SUPPORTING INFORMATION

Artificial neural network-based QSAR model for predicting degradation techniques of pharmaceutical contaminants in water bodies with experimental verification.

Jhon-Alex Gonzalez-Amaya,[1] Andrea-Nadith Niño-Colmenares,[1] Andrés-Felipe Cárdenas Rodríguez [1] and James Guevara-Pulido*[1]
INQA Research Group,[1] Química Farmacéutica,[1] Universidad El Bosque,[1] Bogotá D.C, Colombia.
*joguevara@unbosque.edu.co

QSAR MODELS

Table 1. Models performed in validation of the Ozonization method (Met 1).

Models	**Selected descriptors**							**Nodes**	**R^2**
M1Met1	ALogP	ALogp2	ATS8v	VE3_Dt	MDEC-13	MLFER_A	VE2_D	100	0.314
M2Met1	ALogP	ALogp2	ATS8v	VE3_Dt	MDEC-13	MLFER_A	VE2_D	500	0.467
M3Met1	ALogP	MDEC-13	AMR	C3SP2	SpMax_Dt	GGI9	MDEC-23	200	0.688
M4Met1	ALogP	MDEC-13	AMR	C3SP2	SpMax_Dt	GGI9	MDEC-23	500	0.720
M5Met1	VE1_DzZ	nAtom	nAtomP	TopoPSA	fragC	-	-	300	0.752
M6Met1	ALogp2	VE1_DzZ	nAtomP	TopoPSA	fragC	-	-	200	0.729
M7Met1	ALogp2	VE1_DzZ	nAtomP	TopoPSA	fragC			500	0.650

Figure 1. Heatmap of model M6Met1

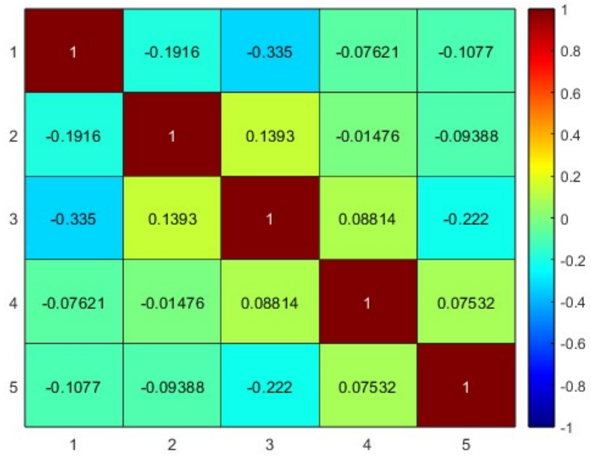


Figure 2. Coefficient of determination (R^2) obtained in validated model M6Met1

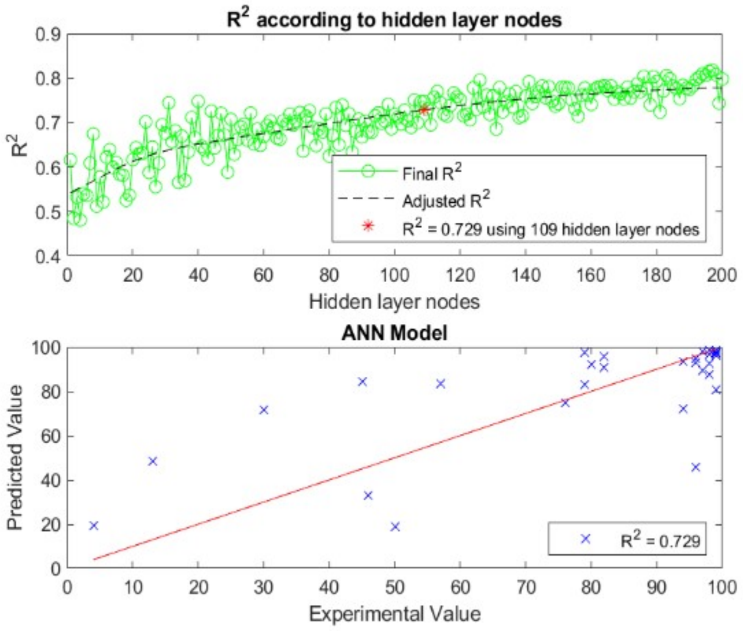


Figure 3. Comparison of experimental data vs. predicted data from model M6Met1

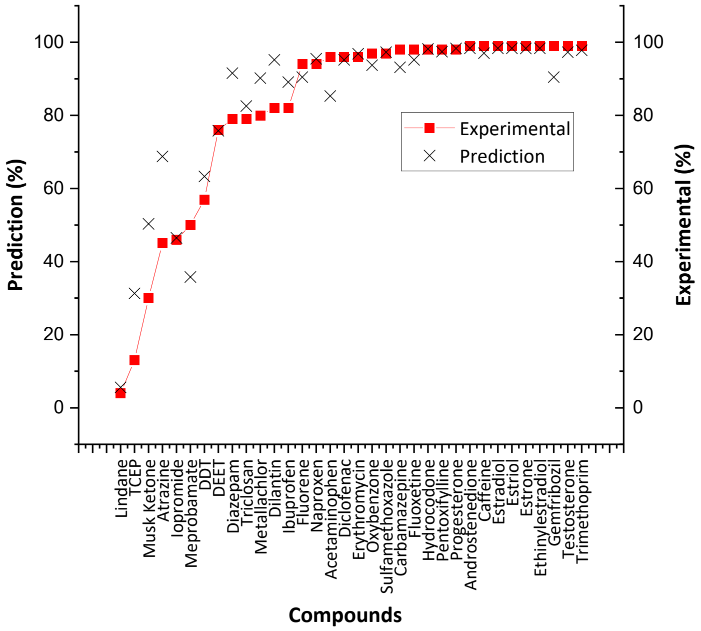
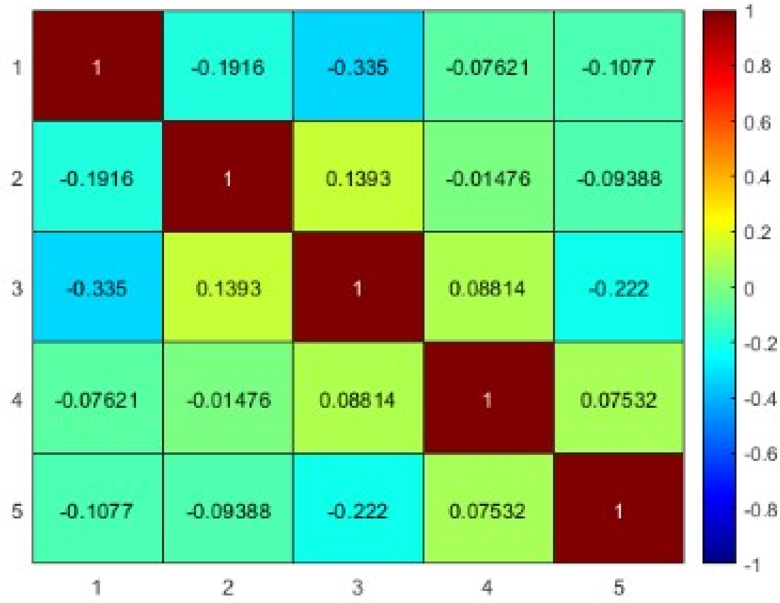


Table 2. Models performed in validation of the Ozonization + H_2O_2 method (Met 2)

Models	**Selected descriptors**					**Nodes**	**R^2**
M1Met2	ALogp2	VE1_DzZ	nAtomP	TopoPSA	fragC	500	0.816

Figure 4. Heatmap of model M1Met2



	1	2	3	4	5
1	1	-0.1916	-0.335	-0.07621	-0.1077
2	-0.1916	1	0.1393	-0.01476	-0.09388
3	-0.335	0.1393	1	0.08814	-0.222
4	-0.07621	-0.01476	0.08814	1	0.07532
5	-0.1077	-0.09388	-0.222	0.07532	1

Figure 5. Coefficient of determination (R^2) obtained in validated model M1Met2

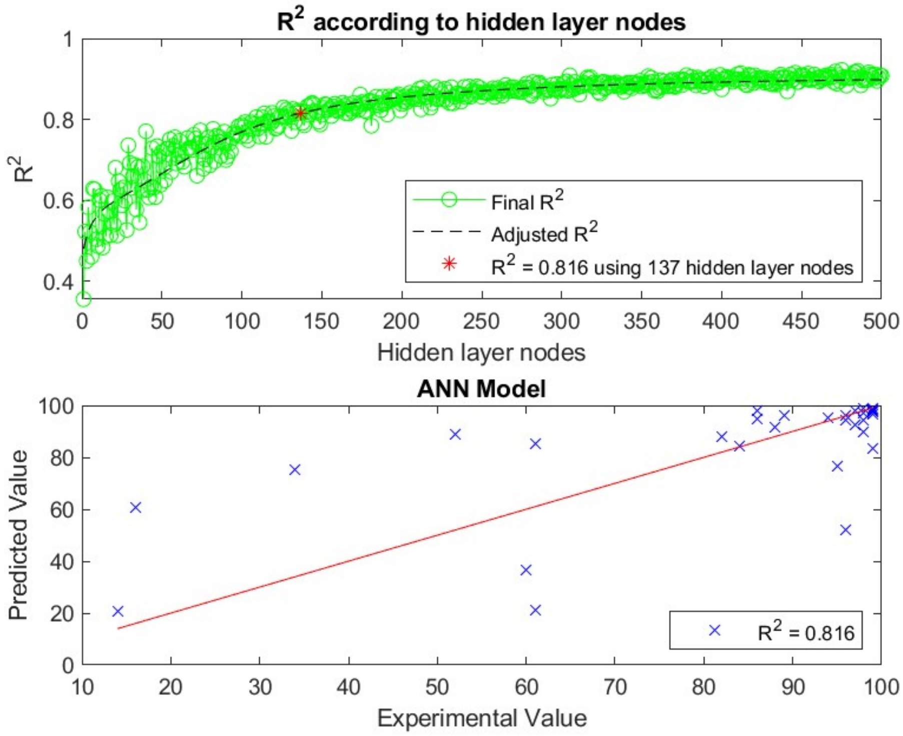


Figure 6. Comparison of experimental data vs. predicted data from model M1Met2

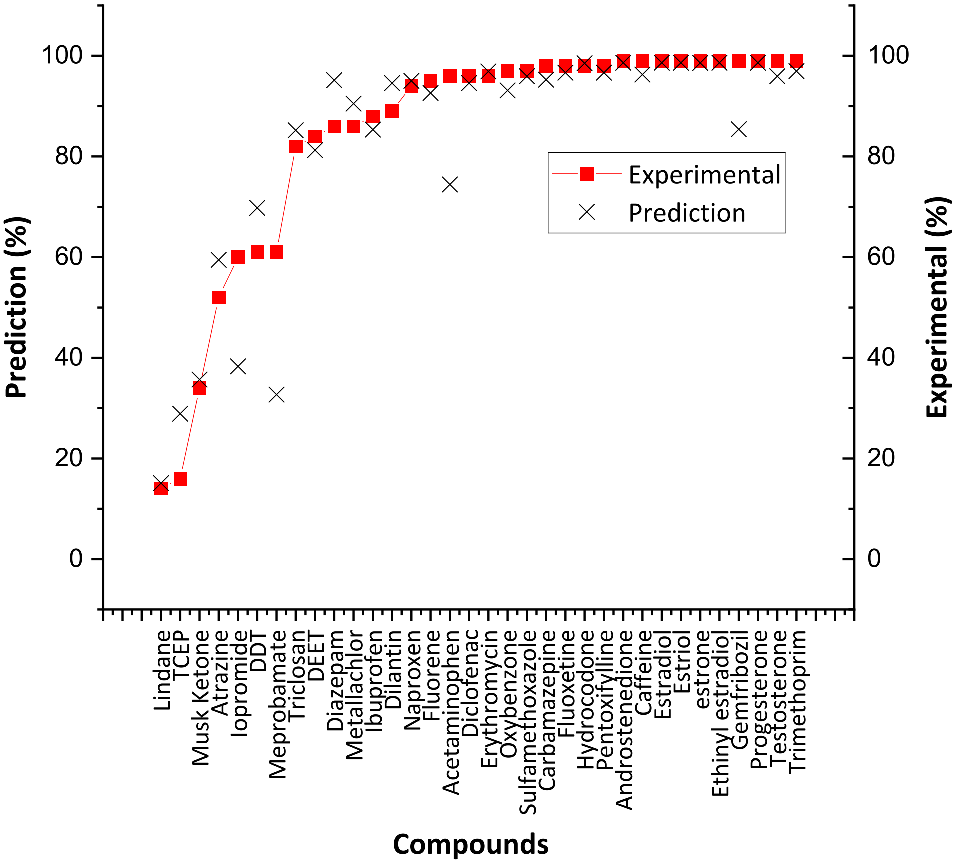


Table 3. Models performed in validation of the activated carbon method (Met 3)

Models	**Selected descriptors**					**Nodes**	**R^2**
M1Met3	ALogP	ALogp2	nAtomLC	VE3_Dt	WTPT-5	200	0.491
M2Met3	ALogP	fragC	WTPT-4	VE3_Dt	XLogP	250	0.492
M3Met3	ALogP	fragC	WTPT-4	VE3_Dt	MDEC-13	300	0.598
M4Met3	ALogP	ALogp2	nAtomLC	VE3_Dt	WTPT-5	300	0.490
M5Met3	ALogP	ALogp2	nAtomLC	VE3_Dt	BCUTw-1h	300	0.687
M6Met3	ALogP	ALogp2	nAtomLC	VE3_Dt	BCUTw-1h	1000	0.689
M7Met3	ALogP	fragC	WTPT-4	VE3_Dt	MDEC-13	800	0.588
M8Met3	ALogP	ATS3v	nAtomLC	VE3_D	BCUTw-1h	500	0.158
M9Met3	ALogp2	fragC	XLogP	VE3_Dt	BCUTw-1h	500	0.574
M10Met3	ALogp2	fragC	XLogP	VE3_Dt	BCUTw-1h	800	0.585
M11Met3	TPSA	TopoPSA	ATS0i	ALogp2	fragC	300	0.183
M12Met3	ALogP	apol	nAtomLC	VE3_Dt	BCUTw-1h	300	0.514
M13Met3	nAtomLC	BCUTw-1h	ALogP	VE3_Dt	ALogp2	1500	0.694

M14Met3	ALogP	apol	nAtomLC	VE3_Dt	BCUTw-1h	800	0.588
M15Met3	nAtomLC	BCUTw-1h	ALogP	VE3_Dt	ALogp2	2500	0.647
M16Met3	nAtomLC	BCUTw-1h	ALogP	VE3_Dt		800	0.444
M17Met3	AMR	nAtom	nAtomP	TopoPSA		300	0.729*
M18Met3	ALogP	apol	nAtomLC	VE3_Dt		500	0.361
M19Met3	VE3_Dt	nAtom	nAtomP	TopoPSA	FragC	300	1.124*
M20Met3	VE3_Dt	nAtom	nAtomP	TopoPSA		300	0.882*
M21Met3	VE1_DzZ	nAtom	nAtomP	TopoPSA		300	0.679
M22Met3	fragC	VE3_Dt	nAtom	nAtomP	TopoPSA	300	1.122*
M23Met3	nAtomP	TopoPSA	fragC	VE3_Dt	VE1_DzZ	150	0.935*
M24Met3	nAtomP	TopoPSA	fragC	VE3_Dt	VE1_DzZ	800	0.468
M25Met3	nAtomP	TopoPSA	fragC	VE3_Dt	VE1_DzZ	300	0.977*
M26Met3	nAtomP	TopoPSA	fragC	VE3_Dt		100	0.944*
M27Met3	nAtomP	TopoPSA	fragC	VE1_DzZ		100	0.724*
M28Met3	nAtomP	TopoPSA	fragC	VE3_Dt		500	0.970

*Note: * Values not taken into account due to high collinearity (Pearson)*

Figure 7. Heatmap of model M28Met3

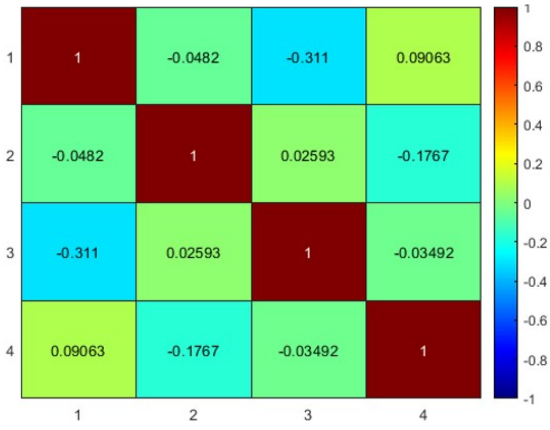


Figure 8. Coefficient of determination (R^2) obtained in validated model M28Met3

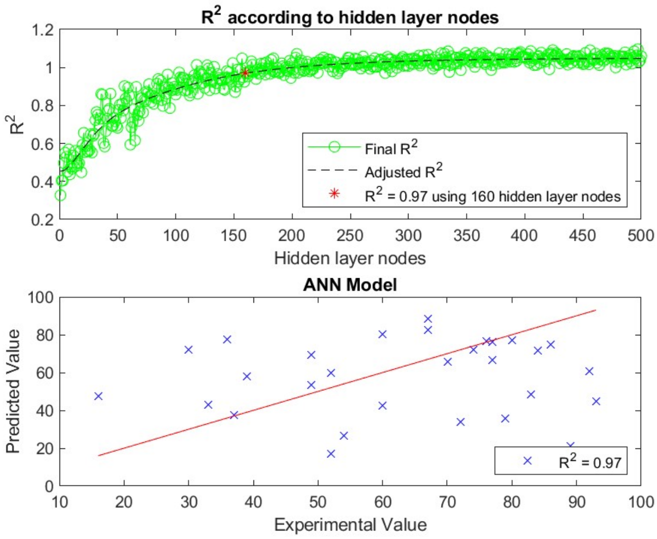


Figure 9. Comparison of experimental data vs. predicted data from model M28Met3

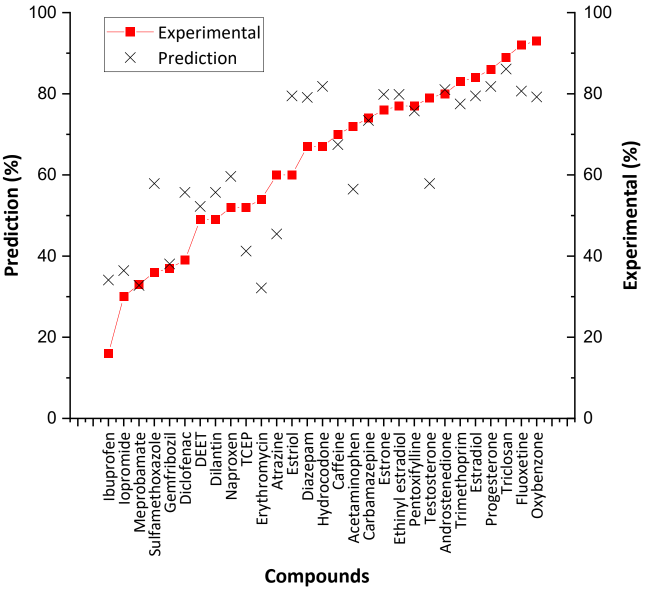


Table 4. Models performed in validation of the UV radiation method (Met 4)

Models	**Selected descriptors**				**Nodes**	**R^2**
M1Met4	VE1_DzZ	nAtomP	ALogP	AATS8v	100	0.495
M2Met4	VE1_DzZ	nAtomP	TopoPSA	AATS8v	100	0.744
M3Met4	VE1_DzZ	nAtomP	TopoPSA	AATS8v	700	0.760

Figure 10. Heatmap of model M3Met4

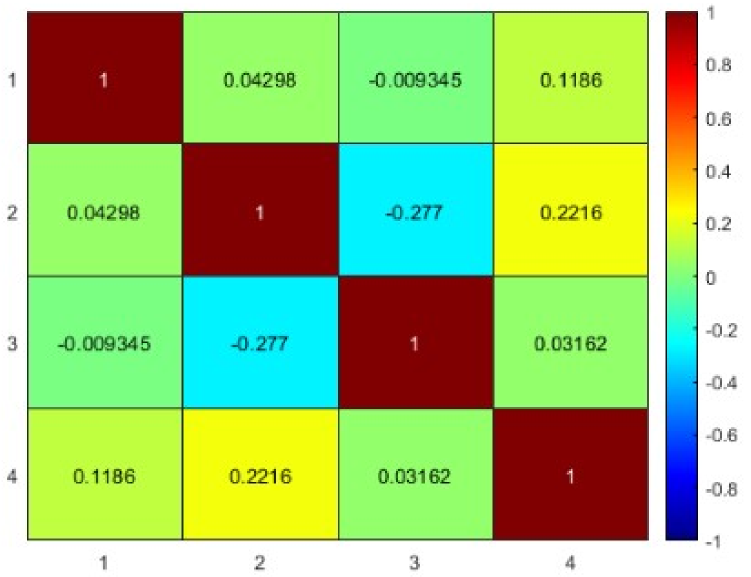


Figure 11. Coefficient of determination (R^2) obtained in validated model M3Met4

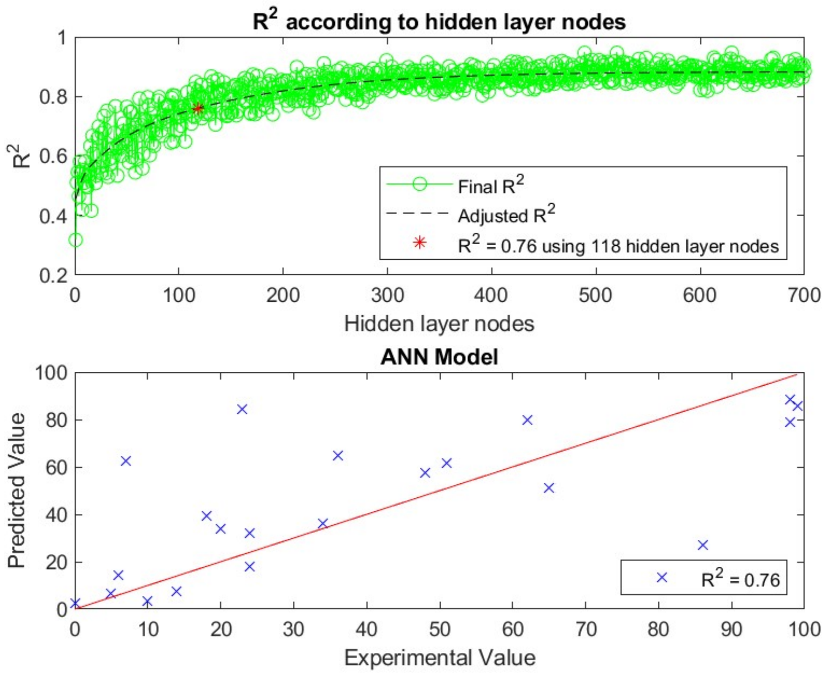


Figure 12. Comparison of experimental data vs. predicted data from model M3Met4

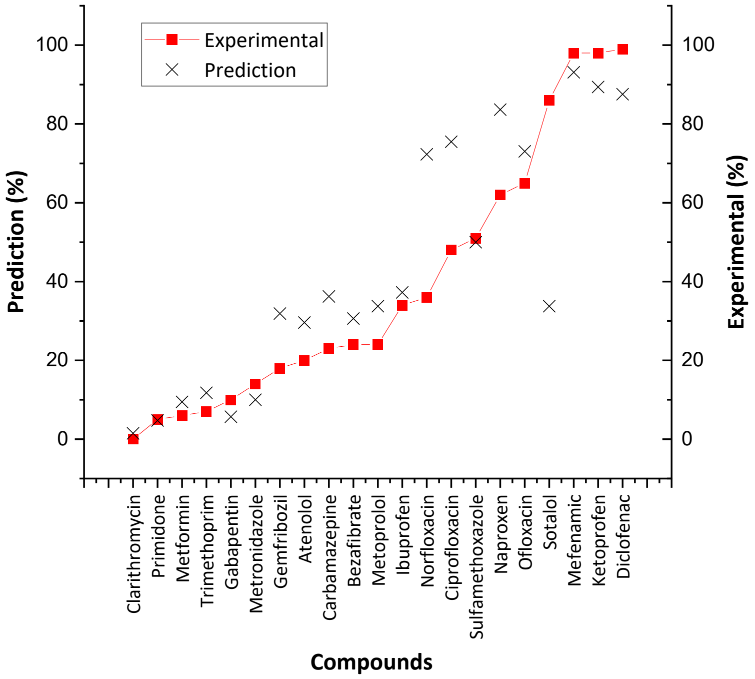


Table 5. Models performed in validation of the Dark-Fenton method (Met 5)

Models	Selected descriptors				Nodes	R^2
M1Met5	ALogP	fragC	VE1_DzZ	nAtomP	100	0.236
M2Met5	ALogP	fragC	VE1_DzZ	nAtomP	500	0.259
M3Met5	fragC	ALogP	nAtomP	VE1_DzZ	150	0.241
M4Met5	nAtomP	VE1_DzZ	fragC	ALogP	150	0.242
M5Met5	ALogP	nAtomP	VE3_Dt	VE1_DzZ	100	0.333
M6Met5	ALogP	nAtomP	AMR	nN	100	0.522
M7Met5	VE1_DzZ	nAtomP	TopoPSA	AATS8v	100	0.735
M8Met5	ALogP	nAtomP	AMR	nN	500	0.519
M9Met5	VE1_DzZ	nAtomP	TopoPSA	AATS8v	500	0.774
M10Met5	VE1_DzZ	nAtomP	TopoPSA	AATS8v	850	0.813
M11Met5	VE1_DzZ	nAtomP	TopoPSA	BCUTp-1h	100	0.406
M12Met5	VE1_DzZ	nAtomP	TopoPSA	GATS4m	200	0.255
M13Met5	VE1_DzZ	nAtomP	TopoPSA	AATSC1c	200	0.248
M14Met5	VE1_DzZ	nAtomP	TopoPSA	AATS4e	200	0.376
M15Met5	VE1_DzZ	nAtomP	TopoPSA	BCUTp-1h	850	0.410
M16Met5	VE1_DzZ	nAtomP	TopoPSA	AATS4e	850	0.371

Figure 13. Heatmap of model M10Met5

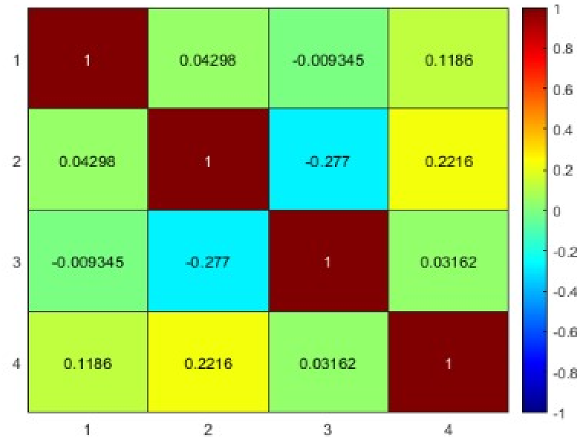


Figure 14. Coefficient of determination (R^2) obtained in validated model M10Met5

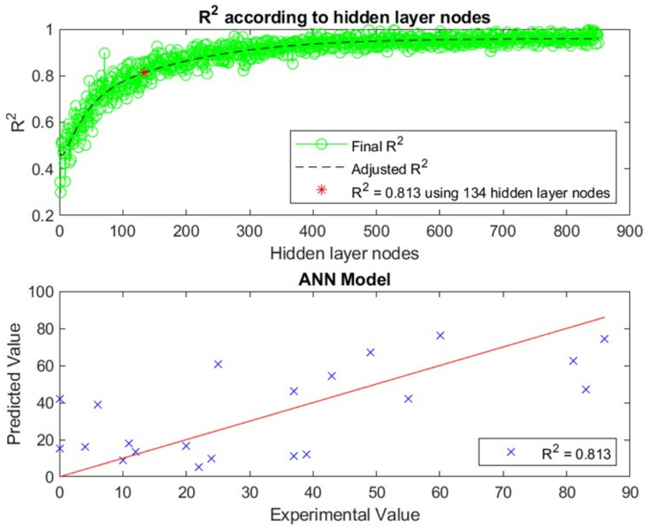


Figure 15. Comparison of experimental data vs. predicted data from model M10Met5

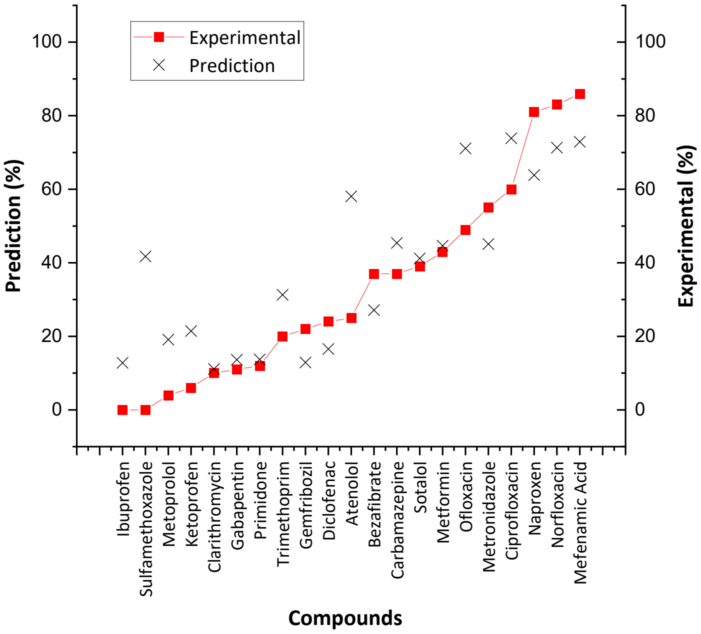
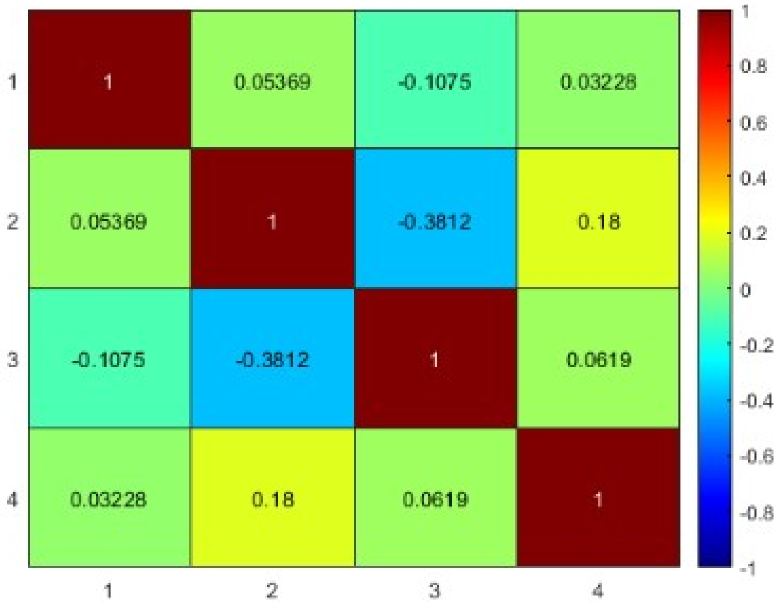


Table 6. Models performed in validation of the Photo-Fenton + H_2O_2 method (Met 6)

Modelos	Descriptores seleccionados						Nodos	R^2
M1Met6	ALogp2	R_TpiPCTPC	fragC	bpol	VE1_DzZ	-	250	0.559
M2Met6	ALogp	ALogp2	R_TpiPCTPC	fragC	bpol	VE1_DzZ	250	0.266
M3Met6	ALogP	bpol	Sv	fragC	McGowan_Volume	MLFER_A	200	0.462
M4Met6	ALogP	bpol	Sv	fragC	R_TpiPCTPC	-	200	0.574
M5Met6	ALogP	bpol	Sv	fragC	R_TpiPCTPC	-	500	0.583
M6Met6	ALogP	ALogp2	SP-7	Mi	FragC	HybRatio	250	0.324
M7Met6	ALogP	VE1_DzZ	bpol	JGI6	nBondsD	-	250	0.225
M8Met6	VE1_DzZ	nAtom	nAtomP	TopoPSA	fragC	-	300	0.819
M9Met6	AMR	nAtomP	TopoPSA	ALogP	VE1_DzZ	-	300	0.369
M10Met6	VE1_DzZ	nAtomP	fragC	ALogP	nN		300	0.409
M11Met6	VE1_DzZ	nAtomP	fragC	ALogP	-	-	200	0.485
M12Met6	VE1_DzZ	ALogP	nAtomP	Mi	MLogP		200	0.393
M13Met6	VE1_DzZ	nAtomP	TopoPSA	fragC	-	-	200	0.697
M14Met6	VE1_DzZ	nAtomP	fragC	ALogP	-	-	800	0.472
M15Met6	ALogP	VE1_DzZ	nAtomP	fragC	-	-	100	0.487
M16Met6	ALogP	VE1_DzZ	nAtomP	fragC	-	-	600	0.487
M17Met6	VE1_DzZ	ALogP	nAtomP	Mi	MLogP	-	600	0.391
M18Met6	ALogp2	VE1_DzZ	nAtomP	TopoPSA	-	-	100	0.172
M19Met6	ALogP	VE1_DzZ	nAtomP	fragC	-	-	200	0.489
M20Met6	VE1_DzZ	nAtomP	TopoPSA	AATS8v	-	-	100	0.771

Figure 16. Heatmap of model M20Met6



	1	2	3	4
1	1	0.05369	-0.1075	0.03228
2	0.05369	1	-0.3812	0.18
3	-0.1075	-0.3812	1	0.0619
4	0.03228	0.18	0.0619	1

Figure 17. Coefficient of determination (R^2) obtained in validated model M20Met6

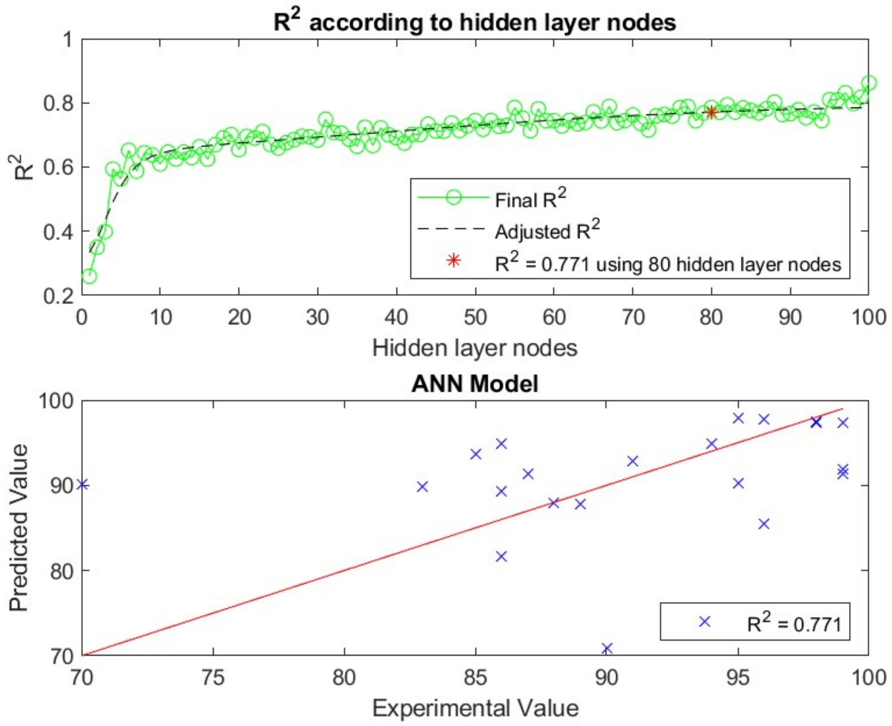


Figure 18. Comparison of experimental data vs. predicted data from model M20Met6

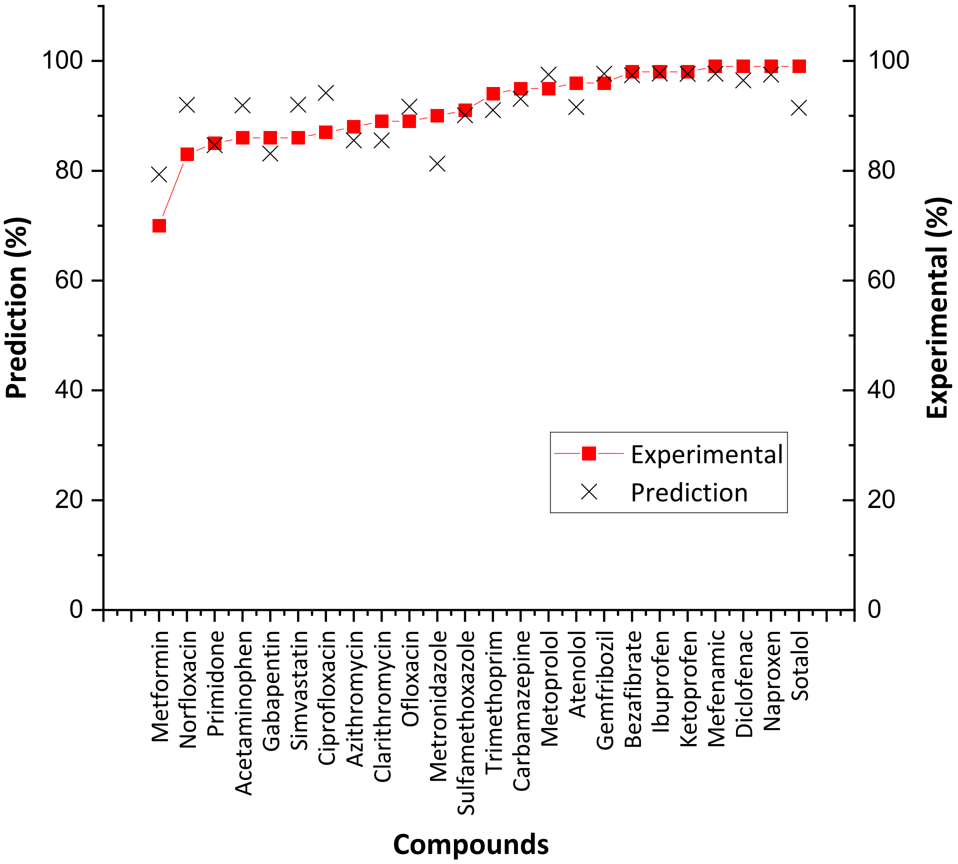
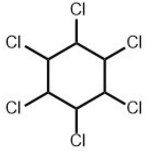
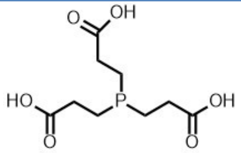
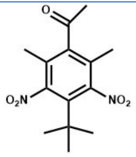
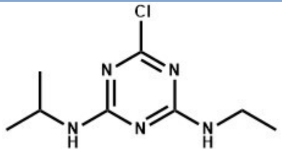
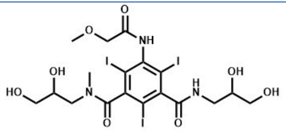
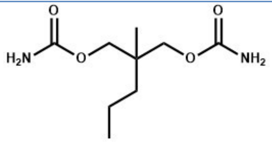
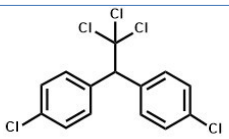
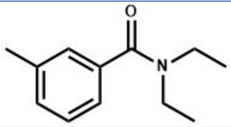
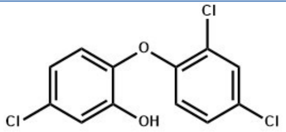
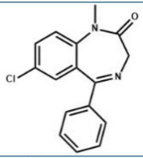
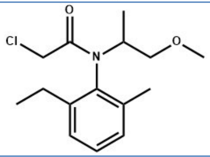
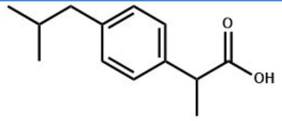
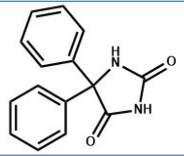
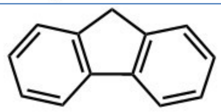
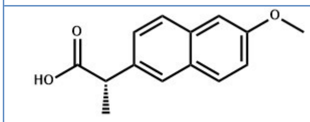
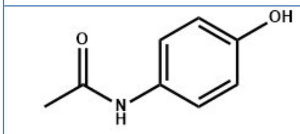
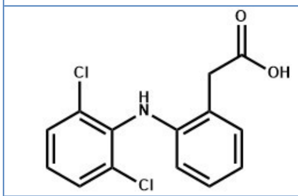
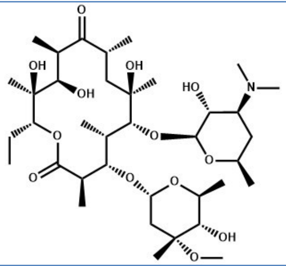
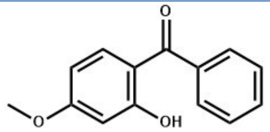
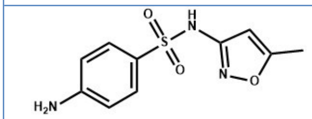
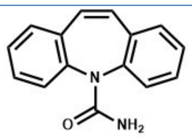
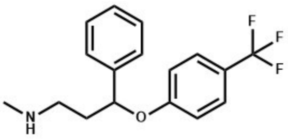
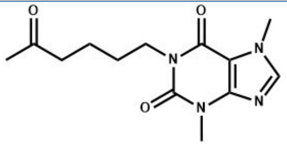
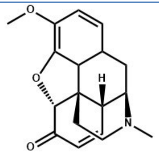
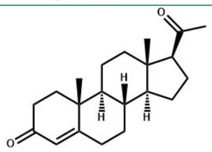
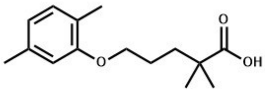
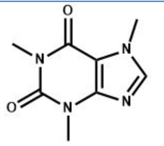
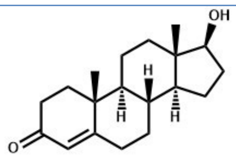
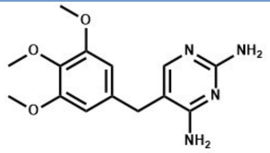
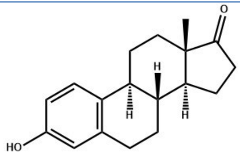
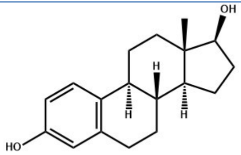
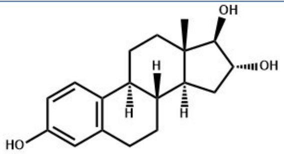
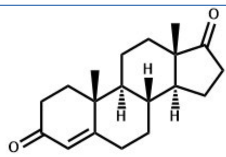
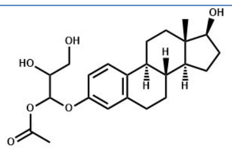


Table 7 Molecules used to create the M6Met1 model.

Molecules	Experimental degradation percentage	Degradation percentage ANN-INQA Predicted
	4	6
	13	31
	30	36
	45	46
	46	50
	50	63
	57	69
	76	76
	79	83
	79	85
	80	89

	82	90
	82	90
	94	90
	94	92
	96	93
	96	94
	96	95
	97	95
	97	95
	98	95
	98	97

	98	97
	98	97
	98	97
	99	97
	99	98
	99	98
	99	98
	99	98
	99	98
	99	98
	99	98
	99	98

Laboratorios del Departamento de Química y Programa de Química Farmacéutica

Manuscript SI

<Sample Information>

Sample Name	: Cefalexina		
Sample ID	: Cefalexina 38.8		
Data Filename	: CE-5.lcd		
Method Filename	: Quercetina nueva columna.lcm		
Batch Filename	:		
Vial #	: 1-1	Sample Type	: Unknown
Injection Volume	: 10 uL		
Date Acquired	: 17/01/2024 11:56:32	Acquired by	: System Administrator
Date Processed	: 18/01/2024 10:30:13	Processed by	: System Administrator

<Chromatogram>

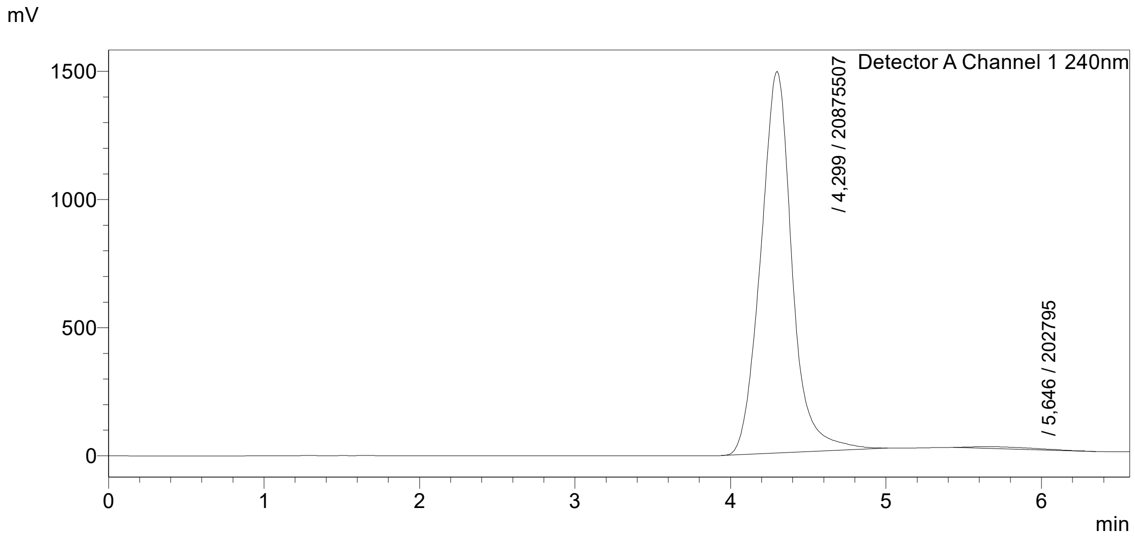


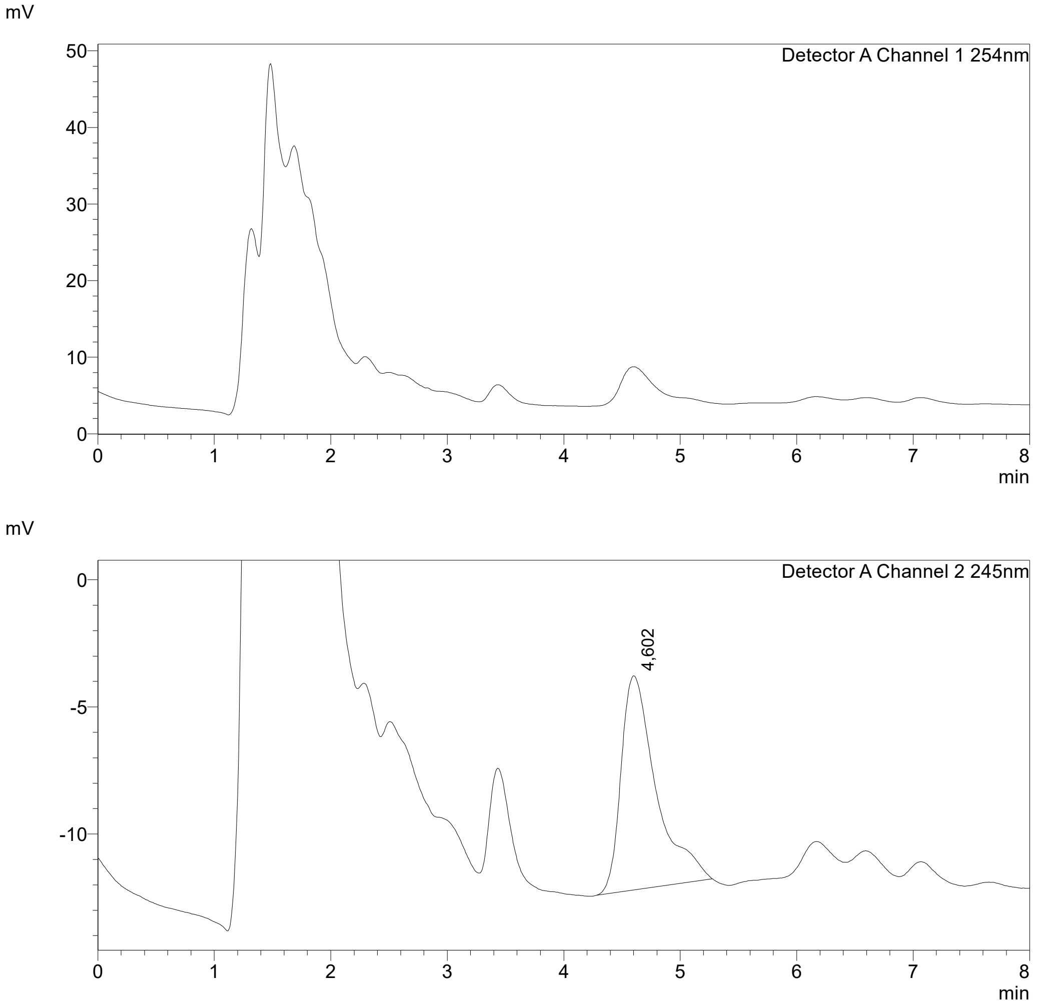
Figure 19

SHIMADZU LabSolutions Analysis Report

<Sample Information>

Sample Name	: CEFA OZONO		
Sample ID	: Cefa - 92-8-AACN-1.5		
Data Filename	: Cefa - 92-8-AACN-1.5.lcd		
Method Filename	: Cefa - 92-8-AACN-1.5.lcm		
Batch Filename	:		
Vial #	: 1-46	Sample Type	: Unknown
Injection Volume	: 10 uL		
Date Acquired	: 26/01/2024 11:51:31	Acquired by	: System Administrator
Date Processed	: 26/01/2024 11:59:32	Processed by	: System Administrator

<Chromatogram>



<Peak Table>

Detector A Channel 1 254nm

Figure 20

Development of artificial neural network (ANN)

An artificial neural network (ANN) with backpropagation was built, where each input node received a particular molecular descriptor, and the output node generated the predicted response variable (degradation%) (Figure 21). The number of nodes in the hidden layer was modified after each computation, considering the leave-one-out cross-validation method and the determination coefficient (R2) to determine the best-fitting model. Thus, the number of hidden nodes that yielded the best-fitting model (R2 closest to 1) was chosen for the final prediction.

The following section will describe in greater detail the structure of the proposed network, the validation method, and the validation metrics used.

Neural Network Structure

The proposed neural network (Figure 21) features an initializing algorithm followed by feed-forward and backpropagation algorithms.

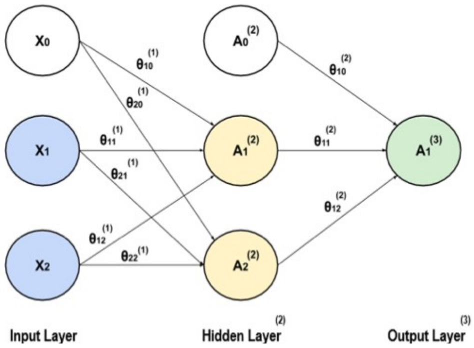


Figure 21 Input Descriptors Output degradation %

Step 1: Initialization:

- Define weights (θ) and bias nodes (X_0, A_0)
- Initialize weights using a random number, different than zero.
- Set bias nodes equal to 1

Step 1: Initialization:

Figure 1. Structure of Neural Network

- Define weights (θ) and bias nodes (X_0, A_0)
- Initialize weights using a random number, different than zero.
- Set bias nodes equal to 1

Step 2: Feed-forward:

- Calculate hidden nodes values

$$A_1^{(2)} = g(\theta_{10}^{(1)}X_0 + \theta_{11}^{(1)}X_1 + \theta_{12}^{(1)}X_2 \quad (1\text{-}1)$$

$$A_2^{(2)} = g(\theta_{20}^{(1)}X_0 + \theta_{21}^{(1)}X_1 + \theta_{22}^{(1)}X_2) \quad (1\text{-}2)$$

- Activate hidden nodes using the Sigmoid function

$$g\,(z) = \frac{1}{1+e^{-z}} \qquad (1\text{-}3)$$

Where: $z = \theta_0 X_0 + \theta_1 X_1 + \cdots + \theta_n X_n$

- Calculate output node values.

$$A_1^{(3)} = g(\theta_{10}^{(2)} A_0^{(2)} + \theta_{11}^{(2)} A_1^{(2)} + \theta_{12}^{(2)} A_2^{(2)}) \quad (1\text{-}4)$$

Calculate cost function to determine adjustment of output according to the descriptors (error) and include the regularization parameter (λ) to reduce overfitting.

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} \quad y^{(i)} log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) log\ log\ \left(1 - h_\theta\left(x^{(i)}\right)\right)\right] + \frac{\lambda}{2m}\sum_{j=1}^{n} \quad \theta_j^2 \quad (1\text{-}5)$$

Step 3:

Backpropagation:

Adjust weights (θ) based on error measured between output of the network and input by applying the gradient descent algorithm to the cost function. This step was carried out using equations 1-6 to 1-9, based on Figure 22.

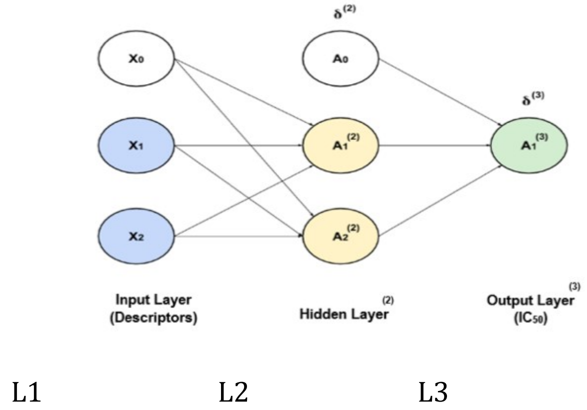


Figure 22

$$\delta_j^{(3)} = A_j^{(3)} - Y_j \quad (1\text{-}6)$$

$$\delta_j^{(2)} = \theta^{(2)}\delta^{(3)}.* g'(\xi^{(2)}) \quad (1\text{-}7)$$

Where:

$$\delta_j^{(L)} = error\ of\ node\ j\ in\ L\ layer$$

$$Y_j = value\ of\ node\ j\ in\ output\ layer$$

$$\xi^{(L)} = \left[1\ z_1^{(L)}\ \ldots\ z_j^{(L)}\right]$$

- Determine the gradient of the cost function

$$\frac{\partial}{\partial\theta_{ij}^{(L)}} J(\theta) = \frac{1}{m}\Delta_{ij}^{(L)}\ if\ j = 0 \quad (1\text{-}8)$$

$$\frac{\partial}{\partial\theta_{ij}^{(L)}} J(\theta) = \frac{1}{m}\left[\Delta_{ij}^{(L)} + \lambda\theta_{ij}^{(L)}\right]\ if\ j \neq 0 \quad (1\text{-}9)$$

Where:

$$\Delta_{ij}^{(L)} \coloneqq \Delta_{ij}^{(L)} + A_j^{(L)}\delta_i^{(L+1)}\ (initialize\ \Delta_{ij}^{(L)} = 0)$$

Validation Method

An exhaustive cross-validation was carried out by the leave-one-out method, in which one instance (molecular descriptors and experimental degradation % of a single molecule) is used as a test set, while all other instances are used as a training set. This process is applied to each molecule of the set; thus, the weight of each cross-validation iteration was determined, and the average of those weights was calculated to predict degradation % values.

Now, the leave-one-out method was selected for validation because the error of estimation didn't vary depending on the data used for the test set and validation, which indicated that the error of estimation was more stable in contrast to other cross-validation methods like k-fold or Montecarlo. Although this method entails a greater computational cost than the other cross-validation methods mentioned, it is commonly used for small data sets such as the one used in this study (36 molecules).