PAPER

Check for updates

Predicting the aggregation number of cationic surfactants based on ANN-QSAR modeling approaches: understanding the impact of molecular descriptors on aggregation numbers†

Behnaz Abdous, S. Maryam Sajjadi [iD] * and Ahmad Bagheri

In this work, a quantitative structure–activity relationship (QSAR) study is performed on some cationic surfactants to evaluate the relationship between the molecular structures of the compounds with their aggregation numbers (AGGNs) in aqueous solution at 25 °C. An artificial neural network (ANN) model is combined with the QSAR study to predict the aggregation number of the surfactants. In the ANN analysis, four out of more than 3000 molecular descriptors were used as input variables, and the complete set of 41 cationic surfactants was randomly divided into a training set of 29, a test set of 6, and a validation set of 6 molecules. After that, a multiple linear regression (MLR) analysis was utilized to build a linear model using the same descriptors and the results were compared statistically with those of the ANN analysis. The square of the correlation coefficient (R^2) and root mean square error (RMSE) of the ANN and MLR models (for the whole data set) were 0.9392, 7.84, and 0.5010, 22.52, respectively. The results of the comparison revealed the efficiency of ANN in detecting a correlation between the molecular structure of surfactants and their AGGN values with a high predictive power due to the non-linearity in the studied data. Based on the ANN algorithm, the relative importance of the selected descriptors was computed and arranged in the following descending order: H-047 > ESpm12x > JGI6> Mor20p. Then, the QSAR data was interpreted and the impact of each descriptor on the AGGNs of the molecules were thoroughly discussed. The results showed there is a correlation between each selected descriptor and the AGGN values of the surfactants.

1. Introduction

Surfactants are among the most versatile chemical products and are widely used in the manufacture of cosmetics, detergents, pharmaceuticals, and in the textile industry, and so on.[1] These materials have two main parts: a hydrophilic group (polar head) and a hydrophobic group (hydrocarbon chain). Based on the nature of the polar head, surfactants can be classified as: anionic, cationic, zwitterionic and non-ionic. Indeed, the amphiphilic structure of surfactants makes them highly suitable for surface activity. Among the surfactants, cationic molecules offer some additional advantages over the others. They show antibacterial properties apart from their surface, a fact which makes them applicable in the synthesis of cationic softeners, retarding agents, lubricants, and in some cases in consumer uses.[2–5]

The solution behaviors of cationic surfactants are commonly estimated using critical micelle concentration (CMC), aggregation number (AGGN) and degree of counter ion binding (α). The AGGN is the average number of surfactant molecules in a micelle unit and practically, the increase in AGGN leads to the formation of micelles which show great potential for use in many applications.[6] For example, micelles with a greater AGGN have a greater capacity to transfer a drug in drug delivery systems or remove hydrocarbon contaminants in wastewater treatment processes.[7] Therefore, measuring and establishing a AGGN is very significant.

There are versatile techniques to determine the AGGN of amphiphilic compounds including stepwise thinning of foam films,[8] freezing point and vapor pressure methods,[9] NMR spectroscopy,[10] static light scattering,[11] small-angle neutron scattering,[12] small angle X-ray scattering,[13] fluorescence probing methods,[14,15] and electron paramagnetic resonance.[16] Some of these are only applicable for AGGN determination at a surfactant concentration equal to CMC which only estimates the micelle AGGN for isolated non-interacting particles. In particular, the static light scattering method determines the AGGN values by calculating the molecular weight of the surfactant

Faculty of Chemistry, Semnan University, Semnan, Iran. E-mail: sajjadi@semnan.ac. ir; Fax: +98-23-33384110; Tel: +98-23-31533192

aggregate at the surfactant CMC. The static light scattering technique is rather complicated as it needs to determine the refractive index increment of the measured surfactant solution independently, and extrapolate the data to the CMC which does not let measuring the concentration dependence of the aggregation number. The small angle neutron scattering method allows the determination of the average micelle AGGN[12] as well as providing information on the micelle shape. However, this technique is not easily available for a routine determination of micelle AGGN because of its complexity and the high cost of the neutron scattering experimental facilities.

Fluorescence probing strategies are commonly applied to estimate micelle AGGN where the estimation is influenced by neither the micellar shape nor by the interactions between the micelles. There are two types of fluorescence strategies: time-resolved fluorescence quenching (TRFQ) and a steady-state fluorescence method. The TRFQ technique calculates the micelle AGGN easily and accurately from the fluorescence decay curves.[17] The steady-state fluorescence measurement has the benefits of conventional spectrophotometers, but it needs the application of single-photon counting equipment, and analysis by suitable non-linear fitting algorithms.

Overall, the fluorescent technique possesses the following advantages over the others: (i) it allows the quantification of AGGN at a given surfactant concentration, and in the presence of additives, (ii) it is not influenced by the phenomena of preferential adsorption, which greatly complicates the interpretation of the results, and (iii) it is applicable to all types of surfactants.[14,15]

The AGGNs of a large number of surfactants have been reported in the literature, based on fluorescence strategy, as these large volumes of data can be combined with modelling techniques to interpret the results, and can even be used to predict the AGGNs of new surfactants. Over the last 50 years, chemometrics has developed a powerful set of multivariate data modeling tools to help the "owner" of data find, plot and interpret statistically reliable patterns of data, and obtain maximal information from the studied system with minimal experimental effort.[18,19]

The QSAR modeling is one of the most versatile computational techniques for predicting the physical and biological properties of molecules, developed over the past decades. This technique has been widely recognized in a variety of fields such as medicinal chemistry, pharmacy, toxicology and material science.[20–22] In fact, QSAR modelling can be used to find the relationship between the structure of chemical compounds, and their physical or biological properties to estimate the properties of new chemical compounds without the need for synthesis and testing. In QSAR analysis studies, molecular descriptors are numerical indices assigned to the molecular structure, and encode some information about the structure. Descriptors are theoretical indices which are computed by mathematical formulas or computational algorithms. The Dragon software is one of best, for finding the descriptors of a molecular structure; and it introduces a large variety of descriptors such as constitutional, topological and 3D-MoRSE

descriptors, walk and path counts, and functional group counts.[23–28]

The predictive ability of a QSAR model is affected by the modeling techniques employed to find the mathematical model between the descriptors and their molecular activities. Basically, there are two general modelling methods used to analyze chemical science data, linear and non-linear. Linear approaches include MLR,[29] principal component regression (PCR)[30] and partial least-squares regression (PLS).[31] Non-linear approaches include ANN,[32–35] the support vector machine algorithm,[36] the self-organizing map (SOM),[37] radial basis functions neural networks (RBF),[38] and multivariate adaptive regression splines.[39]

The ANN methods are known as non-linear learning math systems which construct a mapping of the input and output variables, and then the map is used to predict an unknown output as a function of suitable inputs.[32,40–42] The main advantage of ANNs is that they can combine and incorporate both literature-based and experimental data to solve different problems such as predicting the toxicological and physical properties of surfactants.[43–45]

So far as is known, there is no report on predicting AGGNs of surfactants using linear or non-linear modeling techniques, and here, the non-linear ANN algorithm is proposed as a promising technique for this. A data set including 41 surfactant molecules was selected as a target study, and Dragon software was employed to compute the molecular descriptors of the surfactants and their experimental AGGNs were taken from previously published papers.[46–59]

In this study, firstly, the QSAR analyses of the surfactants were performed using MLR and ANN methods to compare the results of linear and non-linear models, and it was shown that the non-linear ANN model could find a satisfactory relationship between molecular descriptors and their AGGNs. Secondly, because the lengths of the hydrophobic group and polar head group are two important factors which strongly affect the AGGN,[55,56] an explanatory study was conducted to interpret the impact of these factors on AGGNs based on the selected descriptor values such as H-047, ESpm12x, JGI6 and Mor20p.

2. Molecular database and software

In this study, a data set including 41 surfactant molecules was used and these structures are shown in Table 1. We took two points into account when collecting this data set: firstly, all the AGGNs of the molecules were obtained using the same strategy and experimental conditions. They were estimated by fluorescence decay curves of a micelle-solubilized pyrene method in aqueous solutions at 25 °C.[46–59] Secondly, we struggled to collect the surfactant molecules with the same polar head group but a different length of the hydrophobic groups such as the following sets: ($C_{16}TAB$, $C_{14}TAB$, $C_{12}TAB$, $C_{10}TAB$, C_6TAB); and ([$C_{16}MIM$][Br], [$C_{14}MIM$][Br], [$C_{12}MIM$][Br], [$C_{10}MIM$][Br], [C_9MIM][Br]); and ($C_{16}E_2TAB$, $C_{14}E_2TAB$, $C_{12}E_2TAB$, $C_{10}E_2TAB$); and ([BisDec(MIM)$_2$][2Br], [BisOct(MIM)$_2$][2Br], [BisHex(MIM)$_2$] [2Br]). However, the structure of some molecules was only different in their polar head groups such as ($C_{16}TAB$, [$C_{16}MIM$]

RSC Advances

Table 1   The molecular structure and AGGN value of each cationic surfactant used in the ANN–QSAR studies

No.	Symbol	Molecular structure	Predicted AGGN	Experimental AGGN	Set of data	Ref.
1	m-X-3		15.06	16	Training	46
2	EO-2		30.99	31	Validation	46
3	t-B-2		29.16	31	Training	46
4	o-X-2		25.02	25	Test	46
5	BDDAC $C_{21}H_{38}ClN$		25.40	27	Training	46
6	[BisDec(MIM)$_2$] [2Br]		65.82	70	Training	47
7	[BisOct(MIM)$_2$] [2Br]		36.68	39	Training	47
8	[BisHex(MIM)$_2$] [2Br]		15.99	16	Validation	47
9	ValC3LS		76.99	77	Validation	48
10	ProC3LS		41.38	44	Training	48

Table 1 (*Contd.*)
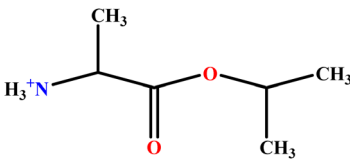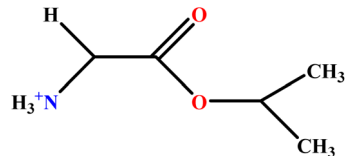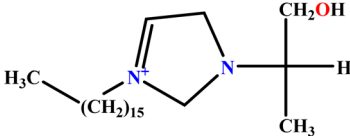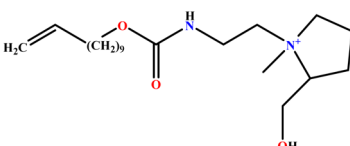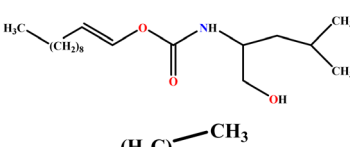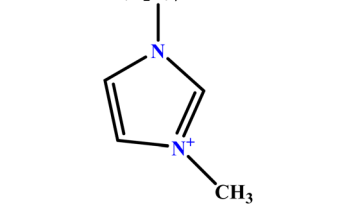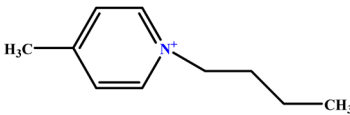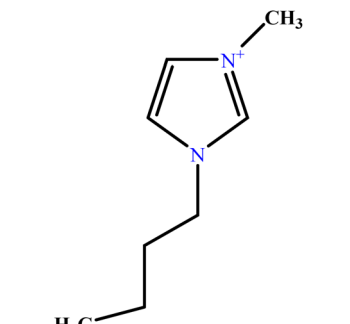
No.	Symbol	Molecular structure	Predicted AGGN	Experimental AGGN	Set of data	Ref.
11	AlaC3LS		76.16	81	Training	48
12	GlyC3LS		94.02	94	Test	48
13	[C$_{16}$hpim]Br		23.52	25	Training	49
14	L-UCPB		89.32	95	Training	50
15	LUCLB		91.20	97	Training	50
16	[C$_8$mim][Cl]		21.64	23	Training	51
17	[C$_4$mpy][Cl]		12.24	13	Training	51
18	[C$_4$mim][Cl]		7.99	8	Validation	51

Table 1 (Contd.)
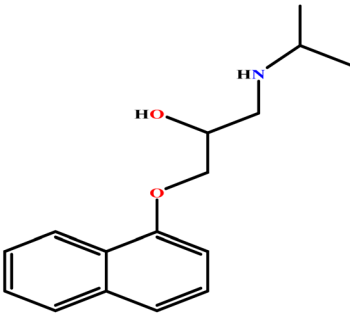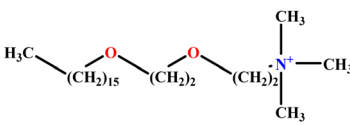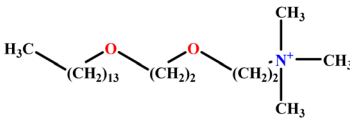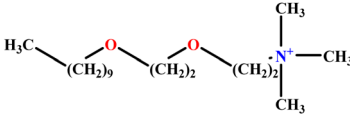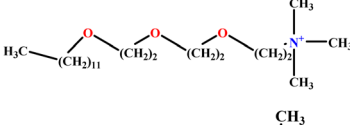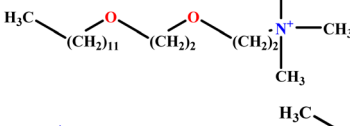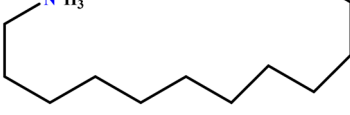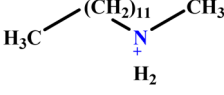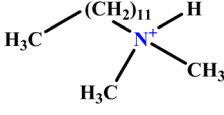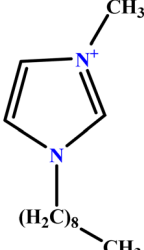
No.	Symbol	Molecular structure	Predicted AGGN	Experimental AGGN	Set of data	Ref.
19	PH		9.41	10	Training	52
20	$C_{16}E_2TAB$		30.10	32	Training	53
21	$C_{14}E_2TAB$		21.64	23	Training	53
22	$C_{10}E_2TAB$		9.02	9	Test	53
23	$C_{12}E_3TAB$		15.06	16	Training	53
24	$C_{12}E_2TAB$		20.70	22	Training	53
25	DAC		107.99	108	Validation	54
26	DMAC		89.02	89	Test	54
27	DDMAC		62.06	66	Training	54
28	[C_9MIM][Br]		42.32	45	Training	56

Table 1 (Contd.)

No.	Symbol	Molecular structure	Predicted AGGN	Experimental AGGN	Set of data	Ref.
29	BHDC		45.14	48	Training	57
30	C$_{12}$DAB		60.02	60	Test	58
31	C$_{16}$TAB		89.32	95	Training	59
32	C$_{14}$TAB		63.94	68	Training	59
33	C$_{12}$TAB		53.60	57	Training	59
34	C$_{10}$TAB		36.68	39	Training	59
35	C$_6$TAB		3.78	4	Training	52
36	CTAC		106.24	113	Training	54
37	[C$_{16}$MIM][Br]		93.08	99	Training	56

Table 1 (*Contd.*)

No.	Symbol	Molecular structure	Predicted AGGN	Experimental AGGN	Set of data	Ref.
38	[C$_{14}$MIM][Br]		74.28	79	Training	56
39	[C$_{12}$MIM][Br]		58.02	58	Test	56
40	[C$_{10}$MIM][Br]		39.99	40	Validation	56
41	CPC		48.90	52	Training	59

[Br], CPC); and (C$_{12}$TAB, [C$_{12}$MIM][Br], C$_{12}$DAB, DDMAC, DAC, DMAC).

2.1 Molecular modeling

The structure of each compound was drawn with GaussView 5.0.8 (Table 1), and optimized by the semi-empirical method, PM6, available in the Gaussian 09 software package,[60] and the optimization goal was to achieve the structures with the lowest energy level. Due to the space limitation, the optimized structure of the molecules are shown in Table S1 (ESI).† Because the molecules were large, we preferred to use semi-empirical PM6 for optimization purposes rather than the density functional theory (DFT) as a quantum mechanical method. Indeed, PM6 can be employed for systems with thousands of atoms while retaining the benefits of the DFT calculations: they are based on a proper physical description of the molecular structure and do not depend on system-specific parameters.[61] Moreover, the computational speed of the PM6 method is more rapid than that of DFT.

Finally, for each optimized structure, the molecular descriptors were computed using the Dragon 5.5-2007 software designed as a user-friendly software.[62] In this software, descriptor calculations are conducted according to these simple steps: firstly, the molecular file obtained from Gaussian is loaded; secondly, the descriptors are selected; thirdly, the descriptors are computed; and fourthly, the calculated descriptors are saved. In this study, the QSAR data obtained were collected in an Excel file (see ESI† for further information). All the calculations were conducted in MATLAB, version 7 (Math Works), and the ANN was performed using the MATLAB Neural Network Toolbox.[63]

3. Artificial neural network

The ANNs are computer programs inspired by the human brain. They have been designed to simulate the processing information in the brain and are widely used in different branches of science such as analytical, physical, organic, inorganic chemistries, and medicinal material sciences.[43–45,64,65] The ANNs obtain their knowledge by finding the patterns and relationships in data through experience.[66] They are made of artificial neurons which are connected with coefficients (weights), constituting the neural structure and organized in layers.

In ANNs, each neutron possesses weighted inputs, transfer function and one output. The behavior of an ANN depends on the transfer functions of its neurons, the learning rule, and the architecture itself. The signal of the neuron is established by the weighed sum of the inputs and passed through the transfer function to create a single output of the neuron. The role of the transfer function is to introduce non-linearity to the network. The ANN algorithm is a two-step processing technique, involving training and validation steps. During training, the weights are optimized until the prediction error

is minimized, and the network gains an acceptable level of accuracy. When the network is trained and tested, it can be applied for predicting the output using new input information.[67]

A variety of types of ANNS have been designed up to now, however, the majority of today's applications apply back-propagation feed-forward ANN (BPFF-ANN).[67] This network consists of at least three layers including input, hidden and output layers. The first one is the input layer which simply serves to enter the input variables, which are the selected descriptors in this investigation. The output layer is the last one where the output variables are handled, here, the number of nodes of this layer is set to one assigning the AGGN of each surfactant. The layers between the input and output ones are called hidden layers, each of which may function independently and may transfer its results to the other one. The most crucial step in designing the ANN is optimizing the number of nodes in the hidden layer, apart from adjusting the weights, as described in the following section.

4. Optimization of ANN

In an ANN algorithm, there is a connector neuron between an input and a hidden layer, as well as between the neurons and the output layer, called the weight (W_{ij}), which represents the "artificial synapses".[68] The input signals (In) are processed in the "body" of the neuron as follows:

$$Z_j = \sum_i^n W_{ij} A_i \tag{1}$$

where Z_j and W_{ij} are the values of the jth hidden neuron and the weight linking the ith input neuron to the jth hidden neuron, respectively. A_j is the value of the ith input neuron, which is a normalized value of the ith independent variable.

In ANN analysis, each variable (input or output values) is rescaled to a new range of values between −1 to +1 as follows:[69]

$$A_i = \frac{X_i - X_{min}}{X_{max} - X_{min}} \times (r_{max} - Xr_{min}) + r_{min} \tag{2}$$

where X_i is i^{th} real variable, A_i is the normalized value of X_i, X_{min} and X_{max} are the minimum and maximum values of X_i, respectively, and r_{min} and r_{max} are attributed to the limits of the range where X_i should be scaled.

In the ANN algorithm, the initialization is conducted with random weights and a different initialization is done to diminish the probability of a convergence to a local minimum. The total data is divided into three sets: training, test and validation. The training set is employed to adjust the weight factors on the ANN, and the test set is used to overcome the over-fitting problem and to find the optimal number of neurons in the hidden layer. The validation set is applied to confirm the actual predictive power of the ANN.

In the BPFF-ANN algorithm, the weights change during each iteration with the aim of minimizing the difference between the actual outputs and the model predicted ones, and the change of each weight can be written as:

$$\Delta W_{ij} + W_{ij} \rightarrow W_{ij}$$

$$\Delta W_{ij} = \eta(t - o)\text{In}_i \tag{3}$$

where, for each sample, t and o are the target and the output value of ANN, respectively, and, η is the learning factor whose role is controlling the amount of weight change at each iteration. The value of η is usually small (*e.g.*, 0.1) and it diminishes, and would have less and less effect as the number of iterations increases.

In the ANN studies, models with fewer variables result in diminishing the complexity of the analysis, preventing overfitting/overtraining and reducing the computational time and improving the prediction power for new samples. Here, firstly, the descriptors of the surfactant molecules with zero values were omitted, and then the descriptors showing a high correlation coefficient with each other were eliminated, and finally based on stepwise regression analysis, four significant descriptors were selected for further analysis (Table 2). These variables had high correlation with the response and less correlation with each other.

The four selected descriptors (Table 2) were applied as input neurons in the ANN modeling, and the AGGN of the surfactant molecule was considered as a neuron in the output layer. The number of hidden layers and their neurons were chosen by optimizing the model in the ANN-Matlab toolbox (Matlab *nntool*) using a BPFF-ANN algorithm. The important network parameters in the toolbox such as topology, number of data values in each classified set (training, validation and test set), and the training algorithm and its parameters are shown in Table 3.

The performance of the ANN model was evaluated based on some statistical parameters such as mean square error (MSE), square of correlation coefficient (R^2), root mean square error (RMSE) introduced in the following equations:[70]

$$\text{MSE} = \frac{\sum_i \left(y_{\text{ANN},i} - y_{\text{exp},i}\right)^2}{n - 1} \tag{4}$$

$$\text{RMSE} = \left(\frac{\sum_{i=1}^n \left(y_{\text{ANN},i} - y_{\text{exp},i}\right)^2}{n - 1}\right)^{\frac{1}{2}} \tag{5}$$

$$R^2 = 1 - \frac{\sum_i \left(y_{\text{ANN},i} - y_{\text{exp},i}\right)^2}{\sum_i \left(y_{\text{ANN},i} - y_m\right)^2} \tag{6}$$

where $y_{\text{ANN},i}$ and $y_{\text{exp},i}$ are predicted, and the experimental value of AGGN for the ith cationic surfactant molecule, respectively, y_m is the mean of y_{exp} in eqn (4)–(6), and, n is the number of molecules in each data set (training, test or validation set).

The main goal in the training step was minimizing the MSE of the test set as data which were not used during the training iterations, a fact which confirmed the ANN ability for the

Table 2 The selected structural descriptors for QSAR analysis

ID	Name	Description	Block
1	JGI6	Mean topological charge index of order 6	2D autocorrelations
2	H-047	H attached to $C^1(sp^3)/C^0(sp^2)$	Atom-centred fragments
3	Mor20p	Signal 20/weighted by atomic polarizability	3D-MoRSE descriptors
4	ESpm12x	Spectral moment 12 from edge adjacency matrix weighted by edge degrees	Edge adjacency indices

Table 3 Network parameters (in the ANN–Matlab toolbox) in the QSAR analysis of the cationic surfactants

Topology	Four inputs, one output and one hidden layer with five neurons ($4 \times 5 \times 1$)
Data	Training set: 70% randomly selected observation data (29 data values)
	Test set: 15% randomly selected observation data (6 data values)
	Validation set: 15% randomly selected observation data (6 data values)
Beginning function	Log–sigmoid
Training algorithm	Levenberg–Marquardt algorithm
Loss function conditions	Minimum MSE
Stopping conditions	The network stops in one of three ways:
	Validation check > 10
	Minimum gradient < 10^{-7}
	Momentum speed > 10^{10}

prediction of the new data. Here, the optimal ANN architecture was achieved according to the minimum value of the MSE and the maximum value of R^2 of the test set. A network ($4 \times 5 \times 1$) was the optimal model whose topology is illustrated in Fig. 1.

The molecules in each data set were analyzed by the optimal ANN algorithm, and their AGGN values were estimated to clarify the prediction ability of this non-linear model. All the results were converted to the original state and plotted *versus* the corresponding experimental AGGNs as shown in Fig. 2. Table 4 shows a summary of statistical parameters such as the values of R^2, MSE and RMSE for training, validation, and test sets using the ANN method. The R^2 values between the experimental and

predicted results reveal that the ANN model was highly efficient for the analysis of the QSAR data studied.

Moreover, the studied data was analyzed by MLR methodology and the results were compared with the ANN strategy to reveal the necessity of employing non-linear modeling in this investigation. Fig. 4 illustrates the MLR coefficients *versus* the descriptors. Some statistical parameters of the MLR model are given in Table 4 and the correlation between the experimental and predicted results of the MLR model are shown in Fig. 3. The compared results showed that ANN is a powerful tool for detecting the relationship between the surfactant molecules and their AGGNs. This could be attributed to the non-linear
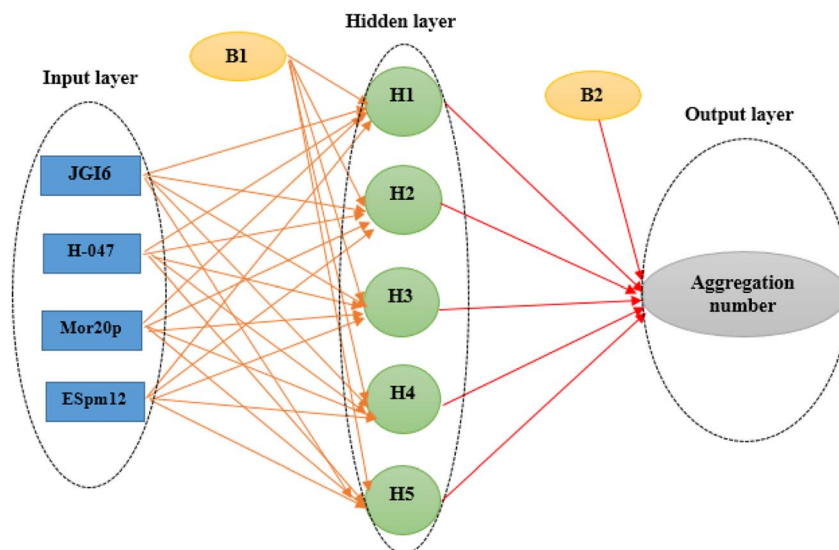

Fig. 1 Artificial neural network architecture in QSAR studies of the cationic surfactants.
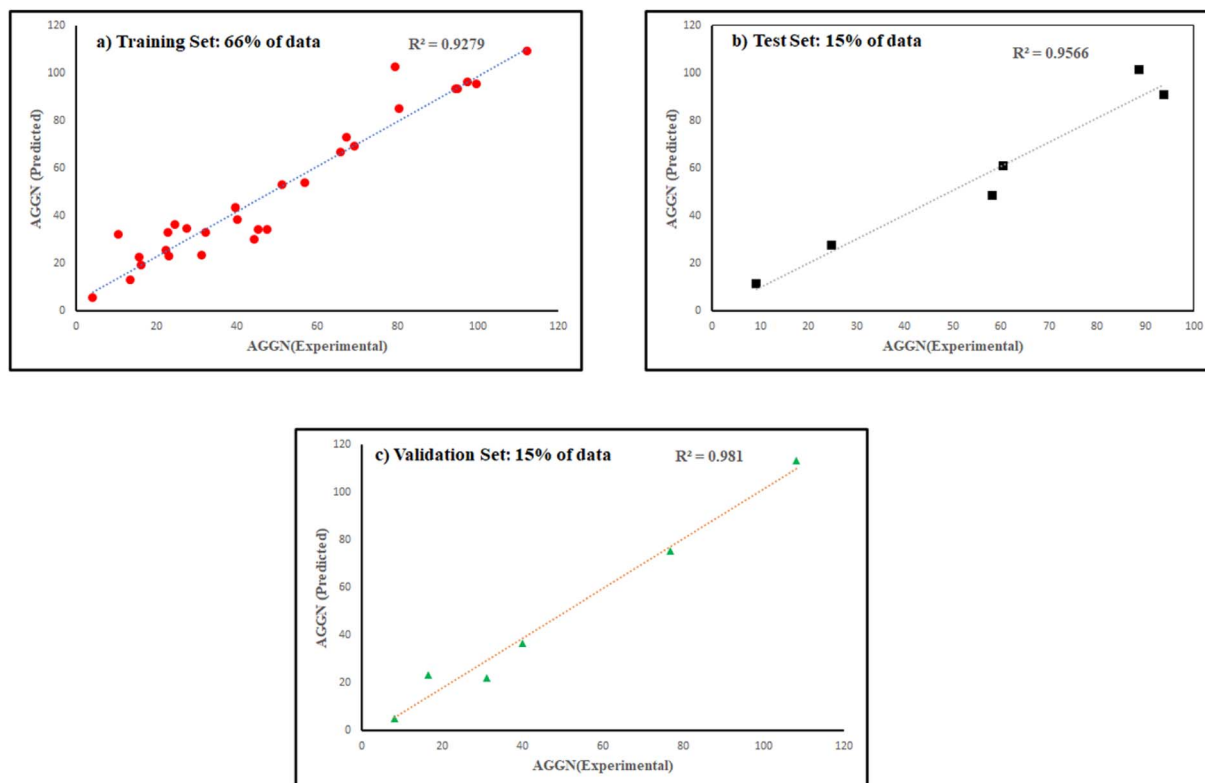
Fig. 2 The plots of predicted AGNNs determined by ANN analysis *versus* experimental AGNNs of cationic surfactants molecules for the three data sets used in the ANN analysis.

relationship between the molecular structures of the surfactants and their AGGNs. To investigate this claim for each selected descriptor, the AGGN values of the surfactants molecules were plotted against the values of the descriptor, as shown in Fig. S1 (ESI).† This figure illustrates the non-linearity in this data, furthermore, Fig. S2 (ESI)† shows the non-linear relationship between the AGGNs of a set of molecules with the same hydrocarbon chain length but different polar head groups (CPC, $[C_{16}MIM]Br$, $C_{16}TAB$, $[C_{16}hpim]Br$, and $C_{16}E_2TAB$).

5. Effect of input variables

The weight values in the ANN network can be employed to estimate the relative importance of each input variable on the output target using the Garson method, a numerical approach, as follows:[71]

$$Q_{md} = \frac{\sum_{n=1}^{h} |w_{mn} v_{nd}| \Big/ \sum_{t=1}^{N} |w_{tn}|}{\sum_{m=1}^{N} \sum_{n=1}^{h} |w_{mn} v_{nd}| \Big/ \sum_{t=1}^{N} |w_{tn}|} \tag{7}$$

where w_{tn} is the weight between the m_{th} input and the n_{th} hidden neuron, and v_{nd} represents the weight between the n_{th} hidden neuron and the d_{th} output target.

In this study, the percentage of influence of the input variables on the AGGNs was estimated by incorporating input-hidden and hidden-output connection weights based on eqn (7), and the results are reported in Table 5. The trend of importance of the input descriptors was in the following order: H-047 > ESpm12x > JGI6> Mor20p.

In MLR analysis, the distribution coefficients of these descriptors were assigned by their importance. Although the importance trend in MLR analysis did not coincide with the

Table 4 Statistical parameters of the ANN and MLR models in the QSAR studies of cationic surfactants

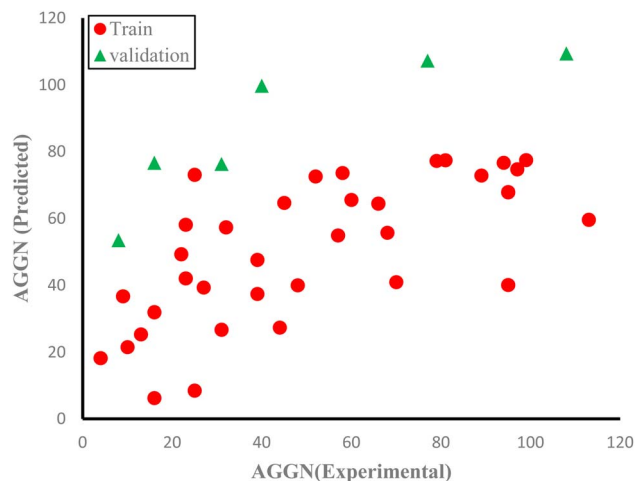| Set of data	R^2		MSE		RMSE	
	ANN	MLR	ANN	MLR	ANN	MLR
Total	0.9392	0.5010	8.7070	507.1145	2.9508	22.5192
Training	0.9256	0.4578	12.4385	528.9620	3.5268	22.9992
Test	0.9526	—	4.80×10^{-4}	—	0.0219	—
Validation	0.9762	0.6053	8.67×10^{-5}	2.4595×10^3	0.0093	49.5936

Fig. 3 The scatterplot of predicted AGNNs by MLR analysis *versus* experimental AGNNs of cationic surfactants molecules in different data sets.
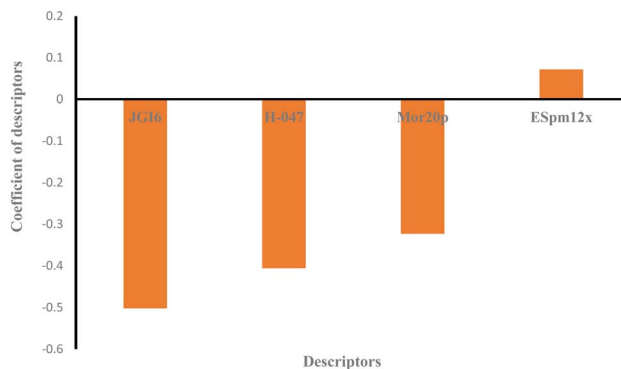

Fig. 4 The plot of coefficients of descriptors in MLR modeling *versus* descriptors' names for QSAR study of the cationic surfactants.

previous trend, the sign of the coefficient gives complementary information about the descriptors. Indeed, the signs can help us interpret the relationship between the AGGNs of the molecules and the descriptors and provide more relevant information, as will be discussed next.

Descriptor H-047 belongs to atom-centered fragments (ACF) class descriptor which shows a structural fragment, H, attached to $C^1(sp^3)/C^0(sp)$, in a molecular structure, where the superscript of C denotes the formal oxidation number of the carbon atom.[72] This oxidation number is the sum of the conventional bond orders with electronegative atoms. The fewer hydrogen atoms that are attached to sp or sp^3 hybridized carbon atoms there are, the higher H-047 descriptor observed.[72] As shown in Fig. 4, this descriptor has a negative effect on the AGGN as expected because fewer hydrogen atoms lead to a higher AGGN. Therefore, the H-047 descriptor recommends fewer hydrogen atoms be attached to sp or sp^3 hybridized carbon atoms to increase the AGGN of the titled compounds. For example, here, increasing the number of hydrogen atoms attached to sp or sp^3 hybridized carbon atoms for DMAC (−0.81 of H-047), DDMAC (−0.63 of H-047), and [C_{16}MIM][Br] (−0.81 of H-047), C_{16}TAB (−0.45 of H-047) molecules causes the AGGN to decrease from 89 to 66, and 99 to 95, respectively.

Table 5 Effective weight matrix of the ANN modeling in the QSAR studies of cationic surfactants

| Input descriptors | | | | | Hidden to |
JGI6	H-047	Mor20p	ESpm12x	Hidden neurons	output
2.8220	−3.1771	−2.1059	−2.2939	H1	2.4575
1.2441	−1.3730	3.3333	0.7633	H2	−1.2424
1.7163	2.5627	−0.8406	−4.1556	H3	−4.2588
1.8402	2.8236	0.8486	−0.2216	H4	−4.9700
2.8215	8.2460	0.7830	−3.1571	H5	3.8160
22.15	38.57	16.78	22.47	Relative importance (%)	

ESpm12x is the spectral moment of the edge-weighted adjacency matrix which is represented by the structural fragments present in the molecules.[73] This descriptor has been widely used for the interpretation of physical and physicochemical properties of alkanes and has presented powerful significant models from the statistical point of view. Indeed, the molecules with higher ESpm12x values belong to the higher length of the hydrocarbon chain.[75] As shown in Fig. 4, this descriptor has a positive effect on the AGGN property, indicating that the ESpm12x is directly related to the AGGN.

JGI6 is a mean topological charge index of order 6 which can assess both the charge transfer between pairs of atoms and the global charge transfer in a molecule.[74] This descriptor represents the total charge transfer between atoms at a topological distance of 6 which are closely related to substitutions at the peripheral molecular sites, and molecular polarity. In a molecule, the higher the charge transfer is, the higher the JGI6 value observed.[74] In this study, the JGI6 descriptor showed a negative impact on AGGN and it is expected that a smaller AGGN will be observed for a molecule with higher polarity and, in turn, a higher charge transfer. For example, both C_{16}TAB (−0.44 of JGI6) and CPC (−0.75 of JGI6) molecules have the same hydrocarbon chain length but in CPC, due to the resonance and charge distribution on the molecule surface, the charge transfer is higher and the molecule has a smaller AGGN.

The Mor20p descriptor expresses the 3D structure of a molecule and encodes information about the polarizability and is similar to the JGI6 descriptor, but the Mor20p value of a molecule is directly related to the polarizability of the molecule. As discussed, previously, the increase in polarizability of a molecule results in the decrease in AGGN. For example, C_{16}TAB (−1 of Mor20p) and CPC (−0.3694 of Mor20p) molecules have the same hydrocarbon chain length but a different polar head group and the Mor20p of CPC molecule is higher than that of C_{16}TAB, and as expected, its AGGN is lower due to its higher polarity.[75]

Overall, it can be concluded that the decreased values for H-047, JGI6 and Mor20p, together with the increased value for the ESpm12x descriptor will provide higher values for the AGGN property of the studied cationic surfactants.

6. Conclusions

In this study, the QSAR study was performed on some cationic surfactants to correlate the molecular structure of the surfactants with their AGGNs. Among more than 3000 molecular descriptors that were considered in generating the QSAR model, four descriptors resulted in a statistically significant model.

The QSAR data was analyzed based on both linear (MLR) and non-linear (ANN) modelling techniques and the results of these methods were compared statistically. A higher R^2 and a lower RMSE of the ANN method were achieved, a fact which supports the efficiency of ANN in detecting relationships between surfactant molecules and their AGGNs with a high predictive power.

In summary, the QSAR-ANN was proposed as a promising technique to predict the AGGNs of surfactants and to obtain extract maximal information about the surfactant systems with minimal experimental effort.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgements

References

1 S. Mahbub, S. Akter, Luthfunnessa, P. Akter, M. A. Hoque, M. A. Rub, D. Kumar, Y. G. Alghamdi, A. M. Asiri and H. Džudžević-Čančar, *RSC Adv.*, 2020, **10**, 14531–14541.

2 E. Jungerman, *Cationic Surfactants*, Marcel Dekker, New York, 1969.

3 C. F. Jesus, A. A. S. Alves, S. M. Fiuza, D. Murtinho and F. E. Antunes, *J. Mol. Liq.*, 2021, **342**, 117389.

4 I. Pacheco-Fernández, R. González-Martín, F. A. e Silva, M. G. Freire and V. Pino, *Anal. Chim. Acta*, 2021, **1143**, 225–249.

5 A. Valls, B. Altava, V. Aseyev, E. García-Verdugo and S. V. Luis, *J. Mol. Liq.*, 2021, **341**, 117434.

6 D. Kumar, N. Azum, M. A. Rub and A. M. Asiri, *J. Mol. Liq.*, 2018, **262**, 86–96.

7 T. Rasheed, S. Shafi, M. Bilal, T. Hussain, F. Sher and K. Rizwan, *J. Mol. Liq.*, 2020, **318**, 113960.

8 H. R. Afifi, S. Mohammadi, A. Mirzaei Derazi, S. Moradi, F. Mahmoudi Alemi, E. Hamed Mahvelati and K. Fouladi Hossein Abad, *J. Mol. Liq.*, 2021, 116808.

9 T. M. Herrington and S. S. Sahi, *Colloids Surf.*, 1986, **17**, 103–113.

10 M. Tornblom, U. Henriksson and M. Ginley, *J. Phys. Chem.*, 1994, **98**, 7041–7051.

11 M. Pisárčik, F. Devínsky and M. Pupák, *Open Chem.*, 2015, **13**, 922–931.

12 P. C. Griffiths, A. Paul, R. K. Heenan, J. Penfold, R. Ranganathan and B. L. Bales, *J. Phys. Chem. B*, 2004, **108**, 3810–3816.

13 F. Reiss-Husson and V. Luzzati, *J. Phys. Chem.*, 1964, **68**, 3504–3511.

14 S. S. Atik, M. Nam and L. A. Singer, *Chem. Phys. Lett.*, 1979, **67**, 75–80.

15 P. Lianos and R. Zana, *J. Phys. Chem.*, 1980, **84**, 3339–3341.

16 N. Lebedeva, R. Zana and B. L. Bales, *J. Phys. Chem. B*, 2006, **110**, 9800–9801.

17 R. Zana, Luminescence Probing Methods, in *Surfactant Solutions*, ed. R. Zana, Dekker, New York, 1987.

18 H. Martens, *J. Chemom.*, 2015, **29**, 563–581.

19 M. A. Rasmussen, Å. Rinnan, A. B. Risum and R. Bro, *J. Chemom.*, 2021, **35**, e3378.

20 K. Ciura, S. Ulenberg, H. Kapica, P. Kawczak, M. Belka and T. Bączek, *J. Pharm. Biomed. Anal.*, 2020, **188**, 113423.

21 L. Lévêque, N. Tahiri, M.-R. Goldsmith and M.-A. Verner, *Comput. Toxicol.*, 2022, **21**, 100211.

22 K. Roy, S. Kar and P. Ambure, *Chemom. Intell. Lab. Syst.*, 2015, **145**, 22–29.

23 P. De and K. Roy, *Eur. J. Med. Chem.*, 2022, **4**, 100035.

24 A. T. Balaban, *SAR QSAR Environ. Res.*, 1998, **8**, 1–21.

25 A. T. Balaban, *From Chemical Topology to Three-Dimensional Geometry*, Springer Science & Business Media, 2006.

26 J. Eichenlaub, P. W. Rakowska and A. Kloskowski, *J. Mol. Liq.*, 2022, **350**, 118511.

27 J. Yao, J. Wen, H. Li and Y. Yang, *J. Hazard. Mater.*, 2022, **423**, 127131.

28 A. E. Comesana, T. T. Huntington, C. D. Scown, K. E. Niemeyer and V. H. Rapp, *Fuel*, 2022, **321**, 123836.

29 L. Si-Hung, Y. Izumi, M. Nakao, M. Takahashi and T. Bamba, *Anal. Chim. Acta*, 2022, **1197**, 339463.

30 B. R. Kowalski and M. B. Seasholtz, *J. Chemom.*, 1991, **5**, 129–145.

31 J. P. M. Andries, G. H. Tinnevelt and Y. Vander Heyden, *Talanta*, 2022, **239**, 123140.

32 A. Badura, J. Krysiński, A. Nowaczyk and A. Buciński, *Chemom. Intell. Lab. Syst.*, 2022, **222**, 104501.

33 H. Chinh Nguyen, F. Alamray, M. Kamal, T. Diana, A. Mohamed, M. Algarni and C.-H. Su, *J. Mol. Liq.*, 2022, **354**, 118888.

34 S. S. Sutar, S. M. Patil, S. J. Kadam, R. K. Kamat, D.-k. Kim and T. D. Dongale, *ACS Omega*, 2021, **6**, 29982–29992.

35 T. D. Dongale, K. P. Patil, S. R. Vanjare, A. R. Chavan, P. K. Gaikwad and R. K. Kamat, *J. Comput. Sci.*, 2015, **11**, 82–90.

36 S. Sahu, M. K. Yadav, A. K. Gupta, V. Uddameri, A. N. Toppo, B. Maheedhar and P. S. Ghosal, *J. Environ. Manage.*, 2022, **302**, 113965.

37 W. F. d. C. Rocha, C. B. d. Prado and N. Blonder, *Molecules*, 2020, **25**, 3025.

38 M. Carlin, T. Kavli and B. Lillekjendlie, *Chemom. Intell. Lab. Syst.*, 1994, **23**, 163–177.

39 K.-T. Fang, Y. Lin and H. Peng, *Chemom. Intell. Lab. Syst.*, 2022, **221**, 104474.

40 M. J. Willis, G. A. Montague, C. Di Massimo, M. T. Tham and A. J. Morris, Artificial neural networks in process estimation and control, *Automatica*, 1992, **28**(6), 1181–1187.

41 T. Davoudizadeh, S. M. Sajjadi and L. Ma'mani, *J. Iran. Chem. Soc.*, 2018, **15**, 1999–2006.

42 T. D. Dongale, P. R. Jadhav, G. J. Navathe, J. H. Kim, M. M. Karanjkar and P. S. Patil, *Mater. Sci. Semicond. Process.*, 2015, **36**, 43–48.

43 D. W. Roberts and J. Costello, *QSAR Comb. Sci.*, 2003, **22**, 220–225.

44 X. Kong, C. Qian, W. Fan and Z. Liang, *J. Mol. Struct.*, 2018, **1156**, 164–171.

45 V. Joshi, M. Kadam and M. Sawant, *J. Surfactants Deterg.*, 2007, **10**, 25–34.

46 L. Wattebled, A. Laschewsky, A. Moussa and J.-L. Habib-Jiwan, *Langmuir*, 2006, **22**, 2551–2557.

47 C. P. Frizzo, I. d. M. Gindri, C. R. Bender, A. Z. Tier, M. A. Villetti, D. C. Rodrigues, G. Machado and M. A. P. Martins, *Colloids Surf., A*, 2015, **468**, 285–294.

48 K. Srinivasa Rao, T. Singh, T. J. Trivedi and A. Kumar, *J. Phys. Chem. B*, 2011, **115**, 13847–13853.

49 X. Li, Y.-A. Gao, J. Liu, L.-q. Zheng, B. Chen, L.-Z. Wu and C. H. Tung, *J. Colloid Interface Sci.*, 2010, **343**, 94–101.

50 S. A. A. Rizvi and S. A. Shamsi, *Anal. Chem.*, 2006, **78**, 7061–7069.

51 T. Singh and A. Kumar, *J. Phys. Chem. B*, 2007, **111**, 7843–7851.

52 V. Mosquera, J. M. Ruso, D. Attwood, M. N. Jones, G. Prieto and F. Sarmiento, *J. Colloid Interface Sci.*, 1999, **210**, 97–102.

53 B. W. Barry and R. Wilson, *Colloid Polym. Sci.*, 1978, **256**, 44–51.

54 A. Malliaris, J. Le Moigne, J. Sturm and R. Zana, *J. Phys. Chem.*, 1985, **89**, 2709–2713.

55 W. Li, H. Xie, Y. Huang, L. Song, Y. Shao and K. Qiu, *J. China Med. Univ.*, 2016, **12**, 134–136.

56 J. Łuczak, J. Hupka, J. Thöming and C. Jungnickel, *Colloids Surf., A*, 2008, **329**, 125–133.

57 R. G. Alargova, I. I. Kochijashky, M. L. Sierra and R. Zana, *Langmuir*, 1998, **14**, 5412–5418.

58 Ž. Medoš, S. Friesen, R. Buchner and M. Bešter-Rogač, *Phys. Chem. Chem. Phys.*, 2020, **22**, 9998–10009.

59 S. E. Anachkov, K. D. Danov, E. S. Basheva, P. A. Kralchevsky and K. P. Ananthapadmanabhan, *Adv. Colloid Interface Sci.*, 2012, **183–184**, 55–67.

60 M. Frisch, G. Trucks, H. B. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, G. Scalmani, V. Barone, B. Mennucci and G. Petersson, *Gaussian 09, revision D. 01*. Gaussian, Inc., Wallingford CT, 2009.

61 J. Řezáč and P. Hobza, *Chem. Phys. Lett.*, 2011, **506**, 286–289.

62 R. Todeschini, V. Consonni, A. Mauri and M. Pavan, *Dragon for windows (software for molecular descriptor calculations), version 5.5*. Talete srl, Milan, Italy, 2006.

63 H. Demuth and M. Beale, *Neural Network Toolbox For Use with Matlab–User'S Guide Version 3.0*, 1993.

64 J. Zupan and J. Gasteiger, *Neural networks in chemistry and drug design*, John Wiley & Sons, Inc, 1999.

65 U. M. R. Paturi, S. Cheruku and N. S. Reddy, *Arch. Comput. Methods Eng.*, 2022, **25**, 3109–3149.

66 A. T. C. Goh, *AIENG*, 1995, **9**, 143–151.

67 V. Arabzadeh and M. R. Sohrabi, *Chemom. Intell. Lab. Syst.*, 2022, **221**, 104475.

68 S. Aber, N. Daneshvar, S. M. Soroureddin, A. Chabok and K. Asadpour-Zeynali, *Desalination*, 2007, **211**, 87–95.

69 F. Despagne and D. Luc Massart, *Analyst*, 1998, **123**, 157R–178R.

70 A. Hammoudi, K. Moussaceb, C. Belebchouche and F. Dahmoune, *Constr. Build. Mater.*, 2019, **209**, 425–436.

71 G. D. Garson, *AI Expert*, 1991, **6**(4), 47–51.

72 B. K. Sharma, K. Sarbhai and P. Singh, *Eur. J. Med. Chem.*, 2010, **45**, 1927–1934.

73 E. Estrada, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 320–328.

74 K. Nikolic and D. Agababa, *J. Mol. Graphics Modell.*, 2009, **28**, 245–252.

75 C. W. Yap and Y. Z. Chen, *J. Pharm. Sci.*, 2005, **94**, 153–168.