Anticipating Dissolution Issues of Sugar-Based Surfactants through a Decision Tree Approach

Théophile Gaudin, Guillaume Fayet, Patricia Rotureau, Isabelle Pezron

Anticipating dissolution issues of sugar-based surfactants through a decision tree approach

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

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: isabelle.pezron@utc.fr

Keywords: Krafft temperature, prediction, decision tree, structure-property relationships

Abstract

Sugar-based surfactants are renewable alternatives to petroleum-based surfactants in many applications such as detergency and cosmetics. However, their molecular structure gives often rise to relatively stable crystals, which can induce difficulties to dissolve them in water. This phenomenon is characterized by the Krafft temperature (T_K), above which the surfactant solubility becomes high enough to induce self-association into micelles. Small changes in the molecular structure can result in large T_K differences, which make rationalization and prediction of T_K challenging. Few models were proposed in literature, but none of them are applicable to sugar-based surfactants. In this paper, we propose two decision tree models to estimate whether sugar-based surfactants exhibit potential dissolution issues at room temperature (i.e. T_K above 25°C or not). The first one, based on descriptors of the whole molecule and including quantum-chemical ones, was able to correctly classify 86% of the surfactants in the validation set. The second one, built from simple structural counts of the polar headgroup and the alkyl chain, could predict the right class for 78% of the surfactants in the validation set. These classification models account for experimentally known trends between the molecular structure and T_K, such as the impact of the polar headgroup size, the alkyl chain length and the presence of an amide linkage. To the end, our models were applied to a practical case to show how they can help designing synthesis campaigns of new surfactants.

1. Introduction

Surfactants are an important category of formulation ingredients (in detergents, cosmetics and foods), notably used for their ability to decrease surface tension of water or to solubilize hydrophobic compounds [1]. For such applications, surfactants are frequently used in the form of micelles. So, the ability of surfactants to dissolve as micelles in water can impact their performance properties [1] or their ability to be purified, notably in pre-formulations steps [2].

The Krafft temperature (T_K) is used to assess this dissolution capacity. It is the temperature at which the solubility limit of the surfactant reaches the critical micelle concentration, above which surfactants self-associate as micelles, rather than hydrated crystals, in the aqueous solution [3]. Below T_K, at thermodynamic equilibrium, surfactants cannot form micelles in solution, which limits their performances. Indeed, the maximum surface activity of surfactants is reached when micelles appear, and surface activity of surfactants is exploited in many applications such as detergents and foams [4]. Moreover, surfactant micelles enable to solubilize hydrophobic compounds in an aqueous medium and this solubilisation ability is used in drug design or cosmetics [4]. Knowledge of T_K is of special importance when considering biobased surfactants which are investigated as substitutes of petroleum-based surfactants [5]. Although non-ionic petroleum-based surfactants usually do not exhibit any T_K, some of their potential bio-based substitutes, in particular sugar-based surfactants, do exhibit one [6].

Sugar-based surfactants are characterized by polar heads made from various sugars such as glucose, fructose or sucrose [6]. Due to the tunability of their molecular structures [7, 6], their production safety, renewability and their biocompatibility [5], they are particularly appealing as substitutes to ethylene oxide derivatives, which require the use of hazardous ethylene oxide for their synthesis and are most often produced from fossil resources. In that context, knowing whether a sugar-based surfactant would exhibit a T_K above ambient temperature (e. g. about 25°C) is especially valuable, since in many applications, surfactants are used at ambient temperature or above [4]. Thus, any method to screen sugar-based surfactants with respect to their ability to be dissolved in water would be beneficial to identify the most promising ones in applications, notably as substitutes to more conventional surfactants.

The prediction of T_K is challenging since the property involves the crystalline state of surfactants. Due to this fact, small variations of the molecular structure can have a large and often non-systematic impact on crystal lattice energies, as also recognized for other properties related to the solid phase such as the melting point [8, 9].

Nevertheless, some structural trends have been pointed out for T_K of sugar-based surfactants based on experimental results. At first, T_K increases with alkyl chain length [10]. The linkage (i.e. the chemical moiety of the polar headgroup connected to the nonpolar chain) structure and stereochemistry also impacts T_K significantly. In particular, amide-linked surfactants generally have a high T_K, whereas their methylamide analogues tend to have a lower T_K [11]. For noncyclic polar heads, T_K increases with the number of alcohol units and the stereoregularity of the alcohols [10]. Increasing size of the polar headgroup was found to decrease T_K in most of the cases [12, 13] but, in some specific cases, can increase [14]. While these trends are already a useful qualitative guide to target water-soluble surfactants, no predictive method is available to help rationalization and pre-screening of sugar-based surfactants prior to any synthesis.

One possible predictive method to access surfactant properties is Quantitative Structure-Property Relationships (QSPR). QSPR models are mathematical relationships developed by correlating between the molecular structure, represented by molecular descriptors, and a target physicochemical property [15]. Some QSPR models have been successfully developed to predict the T_K for some specific anionic surfactant families [16-18]. Huibers [16] proposed a multilinear regression (MLR) model, developed on 43 anionic sodium sulfonates and sulfates and obtained a standard error of 5.3°C. From a training set of 32 sulfonates and sulfates, Jalali-Heravi et al. [17] developed another MLR model, achieving a lower standard error of 4.1°C. At last, Li et al. [18] proposed two QSPR models for T_K. One MLR model, for sulfonates and sulfates, was based on a training set of 46 surfactants, and a standard error of 4.5°C was obtained. Another MLR model, based on a training set of 19 sulfonates and perfluorinated carboxylates, was characterized by a larger standard error of 10.4°C. All models included at least one geometrical descriptor (requiring the 3D structure of the surfactant), or quantum-chemical descriptor (based on a computed electronic structure of the surfactant). Although encouraging correlations were exhibited, none of these four models were validated with an external set, and thus the predictive power of the models remains unknown. To our knowledge, no predictive model was developed for T_K of non-ionic surfactants, and especially sugar-based surfactants.

In this context, this study aims to propose first QSPR models to evidence whether sugar-based surfactants exhibit a T_K above 25°C or not (i.e. whether they exhibit dissolution issues at ambient temperature). A series of decision tree models were developed using several types of descriptors. Finally, their applicative potential as a pre-screening tool to focus synthesis campaigns on the most relevant surfactants was demonstrated for a practical application for which the anticipation of dissolution ability in water was important.

2. Computational details

a. Experimental data sets

In a previous work [19], a dataset of 2626 entries on 24 amphiphilic properties of sugar-based surfactants were gathered. We only considered non-ionic surfactants, in order to study an homogeneous ensemble of molecules. For the same reason, only surfactants with one polar head and one alkyl chain were collected as surfactants with more complex structures may show markedly different behaviour [11]. A particular attention was addressed to amino surfactants, as amine linkers may exhibit a basicity that could imply an equilibrium between the non-ionic and the cationic form of the surfactant in solution. For this reason, entries on such surfactants were retained in the final database only when a tensiometric curve confirmed their behaviour was consistent with analogous non-ionic surfactants like in the work of Boullanger et al. [20]. Among them, both quantitative and non-quantitative T_K data were collected. If quantitative data of T_K were only rarely provided (only 37 data), some authors (e. g., Zhu et al. [21]) reported some visual observation of stirred solutions to notice whether surfactant crystals were dissolving or not. This information can be considered as a qualitative evaluation of T_K. Thus, whenever authors reported a verification of surfactant dissolution, we collected this information as non-quantitative data, and kept only those that identified whether surfactants exhibited or not a T_K value above 25°C. Note that the non-quantitative data collection often came from a posterior interpretation of authors data: we assumed that if a solid surfactant dissolves in the liquid at a given temperature, then it does not exhibit a T_K above it. Otherwise, its T_K was considered as above the considered temperature. In particular, whenever the test was conducted at “room temperature”, “ambient temperature” or “standard conditions”, we considered the test temperature to be 25°C. In addition, the 37 quantitative T_K have also been expressed as non-quantitative values (e.g. 37°C is above 25°C). Indeed, if their number is not large enough to build a reliable predictive model, they can in this way be used in the dataset of non-quantitative data.

A careful data curation was then performed to only keep the most reliable T_K data to develop QSPR models. Indeed, any uncertainty in the training data will propagate into the models. In particular, we sought for indications about the purity of the compounds (e.g. NMR spectra or commercial information), since small impurities can have an important impact on the stability of crystal structures [22], a key factor underlying T_K values.

The final dataset (in Supporting Information, Table S1) contained a total of 152 data, including 37 derived from quantitative T_K. These data were partitioned into a training set of 101 data, used to build the decision trees (two thirds of the data) and a validation set of 51 data (one third of the data), used to estimate the predictive power of the decision trees. The validation set should be at best representative of the chemical diversity of the training set in order to ensure that most molecules of the validation set are representative of the applicability domain of the model. Moreover, both sets should be well balanced in terms of surfactants exhibiting $T_K > 25$°C vs. those that do not. The partition used in this study was obtained randomly and satisfies these two criteria (Table 1).

Fig. 1 is a principal component analysis of the descriptors calculated for this study that shows the chemical space spanned by the training set and the validation set. It can be seen that they are distributed in a similar region of the chemical space, i.e. that the molecules of the validation set are representative of the molecules of the training set.

The whole database is also balanced, with 76 T_K above 25°C out of 152 non-quantitative data, and keeps well balanced in both training and validation sets, as presented in Fig. 2. Indeed, the training set contains 52 surfactants with $T_K > 25$°C out of 101, while the validation set contains 24 such surfactants out of 51. Thus, each set is also well-balanced in terms of experimental T_K classes.

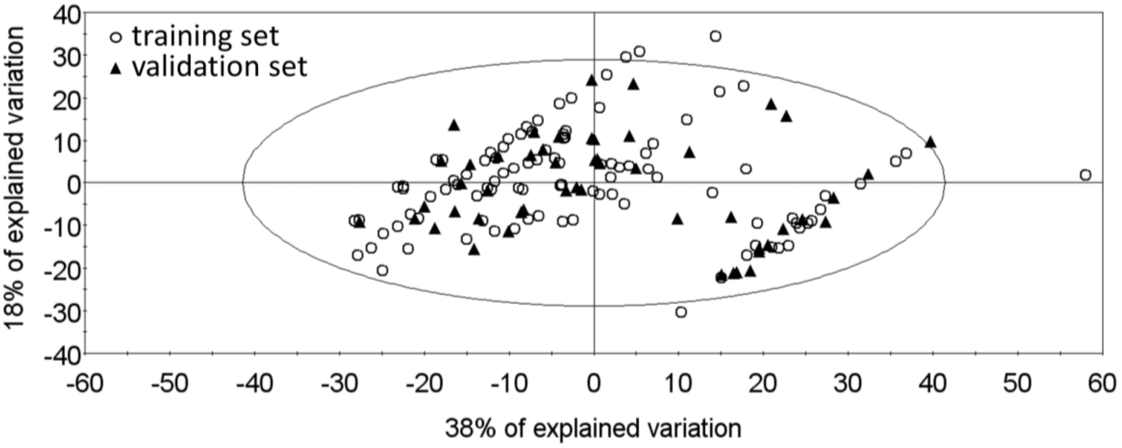


Fig. 1. Repartition of the surfactants belonging to the training and validation sets in the chemical space of the whole data set as defined by principal component analysis based on 896 descriptors

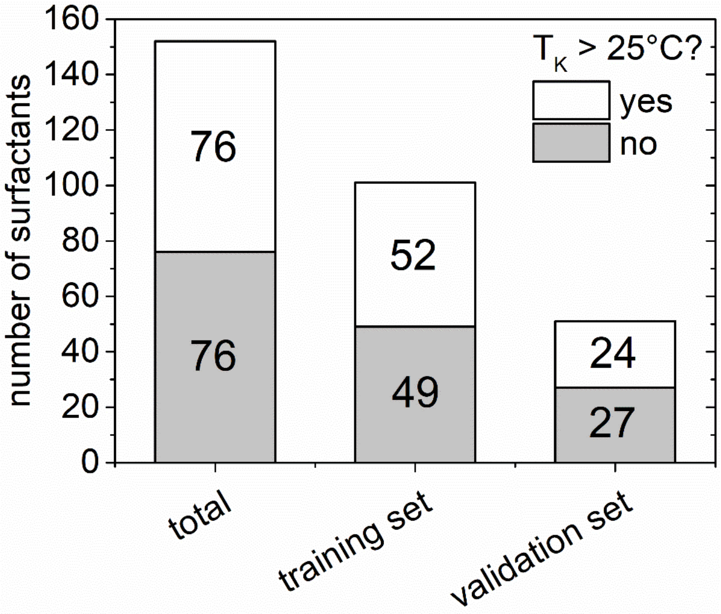


Fig. 2. Distribution of qualitative T_K data for the whole dataset, the training set and the validation set

b. Molecular descriptors

The molecular geometries of the 152 studied sugar-based surfactants of the dataset were optimized using Density Functional Theory (DFT) at B3LYP/6-31+G(d,p) level after preliminary conformation analyses to evidence the most suitable conformation to calculate descriptors. Frequency calculations were also performed at the same level of theory to ensure that the conformation well corresponds to a local minimum on the potential energy surface. This level of calculation has been successfully used in previous works [23] and already used for the development of QSPR models for other properties [24, 25] for this kind of sugar-based surfactants.

The geometries of the 44 hydrophilic (polar heads) and 19 hydrophobic (alkyl chains) fragments constituting the 152 molecules of the dataset were calculated using the same procedure. The separation between the polar headgroup and the alkyl chain was set before the first heteroatom, as in previous works [24, 25]. Then, both fragments were hydrogen-saturated. The Gaussian09 [26] suite of programs was used for all these calculations.

It can be noticed that 28 out of the 152 sugar-based surfactants in Table 1 are in the form of enantiomeric mixtures [27], i.e. surfactants with D and L sugar alcohol polar heads which are difficult to separate due to their identical physical properties, or anomeric mixtures [28], i.e. surfactants with polar heads containing a free anomeric alcohol in two different configurations in aqueous solution, which cannot be separated because a dynamical equilibrium occurs between each other. In all such isomeric mixtures, the different isomers were considered as various 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 calculated structures, about 900 constitutional, topological, geometrical and quantum-chemical descriptors were computed using CODESSA software [29] for each surfactant and each fragment. Additional descriptors were also obtained directly from the quantum-chemical calculations. Descriptors arising from conceptual DFT [30, 31] (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 headgroup and of the first hydrocarbon fragment of the alkyl chain (CH_2 or CH here) were also calculated based on Mulliken [32] and Natural Populations Analyses [33] (as implemented into Gaussian09 software), to take into account the possibility of electron withdrawing from polar headgroup to alkyl chain in surfactants as proposed by Huibers [34].

c. Model development

In this work, decision trees were developed to classify surfactants according to the possibility of T_K > 25°C. As represented in Fig. 3, a decision tree is a model that classifies an instance in a leaf, associated with a predicted class, according to a set of consecutive rules (the nodes). In the context of this study, a node checks whether a given descriptor is above or below a threshold value, and a leaf either predicts a T_K value above 25°C or not for the surfactant. To build decision trees from training set data, the J48 method, a Weka [35] implementation of the C4.5 method [36], was used. This method consists of different steps. At first, for each available descriptor, the threshold enabling to best separate the entire set of molecules between the two classes is calculated and the best descriptor constitutes the first node of the decision tree which separates the surfactants into two new subsets. The same operation is then applied for each new subset until no significant separation between the two classes is obtained. At last, the tree is pruned to avoid over-parameterization by examining whether each node is beneficial to lower classification errors (based on a Bayesian approach).

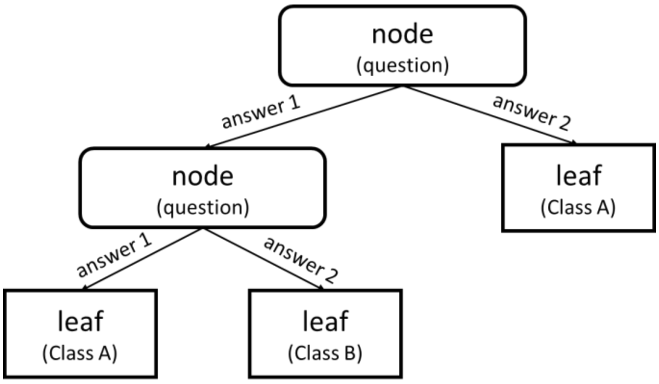


Fig. 3. Schematic representation of a decision tree

d. Model validation

To test the performances and predictive power of decision trees, Cooper statistics [37] were used (cf. Table 1). These statistics are based on the so-called confusion matrix, which summarizes the classification performances issued from a series of predictions.

Confusion matrix

T_K > 25°C ?		Experiment	
		yes	No
Prediction	yes	TP	FN
	no	FP	TN

Classification indicators

TP	True Positive	**TN**	True Negative
FP	False Positive	**FN**	False Negative
Acc	Accuracy	(TP + TN) / (TP + TN FP + FN)	
PP	Positive Predictivity	TP / (TP + FP)	
NP	Negative Predictivity	TN / (TN + FN)	

Table 1. Confusion matrix and classification indicators

In this study, TP is the number of molecules with the correct classification "T_K > 25°C", FP is the number of molecules with an incorrect classification "T_K > 25°C", FN is the number of molecules with an incorrect classification "no T_K > 25°C" and TN is the number of molecules with the correct classification "no T_K > 25°C".

Acc represents a general assessment of the quality of classification for a given set. PP and PN are focused on the performance of the model to classify in a particular class. In our case, PP represents whether the prediction of T_K > 25°C occurrence is reliable, and NP represents whether the opposite prediction is reliable. The closer Acc, PP and NP are to 100%, the higher are the performances of the model.

The quality of fitting of the models was evaluated on these criteria for the predictions performed on the training set. To the end, to assess the predictive power of the models, the surfactants of the validation set, not used to train the models, are classified by the developed trees, and the resulting Acc, PP and NP enabled to assess the predictive power of the new decision trees.

3. Results and discussion

a. Development of classification models

Various decision trees were developed depending on the type of descriptors used. Some models were developed from descriptors of the whole molecule (i for integral), others from fragment descriptors (f). For each of these two approaches, three types of decision trees were developed either including quantum-chemical descriptors (all for all types of descriptors), focusing on constitutional and topological descriptors (ct), or only with constitutional descriptors (c). The six decision trees and their performances are summarized in Table 2 and detailed in Supporting Information (Figs. S1-S6).

Type	n_{desc}	Training			Validation		
		Acc	PP	NP	Acc	PP	NP
i/all	9	95%	98%	93%	86%	92%	81%
i/ct	3	82%	80%	84%	75%	73%	76%
i/c	2	72%	78%	69%	76%	86%	70%
f/all	5	89%	93%	86%	76%	78%	75%
f/ct	3	83%	85%	82%	75%	79%	70%
f/c	4	82%	88%	78%	78%	90%	71%

Table 3. Performances of the six decision trees developed in the present study

Predictive capabilities of the final models range from 75% to 86% in global accuracy (Acc) on the validation set. In most of the cases, the positive predictivity (accuracy of T_K > 25°C prediction) was greater than the negative predictivity (for five out of six models, in training and in validation). Such difference indicates that the obtained decision trees are especially efficient in the identification of non-dissolving surfactants. This trend is even more pronounced for the best models in terms of accuracy in the validation set, i/all and f/c, with differences of 11% and 19%, respectively. So, these models are particularly recommended to identify molecules with dissolution issues.

The decision tree presenting the highest accuracy in prediction was obtained for the 273 integral descriptors of all types (i/all, shown in Fig. 4). Its predictive power was high, with Acc = 86% for the validation set. As for most models, the model tends to be even more reliable to identify surfactants with T_K > 25°C (PP = 92%), as compared to the opposite prediction (NP = 81%). In pre-screening applications, this feature is particularly useful to discard surfactants that would be likely not to dissolve in water.

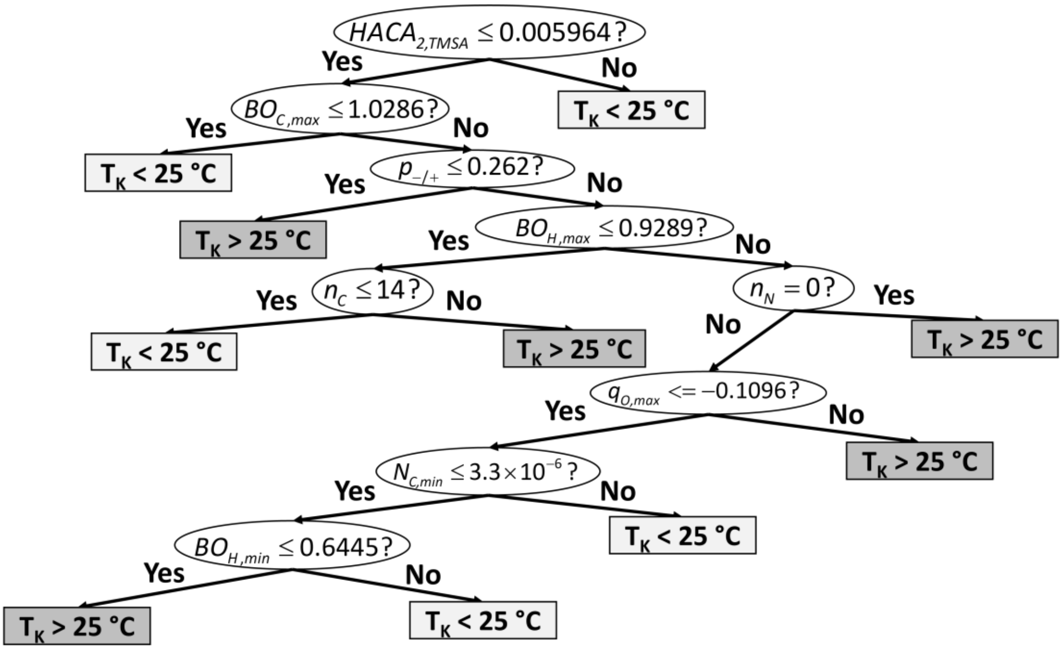


Fig. 4. Decision tree based on integral descriptors for the classification of the surfactants upon their T_K

In this decision tree, $HACA_{2,TMSA}$ is the hydrogen-bond acceptor surface area of order 2 divided by the total molecular surface area and based on Zefirov [38] partial charge model, $BO_{C,max}$ is the maximal bond order of a C atom, $p_{-/+}$ is the polarity parameter (i.e. the difference between maximal and minimal partial charges), $BO_{H,max}$ is the maximal bond order of a H atom, n_C is the number of C atoms, n_N is the number of N atoms, $q_{O,max}$ is the maximal partial charge for a O atom based on Zefirov partial charge model, $N_{C,min}$ is the minimal nucleophilicity index for a C atom, and $BO_{H,min}$ is the minimal bond order for a H atom (> 0.1).

The structure of the decision tree is relatively complex. However, it is consistent with some known experimental trends. In particular, $HACA_{2,TMSA}$, the descriptor at the first node of the decision tree, is related to both the alkyl chain length and the polar headgroup size. Indeed, since alkyl chains do not contain H-acceptors atoms, for a given polar head, $HACA_{2,TMSA}$, is larger for shorter alkyl chains. In addition, when keeping alkyl chain constant, increasing the polar headgroup size usually corresponds, for sugar-based surfactants, to the addition of oxygen and nitrogen atoms, both H-acceptors. Thus, $HACA_{2,TMSA}$, is larger for larger polar heads. Moreover, H-bonding capability is known to influence crystallization. As a consequence, H-acceptor behaviors (as well as H-donor ones) are expected to be relevant to T_K.

The number of N atoms (n_N) is present at another node in the decision tree, at which the surfactants containing N atoms are classified as $T_K > 25°C$. This relates to the generally low solubility observed for surfactants with amide linkages [39].

At last, the number of C atoms (n_C) in the molecule also appears at the end of a branch of the tree. Surfactants of the tested subsets are classified as exhibiting dissolution issues for high n_C. Since C atoms are present in both the alkyl chain and the polar headgroup of the surfactant, this descriptor may reflect the overall size of the surfactant. It is known that the melting point generally increases with the size of the molecule [8], and T_K is also sometimes defined as corresponding to the melting point of the hydrated surfactant [40]. Thus, within a subset, it is not surprising to classify larger surfactants as non-dissolving.

If the other parameters in the tree (minimal and maximal atomic bond orders, atomic nucleophilicity indices and partial charges) are less easily interpretable, they are all related to charge distributions of surfactants inside the head, and notably to possible H-bonding sites on the head (e.g. $q_{O,max}$) that favor crystallization [22].

Among the developed decision trees, another one was evidenced (Fig. 5), owing to its simplicity of use. Based on simple atomic counts of the alkyl chain and the polar headgroup of the molecule, it also presented satisfactory prediction performances with Acc = 78% for the validation set. Once again, predictions of dissolution issues from this decision tree seems particularly reliable (PP = 90%) and the decision tree is more likely relevant to identify surfactants presenting them.

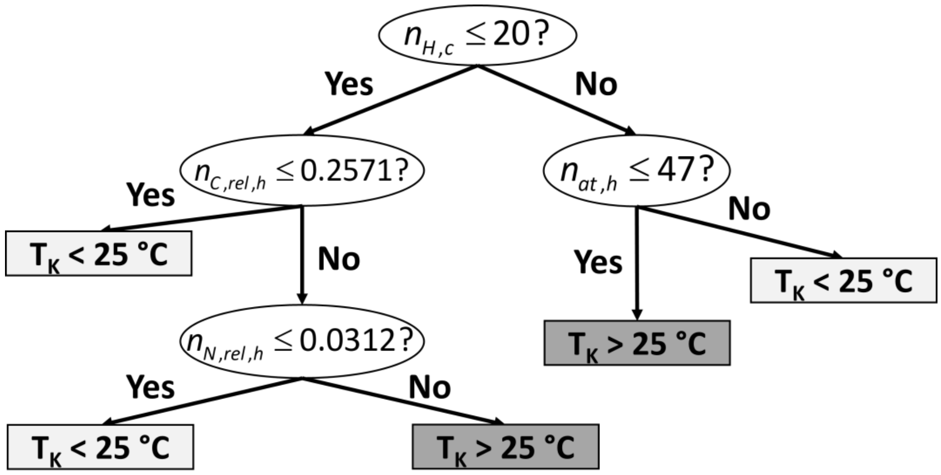


Fig. 5. Decision tree based on fragment descriptors for the classification of the surfactants upon their T_K

In this decision tree, $n_{H,c}$ is the number of H atoms in the alkyl chain, $n_{C,rel,h}$ and $n_{N,rel,h}$ are the relative number of C and N atoms in the polar head, and $n_{at,h}$ is the number of atoms in the polar head. The structural trends involved in the different nodes of the tree are in agreement with those already identified by experimentalists. Furthermore, the decision tree rationalizes these known experimental trends on a series of threshold values of structural descriptors of surfactants.

The first node separates molecules according to their number of hydrogens in the alkyl chain. Since longer chains have more hydrogen atoms, this descriptor is related to the alkyl chain length which is known as a critical structural factor that impacts Krafft temperatures through higher van der Waals interactions, T_K increasing with alkyl chain length [10]. Besides, in the experimental database, 39 of the 54 surfactants with short alkyl chains (containing 20 H atoms or less) dissolve in water at 25°C. On the contrary, the majority of surfactants with long alkyl chains (61 out of the 98 surfactants containing more than 20 H atoms) shows T_K > 25°C.

For the longest alkyl chains (above 20 H atoms, which corresponds to 9 C atoms for saturated alkyl chains), the next node uses the number of atoms in the polar head. This descriptor is related to the size of the polar head, which is another relevant structural factor impacting T_K. It decreases with larger polar heads in most of the cases [11]. Accordingly, polar heads with more than 47 atoms are classified as dissolving at 25°C.

It can be noticed that polar heads with one or two sugar residues often contain less than 47 atoms (for example, maltose contains 43 atoms). Thus, most of the time, surfactants with one or two sugar residues (which constitute most of the studied sugar-based surfactants) and an alkyl chain of 10 atoms or longer are classified as exhibiting T_K > 25°C. This suggests that in our database, sugar-based surfactants with 10 C atoms or more in the alkyl chain are likely to exhibit T_K above 25 °C, which is lower than the threshold of 12 atoms proposed by Marchant et al. [11].

In the case of alkyl chains with 20 H atoms or less (corresponding to a saturated alkyl chain length of 9 C atoms or less), the next node is based on the relative number of C atoms in the polar head. This can be related to the level of polarity of the polar head, as O or N atoms in the polar headgroup decrease the value of the descriptor. The decision tree accounts for the fact that, with a more polar head, a surfactant is more hydrophilic, which can favor dissolution in water. Based on this fact, the decision tree classifies heads with a higher polarity ($n_{C,rel,h} \leq 0.2571$) as dissolving.

At last, surfactants with small alkyl chain and weakly polar headgroup are classified according to the relative number of N atoms in the polar head. At this node, surfactants containing N atoms (generally in amide or amide groups) tend to exhibit dissolution issues whereas surfactants with no N atom as dissolving. This classification is in agreement with the experimental trend observed that sugar-based surfactants with amide linkers that more likely present dissolution issues than their analogues with other linkers [11].

b. Application of Krafft point models to guide synthesis campaigns

Surfactant ability to dissolve in water can be critical in various applications, thus, its anticipation can guide synthesis campaigns. For instance, in recent studies, Lu et al. [41, 42] studied the cytotoxicity of sugar-based surfactants and synthesized a range of molecules characterized by gradual structural modifications (cf. Table 3) to investigate the impact of these modifications on the surfactant/cell interactions. It was therefore essential to know the surfactant physical state, as monomer, micelle or solid, because it may affect the mechanism of biological activity.

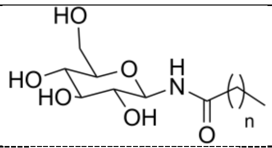
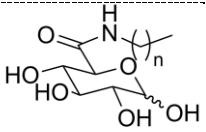
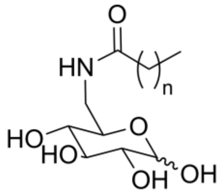
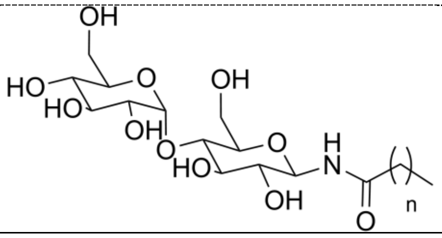
Surfactant Id	structure	experimental dissolution issues at 25°C	predicted T_K > 25°C? qc (Fig. 4)	simple (Fig. 5)
1 (n=6)		no dissolution T_K = 30°C	no	yes
2 (n=8)		no dissolution	yes	yes
3 (n=7)		turbid solution	yes	no
4 (n=9)		turbid solution	yes	yes
5 (n=6)		no dissolution (T_K = 32°C)	no	yes
6 (n=8)		no dissolution	yes	yes
7 (n=6)		dissolution	no	no
8 (n=8)		dissolution	no	no

Table 3. Experimental and calculated T_K for the sugar-based surfactants investigated in the cytotoxicity study [41, 42]

From the analysis of reported studies, dissolution issues are usually expected at or above a chain length of 12 for non-ionic sugar-based amphiphiles [11]. Synthesized molecules were designed with chain lengths of 7 to 10 carbons and, unexpectedly, dissolution issues in water were even observed with 8 carbon alkyl chains for surfactants with one sugar residue [41, 42]. Six of the studied surfactants were either only partially dissolved or formed a turbid solution (Table 2).

At that time, no model was available to estimate the T_K of sugar-based surfactants. In the present work, T_K was estimated from the qualitative models (cf. Table 3). With the two models, all molecules exhibiting dissolution issues were identified, while the ones that did not exhibit such issues (maltose derivatives - surfactants 7 and 8) were also correctly identified. For three of the six non-dissolving molecules contradictory predictions were obtained for the shorter chain, and in two of them, the measured T_K was close to 25°C (30 and 32°C), which suggest that contradictory predictions from both models can be beneficial to identify surfactants with T_K close to 25°C. Moreover, both models correctly predicted dissolution issues for surfactants with 9 to 10 carbon atoms in the alkyl chain and only one sugar residue in the polar headgroup (surfactants 2, 4 and 6). Thus, T_K significantly above 25°C may be expected for these surfactants.

The results show that the predictive models developed for T_K for the non-ionic sugar based surfactants of this study (not present in the dataset used for the development of the model) were able to raise relevant dissolution issues even for derivatives with short alkyl chains. This approach can therefore be used favorably as a pre-screening tool prior to synthesis campaign of new surfactants.

4. Conclusion

The Krafft point is an important surfactant property to describe surfactant ability to dissolve in water. Indeed, when the temperature is below the Krafft point of a surfactant, it cannot be used at its optimal surface activity and solubilization potential. Thus, knowing surfactant T_K helps to estimate their performance in formulations. In this study, we developed new models to qualitatively predict whether sugar-based surfactants exhibit a Krafft point above room temperature (i.e. dissolution issues at ambient temperature). The best model is able to correctly classify 86% of sugar-based surfactants but requires quantum chemical calculations for each tested surfactant. Another, simpler model was evidenced, based on atomic counts in polar heads and alkyl chains. It correctly classifies 78% of sugar-based surfactants. These models identify surfactants exhibiting $T_K > 25°C$ more accurately than those that do not. The descriptors and structures of the decision trees account for different known experimental trends and rationalize them as successions of thresholds, like the increase of the Krafft point with the size of the alkyl chain, its frequent decrease with the size of the polar head, or its increase with the presence of amide linkage. In particular, one of the decision trees emphasizes that many surfactants start to exhibit dissolution issues between chain lengths of 9 and 10 carbon atoms, lower than the literature threshold of 12 carbon atoms. To the end, we illustrated how these models can help to design relevant experimental synthesis campaigns, by pointing out potential dissolution issues. Molecular design applications are also possible for the developed models, by calculating the properties of a large number of combinations from a few polar heads and alkyl chains and raise solubility issues of possible candidates in surfactant formulations.

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).

5. Supporting Information

Dataset used in this word (Table S1), details of the developed decision trees and their performances (Figs. S1-S6).

6. References

1. Rosen MJ, Kunjappu JT. Surfactants and Interfacial Phenomena. 4th ed. John Wiley & Sons, Inc.; 2012.
2. Pedro Pinho S, Almeida Macedo E. Solubility in Food, Pharmaceutical, and Cosmetic Industries. In: Letcher TM, editor. Developments and Applications in Solubility. RSC Publishing; 2007.
3. Tsujii K. Thermodynamic Studies on the Krafft point in Aqueous Surfactant Systems: Osaka University; 1987.
4. Myers D. Surfactant Science and Technology. 3rd ed. Wiley-Interscience; 2006.
5. Kjellin M, Johansson I. Surfactants from Renewable Resources. 1 ed. John Wiley & Sons, Ltd; 2010.
6. Ruiz CC. Sugar-Based Surfactants: Fundamentals and Applications. CRC Press, Taylor & Francis Group; 2009.
7. Dembitsky V. Astonishing diversity of natural surfactants: 1. Glycosides of fatty acids and alcohols. Lipids. 2004;39(10):933-53.
8. Katritzky AR, Kuanar M, Slavov S, Hall CD, Karelson M, Kahn I et al. Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. Chem Rev. 2010;110(10):5714-89. doi:10.1021/cr900238d.
9. Nieto-Draghi C, Fayet G, Creton B, Rozanska X, Rotureau P, de Hemptinne J-C et al. A General Guidebook for the Theoretical Prediction of Physicochemical Properties of Chemicals for Regulatory Purposes. Chem Rev. 2015;115(24):13093-164. doi:10.1021/acs.chemrev.5b00215.
10. Laughlin RG, Scheibel JJ, Fu YC, Wireko FC, Munyon RL. N-Alkanoyl-N-Alkyl-1-Glycamines. In: Holmberg K, editor. Novel Surfactants. Surfactant Science: CRC Press; 2003. p. 1-34.
11. Marchant RE, Anderson EH, Zhu J. Polysaccharide Surfactants: Structure, Synthesis, and Surface-Active Properties. In: Dimitriu S, editor. Polysaccharides, Structural Diversity and Functional Versatility. 2nd ed.: Marcel Dekker; 2005. p. 1055-86.
12. Van Doren HA. Tailor-made carbohydrate surfactants? Systematic investigations into structure-property relationships of N-Acyl N-Alkyl 1-Amino-1-Deoxy-D-Glucitols. Carbohydrates as Organic Raw Materials III. Wiley-VCH Verlag GmbH; 2007. p. 255-72.
13. Burczyk B, Wilk KA, Sokołowski A, Syper L. Synthesis and Surface Properties of N-Alkyl-N-methylgluconamides and N-Alkyl-N-methyllactobionamides. J Colloid Interface Sci. 2001;240(2):552-8. doi:http://dx.doi.org/10.1006/jcis.2001.7704.
14. Hato M. Synthetic glycolipid/water systems. Current Opinion in Colloid & Interface Science. 2001;6(3):268-76. doi:http://dx.doi.org/10.1016/S1359-0294(01)00096-6.
15. Karelson M. Molecular descriptors in QSAR/QSPR. Wiley; 2000.
16. Huibers PDT. Surfactant Self-Assembly, Kinetics and Thermodynamics in Micellar and Microemulsion Systems.: University of Florida; 1996.
17. Jalali–Heravi M, Konouz E. Use of Quantitative Structure–Property Relationships in Predicting the Krafft Point of Anionic Surfactants. IEJMD. 2002;1(8):410–7.
18. Li Y, Xu G, Luan Y, Yuan S, Xin X. Property Prediction on Surfactant by Quantitative Structure-Property Relationship: Krafft Point and Cloud Point. J Disper Sci Technol. 2005;26(6):799-808. doi:10.1081/DIS-200063127.
19. Gaudin T. Développement de modèles QSPR pour la prédiction et la compréhension des propriétés amphiphiles des tensioactifs dérivés de sucre: Ph. D. thesis, Université de Technologie de Compiègne; 2016.
20. Boullanger P, Chevalier Y. Surface Active Properties and Micellar Aggregation of Alkyl 2-Amino-2-deoxy-β-d-glucopyranosides. Langmuir. 1996;12(7):1771-6. doi:10.1021/la950485i.
21. Zhu Y-P, Rosen MJ, Vinson PK, Morrall SW. Surface Properties of N-Alkanoyl-N-methyl Glucamines and Related Materials. J Surfact Deterg. 1999;2(3):357-62.
22. Price SL. Predicting crystal structures of organic compounds. Chem Soc Rev. 2014;43(7):2098-111. doi:10.1039/C3CS60279F.
23. Gaudin T, Rotureau P, Pezron I, Fayet G. Conformations of n-alkyl-α/β-d-glucopyranoside surfactants: Impact on molecular properties. Comp Theor Chem. 2017;1101:20-9. doi:http://dx.doi.org/10.1016/j.comptc.2016.12.020.

24. Gaudin T, Rotureau P, Pezron I, Fayet G. New QSPR Models to Predict the Critical Micelle Concentration of Sugar-Based Surfactants. Ind Eng Chem Res. 2016;55(45):11716-26. doi:10.1021/acs.iecr.6b02890.
25. Gaudin T, Rotureau P, Pezron I, Fayet G. Investigating the impact of sugar based surfactants structure on surface tension at critical micelle concentration with structure-property relationships J Colloid Interface Sci. 2018;516:162-71.
26. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR et al. Gaussian 09, Revision B.01. Wallingford CT2009.
27. Savelli MP, Van Roekeghem P, Douillet O, Cavé G, Godé P, Ronco G et al. 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(2):221-36. doi:http://dx.doi.org/10.1016/S0378-5173(99)00078-2.
28. Maunier V, Boullanger P, Lafont D, Chevalier Y. Synthesis and surface-active properties of amphiphilic 6-aminocarbonyl derivatives of d-glucose. Carbohydr Res. 1997;299(1–2):49-57. doi:http://dx.doi.org/10.1016/S0008-6215(96)00336-9.
29. Codessa, www.semichem.com/codessa/.
30. Chermette H. Chemical reactivity indexes in density functional theory. J Comput Chem. 1999;20:129-54.
31. Geerlings P, De Proft F, Langenaeker W. Conceptual density functional theory. Chem Rev. 2003;103:1793-873.
32. Mulliken RS. Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I. J Chem Phys. 1955;23(10):1833-40. doi:doi:http://dx.doi.org/10.1063/1.1740588.
33. Reed AE, Weinstock RB, Weinhold F. Natural population analysis. J Chem Phys. 1985;83(2):735-46. doi:doi:http://dx.doi.org/10.1063/1.449486.
34. Huibers PDT. Quantum-Chemical Calculations of the Charge Distribution in Ionic Surfactants. Langmuir. 1999;15(22):7546-50. doi:10.1021/la990367l.
35. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor Newsl. 2009;11(1):10-8. doi:10.1145/1656274.1656278.
36. Quinlan JR. C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc.; 1993.
37. Cooper JA, Saracci R, Cole P. Describing the Validity of Carcinogen Screening Tests. Br J Cancer. 1979;39:87-9.
38. Zefirov NS, Kirpichenok MA, Izmailov FF, Trofimov MI. Scheme for the Calculation of the Electronegativities of Atoms in a Molecule in the Framework of Sanderson's Principle. Dokl Akad Nauk SSSR. 1987;296:883-7.
39. Piispanen PS. Synthesis and Characterization of Surfactants Based on Natural Products. Stockholm: Kungl Tekniska Högskolan; 2002.
40. Shinoda K, Tamamishi B-I, Nakagawa T, Isemura T. Colloidal Surfactants - Some physicochemical properties. Physical chemistry. Academic Press; 1963.
41. Lu B. Evaluation of physico-chemical properties of biorefinery-derived amphiphilic molecules and their effects on multi-scale biological models.: PhD. Thesis: Université de Technologie de Compiègne; 2015.
42. Lu B, Vayssade M, Miao Y, Chagnault V, Grand E, Wadouachi A et al. Physico-chemical properties and cytotoxic effects of sugar-based surfactants: Impact of structural variations. Colloids Surf, B. 2016;145:79-86. doi:http://dx.doi.org/10.1016/j.colsurfb.2016.04.044.