei[46], with $RMSE_{IN}$ of 0.32-0.52 (log) for our models compared to 1.11 (log) for the model of Mattei on the same validation set (for consistent comparison), see Supporting Information. For most of the models (eqs. 2,3 and 5-7), $RMSE_{IN}$ is between 0.32 and 0.37 (log), which is relevant regarding experimental variance found in literature as shown in Experimental dataset section.

The quality of the developed QSPR models appears to improve when topological and quantum-chemical descriptors are included in the case of integral descriptors. Indeed, the $RMSE_{IN}$ lowers from 0.52 to 0.37, from the model containing only constitutional descriptors to the model including topological descriptors, and to 0.32 for the model including quantum-chemical descriptors. In the case of fragment-based descriptors, same good results were yielded for all models, with a $RMSE_{IN}$ of 0.36 regardless of the level of complexity of the model even for the one based on constitutional descriptors only. It seems that the integral all types descriptors can well describe polar head and alkyl chain fragments specificities, as suggested by similar predictive powers obtained when considering topological and/or quantum-chemical descriptors in integral-based (eqs. 2, 3) and fragment-based (eqs. 5, 6) models. Besides, same good predictivity has been obtained with the three fragment-based models, indicating that specificities of such surfactants may be also quite well-encoded by simple constitutional descriptors through the fragment-based approach, which gives them a special interest.

Overall, the best model is the model based on integral descriptors including quantum-chemical ones, with $RMSE_{IN}$ of only 0.32 (log). This model (eq. 2) is recommended to access the best CMC

prediction of sugar-based surfactants but it requires preliminary quantum chemical calculations. To access faster and easier predictions, the fragment-based QSPR model including only constitutional descriptors (eq. 7) also reached good performances, with $RMSE_{IN}$ = 0.36 (log). This model may turn out to be a particularly powerful tool for high-throughput molecular screening and design of new surfactants based on CMC performances.

5 Conclusion

A series of new QSPR models were developed and validated to predict the CMC of sugar-based surfactants in order to facilitate the screening and design of such surfactants and formulations that include them, notably in view of replacing petroleum-based surfactants. Different models were developed upon the kinds of descriptors used: calculated for the whole molecule (as classically done) or for hydrophilic/hydrophobic fragments to better account the specificity of surfactants, including quantum-chemical and topological descriptors or only constitutional descriptors to reach easy-to-use models. The new developed models encode simple experimentally observed trends like the decrease of CMC with the increase of alkyl chain length. In particular, a CMC-decreasing role of the sulfur linkage was revealed by several models. Whatever the model, good performances were reached, with most $RMSE_{IN}$ between 0.32 and 0.37 (log). These performances compared favorably with RMSE = 1.11 (log) obtained for the best identified model applicable to sugar-based surfactants, Mattei's model, when tested on the same validation set. The most reliable developed model includes quantum-chemical and topological descriptors computed for the whole molecule with a $RMSE_{IN}$ of 0.32 (log). A simpler fragment-based model was also obtained using only constitutional descriptors with a $RMSE_{IN}$ of 0.36 (log). This simple fragment-based approach is very easy to use based on the only knowledge of the 2D structure of the surfactants. Both models offer good alternative tools to the systematic experimental characterizations in particular at R&D level as they allow accessing to CMC estimates of new surfactants even before synthesis, or to enforce the evaluation of partition coefficients of surfactants in the context of REACH regulation. To the end, the structure-CMC trends evidenced in these QSPR models can even help to identify new high potential surfactant structures.

Specifically, our analyses suggest that, beyond bearing long alkyl chains, sugar-based surfactants with small polar heads, saturated alkyl chains and sulfur linkages are expected to have lower CMCs.

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 CMC values from the model of Mattei and the new QSPR models (Table S1), descriptor values for all surfactants of the dataset (Table S2), definitions of validation metrics (Table S3), details about the new QSPR models (Tables S4-S9), definitions of the descriptors involved in the models, Surfactants modeled in Figure 6 (Figure S1).

References

(1) Rosen, M.J.; Kunjappu, J.T., *Surfactants and Interfacial Phenomena*. 4th ed.: John Wiley & Sons, Inc., 2012.
(2) Dembitsky, V., Astonishing diversity of natural surfactants: 1. Glycosides of fatty acids and alcohols. *Lipids*, **2004**, *39*, 933.
(3) Myers, D., *Surfactant Science and Technology*. 3rd ed.: Wiley-Interscience, 2006.
(4) Eby, C.; Tatum, *Chemistry of Soap, Detergents, & Cosmetics*. United States: Flinn Scientific, Incorporated, 1989.
(5) Fuerstenau, M.C.; Miller, J.D.; Kuhn, M.C., *Chemistry of Flotation*. New York: Society of Mining Engineers, AIME, 1985.
(6) Muller, C.; Maldonado, A.G.; Varnek, A.; Creton, B., Prediction of Optimal Salinities for Surfactant Formulations Using a Quantitative Structure–Property Relationships Approach. *Energy Fuels*, **2015**, *29*, 4281.
(7) Jones, M.N., Surfactants in membrane solubilisation. *Int. J. Pharm.*, **1999**, *177*, 137.
(8) Carpenter, E.P.; Beis, K.; Cameron, A.D.; Iwata, S., Overcoming the challenges of membrane protein crystallography. *Curr. Opin. Struct. Biol.*, **2008**, *18*, 581.
(9) Ruiz, C.C., *Sugar-Based Surfactants: Fundamentals and Applications*. CRC Press, Taylor & Francis Group, 2009.
(10) Shinoda, K.; Yamanaka, T.; Kinoshita, K., Surface Chemical Properties in Aqueous Solutions of Non-ionic Surfactants Octyl Glycol Ether, α-Octyl Glyceryl Ether and Octyl Glucoside. *J. Phys. Chem.*, **1959**, *63*, 648.
(11) Liljekvist, P.; Kjellin, M.; Christer Eriksson, J., The surface pressure effect of pentaoxyethylene and maltoside surfactant head groups. *Adv. Colloid Interface Sci.*, **2001**, *89–90*, 293.
(12) Bazin, H.G.; Polat, T.; Linhardt, R.J., Synthesis of sucrose-based surfactants through regioselective sulfonation of acylsucrose and the nucleophilic opening of a sucrose cyclic sulfate. *Carbohydr. Res.*, **1998**, *309*, 189.
(13) Kjellin, M.; Johansson, I., *Surfactants from Renewable Resources*. 1 ed.: John Wiley & Sons, Ltd, 2010.
(14) Matsumura, S.; Imai, K.; Yoshikawa, S.; Kawada, K.; Uchibor, T., Surface activities, biodegradability and antimicrobial properties of n-alkyl glucosides, mannosides and galactosides. *J. Am. Oil Chem. Soc.*, **1990**, *67*, 996.
(15) Hill, K.; LeHen-Ferrenbach, C., *1. Sugar-Based Surfactants for Consumer Products and Technical Applications*, in *Sugar-Based Surfactants: Fundamentals and Applications*, C.C. Ruiz, Editor. 2009, CRC Press, Taylor & Francis Group.
(16) Rojas, O.J.; Stubenrauch, C.; Lucia, L.A.; Habibi, Y., *Interfacial Properties of Sugar-Based Surfactants*, in *Biobased Surfactants and Detergents: Synthesis, Properties, and Applications*, D. Hayes, et al., Editors. 2009, AOCS Press: Urbana.
(17) Israelachvili, J.N.; Mitchell, D.J.; Ninham, B.W., Theory of self-assembly of hydrocarbon amphiphiles into micelles and bilayers. *J. Chem. Soc., Faraday Trans*, **1976**, *72*, 1525.
(18) Rangel-Yagui, C.O.; Pessoa-Jr, A.; Costa Tavares, L., Micellar solubilization of drugs. *J. Pharm. Pharm. Sci.*, **2005**, *8*, 147.
(19) *Regulation (EC) No 1907/2006 of the European Parliament and of the Council*. European Parliament: Brussels, December 2006.
(20) *Guidance on Information Requirements and Chemical Safety Assessment, − Chapter R.7a:Endpoint Specific Guidance, Version 3*. 2014, European Chemicals Agency (ECHA): Brussels.
(21) Lu, B.; Vayssade, M.; Miao, Y.; Chagnault, V.; Grand, E.; Wadouachi, A.; Postel, D.; Drelich, A.; Egles, C.; Pezron, I., Physico-chemical properties and cytotoxic effects of sugar-based surfactants: Impact of structural variations. *Colloids and Surfaces B: Biointerfaces*, **2016**, *145*, 79.
(22) Esumi, K.; Ueno, M., *Structure-Performance Relationships in Surfactants*. 2nd ed. Surfactant Science Series. 2003.
(23) Mukerjee, P.; Mysels, K.J., *Critical Micelle Concentrations of Aqueous Surfactant Systems*. 1971, National Bureau of Standards: United States.
(24) du Noüy, P.L., An interfacial tensiometer for universal use. *J. Gen. Physiol.*, **1925**, *7*, 625.

(25) Fayet, G.; Rotureau, P., How to use QSPR-type approaches to predict properties in the context of Green Chemistry. *Biofuels, Bioprod. Biorefin.*, **2016**, *in press, DOI: 10.1002/bbb.1723*.
(26) Creton, B., Prediction of Surfactants' Properties using Multiscale Molecular Modeling Tools: A Review. *Oil Gas Sci. Technol. Rev. IFPEN*, **2013**, 1.
(27) Nieto-Draghi, C.; Fayet, G.; Creton, B.; Rozanska, X.; Rotureau, P.; de Hemptinne, J.-C.; Ungerer, P.; Rousseau, B.; Adamo, C., A General Guidebook for the Theoretical Prediction of Physicochemical Properties of Chemicals for Regulatory Purposes. *Chem. Rev.*, **2015**, *115*, 13093.
(28) Stephenson, B.C., *Complementary Use of Computer Simulations and Molecular-Thermodynamic Theory to Model Surfactant and Solubilize Self-Assembly*. 2006, Massachusetts Institute of Technology.
(29) Vishnyakov, A.; Lee, M.-T.; Neimark, A.V., Prediction of the Critical Micelle Concentration of Nonionic Surfactants by Dissipative Particle Dynamics Simulations. *J. Phys. Chem. Lett.*, **2013**, *4*, 797.
(30) Klevens, H.B., Structure and aggregation in dilate solution of surface active agents. *J. Am. Oil Chem. Soc.*, **1953**, *30*, 74.
(31) Huibers, P.D.T.; Lobanov, V.S.; Katritzky, A.R.; Shah, D.O.; Karelson, M., Prediction of Critical Micelle Concentration Using a Quantitative Structure−Property Relationship Approach. 1. Nonionic Surfactants. *Langmuir*, **1996**, *12*, 1462.
(32) Hu, J., A Review on Progress in QSPR Studies for Surfactants. *Int. J. Mol. Sci.*, **2010**, *11*, 1020.
(33) Katritzky, A.R.; Kuanar, M.; Slavov, S.; Hall, C.D.; Karelson, M.; Kahn, I.; Dobchev, D.A., Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chem. Rev.*, **2010**, *110*, 5714.
(34) Nicolotti, O.; Benfenati, E.; Carotti, A.; Gadaleta, D.; Gissi, A.; Mangiatordi, G.F.; Novellino, E., REACH and in silico methods: an attractive opportunity for medicinal chemists. *Drug Discov. Today*, **2014**, *19*, 1757.
(35) Gasteiger, J.; Engel, T., *Chemoinformatics: A Textbook*. Weinheim: Wiley GmbH, 2003.
(36) Dagnélie, P., *Statistique théorique et appliquée,Tomes 1 & 2*. Bruxelles: De Boeck, 2011.
(37) Gasteiger, J.; Zupan, J., Neural Networks in Chemistry. *Angew. Chem. Int. Ed. Engl.*, **1993**, *32*, 503.
(38) Katritzky, A.R.; Pacureanu, L.M.; Slavov, S.H.; Dobchev, D.A.; Karelson, M., QSPR Study of Critical Micelle Concentrations of Nonionic Surfactants. *Ind. Eng. Chem. Res.*, **2008**, *47*, 9687.
(39) Estrada, E., Spectral Moments of the Edge Adjacency Matrix in Molecular Graphs. 1. Definition and Applications to the Prediction of Physical Properties of Alkanes. *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 844.
(40) Varnek, A.; Kireeva, N.; Tetko, I.V.; Baskin, I.I.; Solov'ev, V.P., Exhaustive QSPR Studies of a Large Diverse Set of Ionic Liquids: How Accurately Can We Predict Melting Points? *J. Chem. Inf. Model.*, **2007**, *47*, 1111.
(41) Varnek, A., *Fragment Descriptors in Structure–Property Modeling and Virtual Screening*, in *Chemoinformatics and Computational Chemical Biology*, J. Bajorath, Editor. 2011, Humana Press: Totowa, NJ.
(42) Saunders, R.A.; Platts, J.A., Correlation and prediction of critical micelle concentration using polar surface area and LFER methods. *J. Phys. Org. Chem.*, **2004**, *17*, 431.
(43) Khayamian, T.; Esteki, M.; Abbasi, A., *Application of Wavelet Neural Networks in Multivariate Data Analysis*, in *Progress in Chemometrics Research*, A.L. Pomerantsev, Editor. 2005.
(44) Mozrzymas, A.; Rózycka-Roszak, B., Prediction of critical micelle concentration of nonionic surfactants by a quantitative structure - property relationship. *Comb. Chem. High Throughput Screen.*, **2010**, *13*, 39.
(45) Roy, K.; Kabir, H., QSPR with extended topochemical atom (ETA) indices: Modeling of critical micelle concentration of non-ionic surfactants. *Chem. Eng. Sci.*, **2012**, *73*, 86.
(46) Mattei, M.; Kontogeorgis, G.M.; Gani, R., Modeling of the Critical Micelle Concentration (CMC) of Nonionic Surfactants with an Extended Group-Contribution Method. *Ind. Eng. Chem. Res.*, **2013**, *52*, 12236.
(47) Anoune, N.; Nouiri, M.; Berrah, Y.; Gauvrit, J.-Y.; Lanteri, P., Critical micelle concentrations of different classes of surfactants: A quantitative structure property relationship study. *J. Surfact. Deterg.*, **2002**, *5*, 45.

(48) Gaudin, T.; Lu, H.; Van Hecke, E.; Drelich, A.; Dao, T.T.; Rotureau, P.; Benali, M.; Bonnet, V.; Wadouachi, A.; Pourceau, G.; Fayet, G.; Pezron, I. *Data analysis on sugar-based surfactants: Towards structure-property relationships.* in *10th World Surfactant Congress and Business Convention (CESIO)*. 2015. Haliç Congress Center, Istanbul.
(49) Mahmood, M.E.; Al-Koofee, D.A.F., Effect of Temperature Changes on Critical Micelle Concentration for Tween Series Surfactant. *GJSFR: Chem.*, **2013**, *13*, 1.
(50) Moroi, Y., *Relationship between solubility and micellization of surfactants: The temperature range of micellization*, in *Dispersed Systems*, K. Hummel and J. Schurz, Editors. 1988, Steinkopff. p. 55.
(51) Syper, L.; Wilk, K.A.; Sokołowski, A.; Burczyk, B., *Synthesis and surface properties of N-alkylaldonamides*, in *Trends in Colloid and Interface Science XII*, G.J.M. Koper, et al., Editors. 1998, Steinkopff. p. 199.
(52) Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H.P.; Izmaylov, A.F.; Bloino, J.; Zheng, G.; Sonnenberg, J.L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J.A.; Peralta, J.E.; Ogliaro, F.; Bearpark, M.; Heyd, J.J.; Brothers, E.; Kudin, K.N.; Staroverov, V.N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J.C.; Iyengar, S.S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J.M.; Klene, M.; Knox, J.E.; Cross, J.B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R.E.; Yazyev, O.; Austin, A.J.; Cammi, R.; Pomelli, C.; Ochterski, J.W.; Martin, R.L.; Morokuma, K.; Zakrzewski, V.G.; Voth, G.A.; Salvador, P.; Dannenberg, J.J.; Dapprich, S.; Daniels, A.D.; Farkas; Foresman, J.B.; Ortiz, J.V.; Cioslowski, J.; Fox, D.J., *Gaussian 09, Revision B.01*. 2009: Wallingford CT.
(53) Savelli, M.P.; Van Roekeghem, P.; Douillet, O.; Cavé, G.; Godé, P.; Ronco, G.; Villa, P., 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.*, **1999**, *182*, 221.
(54) Matsumura, S.; Imai, K.; Yoshikawa, S.; Kawada, K.; Uchibori, T., Surface Activities, Foam Suppression, Biodegradability and Antimicrobial Properties of s-Alkyl Glucopyranosides. *J. Jpn. Oil. Chem. Soc.*, **1991**, *40*, 709.
(55) Kjellin, U.R.M.; Claesson, P.M.; Vulfson, E.N., Studies of N-Dodecyllactobionamide, Maltose 6'-O-Dodecanoate, and Octyl-β-glucoside with Surface Tension, Surface Force, and Wetting Techniques. *Langmuir*, **2001**, *17*, 1941.
(56) Boyère, C.; Broze, G.; Blecker, C.; Jérôme, C.; Debuigne, A., Monocatenary, branched, double-headed, and bolaform surface active carbohydrate esters via photochemical thiol-ene/-yne reactions. *Carbohydr. Res.*, **2013**, *380*, 29.
(57) *Codessa*, *www.semichem.com/codessa/*.
(58) Chermette, H., Chemical reactivity indexes in density functional theory. *J. Comput. Chem.*, **1999**, *20*, 129.
(59) Geerlings, P.; De Proft, F.; Langenaeker, W., Conceptual density functional theory. *Chem. Rev.*, **2003**, *103*, 1793.
(60) Mulliken, R.S., Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I. *J. Chem. Phys.*, **1955**, *23*, 1833.
(61) Reed, A.E.; Weinstock, R.B.; Weinhold, F., Natural population analysis. *J. Chem. Phys.*, **1985**, *83*, 735.
(62) Huibers, P.D.T., Quantum-Chemical Calculations of the Charge Distribution in Ionic Surfactants. *Langmuir*, **1999**, *15*, 7546.
(63) Yee, L.C.; Wei, Y.C., *Current Modeling Methods Used in QSAR/QSPR*, in *Statistical Modelling of Molecular Descriptors in QSAR/QSPR.*, M. Dehmer, K. Varmuza, and D. Bonchev, Editors. 2012. p. 1.
(64) Fayet, G.; Rotureau, P.; Joubert, L.; Adamo, C., QSPR modeling of thermal stability of nitroaromatic compounds: DFT vs. AM1 calculated descriptors. *J. Mol. Model.*, **2010**, *16*, 805.
(65) Fayet, G.; Rotureau, P.; Joubert, L.; Adamo, C., Development of a QSPR model for predicting thermal stabilities of nitroaromatic compounds taking into account their decomposition mechanisms. *J. Mol. Model.*, **2010**, *17*, 2443.

(66) OECD, *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] models*. 2007: Paris.
(67) Rücker, C.; Rücker, G.; Meringer, M., y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.*, **2007**, *47*, 2345.
(68) Nicolotti, O.; Carotti, A., QSAR and QSPR Studies of a Highly Structured Physicochemical Domain. *J. Chem. Inf. Model.*, **2006**, *46*, 264.
(69) Tropsha, A.; Gramatica, P.; Gombar, V.K., The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *J. Comb. Sci.*, **2003**, *22*, 69.
(70) Schüürman, G.; Ebert, R.-U.; Chen, J.; Wang, B.; Kühne, R., External Validation, Prediction employing the predictive squared correlation coefficient - test set activity mean vs. training set activity mean. *J. Chem. Inf. Model.*, **2008**, *48*, 2140.
(71) Consonni, V.; Ballabio, D.; Todeschini, R., Comments on the Definition of the Q2 Parameter for QSAR Validation. *J. Chem. Inf. Model.*, **2009**, *49*, 1669.
(72) Lawrence, I.K.L., A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, **1989**, *45*, 255.
(73) Lawrence, I.K.L., Assay Validation Using the Concordance Correlation Coefficient. *Biometrics*, **1992**, *48*, 599.
(74) Roy, K.; Mitra, I.; Kar, S.; Ojha, P.K.; Das, R.N.; Kabir, H., Comparative Studies on Some Metrics for External Validation of QSPR Models. *J. Chem. Inf. Model.*, **2012**, *52*, 396.
(75) Chirico, N.; Gramatica, P., 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.*, **2012**, *52*, 2044.
(76) Jaworska, J.; Nina, N.-J.; Aldenberg, T., QSAR applicability domain estimation by projection of the training set descriptor space: a review. *Altern. Lab. Anim.*, **2005**, *33*, 445.
(77) Gissi, A.; Galadeta, D.; Floris, M.; Olla, S.; Carotti, A.; Novellino, E.; Benfenati, E.; Nicolotti, O., An Alternative QSAR-Based Approach for Predicting the Bioconcentration Factor for Regulatory Purposes. *Altex*, **2014**, *31*, 23.
(78) Zefirov, N.S.; Kirpichenok, M.A.; Izmailov, F.F.; Trofimov, M.I., Scheme for the Calculation of the Electronegativities of Atoms in a Molecule in the Framework of Sanderson's Principle. *Dokl. Akad. Nauk. SSSR*, **1987**, *296*, 883.
(79) Tanford, C., Interfacial free energy and the hydrophobic effect. *Proc. Natl. Acad. Sci.*, **1979**, *76*, 4175.
(80) Marchant, R.E.; Anderson, E.H.; Zhu, J., *Polysaccharide Surfactants: Structure, Synthesis, and Surface-Active Properties*, in *Polysaccharides, Structural Diversity and Functional Versatility*, S. Dimitriu, Editor. 2005, Marcel Dekker. p. 1055.
(81) Gaudin, T.; Rotureau, P.; Pezron, I.; Fayet, G., Conformations of n-alkyl-α/β-D-glucopyranoside surfactants: impact on molecular properties *Submitted to Comp. Theor. Chem.*, **2015**.
(82) Nilsson, F.; Söderman, O.; Johansson, I., Four Different C8G1Alkylglucosides. Anomeric Effects and the Influence of Straight vs Branched Hydrocarbon Chains. *J. Colloid Interface Sci.*, **1998**, *203*, 131.
(83) Antonelli, M.L.; Bonicelli, M.G.; Ceccaroni, G.; Mesa, C.; Sesta, B., Solution properties of octyl-β-D-glucoside. Part 2: Thermodynamics of micelle formation. *Colloid Polym Sci*, **1994**, *272*, 704.
(84) Mesa, C.; Bonincontro, A.; Sesta, B., Solution properties of octyl β-D glucoside. Part 1: Aggregate size. shape and hydration. *Colloid Polym Sci*, **1993**, *271*, 1165.
(85) Walter, A.; Suchy, S.E.; Vinson, P.K., Solubility properties of the alkylmethylglucamide surfactants. *BBA-Biomembranes*, **1990**, *1029*, 67.
(86) Okawauchi, M.; Hagio, M.; Ikawa, Y.; Sugihara, G.; Murata, Y.; Tanaka, M., A Light-Scattering Study of Temperature Effect on Micelle Formation of N-Alkanoyl-N-methylglucamines in Aqueous Solution. *Bull. Chem. Soc. Jpn.*, **1987**, *60*, 2719.
(87) Harada, S.; Sahara, H., Volumetric Behavior of Micellization of Acyl-N-methylglucamide Surfactants in Water. *Langmuir*, **1994**, *10*, 4073.
(88) Ferrer, M.; Comelles, F.; Plou, F.J.; Cruces, M.A.; Fuentes, G.; Parra, J.L.; Ballesteros, A., Comparative Surface Activities of Di- and Trisaccharide Fatty Acid Esters. *Langmuir*, **2002**, *18*, 667.

(89) Becerra, N.; Toro, C.; Zanocco, A.L.; Lemp, E.; Günther, G., Characterization of micelles formed by sucrose 6-O-monoesters. *Colloids Surf., A*, **2008**, *327*, 134.
(90) Söderberg, I.; Drummond, C.J.; Neil Furlong, D.; Godkin, S.; Matthews, B., Non-ionic sugar-based surfactants: Self assembly and air/water interfacial activity. *Colloids Surf., A*, **1995**, *102*, 91.
(91) Lalot, J.; Stasik, I.; Demailly, G.; Beaupère, D.; Godé, P., Synthesis and amphiphilic properties of S-alkylthiopentonolactones and their pentitol derivatives. *J. Colloid Interface Sci.*, **2004**, *273*, 604.
(92) Plusquellec, D.; Brenner-Hénaff, C.; Léon-Ruaud, P.; Duquenoy, S.; Lefeuvre, M.; Wróblewski, H., An Efficient Acylation of Free Glycosylamines for the Synthesis of N-Glycosyl Amino Acids and N-Glycosidic Surfactants for Membrane Studies. *J. Carbohyd. Chem.*, **1994**, *13*, 737.
(93) Minamikawa, H.; Hato, M., Headgroup effects on phase behavior and interfacial properties of β-3,7-dimethyloctylglycoside/water systems. *Chem. Phys. Lipids*, **2005**, *134*, 151.
(94) Boullanger, P.; Chevalier, Y., Surface Active Properties and Micellar Aggregation of Alkyl 2-Amino-2-deoxy-β-d-glucopyranosides. *Langmuir*, **1996**, *12*, 1771.
(95) Burczyk, B.; Wilk, K.A.; Sokołowski, A.; Syper, L., Synthesis and Surface Properties of N-Alkyl-N-methylgluconamides and N-Alkyl-N-methyllactobionamides. *J. Colloid Interface Sci.*, **2001**, *240*, 552.
(96) Zhu, Y.-P.; Rosen, M.J.; Vinson, P.K.; Morrall, S.W., Surface Properties of N-Alkanoyl-N-methyl Glucamines and Related Materials. *J. Surfact. Deterg.*, **1999**, *2*, 357.
(97) Aveyard, R.; Binks, B.P.; Chen, J.; Esquena, J.; Fletcher, P.D.I., Surface and Colloid Chemistry of Systems Containing Pure Sugar Surfactant. *Langmuir*, **1998**, *14*, 4699.
(98) Persson, C.M.; Kjellin, U.R.M.; Eriksson, J.C., Surface Pressure Effect of Poly(ethylene oxide) and Sugar Headgroups in Liquid-Expanded Monolayers. *Langmuir*, **2003**, *19*, 8152.
(99) Milkereit, G.; Garamus, V.M.; Veermans, K.; Willumeit, R.; Vill, V., Structures of micelles formed by synthetic alkyl glycosides with unsaturated alkyl chains. *J. Colloid Interface Sci.*, **2005**, *284*, 704.
(100) Zhang, T.; Marchant, R.E., Novel Polysaccharide Surfactants: The Effect of Hydrophobic and Hydrophilic Chain Length on Surface Active Properties. *J. Colloid Interface Sci.*, **1996**, *177*, 419.
(101) Wilk, K.; Syper, L.; Burczyk, B.; Maliszewska, I.; Jon, M.; Domagalska, B., Preparation and properties of new lactose-based surfactants. *J. Surfact. Deterg.*, **2001**, *4*, 155.
(102) Ericsson, C.A.; Söderman, O.; Garamus, V.M.; Bergström, M.; Ulvenlund, S., Effects of Temperature, Salt, and Deuterium Oxide on the Self-Aggregation of Alkylglycosides in Dilute Solution. 1. n-Nonyl-β-d-glucoside. *Langmuir*, **2004**, *20*, 1401.
(103) Molina-Bolívar, J.A.; Aguiar, J.; Peula-García, J.M.; Ruiz, C.C., Surface Activity, Micelle Formation, and Growth of n-Octyl-β-d-Thioglucopyranoside in Aqueous Solutions at Different Temperatures. *J. Phys. Chem. B*, **2004**, *108*, 12813.

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

surfactant	CMC (mM)	reference
octyl-β-D-glucoside	17	82
	20	14
	24	83
	34	84
nonanoyl-N-methylglucamine	16	85
	21	86
	24	87
dodecyl-D-lactobionamide	1.3	51
	3.4	51
6-O-dodecanoylsucrose	0.25	88
	0.34	89
	0.46	90

Table 2. Dataset of experimental CMC used for the development and validation of QSPR models.

molecule	CMC (mM)	T (°C)	log CMC (M)	set[(a)]	reference
Octyl-D,L-Glycerol	5.8	25	-2.2	T	10
Octyl Glycol	4.9	25	-2.3	T	10
Octyl-β-D-Glucoside	20	25	-1.7	T	14
Octyl-α-D-Mannoside	6.0	25	-2.2	T	14
Octyl-β-D-Galactoside	16	25	-1.8	T	14
Decyl-α-D-Mannoside	0.25	25	-3.6	T	14
Decyl-β-D-Galactoside	0.70	25	-3.1	V	14
Dodecyl-α-D-Mannoside	0.05	25	-4.3	T	14
Dodecyl-β-D-Galactoside	0.20	25	-3.7	T	14
1-O-Butyl-D,L-Xylitol	58	25	-1.2	T	53
1-O-Pentyl-D,L-Xylitol	38	25	-1.4	T	53
1-O-Hexyl-D,L-Xylitol	9.4	25	-2.0	T	53
1-O-Heptyl-D,L-Xylitol	9.2	25	-2.0	V	53
1-O-Octyl-D,L-Xylitol	6.7	25	-2.2	T	53
1-O-Nonyl-D,L-Xylitol	2.1	25	-2.7	T	53
1-O-Pentanoyl-D,L-Xylitol	120	25	-0.9	T	53
1-O-Hexanoyl-D,L-Xylitol	58	25	-1.2	V	53
1-O-Heptanoyl-D,L-Xylitol	10	25	-2.0	T	53
1-O-Octanoyl-D,L-Xylitol	18	25	-1.7	T	53
1-O-Nonanoyl-D,L-Xylitol	4.4	25	-2.4	V	53
1-O-Decanoyl-D,L-Xylitol	1.8	25	-2.7	T	53
S-Butyl-1-Thio-D,L-Xylitol	180	25	-0.7	T	53
S-Pentyl-1-Thio-D,L-Xylitol	46	25	-1.3	T	53
S-Hexyl 1-Thio-L-Xylitol	12	20	-1.9	T	91
S-Octyl 1-Thio-L-Xylitol	1.2	20	-2.9	T	91
S-Decyl 1-Thio-L-Xylitol	0.40	20	-3.4	T	91
S-Hexyl 1-Thio-D-Lyxitol	18	20	-1.7	V	91
S-Octyl 1-Thio-D-Lyxitol	1.8	20	-2.7	V	91
S-Hexyl 1-Thio-L-Ribitol	10.2	20	-2.0	T	91
S-Octyl 1-Thio-L-Ribitol	0.38	20	-3.4	V	91
S-Hexyl 5-Thio-D-Arabinonolactone	6.7	20	-2.2	V	91
S-Octyl 5-Thio-D-Arabinonolactone	0.48	20	-3.3	T	91
S-Decyl 5-Thio-D-Arabinonolactone	0.033	20	-4.5	T	91
S-Hexyl 5-Thio-D-Xylonolactone	5.0	20	-2.3	V	91
S-Octyl 5-Thio-D-Xylonolactone	0.53	20	-3.3	T	91
S-Decyl 5-Thio-D-Xylonolactone	0.023	20	-4.6	T	91
1-Butylhexyl-β-D-Glucoside	15	25	-1.8	T	54
1-Propylheptyl-β-D-Glucoside	12	25	-1.9	V	54
1-Ethyloctyl-β-D-Glucoside	8.5	25	-2.1	T	54
1-Methylnonyl-β-D-Glucoside	4.5	25	-2.3	T	54
Octanoyl-β-D-Galactosylamine	45	25	-1.3	V	92
Octanoyl-β-D-Glucosylamine	70	25	-1.2	T	92
3,7-Dimethyloctyl-β-D-Glucoside	4.0	25	-2.4	T	93

3,7-Dimethyloctyl-β-D-Maltoside	5.3	25	-2.3	T	93
3,7-Dimethyloctyl-β-D-Maltotrioside	5.0	25	-2.3	T	93
2-Amino-2-Deoxy-Octyl-β-D-Glucoside	23	25	-1.6	T	94
2-Amino-2-Deoxy-Nonyl-β-D-Glucoside	7.0	25	-2.2	T	94
N-Decyl-N-Methyl Gluconamide	1.3	25	-2.9	V	95
N-Dodecyl-N-Methyl Gluconamide	0.14	20	-3.8	V	95
N-Tetradecyl-N-Methyl Gluconamide	0.024	20	-4.6	V	95
N-Oleyl-N-Methyl Gluconamide	0.032	20	-4.5	T	95
N-Decyl-N-Methyl Lactobionamide	2.3	20	-2.6	V	95
N-Dodecyl-N-Methyl Lactobionamide	0.25	20	-3.6	T	95
N-Tetradecyl-N-Methyl Lactobionamide	0.036	20	-4.4	V	95
N-Hexadecyl-N-Methyl Lactobionamide	0.0093	20	-5.0	V	95
N-Octadecyl-N-Methyl Lactobionamide	0.0033	20	-5.5	T	95
N-Oleyl-N-Methyl Lactobionamide	0.054	20	-4.3	V	95
Dodecanoyl-N-Methylglyceramine	0.23	25	-3.6	V	96
Dodecanoyl-N-Methylxylamine	0.36	25	-3.4	T	96
Octanoyl-N-Methylglucamine	69	25	-1.2	T	86
Nonanoyl-N-Methylglucamine	21	25	-1.7	V	86
Decanoyl-N-Methylglucamine	6.7	25	-2.2	T	86
Decyl-β-D-Maltoside	2.0	25	-2.7	T	97
Decyl-β-D-Glucoside	2.0	25	-2.7	V	97
Dodecyl-β-D-Maltoside	0.17	22	-3.8	T	98
Oleyl-β-D-Maltoside	0.02	25	-4.7	T	99
Oleyl-β-D-Maltotrioside	0.042	25	-4.4	T	99
[N-(Oleoyl)-2 -Ethylamino]-β-D-Maltoside	0.09	25	-4.0	T	99
Hexyl-D-Maltonamide	83	25	-1.1	V	100
Octyl-D-Maltonamide	5.7	25	-2.2	V	100
Decyl-D-Maltonamide	1.3	25	-2.9	T	100
Dodecyl-D-Maltonamide	0.31	25	-3.5	V	100
Decyl-D-Lactobionamide	1.3	25	-2.9	T	51
6'-O-Dodecanoylmaltose	0.33	22	-3.5	T	55
N-Decanoyl-N-Methyl Lactitolamine	3.3	25	-2.5	T	101
N-Dodecanoyl-N-Methyl Lactitolamine	0.45	25	-3.3	T	101
N-Tetradecanoyl-N-Methyl Lactitolamine	0.068	25	-4.2	T	101
Nonyl-β-D-Glucoside	6.9	20	-2.2	V	102
6-O-[(Hexyloctyl)-3-Propylsulfide)ethanoyl]-D-Mannose	0.0063	25	-5.2	T	56
6-O-Dodecanoylsucrose	0.46	20	-3.3	V	90
6-O-Dodecanoylraffinose	0.95	20	-3.0	T	90
6-O-Dodecanoylstachyose	2.3	20	-2.6	T	90
Octyl-β-D-Thioglucoside	8.5	25	-2.0	V	103

(a) T: training set, V:validation set

Table 3. Validation thresholds, in accordance with Chirico et al.[75]

validation metric	Chirico thresholds[75]
R^2_{EXT}	>0.70
Q^2_{F1}	>0.70
Q^2_{F2}	>0.70
Q^2_{F3}	>0.70
CCC	>0.85
$\overline{r^2_m}$	>0.65
Δr^2_m	<0.20

Table 4. Performances of the new QSPR models compared to the one of Mattei[46].

model	n_{desc}	descriptors	R^2	RMSE (log)	R^2_{IN}	$RMSE_{IN}$ (log)	n_{out}
integral/all types (eq. 2)	4	$^2ACIC, \eta, T_e, {}^1AIC$	0.93	0.31	0.91	0.32	1
integral/constitutional and topological (eq. 3)	4	$^2ACIC, n_O, {}^0AIC, n_S$	0.87	0.41	0.89	0.37	3
integral/constitutional (eq. 4)	4	$n_{rel,C}, n_H, n_{rings}, n_S$	0.82	0.47	0.78	0.52	0
fragments/all types (eq. 5)	3	$TMSA_c, \eta_h, {}^2IC_h$	0.93	0.30	0.88	0.36	1
fragments/constitutional and topological (eq. 6)	2	$^2CIC_c, n_{rel,S,h}$	0.87	0.41	0.89	0.36	0
fragments/constitutional (eq. **7**)	3	$M_{w,c}, n_{rel,S,h}, n_{rel,single,c}$	0.86	0.41	0.88	0.36	0
Mattei et al.[46]	-	*group contributions*	-	-	*0.44*[a] *(0.50)*[b]	*1.11*[a] *(0.93)*[b]	-

a) calculated for the validation set (27 molecules); b) calculated for the whole dataset (83 molecules); n_{desc}: number of descriptors; n_{out}: number of molecules out of AD of the model.

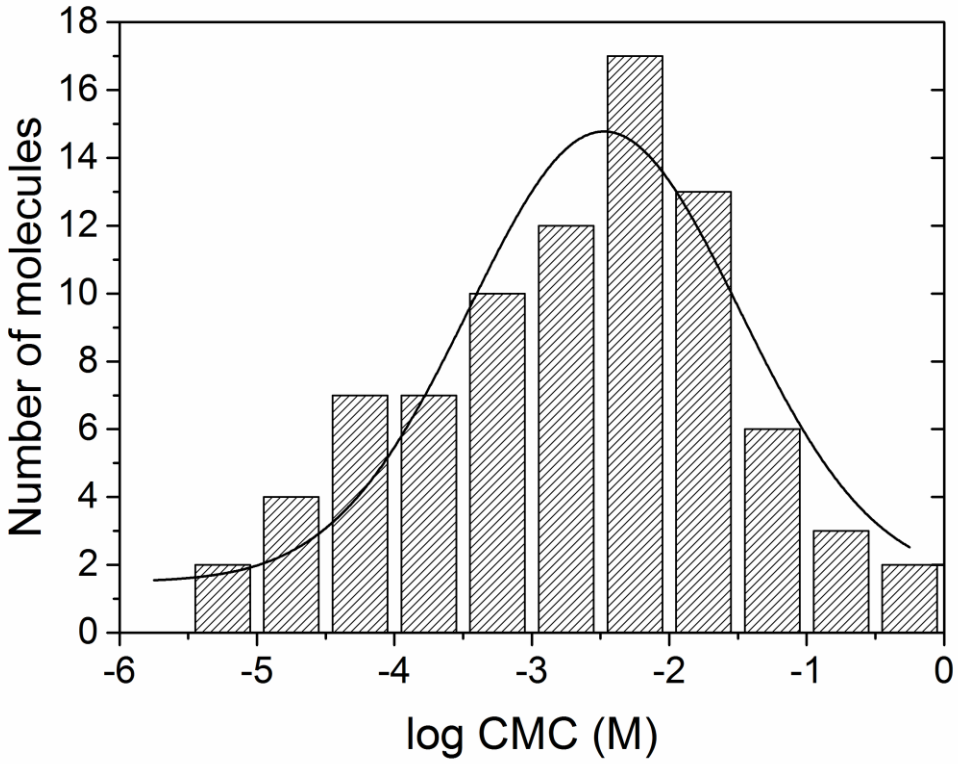

Figure 1. Distribution of log CMC from the dataset presented in Table 2.

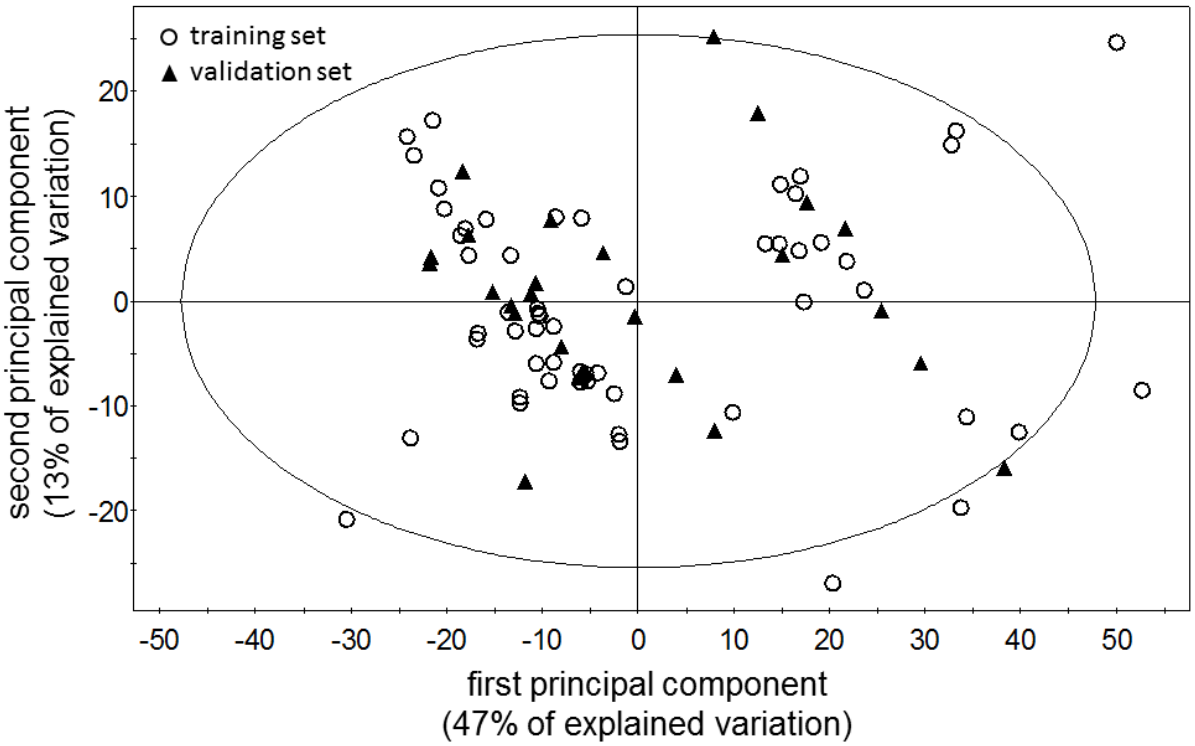

Figure 2. Repartition of the molecules belonging to the training (circles) and the validation (triangles) sets in the chemical space of the whole dataset as defined by Principal Component Analysis based on 953 descriptors

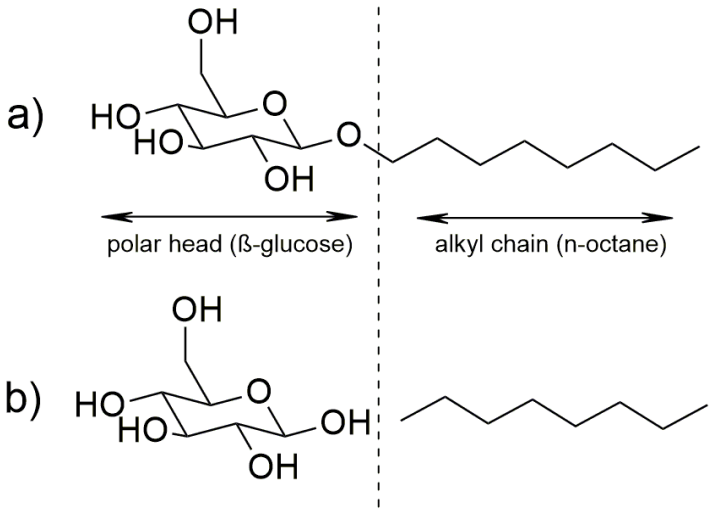

Figure 3. Definition of polar heads and alkyl chains in this study, exemplified for octyl-β-D-glucoside, a) representation of octyl-β-D-glucoside, b) fragments modeled for octyl-β-D-glucoside.

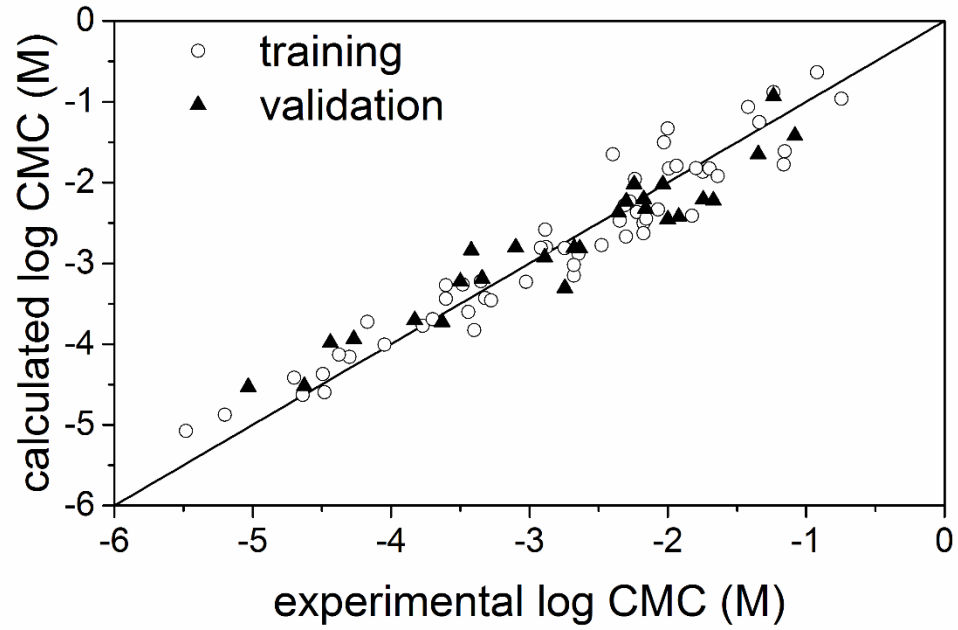

Figure 4. Plot of experimental vs. calculated values for the model based on integral descriptors of all types (eq. 2)

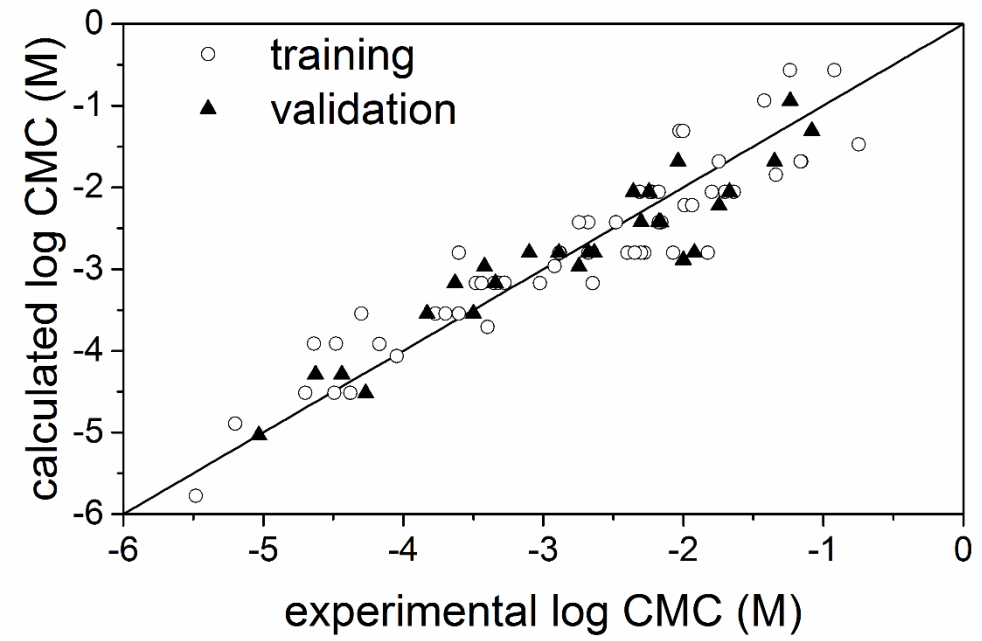

Figure 5. Plot of experimental vs. calculated values for the model based on constitutional fragment descriptors (eq. 7).

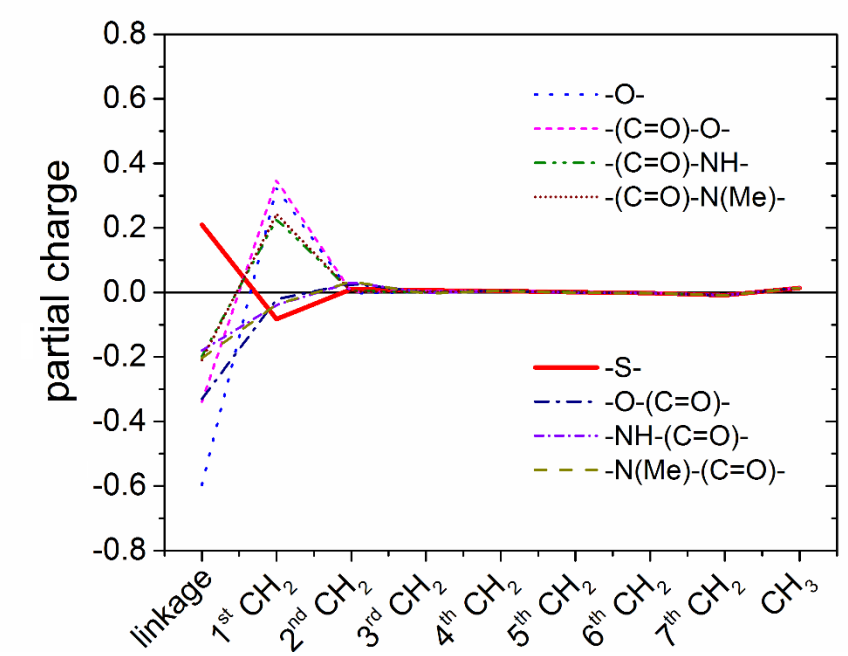

Figure 6. Partial charges from the linkage along the alkyl chain for sugar-based surfactants with different linkages.

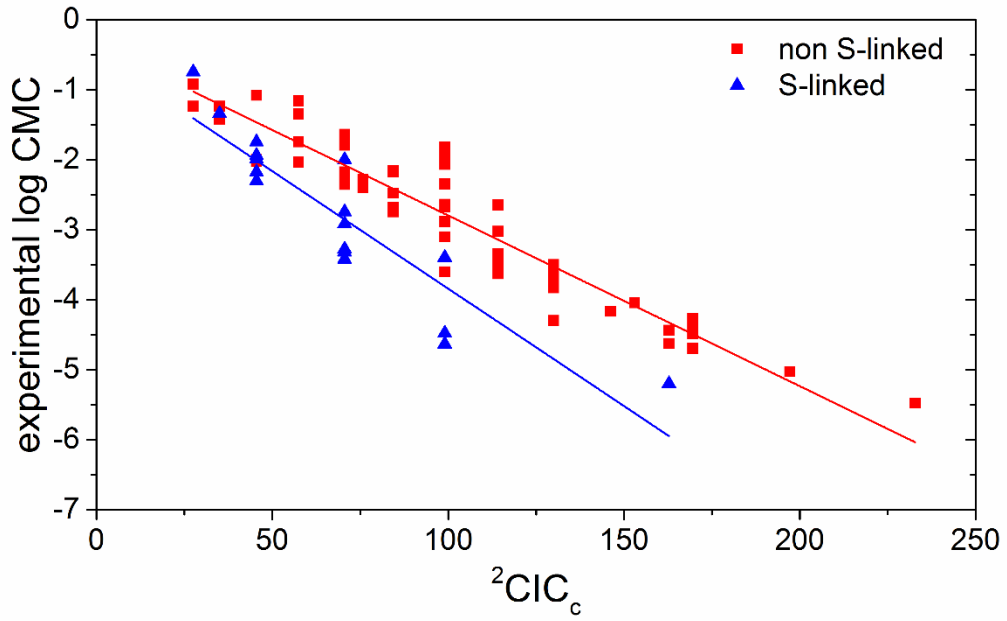

Figure 7. Correlation between Complementary Information Content (order 2) for the alkyl chain fragment and log CMC, separately for S-linked and non S-linked surfactants.