J Comput Aided Mol Des (2008) 22:441–460
DOI 10.1007/s10822-008-9204-9

On the importance of topological descriptors in understanding structure–property relationships

David T. Stanton

Received: 1 October 2007 / Accepted: 20 February 2008 / Published online: 13 March 2008

Abstract It has been generally observed in our work that molecular descriptors derived from a molecular graph theory or topological representation of structure play an important and often key role in many QSAR and QSPR models we have developed. These descriptors do not only provide the means to generate a good fit to the observed data used to train the models, but they also provide information that is needed to generate a clear physical interpretation of the underlying structure–activity or property relationships. In addition, these descriptors provide a conformation-independent method of measuring the key features of molecular structure that affect the observed properties of the molecules. These characteristics are exemplified in a model developed to predict critical micelle concentration (CMC). A model is described that exhibits excellent predictive strength, is independent of conformation of the structures used, and that yields a great deal of detail regarding the underlying structure–property relationship driving the observed CMC.

Keywords Critical micelle concentration · Molecular connectivity · Physical interpretation · QSAR · QSPR · Structure–property relationship · Surfactants · Topological descriptors

Abbreviations

2D	2-Dimensional
3D	3-Dimensional
CMC	Critical micelle concentration
LogCMC	Base-10 logarithm of the CMC
CPSA	Charged partial surface area
HAS	Hydrophobic surface area
PLS	Partial least squares, or projection of latent structures
PRESS	Predicted sum of squared (error)
QSAR	Quantitative Structure–Activity Relationship
QSPR	Quantitative Structure–Property Relationship
SIR	Structure information representation
SPR	Structure–property relationship
VIF	Variance inflation factor

D. T. Stanton (✉)
Corporate Research, Modeling and Simulations Department, Procter & Gamble, Miami Valley Innovation Center, 11810 East Miami River Road, Cincinnati, OH 45252, USA
e-mail: stanton.dt@pg.com

Introduction

The purpose of a molecular descriptor in a quantitative structure–activity and, more broadly, a structure–property relationship (QSAR and QSPR, respectively) application is to provide a measure of a particular feature of the structure of the compounds being studied. The goal is simply to measure the feature in question as accurately and unambiguously as possible. Several different representations of molecular structure are often used, each providing a unique perspective on the nature of a molecule, in order to assemble a diverse set of measures of molecular structure. A subsequent statistical analysis is used to identify the subset of descriptors that maximally explain the variance in the observed property or reactivity of interest. The physical interpretation of the model is arrived at by an examination of the changes in key structural features identified by the descriptors in the context of model training set [1]. As such, there is no requirement that a descriptor has with it any preexisting physical interpretation related to the property being studied. This is why one particular

descriptor can play very different roles in models for different properties. The term *Structure Information Representation* (SIR) has been used to capture the notion that the primary role of a molecular descriptor is to provide information about the molecular structure, which is subsequently interpreted in the context of the structures being examined [2–4].

The value and utility of topological descriptors in QSAR and QSPR applications has been criticized [5]. However, our experience has been that topological descriptors are not only useful in generating good fitting internally and externally validated models, they often make the greatest statistical contribution to the model and also provide a high degree of detail with regard to how changes in the molecular structure relate to differences in observed activities or properties of the compounds being studied. An additional important characteristic of topological descriptors that is sometimes overlooked is their independence of structural conformation. This conformational independence is particularly important in the study of molecules that are flexible and when the proper conformation of the molecules is not well defined. As we consider the rebirth of QSAR as a discipline (QSAR Reborn, a symposium honoring Dr. Philip Magee, 234th Nation Meeting of the American Chemical Society, Boston, MA. August 19–23, 2007), it seems appropriate to revisit the class of descriptors that are derived from a molecular topological representation of structure.

The work described here illustrates the importance of topological descriptors for generating QSPR models that are predictive and that provide clear and detailed information regarding the underlying structure–property relationships useful for molecular design purposes. The property of interest here is the critical micelle concentration, or CMC, of anionic surfactants. A micelle is a colloidal-sized cluster of amphiphilic (surfactant) molecules in solution [6]. In the case of aqueous solutions, surfactants form micelles with the nonpolar hydrophobic portions of the molecule, or tail, oriented toward the center of the cluster and the polar portions, or head group, oriented toward the solvent. At low concentrations, too few individual surfactant molecules are available to achieve an effective elimination of the hydrocarbon-water interface [7]. However, as the concentration of the surfactant is increased, a point is reached where there are sufficient numbers of surfactant molecules available to begin forming micelles (see Fig. 1). The concentration at which micelles begin to form defines the CMC. The CMC of a surfactant is an important defining property of a surfactant relating to its surface tension or interfacial tension reduction and detergency. This particular property is well understood, and the underlying structure–property relationship is clearly defined as a balance of attractive and repulsive forces in a solution of amphiphiles [7]. In aqueous solutions there is an attractive force between the hydrophobic portions of the amphiphile, a negative affect of the disruption of structure of water by the tail groups, and in the case of ionic surfactants there is the repulsive force between head groups of like charge. Micelles form as a result to minimize the negative and repulsive forces and maximize the attractive forces. Structural features that affect the size and shape of a micelle formed in aqueous solution are the volume occupied by the hydrophobic group, the length of the hydrophobic group, and cross-sectional area of the hydrophilic group. Thus, CMC was selected as the subject for this study because it is a relatively simple property with a well defined structure–property relationship, and it allows for the clear illustration of two important characteristics of topological descriptors: their ability to provide a high degree of detail regarding the structure–property relationship, and the importance of their conformation independence.

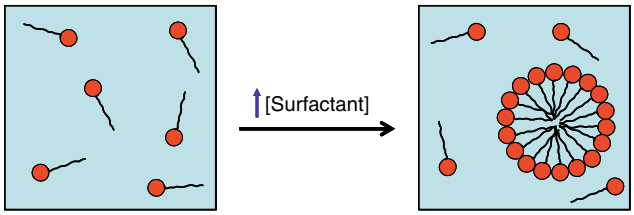

Fig. 1 A solution of surfactant below the critical micelle concentration contains only free surfactant (left). As the concentration of surfactant is increased, the critical micelle concentration is reached and micelles begin to form (right)

Experimental

Data set

The data used in this study involving 175 anionic surfactants is provided in Table 1 and was drawn from several sources. The source of each entry is also provided in Table 1. Since the main sources of data were compilations from several primary sources, many of the observations were verified in the original literature. The molecules involved were all anionic surfactants for which sodium was the counter ion. The CMC values used were observed at 40 °C in pure water, or were observed at 25 °C and adjusted to 40 °C using the method described by Huibers, et al. [11]. The logarithm (base 10) of the CMC (mol/L) was as used as the dependent variable in all subsequent modeling work.

Structure entry and preparation

Structures for all 175 surfactant molecules were first assembled as 2D sketches using ChemDraw (version 9.0.1,

Table 1 Identifiers and the observed and computed logCMC values for the 175 surfactants used in model development and testing

Structure ID	Chemical name[a]	Observed $\log_{10}$(CMC)	Modeling set assignment	Computed $\log_{10}$(CMC) ADAPT model	Computed $\log_{10}$(CMC) Molconn model	Source (Ref. #)
Surf_022	Decane-1-sulfonic acid	−1.40	Training	−1.46	−1.61	[8]
Surf_023	Dodecane-1-sulfonic acid	−1.97	Training	−2.00	−2.19	[8, 9]
Surf_024	Hexadecyl hydrogen sulfate	−3.26	Training	−3.20	−3.18	[8, 9, 10]
Surf_025	Hexadecan-4-yl hydrogen sulfate	−2.77	Training	−2.96	−2.90	[8, 9, 10]
Surf_026	Hexadecan-6-yl hydrogen sulfate	−2.63	Training	−2.81	−2.78	[8, 9, 10]
Surf_029	Tetradecyl hydrogen sulfate	−2.64	Training	−2.61	−2.55	[8, 10]
Surf_030	Decyl hydrogen sulfate	−1.48	Training	−1.53	−1.37	[8, 9]
Surf_031	2-Heptylnonyl hydrogen sulfate	−2.52	Training	−2.65	−2.44	[8]
Surf_033	Pentadecyl hydrogen sulfate	−2.92	Training	−2.90	−2.82	[8]
Surf_034	2-Hexyldecyl hydrogen sulfate	−2.64	Training	−2.64	−2.72	[8]
Surf_035	Octane-1-sulfonic acid	−0.79	Training	−0.96	−0.98	[8, 9]
Surf_036	Tetradecane-1-sulfonic acid	−2.60	Ext. Pred.	−2.57	−2.72	[8]
Surf_037	Dodecyl hydrogen sulfate	−2.06	Training	−2.05	−1.93	[8, 9, 10]
Surf_038	Octyl hydrogen sulfate	−0.86	Training	−1.06	−0.78	[8, 9, 10]
Surf_039	2-Ethyltetradecyl hydrogen sulfate	−3.05	Training	−2.90	−2.86	[8]
Surf_040	2-Butyldodecyl hydrogen sulfate	−2.82	Training	−2.70	−2.76	[8]
Surf_041	Tridecyl hydrogen sulfate	−2.37	Training	−2.33	−2.22	[8]
Surf_043	2-Propyltridecyl hydrogen sulfate	−2.96	Training	−2.72	−2.80	[8]
Surf_044	2-Methylpentadecyl hydrogen sulfate	−3.10	Training	−2.91	−2.91	[8]
Surf_045	2-Pentylundecyl hydrogen sulfate	−2.70	Ext. Pred.	−2.79	−2.74	[8]
Surf_055	2,2,3,3,4,4,5,5,6,6,7,7,8,8,8-Pentadecafluorooctanoic acid	−2.01	Outlier[b]	−0.86	−2.04	[9]
Surf_058	2-(2-(Dodecyloxy)ethoxy)ethyl hydrogen sulfate	−2.55	Training	−2.50	−2.51	[11]
Surf_060	Hexane-1-sulfonic acid	−0.50	Training	−0.53	−0.44	[11]
Surf_061	Hexadecane-1-sulfonic acid	−3.13	Training	−3.17	−3.37	[11]
Surf_062	(E)-Dodec-1-ene-1-sulfonic acid	−1.89	Training	−1.85	−1.91	[11]
Surf_063	(E)-Tetradec-1-ene-1-sulfonic acid	−2.57	Training	−2.41	−2.50	[11]
Surf_064	(E)-Hexadec-1-ene-1-sulfonic acid	−3.22	Training	−2.99	−3.10	[11]
Surf_065	(E)-Octadec-1-ene-1-sulfonic acid	−3.75	Training	−3.60	−3.70	[11]
Surf_066	Dodecane-3-sulfonic acid	−1.73	Training	−1.70	−1.86	[11]
Surf_067	Dodecane-4-sulfonic acid	−1.64	Training	−1.60	−1.58	[11]
Surf_068	Dodecane-5-sulfonic acid	−1.55	Training	−1.66	−1.53	[11]
Surf_070	Dodecane-6-sulfonic acid	−1.44	Training	−1.63	−1.42	[11]
Surf_072	4-Heptylbenzenesulfonic acid	−1.58	Training	−1.47	−1.35	[11]
Surf_073	4-Octylbenzenesulfonic acid	−1.91	Training	−1.73	−1.78	[11]
Surf_074	4-(Nonan-3-yl)benzenesulfonic acid	−1.97	Training	−1.86	−1.83	[11, 12]
Surf_075	4-(Decan-2-yl)benzenesulfonic acid	−2.30	Ext. Pred.	−2.18	−2.14	[11, 13]
Surf_076	4-(Decan-3-yl)benzenesulfonic acid	−2.20	Training	−2.15	−2.07	[11, 12, 13]
Surf_077	4-(Decan-5-yl)benzenesulfonic acid	−2.05	Training	−1.98	−2.07	[11, 13]
Surf_078	4-(Undecan-2-yl)benzenesulfonic acid	−2.72	Training	−2.48	−2.44	[11, 14]
Surf_080	4-(Dodecan-3-yl)benzenesulfonic acid	−2.61	Training	−2.77	−2.66	[11, 12]
Surf_082	4-(Dodecan-6-yl)benzenesulfonic acid	−2.59	Training	−2.43	−2.64	[11]
Surf_083	4-(Tridecan-2-yl)benzenesulfonic acid	−3.21	Training	−3.09	−3.03	[11, 14]
Surf_084	4-(Pentadecan-2-yl)benzenesulfonic acid	−3.58	Training	−3.72	−3.60	[11, 14]
Surf_085	Octadecyl hydrogen sulfate	−3.79	Training	−3.80	−3.72	[10, 11]
Surf_086	Tetradecan-3-yl hydrogen sulfate	−2.37	Training	−2.33	−2.34	[9, 10]
Surf_087	Tetradecan-4-yl hydrogen sulfate	−2.29	Training	−2.31	−2.28	[9, 10]

Table 1 continued

Structure ID	Chemical name[a]	Observed $\log_{10}$(CMC)	Modeling set assignment	Computed $\log_{10}$(CMC) ADAPT model	Computed $\log_{10}$(CMC) Molconn model	Source (Ref. #)
Surf_088	Tetradecan-5-yl hydrogen sulfate	−2.17	Training	−2.19	−2.23	[10, 11]
Surf_090	Tetradecan-7-yl hydrogen sulfate	−2.01	Ext. Pred.	−1.90	−2.19	[10, 11]
Surf_091	Pentadecan-3-yl hydrogen sulfate	−2.66	Training	−2.64	−2.66	[10, 11]
Surf_097	Hexadecan-8-yl hydrogen sulfate	−2.37	Training	−2.56	−2.75	[10, 11]
Surf_101	Octadecan-4-yl hydrogen sulfate	−3.35	Ext. Pred.	−3.61	−3.46	[10, 11]
Surf_102	Octadecan-6-yl hydrogen sulfate	−3.14	Training	−3.52	−3.37	[10, 11]
Surf_103	Nonadecan-10-yl hydrogen sulfate	−3.03	Training	−3.59	−3.21	[10, 11]
Surf_104	Nonadecan-5-yl hydrogen sulfate	−3.48	Training	−3.90	−3.70	[10, 11]
Surf_105	Pentadecane-1-sulfonic acid	−3.14	Training	−2.87	−3.05	[11]
Surf_106	Heptadecane-1-sulfonic acid	−3.63	Training	−3.47	−3.65	[11]
Surf_107	Pentadecane-8-sulfonic acid	−2.25	Training	−2.56	−2.03	[15]
Surf_108	4-(Dodecan-2-yl)benzenesulfonic acid	−2.64	Training	−2.78	−2.61	[11, 16]
Surf_109	4-(Dodecan-4-yl)benzenesulfonic acid	−2.72	Training	−2.74	−2.63	[11, 17]
Surf_110	Undecyl hydrogen sulfate	−1.79	Training	−1.78	−1.66	[11]
Surf_112	2-(Dodecyloxy)ethyl hydrogen sulfate	−2.39	Training	−2.27	−2.25	[11]
Surf_113	3-Methoxydodecane-1-sulfonic acid	−2.12	Training	−1.99	−2.05	[11]
Surf_114	3-Ethoxydodecane-1-sulfonic acid	−2.30	Training	−2.12	−2.21	[11]
Surf_115	3-Propoxydodecane-1-sulfonic acid	−2.42	Training	−2.36	−2.44	[11]
Surf_116	3-Isopropoxydodecane-1-sulfonic acid	−2.46	Ext. Pred.	−2.21	−2.32	[11]
Surf_117	3-Butoxydodecane-1-sulfonic acid	−2.82	Training	−2.55	−2.65	[11]
Surf_118	3-(Hexyloxy)dodecane-1-sulfonic acid	−3.19	Training	−2.80	−3.23	[11]
Surf_119	3-(Octyloxy)dodecane-1-sulfonic acid	−3.92	Outlier[b]	−3.01	−3.80	[11]
Surf_120	3-(2-Ethylhexyloxy)dodecane-1-sulfonic acid	−3.50	Training	−3.24	−3.55	[11]
Surf_121	3-Phenoxydodecane-1-sulfonic acid	−2.71	Training	−2.54	−2.64	[11]
Surf_123	3-Oxododecane-1-sulfonic acid	−1.54	Ext. Pred.	−1.82	−1.71	[11]
Surf_124	3-Hydroxydodecane-1-sulfonic acid	−1.61	Training	−1.77	−1.83	[11]
Surf_125	3-Hydroxytetradecane-1-sulfonic acid	−2.20	Training	−2.32	−2.42	[11]
Surf_126	3-(2-Hydroxyethyl)tetradecane-1-sulfonic acid	−3.45	Ext. Pred.	−2.75	−2.83	[11]
Surf_127	3-(2-(2-Hydroxyethoxy)ethoxy)tetradecane-1-sulfonic acid	−2.92	Training	−2.54	−2.74	[11]
Surf_128	3-Phenoxytetradecane-1-sulfonic acid	−3.64	Training	−3.23	−3.25	[11]
Surf_129	3-(2,4,6-Trichlorophenoxy)tetradecane-1-sulfonic acid	−4.79	Outlier[c]	−4.29	−3.04	[11]
Surf_130	3-(Dimethylamino)tetradecane-1-sulfonic acid	−2.97	Training	−2.71	−3.21	[11]
Surf_131	3-(Propylamino)tetradecane-1-sulfonic acid	−3.20	Ext. Pred.	−3.15	−3.39	[11]
Surf_132	3-(Butylamino)tetradecane-1-sulfonic acid	−3.71	Training	−3.40	−3.68	[11]
Surf_133	3-(Morpholino)tetradecane-1-sulfonic acid	−3.11	Ext. Pred.	−2.90	−3.02	[11]
Surf_134	3-(Piperidino)tetradecane-1-sulfonic acid	−3.31	Training	−3.60	−3.56	[11]
Surf_135	3-Oxotetradecane-1-sulfonic acid	−2.17	Training	−2.35	−2.28	[11]
Surf_136	3-Hydroxyhexadecane-1-sulfonic acid	−2.84	Training	−2.90	−3.00	[11]
Surf_137	3-Methoxyhexadecane-1-sulfonic acid	−3.47	Training	−3.17	−3.24	[11]
Surf_138	3-Propoxyhexadecane-1-sulfonic acid	−4.09	Training	−3.64	−3.62	[11]
Surf_139	3-Butoxyhexadecane-1-sulfonic acid	−4.46	Training	−3.90	−3.87	[11]
Surf_140	3-Oxohexadecane-1-sulfonic acid	−2.74	Training	−2.93	−2.86	[11]
Surf_141	3-Hydroxyoctadecane-1-sulfonic acid	−3.42	Training	−3.50	−3.59	[11]
Surf_142	2-(Decyloxy)ethanesulfonic acid	−1.79	Training	−1.65	−1.88	[11]
Surf_143	1-Hydroxytetradecane-2-sulfonic acid	−1.79	Training	−2.17	−2.12	[11]

Table 1 continued

Structure ID	Chemical name[a]	Observed $\log_{10}$(CMC)	Modeling set assignment	Computed $\log_{10}$(CMC) ADAPT model	Computed $\log_{10}$(CMC) Molconn model	Source (Ref. #)
Surf_144	1-Hydroxyhexadecane-2-sulfonic acid	−2.43	Training	−2.77	−2.71	[11]
Surf_145	2-(Hexyloxy)-2-oxoethanesulfonic acid	−0.74	Training	−0.69	−0.64	[11]
Surf_146	2-(Octyloxy)-2-oxoethanesulfonic acid	−1.14	Training	−1.14	−1.18	[11]
Surf_147	2-(Decyloxy)-2-oxoethanesulfonic acid	−1.62	Training	−1.64	−1.74	[11]
Surf_148	3-Oxo-3-(tetradecyloxy)propane-1-sulfonic acid	−3.05	Training	−3.01	−3.04	[11]
Surf_149	1-Methoxy-1-oxotridecane-2-sulfonic acid	−1.99	Training	−2.02	−2.20	[11]
Surf_150	1-Methoxy-1-oxopentadecane-2-sulfonic acid	−2.58	Training	−2.63	−2.78	[11]
Surf_151	1-Methoxy-1-oxoheptadecane-2-sulfonic acid	−3.40	Training	−3.25	−3.37	[11]
Surf_152	1-Ethoxy-1-oxoheptadecane-2-sulfonic acid	−3.51	Training	−3.45	−3.44	[11]
Surf_153	1-Oxo-1-propoxyheptadecane-2-sulfonic acid	−3.97	Ext. Pred.	−3.71	−3.61	[11]
Surf_154	1-Methoxy-1-oxononadecane-2-sulfonic acid	−4.00	Training	−3.87	−3.97	[11]
Surf_155	1-Ethoxy-1-oxononadecane-2-sulfonic acid	−4.11	Training	−4.10	−4.03	[11]
Surf_156	1-Oxo-1-propoxynonadecane-2-sulfonic acid	−4.90	Training	−4.37	−4.20	[11]
Surf_157	1-Isopropoxy-1-oxononadecane-2-sulfonic acid	−4.57	Ext. Pred.	−4.29	−4.23	[11]
Surf_158	1,4-Bis(2-ethylhexyloxy)-1,4-dioxobutane-2-sulfonic acid	−2.57	Training	−2.49	−2.49	[11]
Surf_159	1,4-Dibutoxy-1,4-dioxobutane-2-sulfonic acid	−0.66	Training	−0.88	−0.87	[11]
Surf_160	1,4-Dioxo-1,4-bis(pentyloxy)butane-2-sulfonic acid	−1.24	Training	−1.30	−1.33	[11]
Surf_161	1,4-Bis(hexyloxy)-1,4-dioxobutane-2-sulfonic acid	−1.82	Training	−1.77	−1.82	[11]
Surf_162	1,4-Bis(octyloxy)-1,4-dioxobutane-2-sulfonic acid	−3.13	Training	−2.96	−2.87	[11]
Surf_163	Tetradecan-2-yl hydrogen sulfate	−2.48	Training	−2.38	−2.48	[9, 10, 18]
Surf_164	Pentadecan-2-yl hydrogen sulfate	−2.77	Training	−2.69	−2.78	[10, 18]
Surf_165	Pentadecan-5-yl hydrogen sulfate	−2.47	Training	−2.59	−2.52	[9, 10, 18]
Surf_166	Pentadecan-8-yl hydrogen sulfate	−2.18	Training	−2.56	−2.16	[9, 10,18]
Surf_167	Heptadecan-2-yl hydrogen sulfate	−3.31	Training	−3.31	−3.38	[10, 18]
Surf_168	Heptadecan-9-yl hydrogen sulfate	−2.63	Training	−3.22	−2.68	[10,18]
Surf_169	Octanoic acid	−0.43	Training	−0.11	−0.21	[9]
Surf_170	Octan-2-yl hydrogen sulfate	−0.74	Training	−0.66	−0.76	[9, 10]
Surf_171	Decan-2-yl hydrogen sulfate	−1.31	Training	−1.22	−1.32	[10]
Surf_172	Undecan-3-yl hydrogen sulfate	−1.54	Training	−1.34	−1.50	[9, 10]
Surf_173	Undecan-6-yl hydrogen sulfate	−1.08	Training	−1.28	−1.16	[9, 10]
Surf_174	Tridecan-2-yl hydrogen sulfate	−2.19	Ext. Pred.	−2.09	−2.19	[9, 10]
Surf_175	Tridecan-7-yl hydrogen sulfate	−1.71	Ext. Pred.	−1.90	−1.65	[9, 10]
Surf_176	Dodecane-2-sulfonic acid	−1.83	Training	−1.81	−1.71	[11]
Surf_177	Heptadecan-2-yl hydrogen sulfate	−3.31	Training	−3.31	−3.39	[10, 11]
Surf_178	Octadecan-2-yl hydrogen sulfate	−3.59	Training	−3.62	−3.68	[10, 11]
Surf_179	Nonacosan-15-yl hydrogen sulfate	−4.10	Outlier[b,c]	−7.01	−5.93	[10]
Surf_253	Hexyl hydrogen sulfate	−0.36	Training	−0.64	−0.25	[19]
Surf_254	Heptyl hydrogen sulfate	−0.65	Ext. Pred.	−0.84	−0.51	[19]
Surf_255	Nonyl hydrogen sulfate	−1.21	Ext. Pred.	−1.28	−1.07	[19]
Surf_256	4-Decylbenzenesulfonic acid	−2.51	Training	−2.30	−2.34	[19]
Surf_257	4-Dodecylbenzenesulfonic acid	−2.92	Training	−2.90	−2.97	[19]
Surf_258	4-(Undecan-3-yl)benzenesulfonic acid	−2.50	Training	−2.46	−2.42	[19]
Surf_259	4-(Undecan-4-yl)benzenesulfonic acid	−2.40	Training	−2.41	−2.37	[19]
Surf_260	4-(Undecan-5-yl)benzenesulfonic acid	−2.30	Training	−2.31	−2.36	[19]
Surf_261	4-(Undecan-6-yl)benzenesulfonic acid	−2.25	Training	−2.05	−2.13	[19]
Surf_262	4-(Dodecan-5-yl)benzenesulfonic acid	−2.57	Training	−2.66	−2.65	[19]

Table 1 continued

Structure ID	Chemical name[a]	Observed $\log_{10}$(CMC)	Modeling set assignment	Computed $\log_{10}$(CMC) ADAPT model	Computed $\log_{10}$(CMC) Molconn model	Source (Ref. #)
Surf_263	4-(Tridecan-3-yl)benzenesulfonic acid	−3.00	Ext. Pred.	−3.08	−2.99	[19]
Surf_264	4-(Tridecan-4-yl)benzenesulfonic acid	−2.90	Training	−3.06	−2.96	[19]
Surf_265	4-(Tridecan-5-yl)benzenesulfonic acid	−2.78	Training	−3.01	−2.94	[19]
Surf_266	4-(Tridecan-6-yl)benzenesulfonic acid	−2.70	Training	−2.82	−2.93	[19]
Surf_267	4-(Tridecan-7-yl)benzenesulfonic acid	−2.60	Training	−2.47	−2.67	[19]
Surf_268	4-(Tetradecan-2-yl)benzenesulfonic acid	−3.39	Training	−3.41	−3.33	[19]
Surf_269	4-(Tetradecan-3-yl)benzenesulfonic acid	−3.28	Training	−3.40	−3.25	[19]
Surf_270	4-(Tetradecan-4-yl)benzenesulfonic acid	−3.15	Training	−3.39	−3.26	[19]
Surf_271	4-(Tetradecan-5-yl)benzenesulfonic acid	−3.05	Training	−3.36	−3.24	[19]
Surf_272	4-(Tetradecan-6-yl)benzenesulfonic acid	−2.90	Training	−3.21	−3.23	[19]
Surf_273	4-(Tetradecan-7-yl)benzenesulfonic acid	−2.80	Training	−2.93	−3.22	[19]
Surf_274	2-(2-(Decyloxy)ethoxy)ethyl hydrogen sulfate	−1.93	Training	−1.92	−1.91	[19]
Surf_275	3,6,9,12-Tetraoxatetracosyl hydrogen sulfate	−2.79	Training	−2.98	−2.97	[19]
Surf_276	2-(Tetradecyloxy)ethyl hydrogen sulfate	−2.87	Ext. Pred.	−2.85	−2.85	[19]
Surf_277	2-(2-(Tetradecyloxy)ethoxy)ethyl hydrogen sulfate	−3.09	Training	−3.09	−3.10	[19]
Surf_278	3,6,9,12-Tetraoxahexacosyl hydrogen sulfate	−3.18	Training	−3.59	−3.57	[19]
Surf_279	2-(2-(Hexadecyloxy)ethoxy)ethyl hydrogen sulfate	−3.64	Ext. Pred.	−3.70	−3.71	[19]
Surf_280	3-(Octyloxy)-3-oxopropane-1-sulfonic acid	−1.31	Training	−1.38	−1.28	[19]
Surf_281	3-(Decyloxy)-3-oxopropane-1-sulfonic acid	−1.89	Ext. Pred.	−1.89	−1.87	[19]
Surf_282	3-(Dodecyloxy)-3-oxopropane-1-sulfonic acid	−2.52	Ext. Pred.	−2.44	−2.45	[19]
Surf_283	7,7,8,8,9,9,10,10,10-Nonafluorodecan-4-yl hydrogen sulfate	−1.70	Training	−1.65	−1.68	[19]
Surf_284	7,7,8,8,9,9,10,10,11,11,12,12,12-Tridecafluorododecan-4-yl hydrogen sulfate	−2.50	Training	−2.46	−2.64	[19]
Surf_285	7,7,8,8,9,9,10,10,11,11,12,12,13,13,14,14,14-Heptadecafluorotetradecan-4-yl hydrogen sulfate	−3.42	Training	−3.28	−3.69	[19]
Surf_286	9,9,10,10,11,11,12,12,12-Nonafluorododecan-6-yl hydrogen sulfate	−2.14	Training	−2.14	−1.95	[19]
Surf_287	9,9,10,10,11,11,12,12,13,13,14,14,14-Tridecafluorotetradecan-6-yl hydrogen sulfate	−3.02	Training	−3.04	−2.89	[19]
Surf_288	9,9,10,10,11,11,12,12,13,13,14,14,15,15,16,16,16-Heptadecafluorohexadecan-6-yl hydrogen sulfate	−3.90	Training	−3.90	−3.89	[19]
Surf_289	1,1,1,2,2,3,3,4,4-Nonafluorotetradecan-7-yl hydrogen sulfate	−2.49	Training	−2.68	−2.35	[19]
Surf_290	11,11,12,12,13,13,14,14,15,15,16,16,16-Tridecafluorohexadecan-8-yl hydrogen sulfate	−3.40	Training	−3.59	−3.25	[19]
Surf_291	11,11,12,12,13,13,14,14,15,15,16,16,17,17,18,18,18-Heptadecafluorooctadecan-8-yl hydrogen sulfate	−4.25	Training	−4.48	−4.23	[19]
Surf_292	(5-(Heptylamino)-3,4-dihydroxy-6-methoxytetrahydro-2H-pyran-2-yl)methyl hydrogen sulfate	−1.60	Training	−1.64	−1.63	[19]
Surf_293	(3,4-Dihydroxy-6-methoxy-5-(undecylamino)tetrahydro-2H-pyran-2-yl)methyl hydrogen sulfate	−2.77	Ext. Pred.	−2.76	−2.76	[19]
Surf_294	(3,4-Dihydroxy-6-methoxy-5-(pentadecylamino) tetrahydro-2H-pyran-2-yl)methyl hydrogen sulfate	−3.92	Training	−3.95	−3.93	[19]
Surf_295	(S)-2-(3-Benzyl-4-oxo-4-(pentyloxy)butanoyl) benzenesulfonic acid	−1.82	Training	−1.84	−2.12	[19]
Surf_296	(S)-2-(3-Benzyl-4-(octyloxy)-4-oxobutanoyl) benzenesulfonic acid	−2.82	Training	−2.76	−2.91	[19]

Table 1 continued

Structure ID	Chemical name[a]	Observed $\log_{10}$(CMC)	Modeling set assignment	Computed $\log_{10}$(CMC) ADAPT model	Computed $\log_{10}$(CMC) Molconn model	Source (Ref. #)
Surf_297	(S)-2-(3-Benzyl-4-(decyloxy)-4-oxobutanoyl) benzenesulfonic acid	−3.45	Training	−3.48	−3.48	[19]
Surf_298	(S)-2-(3-Benzyl-4-(dodecyloxy)-4-oxobutanoyl) benzenesulfonic acid	−4.16	Training	−4.19	−4.06	[19]
Surf_299	(S)-2-(3-Benzyl-4-oxo-4-(tetradecyloxy)butanoyl) benzenesulfonic acid	−4.76	Training	−4.90	−4.64	[19]
Surf_300	(S)-2-(4-(Dodecyloxy)-3-methyl-4-oxobutanoyl) benzenesulfonic acid	−3.35	Training	−3.43	−3.33	[19]

Computed values for training set members are the fitted values obtained from the regression analysis. Computed values for the external prediction set members (Ext. Pred.) and the modeling outliers represent the predicted values obtained using the corresponding model. The source of the observed CMC data is also provided

[a] Name generated using ChemFinder, ver-9.0, CambridgeSoft

[b] Outlier from ADAPT data set model development

[c] Outlier from Molconn data set model development

CambridgeSoft), and ChemFinder (version 9.0.1, CambridgeSoft). Since the CMC for all the surfactants were experimentally determined using the sodium salt of the acid, it was not necessary to include the head-group charge as part of modeling step, since it does not change to a significant degree. Thus, all the structures were entered and used as the neutral form of the acid. The structures were saved as a 2D MACCS SDF file. The SDF file was imported into Sybyl (version 7.2, Tripos Associates). Initial 3D conformations were generated using Concord, followed by strain-energy optimization using the Tripos force field including electrostatic terms and a water dielectric. The partial atomic charges needed for the force field calculations were computed using the Gasteiger-Huckel [20] method in Sybyl. The structures were then exported in the form of a Sybyl MOL file for subsequent descriptor calculations.

Descriptor calculations

Two separate sets of descriptors were computed for all 175 structures, each set was used in a separate model-development exercise. One set included 233 topological descriptors computed using MolconnZ (Ver 3.50, Hall Associates) and using the 2D structures from the SDF file. This will be referred to as the *Molconn* set in subsequent work. A second diverse set of 175 descriptors was computed using ADAPT [21, 22]. These descriptors were chosen to capture broad range of topological, geometric, and electronic structural features. The topological descriptors [23] were included to capture detailed information concerning molecular shape and complexity and have the added advantage of being independent of conformation. Additional conformation-independent information was expressed as counts of specific structural fragments (i.e., counts of carbon and heteroatoms, counts of single, double, triple, and aromatic bonds, etc.). Geometric descriptors provide measures of conformation-dependent shape characteristics of structure, such as surface area and volume [24], molecular length, width, and thickness [25, 26] whereas electronic descriptors provide information concerning the distribution of charge in the molecule [27]. Additionally, some descriptors employ structural representations that capture two or more of these structural feature types (e.g., surface area and partial atomic charge). This class of descriptors is represented by the CPSA descriptors [28, 29] and the related hydrophobic surface area (HSA) descriptors [30] that have been shown to be useful in past studies. The partial charges used in the calculation of the CPSA and related descriptors were those obtained using the Gasteiger-Huckel method during the strain-energy optimization step in Sybyl. These descriptors will be referred to as the *ADAPT* set in subsequent work.

Model development and validation

Models for both descriptor sets were generated using the same methods. Model development began with the selection of a subset of the structures to be used as an external prediction set for both descriptor set models. The subset was selected to mimic the distribution of CMC values of the whole data set. This was done by first sorting all 175 observations in order of increasing logCMC. Each of the sorted observations was assigned an integer value sequentially in the range of 1–8. The data set was again sorted in increasing order based on the assigned integers. The set of 22 observations assigned the value 4 were arbitrarily selected to act as the external prediction (test) set. The remaining 153 observations were assigned to the model training set. The descriptors were analyzed in a process

termed *objective feature selection* [31], where descriptors showing little variation (<10% identical values) were set aside. Additionally, remaining descriptors yielding large pair-wise correlation values (Pearson correlation coefficient ≥ 0.93) were also identified, and one descriptor of the pair was set aside. A record of the descriptors set aside by the correlation test was maintained, and these descriptors were reexamined by exchanging correlated descriptors in the models to determine if any of the descriptors held out were more useful. Following descriptor analysis, models were developed using both simulated annealing [32] and genetic algorithm [33] methods. The results of both methods were examined, and models yielding the smallest root mean squared (RMS) error were considered for subsequent analysis. Internal validation statistics used to evaluate models include the overall-F test [34], the partial-F test [35], variance inflation factor or VIF [36], and PLS PRESS test [37]. The fit and residual plots were also visually examined for any evidence of outlying observations or bias in the model. Lastly, the model was subjected an external validation test by predicting the logCMC values for the 22 observations in the external prediction set.

Conformation analysis of selected surfactants

Conformational analysis for selected surfactant structures was carried out using Spartan '04 (Build 124int9e) for Linux. A conformational search was performed using the Monte Carlo method and used the MMFF forcefield with aqueous correction. A limit of 100 unique conformers was obtained for each structure with a strain energy within 10 kcal/mol of the minimum found for each structure. All conformers were exported to Sybyl in order to compute the Gasteiger-Huckel partial atomic charges needed for subsequent use in ADAPT.

Results and discussion

ADAPT descriptor set model

A good quality model was obtained for 150 of the original 153 observations. The other 3 observations were detected as statistical outliers and are discussed separately. The final model used 5 terms, and yielded a very good fit to the observed logCMC values with $R^2 = 0.951$ and $s = 0.201$. The details of the model are provided in Table 2. The model performed well with respect to all the internal validation statistics. In addition, the model also performed very well in prediction of the logCMC values for the 22 external prediction set structures. The correlation of the predicted and observed logCMC values for the external prediction set is shown in Fig. 2 (Pearson correlation coefficient (r) = 0.974). The computed values for both the training and external prediction set structures obtained using the model are provided in Table 1.

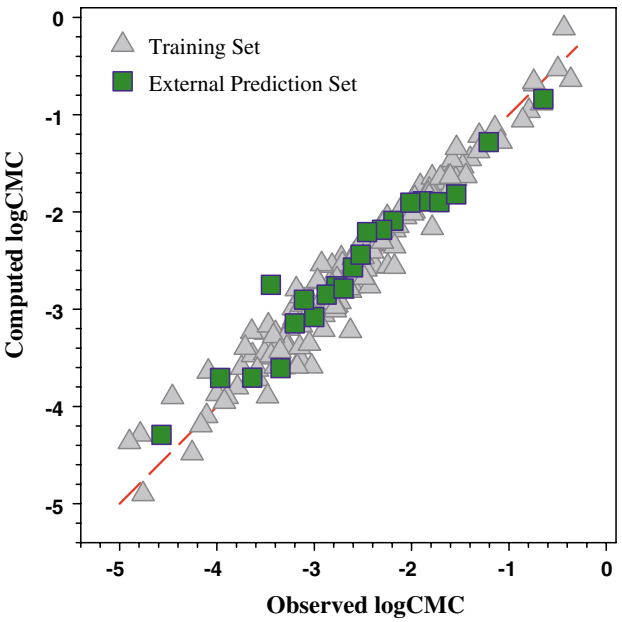

Fig. 2 Results of the prediction of the logCMC for the 22-observation external prediction set based on the ADAPT descriptor set model (r = 0.974). The results for the training set are show as reference

An examination of the model shows that it incorporates a diverse set of descriptors. The RSAM descriptor [38] measures the solvent-accessible surface area of hydrogen-bond acceptor groups in the structure. The MOMH-4 descriptor is geometric descriptor and measures the ratio of the first and the second major moments of inertia of the

Table 2 Details of the model developed using the ADAPT descriptor set

Descriptor	Regression coefficient	Standard error of coefficient	T-value	Variance inflation factor
RSAM	−6.65	0.916	−7.27	3.67
V6P	−0.798	0.0968	−8.25	3.74
MOMH-4	1.60	0.208	7.67	1.45
2SP3	−0.0839	0.00831	−10.1	2.31
FPHS-2	−0.513	0.0224	−22.9	3.58
Y-Intercept	1.40	0.380	3.69	N/A

$R^2 = 0.951$, $Q^2_{LOO} = 0.947$, $s = 0.201$, F-value = 562.17, $N = 150$

structure [26]. The FPHS-2 descriptor is one of the set of HSA descriptors [30]. This particular variant is the type-2 positive hydrophobic surface area descriptor which measures the amount of positive (hydrophobic) solvent-accessible surface area of the structure weighted by the sum of the positive contributions to logP (positive Crippen hydrophobicity fragment constants [39]). The 2SP3 descriptor is a simple count of the occurrences of a sp^3-hybridized carbon atom bonded to two other carbon atoms (i.e., a methylene group). Lastly, the V6P descriptor is the valence-corrected 6th-order path molecular connectivity index [40, 41] which measures characteristics of the structures related to substructure of 6 contiguous bonds. The first three are all conformation-dependent descriptors while the last two are conformation-independent descriptors. As is typically observed, topological descriptors are present in the model and they are found to play a major role in providing measures of the changes in structural features that correlated to differences in the observed property (the structure–property relationship). Results of the analysis of the model using PLS [1] indicates that 82.0% of the variance in the observed logCMC values is accounted for by the first PLS component, and that the V6P descriptor makes the second-largest contribution of information in that component (27.2%) (data not shown). If the two conformation-independent descriptors are considered together, they account for 47.3% of component-1. These descriptors are providing measures of structure related to the length and branching characteristics of the hydrophobic portion of the surfactant molecules that form the core of the micelle, and as previously noted, the length and shape of the hydrophobic group are two of the factors that affect micelle size and shape.

The characteristics of this model provided an opportunity for improvement. As already noted, three of the five descriptors in the model (RSAM, MOMH-4, and FPHS-2) are sensitive to differences in molecular shape. This raises the question of what an appropriate conformation for a surfactant molecule might be. The method employed to generate 3D atomic coordinates in the present case involved using Concord to compute the initial conformation followed by minimization of the strain-energy using molecular mechanics. This generally results in an extended or all-trans configuration for the hydrocarbon chains of the surfactant. While this is certainly a low-energy conformation, it is not clear if such a conformation is relevant for modeling this property. While a micelle is often represented in cartoon form with all the surfactant molecules in an extended conformation as shown Fig. 1, the hydrophobic chains forming the core of the micelle are considered to be disordered with the arrangements of molecules resembling what would be found in bulk hydrocarbon liquid [42]. The shapes of surfactant molecules in a micelle is more clearly illustrated in Fig. 3 which shows the results of a molecular dynamics simulation of a micelle of sodium dodecyl sulfate (SDS) in water (K. Anderson, personal communication, August 13, 2007). Several individual surfactant monomers extracted from the simulated micelle show the variety of conformations that are achieved by the surfactants in the cluster. While some are nearly fully extended, others are highly kinked. Because of the variety of conformations observed in this simulation, it was of interest to determine the degree of sensitivity of the present model to changes in the conformation of different types of surfactants.

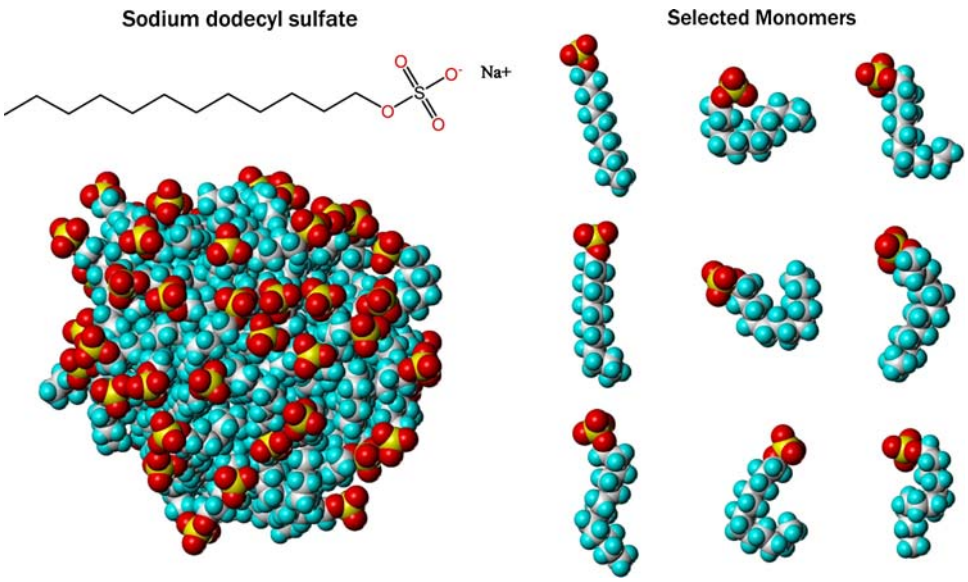

Fig. 3 Image of a micelle formed by sodium dodecyl sulfate in water (water molecules not shown for clarity) based on atomistic molecular dynamics. Several isolated surfactant monomer structures were extracted from the simulated micelle to illustrate the degree of conformation flexibility observed

Surfactant conformation analysis

Four surfactant structures were selected for a conformation analysis. Two linear surfactants (Surf_037 and Surf_257) and two branched surfactants (Surf_070 and Surf_082) were selected that also sampled aromatic and aliphatic head-group types. The conformational search was performed as previously described.

The ADAPT descriptor set model was used to compute the logCMC for all 100 conformers obtained from the conformation search conducted using each of the four test structures. The results of the calculations are shown in Fig. 4. In each case, the computed logCMC values cover a range of at least one order of magnitude. In general, the extended conformers exhibit the lowest strain energy and also yield the lowest computed logCMC values, and the computed values of logCMC increase as the structure becomes more kinked. The results obtained for the conformations of these structures that were used to build the model are also indicated for each test structure in Fig. 4. For three of the four structures the lower energy conformations yield the most accurate computed logCMC values. However, the results for Surf_082 shows that the lowest energy conformers yield a computed logCMC that are about 0.5 log units less accurate than the one used to develop the model. This is a little over twice the magnitude of the standard deviation of regression for the model, making it a significant difference.

While the present model was found to be statistically valid, yielding very good results in external prediction and is based on descriptors that provide an explanation of the underlying structure–property relationship that is consistent with empirical observations, the results of the conformation analysis experiment show that this model can produce a wide range of logCMC values if the method used to generate the 3D atomic coordinates differs from that used to develop the model. In addition to the uncertainty this adds to the predicted logCMC values, it also decreases the confidence of future users of the model who can obtain different computed values for the same structure. Thus, an alternative model was sought that would be independent of the conformation structures involved.

Molconn descriptor set model

Using the same variable selection and model development methods already described, a new model consisting entirely of molecular topology-based, conformation-independent descriptors was developed using the same training and external prediction set selections used to develop the ADAPT set model. A new 7-term model was obtained that yielded similarly good fit to the observed logCMC values ($R^2 = 0.963$, $s = 0.173$). The final training set for the

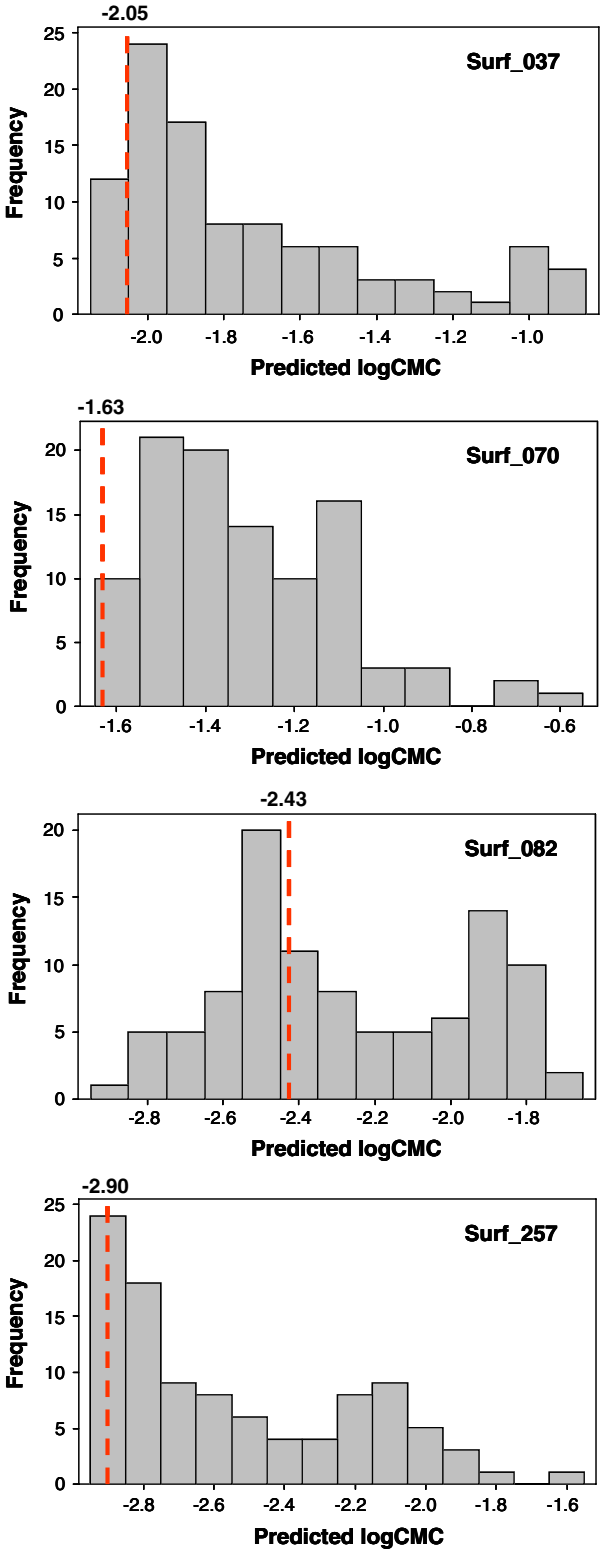

Fig. 4 Bar graphs indicating the range of computed logCMC values obtained using the conformation-sensitive ADAPT descriptor set model for 100 strain energy-optimized conformers for 4 example surfactants: Surf_037, Surf_070, Surf_082, and Surf_257. The vertical dashed line indicates the logCMC value obtained for the conformation of each surfactant used to develop the ADAPT model

model included 151 of the 153 structures originally assigned to the set. The remaining 2 observations were set aside as statistical outliers and are discussed separately. The details of the model are provided in Table 3. The model performed well with respect to all but one of the internal validation statistics. The VIF, a measure of multicollinearity, yielded a value of is 22.2 for descriptor dx0. This is high compared to the general rule of thumb that suggests VIF values should be 10.0 or less. However, our experience has been that if the training set is large ($N > 100$), VIF values in excess of 10.0 can be tolerated without having an adverse effect on either the predictive strength or physical interpretation of the model. This is certainly true for the new model, which performed very well in external prediction. The correlation of the predicted and observed logCMC values for the model is shown in Fig. 5. The computed values of logCMC for the training and prediction set structures are provided in Table 1.

A comparison of the fitted logCMC values for the training set obtained using both models indicates that the two models yield very similar results (Pearson correlation coefficient (r) = 0.980). A similar comparison was made of the external prediction results for the two models which also showed a high degree of correlation of the results (Pearson correlation coefficient (r) = 0.986). The comparisons suggest that the descriptors in the models are equally good at measuring the key changes in molecular structure that are responsible for the differences in the observed logCMC values, and that conformation information is not required to do so.

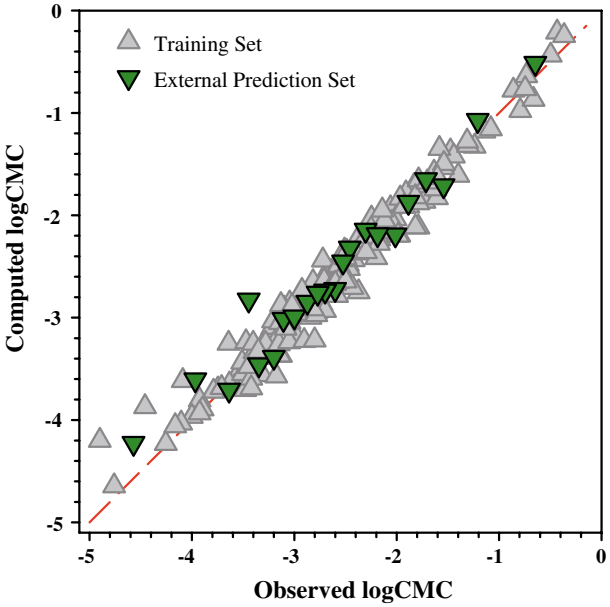

Fig. 5 Results of the prediction of the logCMC for the 22-observation external prediction set based on the Molconn descriptor set model (r = 0.978). The results for the training set are show as reference

Physical interpretation of the *Molconn* model

The definitions for the seven topological descriptors used in the Molconn descriptor set model are provided in Table 4. Physical interpretation of the model was accomplished using the PLS method described previously. The overall results of the PLS analysis are shown in Table 5. While PLS shows that 7 components are validated, 94.1% of the variance in the observed logCMC values is explained in the first 4 components. Thus, interpretation of the model will focus on each of these 4 components in turn. Values of the descriptor weights for each of the first 4 components are provided in Table 6. The squared x-weight values (Table 6b) provide a measure of the contribution of a given descriptor to a component, and the original x-weight values (Table 6a) provide the sign of the weight indicating the direction of the relationship between the descriptor and the dependent property for that component. In order to accomplish the interpretation, it is necessary to examine the PLS score plots and examine the structures of molecules that are the focus of each component with respect to the descriptors that are highly weighted in each component. This provides the details of the structure–property relationships that are captured in the model.

Table 3 Details of the model developed using the Molconn descriptor set

Descriptor	Regression coefficient	Standard error of coefficient	T-value	Variance inflation factor
xvc4	−7.71	1.28	−6.02	6.18
dx0	−1.44	0.105	−13.7	22.2
SssCH2	−0.177	0.00543	−32.6	4.56
SaaCH	−0.123	0.00764	−16.1	3.96
knotpv	−0.103	0.0137	−7.50	7.49
nclass	−0.0436	0.00579	−7.54	3.74
O-Count	0.0864	0.0149	5.80	1.70
Y-Intercept	1.33	0.0898	14.8	N/A

$R^2 = 0.963$, $Q^2_{LOO} = 0.953$, s = 0.173, F-value = 538.59, N = 151

Table 4 Definitions of the descriptors used in the Molconn descriptor set model

Descriptor label	Definition
dx0	0th Order difference molecular connectivity index[a]
xvc4	Valence-corrected 4th-order cluster molecular connectivity index[b]
knotpv	Topological complexity index defined as the difference of xvc3 (valence corrected 3rd-order cluster molecular connectivity index) and xvpc4 (valence-corrected 4th-order path-cluster molecular connectivity index)[c]
SssCH2	Sum of electrotopological state indices for methylene carbons[d]
SaaCH	Sum of electrotopological state indices for unsubstituted aromatic[d] carbons
nclass	Count of the number of types of connectivity classes identified in a structure
O-count	Count of oxygen atoms

[a] See Reference [43]
[b] See References [40, 41]
[c] See Reference [44]
[d] See Reference [45]

Table 5 Summary of the results of the partial least squares (PLS) analysis of the Molconn descriptor set model

Components	X Variance[a]	Error sum of squares	Cumulative R^2	PRESS	Cumulative cross-validated R^2 (Q^2)[b]
1	0.189	25.6	0.781	29.7	0.746
2	0.424	13.7	0.883	13.8	0.882
3	0.792	11.2	0.904	13.4	0.885
4	0.817	6.86	0.941	9.84	0.916
5	0.963	6.45	0.945	9.22	0.921
6	0.975	4.45	0.962	4.91	0.958
7	1.00	4.27	0.963	4.76	0.959

PLS validates all 7 components
[a] Cumulative fraction of X-variance used to explain Y
[b] Computed using a leave-one-out (LOO) cross-validation procedure

Table 6 Details from the PLS analysis of the Molconn data set model. (a) PLS x-weight values for the first 4 components of the Molconn descriptor set model; (b) Squared PLS x-weight values for the first 4 components of the Molconn descriptor set model

Descriptor label	x-Weight component-1	x-Weight component-2	x-Weight component-3	x-Weight component-4
(a) PLS x-weight values				
xvc4	−0.179	−0.015	−0.617	−0.257
dx0	−0.129	0.062	−0.596	−0.213
SssCH2	−0.530	−0.577	0.180	−0.440
SaaCH	−0.185	0.217	0.110	−0.384
knotpv	0.096	−0.115	0.417	−0.360
nclass	−0.790	0.273	0.132	0.461
O-Count	−0.046	0.727	0.165	−0.453
Descriptor Label	Squared x-weight component-1	Squared x-weight component-2	Squared x-weight component-3	Squared x-weight component-4
(b) Squared PLS x-weight values				
xvc4	0.032	0.000	0.381	0.066
dx0	0.017	0.004	0.356	0.045
SssCH2	0.281	0.333	0.033	0.194
SaaCH	0.034	0.047	0.012	0.148
knotpv	0.009	0.013	0.174	0.129
nclass	0.624	0.075	0.017	0.213
O-Count	0.002	0.528	0.027	0.205

Component-1

Component-1 explains 78.1% of the variance in the model and represents by far the most important structure–property trend in the model. Two descriptors are highly weighted in this component. The nclass descriptor contributes 62.4% of the information in this component and takes a negative weight indicating that increases in the value of this descriptor are correlated with a decrease in logCMC. The other important descriptor in this component is SssCH2, which provides an additional 28.1% of the information in this component (90.5% cumulative) and also takes a negative weight indicating an increase in this descriptor value is also correlated with a decrease in logCMC. The nclass descriptor acts in this instance as a measure of the complexity of the structure. Each type of topological substructure is considered a class. For example, a first-order path, a second-order path, and a third-order path are each considered a separate topological class. So, the nclass descriptor is simply a count of the number of types of topological classes identified in each structure. As the size and complexity of the structure increases, the value of nclass increases. The SssCH2 descriptor is one of the electrotopological state descriptors [45] designed specifically to provide a measure of the number of occurrences and environment of methylene ($–CH_2–$) groups. The role of these two descriptors is to show that the length and complexity of the hydrophobic tail groups are the primary structural features that determine the CMC for a molecule. This is illustrated in the score plot for component-1 (plot-A) shown in Fig. 6. Points representing structures that are the focus of this component fall generally on the diagonal of the plot, and are identified as cluster-a. The descriptor values for structures at the upper end of the diagonal have low values for both nclass and SssCH because the structures are shorter and simpler, as shown in Fig. 7a. These compounds have high logCMC values because they disrupt the structure of bulk water less, so higher concentrations are needed before micelles form. Structures represented by points at the lower end of the diagonal have high values for both nclass and SssCH2 because they are much longer and more complex (see Fig. 7b), resulting in lower logCMC values. The length and nature of the hydrophobic tail groups for these molecules cause them to disrupt the solvent structure much more so that association with other surfactant molecules is thermodynamically favored resulting in a much lower critical micelle concentration.

The PLS Y-score for a given structure tends toward zero once the observed property for that observation mathematically explained. The structure–property trend described for component-1 explains the observed property for 93 (61.6%) of the 151 structures in the training set, which form the cluster of points with X and Y-scores

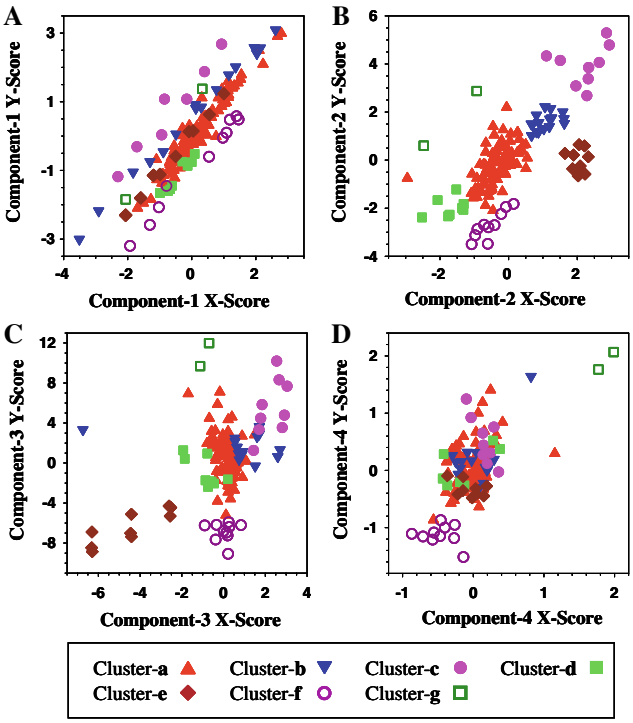

Fig. 6 Score plots for PLS components 1–4 (plots **a–d**, respectively). Points representing the structures that are the focus of the structure–property relationships (SPRs) grouped in clusters a–g

tending toward zero in components 2 through 4 as observed in Fig. 6. This means that there are aspects of molecular structure that are not accounted for in component-1 for the remaining 58 structures that need to be explained. The model accomplishes this in the subsequent components.

Component-2

Component-2 explains an additional 10.2% of the variance in observed logCMC (88.3% cumulative). Two descriptors are highly weighted in this component. Once again, SssCH2 plays an important role, providing 33.3% of the information in the component. However, the primary descriptor is O-count, a simple count of oxygen atoms, which accounts for 52.8% of the information in component-2 (86.1% cumulative). O-count takes a positive weight in this component indicating that increasing values of this descriptor correlate with increases in observed logCMC. The SssCH2 descriptor takes a negative weight indicating, as before, that increasing values of this descriptor are correlated with decreasing values of observed logCMC. The purpose of this trend is to correct for differences in the polar head groups of the surfactants, and the model uses a count of oxygen atoms to measure these differences. The SssCH2 descriptor continues to play the role of accounting for differences in hydrophobic chain

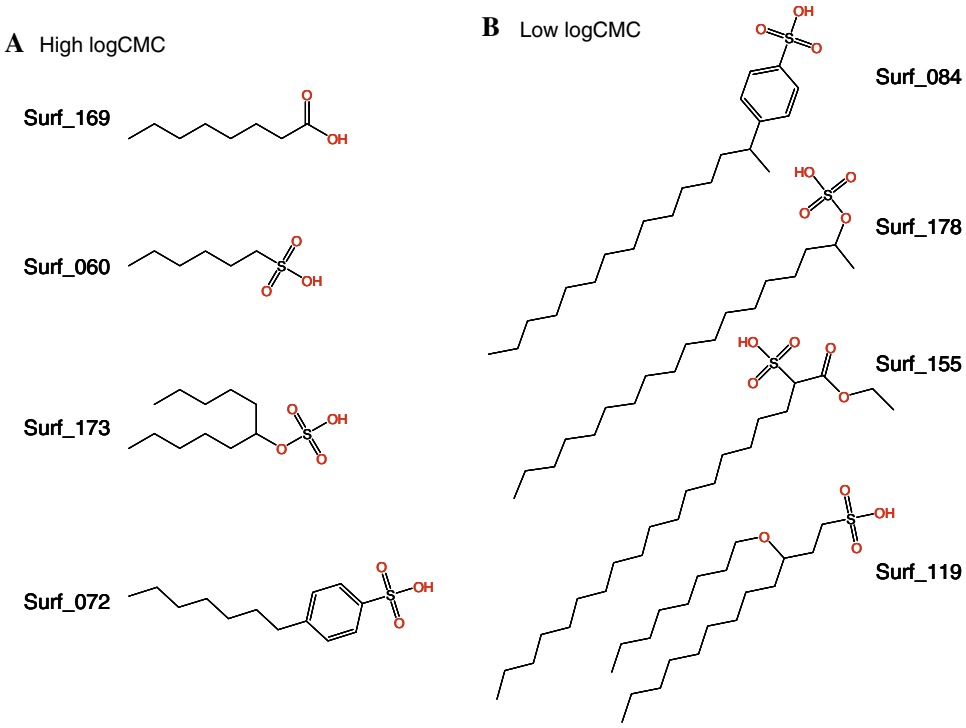

Fig. 7 Example structures for cluster-a in component-1: (**a**) structures from the upper right portion of the diagonal cluster which exhibit higher logCMC values and (**b**) structures from the lower left portion of the diagonal cluster which exhibit lower logCMC values

length, which remains the key factor explaining differences in observed logCMC for a given class of surfactants. The structure–property trend is clearly visible in the score plot for component-2, which is broken down by surfactant type for clarity. The score plot for component-2 (Fig. 6, plot B) shows a cluster of 16 surfactants (cluster-b) representing structures with slightly higher logCMC values than are accounted for by component-1. The structure–property relationship for these materials is parallel to that for the materials explained by component-1, but one aspect of the structure is underdetermined by that trend. The model identifies the difference as being the composition of the polar head group. Structures of some example materials from this cluster are shown in Fig. 8. The polar head groups are much larger and more complex, which pack less well at the surface of the micelle, resulting in a higher observed logCMC. Another set of 9 materials with even larger and more complex polar head groups forms another cluster (cluster-c) in the component-2 score plot. Examples of the structures of these materials are shown in Fig. 9. The polar head groups for these surfactants are very large and sometimes occupy a central position in the molecule, both features lead to an increase in the observed logCMC due to poorer packing of the head groups and to shorter effective length of the hydrophobic tail. The role of the hydrophobic tail group remains the same, once again in parallel with the trend observed in component-1. Another cluster is observed in component-2 (cluster-d) with logCMC values that are at the lower end of the scale. This cluster of points represents 8 structures with very simple and compact polar

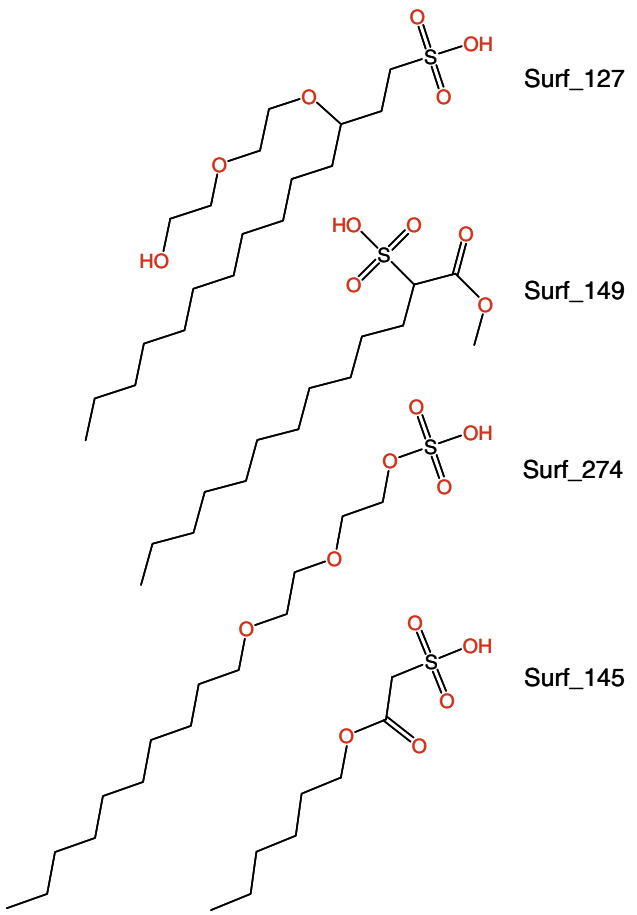

Fig. 8 Example structures representing the surfactants related to cluster-b in the score plot for component-2

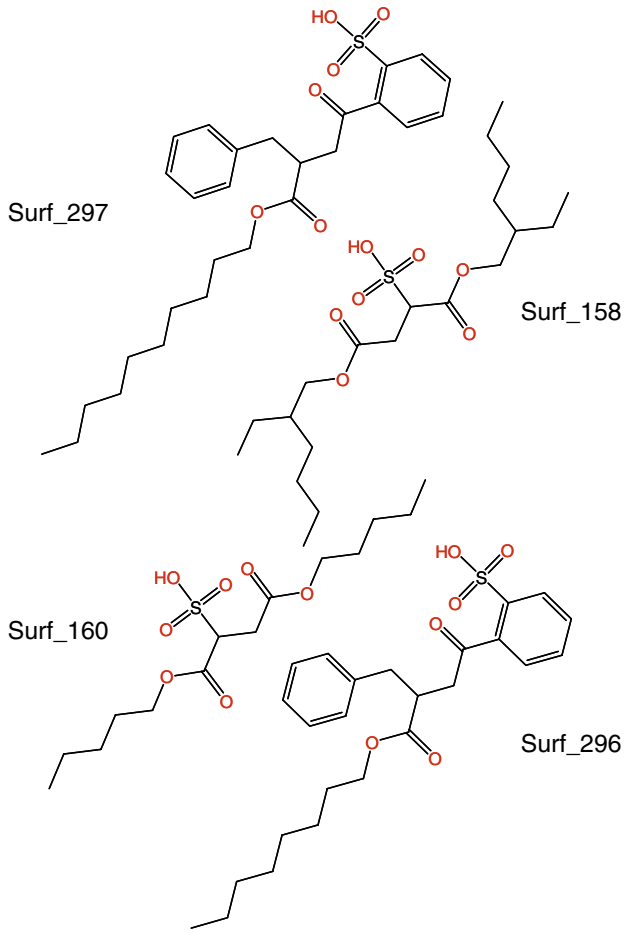

Fig. 9 Example structures representing the surfactants from cluster-c in the score plot for component-2

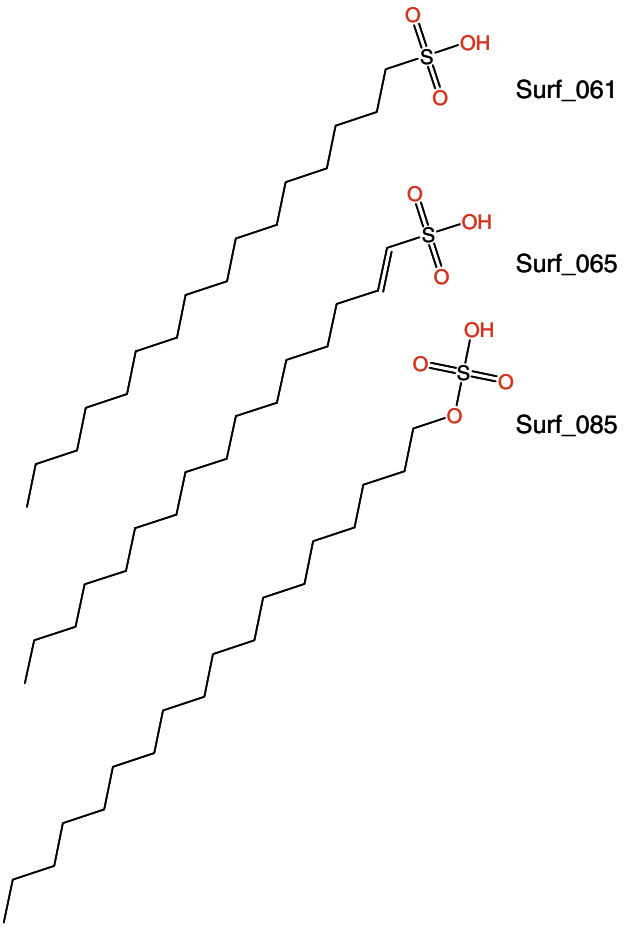

Fig. 10 Example structures representing the surfactants from cluster-d in the score plot for component-2

head groups. Examples of these structures are shown in Fig. 10. This allows greater packing of surfactant molecules in a micelle, resulting in the reduction of the observed logCMC. The role of the length of the hydrophobic tail group parallels the trend observed in component-1. Thus, it is clear that the role of component-2 is to allow the model to account for differences in the nature of the polar head groups for these materials.

Component-3

Component-3 accounts for an additional 2.1% of the variance in the observed logCMC values (90.4% cumulative). While this is a small amount, the model is accounting for an important aspect of the structures of some particular surfactants. Three descriptors provide most of the information for this component. The xvc4 descriptor provides 38.1% of the information, the dx0 descriptor provides an additional 35.6%, and knotpv descriptor provides 17.4% more (91.1% cumulative). The knotpv descriptor takes a positive weight in the component, while the other two take negative PLS weights. In this component, the model is taking into account some unusual features in some of the surfactant structures that have a large affect on observed logCMC. The xvc4 descriptor plays a key role in capturing a difference in the hydrophobic tail groups for one particular set of surfactants. These materials form a cluster of 9 points (cluster-e) that is visible in the component-2 score plot below the diagonal, indicating that the component is over-estimating the logCMC for these materials. This over-estimation is corrected in component-3 indicated by the movement of cluster-e points to the lower left quadrant of the component-3 score plot (Fig. 6, plot C). This correction toward lower logCMC values is primarily due to information provided by the xvc4 descriptor. This descriptor measures the number and environment of an atom that is bonded to four other non-hydrogen atoms, called a 4th-order cluster. This particular version of the descriptor includes a valence correction, indicating the descriptor can discriminate between atom types. An examination of the example structures representing these materials shown in Fig. 11 clearly indicates the key structural feature the

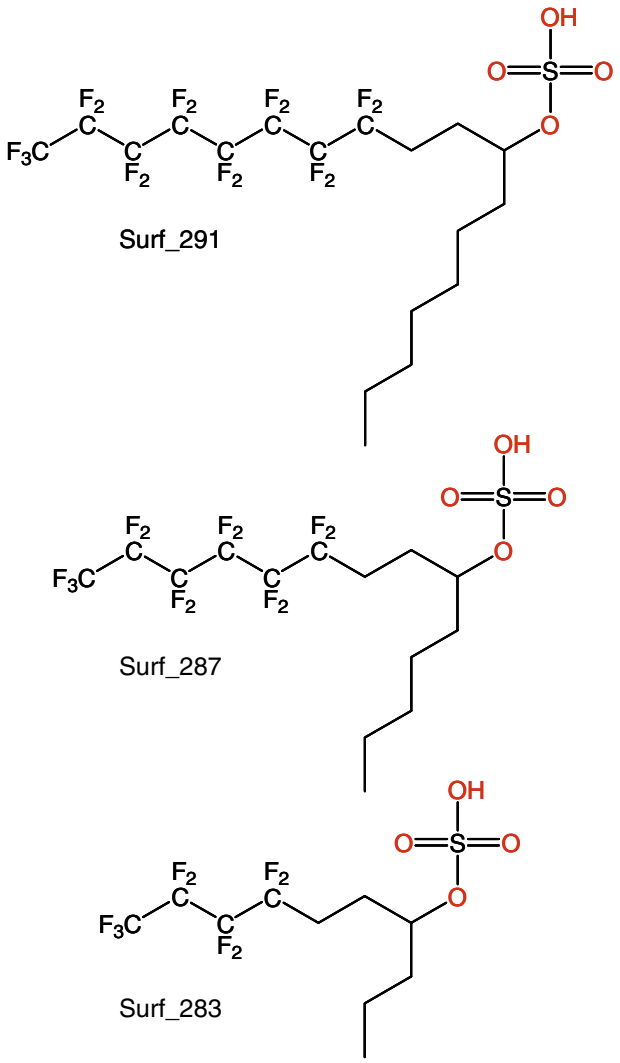

Fig. 11 Example structures representing the surfactants from cluster-e in the score plot for component-3

model is accounting for. All of the surfactants in question contain floromethylene groups in the hydrophobic tail. The topological treatment of molecular structure uses hydrogen-suppressed graphs, so hydrogen atoms are not considered when the counting of attached atoms. Thus the carbon of a methylene group has only two attached atoms, where a fluoromethylene group has four. The xvc4 descriptor is directly indicating the key structural features of the molecule that model needs to account for in capturing this part of the structure–property relationship. A floromethylene group is more hydrophobic than a corresponding methylene group [46]. As a result, the CMC of a perfluormethyl surfactant is similar to that of an ordinary surfactant with a tail group length of 1.5 times the length of that for the perfluoromethyl surfactant [47]. The xcv4 descriptor allows the model to account for this difference in the 9 fluorocarbon surfactants in the presence of the 142 other non-fluorocarbon surfactants in the training set. The model also makes additional corrections for two other types of surfactants in this component. The knotpv and dx0 descriptors measure the special structural features of the polar head groups of several surfactants. These materials are represented in the cluster of 9 points (cluster-c) in the component-3 score plot. Example structures for this cluster are shown in Fig. 12. In one set, the head group is composed of a compact sulfate group and a linear polyoxyethylene chain. The methylene groups in the polyoxyethylene chain do not provide the same degree of hydrophobicity as those in a typical hydrophobic tail group because of the presence of the polar oxygen atoms. This results in an increase in the observed logCMC for surfactants of similar length but containing only non-polar groups. The model uses the dx0 descriptor to help detect and measure this difference. The other set of structures are materials that were a focus of component-2 on the basis of the count of oxygen atoms. That structure–property relationship accounted for the increase in the size and complexity of the polar head group related to oxygen atoms. However, these materials also incorporate a benzyl

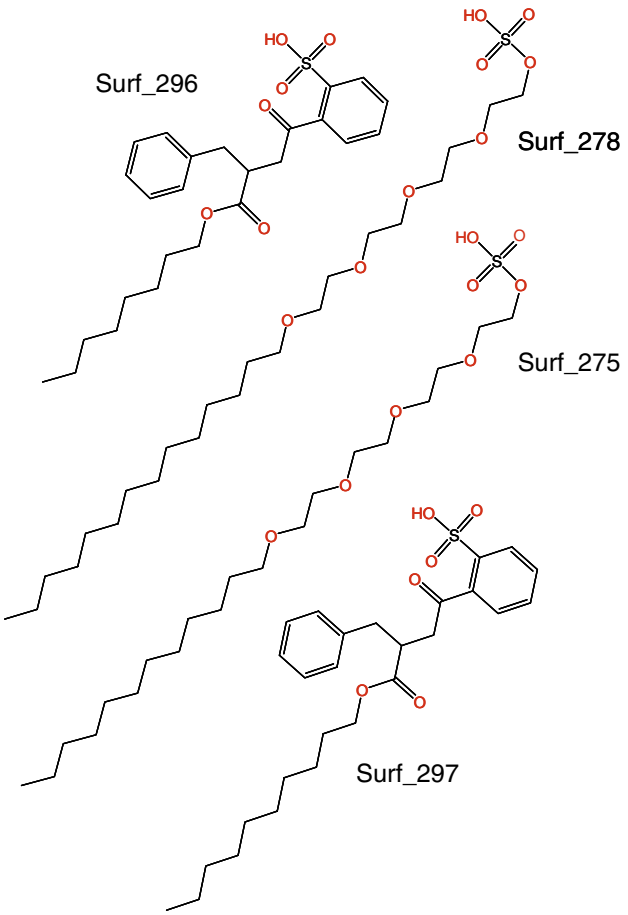

Fig. 12 Example structures representing the surfactants from cluster-c in the score plot for component-3

group in close proximity to the charged head group which increases the steric bulk of the head group resulting in poorer head group packing and an increased logCMC. The knotpv descriptor provides the measure of this feature and allows the model to account for the difference, in addition to the correction made previously for the size of the head group measured using a count of oxygen atoms observed in component-2.

Component-4

This component accounts for an additional 3.7% of the variance in observed logCMC, for a total of 94.1%. As was the case with component-3, the overall contribution to the model is small, but this component captures important information about features of the molecular structure of particular surfactants that have not yet been accounted for. In this case, three descriptors provide the bulk of the information needed. The nclass descriptor provides the largest contribution (21.3%), followed by the O-count descriptor (20.5%), and the SssCH2 descriptor (19.4%). A particularly interesting observation is that the weights for both the nclass and O-count descriptors take opposite signs in component-4 compared to prior components in which they were highly weighted. This type of observation is a unique outcome provided by the use to the PLS analysis that a simple examination of the model regression coefficients could not provide. In this component, the nclass descriptor takes a positive weight which indicates that for this component increasing values of nclass correlate with increasing values of logCMC. The O-count descriptor takes a negative weight in component-4, indicating that increased values on O-count are correlated with decreasing logCMC values. The SssCH2 descriptor also takes a negative weight and performs similarly. Points representing the key surfactants are highlighted in the scope plot for component-4 (Fig. 6, plot D). An addition slight correction is provided by this component for a set of 10 branched surfactants (cluster-f) that have longer hydrophobic tail groups, leading to lower logCMC values than other similarly branched but shorter surfactants (see Fig. 13). Component-4 also provides a correction for the composition of the polar head groups of two unique surfactants that incorporate a pyranose ring. These two materials are identified as cluster-g in the score plot for component-4. The structures of these two materials are provided in Fig. 14. The correction is provided primarily by the nclass descriptor which can account for the large size and complexity of the head group, leading to an increase in the observed logCMC. It is interesting to note that only three examples of this class of surfactant were included in this study. The two shown in Fig. 14 (Surf_292 and Surf_294) were included in the training set for the model, while the

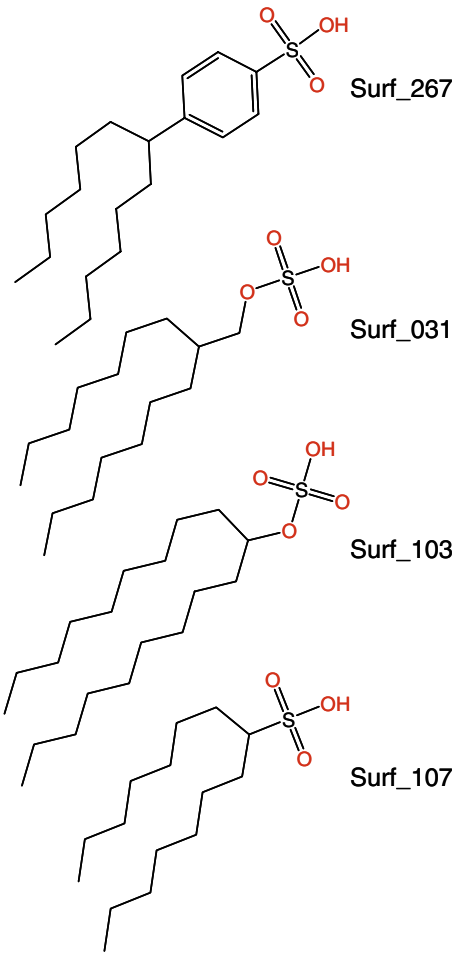

Fig. 13 Example structures representing the surfactants from cluster-f in the score plot for component-4

third had been set aside as part of the external prediction set. The third example (Surf_293) has a hydrophobic chain length of 11 carbon atoms, where the two training set materials had chains containing 7 and 15 carbons atoms. Even though there are only the two examples of this surfactant class in the training set, the model based on the topological descriptors captures so much detail regarding the role of the hydrophobic tail and the polar head group that the prediction error for Surf_293 is only very small with a value of −0.0128 log units.

Examination of the outliers

A small number of observations were detected as statistical outliers during development of both the ADAPT and Molconn models. These observations are identified in Table 1. Identification of the outliers was accomplished either by a simple examination of the residual plots for the models, or using robust regression analysis [48]. A set of 3 outliers were detected during development of the ADAPT descriptor set model, and 2 were detected during

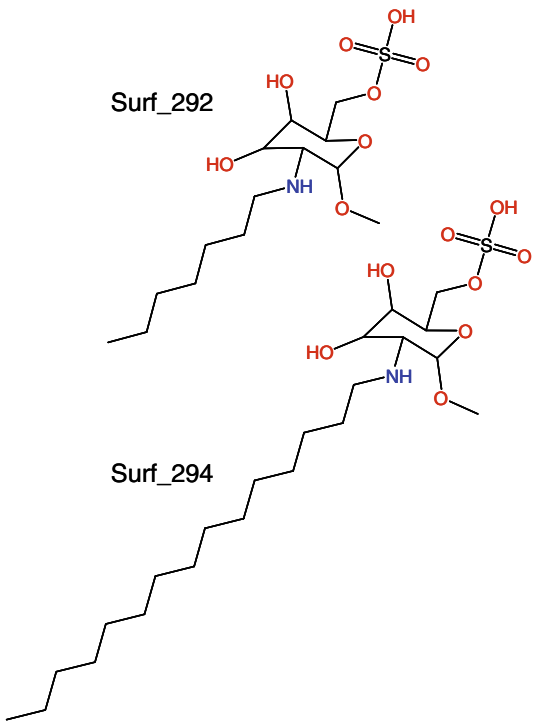

Fig. 14 Example structures representing the surfactants from cluster-g in the score plot for component-4

development of the Molconn model. Only one observation, Surf_179, was found to be an outlier in both analyses. The observed CMC for this surfactant was verified in the original literature and was found in agreement to the reported value. The leverage value [49] for this observation is large in both models, suggesting that this particular observation is significantly different from the other materials in the data set. It is the largest of any of the branched alkyl sulfate surfactants included in this study, with each branch being 14-carbon atoms in length. Since this observation is an outlier in both models and the leverage for this observation is large for both analyses, it is reasonable to conclude that there is some aspect of this structure that is not sufficiently represented if the data set as a whole preventing proper measurement of that feature. Another possibility is that the aqueous solubility of the compound is limited and is interfering with an accurate measurement of the CMC.

Two other outliers were detected during the development of the ADAPT model. One was Surf_055. This surfactant has a high leverage in the model, indicating it is unique compared to the rest of the data set. This particular material is only one of two carboxylic acid surfactants in the data set, and it is the only perfloro example. The other ADAPT data set outlier is Surf_119. This observation also has a large leverage value, suggesting that it is unique in some fashion that the model is unable to account for. This particular material is a branched sulfonic acid surfactant that contains an ether oxygen in one of the branches. The proximity of this oxygen to the head group may be interpreted by the model as making the head group larger, since the computed logCMC is higher than the observed value. This particular observation has a low leverage in the Molconn model, suggesting that the topological descriptors are performing better at capturing information regarding this feature.

There is only one other outlier that is unique to the Molconn model, Surf_129. This material is the only chlorine-containing surfactant in the data set, and it has a high leverage for this model. Thus, in the context of the descriptors in the Molconn model, this material appears to be unique. However, it is not an outlier with respect to the ADAPT model, suggesting the impact of the chlorine atoms on the CMC of this material is appropriately accounted for by the ADAPT model.

Conclusions

This work has clearly illustrated two of the most important characteristics of the general class of topological molecular descriptors in a QSPR application: their independence of the conformation of molecular structure, and the high degree of detail they provide regarding the underlying structure–property relationship. The model based on the topological descriptors has been shown to be as accurate in prediction of logCMC values as the model that included the conformation-dependent descriptors, indicating that the topological descriptors are correctly capturing the important information regarding molecular size and shape of these very flexible molecules. This means that the conventional step of generating 3D atomic coordinates can be eliminated without loss of utility of the model. It also eliminates the need to define which conformation is most important, with the result that the model yields exactly the same logCMC prediction regardless of the way the structure is entered into the computer or how the conformation is optimized.

However, the most important aspect of the topological descriptor-based model is the high degree of structure–property relationship detail it provides. The role of the size and nature of the hydrophobic tail is clearly the dominant factor in determining the CMC. The model shows that CMC is essentially linearly related to chain length over the range examined by this training set, an outcome that is consistent with current knowledge. Long unbranched tail groups yield decreased CMCs, and short chains yield increased CMCs. Structural modifications such as branching of the hydrophobic tail and the addition of fluorine are clearly accounted for. The size and nature of the polar head group is also accurately captured. Smaller and more compact head groups yield decreased CMCs, larger and more complex head groups yield increased CMCs.

A practical way of determining if the structure–property relationship (SPR) derived from a model is correct is to design new structures based on that SPR, and then determine if these new structures behave as predicted. The external prediction set results show the predictive strength of the model. However, it seems clear that one could use the SPR information to successfully modify existing surfactants in such a way that will move the CMC in the desired direction. It is also likely that new classes of surfactants could be designed to have CMC values in a desired range.

These observations regarding the SPR have all been made previously by others, which was the reason that the CMC was selected for this study in the first place. The goal was to show that the topological descriptors do provide proper physically interpretable measures of molecular structure (a SIR) that are useful for molecular design. By using a property that was already generally understood, it was possible to show how the topological descriptors work to capture the key structural information needed to reproduce the same structure–property relationship interpretation. The results also show that a preexisting physical meaning is not required for a descriptor to be useful in a structure–property relationship modeling or molecular design application. We have observed similar results for many other physical and biological properties as well. Thus, this work suggests that as interest in QSAR and QSPR methods is rekindled, special attention should be paid to the inclusion of topological descriptors in such studies.

Acknowledgements The author wishes to thank Dr. M. Lynch of Procter & Gamble for providing access to the Mukerjee and Mysels compilation of CMC data, and also Dr. K. Anderson of Procter & Gamble for providing the result from the molecular dynamics simulation of sodium dodecyl sulfate.

References

1. Stanton DT (2003) J Chem Inf Comput Sci 43:1423
2. Hall LH (2004) Chem Biodiv 1:183
3. Hall LH, Hall LM (2005) SAR QSAR in Environ Res 16:13
4. Kier LB, Hall LH (2005) Chem Biodiv 2:1428
5. Kubinyi H (1993) QSAR: Hansch analysis and related approaches. VCH, New York, pp 50–53
6. Rosen MJ (1989) Surfactants and interfacial phenomena. Wiley, New York, p 108
7. Tanford C (1973) The hydrophobic effect: formation of micelles and biological membranes. Wiley, New York, p 43
8. Rosen MJ (1989) Surfactants and interfacial phenomena. Wiley, New York, pp 116–132
9. Mukerjee P, Mysels KJ (1971) Critical micelle concentrations of aqueous surfactant systems, National Standard Reference Data Service, United States National Bureau of Standards, Washington, DC, pp 51–65
10. Evans HC (1956) J Chem Soc 579
11. Huibers PDT, Lobanov VS, Katritzky AR, Shah DO, Kaeelson M (1997) J Colloid Interface Sci 187:113
12. van Os NM, Daane GJ, Bolsman TABM (1988) J Colloid Interface Sci 123:267
13. van Os NM, Daane GJ, Bolsman TABM (1987) J Colloid Interface Sci 115:402
14. Gershman JW (1957) J Phys Chem 61:581
15. Fenghänel E, Ortman W, Behrmann K, Willscher S (1987) J Phys Chem 91:3700
16. Schick MJ, Fowkes FM (1957) J Phys Chem 61:1062
17. Lianos P, Lang J (1983) J Colloid Interface Sci 96:222
18. Jalali-Heravi M, Konouz E (2000) J Surfactants Deterg 3:47
19. Katrizky AR, Pacureanu L, Dobchev D, Karelson M (2007) J Chem Inf Model 47:782
20. Gasteiger-Huckel partial atomic charges are calculated using the Gasteiger-Marsili method to calculate the σ-electron contributions and the Huckel method for calculating the π-electron contributions, *Sybyl Version 6.3 Force Field Manual*, Tripos, St. Louis, MO, USA, 1996, p 290
21. Stuper AJ, Jurs PC (1976) J Chem Inf Comput Sci 2:99
22. Jurs PC, Chou JT, Yuan M (1979) In: Olson RC, Christoffersen RE (eds) Computer-assisted drug design. American Chemical Society, Washington DC, pp 103–129
23. Ivanciuc O, Balaban AT (1999) In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, The Netherlands, pp 59–167
24. Pearlman RS (1980) In: Yalkowsky SH, Sinkula AA, Valvani SC (eds) Physical chemical properties of drugs. Marcel Dekker, New York
25. Brugger WE, Stuper AJ, Jurs PC (1976) J Chem Inf Comput Sci 16:105
26. Todeschini R, Consonni V (2000) In: Mannhold R, Kubinyi H, Timmerman H (eds) Handbook of molecular descriptors. Wiley-VCH, Weinheim, Federal Republic of Germany, p 352
27. Dixon SL, Jurs PC (1992) J Comput Chem 13:492
28. Stanton DT, Jurs PC (1990) Anal Chem 62:2323
29. Stanton DT, Dimitrov S, Grancharov V, Mekenyan OG (2002) SAR QSAR Environ Res 13:341
30. Stanton DT, Mattioni B, Knittel JJ, Jurs PC (2004) J Chem Inf Comput Sci 44:1010
31. Stanton DT (2000) J Chem Inf Comput Sci 40:81
32. Sutter JM, Jurs PC (1995) Data Handl Sci Tech 15:111
33. Luke BT (1996) In: Devillers J (ed) Genetic algorithms in molecular modeling. Academic Press, New York NY, p 35–66
34. Kutner MH, Nachtshein CJ, Neter J, Li W (2005) Applied linear statistical models, 5th edn. McGraw-Hill Irwin, New York, p 266
35. Kutner MH, Nachtshein CJ, Neter J, Li W (2005) Applied linear statistical models, 5th edn. McGraw-Hill Irwin, New York, p 268
36. Kutner MH, Nachtshein CJ, Neter J, Li W (2005) Applied linear statistical models, 5th edn. McGraw-Hill Irwin, New York, pp 408–410
37. Geladi P, Kowalski BR (1986) Anal Chim Acta 185:1
38. Stanton DT, Egolf LM, Jurs PC (1992) J Chem Inf Comput Sci 32:306
39. Wildman SA, Crippen GM (1999) J Chem Inf Comput Sci 39:868
40. Kier LB, Hall LH (1976) Molecular connectivity in chemistry and drug research. Academic, New York
41. Kier LB, Hall LH (1986) Molecular connectivity in structure–activity analysis. Wiley, New York
42. Tanford C (1973) The hydrophobic effect: formation of micelles and biological membranes. Wiley, New York, p 36
43. Kier LB, Hall LH (1991) Quant Struct-Act Relat 10:134
44. Hall LH, Kellogg GE, Molconn-Z 3.50 Users Guide, EduSoft, 1999, Appendix II. Retrieved from http://www.edusoft-lc.com/molconn/manuals/350/appII.html, 30/9/2007

45. Kier LB, Hall LH (1999) Molecular structure description: the electrotopological state. Academic Press, London
46. Lin IJ, Moudgil BM, Somasundaran P (1974) Colloid Polym Sci 252:407
47. Shinoda K, Hato M, Hayashi T (1972) J Phys Chem 76:909
48. Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley, New York
49. Kutner MH, Nachtshein CJ, Neter J, Li W (2005) Applied linear statistical models, 5th edn. McGraw-Hill Irwin, New York, pp 398–400