HOA SEN UNIVERSITY
WORLD CLASS EDUCATION

Tạp chí Khoa học Trường Đại học Hoa Sen

https://www.hoasen.edu.vn/qlkh/ tapchidhhs.vn

UTILIZING ARTIFICIAL NEURAL NETWORKS TO EVALUATE THE AQUEOUS SOLUBILITY OF DRUG-LIKE COMPOUNDS

Duong Quang Trung[1], Nguyen Minh Quang*[2]

[1]*Faculty of Pharmacy, Ho Chi Minh City University of Technology, Ho Chi Minh City, Viet Nam*

[2]*Faculty of Chemical Engineering, Industrial University of Ho Chi Minh*

Article information

Received: 12/2023

Accepted: 1/2024

Available online: 3/2024

ABSTRACT

This study has propelled the evolution of Quantitative Structure-Property Relationship (QSPR) models for forecasting the aqueous solubility of drug-like compounds. Through the amalgamation of multivariate regression and neural network techniques, the investigation employed the backward algorithm to meticulously select 2D and 3D molecular descriptors, culminating in the creation of an optimal $QSPR_{MLR}$ model with k = 23. The artificial neural network regression model ($QSPR_{ANN}$), derived from chosen descriptors of the multivariable linear regression model ($QSPR_{MLR}$), exhibited heightened predictive prowess for logS values in both validation and prediction cohorts, yielding SE values of 0.786 and 0.808, respectively. The $QSPR_{ANN}$ significantly elevated the overall predictability of the multivariate regression model. Statistical evaluations of the $QSPR_{ANN}$ model unveiled SE = 0.699, R^2_{train} = 0.918, and Q^2_v = 0.878. The predicted logS values from the $QSPR_{ANN}$ model harmonize well with experimental data, validating the reliability and precision of the developed model.

Nghiên cứu này đã thúc đẩy sự phát triển của các mô hình quan hệ tính chất và cấu trúc định lượng (QSPR) để dự báo khả năng hòa tan trong nước của các hợp chất giống thuốc. Thông qua việc kết hợp hồi quy đa biến và kỹ thuật mạng thần kinh nhân tạo, nghiên cứu đã sử dụng thuật toán ngược để chọn lọc tỉ mỉ các bộ mô tả phân tử 2D và 3D, kết quả là mô hình $QSPR_{MLR}$ tối ưu với k = 23. Mô hình hồi quy phi tuyến mạng thần kinh nhân tạo ($QSPR_{ANN}$) được xây dựng bắt nguồn từ các mô tả đã chọn của mô hình hồi quy tuyến tính đa biến ($QSPR_{MLR}$), kết quả cho thấy khả năng dự đoán nâng cao đối với các giá trị logS trong cả bộ dữ liệu đánh giá và bộ dự đoán, với các giá trị SE lần lượt là 0,786 và 0,808. Mô hình $QSPR_{ANN}$ cũng đã nâng cao đáng kể khả năng dự đoán tổng thể của mô hình hồi quy đa biến. Đánh giá thống kê về mô hình $QSPR_{ANN}$ cho các giá trị khá tốt như SE = 0,699, R^2_{train} = 0,918 và Q^2_v = 0,878. Các giá trị logS dự đoán từ mô hình $QSPR_{ANN}$ phù hợp với dữ liệu thực nghiệm, xác nhận độ tin cậy và độ chính xác của mô hình đã phát triển.

Keywords: 2D and 3D descriptor; QSPR model; multivariate regression; aqueous solubility..

1. INTRODUCTION

The water solubility of a chemical compound stands as a pivotal property, wielding the potential to shape its biological activity and dictate its dispersion within the body. When a chemical compound grapples with inadequate solubility, it often emerges as a significant contributor to setbacks encountered in the advanced phases of drug development (Cheuk et al., 2020). Detecting and removing potential pharmacokinetics with insufficient solubility in the early stages constitutes a critical facet of drug discovery and development (Eric et al., 2012). Therefore, it is crucial to recognize this stage at an early juncture. Ideally, the removal of compounds lacking sufficient solubility should be anticipated and executed before embarking on drug synthesis (Savjani et al., 2012). Predicting solubility hinges entirely on computational techniques and solubility prediction methods.

* Corresponding author: Nguyen Minh Quang

Email: nguyenminhquang@iuh.edu.vn

In recent times, substantial endeavors have been dedicated to formulating robust mathematical models, enabling swift predictions of the aqueous solubility of compounds and resulting in a myriad of published studies. Numerous approaches for calculating the solubility of valuable chemicals have been introduced (Akbari et al., 2020). Various application approaches, encompassing both linear and nonlinear regression, have been successfully devised and utilized in conjunction with diverse structural representations. Despite considerable advancements and breakthroughs in implementing innovative modeling approaches, a variety of methods and descriptions of varying complexities persist (Cao et al., 2021). The performance methodologies of the majority of mathematical models identified in the literature are still modest, facing several challenges in drug synthesis, especially concerning diverse drug molecular structures.

Several factors play a role in the less-than-ideal predictions of compound solubility: (a) inadequacies in training data sets, which lack both drug-like and structurally diverse compounds; (b) challenges with experimental data collections, such as high experimental error, inconsistent procedures for measuring solubility, and reliance on kinetics instead of equilibrium; (c) insufficiently accurate representation of the effects of substances in different states; (d) validation of solubility models that aren't directly related to pharmacological properties.

A point under constant debate is whether the primary limitation affecting the performance of solubility prediction models lies in the quality of experimental data (Savjani et al., 2012). To tackle this issue, obtaining a larger volume of high-quality solutes may be necessary. Establishing a precise experimental dataset involves evaluating the consistency of results generated by the predictive model, a concept emphasized in various studies. The collected data should undergo standardization within a single laboratory, forming an initial training set that includes consistently defined experiments featuring diverse drug-like structures and known intrinsic solubility values. This approach has the potential to improve the model's performance, crafting a more suitable dataset for the development of predictive models.

Recent showcases of prediction model effectiveness have emerged through discoveries in the aqueous solubility challenge. Developing solubility prediction models with a dataset consisting of consistently defined experimental data remains a task that is inherently complex (Eric et al. 2012). Furthermore, researchers may utilize a variety of modeling techniques across a comprehensive solubility dataset. Insights from model and data challenges provide a unique perspective on the performance of all models, incorporating both linear and nonlinear approaches. Intriguingly, there are currently no universally established methods, highlighting a lack of consensus in the literature regarding the effectiveness of linear versus nonlinear models. Some authors tend to favor linear models, considering them more interpretable (Akbari et al., 2020). But some other work has shown that nonlinear methods can yield better predictability models.

In fact, some famous pharmaceutical companies in the world have used published research on QSPR and QSAR models as reference materials for drug discovery and development such as AstraZeneca, Boehringer Ingelheim, and GlaxoSmithKline. Specifically, AstraZeneca used QSPR models to prioritize compounds for preclinical testing, leading to the discovery of Brilinta (ticagrelor), a blood thinner for preventing heart attacks; Boehringer Ingelheim employed QSPR models t`o optimize the solubility of a compound, resulting in the development of Spiriva (tiotropium bromide), a long-acting inhaled medication for chronic obstructive pulmonary disease; GlaxoSmithKline leveraged QSPR models to assess the genotoxicity of potential drug candidates, reducing the need for animal testing (André et al., 2022).

When evaluating the predictive abilities of models, inherent distinctions arise between those derived from linear methods and nonlinear methods such as artificial neural networks (ANNs). The a`pplication of ANN models has shown limited potential for the efficacy seen in accepted models. (Eric et al. 2012; Savjani et al., 2012). ANN models possess reduced interpretability, frequently earning them the characterization of "black box" models. In numerous cases, the role of individual descriptors in a model created using specific ANN algorithms remains undisclosed, adding complexity to the interpretation of the model.

To tackle the challenge with ANN models, certain authors suggest adopting the concept of "local descriptor sensitivity." This approach entails assigning each descriptor a measure of its significance, proposing that the models' sensitivity to changes in the values of individual descriptors should be assessed independently based on specific characteristics (Akbari et al., 2020; Cao et al., 2021). The model captures a portion of the chemical space around the examined structure at a particular

point. This approach allows for the local assessment of the impact of each descriptor. Another tactic focuses on improving the informativeness of an ANN model by gauging the significance of descriptors in clarifying the relative influence of each one. It's acknowledged that not all ANN algorithms are equivalent. "Black box" ANN models can be supplemented with different types of ANN models that aid in data analysis (Eric et al. 2012). By employing component clustering evaluation, one can unveil the weight levels associated with distinct molecular descriptor symbols.

In this study, we present the development of robust Quantitative Structure-Property Relationship (QSPR) models designed to predict the solubility of drug-like molecules. Our approach combines regression and Artificial Neural Network (ANN) techniques, with algorithms automatically searched and adjusted to determine the relative importance of descriptors. The incorporation of these algorithms significantly enhances applicability, allowing for a detailed interpretation of descriptor contributions, which is crucial for achieving a high level of effectiveness in the QSPR models. Additionally, our QSPR modeling technique is well-suited for generating simpler and faster models. The synergy of regression techniques and ANN is demonstrated to streamline the modeling process by providing insights into the factors governing aqueous solubility.

2. MATERIALS AND METHOD

2.1. Data set

The dataset employed in this study was sourced from the same ADME database, comprising 1290 compounds with structural similarities, complemented by logS solubility data (Hou at el., 2004). The data were obtained through the same experimental procedure. In the logS database, water solubility is expressed in logS, where S represents solubility at 20-25°C in mol/L, forming the basis for our model construction. The information of Tetko was utilized in this procedure, and the study's database was randomly chosen from a pool of 902 chemicals (Hou at el., 2004). The SMILES flat-file representation of the dataset was transformed into an SDF structured data file (Wang at el., 2007). Solubility measurements within the dataset are gathered from diverse literature references, following specific criteria: (a) evaluation of drug-like compounds at room temperature; (b) inclusion of solubility values with intrinsic values roughly equivalent at 25°C (Atkin at el., 2012)..

2.2. Computation of Molecular Descriptors

The MM+ molecular-mechanic approach was used to create and geometrically optimize each structure. After that, configurations were optimized using the semi-empirical PM3 quantization approach until the ideal structures were reached. Nine hundred and two molecules have all two and three dimensional structural molecular descriptors computed (Tat et al., 2009; QSARIS, 2001). Five different types of molecular descriptors have been computed: 3D spatial structure, geometric structure, topological descriptors, and electrostatic potential descriptors. As an extra descriptor, the water-octanol partition coefficient (logP) was also computed. As a results, there are 240 molecular descriptions in all.

A heuristic technique was utilized to identify and eliminate less influential molecular characteristics. This approach has found widespread adoption in various studies for descriptor selection and the construction of linear models (QSARIS Reference Guide, 2000). The heuristic approach facilitates the elimination of descriptors with missing values and/or those showing low or zero variance. A descriptor is removed if its single-parameter correlation coefficient is established as statistically insignificant ($R^2 < 0.1$ or F-test value < 1.0). Descriptor pairs with the highest F-values are identified as new working sets and systematically merged to create three-parameter correlations. This process is iterated until the desired number of descriptors is achieved. Integrated addictiveness emphasizes closely linked descriptors ($R^2 > 0.8$). The sum of retained descriptors is determined based on the probability p-value of significance, resulting in the optimal correlation model. The optimal number of input descriptors is determined by selecting descriptors from the regression technique, evaluated based on correlation values. This comprehensive approach is elucidated for predicting the solubility of compounds during the model search.

2.3. Data set division

The dataset undergoes partitioning into training sets, validation sets, and test sets using a random sampling technique to build the QSPR models. The

initial dataset is divided into a 70% training set with 601 compounds, a 15% validation set containing 150 compounds, and a 15% test set consisting of 151 compounds. The development of QSPRANN models involves supervised training, incorporating all molecular input descriptors derived from the molecular descriptors screened by the regression algorithm (Quang at el., 2019). To validate the efficacy of the QSPR models, the statistical data set results are examined.

2.4. Computational Method

2.4.1. Standard Least Squares Model

Standard least-squares modeling is performed to create a model that conforms to various standard data models, including mixed multiple regression methods (Montgomery at el., 2001; Dehmer at et., 2012). The standard least squares model properties are utilized to build linear models for continuous response data through the least squares method. Visual statistical tools, graphs, and surface plots reinforce the outcomes of regression analysis. These intuitive statistical properties not only complement but also facilitate rapid model quality assessment. Additionally, they contribute to optimizing specific effect estimates for each descriptor.

2.4.2. Neural network model

The neural network model allows for the development of models for nonlinear datasets using nodes and layers. It aids in illustrating the relationship between input molecular descriptors and response variables within the dataset (Cao et al., 2021). At its essence, a neural network consists of a fully connected multilayer perceptron with one or two layers. Utilizing a neural network entails predicting one or more response variables by applying an activation function to the input variables. Neural network models excel as predictive models in situations where there is no critical requirement to intricately describe the functional form of the response surface (Quang at el., 2019). The neural network model utilizes the validation method to customize the dataset, employing techniques such as:

Holdback sampling

The neural network model is crafted by randomly dividing the initial dataset into training and validation sets. The training set consists of retained data, while the validation set is formed by excluding data from the original dataset.

K-fold sampling

In this approach, the initial data is randomly split into K smaller datasets. Each of these sub-datasets validates the neural network model against the remaining data, leading to the aggregation of K models. The ultimate model derived emphasizes the most favorable validation statistics

3. RESULTS AND DISCUSSION

3.1. Building $QSPR_{MLR}$ model

To minimize experimental errors in logS, we gathered the dataset from a single source. Analyzing the data distribution using the standard Gaussian distribution revealed that the density distribution of logS data for drug-like substances predominantly falls within the range of -11.62 to 1.58, as illustrated in Figure 1. This dataset is well-suited for constructing a multivariate regression model. To build an effective $QSPR_{MLR}$ model, it's essential to partition the dataset into a 70% training set, a 15% validation set, and a 15% test set. In this context, the Agglomerative Hierarchical Clustering method is employed to create similar groups of logS based on the dendrogram method (Quang at el., 2019).

The dataset is divided into a training set of 601 substances, a validation group with 150 substances, and the remaining substances forming the test group. The logS values of these substances are employed in the development of the $QSPR_{MLR}$ model, as outlined in Table 1. Constructing the $QSPR_{MLR}$ models involves using drug-like substances from the training group. Throughout the modeling process, back elimination and forward algorithms are applied to select molecular descriptors from the training dataset, which encompasses 240 2D and 3D molecular descriptors.

The chosen $QSPR_{MLR}$ models encompass a range of 1 to 23 molecular descriptors, and Table 1 details the most crucial 2D and 3D molecular descriptors selected, along with their statistical contributions evaluated based on important effects. Several 2D and 3D descriptors consistently appear in $QSPR_{MLR}$ models, underscoring their significance. Notably, descriptors such as x0, SssCH2, MaxNeg, SsCl, SaaCH, SdS, SdsCH, SsI, SsCH3, SsBr, SddssS, SdssS, SHBint4_Acnt, SaasC_acnt, SHBint5, SsNH2, SdaaN, SssNH, SdsN, SsssCH_acnt, SpcPolarizability, SssO, SsOH, and SsssN play a crucial role. Molecular descriptors x0, SssCH2, MaxNeg, SsCl, SaaCH, and SdS exhibit high t-ratio values, indicating their significance in the models.

$$\begin{aligned}\log S = {} & -1.109 - 0.270\times x_1 - 0.235\times x_2 - 4.979\times x_3 - 0.112\times x_4 - 0.261\times x_5 - 0.092\times x_6 - 0.483\times x_7 - 0.096\times x_8 \\ & - 2.004\times x_9 - 0.103\times x_{10} + 0.433\times x_{11} + 0.124\times x_{12} + 12.129\times x_{13} + 0.055\times x_{14} + 0.459\times x_{15} + 0.037\times x_{16} \\ & + 0.064\times x_{17} + 0.099\times x_{18} + 0.040\times x_{19} - 0.015\times x_{20} - 0.085\times x_{21} - 0.155\times x_{22} - 0.203\times x_{23} - 0.098\times x_{24}\end{aligned} \quad (1)$$

$R^2 = 0.885$; $R^2_{Adj} = 0.882$; $Q^2 = 0.835$; RMSE = 0.710; $F_{rat} = 282.261$; $F_{sig} = 0.0001$; DF = 901; p-values in range 0.0000 to 0.0063 at the confidence level $\alpha = 0.05$ for the regression coefficients.

Table 1: The quality of $QSPR_{MLR}$ model and the effects of descriptors are sorted by descending

Term	Descriptor	Parameter Quality				Important Effect		
		Coeff.	Std Error	t-Ratio	Prob>|t|	Term	Log Worth	Effect
C	Constant	-1.109	0.162	-6.850	<.0001			
x_1	x0	-0.270	0.013	-20.560	<.0001	x_1	76.217	
x_2	SssCH2	-0.235	0.013	-18.380	<.0001	x_2	63.343	
x_3	MaxNeg	-4.979	0.285	-17.470	<.0001	x_3	58.143	
x_4	SsCl	-0.112	0.007	-16.790	<.0001	x_4	54.353	
x_5	SaaCH	-0.098	0.009	-10.380	<.0001	x_5	45.384	
x_6	SdS	-0.261	0.026	-10.010	<.0001	x_6	23.142	
x_7	SdsCH	-0.092	0.010	-9.390	<.0001	x_7	21.673	
x_8	SsI	-0.483	0.061	-7.940	<.0001	x_8	19.292	
x_9	SsCH3	-0.096	0.015	-6.420	<.0001	x_9	14.219	
x_{10}	SsBr	-0.203	0.032	-6.390	<.0001	x_{10}	13.105	
x_{11}	SddssS	-0.155	0.032	-4.810	<.0001	x_{11}	11.424	
x_{12}	SdssS	-2.004	0.561	-3.580	0.0004	x_{12}	9.665	
x_{13}	SHBint4_Acnt	-0.103	0.030	-3.410	0.0007	x_{13}	9.579	
x_{14}	SaasC_acnt	-0.085	0.025	-3.390	0.0007	x_{14}	7.036	
x_{15}	SHBint5	-0.015	0.006	-2.740	0.0063	x_{15}	6.274	
x_{16}	SsNH2	0.040	0.013	3.230	0.0013	x_{16}	5.874	
x_{17}	SdaaN	0.433	0.109	3.980	<.0001	x_{17}	5.753	
x_{18}	SssNH	0.099	0.025	4.050	<.0001	x_{18}	4.249	
x_{19}	SdsN	0.064	0.013	4.870	<.0001	x_{19}	4.131	
x_{20}	SsssCH_acnt	0.124	0.025	5.050	<.0001	x_{20}	3.434	
x_{21}	SpcPolarizability	12.129	2.251	5.390	<.0001	x_{21}	3.163	
x_{22}	SssO	0.055	0.008	7.040	<.0001	x_{22}	3.142	
x_{23}	SsOH	0.037	0.005	7.600	<.0001	x_{23}	2.886	
x_{24}	SsssN	0.459	0.030	15.130	<.0001	x_{24}	2.200	

These molecular descriptors hold significant importance in the $QSPR_{MLR}$ model. The process of selecting the optimal $QSPR_{MLR}$ model (1), consisting of 23 molecular descriptors, is based on crucial statistical values such as R^2, R^2_{adj}, Q^2, and standard errors, as detailed in Table 1. The chosen $QSPR_{MLR}$ model forms the basis for building the $QSPR_{ANN}$ model with $k = 23$, representing the optimal configuration.

Utilizing the optimal $QSPR_{MLR}$ model (1) with 23 descriptors, as delineated in Table 1, allows for the assessment of the significant effects attributed to each descriptor. Log worth values provide insights into the substantial contributions made by individual descriptors. Through the cross-validation process, it is demonstrated that this constructed model can adeptly predict logS values. The $QSPR_{MLR}$ model effectively characterizes the training set, underscoring its statistical significance.

The $QSPR_{MLR}$ model with $k = 23$ exhibits robust predictability, as corroborated by the data presented in Table 1 and Figure 1, confirming its statistical appropriateness. Figure 1 visually depicts the correlation between experimental and calculated logS values derived from the $QSPR_{MLR}$ model (k = 23), with molecular descriptors organized by descending effect values as detailed in Table 1.

The computational outcomes presented in Table 1, which emphasize the noteworthy contribution levels of 2D and 3D molecular descriptors within the $QSPR_{MLR}$ model, clearly reveal the quantitative impact on each drug-like structure. This revelation carries significant implications for the design of novel drug molecules aimed at enhancing solubility. The standard error (SE) value functions as a tool to validate predictive outcomes by comparing the QSPR model's predictions with the experimental values (Tat at el., 2009; QSARIS, 2001):

$$SE = \sqrt{\sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{N - k - 1}} \quad (2)$$

Here yi and ŷi are experimental and calculated values logS; N is the number of experimental values; k is the number of descriptors in the $QSPR_{MLR}$ model.

Molecular descriptors like x0, SssCH2, MaxNeg, SsCl, SaaCH, and SdS exert influence on the Logworth values of logS, as indicated by their significant t-ratio values. The specific effects of these molecular descriptors are elaborated in Table 1.

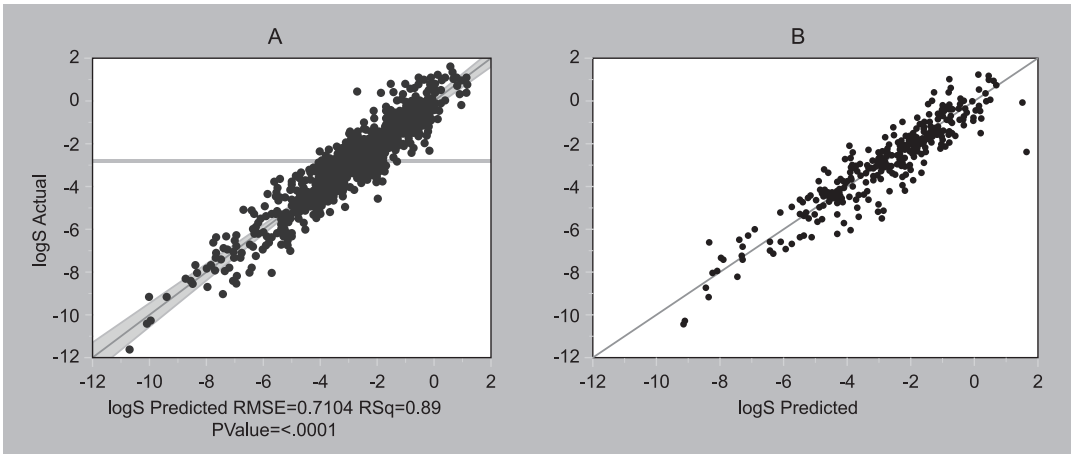
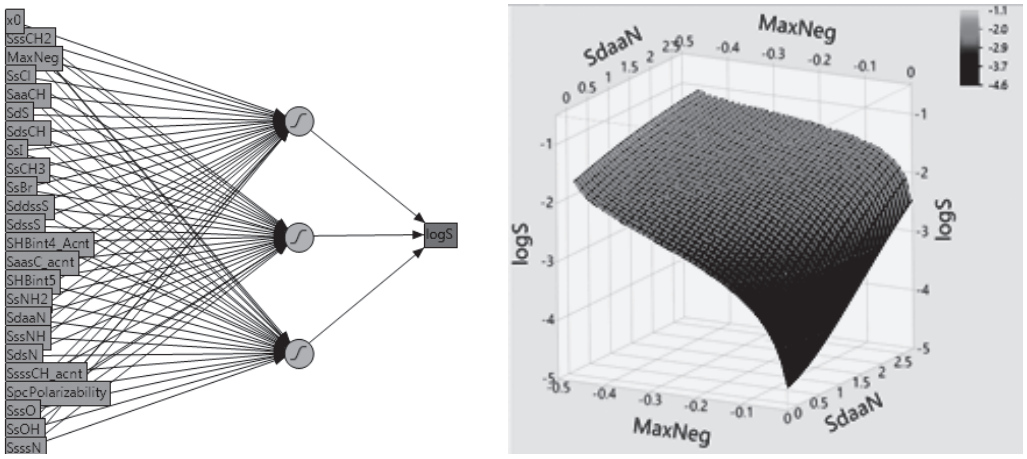


Figure 1. The correlation between experimental and calculated logS values derived from the $QSPR_{MLR}$ The correlation between experimental and calculated logS values derived from the $QSPR_{MLR}$ model ($k = 23$); A) the correlation of training set; B) the correlation of validation set.

3.2. Building $QSPR_{ANN}$ model

Creating a $QSPR_{ANN}$ model involves designing a neural network architecture with three layers, illustrated in Figure 2. The input layer is equipped with neurons corresponding to the number of molecular descriptors selected in equation (1). The hidden layer encompasses three neurons, and the output layer consists of one neuron representing the response value logS. All nodes in the hidden layer utilize the TanH transfer function, and the Sigmoid function is applied based on the number of nodes for each activation type. A learning rate of 0.1 is set. The network training process comprises 10,000 iterations for both the training set with 601 compounds and the validation set with 301 substances.



A B

Figure 2. A) The three-layer neural network model I(23)-HL(3)-O(1);.
B) the influence of molecular descriptors for logS values

It is essential to ascertain the optimal number of hidden layers and the required hidden neurons (*m*). In order to streamline the learning process and reduce complexity and noise in the neural network, we designed a neural network model I(23)-HL(*m*)-O(1). The quantity of neurons (*m*) on the hidden layer HL(*m*) can be determined using the relative rule proposed by Stathakis and Huang (Stathakis, 2009; Huang, 2003):

$$m = \sqrt{(x+60)/N} + \sqrt{N/(x+60)} \quad (3)$$

Here x output neurons; *m* the number of hidden neurons; N samples were used to train the neural network. In our study, $x = 1$, $N = 601$ training samples account for 70% of the data set. The number of neurons m on the hidden layer determined is three neurons. The neural network structure I(23)-HL(3)-O(1) was used for this study.

Table2: The statistical values resulting from the training and validation process of the $QSPR_{ANN}$ model I(23)-HL(3)-O(1)

Training results			Validation results	
Measures	**Value**		**Measures**	**Value**
R2	0.919		Q2v	0.878
RASE	0.657		RASE	0.756
Mean Abs Dev	0.448		Mean Abs Dev	0.548
-LogLikelihood	535.438		-LogLikelihood	328.615
SSE	259.267		SSE	171.883
Sum Freq	601		Sum Freq	301

Building the $QSPR_{ANN}$ model involves utilizing the 23 molecular descriptors from $QSPR_{MLR}$ model (1). The neural network architecture I(23)-HL(3)-O(1) is depicted in Figure 2A, with the input layer I(23) comprising x0, SssCH2, MaxNeg, SsCl, SaaCH, SdS, SdsCH, SsI, SsCH3, SsBr, SddssS, SdssS, SHBint4_Acnt, SaasC_acnt, SHBint5, SsNH2, SdaaN, SssNH, SdsN, SsssCH_acnt, SpcPolarizability, SssO, SsOH, and SsssN. The output layer O(1) includes a neuron representing the solubility value logS. The neural network undergoes training using the Holdback method with a holdback proportion parameter of 0.3333. Employing an error back-propagation algorithm, the MAD values for the training and validation sets are 0.448 and 0.548, respectively.

The $QSPR_{ANN}$ model exhibits superior predictability for the validation set compared to the $QSAR_{MLR}$ model, as depicted in Table 2, Figure 1, and Figure 3. Predicted logS values from the $QSPR_{ANN}$ model predominantly align with or closely approach the 95% confidence boundary. Furthermore, the correlation coefficients for the $QSPR_{ANN}$ model stand at R^2 of 0.919 and Q^2 of 0.878, indicating high confidence levels in its predictions. The QSPRANN model I(23)-HL(3)-O(1) robustly predicts logS values, making it applicable for drug-like substances in the training, validation, and test sets. Specifically, it reliably predicts logS values for newly designed anti-SARS-CoV-2 or anticancer substances, outperforming the $QSPR_{MLR}$ model, which exhibits higher prediction errors as indicated in Table 2.

In this context, we underscore the significance of drug-like substances in the development of diverse novel compounds. Current drug design strategies, focused on aqueous solubility, facilitate the creation of drugs with a multitude of activities. To expedite the virtual screening process from extensive databases, this study utilizes the $QSPR_{MLR}$ and $QSPR_{ANN}$ models alongside docking simulations to predict logS values for potential new anti-SARS-CoV-2 drugs. The $QSPR_{ANN}$ model I(23)-HL(3)-O(1) emerges as a valuable tool for predicting logS values for these newly designed substances, offering efficiency in the drug development pipeline.

As widely recognized, the interaction between a molecule and a protein receptor is inherently influenced by its spatial configuration. To thoroughly evaluate the influence of molecular structures, we have successfully compiled a database that incorporates both 2D and 3D molecular descriptors. In certain prior investigations concerning the development of SARS-CoV-2 inhibitors, 2D descriptors were employed to construct a 2D-QSAR model as proposed by Kumar and Roy (2020), Bobrowskia (Bobrowskia et al., 2020), and Amin (Amin et al., 2020). The 2D-QSAR models enable the interpretation and rapid prediction of SAR-CoV-2 inhibition for a derivative through a linear regression model (MLR) (Kumar and Roy, 2020; Ghosh at el., 2021; Bobrowsk at el., 2020; Amin at el., 2021). These 2D-QSAR models have shown success in predicting and designing n-Pyridines and n-Thiophenes derivatives that inhibit SARS-CoV (Kumar and Roy, 2020). While 2D parameters illustrate the flatness of a molecule, it's crucial to acknowledge that molecules can rotate around single bonds, introducing 3D structural properties. Hence, this investigation delves into an exhaustive array of molecular descriptors that encompass both 2D and 3D aspects.

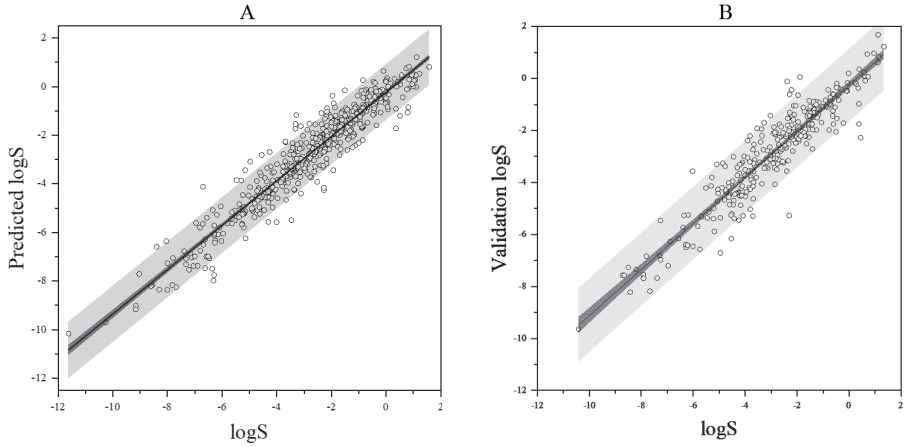


Figure 3. The predictability of the QSPRANN model for the training and validation set

4. CONCLUSION

We have adeptly crafted QSPRMLR and QSPRANN models for the accurate prediction of the aqueous solubility logS in drug-like substances. These models demonstrate strong predictive capabilities, utilizing thoughtfully selected 2D and 3D molecular descriptors, validated through standard statistical measures. Besides, QSPR models were developed using experimental data set with descriptors by using PM3 semi-empirical quantum calculations. The process of building QSPR models was carefully evaluated based on statistical indicators such as R2train, Q2v, and SE. In which, models QSPRMLR and QSPRANN I(23)-HL(3)-O(1) met the requirements for practical applicability. The results obtained from this work allow predicting and experimentally orienting the synthesis of new derivatives with good solubility. This achievement lays a robust foundation for guiding the development of cutting-edge drugs. The built model has the great advantage of ensuring regression parameters, thereby evaluating well the ability to predict the development of new chemicals. However, the weakness of the model is that there are too many variables, which also reduces the ability to select variables to guide the development of new derivatives.

REFERENCES

1. André M. D. O., Mithun R. & Chukwuebuka E. (2022). In Drug Discovery Update, Computer Aided Drug Design (CADD): From Ligand-Based Methods to Structure-Based Approaches, Chapter 4 - Quantitative structure-activity relationships (QSARs), Elsevier, 101-123, https://doi.org/10.1016/B978-0-323-90608-1.00007-1.
2. Akbari, F., Didehban, K., & Farhang, M. (2020). Solubility of solid intermediate of pharmaceutical compounds in pure organic solvents using semi-empirical models, European Journal of Pharmaceutical Sciences, 143, 105209.
3. Amin, S. K., Banerjee, S., Singh, S., Qureshi, I. A., Gayen S., & Jha, T., (2021). First structure–activity relationship analysis of SARS-CoV-2 virus main protease (Mpro) inhibitors: an endeavor on COVID-19 drug discovery, Mol. Divers., 25(3):1827-1838.
4. Atkins, P. & Paula J. D. (2012). Physical Chemistry. W. H. Freeman, Sixth Edition, USA.
5. Bobrowski, T., Alves, V., Melo-Filho, C. C., Korn, D., Auerbach, S. S., Schmitt, C., Muratov, E., & Tropsha, A., (2020). Computational Models Identify Several FDA Approved or Experimental Drugs as Putative Agents Against SARS-CoV-2., Chemrxiv, doi: 10.26434/chemrxiv.12153594.
6. Cao, Y., Khan, A., Zabihi, S., & Albadarin, A. B. (2021). Neural simulation and experimental investigation of Chloroquine solubility in supercritical solvent, Journal of Molecular Liquids, 333, 115942.
7. Cheuk, D., Svärd, M., & Rasmuson, Å. C. (2020). Thermodynamics of the Enantiotropic Pharmaceutical Compound Benzocaine and Solubility in Pure Organic Solvents, Journal of Pharmaceutical Sciences, 109, 3370-3377.
8. Dehmer, M., Varmuza, K., & Bonchev, D., (2012) Statistical Modelling of Molecular Descriptors in QSAR/QSPR, Wiley-VCH Verlag & Co. KGaA, Weinheim, Germany.
9. Eric, S., Kalinica, M., Popovic, A., Zloh, M. & Kuzmanovski, I. (2012). Prediction of aqueous solubility of drug-like molecules using a novel algorithm for automatic adjustment of relative importance of descriptors implemented in counter-propagation artificial neural networks, International Journal of Pharmaceutics, 437, 232-241.
10. Ghosh, K., Amin, S. A., Gayen, S., & Jha, T., (2020). Chemical-informatics approach to COVID-19 drug discovery: Exploration of important fragments and data mining based prediction of some hits from natural origins as main protease (Mpro) inhibitors, J. Molecular Structure, 1224, 129026, https://doi.org/10.1016/j.molstruc.2020.129026.
11. Hou, T., Xia, K., Zhang, W., & Xu, X. (2004). ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach, Journal of Chemical Information and Computer Sciences, 44, 266-275.
12. Huang, G. B., (2003). Learning capability and storage capacity of two-hidden-layer feed-forward networks, IEEE Transactions on Neural Networks, 14, 274-281.
13. Kumar V. & Roy, K. (2020). Development of a simple, interpretable and easily transferable QSAR model for quick screening antiviral databases in search of novel 3C-like protease (3CLpro) enzyme inhibitors against SARS-CoV diseases, SAR and QSAR in environmental research, 31(7), 511-526, 16 Jun 2020. https://doi.org/10.1080/1062936X.2020.1776388.
14. Montgomery, D. C., Peck, E. A., & Vining, C. G., (2001). Introduction to Linear Regression Analysis Third Edition, Wiley-Interscience, New York, USA.
15. QSARIS 1.1 (2001). Statistical Solutions Ltd., USA, 2001.
16. QSARIS Reference Guide (2000). Statistical Analysis and Molecular Descriptors. Academic Press, San Diego, USA, 2000.
17. Quang, N. M., Mau, T. X., Nhung, N. T. A., An, T. N. M., & Tat, P. V., (2019). Novel QSPR modeling of stability constants of metalthiosemicarbazone complexes by hybrid multivariate technique: GA-MLR, GA-SVR and GA-ANN, J. Molecular Structure, 1195, 95-109.
18. Savjani, K. T., Gajjar, A. K., & Savjani, J. K. (2012). Drug Solubility: Importance and Enhancement Techniques, International Scholarly Research Network, 195727.
19. Stathakis, D., (2009). How many hidden layers and nodes?, International Journal of Remote Sensing, 30(8), 2133–2147.
20. Tat, P. V (2009). Development of QSAR and QSPR. Publisher of Natural sciences and Technique, Hanoi, 2009.
21. Wang, J., Krudy, G., Hou, T., Holland, G., & Xu, X., (2007). Development of reliable aqueous solubility models and their application in drug-like analysis, Journal of Chemical Information