Advanced Pharmaceutical Bulletin

Adv Pharm Bull, 2023, 13(4), 784-791
doi: 10.34172/apb.2023.078
https://apb.tbzmed.ac.ir

Original Article

A Quantitative Structure-Activity Relationship for Human Plasma Protein Binding: Prediction, Validation and Applicability Domain

Affaf Khaouane*, **Samira Ferhat**, **Salah Hanini**

Laboratory of Biomaterial and transport Phenomena (LBMPT), University of Médéa, pole urbain, 26000, Médéa, Algeria.

Article info

Article History:
Received: May 24, 2022
Revised: January 23, 2023
Accepted: April 24, 2023
epublished: April 29, 2023

Keywords:
Quantitative structure-activity relationship, Artificial neural network, Prediction, Protein-binding

Abstract

Purpose: The purpose of this study was to develop a robust and externally predictive in silico QSAR-neural network model for predicting plasma protein binding of drugs. This model aims to enhance drug discovery processes by reducing the need for chemical synthesis and extensive laboratory testing.

Methods: A dataset of 277 drugs was used to develop the QSAR-neural network model. The model was constructed using a Filter method to select 55 molecular descriptors. The validation set's external accuracy was assessed through the predictive squared correlation coefficient Q2 and the root mean squared error (RMSE).

Results: The developed QSAR-neural network model demonstrated robustness and good applicability domain. The external accuracy of the validation set was high, with a predictive squared correlation coefficient Q2 of 0.966 and a root mean squared error (RMSE) of 0.063. Comparatively, this model outperformed previously published models in the literature.

Conclusion: The study successfully developed an advanced QSAR-neural network model capable of predicting plasma protein binding in human plasma for a diverse set of 277 drugs. This model's accuracy and robustness make it a valuable tool in drug discovery, potentially reducing the need for resource-intensive chemical synthesis and laboratory testing.

Introduction

Many drugs interact with plasma or other molecules, such as DNA, to form a drug-molecule complex. The process is called protein binding, more specifically the binding of drugs to proteins. The bond drug remains in the bloodstream while the unbound component can be metabolized or excreted to become the active component.[1] In short, protein-binding process is defined as the formation of complexes: hydrogen bonding, hydrophilic bonding, ionic bonding, Vander Walls bonding, and covalent bonding.

The binding of drugs to proteins can be reversible or irreversible.[2,3] Irreversible drug-protein binding is the result of chemical activation of a drug tightly binding to a protein or macromolecule through a covalent chemical bond. Irreversible drug binding is responsible for some types of drug toxicity that can occur over a long period of time.[4] Reversible drug- protein binding means that the drug binds to weaker chemical bound, such as hydrogen bonds or Vander Waals forces. At low drug concentrations, most of the drug is bound to the protein, while at high drug concentrations, the protein is bound to the sites to saturate, leading to a rapid increase in the free drug concentration. Therefore plasma protein binding plays a key role in drug therapy as it affects the pharmacokinetics and pharmacodynamics of the drug as it is often directly related to the concentration of free drug in plasma.[5,6]

The construction of in silico models that establish a mathematical relationship between the molecular structure and the properties of interest is an important step in drug discovery as it avoids chemical synthesis and expansive and lengthy ones laboratory tests reduced.[7,8]

In recent years, several QSAR models have been developed to predict plasma protein binding and powerful plasma protein binding prediction algorithms are used, such as support vector machines and their derivatives,[9-11] the random forest,[12] neural networks,[13,14] and gradient boosting decision trees.[15] In 2017, Sun et al constructed QSAR models using six machine-learning algorithms with 26 molecular descriptors.[16] Kumar et al presented in 2018 a systematic approach using support vector machine, artificial neural network, K-nearest neighbor, probabilistic neural network, partial least square, and linear discriminant analysis for a diverse dataset of 735 remdies.[17] Yuan et al. published a global quantitative structure-activity relationships (QSAR) model for plasma

***Corresponding Author:** Affaf Khaouane, Email: affoufa80@gmail.com

protein-binding in 2020, and developed a novel strategy to construct a robust QSAR model for predicting plasma protein-binding.[18] Altae-Tran et al introduced deep–learning healthcare techniques successfully predicting drug activity and structure.[19] Wallach and his co-authors introduced AtomNet, known as the first structure-based deep convolutional neural network, to predict small molecule bioactivity for drug discovery applications.[20]

This work uses a systematic methodology based on QSAR, Filter method, and feed-forward neural network (FFNN) to predict plasma protein binding for 277 molecules. Filter method, known as the most popular feature selection technique, was used to reduce the descriptors. A feed forward neural network was then used to predict plasma protein-binding from the extracted descriptors.

Materials and Methods

A five-step process was employed to predict the plasma protein-binding, as shown in Figure 1: (1) data set collection, (2) molecular descriptors generation, (3) selection of relevant descriptors by a filter method, (4) FFNN modeling, (5) validation of models.

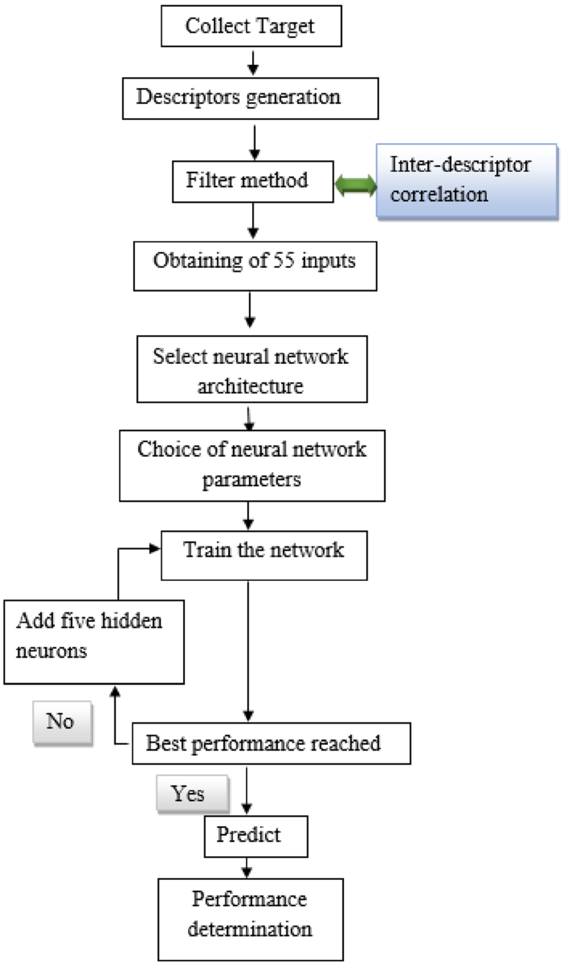

Figure 1. Flow sheet of the procedure followed

Data set collection

The experimental data values of protein-binding of the 277 drugs used in this study were selected from the pharmacological basis of the therapeutics handbook[21] and the handbook of clinical drug data.[22] Chemical names and experimental protein-binding values are presented in Supplementary file 1. This dataset was divided into two parts. The first one with 235 plasma protein-binding values, dedicated to develop the QSAR model. The second included 42 elements left for the external validation. The data was partitioned using holdout cross-validation.

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.[23] SMILES is a standard for specifying the structure of chemical species that takes the form of a line notation.[24] Table 1 lists 1666 descriptors that were sorted into twenty categories using the SMILES scripts for the 277 drugs. The E-Dragon online programs,[25] also known as the electronic remote version of the well-known software DRAGON created by the Milano Chemometrics and QSAR Research Group by Prof. R. Todeschini, were used to collect all descriptors. In Supplementary file 2, the name and number of calculated descriptors are presented.

Table 1. Number of calculated descriptors and their categories

Descriptors category	Number
Constitutional descriptors	48
Topological descriptors	119
Walk and pathcounts	47
Connectivity indices	33
Information indices	47
2D autocorrelations	96
Edgeadjacency indices	107
Burdeneigen value descriptors	64
Topological charge indices	21
Eigen value based indices	44
Randic molecular profiles	41
Geometrical descriptors	74
RDF descriptors	150
3D-morse descriptors	160
WHIM descriptors	99
GETAWAY descriptors	197
Functional group counts	154
Atomcentred fragments	120
Charge descriptors	14
Molecularproperties	31
Total	**1666**

Selection of relevant descriptors

Feature selection techniques are applied to decrease the number of elements in the dataset by choosing features that will give us better accuracy with less data.[26-28] It also reduces the overfitting and the overtraining risk.[29] Feature selection methods are widely available in the literature. The characteristics, advantages, and disadvantages of the three main strategies that can be used for the selection of relevant descriptors are reported in Table 2.[30]

The following procedure was used to reduce the number of molecular descriptors[31]:

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.[32]

4. Matrices of the pairwise linear correlation between each pair of the column in the input matrices were calculated via MATLAB.[33] Additionally, every variable that has a correlation coefficient $R > 0.75$ were removed. For more robustness of the model, the variance inflation factor VIF whose equation is as follows was calculated:

$$VIF_i = \frac{1}{1 - R_i^2} \quad (1)$$

Where R_i^2 is the squared correlation coefficient between the ith descriptor and the others. All descriptors with $VIF > 5$ were eliminated from the model.[34]

Model development

For the purpose of predicting the plasma protein-binding, the selected descriptors were used as inputs in FFNN. There are different approaches to discover the number of hidden neurons required for a modeling task explained in detail in a review named methods of selecting the number of hidden nodes in Artificial Neural Networks review.[35] In this work, the following steps were used to choose the number of neurons in the hidden layer[36]:

1. Initially, only five hidden neurons were taken.
2. The FFNN is trained until the mean square error does no longer seem to improve.
3. At this moment, five neurons are added to the hidden layer, each with randomly initialized weights, and resumed training.
4. The steps 2 and 3 are repeated until a termination criterion has been satisfied.

The mathematical equation of the model used for the prediction of protein binding is:

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

xi ($i = 1...p$) is the input that corresponds to the number of data included in the training of the ANN, i from 1 to 15, $wij(i = 1...p, j = 1...k)$ are weights from input to hidden layer, $b\,j$ $(j = 1...k)$ are biases of the neurons in the hidden layer, $k = 40$ for filter 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.

Model validation

We established internal and external validation criteria to assess the QSAR models' generalizability and predictive power. The following statistical parameters were used in our investigation to evaluate the models' efficacy: the mean squared error (MSE), correlation coefficient (R), predictive squared correlation coefficient (Q^2), and coefficient of determination (R^2) values.

$$R^2 = 1 - \frac{RSS}{SS} \quad (3)$$

Table 2. Feature selection methods and their advantages and disadvantages

Feature selection with filter methods	Feature selection with wrapper methods	Feature selection with embedded methods
Relevance of the features is calculated by considering the intrinsic properties of the data.	Wrapper methods select a subset of relevant features using a learning algorithm.	Includes the classifier construction for the optimal feature selection.
Use feature relevance score to select the top rank features.	Conduct search in the space of possible parameters.	Like wrapper approaches, these methods are specific to a given learning algorithm.
Examples	**Examples**	**Examples**
Information gain Correlation coefficient scores Chi squared test T-test	Genetic search Sequential forward selection Sequential backward elimination	Decision tree Weighted Naive Bayes SVM
Advantages	**Advantages**	**Advantages**
Can scale to high-dimensional data sets Fast and computationally inexpensive in comparison to wrapper method	Considers features dependencies Interaction with classifier Simple to implement	Classifier interaction Considers feature dependencies
Disadvantages	**Disadvantages**	**Disadvantages**
No interaction with the classifier Univariate feature selection methods do not consider feature dependencies/ redundancy	Higher risk of overfitting Selection based on classifier Computationally intensive	Classifier dependencies

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

$$Q^2 = 1 - \frac{PRESS}{SS} \quad (5)$$

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

Results and Discussion

The results obtained from the selection of the most important descriptors using the correlation coefficient *R* and the variance inflation factor *VIF* showed that 55 descriptors seemed to be the most appropriate. The calculated *VIFs* among the values of the selected descriptors are less than five, indicating that multicollinearity between the selected descriptors is acceptable. To get an overview of the correlation structure we used a heatmap to highlight what is important (Figure 2). Table 3 shows the *VIF* values for the selected descriptors and their meanings.

We followed the above-mentioned procedure to determine the required number of hidden neurons. The best model's accuracy was assessed using the *R*(all), *MSE*(validation), R^2_{train}, and Q^2 criteria. The best model was chosen based on the maximum *R*(all), R^2_{train}, and Q^2 and the lowest MSE (validation).[31,37] Table 4 shows 10 network models developed. The results obtained show that network eight with 40 neurons is the best model with *R* (all) = 0.990, R^2_{train} = 0.981, Q^2 = 0.989, and *MSE* (validation) = 0.002. The best performance of the model had a topology of (55-40-1): 55 input nodes, one hidden layer with 40 nodes having the hyperbolic tangent as a transfer function, and one output layer with an identity function. The neural networks were implemented using Neural Network Toolbox for MATLAB.[33] Figure 3 shows the predicted protein-binding values versus the experimental ones for the training and validation sets. The results show a close correlation between predicted and observed plasma protein-binding. The network type used is a Feed-Forward Network with the Levenberg-Marquardt backpropagation training function and gradient descent with momentum weight and bias learning function and the data was partitioned using holdout cross-validation. The difference between R^2_{train} and Q^2 was equal to 0.008. this difference did not exceed 0.3 indicating the robustness of the model.[38]

In order to investigate the predictability and performance of the model developed in this work, a statistical evaluation is carried out, as shown in Table 5. The model's robustness is demonstrated by the fact that the internal validation's statistical coefficients are all acceptable and satisfactory (lowest *MSE*, *RMSE*, and *MAE*, as well as high R^2_{train}, Q^2, $R^2_{adjusted}$). External validation parameters were also used to evaluate the model's quality. We can say that this model stands out due to its high predictive power. The excellent Q^2 value is greater than 0.9.[38]

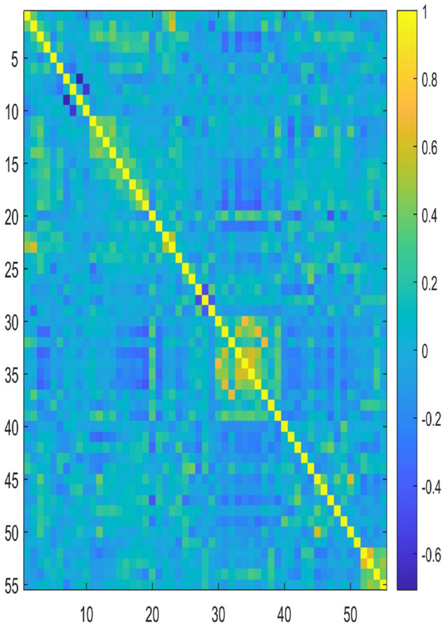

Figure 2. Heatmap of the correlation matrix for Filter method

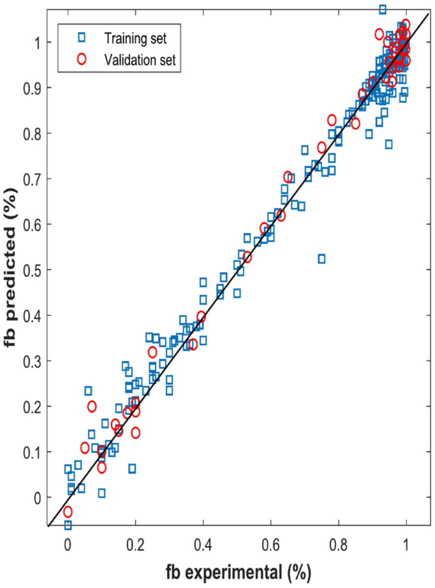

Figure 3. Comparison between experimental and predicted values for training and validation sets

Comparison between models from literature

We made a comparison between the few models reported in the literature with our developed model for the prediction of the binding of drugs to plasma proteins (Table 6). The evaluation of the advantages and disadvantages of these methods is quite difficult (each study used different data sets and different modeling approaches). We can see that the statistical parameters of our study exceed the models

Table 3. The VIF values for the selected descriptors by filter method

Descriptor	Type	Description	VIF
nX	Constitutional descriptors	number of halogen atoms maximal	2.6615
MAXDN	Topological descriptors	Electrotopological negative variation	4.4004
MAXDP	Topological descriptors	maximal electrotopological positive variation	3.5133
PJI2	Topological descriptors	2D Petitjean shape index	1.7543
Lop	Topological descriptors	Lopping centric index	1.9696
MATS1m	2D autocorrelations	Moran autocorrelation - lag 1 / weighted by atomic masses	2.9735
MATS2m	2D autocorrelations	Moran autocorrelation - lag 2 / weighted by atomic masses	4.0324
MATS4m	2D autocorrelations	Moran autocorrelation - lag 4 / weighted by atomic masses	2.6631
GATS2m	2D autocorrelations	Geary autocorrelation - lag 2 / weighted by atomic masses	3.3624
GATS4m	2D autocorrelations	Geary autocorrelation - lag 4 / weighted by atomic masses	2.7846
JGI2	Topological charge indices	Mean topological charge index of order2	2.4942
JGI3	Topological charge indices	Mean topological charge index of order3	3.5132
JGI4	Topological charge indices	Mean topological charge index of order4	1.8913
JGI5	Topological charge indices	Mean topological charge index of order5	2.3317
JGI6	Topological charge indices	Mean topological charge index of order6	1.9848
JGI7	Topological charge indices	Mean topological charge index of order7	2.0805
JGI8	Topological charge indices	Mean topological charge index of order8	1.6600
JGI9	Topological charge indices	Mean topological charge index of order9	2.0077
JGI10	Topological charge indices	Mean topological charge index of order10	1.8634
FDI	Geometrical descriptors	Folding degree index	2.7337
PJI3	Geometrical descriptors	3D Petitjean shape index	1.8725
DISPm	Geometrical descriptors	d COMMA2 value / weighted by atomic masses	2.3418
DISPe	Geometrical descriptors	d COMMA2 value / weighted by atomic Sanderson electronegativities	4.0247
Mor04m	3D-MoRSE descriptors	3D-MoRSE - signal 04 / weighted by atomic masses	2.4145
Mor12m	3D-MoRSE descriptors	3D-MoRSE - signal 12 / weighted by atomic masses	3.0770
Mor17m	3D-MoRSE descriptors	3D-MoRSE - signal 17 / weighted by atomic masses	1.8941
Mor26m	3D-MoRSE descriptors	3D-MoRSE - signal 26 / weighted by atomic masses	2.3693
Mor28m	3D-MoRSE descriptors	3D-MoRSE - signal 28 / weighted by atomic masses	2.5970
Mor31m	3D-MoRSE descriptors	3D-MoRSE - signal 31 / weighted by atomic masses	2.7935
G2u	WHIM descriptors	2st component symmetry directional WHIM index / unweighted	2.5765
G2m	WHIM descriptors	2st component symmetry directional WHIM index / weighted by atomic masses	2.7232
E2m	WHIM descriptors	2nd component accessibility directional WHIM index / weighted by atomic masses	2.9830
G2v	WHIM descriptors	2st component symmetry directional WHIM index / weighted by atomic van der Waals volumes	2.6210
G2e	WHIM descriptors	2st component symmetry directional WHIM index / weighted by atomic Sanderson electronegativities	3.5147
G2p	WHIM descriptors	2st component symmetry directional WHIM index / weighted by atomic polarizabilities	2.8308
G2s	WHIM descriptors	2st component symmetry directional WHIM index / weighted by atomic electrotopological states	2.9183
E2s	WHIM descriptors	2nd Component accessibility directional WHIM index / weighted by atomic electrotopological states	3.0706
ISH	GETAWAY descriptors	Standardized information content on the leverage equality	1.7718
HATS4m	GETAWAY descriptors	Leverage-weighted autocorrelation of lag 4 / weighted by atomic masses	3.2173
C-005	Atom-centred fragments	Atom-centred fragments	2.0372
C-006	Atom-centred fragments	Atom-centred fragments	2.3826
C-008	Atom-centred fragments	Atom-centred fragments	2.7744
C-025	Atom-centred fragments	Atom-centred fragments	2.4806
C-026	Atom-centred fragments	Atom-centred fragments	3.2520
C-040	Atom-centred fragments	Atom-centred fragments	3.5563
H-048	Atom-centred fragments	Atom-centred fragments	1.9890
H-052	Atom-centred fragments	Atom-centred fragments	2.3870

Table 3. Continued.

Descriptor	Type	Description	VIF
O-057	Atom-centred fragments	Atom-centred fragments	2.2844
O-060	Atom-centred fragments	Atom-centred fragments	2.7103
N-072	Atom-centred fragments	Atom-centred fragments	2.5844
N-075	Atom-centred fragments	Atom-centred fragments	2.0612
Inflammat-80	Molecularproperties	Ghose-Viswanadhan-Wendoloskiantiinflammatory at 80% (drug-like index)	2.6589
Hypertens-80	Molecularproperties	Ghose-Viswanadhan-Wendoloski antihypertensive at 80% (drug-like index)	2.8495
Hypnotic-80	Molecular properties	Ghose-Viswanadhan-Wendoloski hypnotic at 80% (drug-like index)	2.4199
Neoplastic-50	Molecular properties	Ghose-Viswanadhan-Wendoloski antineoplastic at 50% (drug-like index)	1.8707

published previously. Our model gives a high R^2, Q^2, $R^2_{adjusted}$ and lowest MSE, RMSE, MAE. According to these results, our model can be used for predicting plasma protein binding for new drugs saving amounts of money and time.

Applicability domain

A clearly defined applicability domain is recommended as the principle in OECD[41] guidelines. In this work, we analyzed the domain of applicability with different approaches reported in Table 7 with the results. The proposed approaches' algorithm and method can be found in the literature.[42,43]

The number of samples inside the applicability domain varied depending on the method used. Euclidean distance (95 percentile) and Classical KNN (Euclidean distance, k = 5) identified two test samples out of the domain of applicability. KNN (Euclidean distance k = 25) showed one of the test samples out of the applicability domain. Bounding box considered 03 test samples out of the applicability domain as shown in Figure 4. Although our points are far from the rest of the observations, they are close to the regression fitted line because they have a small residual, we speak of good leverage points. These results show that the model can be used to predict plasma protein binding for new compounds that have not been tested.

Conclusion

In this study, we constructed a QSAR model to predict 277 human plasma protein binding. The feature selection strategy by a Filter method has produced 55 inputs, which were used to train a FFNN for predictions. Examination of the estimates of external and internal criteria indicated that the QSAR model developed 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 0.966 and 0.063 respectively. 98.30% of the external validation set is correctly predicted. According to the OECD principle, we can say that this QSAR 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.

Table 4. Selected criteria of the different multi-layer perceptron for Filter method

Number of hidden neurons	*R*(all)	R^2_{train}	Q^2	MSE (validation)
5	0.849	0.743	0.707	0.039
10	0.857	0.729	0.664	0.032
15	0.870	0.774	0.743	0.031
20	0.872	0.780	0.714	0.038
25	0.917	0.839	0.780	0.026
30	0.957	0.918	0.882	0.020
35	0.955	0.953	0.818	0.024
40	0.990	0.981	0.989	0.002
45	0.944	0.901	0.875	0.014
50	0.832	0.694	0.714	0.027

Table 5. External and internal criteria of the model

Parameters	Value
Internal validation	
R (all)	0.991
R^2_{train}	0.981
Q^2	0.989
MSE	0.002
MAE	0.028
RMSE	0.039
$R^2_{adjusted}$	0.989
External validation	
R	0.983
Q^2	0.966
MSE	0.004
MAE	0.042
RMSE	0.063

Authors' Contribution

Conceptualization: Affaf Khaouane, Samira Ferhat, Salah Hanini.
Data curation: Affaf Khaouane, Samira Ferhat, Salah Hanini.
Formal analysis: Affaf Khaouane, Samira Ferhat, Salah Hanini.
Funding acquisition: Affaf Khaouane, Samira Ferhat, Salah Hanini.
Investigation: Affaf Khaouane, Samira Ferhat, Salah Hanini.
Methodology: Affaf Khaouane, Samira Ferhat, Salah Hanini.
Project administration: Affaf Khaouane, Samira Ferhat, Salah Hanini.
Resources: Affaf Khaouane, Samira Ferhat, Salah Hanini.

Table 6. Comparison with literature

Method	*MAE*	*R²*	*R*	*MSE*
Suggested method (Filter method)	Train 0.0313 Validation 0.0284 Test 0.0423	Train 0.981 Validation 0.989 Test 0.966	Train 0.991 Validation 0.995 Test 0.983	Train 0.031 Validation 0.028 Test 0.042
Yuan et al[18]	Test 0.076			
Sun et al[16]	Test 0.126			
Kumar et al[17]				Train 0.869 Test 0.8881
Li et al[39]		Train 0.86		
Ghafourian et al[12]	Train 13.25 Validation 14.96	Train 0.717 Validation 0.646	Train 0.681 Validation 0.641	
Moda et al[40]		Test 0.91		

Table 7. Applicability domain for Filter method

Approach	Test inside AD	Test outside AD
Bounding box (PCA)	39	3
Euclidan distance (95 percentile)	40	2
Classical KNN (Euclidean distance, k=5)	40	2
KNN (Euclidean distance, k=25)	41	1

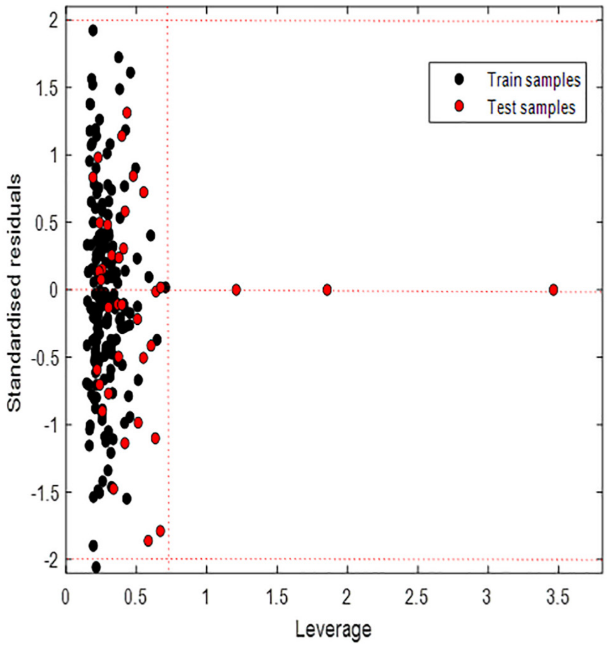

Figure 4. Plot of the residuals for calculated values of fb versus their experimental values for training and test sets

Software: Affaf Khaouane, Samira Ferhat, Salah Hanini.
Supervision: Affaf Khaouane, Samira Ferhat, Salah Hanini.
Validation: Affaf Khaouane, Samira Ferhat, Salah Hanini.
Visualization: Affaf Khaouane, Samira Ferhat, Salah Hanini.
Writing–original draft: Affaf Khaouane, Samira Ferhat, Salah Hanini.
Writing–review & editing: Affaf Khaouane, Samira Ferhat, Salah Hanini.

Competing Interests

None to be declared.

Ethical Approval

Not applicable.

Supplementary Files

Supplementary file 1. Chemical names and experimental protein-binding values.
Supplementary file 2. The name and number of calculated descriptors.

References

1. Luscombe D, Nicholls PJ. Processes of drug handling by the body. In: Smith HJ, ed. *Smith and Williams' Introduction to the Principles of Drug Design and Action*. 3rd ed. Overseas Publishers Association; 1998. p. 1-31.
2. Schmidt S, Gonzalez D, Derendorf H. Significance of protein binding in pharmacokinetics and pharmacodynamics. *J Pharm Sci* 2010;99(3):1107-22. doi: 10.1002/jps.21916
3. Zhivkova ZD. Quantitative structure - pharmacokinetics relationships for plasma protein binding of basic drugs. *J Pharm Pharm Sci* 2017;20(1):349-59. doi: 10.18433/j33633
4. Singh J, Petter RC, Baillie TA, Whitty A. The resurgence of covalent drugs. *Nat Rev Drug Discov* 2011;10(4):307-17. doi: 10.1038/nrd3410
5. 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
6. Bohnert T, Gan LS. Plasma protein binding: from discovery to development. *J Pharm Sci* 2013;102(9):2953-94. doi: 10.1002/jps.23614
7. 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
8. Sliwoski G, Kothiwale S, Meiler J, Lowe EW Jr. Computational methods in drug discovery. *Pharmacol Rev* 2014;66(1):334-95. doi: 10.1124/pr.112.007336
9. 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
10. 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. *Pharm Res* 2006;23(1):41-8. doi: 10.1007/s11095-005-8716-4
11. 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
12. 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
13. 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

14. Fu XC, Wang GP, Gao JQ, Zhan SY, Liang WQ. Prediction of plasma protein binding of cephalosporins using an artificial neural network. *Pharmazie* 2007;62(2):157-8. doi: 10.1691/ph.2007.2.5735
15. Deng L, Sui Y, Zhang J. XGBPRH: prediction of binding hot spots at protein-RNA interfaces utilizing extreme gradient boosting. *Genes (Basel)* 2019;10(3):242. doi: 10.3390/genes10030242
16. 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. *ChemMedChem* 2018;13(6):572-81. doi: 10.1002/cmdc.201700582
17. 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
18. 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
19. 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
20. Wallach I, Dzamba M, Heifets A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. arXiv [Preprint]. doi: https://doi.org/10.48550/arXiv.1510.02855.
21. Brunton LL, Knollmann BC. Goodman & Gilman's: The Pharmacological Basis of Therapeutics. 12th ed. McGraw Hill; 2010.
22. Anderson PO, Knoben JE, Troutman WG. *Handbook of Clinical Drug Data*. McGraw-Hill; 2002.
23. National Institutes of Health (NIH). PubChem. Available from: https://pubchemdocs.ncbi.nlm.nih.gov/.
24. 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
25. Vcclab. Virtual Computational Chemistry Laboratory. 2005. Available from: http://www.vcclab.org/.
26. Blum AL, Langley P. Selection of relevant features and examples in machine learning. *Artificial Intelligence* 1997; 97: 245-271. doi: 10.1016/S0004-3702(97)00063-5
27. 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
28. 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
29. 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
30. 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
31. Hamadache M, Benkortbi O, Hanini S, Amrane A, Khaouane L, Si Moussa C. 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
32. Stat Soft. STATISTICA, version 8.0. Available from: https://www.statsoft.com/.
33. MathWorks. MATLAB R2019b. Available from: https://www.mathworks.com/products/new_products/release2019b.html.
34. 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(7):754-67. doi: 10.4236/ojs.2015.57075
35. 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.
36. Kubat M. *An Introduction to Machine Learning*. Cham, Switzerland: Springer; 2017.
37. 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):471-89. doi: 10.1080/1062936x.2017.1331467
38. 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
39. 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
40. 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
41. OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*. Paris: OECD Publishing; 2014.
42. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* 2012;17(5):4791-810. doi: 10.3390/molecules17054791
43. Roy K, Kar S, Ambure P. On a simple approach for determining applicability domain of QSAR models. *Chemometr Intell Lab Syst* 2015;145:22-9. doi: 10.1016/j.chemolab.2015.04.013