December 2022. Volume 8. Number 4

Original Article

A Novel Methodology for Human Plasma Protein Binding: Prediction, Validation, and Applicability Domain

Affaf Khaouane[1*], Samira Ferhat[1], Salah Hanini[1]

1. Laboratory of Biomaterial and Transport Phenomena, University of Médéa, Médéa, Algeria.

*** Corresponding Author:**
Affaf Khaouane, PhD.
Address: Laboratory of Biomaterial and Transport Phenomena, University of Médéa, Médéa, Algeria.
Phone: +213 (791) 79 32 84
E-mail: affoufa80@gmail.com

Article info:
Received: 11 Apr 2022
Accepted: 01 Jun 2022

Keywords:
Quantitative structure-activity relationship, Convolutional neural network, Feed-forward neural network, Protein binding.

ABSTRACT

Background: Plasma protein binding is a key component in drug therapy as it affects the pharmacokinetics and pharmacodynamics of drugs.

Objectives: This study aimed to predict the fraction of plasma protein binding.

Methods: A quantitative structure-activity relationship, convolutional neural network, and feed-forward neural network (QSAR-CNN-FFNN) methodology was used. CNN was used for feature selection, which is known as a difficult task in QSAR studies. The values of the descriptors acquired without the preprocessing procedures were rearranged into matrices, and features from a deep fully connected layer of a pre-trained CNN (ALEXNET) were extracted. Then, the latest features learned from the CNN layers were flattened out and passed through an FFNN to make predictions.

Results: The external accuracy of the validation set (Q^2=0.945, RMSE=0.085) showed the performance of this methodology. Another extremely favorable circumstance of this method is that it does not take a lot of time (only a few minutes) compared to the QSAR-Wrapper-FFNN method (days of hard work and concentration) and it automatically gives us the characteristics that are the best representations of our input.

Conclusion: We can say that this model can be used to predict the fraction of human plasma protein binding for drugs that have not been tested to avoid chemical synthesis and reduce expansive laboratory tests.

Citation Khaouane A, Ferhat S, Hanini S. A Novel Methodology for Human Plasma Protein Binding: Prediction, Validation, and Applicability Domain. Pharmaceutical and Biomedical Research. 2022; 8(4):311-322. http://dx.doi.org/10.32598/PBR.8.4.1086.1

http://dx.doi.org/10.32598/PBR.8.4.1086.1

Introduction

Building a quantitative structure-activity relationship (QSAR) model, known as in silico model, is an important step in drug discovery. It allows us to avoid chemical synthesis and reduce expansive and tedious laboratory tests [1, 2]. Different QSAR approaches have been developed over the past few years [3-5]. These approaches can determine the quantitative relations between the variation in the values of calculated descriptors and the biological activity of a series of chemical molecules [6]. Thus, once a correlation has been established, it can be used to predict the property or biological activity of new structures [7, 8]. In QSAR studies, selecting a few relevant descriptors from a large number of features is a difficult task. Feature selection techniques are applied to reduce the number of attributes in the dataset by choosing features that will give us better accuracy with fewer data [9-11]. It also reduces the overfitting and the overtraining risk [12]. Feature selection methods can be divided into three categories: filter, wrapper, and embedded methods. Filters are applied independently of the mapping method used. They are executed before the mapping, to reduce the number of descriptors, following some objective criteria [13]. The wrapper techniques are used to select the optimum subset of features based on the error reduction of classifier algorithms. Wrapper methods perform better than filter methods but are more expensive and time-consuming [14]. Features selected with embedded or hybrid methods are sensitive to the structure of the underlying classifiers. Thus, in most cases, the features selected by one embedded method might not be suitable for others [15].

Plasma protein binding plays a key role in drug therapy that affects the pharmacokinetics and pharmacodynamics of the drug, as they are often directly related to the free drug concentration in plasma [16, 17]. In recent years, several QSAR-artificial intelligence models have been developed to predict plasma protein binding, such as support vector machines and their derivatives [18-20], random forest [21], neural networks [22, 23], and gradient-boosting decision trees [24]. In 2017, Sun et al. constructed QSAR models by six machine-learning algorithms with 26 molecular descriptors [25]. Kumar et al. in 2018 presented a systematic approach using a support vector machine, artificial neural network, K-nearest neighbor, probabilistic neural network, partial least square, and linear discriminant analysis to a diverse dataset of 735 drugs [26]. Yuan et al. in 2020 published a global QSAR model for plasma protein binding and developed a novel strategy to construct a robust QSAR model for plasma protein binding prediction [27]. Recently, deep learning algorithms have attracted the attention of scientists and become an important option for pharmaceutical research. Ramsundar et al. presented deep learning techniques for healthcare that efficiently predict drug activity and structure [28]. Wallach et al. introduced AtomNet, known as the first structure-based deep convolutional neural network (CNN), to predict the bioactivity of small molecules for drug discovery applications [29].

In this work, a novel combined methodology based on QSAR, CNN, and a feed-forward neural network (FFNN) was used to predict plasma protein binding for 277 molecules. The CNN, known to be the most popular algorithm for deep learning, was used for feature selection as an alternative to the wrapper method. The FFNN was then used for the prediction of the plasma protein binding from the extracted features.

Material and Metods

In order to predict the plasma protein binding, a methodology based on five steps was used as summarized in Figure 1: (1) data set collection, (2) molecular descriptors generation, (3) selection of relevant descriptors by a wrapper method and a CNN method, (4) FFNN modeling, and (5) validation of models.

Data set collection

The experimental data of protein binding of the 277 drugs used in this study were selected from the pharmacological basis of the therapeutics handbook [30] and the handbook of clinical drug data [31]. Chemical names and experimental protein binding values are presented in Supplemental Material 1.

Molecular descriptors generation

The numerical representation of molecular structure was assessed in terms of molecular descriptors. The SMILES script (simplified molecular-input line-entry system) required to calculate descriptors was extracted from the open-access database PubChem [32]. SMILES is a specification in the form of a line notation describing the structure of chemical species [33]. The SMILES scripts for the 277 drugs were used to generate 1666 descriptors divided into 20 categories: *(i)* constitutional descriptors, *(ii)* topological descriptors, *(iii)* walk and path counts; *(iv)* connectivity indices; *(v)* index information; *(vi)* 2Dautocorrelations; *(vii)* edge adjacency indices; *(viii)* burden eigenvalue

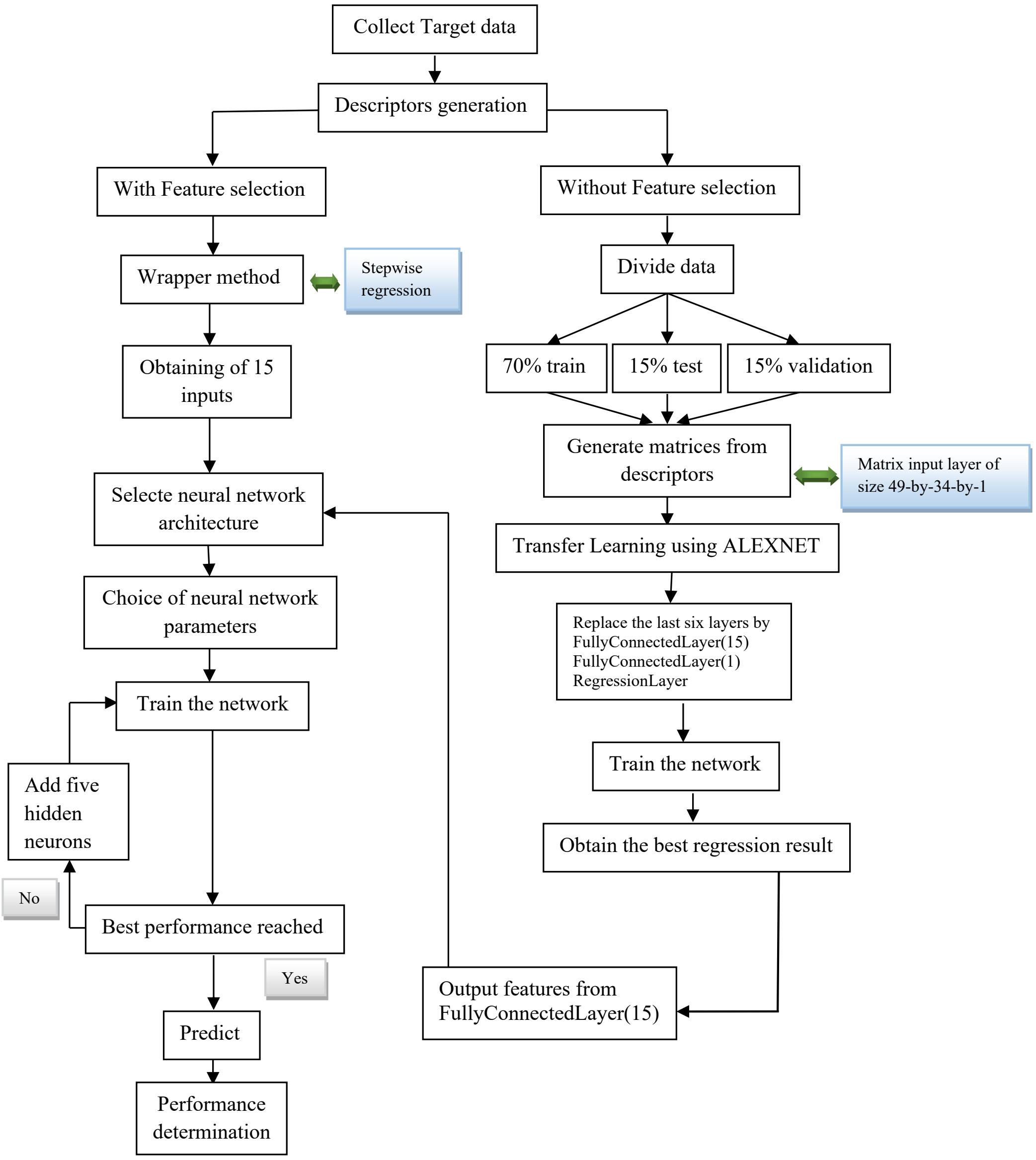

Collect Target data
Descriptors generation
With Feature selection
Without Feature selection
Wrapper method
Stepwise regression
Divide data
Obtaining of 15 inputs
70% train
15% test
15% validation
Generate matrices from descriptors
Matrix input layer of size 49-by-34-by-1
Selecte neural network architecture
Transfer Learning using ALEXNET
Choice of neural network parameters
Replace the last six layers by FullyConnectedLayer(15) FullyConnectedLayer(1) RegressionLayer
Train the network
Add five hidden neurons
Train the network
Obtain the best regression result
No
Best performance reached
Yes
Output features from FullyConnectedLayer(15)
Predict
Performance determination

Figure 1. Flow sheet of the procedure followed

descriptors; *(ix)* topological charge indices; *(x)* eigenvalue-based indices; *(xi)* Randic molecular profiles; *(xii)* geometrical descriptors; *(xiii)* RDF descriptors; *(xiv)* 3D-MoRSE descriptors; *(xv)* WHIM descriptors; *(xvi)* GETAWAY descriptors; *(xvii)* functional group counts; *(xviii)* atom-centered fragments; *(xix)* charge descriptors; and *(xx)* molecular properties.

All descriptors were obtained through the E-Dragon online programs [34], which is known as the electronic remote version of the well-known software DRAGON developed by the Milano Chemometrics and QSAR Research Group by Prof. R. Todeschini. The name and number of calculated descriptors are reported in Supplemental Material 2.

Selection of relevant descriptors

The selection of the most efficient descriptors, which is an important step in QSAR modeling, was made by two techniques in this work.

Wrapper method

The number of molecular descriptors was reduced by the following procedure [35]:

1) Descriptors having constant values (min=max) were eliminated.

2) Quasi-constant descriptors (1st quartile 25%=2nd quartile 75%) were removed.

3) Descriptors with standard relative deviation *RSD*<0.05 were deleted.

The three steps above were performed using STATISTICA software [36].

4) Matrices of the pairwise linear correlation between each pair of columns in the input matrices were calculated via MATLAB [37]. In addition, all variables with a correlation coefficient R>0.75 were removed. For more robustness of the model, the variance inflation factor *VIF* was calculated (Equation 1):

$$1.\ VIF_{\iota}=1/(1-R_{\iota}^{2})$$

, Where is the squared correlation coefficient of the ith descriptor versus the remaining ones. All the descriptors with VIF>5 were eliminated from the model [38].

5) The remaining descriptors were analyzed by stepwise regression in STATISTICA software [36]. Only 15 descriptors were selected and used as input variables in the FFNN models.

CNN method

CNNs are one of the most popular algorithms for deep learning, and they are mostly used for image classification. Computers do not see an image as an image, but as an array of pixels and the whole process in CNNs depends on it. Therefore, we provided our CNN network with previously digitalized data as input. Because for deep learning, in image classification, we do not have to understand which features are the best representations of the object, they are taught to us.

In our approach, the 1666 descriptors calculated with E-DRAGON were transformed for each drug into matrices of size 49-by-34-by-1 dimension in MATLAB because CNN requires matrices as inputs [39]. The latest features learned from the CNN layers were flattened out and passed through an FFNN to make predictions.

FFNN modeling

QSAR-wrapper-FFNN

The selected descriptors by the wrapper technique were set as input in an FFNN. So far, there is no clear answer to specify the number of hidden neurons required for a modeling task and it remains an open question. Different approaches are discovering this number, explained in detail in reviews, including methods of selecting the number of hidden nodes in the artificial neural networks review [40]. To decide the number of neurons in the hidden layer, the procedure summarized below was used [41]:

1) In the beginning, only a few hidden neurons (five) were used.

2) The network was trained until the mean square error no longer seemed to improve.

3) At this moment, a few (five) neurons were added to the hidden layer, each with randomly initialized weights, and resumed training.

4) The previous two steps were repeated until a termination criterion has been satisfied.

The mathematical equation of the model, for the prediction of protein binding, is shown below (Equation 2):

$$2.\ fb=\sum_{j=1}^{k} w2j\left(\frac{(\exp(\sum_{i=1}^{P} xi+wij+bj)-\exp(-\sum_{i=1} xi+wij+bj)}{(\exp(\sum_{i=1}^{P} xi+wij+bj)+\exp(-\sum_{i=1} xi+wij+bj)}\right)+b$$

x i (i=1…p) is the input that corresponds to the number of data included in the training of the ANN, I from 1 to 15, wij (i=1…p, j=1…k) weight from input to hidden layer, b j (j=1…k) are biases of the neurons in the hidden layer, k=45 for Wrapper method, w2j (j=1…k) are weights from the hidden to the output layer, b is the bias of the output neuron, and fb is the output.

QSAR-CNN-FFNN

CNN is a powerful machine learning technique. Training a CNN requires a large collection of diverse data, which is not an easy task [42]. However, there is an easy way, we can use a pre-trained CNN, which saves a huge amount of time and effort because fine-tuning a network with transfer learning is usually much faster and easier than training a network from scratch [43]. In this work, transfer learning using the ALEXNET CNN was applied to perform feature extraction and then feature reduction with adjustment in the last six layers by new ones. The CNN model was used to extract representative feature vectors from the penultimate fully connected layer, and then these ”deep features” were used for training an FFNN. The choice of dimension 15 for the feature reduction is to perform a comparison with the features selected by wrapper methods. The details of the 22 layers are reported in Supplemental Material 3. To train the network, holdout cross-validation was used to divide the data randomly 70% for training, 15% for validation, and 15% for testing.

Validation of the models

To determine both the generalizability and the actual predictive capacity of the QSAR models, internal and external validation criteria were used for the validation of the models. The statistical parameters used in our study to check the performance of the models were the coefficient of determination (R^2), the correlation coefficient (R), the predictive squared correlation coefficient (Q^2), and the mean squared error values (MSE) (Equations 3-5):

3. R^2=1-RSS/SS

4. $MSE=\sum_{i}^{n}=1 \frac{(y_{i}^{pred}-y_i)^2}{2}$

5. Q^2=1-PRESS/SS

The residual sum of squares (RSS) is the difference between the fitted values and the observed values. It refers to the difference between the observation and their mean. The predictive RSS is the difference between the predictions and the observations.

Results

QSAR-Wrapper-FFNN method

The results obtained from the selection of the most important descriptors by the wrapper method, using the correlation coefficient R and the variance inflation factor VIF, showed that 15 descriptors seemed to be the most appropriate. The calculated *VIF*s among the values of the selected descriptors were less than five, indicating that multicollinearity between the selected descriptors is acceptable. Supplemental Material 4 shows the *VIF* values for the selected descriptors and their meanings. To specify the number of hidden neurons required, the procedure detailed above was followed. R (all), MSE (validation), and (validation) criteria were employed for the evaluation of the accuracy of the best model. The best model was chosen according to the minimum MSE (validation) and the maximum R (all), and (validation) [35, 44].

Table 1 shows ten network models developed after the wrapper approach. The results obtained showed that network nine with 45 neurons was the best model with R (all)=0.935, R^2_{train}=0.875, Q^2=0.871, and MSE (validation)=0.015. The best performance of the model had a topology of 15-[45]-1:15 input nodes, one hidden layer with 45 nodes having the hyperbolic tangent as a transfer function, and one output layer with an identity transfer function. The neural networks were implemented using the neural network toolbox for MATLAB [37]. Figure 2 shows a comparison between the experiment and predicted plasma protein binding values for training, validation, and testing sets. The results showed a close correlation between predicted and observed plasma protein binding. The FFNN models were trained with the Levenberg-Marquardt backpropagation training function and gradient descent with momentum weight and bias learning function, and the data were partitioned

Table 1. Selected criteria of the different feed-forward neural networks (FFNNs) obtained by the quantitative structure-activity relationship (QSAR)-Wrapper-FFNN methodology

Number of Hidden Neurons	R (all)	R^2_{train}	Q^2	Mean Squared Error (Validation)
5	0.754	0.539	0.632	0.048
10	0.830	0.679	0.728	0.033
15	0.838	0.694	0.756	0.025
20	0.840	0.724	0.651	0.038
25	0.849	0.712	0.712	0.035
30	0.894	0.810	0.780	0.025
35	0.855	0.707	0.764	0.029
40	0.907	0.814	0.871	0.015
45	**0.935**	**0.875**	**0.871**	**0.015**
50	0.903	0.806	0.821	0.023

Table 2. Selected criteria for the different feed-forward neural networks (FFNNs) obtained from the quantitative structure-activity relationship (QSAR)- convolutional neural network (CNN)-FFNN method

Number of Hidden Neurons	R (all)	R^2_{train}	Q^2	Mean squared error (Validation)
5	0.882	0.812	0.701	0.040
10	0.908	0.832	0.798	0.009
15	0.913	0.856	0.760	0.028
20	0.933	0.857	0.948	0.009
25	0.910	0.834	0.830	0.022
30	0.905	0.830	0.775	0.022
35	0.960	0.905	0.964	0.005
40	0.922	0.832	0.961	0.007
45	0.940	0.874	0.899	0.012
50	0.922	0.830	0.841	0.009

PBR

using holdout cross-validation. The difference between R^2_{train} and Q^2 was equal to 0.004. This difference did not exceed 0.3 indicating the robustness of the model [45].

QSAR-CNN-FFNN method

Regarding the QSAR-CNN-FFNN, the results obtained (Table 2) showed that network seven was the best model with R (all)=0.960, R^2_{train}=0.905, Q^2=0.964, and MSE (validation)=0.005. The best performance of the model had a topology of 15-[35]-1:15 input nodes, one hidden layer with 35 nodes having the hyperbolic tangent as a transfer function, and an identity transfer function in the output layer. Figure 3 shows a comparison between experimental and predicted plasma protein binding for training, validation, and testing sets. The difference between R^2_{train} and Q^2 is equal to 0.059. This difference is less than 0.3 indicating the robustness of the model [45]. For QSAR-CNN-FFNN, the correla-

Table 3. External and internal criteria of the two models

	Parameters	QSAR-Wrapper-FFNN Method	QSAR-CNN-FFNN Method
Internal validation	R (all)	0.935	0.960
	R^2_{train}	0.875	0.905
	Q^2	0.871	0.964
	MSE	0.015	0.005
	MAE	0.090	0.054
	RMSE	0.122	0.069
	$R^2_{adjusted}$	0.874	0.963
External validation	R	0.933	0.982
	Q^2	0.870	0.945
	MSE	0.016	0.007
	MAE	0.094	0.068
	RMSE	0.129	0.085

PBR

Table 4. Comparison with the literature

Method/Ref.	MAE	R^2	R	MSE
Suggested method (QSAR-Wrapper-FFNN)	Train 0.080 Validation 0.090 Test 0.094	Train 0.875 Validation 0.871 Test 0.870	Train 0.935 Validation 0.933 Test 0.933	Train 0.014 Validation 0.015 Test 0.017
Suggested method (QSAR-CNN-FFNN)	Train 0.066 Validation 0.054 Test 0.068	Train 0.905 Validation 0.964 Test 0.945	Train 0.951 Validation 0.972 Test 0.982	Train 0.011 Validation 0.005 Test 0.007
Yuan et al. [27]	Test 0.076			
Sun et al. [25]	Test 0.126			
Kumar et al. [26]				Train 0.869 Test 0.8881
Li et al. [49]		Train 0.86		
Ghafourian et al. [21]	Train 13.25 Validation 14.96	Train 0.717 Validation 0.646	Train 0.681 Validation 0.641	
Moda et al. [50]	-	Test 0.91		

tion among each pair of our variables was verified and the results are presented in Supplemental Materials 5. In order to obtain an overview of the correlation structure, a heatmap was used to highlight what is important (Figure 4). The question now is how much is too much correlation between our neural network inputs. It seems that the general rule of thumb is that if the simple correlation coefficient between two regressors is greater than 0.8 or 0.9, multicollinearity is a serious problem [46]. However, with all the rules of thumb, we get in the statistics, there is nothing black or white about this. In our case, a prediction with a correlation coefficient of 0.8 does not appear fatal for our regression model. We can say that the above results served our theory, where more descriptors are valuable and welcome because we accept that by removing some of the features we are also discarding some of the data we have about the problem.

Discussion

Our results described for the first time the use of CNNs as a feature extraction method in QSAR studies for a set of descriptor matrices instead of local pattern extraction from images. One of the huge favorable circumstances of this method is that it does not take much time (only a few minutes) compared to the wrapper method (which took us days and concentration). It also automatically provides us with features that are the best representation of the non-image input, because in this work, we simply constructed the CNN inputs by resizing the 1666 features to 49×34 matrices. We knew that CNNs are used on data that has spatial features, and the fact that it gave us good features to use is that the CNN used discovered that our data has some sort of spatial features. We wanted to know what this deep CNN actually saw, and how it understands the inputs, we feed them. By visualizing the digital values of our matrices, grayscale images were obtained, giving 255 possible different shades of grey

Table 5. Applicability domain of the new method

Approaches	Test Inside AD	Test Outside AD
Bounding box	40	2
Leverage	41	1
Euclidean distance (95 percentile)	41	1
Classical KNN (euclidean distance, k=5)	42	0
KNN (euclidean distance (k=25))	42	0

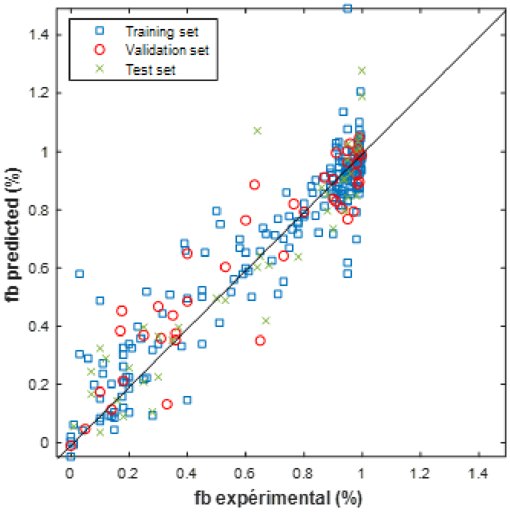


Figure 2. Comparison between experimental and calculated plasma protein binding predicted by the QSAR-Wrapper-FFNN technique.

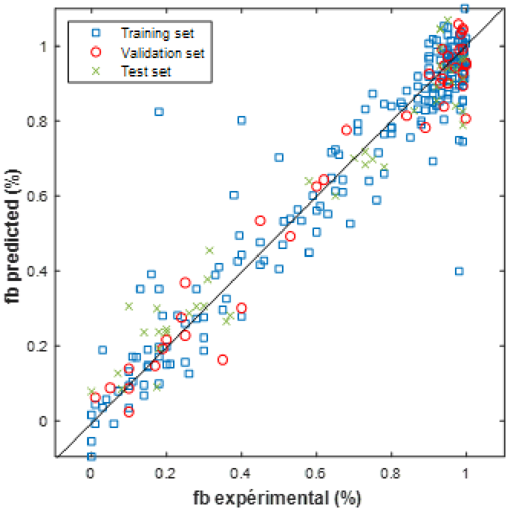


Figure 3. Comparison between experimental and calculated values for plasma protein binding predicted by the QSAR-CNN-FFNN technique.

from black to white. Figure 5 shows a sample image of 3-carboxy-4-methyl-5-propyl-2-furanpropionic acid obtained. Numeric values below zero go to zero and those over one go to 255. Values between zero and one are shades of grey, with values between zero and 255. By taking the values of our descriptors and mapping them to the corresponding color suddenly, we can see the structure within our data. We can say that we have made an interface between our data and our brain by visualizing descriptors, in which physical and chemical information about molecules becomes accessible in a new form of fingerprints. They turn into storytellers of molecules secret in just an image.

Comparison of the two models

A statistical evaluation presented in Table 3 was conducted to compare the performance and quality of the predictions by the FFNN of the two models developed for this work. The statistical coefficients of the internal validation for the two models are all acceptable and satisfactory (high R^2_{train}, Q^2 (validation), $R^2_{adjusted}$ and lowest MSE, RMSE, and MAE) and consequently, these models are robust. The quality of the models was also evaluated in terms of external validation criteria. For the wrapper feature extraction method, a value of $Q^2>0.5$ is considered satisfactory and for the QSAR-CNN feature extraction method, the value of $Q^2>0.9$ is considered excellent [45]. We can say that the two models are distinguished by excellent predictive power.

Comparison between models from the literature

A comparison was made between the few models reported in the literature with those developed in this work for the prediction of the binding of drugs to plasma proteins (Table 4). Yawen Yuan et al. constructed a QSAR model for predicting plasma protein binding based on a large training set comprising more than 5000 compounds and proposed a new strategy for constructing models for different binding levels. Lixia Sun et al. de-

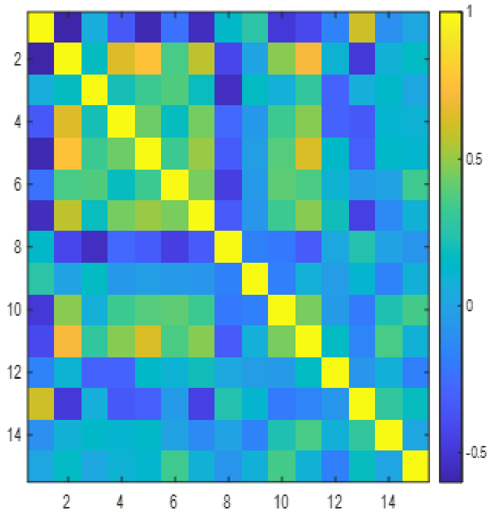
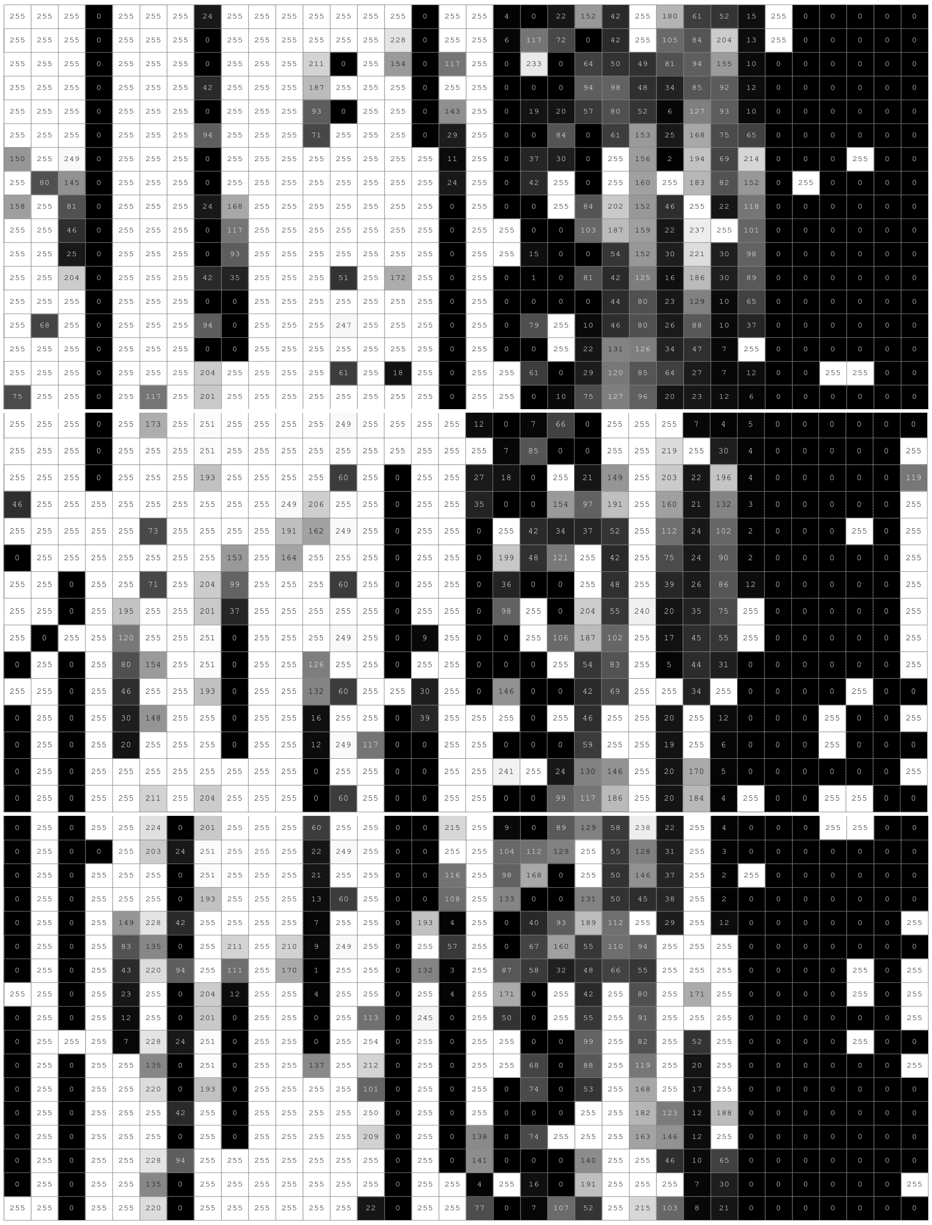


Figure 4. Heatmap of the correlation matrix for the QSAR-CNN-FFNN method.



255	255	255	0	255	255	255	24	255	255	255	255	255	255	255	0	255	255	4	0	22	152	42	255	180	61	52	15	255	0	0	0	0	0
255	255	255	0	255	255	255	0	255	255	255	255	255	255	228	0	255	255	6	117	72	0	42	255	105	84	204	13	255	0	0	0	0	0
255	255	255	0	255	255	255	0	255	255	255	211	0	255	154	0	117	255	0	233	0	64	50	49	81	94	155	10	0	0	0	0	0	0
255	255	255	0	255	255	255	42	255	255	255	187	255	255	255	0	255	255	0	0	0	94	98	48	34	85	92	12	0	0	0	0	0	0
255	255	255	0	255	255	255	0	255	255	255	93	0	255	255	0	143	255	0	19	20	57	80	52	6	127	93	10	0	0	0	0	0	0
255	255	255	0	255	255	255	94	255	255	255	71	255	255	255	0	29	255	0	0	84	0	61	153	25	168	75	65	0	0	0	0	0	0
150	255	249	0	255	255	255	0	255	255	255	255	255	255	255	255	11	255	0	37	30	0	255	156	2	194	69	214	0	0	0	255	0	0
255	80	145	0	255	255	255	0	255	255	255	255	255	255	255	255	24	255	0	42	255	0	255	160	255	183	82	152	0	255	0	0	0	0
158	255	81	0	255	255	255	24	168	255	255	255	255	255	255	255	0	255	0	0	255	84	202	152	46	255	22	118	0	0	0	0	0	0
255	255	46	0	255	255	255	0	117	255	255	255	255	255	255	255	0	255	255	0	0	103	187	159	22	237	255	101	0	0	0	0	0	0
255	255	25	0	255	255	255	0	93	255	255	255	255	255	255	255	0	255	255	15	0	0	54	152	30	221	30	98	0	0	0	0	0	0
255	255	204	0	255	255	255	42	35	255	255	255	51	255	172	255	0	255	0	1	0	81	42	125	16	186	30	89	0	0	0	0	0	0
255	255	255	0	255	255	255	0	0	255	255	255	255	255	255	255	0	255	0	0	0	0	44	80	23	129	10	65	0	0	0	0	0	0
255	68	255	0	255	255	255	94	0	255	255	255	247	255	255	255	0	255	0	79	255	10	46	80	26	88	10	37	0	0	0	0	0	0
255	255	255	0	255	255	255	0	0	255	255	255	255	255	255	255	0	255	0	0	255	22	131	126	34	47	7	255	0	0	0	0	0	0
255	255	255	0	255	255	255	204	255	255	255	255	61	255	18	255	0	255	255	61	0	29	120	85	64	27	7	12	0	0	255	255	0	0
75	255	255	0	255	117	255	201	255	255	255	255	255	255	255	255	0	255	255	0	10	75	127	96	20	23	12	6	0	0	0	0	0	0

255	255	255	0	255	173	255	251	255	255	255	255	249	255	255	255	255	12	0	7	66	0	255	255	255	7	4	5	0	0	0	0	0	0
255	255	255	0	255	255	255	251	255	255	255	255	255	255	255	255	255	255	7	85	0	0	255	255	219	255	30	4	0	0	0	0	0	255
255	255	255	0	255	255	255	193	255	255	255	255	60	255	0	255	255	27	18	0	255	21	149	255	203	22	196	4	0	0	0	0	0	119
46	255	255	255	255	255	255	255	255	255	249	206	255	255	0	255	255	35	0	0	154	97	191	255	160	21	132	3	0	0	0	0	0	255
255	255	255	255	255	73	255	255	255	255	191	162	249	255	0	255	255	0	255	42	34	37	52	255	112	24	102	2	0	0	0	255	0	255
0	255	255	255	255	255	255	255	153	255	164	255	255	255	0	255	255	0	199	48	121	255	42	255	75	24	90	2	0	0	0	0	0	255
255	255	0	255	255	71	255	204	99	255	255	255	60	255	0	255	255	0	36	0	0	255	48	255	39	26	86	12	0	0	0	0	0	255
255	255	0	255	195	255	255	201	37	255	255	255	255	255	0	255	255	0	98	255	0	204	55	240	20	35	75	255	0	0	0	0	0	255
255	0	255	255	120	255	255	251	0	255	255	255	249	255	0	9	255	0	0	255	106	187	102	255	17	45	55	255	0	0	0	0	0	255
0	255	0	255	80	154	255	251	0	255	255	126	255	255	0	255	255	0	0	0	255	54	83	255	5	44	31	0	0	0	0	0	0	255
255	255	0	255	46	255	255	193	0	255	255	132	60	255	255	30	255	0	146	0	0	42	69	255	255	34	255	0	0	0	0	255	0	0
0	255	0	255	30	148	255	255	0	255	255	16	255	255	0	39	255	255	255	0	255	46	255	255	20	255	12	0	0	0	255	0	0	255
0	255	0	255	20	255	255	255	0	255	255	12	249	117	0	0	255	255	0	0	0	59	255	255	19	255	6	0	0	0	255	0	0	0
0	255	0	255	255	255	255	255	255	255	255	0	255	255	0	0	255	255	241	255	24	130	146	255	20	170	5	0	0	0	0	0	0	255
0	255	0	255	255	211	255	204	255	255	255	0	60	255	0	0	255	255	0	0	99	117	186	255	20	184	4	255	0	0	255	255	0	0

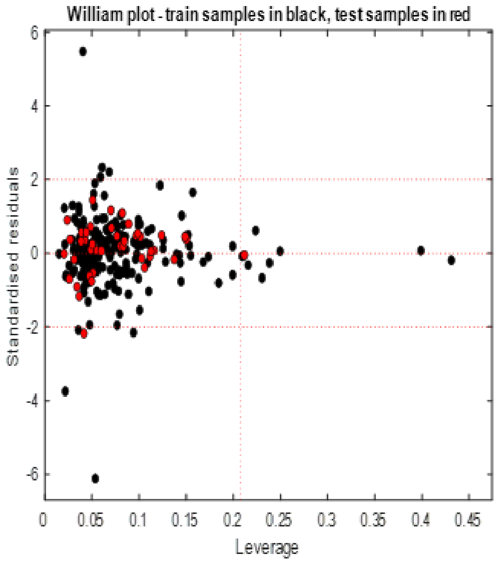
0	255	0	255	255	224	0	201	255	255	255	60	255	255	0	0	215	255	9	0	89	129	58	238	22	255	4	0	0	0	255	255	0	0
0	255	0	0	255	203	24	251	255	255	255	22	249	255	0	0	255	255	104	112	129	255	55	128	31	255	3	0	0	0	0	0	0	0
0	255	0	255	255	255	0	251	255	255	255	21	255	255	0	0	116	255	98	168	0	255	50	146	37	255	2	255	0	0	0	0	0	0
0	255	0	255	255	255	0	193	255	255	255	13	60	255	0	0	108	255	133	0	0	131	50	45	38	255	2	0	0	0	0	0	0	0
0	255	0	255	149	228	42	255	255	255	255	7	255	255	0	193	4	255	0	40	93	189	112	255	29	255	12	0	0	0	0	0	0	255
0	255	0	255	83	135	0	255	211	255	210	9	249	255	0	255	57	255	0	67	160	55	110	94	255	255	255	0	0	0	0	0	0	0
0	255	0	255	43	220	94	255	111	255	170	1	255	255	0	132	3	255	87	58	32	48	66	55	255	255	255	0	0	0	0	255	0	255
255	255	0	255	23	255	0	204	12	255	255	4	255	255	0	255	4	255	171	0	255	42	255	80	255	171	255	0	0	0	0	255	0	255
0	255	0	255	12	255	0	201	0	255	255	0	255	113	0	245	0	255	50	0	255	55	255	91	255	255	255	0	0	0	0	0	0	255
0	255	255	255	7	228	24	251	0	255	255	0	255	254	0	255	0	255	255	0	0	99	255	82	255	52	255	0	0	0	0	255	0	0
0	255	0	255	255	135	0	251	0	255	255	137	255	212	0	255	0	255	255	68	0	88	255	119	255	20	255	0	0	0	0	0	0	255
0	255	0	255	255	220	0	193	0	255	255	255	255	101	0	255	0	255	0	74	0	53	255	168	255	17	255	0	0	0	0	0	0	0
0	255	0	255	255	255	42	255	0	255	255	255	255	250	0	255	0	255	0	0	0	255	255	182	123	12	188	0	0	0	0	0	0	0
0	255	0	255	255	255	0	255	0	255	255	255	255	209	0	255	0	138	0	74	255	255	255	163	146	12	255	0	0	0	0	0	0	0
0	255	0	255	255	228	94	255	255	255	255	255	255	255	0	255	0	141	0	0	0	140	255	255	46	10	65	0	0	0	0	0	0	0
0	255	0	255	255	135	0	255	255	255	255	255	255	255	0	255	255	4	255	16	0	191	255	255	255	7	30	0	0	0	0	0	0	255
255	255	0	255	255	220	0	255	255	255	255	255	255	22	0	255	255	77	0	7	107	52	255	215	103	8	21	0	0	0	0	0	0	0

PBR

Figure 5. Image of the input matrices of 3-Carboxy-4-methyl-5-propyl-2-furanpropionic acid with the corresponding numerical values of shade of gray.

veloped QSAR models by six machine learning algorithms with 26 molecular descriptors for predicting the protein-bound fractions of 967 pharmaceuticals. The neural network model yielded an MAE error of 0.126. Rajnish Kumar et al. presented six machine learning algorithms for QSAR models using a database containing 735 drugs. The model with the artificial neural network gave an MSE train=0.869 and MSE val=0.881. Haiyan Li et al. developed a new method called QSAR Plasma Protein Interaction QSAR Analysis (PPI-QSAR) with a database of 65 antibiotics, providing R^2_{train}=0.86 and Q^2=0.72. Taravat Ghafourian et al. collected a database of 794 drugs and used four data mining tools; the best model was boosted trees providing an error MAE train=13.25 and an MAE=14.96. Moda Tiago et al. developed a hologram quantitative structure-activity relationship (HQSAR) on a series of 312 drugs, and they obtained a Q^2=0.72 and R^2=0.91.

We were not interested in evaluating the advantages and disadvantages of these methods because it is quite difficult (each study used different data sets and different modeling approaches) and because the objective of our work was to prove that our new feature selection method



PBR

Figure 6. Plot of the residuals for calculated values of plasma protein binding from the new approach versus their experimental values for training and test sets.

is as effective as these old methods. According to these results, the models developed in this work performed as well as these methods.

Applicability domain

The third principle of the OECD guidelines [47] recommends a defined applicability domain. In this work, the domain of applicability was analyzed with different approaches reported in Table 5. The algorithm and methodology of the proposed approaches are available in the literature [48]. The number of samples inside the applicability domain varied depending on the method used: the bounding box considered two test samples out of the applicability domain, while leverage and Euclidean distance (95 percentile) identified one test sample out of the domain of applicability as shown in Williams plot (Figure 6). Despite its distance from the other observations, our point is close to the regression fitted line since it has a small residual. Classical KNN (Euclidean distance, k=5) and KNN (Euclidean distance (k=25)) showed none of the test samples out of the applicability domain. These results show that we can use the model to predict plasma protein binding for new compounds that have not been tested.

Conclusion

Two different approaches to feature selection in machine learning were tested. These feature selection strategies were followed to produce inputs, which were used in an FFNN to make predictions. Examination of the estimates of external and internal criteria revealed that the QSAR model developed by the new feature selection method is robust, externally predictive, and distinguished by a good applicability domain. The external accuracy of the validation set was calculated by the Q^2 and RMSE, which are equal to 0.945 and 0.085, respectively indicating that the accuracy of NN trained with extracted features from CNNs is slightly better than the accuracy of NN trained with features obtained by the current predictors. This investigation showed that extracting features using CNNs takes less time (only a few minutes) and summarizes much of the information contained in the original features. Contrary to feature selection techniques, we do not have to figure out which features are the best representations of the input, they are learned for us. In the end, according to the OECD principle, we can say that we can use this QSAR-CNN-FFNN model to predict the fraction of plasma protein binding to human plasma for drugs that have not been tested to avoid chemical synthesis and reduce expansive laboratory tests.

Ethical Considerations

Compliance with ethical guidelines

There were no ethical considerations to be considered in this research.

Funding

The authors declare no conflict of interest.

Authors' contributions

Study design: Salah Hanini; Investigation, data collection, and data analysis: Khaouane Affaf; Supervision and Writing–original draft: Samira Ferhat; Writing–review & editing: all authors.

Conflict of interest

The authors declared no conflict of interest.

Acknowledgments

The authors thank the Laboratory of Biomaterial and Transport Phenomena (LBMPT), University of Médéa, for providing the facilities to conduct this study.

References

[1] Sliwoski G, Kothiwale S, Meiler J, Lowe EW Jr. Computational methods in drug discovery. Pharmacol Rev. 2013; 66(1):334-95. [DOI:10.1124/pr.112.007336] [PMID] [PMCID]

[2] Lin X, Li X, Lin X. A review on applications of computational methods in drug screening and design. Molecules. 2020; 25(6):1375. [DOI:10.3390/molecules25061375] [PMID] [PMCID]

[3] Kubinyi H. QSAR and 3D QSAR in drug design Part 1: methodology. Drug Discovery Today. 1997; 2(11):457-67. [DOI:10.1016/S1359-6446(97)01084-2]

[4] Khan MT. Predictions of the ADMET properties of candidate drug molecules utilizing different QSAR/QSPR modelling approaches. Curr Drug Metab. 2010; 11:285-95. [DOI:10.2174/138920010791514306] [PMID]

[5] Zhang L, Tan J, Han D, Zhu H. From machine learning to deep learning: Progress in machine intelligence for rational drug discovery. Drug Discov Today. 2017; 22(11):1680-5. [DOI:10.1016/j.drudis.2017.08.010] [PMID]

[6] Rasulev BF, Abdullaev ND, Syrov VN, Leszczynski J. A quantitative structure-activity relationship (QSAR) study of the antioxidant activity of flavonoids. QSAR Comb Sci. 2005; 24:1056-65. [DOI:10.1002/qsar.200430013]

[7] Nirmalakhandan N, Speece RE. ES & T critical review: Structure-activity relationships. Quantitative techniques for predicting the behavior of chemicals in the ecosystem. Environ Sci Technol. 1988; 22:606-15. [DOI:10.1021/es00171a002]

[8] Grover M, Singh B, Bakshi M, Singh S. Quantitative structure-property relationships in pharmaceutical research-Part 1. Pharm Sci Technol. 2000; 3(2):50-57. [DOI:10.1016/S1461-5347(99)00214-X]

[9] Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinformatics. 2015; 2015:198363. [DOI:10.1155/2015/198363] [PMID] [PMCID]

[10] Langley P. Selection of relevant features in machine learning. Palo Alto: Institute for the Study of Learning and Expertise; 1994. [DOI:10.21236/ADA292575]

[11] Dash M, Liu H. Feature selection for classification. Intell Data Anal. 1997; 1(1-4):131-56. [DOI:10.1016/S1088-467X(97)00008-5]

[12] Tetko IV, Livingstone DJ, Luik AI. Neural network studies. 1. Comparison of overfitting and overtraining. J Chem Inf Comput Sci. 1995:35(5):826–33. [DOI:10.1021/ci00027a006]

[13] Dudek AZ, Arodz T, Gálvez J. Computational methods in developing quantitative structure-activity relationships (QSAR): A review. Comb Chem High Throughput Screen. 2006; 9(3):213-28. [DOI:10.2174/138620706776055539] [PMID]

[14] Ghasemi F, Mehridehnavi A, Pérez-Garrido A, Pérez-Sánchez H. Neural network and deep-learning algorithms used in QSAR studies: Merits and drawbacks. Drug Discov Today. 2018; 23(10):1784-90. [DOI:10.1016/j.drudis.2018.06.016] [PMID]

[15] Danishuddin, Khan AU. Descriptors and their selection methods in QSAR analysis: Paradigm for drug design. Drug Discov Today. 2016; 21(8):1291-302. [DOI:10.1016/j.drudis.2016.06.013] [PMID]

[16] Bohnert T, Gan LS. Plasma protein binding: From discovery to development. J Pharm Sci. 2013; 102(9):2953-94. [DOI:10.1002/jps.23614] [PMID]

[17] Trainor GL. The importance of plasma protein binding in drug discovery. Expert Opin Drug Discov. 2007; 2(1):51-64. [DOI:10.1517/17460441.2.1.51] [PMID]

[18] Zsila F, Bikadi Z, Malik D, Hari P, Pechan I, Berces A, et al. Evaluation of drug-human serum albumin binding interactions with support vector machine aided online automated docking. Bioinformatics. 2011; 27(13):1806-13. [DOI:10.1093/bioinformatics/btr284] [PMID]

[19] Guo Y, Yu L, Wen Z, Li M. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res. 2008; 36(9):3025-30. [DOI:10.1093/nar/gkn159] [PMID] [PMCID]

[20] Zhao C, Zhang H, Zhang X, Zhang R, Luan F, Liu M, et al. Prediction of milk/plasma drug concentration (M/P) ratio using support vector machine (SVM) method. Pharmaceutical Res. 2006; 23(1):41-8. [DOI:10.1007/s11095-005-8716-4] [PMID]

[21] Ghafourian T, Amin Z. QSAR models for the prediction of plasma protein binding. Bioimpacts. 2013; 3(1):21-7. [DOI:10.5681/bi.2013.011] [PMID] [PMCID]

[22] Fu X, Wang G, Gao J, Zhan S, Liang W. Prediction of plasma protein binding of cephalosporins using an artificial neural network. Pharmazie. 2007; 62(2):157-8. [PMID]

[23] Turner JV, Maddalena DJ, Cutler DJ. Pharmacokinetic parameter prediction from drug structure using artificial neural networks. Int J Pharm. 2004; 270(1-2):209-19. [DOI:10.1016/j.ijpharm.2003.10.011] [PMID]

[24] Deng L, Sui Y, Zhang J. XGBPRH: Prediction of binding hot spots at protein-RNA interfaces utilizing extreme gradient boosting. Genes. 2019; 10(3):242. [DOI:10.3390/genes10030242] [PMID] [PMCID]

[25] Sun L, Yang H, Li J, Wang T, Li W, Liu G, et al. In silico prediction of compounds binding to human plasma proteins by QSAR models. Chem Med Chem. 2018; 13(6):572-81. [DOI:10.1002/cmdc.201700582] [PMID]

[26] Kumar R, Sharma A, Siddiqui MH, Tiwari RK. Prediction of drug-plasma protein binding using artificial intelligence based algorithms. Comb Chem High Throughput Screen. 2018; 21(1):57-64. [DOI:10.2174/1386207321666171218121557] [PMID]

[27] Yuan Y, Chang S, Zhang Z, Li Z, Li S, Xie P, et al. A novel strategy for prediction of human plasma protein binding using machine learning techniques. Chemometr Intell Lab Syst. 2020; 199:103962. [DOI:10.1016/j.chemolab.2020.103962]

[28] Altae-Tran H, Ramsundar B, Pappu AS, Pande V. Low data drug discovery with one-shot learning. ACS Cent Sci. 2017; 3(4):283-93. [DOI:10.1021/acscentsci.6b00367] [PMID] [PMCID]

[29] Wallach I, Dzamba M, Heifets A. AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXi v:1510.02855v1. 2015; 1-11. [DOI:10.48550/arXiv.1510.02855]

[30] Brunton L, Lazo J, Parker K. Goodman & Gilman's the pharmacological basis of therapeutics. New York: McGraw Hill Professional; 2005. [Link]

[31] Troutman WG. Handbook of clinical drug data. New York: McGraw-Hill; 2002. [Link]

[32] PubChem. National Institutes of Health (NIH). https://pubchemdocs.ncbi.nlm.nih.gov/

[33] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988; 28(1):31–6. [DOI:10.1021/ci00057a005]

[34] Virtual Computational Chemistry Laboratory (VCCLAB). http://www.vcclab.org/

[35] Hamadache M, Benkortbi O, Hanini S, Amrane A, Khaouane L, Moussa CS. A quantitative structure activity relationship for acute oral toxicity of pesticides on rats: Validation, domain of application and prediction. J Hazard Mater. 2016; 303:28-40. [DOI:10.1016/j.jhazmat.2015.09.021] [PMID]

[36] Stat Soft Inc. STATISTICA, Version 8 [Internet]. 2007 [Updated November 2022]. Available from: **[Link]**

[37] Math Works MATLAB R2019b. MathWorks, Inc. [Internet]. 2022 [Updated November 2022]. Available from: [Link]

[38] Akinwande MO, Dikko HG, Samson A. Variance inflation factor: As a condition for the inclusion of suppressor variable (s) in regression analysis. "Open J Stat. 2015; 5:754-67. [DOI:10.4236/ojs.2015.57075]

[39] Bazgir O, Zhang R, Dhruba SR, Rahman R, Ghosh S, Pal R. Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. Nat Commun. 2020; 11:4391. [DOI:10.1038/s41467-020-18197-y] [PMID] [PMCID]

[40] Panchal FS, Panchal M. Review on methods of selecting number of hidden nodes in artificial neural network. Int J Comput Sci Mob Computing. 2014; 3(11):455-64. [Link]

[41] Kubat M. An introduction to machine learning. Cham: Springer; 2017. [DOI:10.1007/978-3-319-63913-0]

[42] Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans Med Imaging. 2016; 35(5):1299-312 [DOI:10.1109/TMI.2016.2535302] [PMID]

[43] Schwarz M, Schulz H, Behnke S. RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features. Paper presented at: 2015 IEEE International Conference on Robotics and Automation (ICRA). 26-30 May 2015; Seattle, USA. [DOI:10.1109/ICRA.2015.7139363]

[44] Bitam S, Hamadache M, Hanini S. QSAR model for prediction of the therapeutic potency of N-benzylpiperidine derivatives as AChE inhibitors. SAR QSAR Environ Res. 2017; 28(6):471-49. [DOI:10.1080/1062936X.2017.1331467] [PMID]

[45] Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect. 2003; 111(10):1361-75. [DOI:10.1289/ehp.5758] [PMID] [PMCID]

[46] Midi H, Sarkar SK, Rana S. Collinearity diagnostics of binary logistic regression model. J Interdiscip Math. 2010; 13:253-67. [DOI:10.1080/09720502.2010.10700699]

[47] Organisation for Economic Co-operation and Development (OECD). Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models., Paris: OECD Publishing; 2014. [Link]

[48] Roy K, Kar S, Ambure P. On a simple approach for determining applicability domain of QSAR models. Chemom Intell Lab Syst. 2015; 145:22-9. [DOI:10.1016/j.chemolab.2015.04.013]

[49] Li H, Chen Z, Xu X, Sui X, Guo T, Liu W, et al. Predicting human plasma protein binding of drugs using plasma protein interaction QSAR analysis (PPI-QSAR). Biopharm Drug Dispos. 2011; 32(6):333-42. [DOI:10.1002/bdd.762] [PMID]

[50] Moda TL, Montanari CA, Andricopulo AD. *In silico* prediction of human plasma protein binding using hologram QSAR. Lett Drug Des Discov. 2007; 4(7):502-9. [DOI:10.2174/157018007781788480]