Estimating the adsorption efficiency of sugar-based surfactants from QSPR models

Théophile Gaudin, Patricia Rotureau, Isabelle Pezron, Guillaume Fayet

Estimating the adsorption efficiency of sugar-based surfactants from QSPR models

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; fax: +33(0)344556565

Abstract

Adsorption efficiency, measured as the surfactant concentration at which the surface tension of the aqueous solution decreases by 20 mN/m, characterizes their affinity for surfaces and interfaces, which is crucial for a cost-effective use of surfactants. In this article, the first Quantitative Structure Property Relationship models to predict this efficiency were proposed based on a dataset of 82 diverse sugar-based surfactants and using different types of molecular descriptors. Finally, an easy-to-use model was evidenced with good predictivity assessed on an external validation set. Moreover, it is based on a series of fragment descriptors accounting for the different structural trends affecting the efficiency of sugar-based surfactants. Due to its predictive capabilities and to the structure-property trends it involves, this model open perspectives to help the design of new sugar-based surfactants, notably to substitute petroleum-based ones.

Keywords: QSPR; sugar-based surfactants; efficiency; bio-based compounds

1 Introduction

In the context of development and use of more environmentally-friendly products, efforts are nowadays in progress towards the design of new surfactants issued from renewable resources to substitute petroleum-based surfactants that constitute most part of the market. Among them, sugar-based surfactants are an important subfamily of bio-based surfactants (contributing up to 40% of biosurfactant consumption (Anbu, 2017)), characterized by their polar head constituted by carbohydrates such as glucose, maltose or sucrose, and their derivatives. For this reason, sugar-based surfactants can be obtained from renewable resources such as starch (Kjellin & Johansson, 2010), and are often biocompatible and easily biodegradable (Matsumura, Imai, Yoshikawa, Kawada, & Uchibor, 1990). So, they are commonly considered among the most promising alternatives to conventional petroleum-based non-ionic surfactants (Hill & LeHen-Ferrenbach, 2009), particularly regarding soft detergents or personal care products, cosmetics and pharmaceutical formulations (Rojas, Stubenrauch, Lucia, & Habibi, 2009).

The main performance characteristics of surfactants that are used to select surfactants in application formulations relate to their effectiveness and their efficiency. The target effectiveness, *i.e.* the maximum performance to reach, is the maximum lowering of the surface tension in aqueous solution. This maximum is almost reached when the concentration in surfactants favors their aggregation to form micelles in the solution. This concentration is the critical micelle concentration (CMC). So, effectiveness is characterized by measuring the surface tension at CMC, denoted γ_{CMC} (Rosen, 1976).

Efficiency (Rosen, 1974) represents the amount of surfactant needed to reach a given performance. In surfactant solutions, two phenomena are in competition, adsorption at solvent/air interface and aggregation into micelles. While CMC is more related to aggregation, adsorption efficiency (commonly simply named efficiency) is measured as the quantity of surfactants needed to decrease the surface tension of the aqueous solution by 20 mN/m, in negative logarithmic unit by convention (-log C_{20}, denoted pC_{20}) (Rosen & Kunjappu, 2012). This level of reduction of surface tension is expected as generally sufficient to characterize a quasi-saturated interface (Rosen, 1974).

In applications, a high efficiency enables to reduce the amount of surfactant used (Fisher, Zeringue, & Feuge, 1977; Rosen, 1974). Indeed, using efficient surfactants allows to limit the amount of surfactant in formulation for a same target application and, as a consequence, limit their cost and, in some cases, avoid their possible toxicity.

If CMC and pC_{20} characterize different aspects of the behavior of surfactants in solution (aggregation for CMC and adsorption for pC_{20}), they are both related to the hydrophobic effect (Tanford, 1979), that minimizes the contact between solvent and alkyl chains. So, it is not surprising to note that some common trends have been highlighted between the molecular structure and both the CMC (in log) and the pC_{20} in the experimental literature (Rosen & Kunjappu, 2012). For instance, in general, pC_{20} was observed to increase with alkyl chain length (Rosen, 1974; Zhu, Rosen, Vinson, & Morrall, 1999), and to slightly decrease with polar head size, which also affects CMC (Crook, Fordyce, & Trebbi, 1963; Myers, 2006; Rosen & Kunjappu, 2012).

The relationship between the molecular structure and pC_{20} was quantified by Rosen (Rosen, 1974) under a thermodynamic formalism by considering the free energies of transfer of the CH_n groups and of the polar head from solvent to interface. This approach has notably been used to evidence structural trends of a series of conventional surfactants (without any sugar-based surfactants), like the increase of efficiency with the length of the alkyl chain or its decrease in the case of ionic surfactants.

To reach predictive models, in particular for sugar-based surfactants, the quantitative structure-property relationship approach represents a relevant predictive method that already demonstrated success for physico-chemical properties (Katritzky et al., 2010; Nieto-Draghi et al., 2015) notably for surfactants (Creton, 2013; Hu, 2010) and, in particular, sugar-based ones (Gaudin, Rotureau, Pezron, & Fayet, 2016, 2018). But up to now, none were dedicated to the adsorption efficiency of surfactants.

In this study, a series of six QSPR models dedicated to the adsorption efficiency of sugar-based surfactants were developed based on different types of descriptors to access models accounting at best for the different structural trends influencing the adsorption efficiency of sugar-based surfactants. The inclusion of quantum chemical descriptors allows to access to physically meaningful description of

molecular structures, notably in terms of charge distributions, whereas topological and, furthermore, constitutional descriptors are easier to calculate. Moreover, fragment-based descriptors were introduced. Computed separately for the polar head and the alkyl chain of the surfactant, they are particularly representative of the specificity of the surfactants' structure and may allow to distinguish the specific impacts of polar heads and alkyl chains on their efficiency.

2 Computational details

2.1 Experimental dataset

The experimental data used in this paper were issued from a large data collection on the properties of sugar-based surfactants (Gaudin, 2016). More than 2500 data on different amphiphilic properties (including, notably, critical micelle concentrations (CMC), surface tensions at CMC, Krafft points and adsorption efficiencies) for more than 600 different sugar-based surfactants were extracted from literature constituting the largest database for this family of compounds, to the best of our knowledge.

The 172 gathered data related to efficiency were analyzed in detail to select only the most reliable ones for the development and the validation of QSPR models. Indeed, the performances of QSPR models are critically dependent on the number and quality of the data employed for its development. For this reason, all the selected data were obtained using a robust protocol and under homogeneous conditions. Indeed, all selected pC_{20} were measured near room temperature (*i.e.* between 20°C and 25°C), since temperature can strongly influence pC_{20} values (Mahmood & Al-Koofee, 2013).

Moreover, all pC_{20} values collected in the dataset (cf. Table 1) were measured by tensiometry (using, e. g. the Wilhelmy plate method (Le Neindre, 1993) or the Du Noüy Ring (du Noüy, 1925)) upon the following general principle. Aqueous solutions at various concentrations of surfactants are prepared and the measured surface tension are plotted versus concentration in logarithmic scale. This curve enables the determination of various properties, like the CMC, γ_{CMC}, and the efficiency. In the case of efficiency, the concentration at which the surface tension has decreased by 20 mN/m with respect to the pure water is extracted, and then, adsorption efficiency is calculated as the negative logarithmic transformation of this concentration, pC_{20}.

<< TABLE 1 >>

The collected data allow to figure out the experimental variance that can be found in literature for a single molecule by different experimentalists. As shown in Table 2, this variance reaches 0.7 (log unit) for 1-O-octanoyl-D,L-xylitol.

<< TABLE 2 >>

After data curation, the final dataset (in Table 1) was constituted of 82 pC_{20} of diverse sugar-based surfactants, 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 pC_{20} data is relatively homogeneous, with data ranging between 1.70 and 6.46 (log unit) and a maximum between 3 and 3.5 (log unit), as shown in Figure 1.

<< FIGURE 1 >>

To perform an external validation of the model (to evaluate its predictive power), the dataset was divided into two parts taking care of their similarity in terms of property distribution and chemical diversity to ensure that the validation set is at best representative of the applicability domain of the model. The partition was performed by a property-ranged approach to ensure similar distributions of efficiency values in both sets. Surfactants were classified by increasing order of pC_{20} and the molecules of the validation set were selected regularly to obtain a 1:3 partition, *i.e.* by selecting the 2^{nd}, 5^{th}, 8^{th} molecule, etc. Hence, a training set of 55 surfactants was selected for the development of the model and a validation set of 27 surfactants was kept apart for the external validation. The similarity between both sets in terms of chemical diversity was also checked based on a principal component analysis on all computed descriptors. The molecules of both sets revealed well-distributed in the global chemical space of the investigated surfactants, as shown in Figure 2.

<< FIGURE 2 >>

2.2 *Molecular descriptors*

The molecular structures of the 82 studied sugar-based surfactants of the dataset were optimized from Density Functional Theory (DFT) at B3LYP/6-31+G(d,p) level with Gaussian 09 software (Frisch et al., 2009) after preliminary conformation analyses to identify the most stable conformation (Gaudin, Rotureau, Pezron, & Fayet, 2017). Frequency calculations were performed at same level to ensure they correspond to local minima, *i.e.* presenting no imaginary frequency. This level of calculation was found relevant for sugar-based surfactants to evidence the influence of their molecular structure on a series of molecular descriptors in a previous work (Gaudin et al., 2017) and it was successfully used in previous QSPR study targeting their CMC (Gaudin et al., 2016).

The structures of the 35 polar heads and 18 alkyl chains constituting these 82 surfactants were also computed on the same scheme. These fragments were separated before the first heteroatom and each fragment was saturated in hydrogen, as illustrated in Figure 3.

<< FIGURE 3 >>

Based on these structures, *i.e.* for each surfactant and each fragment, more than 300 descriptors were computed using CODESSA software . These descriptors are of four types:

- constitutional descriptors, related to the number of specific types of atoms, functional groups and bonds (*e.g.* number of carbon, of single bonds);
- topological descriptors, that characterize the atomic connectivity including information about the size, composition and degree of branching (*e.g.* Wiener index);
- geometrical descriptors, related to the 3D molecular structure (*e.g.* molecular volume, bond length);
- quantum-chemical descriptors, including atomic charges, electronic and binding information.

Within the investigated dataset, some sugar-based surfactants were enantiomeric mixtures (Savelli et al., 1999), for D and L sugar alcohol polar heads, or anomeric (Boyère, Broze, Blecker, Jérôme, & Debuigne, 2013) mixtures, for polar heads containing a free anomeric alcohol. In these cases, the different isomers were considered as conformations of the same compound. The different isomers

(possibly in different conformations) were calculated at B3LYP/6-31+G(d,p) level and the most stable one was used to compute the molecular descriptors.

2.3 *Model development and validation*

In this study, multi-linear regressions (MLR) were developed to favor a simpler use for predictions and an easier interpretation of the role of each descriptor in the predictive power of the model from both a statistical and phenomenological point of view.

The selection of the descriptors included into the models was performed using the Best MultiLinear Regression (BMLR) approach, as implemented in CODESSA software . This algorithm was described in details and successfully used in previous works (Fayet, Rotureau, Joubert, & Adamo, 2010a, 2010b) notably for the CMC of sugar-based surfactants (Gaudin et al., 2016). The final model was chosen as the best compromise between correlation and number of descriptors to avoid against any over-parameterization.

To evaluate the performances of models, a series of internal and external validations were performed. 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.

The robustness of the models was assessed by using leave-one-out (LOO) and leave-many-out (LMO) cross-validations. The Q^2_{LOO}, Q^2_{3CV}, Q^2_{7CV} and Q^2_{10CV} coefficients issued at various levels of partition (for LOO, 3-fold, 7-fold and 10-fold cross-validations, respectively) were checked to be as close as possible to R^2 and close one to each other.

A Y-scrambling test was performed to ensure that models were not issued from chance correlations. 500 random permutations of experimental property values were performed. New models were refitted (Lindgren, Hansen, Karcher, Sjöström, & Eriksson, 1996) at each permutation and the average (R^2_{YS}) and standard deviation (SD_{YS}) in the R^2 of the new models were calculated. In absence of chance

correlation, low values of R^2_{YS} are expected. As proposed by Rücker (Rücker, Rücker, & Meringer, 2007), R^2_{YS} should be superior to 2.3 SD_{YS} for a model to be considered as not issued from chance correlations with a 99% confidence level.

Then, an external validation was performed to evaluate the predictive power of the selected models, on the validation set of 27 surfactants. The coefficient of determination R^2_{EXT}, the mean absolute error MAE_{EXT} and the root mean square error $RMSE_{EXT}$ were calculated between experimental and predicted values of efficiency. A series of additional external validation metrics were also used (listed in Table 3), for which thresholds values have been proposed by Chirico et al. (Chirico & Gramatica, 2012) to estimate that a QSPR model is reliable.

<< TABLE 3 >>

In a last step, the applicability domain (AD), *i.e.* the domain in which a prediction is valid, was defined. As a QSPR model is issued from a similarity principle, it is expected to be reliable for molecules similar to those used to fit it. So, the AD of each model has been defined by the range of values of the calculated descriptors and of the experimental property in the training set. To evidence the actual performances of the model inside its AD, all external validation criteria were calculated again considering only the molecules of the validation set belonging to the applicability domain (R^2_{IN}, MAE_{IN}, $RMSE_{IN}$, $Q^2_{F1,IN,}$ $Q^2_{F2,IN,}$ $Q^2_{F3,IN,}$ CCC_{IN}, $\overline{r^2_{m,IN}}$, $\Delta r^2_{m,IN}$).

3 Results

As explained in the previous sections, different types of descriptors were developed from the simplest constitutional counts to quantum chemical descriptors. Moreover, some descriptors were calculated from the whole molecular structure whereas others were related to the alkyl chain or to the polar head. Finally, six new QSPR models were developed in this study for pC_{20}. Three models were based on integral descriptors: 1) based on all types of descriptors, 2) limited to topological and constitutional descriptors or 3) focused on constitutional descriptors to favor simpler models. On the same scheme, three other models were based on fragment descriptors.

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. 1) was found as the best compromise between correlation and number of descriptors among the 17 equations sorted out by the BMLR method:

$$pC_{20} = 1.952\ {}^{2}ACIC - 77.58\ HACA_{2,TMSA} - 113.8\ N_{O,avg} - 0.302\ n_{rings} + 2.80 \quad (1)$$

with ${}^{2}ACIC$ the Average Complementary Information Content of order 2, $HACA_{2,TMSA}$ the area-weighted surface charge of hydrogen bonding acceptor atoms divided by the total molecular surface area, based on the Mulliken partial charges (Mulliken, 1955), $N_{O,avg}$ the average nucleophilic reactivity index for a O atom and n_{rings} the number of rings in the molecule.

The model is characterized by good fitting (R^2 = 0.90, RMSE = 0.37 (log unit)) and robustness (Q^2_{CV} = 0.88, Q^2_{10CV} = 0.87, Q^2_{7CV} = 0.87, Q^2_{3CV} = 0.86). From the Y-scrambling test, the criterion of Rücker (Rücker et al., 2007) revealed satisfied: $R^2 - R^2_{YS} = 0.81 > 2.3\ SD_{YS} = 0.12$. So, the model was not issued from a chance correlation.

When tested in external validation, a good predictivity was obtained for this model with $RMSE_{IN}$ = 0.49. In particular, it fulfilled all the criteria of Chirico et al. ($R^2_{IN} = 0.83 > 0.70$, $Q^2_{F1,IN} = 0.82 > 0.70$, $Q^2_{F2,IN} = 0.82 > 0.70$, $Q^2_{F3,IN} = 0.83 > 0.70$, $CCC_{IN} = 0.89 > 0.85$, $\overline{r^2_m}_{IN} = 0.70 > 0.65$, $\Delta r_m{}^2_{IN} = 0.16 < 0.20$) in its applicability domain.

It has to be noticed that the only surfactant of the validation set that felt out of AD was octyl glycol (Figure 4) due to a slightly too low $HACA_{2,TMSA}$ (with $9.36 \cdot 10^{-3}$ whereas the AD ranged between $1.19 \cdot 10^{-2}$ and $4.04 \cdot 10^{-2}$) and a too high value of $N_{O,avg}$ ($2.97 \cdot 10^{-2}$ vs. an AD between $1.95 \cdot 10^{-6}$ and $2.01 \cdot 10^{-2}$). This molecule has the smallest polar head of the dataset (notably when compared to octyl-β-D-glucoside in Figure 3), with only two hydroxyl groups, which explains its finding at the limits of the

AD. However, the predicted pC_{20} for this surfactant presented a deviation of 0.39 (log unit), *i.e.* within the standard error in the validation set ($RMSE_{IN}$ = 0.49 (log unit)).

<< FIGURE 4 >>

3.1.2 Models with topological and constitutional descriptors

The use of the only 74 constitutional and topological descriptors led to a five-parameter model (Eq. 2) as best compromise between correlation and number of descriptors among the 7 equations sorted out by the BMLR method:

$$pC_{20} = 1.093\, n_S - 9.64\, {}^2ASIC + 0.224\, {}^1\chi^v - 0.796\, n_{rings} + 42.89\, n_{C,rel} - 4.10 \qquad (2)$$

with n_S the number of S atoms, ${}^2ASIC$ the Average Structural Information Content of order 2, ${}^1\chi^v$ the Kier & Hall index of order 1, n_{rings} the number of rings, and $n_{C,rel}$ the relative number of C atoms.

The model was well fitted for the training molecules (R^2 = 0.92, RMSE = 0.34 (log unit)) and revealed robust from LOO and LMO cross-validations with Q^2_{nCV} = 0.87-0.90. The Y-scrambling demonstrated that it did not originate from a chance correlation according to Rücker's criteria: $R^2 - R^2_{YS} = 0.83 > 2.3\ SD_{YS} = 0.12$.

The predictivity of Eq. 2 was found better to that of Eq. 1 with $RMSE_{IN}$ = 0.41 (log unit) and better external validation metrics (R^2_{IN} = 0.88, $Q^2_{F1,IN} = Q^2_{F2,IN}$ = 0.86, $Q^2_{F3,IN}$ = 0.88, CCC_{IN} = 0.93, $\overline{r^2_m}_{IN}$ = 0.83 $\Delta r_m^2{}_{IN}$ = 0.01), for the 25 molecules of the validation set that belong to AD.

Two molecules were found out of AD. For the S-butyl-1-thio-D,L-xylitol, the relative number of carbon $n_{C,rel}$ was 0.265 for an AD between 0.270 and 0.313 and the error in prediction for this molecule (0.52) was only slightly higher than $RMSE_{IN}$. For the octyl glycol, ${}^1\chi^v$ = 5.1 is lower than the range of the AD at [5.3;20.0] and predicted pC_{20} was overestimated by 0.80, as shown in Table 1.

3.1.3 Model with constitutional descriptors

The five-parameter model in Eq. 3 was chosen among the six equations sorted out by the BMLR method when focusing on the only 36 constitutional descriptors of the entire surfactants:

$$pC_{20} = 46.45\, n_{C,rel} - 0.802\, M_{w,rel} + 8.52 \cdot 10^{-2}\, n_H + 1.452\, n_S - 0.534\, n_{rings} - 7.68 \qquad (3)$$

with $n_{rel,C}$ the relative number of C atoms, $M_{w,rel}$ the relative molecular weight, n_H the number of H atoms, n_S the number of S atoms and n_{rings} the number of rings.

If the goodness of fit of this model is slightly lower than Eqs. 1 and 2 with $R^2 = 0.88$ and RMSE = 0.41 (log unit), it proved robust ($Q^2_{nCV} = 0.83$-0.85). The Y-scrambling ensured that the model was not issued from a chance correlation with low values of R^2 for the models obtained after randomization ($R^2_{YS} = 0.09$; $SD_{YS} = 0.05$).

Once again, the predictive power revealed slightly lower than Eqs. 1 and 2 with $RMSE_{IN} = 0.45$ (log unit). However, all criteria of Chirico are satisfied ($R^2_{IN} = 0.84$, $Q^2_{F1} = 0.83$, $Q^2_{F2} = 0.83$, $Q^2_{F3} = 0.86$, CCC = 0.91, $\overline{r^2_m} = 0.77$ and $\Delta r_m^2 = 0.12$). The two molecules out of AD are the same than for Eq. 2, based on the same criteria (S-butyl-1-thio-D,L-xylitol due to $n_{C,rel} = 0.265$ vs. [0.270;0.313] and octyl glycol due to $M_{w,rel} = 5.13$ vs. [5.38;6.96]) and for both of them pC_{20} the overestimation was of 0.69 and 0.55, respectively.

3.2 *Models focused on fragment-based descriptors*

3.2.1 Model with all types of descriptors

A total of 627 fragment-based descriptors were calculated when considering the descriptors of the polar head and of the alkyl chain. From this large set of descriptors, the five-parameter model in Eq. 4 was selected as giving the best compromise between correlation and number of descriptors among the 13 equations sorted out by the BMLR method:

$$pC_{20} = -4.97\ FHBSA_h + 1.60\cdot10^{-2}\ {}^2CIC_c - 1.616\ RNCS_c - 6.31\cdot10^{-2}\ {}^0SIC_h - 2.95\ {}^2ASIC_c + 10.22 \quad (4)$$

with $FHBSA_h$ the fractional H-bonding surface area of the polar head based on Mulliken partial charges (Mulliken, 1955), ${}^2CIC_c$ the complementary information content of order 2 of the alkyl chain, $RNCS_c$ is the relative negatively charged surface area of the alkyl chain based on Zefirov partial charges (Zefirov, Palyulin, Oliferenko, Ivanova, & Ivanov, 2001), ${}^0SIC_h$ is the structural information content of order 0 of the polar head and ${}^2ASIC_c$ is the average structural information content of order 2 for the alkyl chain.

An excellent fitting was obtained with the training set data with a coefficient of determination $R^2 = 0.95$ and a low error (RMSE = 0.28 (log unit)), the model appears to be robust ($Q^2_{CV} = Q^2_{10CV} = Q^2_{3CV} = 0.93$, $Q^2_{7CV} = 0.92$) and not issued from chance correlation as evidenced by Y-scrambling ($R^2_{YS} = 0.09$ and $SD_{YS} = 0.05$).

This model demonstrated a good predictive power satisfying all criteria of Chirico with $R^2_{IN} = 0.87$, $Q^2_{F1,IN} = 0.85$, $Q^2_{F2,IN} = 0.85$, $Q^2_{F3,IN} = 0.86$, $CCC_{IN} = 0.91$, $\overline{r^2_m}_{IN} = 0.70$, $\Delta r_m^2{}_{IN} = 0.15$ and an error of $RMSE_{IN} = 0.44$ (log unit) in its applicability domain.

5 surfactants of the validation set were found out of AD of this model. S-Hexyl, S-Octyl and S-Decyl 5-Thio-D-Xylonolactone were found nearly on the threshold of the AD in terms of $FHBSA_h$ with 0.7276 for these 3 surfactants for an AD between 0.7283 and 0.9929. So, it is not surprising to note that these molecules were correctly predicted with deviations between experimental and predicted efficiencies of 0.05 to 0.42 (*i.e.* within the $RMSE_{IN}$ level at 0.44 (log unit)). Once again, octyl glycol is found out of AD due to its value of ${}^0SIC_h$ (4.1 vs. AD = [5.2;22.8]) but the predicted value of pC_{20} was finally close to experiment (3.51 (log unit) vs. $pC_{20,exp} = 3.19$ (Shinoda, Yamanaka, & Kinoshita, 1959)). At last, a pC_{20} of 1.67 (log unit) was predicted for 1-O-hexanoyl-D,L-xylitol, *i.e.* minor to the AD (1.70 to 6.46 (log unit)) but this prediction remains within the expected uncertainty of the model, when compared to the experimental value of pC_{20} at 2.10 (log unit) (Savelli et al., 1999).

3.2.2 Topological/Constitutional descriptors

To avoid quantum chemical calculations, the second fragment-based model focused on the 150 constitutional and topological descriptors. A four-parameter model (Eq. 5) was finally selected.

$$pC_{20} = 1.129\ n_{S,h} + 0.753\ {}^{2}\chi_{c} - 6.78\ {}^{2}ABIC_{c} - 3.26 \cdot 10^{-2}\ {}^{1}SIC_{h} + 4.74 \quad (5)$$

with $n_{S,h}$ the number of S atoms in the polar head, ${}^{2}\chi_{c}$ the Randic index of order 2 in the alkyl chain, ${}^{2}ABIC_{c}$ the Average Bonding Information Content of order 2 in the alkyl chain and ${}^{1}SIC_{h}$ the Structural Information Content of order 1 in the polar head.

Based on four parameters, this model presents good correlation ($R^2 = 0.92$, RMSE = 0.33 (log unit)) and robustness ($Q^2_{CV} = 0.91$, $Q^2_{10CV} = Q^2_{7CV} = 0.90$, $Q^2_{3CV} = 0.89$). The Y-scrambling test ensures that the model was not obtained by chance correlation considering the Rücker criterion at a 99% confidence level: $R^2 - R^2_{YS} = 0.85 > 2.3SD_{YS} = 0.11$.

This model exhibits also good predictivity in its applicability domain, in which all the molecules of the validation set were included. Once again it fulfilled Chirico criteria ($R^2_{IN} = 0.87$, $RMSE_{IN} = 0.42$ (log unit), $Q^2_{F1} = Q^2_{F2} = Q^2_{F3} = 0.87$, CCC = 0.92, $\overline{r^2_m} = 0.75$, $\Delta r_m^2 = 0.13$). The predictivity of this model including constitutional and topological descriptors is similar to the quantum chemical model ($RMSE_{IN}$ = 0.42 vs. 0.44 (log unit) for Eq. 4).

3.2.3 Model based on constitutional descriptors

At last, focusing on the 72 fragment-based constitutional descriptors, a four-parameter model (Eq. 6) was found to be the best compromise between correlation and number of descriptors among the 8 equations proposed by the BMLR method.

$$pC_{20} = 26.36\ n_{S,rel,h} + 2.70 \cdot 10^{-2}\ M_{w,c} - 0.199\ n_{rings,h} + 58.77\ n_{single,rel,c} - 58.71 \quad (6)$$

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, $n_{rings,h}$ the number of rings in the polar head and $n_{rel,single,c}$ the relative number of single bonds in the alkyl chain.

This model is well-fitted with the training set surfactants (R^2 = 0.91, RMSE = 0.35 (log unit)), and presents a good robustness (Q^2_{CV} = 0.89, Q^2_{10CV} = Q^2_{7CV} = Q^2_{3CV} = 0.88). The Y-scrambling confirms that it is not issued from chance correlation: R^2_{YS} = 0.08 and SD_{YS} = 0.11.

At last, Eq. 6 demonstrates similar predictive performances than more complex fragment-based models (Eqs. 4-5). As shown in Figure 5, the errors observed for the molecules of the validation set within AD are low ($RMSE_{IN}$ = 0.43 (log unit)) and all validation metrics fit with the Chirico criteria (R^2_{IN} = 0.85, $Q^2_{F1,IN}$ = $Q^2_{F2,IN}$ = 0.84, $Q^2_{F3,IN}$ = 0.87, CCC_{IN} = 0.91, $\overline{r^2_m}_{IN}$ = 0.73, $\Delta r_m^2{}_{IN}$ = 0.15) within its applicability domain.

<< FIGURE 5 >>

Only one surfactant of the validation set revealed out of AD, N-octadecyl-N-methyl lactobionamide, which presents the longest saturated alkyl chain of the dataset (containing 18 C atoms) leading to a predicted pC_{20} slightly higher than the AD range (6.74 (log unit) > 6.46 (log unit)). However, the observed error in prediction for this molecule, 0.40 (log unit), is in line with the error obtained in the validation set ($RMSE_{IN}$ = 0.43 (log unit)).

Finally, this model is particularly appealing since it is based on very simple descriptors and remains reliable enough for good-quality and fast estimation of pC_{20} of sugar-based surfactants in the perspective of molecular screening or discovery, notably for formulation specialists.

4 Discussion

A summary of the six models developed in this study is proposed in table 4 and the applicability domain associated to the descriptors they used are provided in Table 5. These models present good predictive powers with standard errors between 0.41 and 0.49 for the validation set, which is reasonable compared to experimental deviations observed in literature for a single surfactant in Table 2 and, as indicated in previous section, they fulfilled all the Chirico criteria.

<< TABLE 4 >>

<< TABLE 5 >>

When looking at the descriptors included in the models, they are related to four main structural trends: the alkyl chain length, the size of the polar head, the presence of a S linkage and the presence of a double bond in the alkyl chain.

In addition to the molecular weight of the alkyl chain (in Eq. 6), two other descriptors are related to the length of the alkyl chain. The Randic index of order 2 in the alkyl chain ${}^{2}\chi_{c}$ and the information content descriptors ${}^{2}CIC_{c}$ are two topological descriptors of the alkyl chain that relate to the size of this fragment. Besides, ${}^{2}\chi_{c}$ has been also evidenced to increase with alkyl chain length in previous work (Gaudin et al., 2016). In all cases, these descriptors present the largest t-test. This is in agreement with experimental knowledge, since increasing alkyl chain length has been identified as the main factor increasing the efficiency of surfactants due to a higher affinity of the surfactant for interfaces (Rosen, 1974).

Four of the QSPR models include descriptors related to the polar head size. The first ones are the number of rings in the whole molecule, and in the polar head only, $n_{rings,h}$. In practice, they correspond to a unique descriptor since none of the alkyl chains of the investigated surfactants contain rings. This descriptor is directly linked to the number of sugar residues. Indeed, in the data set, surfactants with more than one sugar residue contain at least one ring in its polar head and all acyclic molecules only contain one sugar residue. So, no ring in the polar heads is equivalent to a small polar head, and the more the number of rings is large, the more the polar head is large. The second ones are Structural Information Content descriptors, ${}^{0}SIC_{h}$ and ${}^{1}SIC_{h}$, two topological indices that also increase with polar head size, as shown in Figure 6. In all cases, the regression coefficients associated to these descriptors are negative, which means that the size of the polar head decreases the efficiency of sugar-based surfactants. This trend is again meaningful: because of the low affinity of alkyl chains with water (the so-called hydrophobic effect (Tanford, 1979)), the more alkyl chains are adsorbed at the surface, the more the surface tension decreases (Rosen, 1974). At equal chain length and with bigger polar heads, a lower amount of alkyl chains per unit surface is adsorbed at saturation due to sterical hindrance between polar heads. Thus, at a given concentration, the associated decrease of energy per unit surface (or surface tension decrease) is

expected to be lower with bigger polar heads, which implies a higher concentration of molecules to achieve a 20 mN/m surface tension decrease, i. e. a lower pC_{20}.

<< FIGURE 6 >>

Then, the (relative) number of S atoms (n_S and $n_{S,rel}$) appears in four of the developed models. In the dataset, the sulfur is always found in the linkage. Sulfur linkage has been evidenced to add an hydrophobic contribution in surfactants (in addition to the carbon chain) either based on theoretical calculations, in a previous study (Gaudin et al., 2016), and by experimentalists (Marchant, Anderson, & Zhu, 2005). In these new models, these descriptors always positively contribute to the calculated efficiency, which is in line with the mentioned experiments and calculations suggesting that sulfur linkage increases efficiency.

The last trend encountered into the models is the presence of a double-bond in the alkyl chain. Indeed, $n_{single,rel,c}$ is lower (minor to 1) for surfactants containing one double-bond in the alkyl chain. In Eq. 6, $n_{single,rel,c}$ has a positive regression coefficient which is relevant. Indeed, as experimentally observed (Myers, 2006; Rosen & Kunjappu, 2012), surfactants with unsaturated alkyl chains have lower efficiency due to the increased hydrophilicity of their chain.

Finally, it can be noticed that the physical factors reflected by the descriptors included in these models for pC_{20} are the same than those found for CMC in previous work (Gaudin et al., 2016). This is in accordance with the observed correlation between pC_{20} and CMC as evidenced for the 24 common molecules of both validation sets of the present article and the previous one for CMC (in Figure 7).

When comparing the different models developed in this study, it is interesting to note that simple models only based on constitutional descriptors performed as well as, and even slightly better, than models including quantum-chemical descriptors, with $RMSE_{IN}$ of 0.45 vs. 0.49 (log unit) for integral descriptors and 0.43 vs. 0.44 (log unit) for fragment descriptors. Thus, the only 2D structure of sugar-based surfactants seems sufficient to access reliable pC_{20} estimates.

Finally, among the developed models, the simple fragment-based constitutional model (Eq. 6) represents the most relevant one. Its performances are very high with an error of only 0.43 (log unit) (in terms of $RMSE_{IN}$), which is reasonable compared to the observed deviations in experimental data found in literature (see Table 2). Moreover, it is very easy to use, requiring knowledge of the only 2D-structure of surfactants.

This model even demonstrated similar deviations in prediction than the relationship between pC_{20} and CMC, as expressed in Eq. 7, according to the Abbott criterion (Abbott, 2016) considering that the concentration to decrease the surface tension of water by 20mN/m (*i.e.* C_{20}) is ten times lower than CMC, which was highlighted as relevant for sugar-based surfactants as illustrated in Figure 7.

$$pC_{20} = -\log CMC + 1 \quad (7)$$

<< FIGURE 7 >>

Indeed, when applied to the validation set molecules, Eq. 7 did not offer more reliable estimates than the new QSPR model with a mean absolute error of 0.30 (log unit) vs. 0.31 (log unit) for the QSPR model (as shown in Table 6). This enforces the interest of the new QSPR model that can be applied without knowledge of any experimental data and can then be applied in an in-silico design or screening strategy before any experimental characterization and even synthesis.

<< TABLE 6 >>

5 Conclusion

This paper presents the first QSPR models dedicated to the prediction of efficiency of surfactants. Based on different types of descriptors, a series of new QSPR models were developed and validated focusing on sugar-based surfactants in the perspective of their use to guide the design of bio-based formulations in substitution to petroleum-based surfactants. The final proposed model is very simple, based only on constitutional descriptors of the polar head and of the alkyl chain of surfactants. A good predictive power was highlighted with a $RMSE_{IN}$ of 0.43 (log unit) evaluated on an external validation set. This simple fragment-based approach is very easy to use based on the only knowledge of the 2D structure of the

surfactants and represents a powerful alternative to the systematic experimental campaigns in particular at early R&D stages as they allow to access to efficiency estimates of new surfactants even before synthesis. This enables to identify the surfactants that could be used for target applications at the lowest concentrations. To the end, the structure-efficiency trends encountered in these models can guide the design of new high potential surfactant's structures. Indeed, a similarity was found between the molecular factors that affect the critical micelle concentration and efficiency. 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 higher efficiencies.

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).

References

Abbott, S. CMC and Γ. Retrieved 24/11/17, 2016, from www.stevenabbott.co.uk/practical-surfactants/cmc.php

Abbott, S. (2016). *Surfactant Science: Principles and Practice*.

Agrawal, V. K., & Khadikar, P. V. (2001). QSAR prediction of toxicity of nitrobenzenes. *Bioorganic & Medicinal Chemistry, 9*(11), 3035-3040.

Anbu, S. (2017). Procuring sugar surfactants is a cleaner and greener alternative to synthetic surfactants. *Beroe - Industries: Chemical, Personal Products*. https://www.beroeinc.com/article/sugar-surfactants/

Aveyard, R., Binks, B. P., Chen, J., Esquena, J., & Fletcher, P. D. I. (1998). Surface and Colloid Chemistry of Systems Containing Pure Sugar Surfactant. *Langmuir, 14*, 4699-4709.

Boullanger, P., & Chevalier, Y. (1996). Surface Active Properties and Micellar Aggregation of Alkyl 2-Amino-2-deoxy-β-d-glucopyranosides. *Langmuir, 12*(7), 1771-1776. doi: 10.1021/la950485i

Boyère, C., Broze, G., Blecker, C., Jérôme, C., & Debuigne, A. (2013). Monocatenary, branched, double-headed, and bolaform surface active carbohydrate esters via photochemical thiol-ene/-yne reactions. *Carbohydrate Research, 380*, 29-36. doi: http://dx.doi.org/10.1016/j.carres.2013.07.003

Burczyk, B., Wilk, K. A., Sokołowski, A., & Syper, L. (2001). Synthesis and Surface Properties of N-Alkyl-N-methylgluconamides and N-Alkyl-N-methyllactobionamides. *Journal of Colloid and Interface Science, 240*(2), 552-558. doi: http://dx.doi.org/10.1006/jcis.2001.7704

Chirico, N., & Gramatica, P. (2012). Real External Predictivity of QSAR models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection. *Journal of Chemical Information and Modeling, 52*(8), 2044-2058.

Consonni, V., Ballabio, D., & Todeschini, R. (2009). Comments on the Definition of the Q^2 Parameter for QSAR Validation. *Journal of Chemical Information and Modeling, 49*(7), 1669-1678.

Creton, B. (2013). Prediction of Surfactants' Properties using Multiscale Molecular Modeling Tools: A Review. *Oil. Gas. Sci. Technol. Rev. IFPEN*, 1-14.

Crook, E. H., Fordyce, D. B., & Trebbi, G. F. (1963). Molecular Weight Distribution of Nonionic Surfactants. I. Surface and Interfacial Tension of Normal Distribution and Homogeneous p,t-Octylphenoxyethoxyethanols (OPE's). *Journal of Physical Chemistry, 67*(10), 1987.

du Noüy, P. L. (1925). An interfacial tensiometer for universal use. *The Journal of General Physiology, 7*(5), 625-631. doi: 10.1085/jgp.7.5.625

Ericsson, C. A., Söderman, O., Garamus, V. M., Bergström, M., & Ulvenlund, S. (2004). Effects of Temperature, Salt, and Deuterium Oxide on the Self-Aggregation of Alkylglycosides in Dilute Solution. 1. n-Nonyl-β-d-glucoside. *Langmuir, 20*(4), 1401-1408. doi: 10.1021/la035613e

Fayet, G., Rotureau, P., Joubert, L., & Adamo, C. (2010a). Development of a QSPR model for predicting thermal stabilities of nitroaromatic compounds taking into account their decomposition mechanisms. *Journal of Molecular Modeling, 17*(10), 2443-2453. doi: 10.1007/s00894-010-0908-0

Fayet, G., Rotureau, P., Joubert, L., & Adamo, C. (2010b). QSPR modeling of thermal stability of nitroaromatic compounds: DFT vs. AM1 calculated descriptors. *Journal of Molecular Modeling, 16*(4), 805-812. doi: 10.1007/s00894-009-0634-7

Ferrer, M., Comelles, F., Plou, F. J., Cruces, M. A., Fuentes, G., Parra, J. L., & Ballesteros, A. (2002). Comparative Surface Activities of Di- and Trisaccharide Fatty Acid Esters. *Langmuir, 18*(3), 667-673. doi: 10.1021/la010727g

Fisher, G. S., Zeringue, H. J., & Feuge, R. O. (1977). Surface activity of sucrose palmitates. *Journal of the American Oil Chemists' Society, 54*(2), 59-61. doi: 10.1007/bf02912390

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., . . . Fox, D. J. (2009). Gaussian 09, Revision B.01. Wallingford CT.

Gaudin, T. (2016). *Développement de modèles QSPR pour la prédiction et la compréhension des propriétés amphiphiles des tensioactifs dérivés de sucre.* (PhD), Ph. D. thesis, Université de Technologie de Compiègne.

Gaudin, T., Rotureau, P., Pezron, I., & Fayet, G. (2016). New QSPR Models to Predict the Critical Micelle Concentration of Sugar-Based Surfactants. *Industrial & Engineering Chemistry Research, 55*(45), 11716-11726. doi: 10.1021/acs.iecr.6b02890

Gaudin, T., Rotureau, P., Pezron, I., & Fayet, G. (2017). Conformations of n-alkyl-α/β-d-glucopyranoside surfactants: Impact on molecular properties. *Computational and Theoretical Chemistry, 1101*, 20-29. doi: http://dx.doi.org/10.1016/j.comptc.2016.12.020

Gaudin, T., Rotureau, P., Pezron, I., & Fayet, G. (2018). Investigating the impact of sugar based surfactants structure on surface tension at critical micelle concentration with structure-property relationships *Journal of Colloid and Interface Science, 516*, 162-171.

Hill, K., & LeHen-Ferrenbach, C. (2009). 1. Sugar-Based Surfactants for Consumer Products and Technical Applications. In C. C. Ruiz (Ed.), *Sugar-Based Surfactants: Fundamentals and Applications*: CRC Press, Taylor & Francis Group.

Hu, J. (2010). A Review on Progress in QSPR Studies for Surfactants. *International Journal of Molecular Sciences, 11*, 1020-1047.

Katritzky, A. R., Kuanar, M., Slavov, S., Hall, C. D., Karelson, M., Kahn, I., & Dobchev, D. A. (2010). Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chemical Reviews, 110*(10), 5714-5789.

Kjellin, M., & Johansson, I. (2010). *Surfactants from Renewable Resources* (1 ed.): John Wiley & Sons, Ltd.

Lalot, J., Stasik, I., Demailly, G., Beaupère, D., & Godé, P. (2004). Synthesis and amphiphilic properties of S-alkylthiopentonolactones and their pentitol derivatives. *Journal of Colloid and Interface Science, 273*(2), 604-610. doi: http://dx.doi.org/10.1016/j.jcis.2004.01.020

Le Neindre, B. (1993). Tensions superficielles et interfaciales. *Techniques de l'ingénieur*.

Lindgren, F., Hansen, B., Karcher, W., Sjöström, M., & Eriksson, L. (1996). Model validation by permutation tests: Applications to variable selection. *Journal of Chemometrics, 10*(5-6), 521-532.

Mahmood, M. E., & Al-Koofee, D. A. F. (2013). Effect of Temperature Changes on Critical Micelle Concentration for Tween Series Surfactant. *Global Journal of Science Frontier Research Chemistry, 13*(4), 1-7.

Marchant, R. E., Anderson, E. H., & Zhu, J. (2005). Polysaccharide Surfactants: Structure, Synthesis, and Surface-Active Properties. In S. Dimitriu (Ed.), *Polysaccharides, Structural Diversity and Functional Versatility* (2nd ed., pp. 1055-1086): Marcel Dekker.

Matsumura, S., Imai, K., Yoshikawa, S., Kawada, K., & Uchibor, T. (1990). Surface activities, biodegradability and antimicrobial properties of n-alkyl glucosides, mannosides and galactosides. *Journal of the American Oil Chemists' Society, 67*(12), 996-1001. doi: 10.1007/BF02541865

Milkereit, G., Garamus, V. M., Veermans, K., Willumeit, R., & Vill, V. (2005). Structures of micelles formed by synthetic alkyl glycosides with unsaturated alkyl chains. *Journal of Colloid and Interface Science, 284*(2), 704-713. doi: http://dx.doi.org/10.1016/j.jcis.2004.10.039

Minamikawa, H., & Hato, M. (2005). Headgroup effects on phase behavior and interfacial properties of β-3,7-dimethyloctylglycoside/water systems. *Chemistry and Physics of Lipids, 134*(2), 151-160.

Mulliken, R. S. (1955). Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I. *The Journal of Chemical Physics, 23*(10), 1833-1840. doi: doi:http://dx.doi.org/10.1063/1.1740588

Myers, D. (2006). *Surfactant Science and Technology* (I. John Wiley & Sons Ed. 3rd ed.): Wiley-Interscience.

Nieto-Draghi, C., Fayet, G., Creton, B., Rozanska, X., Rotureau, P., De Hemptinne, J.-C., . . . Adamo, C. (2015). A General Guidebook for the Theoretical Prediction of Physico-Chemical Properties of Chemicals for Regulatory Purposes. *Chemical Reviews, 115*(24), 13093-13164.

OECD. (2007). *Guidance Document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models*. (69). Organisation for Economic Co-operation and Development (OECD).

Ojha, P. K., Mitra, I., Das, R. N., & Roy, K. (2011). Further exploring rm2 metrics for validation of QSPR models. *Chemometrics and Intelligent Laboratory Systems, 107*(1), 194-205. doi: http://dx.doi.org/10.1016/j.chemolab.2011.03.011

Persson, C. M., Kjellin, U. R. M., & Eriksson, J. C. (2003). Surface Pressure Effect of Poly(ethylene oxide) and Sugar Headgroups in Liquid-Expanded Monolayers. *Langmuir, 19*(20), 8152-8160. doi: 10.1021/la026943m

Piao, J., Kishi, S., & Adachi, S. (2006). Surface tensions of aqueous solutions of 1-O-monoacyl sugar alcohols. *Colloids and Surfaces A: Physicochemical and Engineering Aspects, 277*(1–3), 15-19. doi: http://dx.doi.org/10.1016/j.colsurfa.2005.10.053

Piasecki, A., & Pilakowska-Pietras, D. (2007). Synthesis and Properties of Functionalized Alkylaldonamides. *Journal of Surfactants and Detergents, 10*, 125-130.

Plusquellec, D., Brenner-Hénaff, C., Léon-Ruaud, P., Duquenoy, S., Lefeuvre, M., & Wróblewski, H. (1994). An Efficient Acylation of Free Glycosylamines for the Synthesis of N-Glycosyl Amino Acids and N-Glycosidic Surfactants for Membrane Studies. *Journal of Carbohydrate Chemistry, 13*(5), 737-751.

Rojas, O. J., Stubenrauch, C., Lucia, L. A., & Habibi, Y. (2009). Interfacial Properties of Sugar-Based Surfactants. In D. Hayes, D. Kitamoto, D. Solaiman & R. Ashby (Eds.), *Biobased Surfactants and Detergents: Synthesis, Properties, and Applications* (pp. 457-480). Urbana: AOCS Press.

Rosen, M. J. (1974). Relationship of Structure to Properties in Surfactants: II. Efficiency in Surface or Interfacial Tension Reduction. *Journal of the American Oil Chemists' Society, 51*, 461-465.

Rosen, M. J. (1976). The relationship of structure to properties in surfactants. IV. Effectiveness in surface or interfacial tension reduction. *Journal of Colloid and Interface Science, 56*(2), 320-327. doi: http://dx.doi.org/10.1016/0021-9797(76)90257-5

Rosen, M. J., & Kunjappu, J. T. (2012). *Surfactants and Interfacial Phenomena* (I. John Wiley & Sons Ed. 4th ed.): John Wiley & Sons, Inc.

Roy, K., Mitra, I., Kar, S., Ojha, P. K., Das, R. N., & Kabir, H. (2012). Comparative Studies on Some Metrics for External Validation of QSPR Models. *Journal of Chemical Information and Modeling, 52*(2), 396-408.

Rücker, C., Rücker, G., & Meringer, M. (2007). Y-Randomization and Its Variants in QSPR/QSAR. *Journal of Chemical Information and Modeling, 47*(6), 2345-2357.

Savelli, M. P., Van Roekeghem, P., Douillet, O., Cavé, G., Godé, P., Ronco, G., & Villa, P. (1999). 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). *International Journal of Pharmaceutics, 182*(2), 221-236. doi: http://dx.doi.org/10.1016/S0378-5173(99)00078-2

Schüürmann, G., Ebert, R. U., Chen, J., Wang, B., & Kühne, R. (2008). External Validation and Prediction Employing the Predictive Squared Correlation Coefficient — Test Set Activity Mean vs Training Set Activity Mean. *Journal of Chemical Information and Modeling, 48*(11), 2140-2145.

Shinoda, K., Yamanaka, T., & Kinoshita, K. (1959). Surface Chemical Properties in Aqueous Solutions of Non-ionic Surfactants Octyl Glycol Ether, α-Octyl Glyceryl Ether and Octyl Glucoside. *The Journal of Physical Chemistry, 63*(5), 648-650. doi: 10.1021/j150575a003

Silva, F. V. M., Goulart, M., Justino, J., Neves, A., Santos, F., Caio, J., . . . Rauter, A. P. (2008). Alkyl deoxy-arabino-hexopyranosides: Synthesis, surface properties, and biological activities. *Bioorganic & Medicinal Chemistry, 16*, 4083-4092.

Söderberg, I., Drummond, C. J., Neil Furlong, D., Godkin, S., & Matthews, B. (1995). Non-ionic sugar-based surfactants: Self assembly and air/water interfacial activity. *Colloids and Surfaces A: Physicochemical and Engineering Aspects, 102*, 91-97. doi: http://dx.doi.org/10.1016/0927-7757(95)03250-H

Syper, L., Wilk, K. A., Sokołowski, A., & Burczyk, B. (1998). 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* (Vol. 110, pp. 199-203): Steinkopff

Tanford, C. (1979). Interfacial free energy and the hydrophobic effect. *Proc. Natl. Acad. Sci., 76*(9), 4175-4176.

Tropsha, A., Gramatica, P., & Gombar, K. V. (2003). The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR & Combinatorial Science, 22*(1), 69-77.

Waltermo, Å., Claesson, P. M., & Johansson, I. (1996). Alkyl Glucosides on Hydrophobic Surfaces Studied by Surface Force and Wetting Measurements. *Journal of Colloid and Interface Science, 183*(2), 506-514. doi: http://dx.doi.org/10.1006/jcis.1996.0574

Wilk, K., Syper, L., Burczyk, B., Maliszewska, I., Jon, M., & Domagalska, B. (2001). Preparation and properties of new lactose-based surfactants. *Journal of Surfactants and Detergents, 4*(2), 155-161. doi: 10.1007/s11743-001-0169-1

Zefirov, N. S., Palyulin, V. A., Oliferenko, A. A., Ivanova, A. A., & Ivanov, A. A. (2001). Method for the Construction of Universal Structure–Property Relationship Models using the Example of Normal Boiling Temperature for a Wide Set of Organic Compounds. *Doklady Chemistry, 381*(4), 356-358.

Zhang, T., & Marchant, R. E. (1996). Novel Polysaccharide Surfactants: The Effect of Hydrophobic and Hydrophilic Chain Length on Surface Active Properties. *Journal of Colloid and Interface Science, 177*(2), 419-426. doi: http://dx.doi.org/10.1006/jcis.1996.0054

Zhu, Y.-P., Rosen, M. J., Vinson, P. K., & Morrall, S. W. (1999). Surface Properties of N-Alkanoyl-N-methyl Glucamines and Related Materials. *Journal of Surfactants and Detergents, 2*(3), 357-362.

Table 1. Experimental and calculated pC_{20} values from and the new QSPR models

surfactant	exp pC_{20}	T (°C)	ref	predicted pC_{20}					
				eq. 1	eq. 2	eq. 3	eq. 4	eq. 5	eq. 6
TRAINING SET									
Octanoyl-β-D-Glucosylamine	1.70	25	(Plusquellec et al., 1994)[a]	2.35	2.20	2.55	2.18	2.63	2.57
1-O-Pentanoyl-D,L-Xylitol	1.90	25	(Savelli et al., 1999)	1.81	1.57	1.68	1.53	1.63	1.63
Octanoyl-β-D-Galactosylamine	2.00	25	(Plusquellec et al., 1994)[a]	2.32	2.20	2.55	2.58	2.63	2.57
S-Pentyl-1-Thio-D,L-Xylitol	2.20	25	(Savelli et al., 1999)	2.95	2.79	3.02	2.24	3.03	3.21
Hexyl-D-Maltonamide	2.50	25	(Zhang & Marchant, 1996)	1.86	2.18	2.19	2.10	1.94	2.19
1-O-Octanoyl-D,L-Xylitol	2.80	25	(Savelli et al., 1999)	2.98	3.01	3.00	2.93	2.73	2.77
1-O-Heptyl-D,L-Xylitol	2.90	25	(Savelli et al., 1999)	2.96	2.87	2.83	2.42	2.86	2.77
3,7-Dimethyloctyl-β-D-Maltotrioside	3.00	25	(Minamikawa & Hato, 2005)[a]	3.39	3.36	3.32	3.20	3.13	3.31
N-Decanoyl-N-Methyl Lactitolamine	3.06	25	(Wilk et al., 2001)	3.85	3.67	3.86	3.11	3.21	3.33
Octyl-D-Maltonamide	3.10	25	(Zhang & Marchant, 1996)	2.58	2.90	2.91	2.72	2.84	2.95
Octyl-D,L-Glycerol	3.19	25	(Shinoda et al., 1959)[a]	3.10	3.52	3.52	3.37	3.39	3.15
1-O-Heptanoyl-D,L-Xylitol	3.20	25	(Savelli et al., 1999)	2.54	2.52	2.61	3.12	2.27	2.39
S-Hexyl 1-Thio-D-Lyxitol	3.25	20	(Lalot, Stasik, Demailly, Beaupère, & Godé, 2004)	3.45	3.37	3.53	3.65	3.46	3.59
N-Decyl-N-Hydroxymethyl Gluconamide	3.30	22	(Piasecki & Pilakowska-Pietras, 2007)[a]	4.18	3.95	3.93	3.81	3.86	3.91
Nonyl-β-D-Glucoside	3.30	20	(Ericsson, Söderman, Garamus, Bergström, & Ulvenlund, 2004)[a]	3.44	3.23	3.18	3.35	3.65	3.33
3,7-Dimethyloctyl-β-D-Maltoside	3.31	25	(Minamikawa & Hato, 2005)[a]	3.17	3.06	3.28	3.45	3.35	3.51
S-Hexyl 5-Thio-D-Arabinonolactone	3.36	20	(Lalot et al., 2004)	3.50	3.46	3.62	3.79	3.44	3.66
6-O-Dodecanoylstachyose	3.40	20	(Söderberg, Drummond, Neil Furlong, Godkin, & Matthews, 1995)[a]	3.88	4.01	3.76	3.47	3.60	3.49
N-Octyl Glucoheptonamide	3.40	25	(Syper, Wilk, Sokołowski, & Burczyk, 1998)[a]	2.80	2.91	2.86	2.95	3.08	3.15
1-O-Nonyl-D,L-Xylitol	3.40	25	(Savelli et al., 1999)	3.81	3.85	3.61	3.42	3.72	3.53
N-Decyl-N-Hydroxymethyl Glucoheptonamide	3.50	22	(Piasecki & Pilakowska-Pietras, 2007)[a]	3.69	4.05	3.93	3.73	3.82	3.91
Decyl-D-Maltonamide	3.60	25	(Zhang & Marchant, 1996)	3.21	3.59	3.55	3.59	3.66	3.71
6-O-Dodecanoylraffinose	3.60	20	(Söderberg et al., 1995)[a]	3.78	3.85	3.66	3.69	3.75	3.69
S-Hexyl 1-Thio-L-Xylitol	3.66	20	(Lalot et al., 2004)	3.44	3.37	3.53	3.69	3.46	3.59
Decyl-β-D-Maltoside	3.70	25	(Aveyard, Binks, Chen, Esquena, & Fletcher, 1998)[a]	3.89	3.60	3.28	3.78	3.81	3.51
Decyl-D-Lactobionamide	3.73	25	(Syper et al., 1998)[a]	3.70	3.59	3.55	3.57	3.66	3.71

Undecanoyl-N-Methylglucamine	3.82	25	(Zhu et al., 1999)	4.13	4.07	4.24	4.17	3.89	3.91
1-O-Decanoyl-D,L-Xylitol	3.90	25	(Savelli et al., 1999)	3.79	3.94	3.71	3.92	3.59	3.53
1-O-Decyl-D,L-Xylitol	4.00	25	(Savelli et al., 1999)	4.19	4.30	3.96	4.10	4.12	3.91
Dodecyl-D-Lactobionamide	4.20	25	(Syper et al., 1998)[a]	4.29	4.24	4.14	4.35	4.42	4.47
6-O-Dodecanoylsucrose	4.30	20	(Söderberg et al., 1995)[a]	3.82	3.82	3.67	4.27	4.05	3.89
6-O-Dodecanoylglucose	4.40	20	(Söderberg et al., 1995)[a]	4.28	4.19	3.93	4.70	4.32	4.09
Dodecanoyl-N-Methylglucamine	4.40	25	(Zhu et al., 1999)	4.51	4.46	4.55	4.35	4.28	4.29
Dodecanoyl-N-Methylxylamine	4.43	25	(Zhu et al., 1999)	4.35	4.41	4.60	4.43	4.32	4.29
N-Dodecyl-N-Methyl Lactobionamide	4.45	20	(Burczyk, Wilk, Sokołowski, & Syper, 2001)	4.50	4.35	4.44	4.42	4.39	4.47
Dodecanoyl-N-Methylglyceramine	4.70	25	(Zhu et al., 1999)	4.38	4.61	4.81	4.63	4.43	4.29
Dodecyl-β-D-Maltoside	4.70	22	(Persson, Kjellin, & Eriksson, 2003)[a]	4.46	4.20	3.87	4.56	4.57	4.27
N-Dodecyl-N-Methyl Gluconamide	4.78	20	(Burczyk et al., 2001)	4.85	4.60	4.55	4.71	4.67	4.67
[N-(Oleoyl)-2 -Ethylamino]-β-D-Maltoside	4.80	25	(Milkereit, Garamus, Veermans, Willumeit, & Vill, 2005)[a]	5.04	5.04	5.57	4.98	4.82	4.92
N-Dodecyl-N-Methyl Glucoheptonamide	4.90	25	(Syper et al., 1998)[a]	4.89	4.66	4.54	4.66	4.63	4.67
Tridecanoyl-N-Methylglucamine	5.00	25	(Zhu et al., 1999)	4.85	4.83	4.85	4.95	5.01	5.05
S-Octyl 1-Thio-D-Lyxitol	5.15	20	(Lalot et al., 2004)	4.41	4.46	4.41	4.28	4.36	4.35
N-Oleyl-N-Methyl Gluconamide	5.37	20	(Burczyk et al., 2001)	5.79	5.82	6.30	5.48	5.56	5.78
S-Decyl 5-Thio-D-Arabinonolactone	5.40	20	(Lalot et al., 2004)	5.35	5.46	5.08	5.28	5.16	5.18
S-Decyl 1-Thio-L-Xylitol	5.40	20	(Lalot et al., 2004)	5.23	5.41	5.17	5.17	5.19	5.11
N-Tetradecyl-N-Methyl Lactobionamide	5.40	20	(Burczyk et al., 2001)	5.02	4.97	5.01	5.16	5.11	5.23
Oleyl-β-D-Maltoside	5.40	25	(Milkereit et al., 2005)[a]	5.37	5.17	5.48	5.32	5.45	5.38
N-Hexadecanoyl-N-Methyl Lactitolamine	5.46	25	(Wilk et al., 2001)	5.41	5.43	5.51	5.52	5.43	5.61
N-Tetradecyl-N-Methyl Gluconamide	5.55	20	(Burczyk et al., 2001)	5.55	5.33	5.15	5.45	5.39	5.43
Oleyl-β-D-Maltotrioside	5.70	25	(Milkereit et al., 2005)[a]	5.11	5.12	5.37	5.07	5.22	5.18
N-Octadecanoyl-N-Methyl Lactitolamine	5.95	25	(Wilk et al., 2001)	5.92	5.96	6.01	6.24	6.09	6.36
N-Hexadecyl-N-Methyl Gluconamide	6.11	20	(Burczyk et al., 2001)	6.15	6.00	5.70	6.01	6.06	6.18
6-O-[(Tetradecyl-3-Propylsulfide)ethanoyl]-D-Mannose	6.13	25	(Boyère et al., 2013)[a]	6.01	6.46	6.27	6.16	6.41	6.03
6-O-[(Hexyloctyl)-3-Propylsulfide)ethanoyl]-D-Mannose	6.36	25	(Boyère et al., 2013)[a]	5.56	6.14	6.27	6.16	6.41	6.03
N-Octadecyl-N-Methyl Gluconamide	6.46	20	(Burczyk et al., 2001)	6.83	6.59	6.19	6.73	6.72	6.94
VALIDATION SET									
S-Butyl-1-Thio-D,L-Xylitol	1.80	25	(Savelli et al., 1999)	2.60	2.32[b]	2.49[b]	2.09	2.82	2.83

1-O-Hexanoyl-D,L-Xylitol	2.10	25	(Savelli et al., 1999)	2.14	1.99	2.15	1.68[b]	1.84	2.01
2-Amino-2-Deoxy-Octyl-β-D-Glucoside	2.70	25	(Boullanger & Chevalier, 1996)[a]	3.10	2.15	2.77	3.31	3.14	2.95
Octyl-β-D-Glucoside	3.00	25	(Matsumura et al., 1990)[a]	3.03	2.80	2.85	3.17	3.23	2.95
Octyl-α-D-Glucoside	3.10	25	(Matsumura et al., 1990)[a]	2.83	2.80	2.85	3.19	3.23	2.95
Octyl Glycol	3.19	25	(Shinoda et al., 1959)[a]	2.80[b]	3.99[b]	3.74[b]	3.52[b]	3.46	3.15
N-Decyl-N-Methyl Lactobionamide	3.29	20	(Burczyk et al., 2001)	3.93	3.72	3.86	3.63	3.63	3.71
1-O-Nonanoyl-D,L-Xylitol	3.30	25	(Savelli et al., 1999)	3.40	3.50	3.39	3.75	3.17	3.15
S-Hexyl 5-Thio-D-Xylonolactone	3.38	20	(Lalot et al., 2004)	3.56	3.46	3.62	3.80[b]	3.44	3.66
1-O-Octyl-D,L-Xylitol	3.40	25	(Savelli et al., 1999)	3.37	3.38	3.24	3.24	3.30	3.15
3,7-Dimethyloctyl-β-D-Glucoside	3.50	25	(Minamikawa & Hato, 2005)[a]	2.77	2.88	3.51	3.71	3.60	3.71
N-Decyl-N-Methyl Gluconamide	3.60	25	(Burczyk et al., 2001)	4.16	3.79	3.93	3.93	3.91	3.91
S-Hexyl 1-Thio-L-Ribitol	3.70	20	(Lalot et al., 2004)	3.27	3.37	3.53	3.67	3.46	3.59
Decyl-β-D-Glucoside	3.90	25	(Aveyard et al., 1998)[a]	3.81	3.65	3.51	4.03	4.06	3.71
N-Dodecanoyl-N-Methyl Lactitolamine	4.02	25	(Wilk et al., 2001)	4.33	4.28	4.44	3.98	4.00	4.09
S-Octyl 5-Thio-D-Xylonolactone	4.37	20	(Lalot et al., 2004)	4.51	4.53	4.41	4.42[b]	4.33	4.42
Dodecyl-D-Maltonamide	4.40	25	(Zhang & Marchant, 1996)	3.81	4.24	4.14	4.37	4.42	4.47
1-O-Undecyl-D,L-Xylitol	4.50	25	(Savelli et al., 1999)	4.56	4.70	4.27	4.28	4.51	4.29
S-Octyl 5-Thio-D-Arabinonolactone	4.74	20	(Lalot et al., 2004)	4.46	4.53	4.41	4.41	4.33	4.42
N-Tetradecanoyl-N-Methyl Lactitolamine	4.87	25	(Wilk et al., 2001)	4.71	4.89	5.01	4.77	4.73	4.85
N-Oleyl-N-Methyl Lactobionamide	5.09	20	(Burczyk et al., 2001)	5.13	5.42	6.04	5.18	5.27	5.58
S-Decyl 5-Thio-D-Xylonolactone	5.40	20	(Lalot et al., 2004)	5.41	5.46	5.08	5.28[b]	5.16	5.18
S-Octyl 1-Thio-L-Xylitol	5.40	20	(Lalot et al., 2004)	4.35	4.46	4.41	4.31	4.36	4.35
Tetradecanoyl-N-Methylglucamine	5.40	25	(Zhu et al., 1999)	5.18	5.19	5.15	5.14	5.36	5.43
S-Octyl 1-Thio-L-Ribitol	5.60	20	(Lalot et al., 2004)	4.23	4.46	4.41	4.29	4.36	4.35
N-Hexadecyl-N-Methyl Lactobionamide	6.01	20	(Burczyk et al., 2001)	5.51	5.52	5.51	5.72	5.78	5.99
N-Octadecyl-N-Methyl Lactobionamide	6.34	20	(Burczyk et al., 2001)	5.98	6.05	6.01	6.44	6.44	6.74[b]

[a] data extracted from graphs; [b] out of the applicability domain of the model

Table 2. Experimental values of pC_{20} gathered from different sources.

surfactant	C_{20} (mM)	pC_{20} (M)	reference
octyl-β-D-glucoside	2.0*	2.7	(Waltermo, Claesson, & Johansson, 1996)
	2.5*	2.6	(Shinoda et al., 1959)
	3.0	2.5	(Silva et al., 2008)
N-dodecyl-N-methyl lactobionamide	0.035	4.5	(Burczyk et al., 2001)
	0.049*	4.3	(Syper et al., 1998)
6-O-dodecanoylsucrose	0.032	4.5	(Ferrer et al., 2002)
	0.082*	4.1	(Söderberg et al., 1995)
1-O-octanoyl-D,L-xylitol	0.34*	3.5	(Piao, Kishi, & Adachi, 2006)
	1.6	2.8	(Savelli et al., 1999)

*data extracted from graphs

Table 3. External validation metrics and thresholds according to Chirico et al. (Chirico & Gramatica, 2012)

validation metric	Ref.	Chirico criteria
$R^2_{EXT} = 1 - \frac{\sum_n (y_i - y_i)^2}{\sum_n (y_i - \bar{y})^2}$	(OECD, 2007; Schüürmann, Ebert, Chen, Wang, & Kühne, 2008)	$R^2_{EXT} > 0.70$
$Q^2_{F1} = 1 - \frac{\sum_n (y_i - y_i)^2}{\sum_n (y_i - \bar{y}_{TR})^2}$	(Tropsha, Gramatica, & Gombar, 2003)	$Q^2_{F1} > 0.70$
$Q^2_{F2} = 1 - \frac{\sum_n (y_i - y_i)^2}{\sum_n (y_i - \bar{y}_{EXT})^2}$	(Schüürmann et al., 2008)	$Q^2_{F2} > 0.70$
$Q^2_{F3} = 1 - \frac{\left[\sum_n (y_i - y_i)^2\right] / n_{EXT}}{\left[\sum_n (y_i - \bar{y}_{TR})^2\right] / n_{TR}}$	(Consonni, Ballabio, & Todeschini, 2009)	$Q^2_{F3} > 0.70$
$CCC = \frac{2\sum_n (y_i - \bar{y})(y_i - \bar{\hat{y}})}{\sum_n (y_i - \bar{y})^2 + \sum_n (y_i - \bar{\hat{y}})^2 + n(\bar{y} - \bar{\hat{y}})^2}$	(Chirico & Gramatica, 2012)	$CCC > 0.85$
$r_m^2 = R^2_{EXT}\left(1 - \sqrt{R^2_{EXT} - R^2_0}\right)$ $\overline{r^2_m} = \frac{r_m^2 + r'^2_m}{2}$ $\Delta r^2_m = \lvert r_m^2 - r'^2_m \rvert$	(Ojha, Mitra, Das, & Roy, 2011; Roy et al., 2012)	$\overline{r^2_m} > 0.65$ $\Delta r^2_m < 0.20$

n_i is the number of molecules; y_i and $\hat{y}_i$ are the experimental and calculated properties for the molecule i; $\bar{y}$ is the average of experimental properties; $\bar{\hat{y}}$ is the average of calculated properties; R_0^2 is the coefficient of determination forcing the origin for the axis; r_m^2 and r'^2_m are calculated by using experimental data on the ordinate axis and on the abscissa axis, respectively.

Table 4. Summary of the performances of the new QSPR models.

model	n_{desc}	descriptors	R^2	RMSE	R^2_{IN}	$RMSE_{IN}$	n_{out}
integral/all types (Eq. 1)	4	$^2ACIC, HACA_{2,TMSA}, N_{O,avg}, n_{rings}$	0.90	0.37	0.83	0.49	1
integral/constitutional and topological (Eq. 2)	5	n_S, ^2ASIC, $^1\chi^v$, n_{rings}, $n_{C,rel}$	0.92	0.34	0.88	0.41	2
integral/constitutional (Eq. 3)	5	$n_{C,rel}, M_{w,rel}, n_H, n_S, n_{rings}$	0.88	0.41	0.84	0.45	2
fragments/all types (Eq. 4)	5	$FHBSA_h$, $^2CIC_c, RNCS_c$, $^0SIC_h$, $^2ASIC_c$	0.95	0.28	0.87	0.44	5
fragments/constitutional and topological (Eq. 5)	4	$n_{S,h}$, $^2\chi_c$, $^2BSIC_c$, $^1SIC_h$	0.92	0.33	0.87	0.42	0
fragments/constitutional (Eq. 6)	4	$n_{S,rel,h}, M_{w,c}, n_{rings,h}, n_{\mathrm{sin}gle,rel,c}$	0.91	0.35	0.85	0.43	1

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

Table 5 – Applicability domains for the molecular descriptors used in the different models.

Descriptors	AD
Average Complementary Information Content (order 2)	[1.1740;2.9871]
HACA-2/TMSA [Quantum-Chemical PC]	[1.19e-02;4.04e-02]
Avg nucleoph. react. index for a O atom	[1.948e-06;2.010e-02]
Number of rings	[0;4]
Number of S atoms	[0;1]
Average Structural Information content (order 2)	[0.5314;0.7729]
Kier&Hall index (order 1)	[5.3316;20.01470]
Relative number of C atoms	[0.2703;0.3125]
Relative molecular weight	[5.3765;6.9580]
Number of H atoms	[20;66]
FHBSA Fractional HBSA (HBSA/TMSA) [Quantum-Chemical PC] of the polar head	[0.7283;0.9929]
Complementary Information content (order 2) of the alkyl chain	[27.5098;232.8127]
RNCS Relative negative charged SA (SAMNEG*RNCG) [Zefirov's PC] of the alkyl chain	[0.3476;1.7553]
Structural Information content (order 0) of the polar head	[5.1987;22.7905]
Average Structural Information content (order 2) of the alkyl chain	[0.2841;0.5264]
Number of S atoms in the polar head	[0;1]
Randic index (order 2) in the alkyl chain	[1.0000;5.9497]
Average Bonding Information content (order 2) in the alkyl chain	[0.2854;0.5313]
Structural Information content (order 1) in the polar head	[7.9277;38.7452]
Relative number of S atoms in the polar head	[0.0000;0.0556]
Molecular weight of the alkyl chain	[58.1230;254.4983]
Relative number of single bonds in the alkyl chain	[0.9800;1.0000]
Number of rings in the polar head	[0;4]

Table 6. Estimation of pC_{20} based on Abbott's assumption (Eq. 7)

Surfactant	log CMC	pC_{20}			ref
	exp	QSPR (Eq. 6)	Abbott (Eq. 7)	exp	
S-Butyl-1-Thio-D,L-Xylitol	-0.74	2.83	1.74	1.80	(Savelli et al., 1999)
1-O-Hexanoyl-D,L-Xylitol	-1.24	2.01	2.24	2.10	(Savelli et al., 1999)
2-Amino-2-Deoxy-Octyl-β-D-Glucoside	-1.64	2.95	2.64	2.70	(Boullanger & Chevalier, 1996)[a]
Octyl-β-D-Glucoside	-1.70	2.95	2.70	3.00	(Matsumura et al., 1990)[a]
Octyl Glycol	-2.31	3.15	3.31	3.20	(Shinoda et al., 1959)[a]
N-Decyl-N-Methyl Lactobionamide	-2.64	3.71	3.64	3.30	(Burczyk et al., 2001)
1-O-Nonanoyl-D,L-Xylitol	-2.36	3.15	3.36	3.30	(Savelli et al., 1999)
S-Hexyl 5-Thio-D-Xylonolactone	-2.30	3.66	3.30	3.40	(Lalot et al., 2004)
1-O-Octyl-D,L-Xylitol	-2.17	3.15	3.17	3.40	(Savelli et al., 1999)
3,7-Dimethyloctyl-β-D-Glucoside	-2.40	3.71	3.40	3.50	(Minamikawa & Hato, 2005)[a]
N-Decyl-N-Methyl Gluconamide	-2.89	3.91	3.89	3.60	(Burczyk et al., 2001)
S-Hexyl 1-Thio-L-Ribitol	-1.99	3.59	2.99	3.70	(Lalot et al., 2004)
Decyl-β-D-Glucoside	-2.68	3.71	3.68	3.90	(Aveyard et al., 1998)[a]
N-Dodecanoyl-N-Methyl Lactitolamine	-3.35	4.09	4.35	4.00	(Wilk et al., 2001)
S-Octyl 5-Thio-D-Xylonolactone	-3.28	4.42	4.28	4.20	(Lalot et al., 2004)
Dodecyl-D-Maltonamide	-3.50	4.47	4.50	4.40	(Zhang & Marchant, 1996)
S-Octyl 5-Thio-D-Arabinonolactone	-3.32	4.42	4.32	4.70	(Lalot et al., 2004)
N-Tetradecanoyl-N-Methyl Lactitolamine	-4.17	4.85	5.17	4.90	(Wilk et al., 2001)
N-Oleyl-N-Methyl Lactobionamide	-4.27	5.58	5.27	5.10	(Burczyk et al., 2001)
S-Decyl 5-Thio-D-Xylonolactone	-4.64	5.18	5.64	5.40	(Lalot et al., 2004)
S-Octyl 1-Thio-L-Xylitol	-2.92	4.35	3.92	5.40	(Lalot et al., 2004)
S-Octyl 1-Thio-L-Ribitol	-3.42	4.35	4.42	5.60	(Lalot et al., 2004)
N-Hexadecyl-N-Methyl Lactobionamide	-5.03	5.99	6.03	6.00	(Burczyk et al., 2001)
N-Octadecyl-N-Methyl Lactobionamide	-5.48	6.74	6.48	6.30	(Burczyk et al., 2001)
	MAE[b]	0.31	0.30		

[a] data extracted from graphs; [b] MAE = Mean Absolute Error

List of captions

Figure 1. Distribution of pC_{20} values in the dataset.

Figure 2. Repartition of molecules of training and validation sets in the chemical space of the whole dataset as defined by Principal Component Analysis based on 952 molecular descriptors

Figure 3. Optimized structures of octyl-β-D-glucoside and its fragments (polar head and alkyl chain) at B3LYP/6-31+G(d,p) level.

Figure 4. Octyl glycol.

Figure 5. Experimental vs. calculated values for the constitutional fragment-based model (Eq. 6).

Figure 6. Correlation of Structural Information Content (orders 0 and 1) of the polar head with its molecular weight.

Figure 7. Correlation between log CMC and pC_{20} compared with Abbott criterion.

Figure 1. Distribution of pC_{20} values in the dataset.

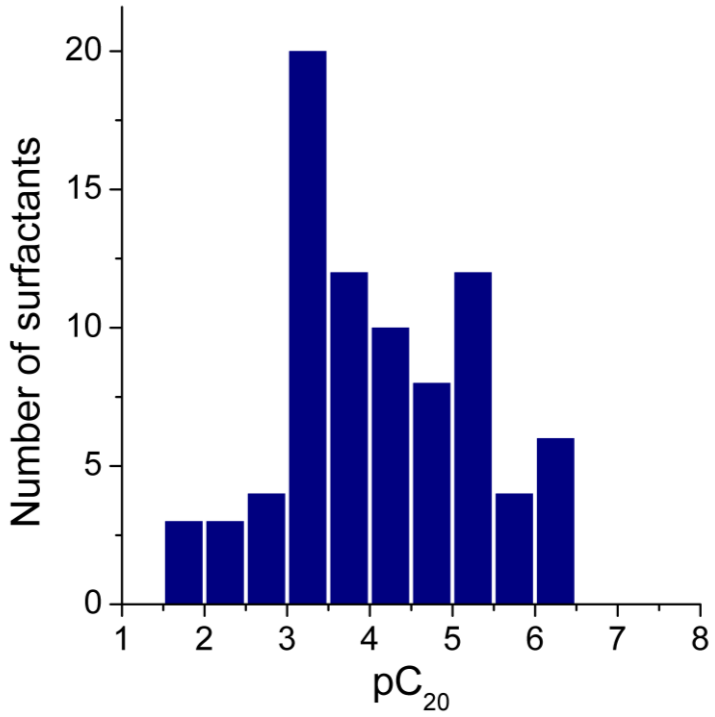

Figure 2. Repartition of molecules of training and validation sets in the chemical space of the whole dataset as defined by Principal Component Analysis based on 952 molecular descriptors

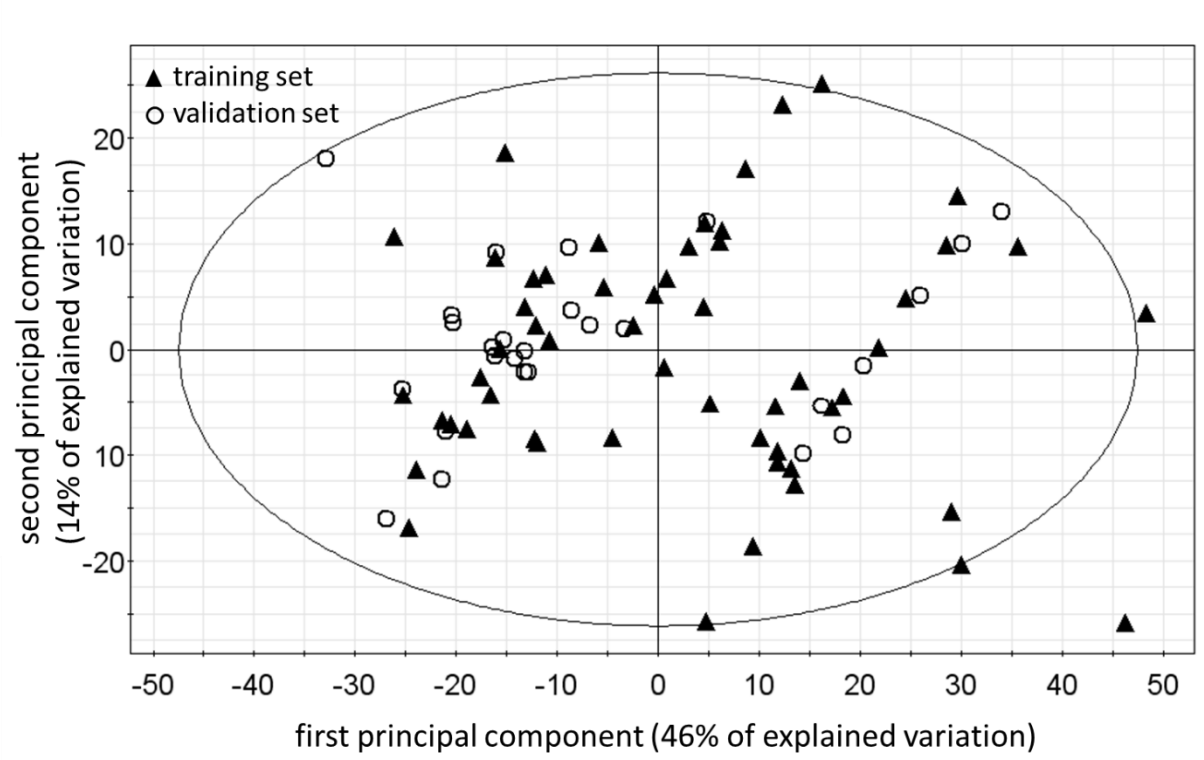

Figure 3. Optimized structures of octyl-β-D-glucoside and its fragments (polar head and alkyl chain) at B3LYP/6-31+G(d,p) level.

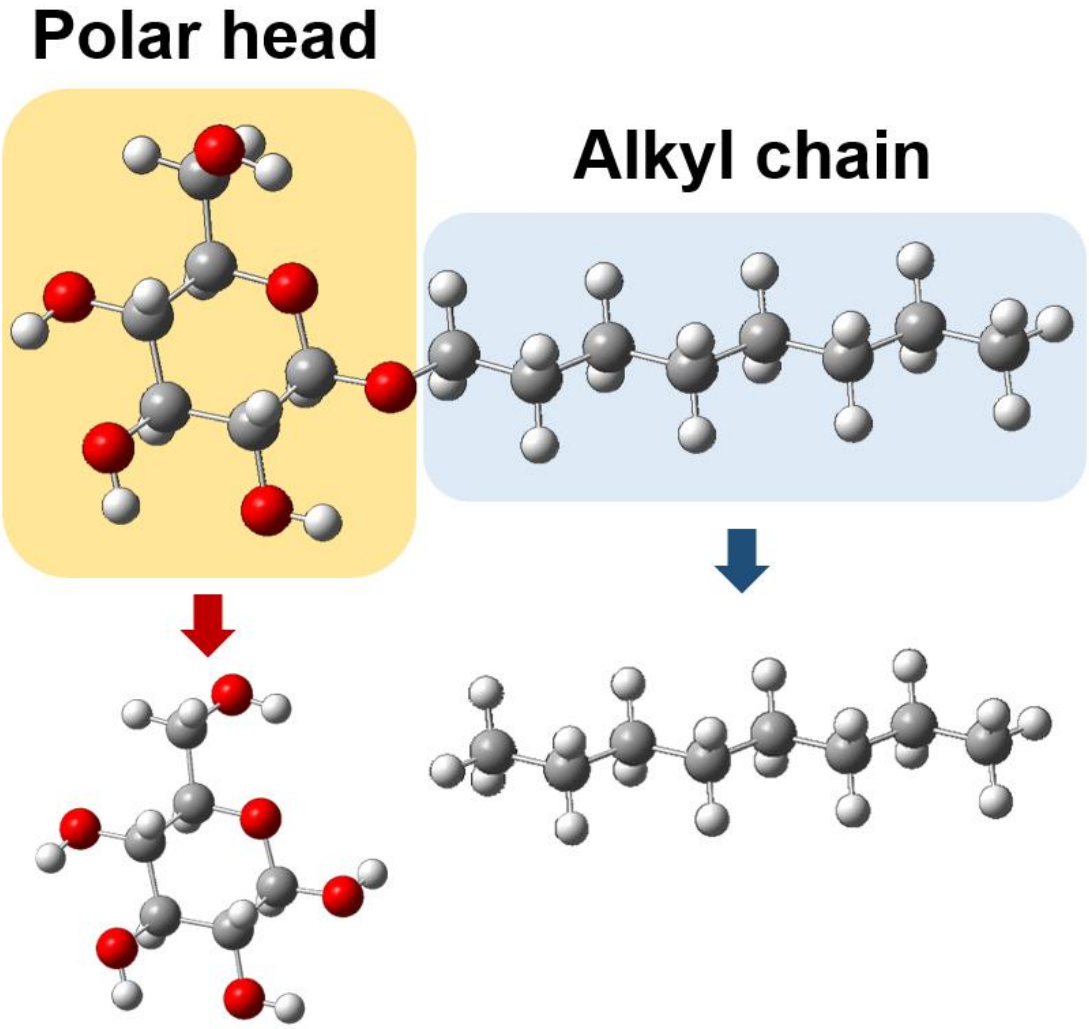

Figure 4. Octyl glycol.

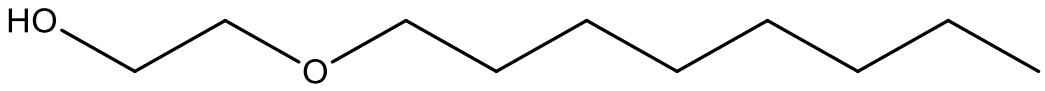

Figure 5. Experimental vs. calculated values for the constitutional fragment-based model (Eq. 6).

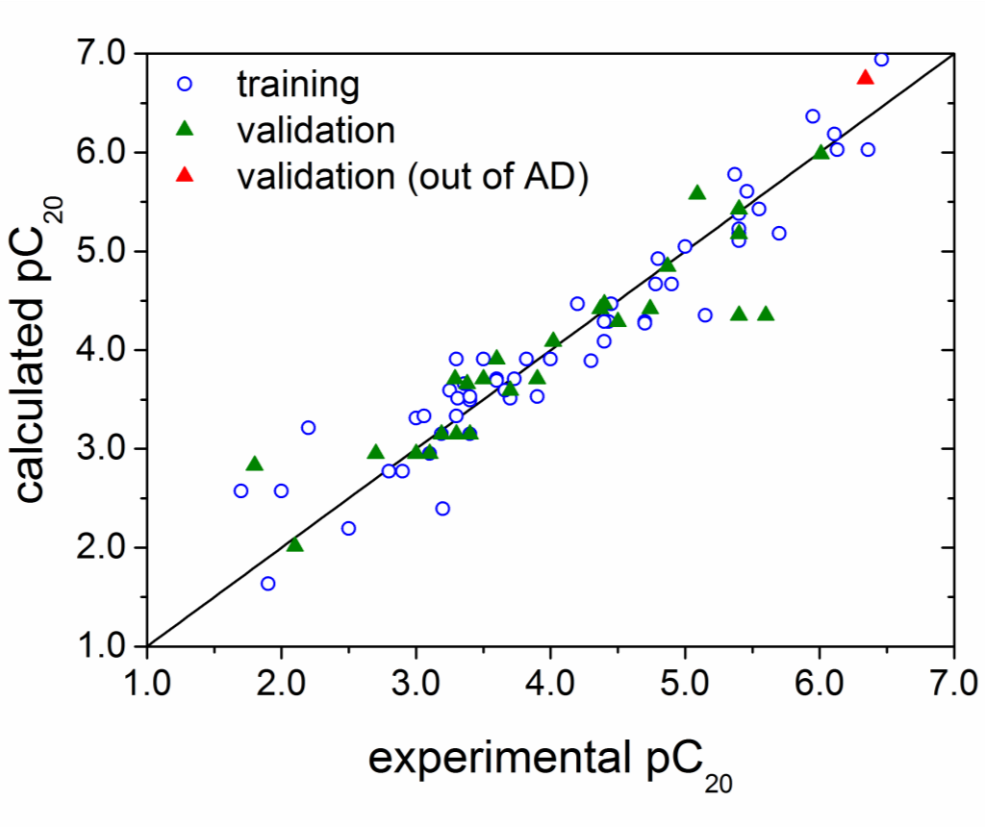

Figure 6. Correlation of Structural Information Content (orders 0 and 1) of the polar head with its molecular weight.

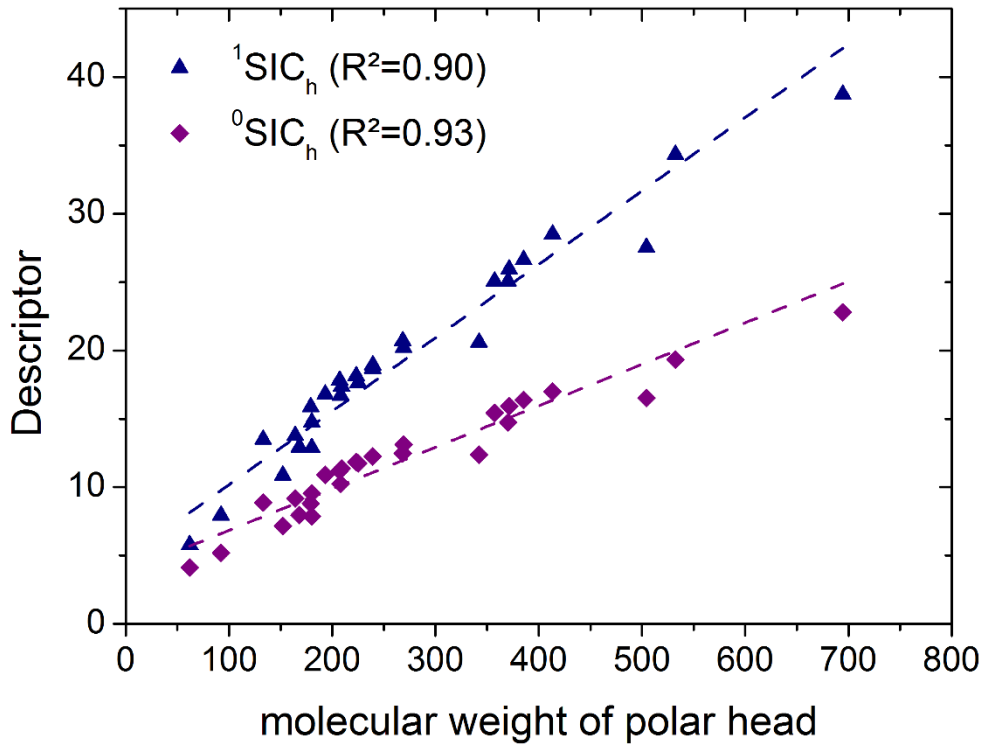

Figure 7. Correlation between log CMC and pC_{20} compared with Abbott criterion.