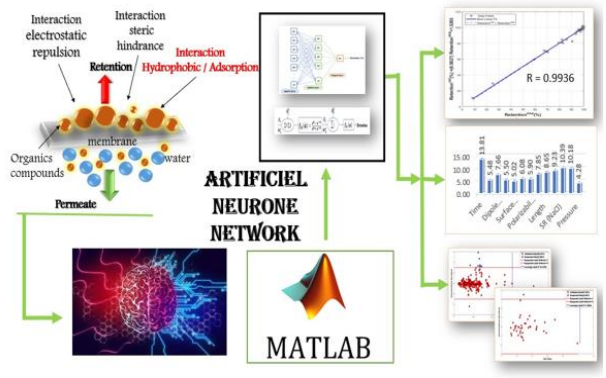
Global NEST Journal, Vol 26, No 5, 05580

Prediction of the effect of adsorption on the retention of organic compounds by NF/RO using QSPR-ANN

Nechoua Merarsi[1]*, Yamina Ammi[1] and Salah Hanini[1]
[1]Laboratory of Biomaterials and Transport Phenomena (LBMPT), University of Medea, Medea 26000, Algeria
Received: 21/11/2023, Accepted: 01/05/2024, Available online: 17/05/2024
*to whom all correspondence should be addressed: e-mail: merarsi.nechoua@gmail.com
https://doi.org/10.30955/gnj.005580

Graphical abstract



Abstract

Understanding the retention of organic compounds (OCs) is critical for membrane applications in water recycling. The objective of this study was to create an optimized model using Artificial Neural Networks for Quantitative Structure-Property Relationship (QSPR-ANN) to predict the effect of adsorption on the retention of organic compounds (OCs) by nanofiltration (NF) and reverse osmosis (RO).

An optimal model (QSPR-ANNoptimal) characterized by a similar structure (13 neurons in the inputs layer, 11 neurons in the hidden layer, and 1 neuron in the output layer) is constructed to predict the effect of adsorption on the retention of organic compounds by membranes. A set of 273 data points was used to test the neural network. the data set was used 70% for training, 15% for validation, and 15% for testing. For the most promising neural network model, the calculated retention values were compared to the experimental retention values, and good correlations were found (the determination coefficient "R^2 = 0.9872" and the root mean squared error "RMSE = 2.2743%" for the test phase). This indicates the good robustness of the established QSPR-ANN model and the possibility of predicting the various parameters that characterize the retention of OCs by RO/NF. Sensitivity analysis revealed that the effect of adsorption retention of organic compounds by reverses osmosis and nanofiltration membranes depends more precisely on two important interactions (hydrophobic/adsorption and steric hindrance).

Keywords: Modelisation; hydrophobic adsorption; interactions; retention; organic compounds; reverses osmosis; nanofiltration; artificial neural networks

1. Introduction

The increasing global utilization of organic compounds (OCs) such as hormones, pesticides, pharmaceuticals, surfactants, and phenolic substances has led to their presence in wastewater effluents, source waters, groundwater, and even treated drinking water. This has given rise to a fresh environmental challenge, prompting significant apprehension among scientists in recent times. Consequently, the removal of OCs has become a subject of great interest. Dolar *et al*. (2013).

Modern methods are utilized to efficiently eliminate OCs. Among these technologies, membrane processes like RO/NF are particularly good at getting rid of OCs, and protecting the environment and human health. NF/RO methods have previously been shown in several investigations to be capable of eliminating OCs. These investigations have demonstrated a universal relationship between retention efficiency and complex solute-membrane interactions. These interactions include hydrophobic adsorption, electrostatic repulsion, and steric hindrance. The properties of the compounds, such as their hydrophobicity, polarity, molecular size, and charge, as well as the characteristics of the membranes, such as porosity, polarity, and electrostatic charges, affect the interactions between solute-membrane. Furthermore, these interactions are greatly influenced by operational filtration parameters are pressure, pH, permeate flux, temperature, recovery, and cross-flow velocity. (Ammi *et al*. 2023; Kim *et al*. 2018; Teychene *et al*. 2020).

(Kiso *et al*. 2001) The adsorption effect plays a key role in the permeation of solutes in practical water treatment procedures. The extended adsorption and accumulation of solutes on membranes can have a profound influence on the efficacy of solute separation. As per the findings of (Comerton *et al*. 2007), the initial retention of OCs through membrane adsorption reaches a point of stability

when equilibrium is achieved. At this juncture, additional mechanisms, such as electrostatic repulsion and steric hindrance, come into play and contribute to the retention of OCs. Surprisingly, once equilibrium is reached, adsorption can exert a detrimental impact on retention. Research has demonstrated that adsorbed compounds can dissolve within the active membrane layer, subsequently diffusing through the polymer, and ultimately dissolving on the permeate side of the membrane. Furthermore, when the compound concentration in the feed water drops below the equilibrium value, these molecules adsorb on the permeate side of the NF/RO. For this reason, studying adsorption is crucial to improving our comprehension of membrane retention processes.

A comprehensive understanding of the solute and membrane properties that influence retention forms the basis for a predictive modeling approach to determine the fate of specific compounds in high-pressure membrane applications. Despite numerous research studies attempting to establish connections between the physicochemical properties of solutes and membranes and solute retention, there remains an ongoing need for systematic and comprehensive efforts to identify key parameters that effectively predict solute separation, as well as a concurrent need for a comprehensive understanding of membrane characteristics to predict interactions between OCs and membranes, ultimately influencing retention. as highlighted by (Bellona *et al*. 2004).

sometimes, real-time analysis can be a time-consuming and laborious task for researchers. Soft computing approaches, such as genetic algorithms, ANN, or fuzzy logic, play an important role in analyzing water engineering problems (water treatment, desalination, and the accurate performance of plants) with minimal space, time, and energy. ANN is a successful soft computing technique that is widely used in chemical engineering research, such as predicting accurate outcomes through appropriate modeling and simulation. It employs a simple mathematical model inspired by the biological analogy of a human brain, it learns from examples of problem datasets and produces meaningful information for performance analysis. It can model and solve linear, nonlinear, and complex systems (Chan *et al*. 2023; Mahadeva *et al*. 2022, 2023).

The literature features a limited quantity of studies attempting to simulate nanofiltration and reverse osmosis processes using artificial neural networks. Nevertheless, only a handful of neural network models exist that can forecast the retention of organic substanc es in reverse osmosis, forward osmosis, and nanofiltration. (Ammi *et al*. 2015, 2018, 2020, 2023; Ammi, Khaouane, *et al.* 2021; Ammi, Hanini, *et al.* 2021; Khaouane *et al.* 2017; Kratbi *et al.* 2023; Libotean *et al.* 2008; Shahmansouri & Bellona, 2013; Yangali-Quintanilla *et al.* 2009).

To the best of our knowledge, this marks the initial endeavor in utilizing QSPR-ANN for forecasting the influence of adsorption on the organic compound retention in NF/RO, as well as assessing its predictive capability. Therefore, the present work aims at the prediction of the effect of adsorption on the retention of OCs by NF/RO using QSPR-ANN. The remainder of this study is structured as follows: Section 2: Artificial Neural Networks, section 3: Modeling Procedure, section 4: Results and Discussion, section 5: Sensitivity Analysis, section 6: Applicability Domain, and section 7: Conclusion.

2. Artificial neural network

Quantitative structure-property relationships (QSPR) is a technique that can predict the properties of chemical/biological systems based on their molecular structure. Relationships are often established using statistical modeling methods, such as artificial neural networks (ANN)Fissa *et al.* (2023).

Artificial neural networks are powerful tools that are often utilized as black-box models due to their exceptional capacity to learn and generalize nonlinear functional relationships between input and output variables. They operate as data-driven adaptive algorithms, capable of learning from training epochs and uncovering subtle functional correlations within the data, even when the underlying relationships between parameters are ambiguous or challenging to define. With a sufficient amount of data, neural networks can effectively tackle problems by treating them as multivariate nonlinear statistical models. The connections within neural networks, known as synapses, have adaptive weights that are adjusted during the learning process and are proportional to the synaptic potential. This adaptability allows neural networks to discover complex patterns and relationships in the data, making them valuable for a wide range of applications in fields like machine learning and artificial intelligence Mohammad *et al.* (2022); Rehab *et al.* (2022).

The most widely used architecture is Multilayer Perceptron (MLP) with only three layers: 1. The first layer is the input layer, responsible for receiving input data, 2. The middle layer(s), referred to as the hidden layer, processes and passes on the information from the input layer, and 3. The final layer, called the output layer, generates the model's output.

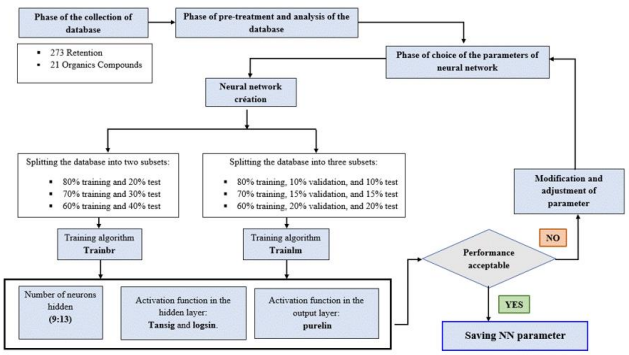


Figure 1. Designing the neural network architecture

The capacity of a neural network to continuously enhance its performance is a fundamental characteristic. With each iteration of the learning process, the network becomes more proficient in understanding and responding to its

environment. In the context of neural networks, 'learning' involves the fine-tuning of connection weights, allowing the network to adapt and make increasingly precise predictions and decisions based on the provided data Mohammad *et al.* (2022); Rehab *et al.* (2022).

3. Modeling procedure

The modeling procedure involved designing and optimizing the neural network architecture, following the steps outlined in Figure 1.

3.1. Data collection, division, pretreatment, and analysis

In this study, we used available data from 4 references from 2009 to 2018 Arsuaga *et al.* (2010); Dolar *et al.*(2013, 2017); Liu *et al.* (2018). The database contains 273 retention data for 21 OCs (pharmaceutical compounds and phenolics). The list of 21 OCs is presented in the Supplementary Data (Table 1).

The selection of input and output variables is based on the hydrophobic/adsorption interaction between the OCs and the membranes (RO/NF). These interactions between solutes and the membrane are determined by the descriptors of the OCs, membrane characteristics, and operating conditions Gur-Reznik *et al.* (2011).

We choose the following inputs:

1 The descriptors of the OCs are molecular weight "Mw", the logarithm of the octanol-water partition coefficient "*log Kow*", dipole moment, molecular length, surface area min, surface area max, polar surface area, and polarizability;

2 the characteristics of the membranes are molecular weight cut-off "MWCO", sodium chloride salt rejection "SR NaCl", and membrane hydrophobicity "contact angle";

3 the operating condition is pressure.

Molecular descriptors (the MW, the *log Kow*, the polar surface area, and the polarizability) were calculated using ChemSpider (Http://Www.Chemspider.Com, n.d.). We calculated the dipole moment of the descriptor and the molecular size of the descriptor (the molecular length, the surface area min, and the surface area max) by two software (hyperChem and Chembio 3D).

The values of the molecular width, the molecular depth, are defined by the following equations (01,02), and the equivalent molecular width “Eqwidth” was calculated by the following equation (03):

$$Width = \frac{1}{2}\sqrt{S_{min}} \quad (1)$$

$$\text{Depth} = \frac{1}{2}\sqrt{S_{max}} \quad (2)$$

$$\text{Eqwidth} = \sqrt{\text{width*depth}} \quad (3)$$

Table 1 displays the minimum (min), maximum (max), mean, and standard deviations (Std) values for both the input and output data.

Table 1. Statistical analysis of inputs and output

	Min	Max	Mean	Std
Temps (h)	0.0000	24.0000	6.5616	7.1859
MW (g mol-1)	94.1100	392.4700	288.0913	70.4984
Dipole moment (Debye)	0.2358	6.3000	4.0203	1.4071
Log K_{ow}	-1.2200	3.4800	1.7275	0.9046
Polar Surface Area (nm^2)	0.2000	1.3000	0.7592	0.2732
Polarizability (nm^3)	0.0112	0.0397	0.0311	0.0065
Length (nm)	0.0970	0.1719	0.1472	0.0146
Surface area min (nm^2)	2.1394	5.1640	4.1329	0.6206
Surface area max (nm^2)	2.4437	7.8549	5.3434	1.2226
MWCO (Dalton)	100.0000	340.0000	185.9341	102.0720
SR ($CaCl_2$) (%)	-	-	-	-
SR(NaCl) (%)	20.5300	98.6400	73.9201	23.4504
Contact angle (°)	20.1000	73.1600	53.5681	14.3834
Pressure (KPa)	1000.0000	4100.0000	1127.4725	523.1549
Retention (%)	7.1713	100.0000	86.0097	20.3951

3.2. Model development

The QSAR-ANN models were developed for the prediction of the effect of adsorption on the retention of OCs by (NF) and (RO) membranes. Each neural network contains 23 and 27 variables (13 neurons in the input layer, 9 and 13 neurons in the hidden layer, and 1 neuron in the output layer).

The collective data in this model were randomly divided into two subsets: training and testing; and also, randomly divided into three subsets: training, validation, and testing.

In the process of creating an ANN model, a typical allocation of 60-80% of the data is designated for training, making it the largest segment of the dataset. The training phase signifies the initial step in constructing an ANN model, during which the network learns and establishes the connections between input and output variables. Complex computations occur at this stage, and the neuron weights are adjusted after each epoch using one of the training algorithms to achieve a high level of accuracy. Different criteria, such as the number of epochs or iterations, and a minimum error threshold, can be configured. After the training phase is completed, the

remaining data is evenly divided between the validation and testing phases. in the validation phase, the validation dataset which includes unseen data, is utilized to evaluate the predictive capabilities of the ANN model. Employing multiple validation checks helps prevent the model from becoming stuck in local minima. In the testing phase, a distinct set of unseen data is used as input to forecast the output parameters, which assesses the model's performance on new, unseen data Jawad *et al*. (2021).

in this work used the training algorithms are the Regularization-Bayesienne"train-BR" and the Levenberg-Marquard "train-LM". The quantity of neurons in the hidden layer varies based on the network's performance throughout the training phase (9 to 13 neurons). The activation functions used in the hidden layer are the tangent hyperbolic (tansig) and logarithmic sigmoid (logsig) and the activation function used in the output layer is the pure-linear (purelin). The selection of the optimal subset division, the number of hidden neurons, the hidden functions, and the output function (Designing the neural network architecture) for a neural network optimal is done by trial and error method. The prediction of the effect of adsorption on the retention of OCs during NF/RO using QSPR-ANN was performed using MATLAB software.

Table 2. Effect of dividing the database with the activation function (tansig) and two training algorithms

	Splitting the database into two subsets (trainbr)				Splitting the database into three subsets (trainlm)		
		R^2	RMSE (%)			R^2	RMSE (%)
Division 01	Total phase 100%: 273 datapoints	0.9720	3.4079	Division 04	Total phase 100%: 273 datapoints	0.9575	4.1991
	Training phase 60%: 164 datapoints	0.9986	0.7735		Training phase 60%: 163 datapoints	0.9732	3.2192
	Validation phase	-	-		Validation phase 20%: 55 datapoints	0.9306	5.1949
	Test phase 40%: 109 datapoints	0.9283	5.3092		Test phase 20%: 55 datapoints	0.9458	5.4609
Division 02	Total phase 100%: 273 datapoints	0.9779	3.0317	Division 05	Total phase 100%: 273 datapoints	0.9746	3.2466
	Training phase 70%: 191datapoints	0.9892	2.1833		Training phase 70%: 191datapoints	0.9710	3.4460
	Validation phase	-	-		Validation phase 15%: 41 datapoints	0.9803	3.1136
	Test phase 30% :82 datapoints	0.9428	4.4154		Test phase 15%: 41 data points	0.9872	2.2743
Division 03	Total phase 100%: 273 datapoints	0.9843	2.5813	Division 06	Total phase 100%: 273 datapoints	0.9817	2.7565
	Training phase 80%: 218 datapoints	0.9976	1.0165		Training phase 80%: 219datapoints	0.9880	2.1326
	Validation phase	-	-		Validation phase 10%: 27 datapoints	0.9752	3.6410
	Test phase 20%: 55 datapoints	0.9368	5.3830		Test phase 10%: 27 datapoints	0.9553	5.1654

4. Results and discussion

In this work, QSPR-ANN was used to construct a nonlinear model for the prediction of the effect of adsorption on the retention of OCs by NF/RO membranes. The performance of the model was assessed using the determination coefficient (R^2) (values above 0.5 are generally considered satisfactory and values above 0.9 are considered excellent) and the root mean squared error (RMSE) was used to determine the modeling error between the experimental and calculated values, with a perfect RMSE when a Lower value, it is defined as follows Sediri *et al.* (2017); Wang *et al.* (2009).

Table 3. Effect of dividing the database with the activation function (logsig) and two training algorithms

	Splitting the database into two subsets (trainbr)				Splitting the database into three subsets (trainlm)		
		R^2	RMSE (%)			R^2	RMSE (%)
Division 01	Total phase 100%: 273 datapoints	0.9706	3.5450	Division 04	Total phase 100%: 273 datapoints	0.9584	4.1779
	Training phase 60%: 164 datapoints	0.9872	2.2732		Training phase 60%: 163 datapoints	0.9912	1.9056
	Validation phase	-	-		Validation phase 20%: 55 datapoints	0.9122	6.2009
	Test phase 40%: 109 datapoints	0.9504	4.8684		Test phase 20%: 55 datapoints	0.9071	6.1179
Division 02	Total phase 100%: 273 datapoints	0.9783	3.0157	Division 05	Total phase 100%: 273 datapoints	0.9631	3.9284
	Training phase 70%: 191datapoints	0.9874	2.3662		Training phase 70%: 191datapoints	0.9890	2.2767
	Validation phase	-	-		Validation phase 15%: 41 datapoints	0.8290	5.8909
	Test phase 30% :82 datapoints	0.9532	4.1518		Test phase 15%: 41 data points	0.8894	6.6263
Division 03	Total phase 100%: 273 datapoints	0.9888	2.1525	Division 06	Total phase 100%: 273 datapoints	0.9549	4.3426
	Training phase 80%: 218 datapoints	0.9918	1.8637		Training phase 80%: 219datapoints	0.9626	3.9661
	Validation phase	-	-		Validation phase 10%: 27 datapoints	0.9586	5.3357
	Test phase 20%: 55 datapoints	0.9791	3.0381		Test phase 10%: 27 datapoints	0.8327	5.8836

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_{i,exp} - Y_{i,cal}\right)^2}{n}} \quad (4)$$

with n is the total number of data points, $Y_{i,\,cal}$ represents the calculated values and $Y_{i,\,exp}$ is the experimental values from the QSPR-ANN models.

Table 2 shows the RMSE and the R^2 obtained for the effect of adsorption on the retention of OCs by NF/RO under the influence of the training algorithm trainbr with the activation function Tansig in the hidden layer: division 1 "164 datapoints for the training data (60%) and 109 datapoints for testing data (40%)", division 2 "191 datapoints for the training data (70%) and 82 datapoints for testing data (30%)", and division 3 "218 datapoints for the training data (80%) and 55 datapoints for testing data (20%)" and with training algorithm trainlm: division 4 "163 datapoints for training data (60%), 55 datapoints for validation data (20%), and 55 datapoints for testing data (20%)", division 5 "191 datapoints for training (70%), 41datapoint for validation data (15%), and 41 datapoints for testing data (15%)", and division 6 "219 datapoints for training data (80%), 27 datapoints for validation data (20%), and 27 datapoints for the testing data(20%)".

Table 3 shows the RMSE and the R^2 obtained for the effect of adsorption on the retention of OCs by NF/RO under the influence of the activation function "Logsig" in the hidden layer with two training algorithms ("trainbr" and "trainlm").

The results of the two tables below show that division 5 is the division optimal with the training algorithm Levenberg-Marquard "train-LM" and activation function hyperbolic tangent sigmoid "Tansig". The QSPR-ANN5 model with the structure optimal (train-LM and activation function Tansig) gives lower errors than the other models (RMSE = 2.2743 and R2 = 0.9872 for the testing phase). We conclude the superiority of the optimal neural networks ($QSPR\text{-}ANN_5$) for modeling the effect of adsorption on the retention of OCs by NF/RO.

Table 4. Structures of the optimized QSPR-ANN model

Training Algorithm	Input layer	Hidden layer		Output layer	
	Neurons numbers	Neurons numbers	Activation function	Neurons numbers	Activation function
Levenberg-Marquard "LM"	13	11	tansig	1	purelin

The structure of the optimized QSPR-ANN for the prediction of the effect of adsorption on the retention of OCs by NF/RO is cited in Figure2, and a more detailed illustration of its architecture is in Table 4.

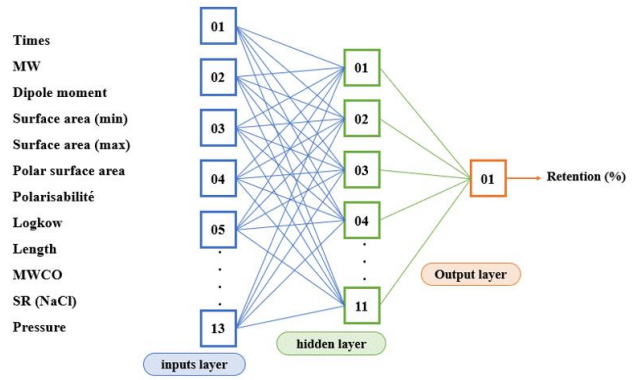


Figure2. Three-layer feed-forward neural network for modeling the

The weight matrices and bias vectors of the QSPR-ANNoptimal model are listed in Supplementary Data (Table 2).

indices w_{ji}^{I} is the input-hidden layer connection weight matrix (11 rows × 13 columns),

b_j^h is the hidden neurons bias column vector (11 rows),

$w_{1.j}^{h}$ is the hidden layer-output connection weight matrix (11 rows × 1 column), b_1^0 is the output neurons bias column vector (1 row).

From the optimized QSPR-ANNoptimal, assimilation of the effect of adsorption on the retention of OCs by the NF/RO be expressed by a mathematical model that incorporates all the inputs *Ei* (time, molecular weight "Mw", dipole moment, surface area min, surface area max, polar surface area, polarizability, *log Kow*, length, MWCO, SR(NaCl), contact angle, and pressure).

The instance outputs Z_j of the hidden layer:

j = 1,2, 3…,11

$$Z_j = f_h\left[\sum_{i=1}^{13} w_{ji}^{I} E_i + b_j^h\right] = \frac{\exp\left(\sum_{i=1}^{13} w_{ji}^{I} E_i + b_j^h\right) - \exp\left(-\sum_{i=1}^{13} w_{ji}^{I} E_i + b_j^h\right)}{\exp\left(\sum_{i=1}^{13} w_{ji}^{I} E_i + b_j^h\right) + \exp\left(-\sum_{i=1}^{13} w_{ji}^{I} E_i + b_j^h\right)} \quad (5)$$

The output "Retention":

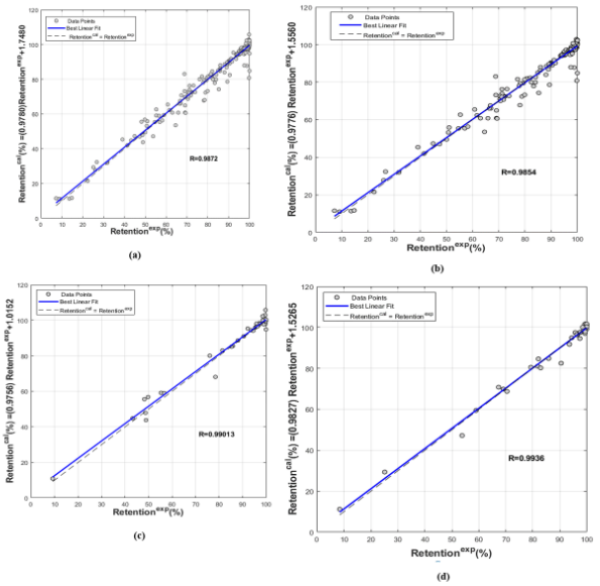
$$\text{Retention} = f_O\left[\sum_{j=1}^{11} w_{1.j}^{h} Z_j + b_1^0\right] = \sum_{j=1}^{11} w_{1.j}^{h} Z_j + b_1^0 \quad (6)$$

The combined equations (05) and (06) lead to the following mathematical formula, which describes the retention assimilation by considering all indices E_i:

$$Retention = \sum_{j=1}^{11} w_{1.j}^{h} \frac{exp\left(\sum_{i=1}^{13} w_{ji}^{I} E_i + b_j^h\right) - exp\left(-\sum_{i=1}^{13} w_{ji}^{I} E_i + b_j^h\right)}{exp\left(\sum_{i=1}^{13} w_{ji}^{I} E_i + b_j^h\right) - exp\left(-\sum_{i=1}^{13} w_{ji}^{I} E_i + b_j^h\right)} + b_1^0 \quad (7)$$

The linear regression's parameters and plot are easily generated with the MATLAB function "postreg" (Figure 3 (a), (b), (c), and (d)). The comparison of the estimated retention values calculated by the QSPR-ANN model with the experimental retention values reveals great agreement between them, with agreed vectors getting closer to the ideal "α=1 (the slope), β=0 (y-intercept), and R=1 (correlation coefficient)": [α, β, R] = [0.9780, 1.7480, 0.9872] for the total phase, [α, β, R] = [0.9776, 1.5560, 0.9854]for the training phase, [α, β, R] = [0.9756, 1.0152, 0.9901] for the validation phase, and [0.9827, 1.5265, 0.9936] for the testing phase respectively.

The errors of the QSPR -ANN optimal for the total phase, the training phase, the validation phase, and the testing phase were calculated to confirm the prediction for the effect of adsorption on the retention of OCs by NF and RO membranes.



Figures 3. Comparison between experimental and calculated retention values for the total (a), training (b), validation (c), and testing phases (d).

The root mean squared error (RMSE), the errors are the mean absolute error (MAE), the standard error of prediction (SEP), residual predictive deviation (RPD), range error ratio (RER), the mean square error (MSE), the mean relative squared error (MRSE), the accuracy factor (Af), and bias factor (Bf).

The error values were obtained with the following equations Dahmani *et al.* (2022):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|(y_{i,exp} - y_{i,cal})\right| \quad (8)$$

$$SEP(\%) = \frac{RMSE}{y_e} \times 100 \quad (9)$$

$$RPD = \frac{SD}{RMSE} \quad (10)$$

$$RER = \frac{max - min}{RMSE} \quad (11)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_{i,exp} - y_{i,cal}\right)^2 \quad (12)$$

$$MRSE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_{i,exp} - y_{i,cal}}{y_{i,exp}}\right)^2 \quad (13)$$

$$A_f = \frac{1}{n}\sum_{i=1}^{n}\left|\log\frac{y_{i,cal}}{y_{i,exp}}\right| \quad (14)$$

$$B_f = \frac{1}{n}\sum_{i=1}^{n}\left(\log\frac{y_{i,cal}}{y_{i,exp}}\right) \quad (15)$$

where n: the total number of data points,

$Y_{i,\,exp}$: the experimental retention value,

$Y_{i,\,cal}$: the calculated retention value,

Y_e: the mean value of experimental data,

SD: the standard deviation of experimental data,

min: the minimum of experimental data,

max: the maximum of experimental data.

Table 5 represents the statistical parameters of the QSPR-ANN optimal model. The determination coefficient (R^2) in both the training and validation phases is quite high, with values of 0.9710 and 0.9803, respectively, indicating excellent agreement between the experimental and calculated results. The determination coefficient (R^2) for the testing phase measures the model's ability to interpolate, and it's impressively high at 0.9872, demonstrating a strong match between experimental and calculated retention. On the flip side, we have embraced the five-level interpretations of Residual Predictive Deviation "RPD" and Range Error Ratio "RER" provided by Viscarra Rossel *et al.* (2006): excellent predictions (RPD and RER > 2.5); good predictions (RPD and RER of 2.0 to 2.5); approximate quantitative predictions (RPD and RER of 1.8 to 2.0); the ability to distinguish between high and low values (RPD and RER of 1.4 to 1.8); and unsuccessful predictions (RPD and RER< 1.40) Ammi *et al.* (2020; Viscarra Rossel *et al.* (2006).

Table 5. Statistical parameters of the QSPR-ANNoptimal model

	Total phase	Training phase	Validation phase	Testing phase
R^2	0.9746	0.9710	0.9803	0.9872
RMSE	3.2466	3.4460	3.1136	2.2743
MAE	1.8235	1.8292	2.0710	1.5494
SEP	3.7747	4.0175	3.6540	2.5870
RER	28.5923	26.9384	29.1745	40.2866
RPD	628.1936	585.2833	701.7060	895.0705
MSE	10.5406	11.8746	9.6943	5.1724
MRSE	7.6634e-06	2.7226e-05	4.3237e-05	5.6940e-07
A_f	1.0064	1.0121	1.0152	1.0017
B_f	0.9936	0.9880	1.0152	0.9983

The RPD = 628.1936 (%) and RER = 28.5923 (%) values of the QSPR-ANNoptimal model are notably higher than 2.5 for the total phase. Furthermore, various other statistical parameters, including MAE, SEP, MSE, MRSE, Af, and Bf, reinforce the model's strong predictive power across the total, training, validation, and testing phases. These results collectively highlight the model's ability to capture the nonlinear relationship between adsorption effects and the retention of OCs by NF/RO.

5. Sensitivity analysis

The analysis of the QSPR-ANNoptimal model establishes the relationship between inputs and outputs. To see the contribution as well as the variation profile of each input variable (time, molecular weight "Mw", dipole moment, surface area min, surface area max, polar surface area, polarizability, *log Kow*, length, MWCO, SR(NaCl), contact angle, and pressure) on the output (retention), sensitivity analysis is often used to study how inputs affect outputs Baghban *et al.* (2017). A “weight” method sensitivity analysis was performed. The method was first proposed by Garson (1991) and repeated by Goh (1995) Gevrey *et al.* (2003). The process of calculating the importance of "weights" is grounded in the following equation, as outlined in the research conducted by Dahmani *et al.* (2022):

IR (relative importance) = Connection Weights of Input-Hidden / Connection Weights of Hidden-Output

This equation provides a measure of the relative importance of the connection weights between the input and hidden layers compared to the connection weights between the hidden and output layers within the neural network.

$$RI_i(\%) = 100 * \frac{\sum_{j=1}^{nj}\frac{\left|w^I * W^H\right|}{\sum_{i=1}^{ni}\left|w^I * W^H\right|}}{\sum_{i=1}^{ni}\sum_{j=1}^{nj}\frac{\left|w^I * W^H\right|}{\sum_{i=1}^{ni}\left|w^I * W^H\right|}} \quad (16)$$

The results of the contributions are presented in Figure 4. The most relevant variables (RI> 5%) that can influence for prediction of the effect of adsorption on the retention of OCs by RO/NF membranes are time, MW, dipole moment, surface area min, surface area max, polar surface area, polarizability, *log Kow*, length, MWCO, SR(NaCl), and contact angle.

Figure 4 shows that the retention of OCs by reverse osmosis and nanofiltration is governed by two important interactions (hydrophobic/adsorption interaction and steric hindrance "sieving effect"). The first interaction

(hydrophobic/adsorption) takes place between hydrophobicity/polarity of OCs "*log Kow* (IR=7.85%), dipole moment (IR = 7.66%), polar surface area (IR = 6.08%), and polarizability (IR = 5.90%)" and hydrophobicity/polarity of membranes "contact angle (IR = 10.18%)". The second interaction steric hindrance "sieving effect" occurs between the parameter steric / size of OCs "length (IR = 7.85%), surface area min (IR = 5.50%), MW (IR = 5.48%), surface area max (IR = 5.02%)" and the parameter steric / size of membrane " MWCO (IR = 9.23%) and SR(NaCl) (IR = 10.38%)". This research work suggests that the OCs retention on the NF/RO strongly depends much more on the time (IR = 13.21%), SR(NaCl), and contact angle.

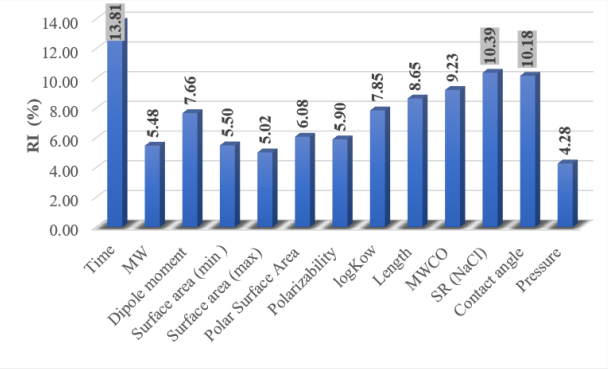


Figure 4. The histograms of the relative importance (RI) of the QSPR-ANN optimal for prediction of the effect of adsorption on the retention of OCs by RO/NF membranes

It is clear that steric/size SR (NaCl) is more suitable for modeling the impact of adsorption on the retention of OCs by RO/NF compared to steric/size MWCO (molecular weight cutoff) (RI (MWCO) = 9.23% and RI (SR "NaCl") = 10.38%). Consequently, characterizing a membrane in terms of the steric/size SR (NaCl) parameter is a simpler and more appropriate approach than using MWCO. These findings align with the results from previous studies by Ammi *et al.* (2020).

The sensitivity analysis using the weight method has effectively determined the true significance of all the variables employed in predicting the impact of adsorption on the retention of OCs by RO/NF. This, in turn, validates the appropriateness of the selected variables utilized in this research study.

6. Applicability domain

The accuracy with which data points are identified has a significant impact on the validity of the model Peter J. Rousseeuw, (2005) . Note that, as previously mentioned, this study used a database, these data points may potentially include errors stemming from laboratory measurements. Outliers are data points that deviate from the general trend of the main data points. therefore, it is important to employ robust outlier detection methods to identify and exclude imprecise experimental data, ultimately enhancing the accuracy of the model. Hosseinzadeh & Hemmati-Sarapardeh, (2014); Mohammadi *et al.* (2012). Corresponding methods often include numerical and graphical algorithms Peter J. Rousseeuw, (2005). In this study, we use the mathematical method of leverage to find outliers. The method first computes the residuals and then creates a hat matrix from the input data points according to Moammadi *et al.* (2012); Peter J. Rousseeuw, (2005):

$$H = X(X^t X)^{-1} X^t \tag{17}$$

Here, X denotes a matrix of dimensions m×n, where n corresponds to the number of inputs layer (rows), m is the model parameters (columns), and t represents the transpose matrix. The Hat values of the data are derived from the main diagonal of the matrix H.

$$Hat = diagonal(H) \tag{18}$$

the Williams plot is created to visually detect suspended data or outliers. The plot illustrates the correlation between Hat indices and standardized cross-validated residuals. These residuals are calculated as the variance between the represented or predicted values and the implemented data.

$$H^* = \frac{3(n+1)}{m} \tag{19}$$

A leverage value (H^*) of three is typically regarded as a 'cut-off' point, accepting points within a range of ±3 standard deviations from the mean (bounded by two horizontal red lines) to encompass 99% of normally distributed data Baghban *et al.* (2017); Hosseinzadeh & Hemmati-Sarapardeh, (2014); Mohammadi *et al.* (2012). the standardized cross-validated residuals are calculated from the data of the retention experimental and that calculated by the model

$$\left(R\_Norm\right)_i = \frac{(Retention_i^{exp} - Retention_i^{cal}}{\sqrt{var(Retention_i^{exp} - Retention_i^{cal})}}, i = 1,...m \tag{20}$$

If the majority of the data points fall within the ranges of 0 ≤ Hat ≤ H* and of − 3 ≤ R_Norm ≤ 3 it indicates that the model development and its predictions occur within the domain of applicability, which leads to a model statistically valid. Thus, we can identify ''Good High Leverage'' points in the domain of 0 ≤ Hat ≤ H* and − 3 ≤ R_Norm ≤3. However, points falling outside this range, with R_Norm <−3 or R_Norm>3 (whether greater or less than the H* value) are classified as model outliers or as "Bad High Leverage" points Baghban *et al.* (2017); Hosseinzadeh & Hemmati-Sarapardeh, (2014); Mohammadi *et al.* (2012).

Figure 5 represents Williams range plot of QSPR-ANN optimal neural model for the total phase. This plot contains 263/273 (96.34%) validated data points (red) and 10/273 (3.66%) suspected data points (blue). The critical leverage value is $H^* = \frac{3(n+1)}{m} = \frac{3(13+1)}{273} = 0.1539$. This indicates that the development of the optimal QSPR-ANN model and its prediction are within bounds leading to the optimal statistically valid neural model. Therefore, we can

affirm that there are "Good Haut Levier" points for the total phase.

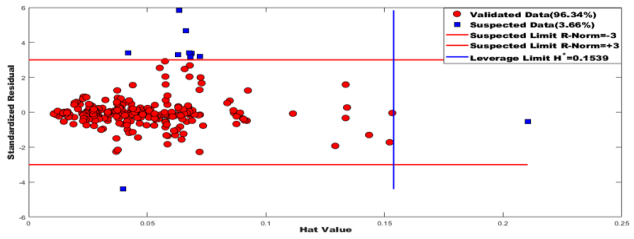


Figure 5. Williams range plot of QSPR-ANN optimal neural model for the total phase.

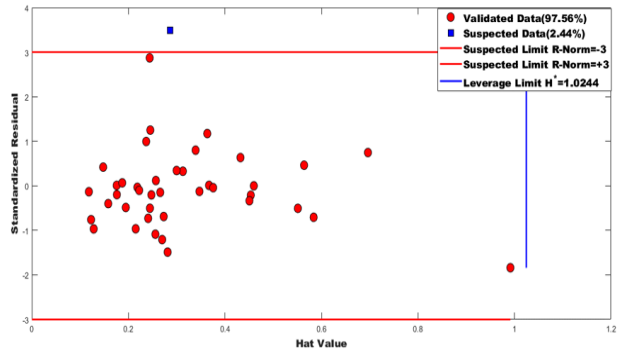


Figure 6. Williams range plot of QSPR-ANN optimal neural model for the testing phase.

Figure 6 represents Williams range plot of QSPR-ANN optimal neural model for testing phase. This plot contains 40/41 (97.56%) validated data points (red) and the blue line vertically and 1/41 (2.44%) suspected data points (blue). The critical leverage value is $H^* = \frac{3(n+1)}{m} = \frac{3(13+1)}{41} = 1.0244$. This indicates that the development of the optimal QSPR-ANN model and its prediction are within bounds leading to the optimal statistically valid neural model. Therefore, we can affirm that there are "Good Haut Levier" points for the test phase.

7. Conclusion

The present paper demonstrates the use of the QSPR-$ANN_{optimal}$ which was developed to predict the effect of adsorption on the retention of OCs by nanofiltration and reverse osmosis. The QSPR-ANN optimal can summarize interactions between the descriptors of OCs are Mw, *log Kow*, dipole moment, molecular length, surface area min, surface area max, polar surface area, and polarizability, the characteristics of the membranes are MWCO, SR NaCl, and contact angle, and the operating conditions is pressure.

An optimal QSPR-ANN is characterized by a structure (13 neurons in the input layer, 11 neurons in the hidden layer, and 1 neuron in the output layer). Training algorithm Levenberg-Marquard “train- LM" with activation function "Tansig " in the hidden layer and "Purlin" in the output layer. QSPR-$ANN_{optimal}$ showed good agreement between calculated and experimental data by the testing phase, with a coefficient of determination “R^2 = 0.9872” and a root mean square error “RMSE = 2.2743%”.

The sensitivity analysis conducted through the weight method successfully identified the true importance of all the utilized variables for the prediction of the effect of adsorption on the retention of OCs by RO/NF which is governed by two important interactions (hydrophobic/adsorption interaction and steric hindrance "sieving effect"), As a result, proves the correctness of the choice of variables appropriateness that were used in this study. The SR(NaCl) may be a possible lump parameter for the prediction of the effect of adsorption on the retention of OCs by NF/RO.

Applicability domain and Diagnostic analysis of the outliers of the optimized neural model (QSPR ANN optimal) demonstrated that both its development and its predictions are performed in the application domain. This substantiates the statistical validity optimal neural model. indicating the presence of "Good High Leverage" points during the test phase.

Abbreviation

OCs	Organic Compounds
QSPR	Quantitative Structure-Property Relationships
ANN	Artificial Neural Networks
RO	Reverse Osmosis
NF	Nanofiltration
MLPs	Multilayer Perceptron
Mw	molecular weight
log Kow	logarithm of the octanol-water partition coefficient
MWCO	molecular weight cut-off
SR NaCl	sodium chloride salt rejection
S_{min}	surface area min
S_{max}	surface area max
Min	minimum
Max	maximum
Mean	means
Std	standard deviations
RMSE	root mean squared error
R^2	determination coefficient
train-BR	Regularization-Bayesienne
train-LM	Levenberg-Marquard,
tansig	tangent hyperbolic
logsig	logarithmic sigmoid
purelin	pure-linear
MAE	mean absolute error
MPE	model predictive error
SEP	the standard error of prediction
RPD	residual predictive deviation
RER	range error ratio
MSE	the mean square error

MRSE	the mean relative squared error
Af	the accuracy factor
Bf	bias factor
IR	relative importance
Exp	experimental
Cal	calculated
W	weights
b	bais

Acknowledgements

The authors gratefully acknowledge the Ministry of Higher Education of Algeria (PRFU Projects N°A16N01UN260120220004) and the group of Laboratory of Biomaterials and Transport Phenomena in the University of Medea for their help throughout this project.

References

Ammi Y., Khaouane L. and Hanini S. (2015). Prediction of the rejection of organic compounds (neutral and ionic) by nanofiltration and reverse osmosis membranes using neural networks. *Korean Journal of Chemical Engineering*, ***32***(11), 2300–2310. https://doi.org/10.1007/s11814-015-0086-y

Ammi Y., Khaouane L. and Hanini S. (2018). A Model Based on Bootstrapped Neural Networks for Modeling the Removal of Organic Compounds by Nanofiltration and Reverse Osmosis Membranes. *Arabian Journal for Science and Engineering*, ***43***(11), 6271–6284. https://doi.org/10.1007/s13369-018-3484–8

Ammi Y., Khaouane L. and Hanini S. (2020). A Comparison of Neural Networks and Multiple Linear Regressions Models to Describe the Rejection of Micropollutants by Membranes. *Kemija u Industriji*, ***69***(3–4), 111–127. https://doi.org/10.15255/kui.2019.024

Ammi Y., Khaouane L. and Hanini S. (2021). Stacked neural networks for predicting the membranes performance by treating the pharmaceutical active compounds. *Neural Computing and Applications*, ***33***(19), 12429–12444. https://doi.org/10.1007/s00521-021-05876-0

Ammi Y., Moussa C.S. and Hanini S. (2023). Machine Learning and Neural Networks for Modelling the Retention of PPhACs by NF/RO. *Journal Kemija u Industriji*, ***72***, 11–12. https://doi.org/https://doi.org/10.21203/rs.3.rs-1120285/v1

Ammi Y., Salah H. and Latifa K. (2021). an artificial intelligence approach for modeling the rejection of anti-inflammatory drugs by nanofiltration and reverse osmosis membranes using kernel support vector machine and neural networks. *Comptes Rendus – Chimie*, ***24***(2), 243–254. https://doi.org/10.5802/crchim.76%3E

Arsuaga J.M., López-Muñoz M.J. and Sotto A. (2010). Correlation between retention and adsorption of phenolic compounds in nanofiltration membranes. *Desalination*, ***250***(2), 829–832. https://doi.org/10.1016/j.desal.2008.11.051

Baghban A., Mohammadi A.H. and Taleghani M.S. (2017). Rigorous modeling of CO2 equilibrium absorption in ionic liquids. *International Journal of Greenhouse Gas Control*, ***58***, 19–41. https://doi.org/10.1016/j.ijggc.2016.12.009

Bellona C., Drewes J.E., Xu P. and Amy G. (2004). Factors affecting the rejection of organic solutes during NF/RO treatment - A literature review. *Water Research*, ***38***(12), 2795–2809. https://doi.org/10.1016/j.watres.2004.03.034

Chan M.K., Shams A., Wang C.C., Lee P.Y., Jahani Y. and Mirbagheri S.A. (2023). Artificial Neural Network Model for Membrane Desalination: A Predictive and Optimization Study. *Computation*, ***11***(3). https://doi.org/10.3390/computation11030068

Comerton A.M., Andrews R.C., Bagley D.M. and Yang P. (2007). Membrane adsorption of endocrine disrupting compounds and pharmaceutically active compounds. *Journal of Membrane Science*, ***303***(1–2), 267–277. https://doi.org/10.1016/j.memsci.2007.07.025

Dahmani A., Ammi Y. and Hanini S. (2022). *Neural network for prediction solar radiation in Relizane region (Algeria) - Analysis study*. ***7***(2), 8–18.

Dolar D., Drašinac N., Košutić K., Škorić I. and Ašperger D. (2017). Adsorption of hydrophilic and hydrophobic pharmaceuticals on RO/NF membranes: Identification of interactions using FTIR. *Journal of Applied Polymer Science*, ***134***(5), 17–21. https://doi.org/10.1002/app.44426

Dolar D., Košutić K. and Ašperger D. (2013). Influence of adsorption of pharmaceuticals onto RO/NF membranes on their removal from water. *Water, Air, and Soil Pollution*, ***224***(1). https://doi.org/10.1007/s11270-012-1377-0

Fissa M.R., Lahiouel Y., Khaouane L. and Hanini S. (2023). Development of QSPR-ANN models for the estimation of critical properties of pure hydrocarbons. *Journal of Molecular Graphics and Modelling*, ***121***(108450). https://doi.org/https://doi.org/10.1016/j.jmgm.2023.108450

Gevrey M., Dimopoulos I. and Lek S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling*, ***160***(3), 249–264. https://doi.org/10.1016/S0304-3800(02)00257-0

Gur-Reznik S., Koren-Menashe I., Heller-Grossman L., Rufel O. and Dosoretz C.G. (2011). Influence of seasonal and operating conditions on the rejection of pharmaceutical active compounds by RO and NF membranes. *Desalination*, ***277***(1–3), 250–256. https://doi.org/10.1016/j.desal.2011.04.029

Hosseinzadeh M. and Hemmati-Sarapardeh A. (2014). Toward a predictive model for estimating viscosity of ternary mixtures containing ionic liquids. *Journal of Molecular Liquids*, ***200***(PB), 340–348. https://doi.org/10.1016/j.molliq.2014.10.033

Jawad J., Hawari A.H. and Javaid Zaidi S. (2021). Artificial neural network modeling of wastewater treatment and desalination using membrane processes: A review. *Chemical Engineering Journal*, ***419***(June 2020), 129540. https://doi.org/10.1016/j.cej.2021.129540

Khaouane L., Ammi Y. and Hanini S. (2017). Modeling the Retention of Organic Compounds by Nanofiltration and Reverse Osmosis Membranes Using Bootstrap Aggregated Neural Networks. *Arabian Journal for Science and Engineering*, ***42***(4), 1443–1453. https://doi.org/10.1007/s13369-016-2320-2

Kim S., Chu K.H., Al-Hamadani Y.A.J., Park C.M., Jang M., Kim D.H., Yu M., Heo J. and Yoon Y. (2018). Removal of contaminants of emerging concern by membranes in water

and wastewater: A review. *Chemical Engineering Journal*, ***335***, 896–914. https://doi.org/10.1016/j.cej.2017.11.044

Kiso Y., Sugiura Y., Kitao T. and Nishimura K. (2001). Effects of hydrophobicity and molecular size on rejection of aromatic pesticides with nanofiltration membranes. *Journal of Membrane Science*, ***192***(1–2), 1–10. https://doi.org/10.1016/S0376-7388(01)00411-2

Kratbi F., Ammi Y. and Hanini S. (2023). Support Vector Machines for Evaluating the Impact of the Forward Osmosis Membrane Characteristics on the Rejection of the Organic Molecules. *Kemija u Industriji*, ***72***(7–8). https://doi.org/10.15255/kui.2022.081

Libotean D., Giralt J., Rallo R., Cohen Y., Giralt F., Ridgway H.F., Rodriguez G. and Phipps D. (2008). Organic compounds passage through RO membranes. *Journal of Membrane Science*, ***313***(1–2), 23–43. https://doi.org/10.1016/j.memsci.2007.11.052

Liu Y. ling Wang X. mao Yang H. wei and Xie Y.F. (2018). Quantifying the influence of solute-membrane interactions on adsorption and rejection of pharmaceuticals by NF/RO membranes. *Journal of Membrane Science*, ***551***(January), 37–46. https://doi.org/10.1016/j.memsci.2018.01.035

Mahadeva R., Kumar M., Goel A., Patole S.P. and Manik G. (2023). A Novel AGPSO3-based ANN Prediction Approach: Application to the RO Desalination Plant. *Arabian Journal for Science and Engineering*, *April*. https://doi.org/10.1007/s13369-023-07631-0

Mahadeva R., Mahendra Kumar Shashikant P.P. and Manik G. (2022). Employing artificial neural network for accurate modeling, simulation and performance analysis of an RO-based desalination process. *Sustainable Computing: Informatics and Systems*, ***35***. https://doi.org/https://doi.org/10.1016/j.suscom.2022.100735

Mohammad A.A., Soudan B., Mahmoud M.S., Sayed E.T., AlMallahi M.N., Inayat A., Radi M.Al. and Olabi A.G. (2022). Progress of artificial neural networks applications in hydrogen production. *Chemical Engineering Research and Design*, ***182***, 66–86. https://doi.org/https://doi.org/10.1016/j.cherd.2022.03.030

Mohammadi A.H., Eslamimanesh A., Gharagheizi F. and Richon D. (2012). A novel method for evaluation of asphaltene precipitation titration data. *Chemical Engineering Science*, ***78***, 181–185. https://doi.org/10.1016/j.ces.2012.05.009

Peter J. Rousseeuw. (2005). *Outlier Diagnostics. Wiley Series in Probability and Statistics*. https://doi.org/10.1002/0471725382.ch6

Rehab I.A., Elsheikh A.H., Mohamed Elasyed A.E. and Mohammed A.A.A. (2022). Chapter one - Basics of artificial neural networks. *Artificial Neural Networks for Renewable Energy Systems and Real-World Applications*, 1–10. https://doi.org/https://doi.org/10.1016/B978-0-12-820793-2.00002-1

Sediri M., Hanini S., Laidi M., Turki S.A., Cherifi H. and Mabrouk H. (2017). Artificial neural networks modeling of dynamic adsorption from aqueous solution. *Moroccan Journal of Chemistry*, ***5***(2), 2–5.

Shahmansouri A. and Bellona C. (2013). Application of quantitative structure-property relationships (QSPRs) to predict the rejection of organic solutes by nanofiltration. *Separation and Purification Technology*, ***118***, 627–638. https://doi.org/10.1016/j.seppur.2013.07.050

Teychene B., Chi F., Chokki J., Darracq G., Baron J., Joyeux M. and Gallard H. (2020). Investigation of polar mobile organic compounds (PMOC) removal by reverse osmosis and nanofiltration: rejection mechanism modelling using decision tree. *Water Science and Technology: Water Supply*, ***20***(3), 975–983. https://doi.org/10.2166/ws.2020.020

Viscarra Rossel R.A., McGlynn R.N. and McBratney A.B. (2006). Determining the composition of mineral-organic mixes using UV-vis-NIR diffuse reflectance spectroscopy. *Geoderma*, ***137***(1–2), 70–82. https://doi.org/10.1016/j.geoderma.2006.07.004

Wang R., Jiang J., Pan Y., Cao H. and Cui Y. (2009). Prediction of impact sensitivity of nitro energetic compounds by neural network based on electrotopological-state indices. *Journal of Hazardous Materials*, ***166***(1), 155–186. https://doi.org/10.1016/j.jhazmat.2008.11.005

Yangali-Quintanilla V., Verliefde A., Kim T.U., Sadmani A., Kennedy M. and Amy G. (2009). Artificial neural network models based on QSAR for predicting rejection of neutral organic compounds by polyamide nanofiltration and reverse osmosis membranes. *Journal of Membrane Science*, ***342***(1–2), 251–262. https://doi.org/10.1016/j.memsci.2009.06.048.