Employing artificial neural networks for the assessment of the aqueous solubility of drug–like substances

Duong Quang Trung¹, Pham Van Tat²

¹Faculty of Pharmacy, Ho Chi Minh City University of Technology, Ho Chi Minh City
²Institute of Pharmaceutical Education and Research, Binh Duong University, Thu Dau Mot City

Corresponding author: Pham Van Tat. Email: pvtat@bdu.edu.vn

Abstract: This research has advanced Quantitative Structure-Property Relationship (QSPR) models for predicting the aqueous solubility of drug-like substances. By integrating multivariate regression and neural network techniques, the study utilized the backward algorithm to strategically select 2D and 3D molecular descriptors, resulting in the development of an optimal QSPRMLR model with k = 23. The artificial neural network regression model (QSPRANN), derived from selected descriptors of the multivariable linear regression model (QSPRMLR), demonstrated enhanced predictive capabilities for logS values in both validation and prediction groups, yielding SE values of 0.786 and 0.808, respectively. The QSPRANN significantly improved the overall predictability of the multivariate regression model. Statistical assessments of the QSPRANN model revealed SE = 0.699, R²train = 0.918, and Q²v = 0.878. The predicted logS values from the QSPRANN model align well with experimental data, confirming the reliability and accuracy of the developed model.

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

1. Introduction

The solubility of a chemical compound in water is a crucial property that can impact its biological activity and influence its distribution within the body. In cases where a chemical compound exhibits poor solubility, it may play a substantial role in the failures observed during the late stages of drug development [1]. Identifying and eliminating potential pharmacokinetics with inadequate solubility at an early stage are crucial aspects of drug discovery and development [2]. Hence, it is imperative
to identify this stage early on. Ideally, the timely elimination of compounds with inadequate solubility is necessary and should be undertaken predictably before initiating drug synthesis [3]. The process of prediction relies solely on computational techniques and methods for predicting solubility.

In recent years, significant efforts have been invested in creating robust mathematical models that facilitate the rapid prediction of compound aqueous solubility, leading to a diverse range of published works. Various methods for calculating the solubility of valuable chemicals have been introduced [4]. Multiple application approaches, incorporating both linear and nonlinear regression, have been effectively developed and employed alongside diverse structural representations. Despite substantial breakthroughs and progress in adopting novel modeling approaches, a range of methods and descriptions of different complexities still coexist [5]. The performance methodologies of most mathematical models identified in the literature remain moderate and encounter numerous obstacles in drug synthesis, particularly for diverse drug molecular structures.

Various factors contribute to the unsatisfactory predictions of compound solubility: (a) training data sets lacking both drug-like and structurally diverse compounds; (b) issues with experimental data collections, including high experimental error, inconsistent procedures for measuring solubility, and the use of kinetics instead of equilibrium; (c) insufficiently representing the effects of substances in different states reliably; (d) confirming solubility models that are unrelated to pharmacological properties.

A widely debated point is that the quality of experimental data stands as the primary limiting factor affecting the performance of modeling processes designed for predicting solubility [3]. To address this, a larger quantity of high-quality solutes may be required. The precise experimental data set can be established by assessing the consistency of results obtained from the predictive model, a notion highlighted in various works. Collected data should be standardized from a single laboratory in an initial training set, encompassing uniformly defined experiments with diverse drug-like structures and known intrinsic solubility values. This approach can enhance the model's performance, creating a more suitable data set for predictive model development.

Recent demonstrations of the performance of prediction models have come from findings in the aqueous solubility challenge. Crafting solubility prediction models using a data set comprised of uniformly defined experimental data remains a task that is not inherently simple [2]. Moreover, researchers might employ diverse modeling techniques across an entire solubility dataset. The findings from model and data challenges offer a distinctive perspective on the performance of all models, encompassing both linear and nonlinear approaches. Interestingly, there are presently no universally proven methods, reflecting a lack of consensus in the literature regarding the efficacy of linear versus nonlinear models. Some authors lean towards linear models, finding them more
interpretable [4]. But some other work has shown that nonlinear methods can yield better predictability models.

When considering the predictive capabilities of models, inherent differences emerge between those derived from linear methods and nonlinear methods like artificial neural networks (ANNs). The utilization of ANN models has demonstrated restricted potential for the effectiveness of accepted models [2,3]. ANN models exhibit lower interpretability, often earning them the label of "black box" models. In many instances, the contribution of individual descriptors in a model developed using certain ANN algorithms remains undisclosed, rendering model interpretation more challenging.

To address this issue with ANN models, some authors propose employing “local descriptor sensitivity”. This involves assigning each descriptor a measure of its importance. The concept suggests that the sensitivity of models to changes in the values of individual descriptors should be evaluated independently based on specific characteristics [2-5]. The model represents a segment of the chemical space surrounding the studied structure at a specific point. Locally determining the influence of each descriptor is achievable through this approach. Another strategy involves enhancing the informability of an ANN model, measuring the descriptor's significance in elucidating the relative influence of each individual descriptor. It is recognized that not all ANN algorithms are equal. "Black box" ANN models can be complemented with various types of ANN models that assist in data analysis [2]. It is possible to discover through component clustering evaluation the weight levels corresponding to different molecular descriptor symbols.

In this study, we introduce the development of robust Quantitative Structure-Property Relationship (QSPR) models for predicting the solubility of drug-like molecules, employing a combination of regression and ANN techniques. The algorithms were automatically searched and adjusted to ascertain the relative importance of descriptors. The algorithms utilized in these QSPR models significantly enhance applicability, enabling a detailed interpretation of descriptor contributions. This proves pivotal in achieving a high level of applicability for the QSPR models. Furthermore, this QSPR modeling technique is well-suited for obtaining simpler and faster models. The combined approach of regression techniques and ANN is demonstrated to facilitate a more efficient model by explaining and analyzing the factors governing aqueous solubility.

2. Materials and method

2.1. Data set

The dataset utilized in this research was sourced from the identical ADME database, encompassing 1290 compounds that share structural similarities and are complemented by logS solubility data [6]. The data were acquired using the identical experimental procedure. In the logS database, water solubility is denoted in logS, with S representing solubility at 20-25°C in mol/L, serving as our foundation for constructing the model. Tetko's information was employed in this process, and the database for this study was randomly selected from a pool of 902.
Employing artificial neural networks for the assessment of the aqueous solubility of drug-like substances

chemicals [6]. The SMILES flat-file representation of the dataset underwent a conversion into an SDF structured data file [7]. The solubility measurements within the dataset are sourced from various literature references, adhering to specific criteria: (a) drug-like compounds evaluated at room temperature; (b) solubility values encompass intrinsic solubility values approximately equivalent at 25 °C [10].

2.2. Molecular descriptors calculation and pre-selection
Every structure was built and geometrically optimized utilizing the MM+ molecular-mechanic method. Subsequently, the semi-empirical PM3 quantization method was employed to optimize the configurations until achieving the optimal structures. The calculation of all 2D and 3D structural molecular descriptors was performed for 902 molecules [12,13]. The calculated molecular descriptors encompass five distinct types: geometric structure, topology descriptors, electrostatic potential descriptors, and 3D spatial structure. Additionally, the water-octanol partition coefficient (log P) was computed as an supplementary descriptor. In total, there are 240 molecular descriptions.

A heuristic technique was employed to select the less impactful molecular characteristics for elimination. This approach has been widely adopted in numerous studies for descriptor selection and the development of linear models [14]. The heuristic approach enables the removal of descriptors with missing values and/or those exhibiting low or zero variance. A descriptor is eliminated if the single-parameter correlation coefficient is established and found statistically insignificant (R2 < 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 form three-parameter correlations. This process is iterated until the desired number of descriptors is attained. The integrated additiveness aligns with closely linked descriptors (R2 > 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 through the selection of 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 is partitioned into training sets, validation sets, and test sets employing a random sampling technique for constructing the QSPR models. The original dataset was segmented into a 70% training set containing 601 compounds, a 15% validation set comprising 150 compounds, and a 15% test set consisting of 151 compounds. The construction of QSPRANN models relies on supervised training, incorporating all molecular-input descriptors derived from the molecular descriptors screened by the regression algorithm [11]. To verify the effectiveness of the QSPR models based on the evaluation statistical data set results.

2.4. Computational Method
2.4.1. Standard Least Squares Model
Standard least-squares modeling is executed to generate a model that adheres to various standard data models, encompassing mixed multiple regression
methods [8,9]. Properties of the standard least squares model are employed to construct linear models for continuous response data using the least squares method. Visual statistical tools, graphs, and surface plots support the results of regression analysis. These intuitive statistical properties serve to complement and facilitate swift model quality assessment. The statistical properties also enable the optimization of certain effect estimates for each descriptor.

2.4.2. Neural network model
The neural network model enables the creation of models for sets of nonlinear data through the utilization of nodes and layers. It facilitates the depiction of the relationship between input molecular descriptors and response variables within the dataset [5]. The core of a neural network comprises a fully connected multilayer perceptron with one or two layers. Employing a neural network involves predicting one or more response variables through an activation function applied to the input variables. Neural network models excel as predictive models when there's no imperative need to intricately describe the functional form of the response surface [11]. The neural network model employs the validation method to tailor the dataset, employing techniques such as:

*Holdback sampling*

The neural network model is created by randomly partitioning the initial dataset into training and validation datasets. The retained data serves as the training set, while the excluded data becomes the validation dataset [15,16].

*K-fold sampling*

This method randomly partitions the original data into K smaller datasets. Each sub-dataset is used to validate the neural network model against the remaining data, resulting in the summation of K models. The final model obtained exhibits the most favorable validation statistics [15,16].

3. Results and discussion
3.1. Building QSPRMLR model
The dataset was gathered from a single source to mitigate experimental error in logS. We assessed the data distribution using the standard Gaussian distribution. Test results revealed that the density distribution of logS data for drug-like substances was concentrated within the range of -11.62 to 1.58, as depicted in Figure 1. This dataset is well-suited for constructing a multivariate regression model. To create an effective QSPR\textsubscript{MLR} model, it is imperative to partition the dataset into a 70% training set, a 15% validation set, and a 15% test set. In this scenario, the Agglomerative Hierarchical Clustering method is employed to generate similar groups of logS based on the dendrogram method [11].

The set of 601 substances is utilized as the training set, while the group of 150 substances constitutes the validation group, and the remaining substances form the test group. LogS values of substances are employed in developing the QSPR\textsubscript{MLR} model, as outlined in Table 1.

The QSPR\textsubscript{MLR} models are constructed from drug-like substances within the training group. Back elimination and forward algorithms are applied in the modeling process to select molecular descriptors from the training dataset, encompassing 240 2D and 3D molecular descriptors. The number of molecular descriptors in the selected QSPR\textsubscript{MLR} models ranges from 1 to 23 molecular
Employing artificial neural networks for the assessment of the aqueous solubility of drug-like substances

descriptors. Table 1 enumerates the most crucial 2D and 3D molecular descriptors selected, with their statistical contributions evaluated based on important effects. Numerous 2D and 3D descriptors consistently appear in QSPR\textsubscript{MLR} models, highlighting their significance. Notably, descriptors such as $x_0$, SssCH\textsubscript{2}, MaxNeg, SsCl, SaaCH, SdS, SdsCH, SsI, SsCH\textsubscript{3}, SsBr, SddssS, SdsS, SHBint4\_Acnt, SaasC\_acnt, SHBint5, SsNH\textsubscript{2}, SdaaN, SssNH, SdsN, SsssCH\_acnt, SpcPolarizability, SssO, SsOH, and SssN play a crucial role. Molecular descriptors $x_0$, SssCH\textsubscript{2}, MaxNeg, SsCl, SaaCH, and SdS exhibit high $t$-ratio values, indicating their significance in the models. [13,14]. These molecular descriptors could be considered the most crucial in the QSPR\textsubscript{MLR} model.

The selection of the best QSPR\textsubscript{MLR} model (1) with 23 molecular descriptors is based on statistical values such as $R^2$, $R^2_{\text{adj}}$, $Q^2$, and standard errors, as outlined in Table 1. The QSPR\textsubscript{MLR} model is chosen to construct the QSPR\textsubscript{ANN} model with $k = 23$, representing an optimal model.

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

With $R^2 = 0.885$; $R^2_{\text{adj}} = 0.882$; $Q^2 = 0.835$; RMSE = 0.710; $F_{\text{rat}} = 282.261$; $F_{\text{sig}} = 0.0001$; 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\textsubscript{MLR} model and the effects of descriptors are sorted by descending

| Term      | Descriptor     | Coeff. | Std Error | t Ratio | Prob>|t| | Log Worth | Effect |
|-----------|----------------|--------|-----------|---------|--------|---------|----------|--------|
| C         | Constant       | -1.109 | 0.162     | -6.850  | <.0001 | x₁      | 76.217  |
| $x_1$     | $x_0$          | -0.270 | 0.013     | -20.560 | <.0001 | $x_1$   | 63.343  |
| $x_2$     | SssCH\textsubscript{2} | -0.235 | 0.013     | -18.380 | <.0001 | $x_2$   | 58.143  |
| $x_3$     | MaxNeg         | -4.979 | 0.285     | -17.470 | <.0001 | $x_3$   | 54.353  |
| $x_4$     | SsCl           | -0.112 | 0.007     | -16.790 | <.0001 | $x_4$   | 45.384  |
| $x_5$     | SaaCH          | -0.098 | 0.009     | -10.380 | <.0001 | $x_{24}$| 23.142  |
| $x_6$     | Sds            | -0.261 | 0.026     | -10.010 | <.0001 | $x_5$   | 21.673  |
| $x_7$     | SdsCH          | -0.092 | 0.010     | -9.390  | <.0001 | $x_6$   | 19.292  |
| $x_8$     | SsI            | -0.483 | 0.061     | -7.940  | <.0001 | $x_7$   | 14.219  |
| $x_9$     | SsCH\textsubscript{3} | -0.096 | 0.015     | -6.420  | <.0001 | $x_8$   | 13.105  |
| $x_{10}$  | SsBr           | -0.203 | 0.032     | -6.390  | <.0001 | $x_{23}$| 11.424  |
| $x_{11}$  | SddssS         | -0.155 | 0.032     | -4.810  | <.0001 | $x_{22}$| 9.665   |
| $x_{12}$  | SdsS           | -2.004 | 0.561     | -3.580  | 0.0004 | $x_9$   | 9.579   |
| $x_{13}$  | SHBint4\_Acnt  | -0.103 | 0.030     | -3.410  | 0.0007 | $x_{10}$| 7.036   |
| $x_{14}$  | SaasC\_acnt    | -0.085 | 0.025     | -3.390  | 0.0007 | $x_{21}$| 6.274   |
| $x_{15}$  | SHBint5        | -0.015 | 0.006     | -2.740  | 0.0063 | $x_{20}$| 5.874   |
Utilizing the optimal QSPRMLR model (1) with 23 descriptors, as outlined in Table 1, enables the determination of the significant effects of each descriptor. The log worth values provide insights into the substantial contributions of individual descriptors. The meanings of molecular descriptors in Table 1 described in references [12,14].

The cross-validation process demonstrates that this constructed model can be judiciously applied to predict logS values. The QSPRMLR model effectively characterizes the training set, showcasing statistical significance. The QSPRMLR model with \( k = 23 \) exhibits robust predictability, as evidenced in Table 1 and Figure 1, affirming its statistical appropriateness. Figure 1 illustrates the correlation between experimental and calculated logS values derived from the QSPRMLR model (\( k = 23 \)), with molecular descriptors arranged by descending effect values in Table 1.

The computation results in Table 1 for significant contribution levels of 2D and 3D molecular descriptors, as presented in the QSPRMLR model, distinctly reveal the quantitative impact on each drug-like structure. This finding holds crucial implications for the design of new drug molecules with enhanced solubility. The standard error SE value [13] can be used to validate the predictive results based on the prediction results from the QSPR model compared with the experimental value:

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

Here \( y_i \) and \( \hat{y}_i \) are experimental and calculated values logS; \( N \) is the number of experimental values; \( k \) is the number of descriptors in the QSPRMLR model.

The Logworth values of logS are influenced by molecular descriptors such as \( x_0, \text{SssCH}_2, \text{MaxNeg}, \text{SsCl}, \text{SaaCH}, \) and \( \text{SdS} \), evident from their substantial t-ratio values. The comparative effects of these molecular descriptors are detailed in Table 1.
Establishing a QSPR ANN model involves constructing a neural network architecture with three layers, as depicted 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, while the output layer consists of one neuron representing the response value logS. The transfer function TanH is applied to all nodes in the hidden layer, and the Sigmoid function is employed based on the number of nodes for each activation type. The learning rate is set at 0.1. The network training process entails 10,000 iterations for both the training set with 601 compounds and the validation set with 301 substances.

Determining the number of hidden layers and the required hidden neurons (m) is crucial. To streamline the learning process and minimize complexity and noise in the neural network, we fashioned a neural network model I(23)-HL(m)-O(1). The quantity of neurons (m) on the hidden layer HL(m) can be established following the relative rule put forth by Huang (2003) [15,16]:

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

(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.

Table 2. The statistical values resulting from the training and validation process of the QSPR\textsubscript{ANN} model \( I(23)-HL(3)-O(1) \)

<table>
<thead>
<tr>
<th>Measures</th>
<th>Training results</th>
<th>Validation results</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R^2 )</td>
<td>0.919</td>
<td>( Q^2_v )</td>
</tr>
<tr>
<td>RASE</td>
<td>0.657</td>
<td>RASE</td>
</tr>
<tr>
<td>Mean Abs Dev</td>
<td>0.448</td>
<td>Mean Abs Dev</td>
</tr>
<tr>
<td>-LogLikelihood</td>
<td>535.438</td>
<td>-LogLikelihood</td>
</tr>
<tr>
<td>SSE</td>
<td>259.267</td>
<td>SSE</td>
</tr>
<tr>
<td>Sum Freq</td>
<td>601</td>
<td>Sum Freq</td>
</tr>
</tbody>
</table>

Constructing the QSPR\textsubscript{ANN} model involves utilizing the 23 molecular descriptors from QSPR\textsubscript{MLR} model (1). The neural network architecture \( I(23)-HL(3)-O(1) \) is illustrated in Figure 2A. The neurons in the input layer \( I(23) \) encompass \( x_0 \), SsCH2, MaxNeg, SsCl, SaaCH, SdS, SdsCH, Ssl, SsCH3, SsBr, SddssS, SdssS, SHBint4_Acnt, SaasC_acnt, SHBint5, SsNH2, SdaaN, SssNH, SdsN, SsCH_acnt, SpcPolarizability, SssO, SsOH, and SsssN. The output layer \( O(1) \) consists of a neuron representing the solubility value \( \log S \). The neural network is trained using the Holdback method with a holdback proportion parameter of 0.3333. Employing an error backpropagation algorithm, the MAD values for the training and validation sets are 0.448 and 0.548, respectively.

The QSPR\textsubscript{ANN} model demonstrates superior predictability for the validation set compared to the QSAR\textsubscript{MLR} model, as illustrated in Table 2, Figure 1, and Figure 3. The predicted \( \log S \) values from the QSPR\textsubscript{ANN} model predominantly fall within or close to the 95% confidence boundary. Additionally, the correlation coefficients for the QSPR\textsubscript{ANN} model stand at \( R^2 \) of 0.919 and \( Q^2 \) of 0.878, indicating high confidence levels in its predictions. The QSPR\textsubscript{ANN} model \( I(23)-HL(3)-O(1) \) is robust in predicting \( \log S \) values, making it applicable for drug-like substances in the training, validation, and test sets. Specifically, it can reliably predict \( \log S \) values for newly designed anti-SARS-CoV-2 or anticancer substances, outperforming the QSPR\textsubscript{MLR} model, which exhibits higher prediction errors as indicated in Table 2.

In this context, we emphasize the significance of drug-like substances in the development of diverse novel compounds. Current drug design strategies, centered on aqueous solubility, facilitate the creation of drugs with a multitude of activities. To expedite the virtual screening process from extensive databases, this study employs the QSPR\textsubscript{MLR} and QSPR\textsubscript{ANN} models in conjunction with docking simulations to predict \( \log S \) values for potential new anti-
Employing artificial neural networks for the assessment of the aqueous solubility of drug-like substances

SARS-CoV-2 drugs. The QSPR$_{\text{ANN}}$ 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 is commonly understood, the interaction of a molecule with a protein receptor is influenced by its spatial configuration. In order to comprehensively assess the impact of molecular structures, we have effectively established a database encompassing both 2D and 3D molecular descriptors. In some previous studies on the development of SARS-CoV-2 inhibitors, 2D descriptors have been used to develop a 2D-QSAR model suggested by V. Kumar et al. (2020) [17], T. Bobrowskia et al. (2020) [19], and Sk.A Amin et al. (2020) [20]. The 2D-QSAR models enable the interpretation and rapid prediction of SAR-CoV-2 inhibition for a derivative through a linear regression model (MLR) [17-20]. These 2D-QSAR models have shown success in predicting and designing nPyridines and nThiophenes derivatives that inhibit SARS-CoV [17].

The study employed a backward algorithm to strategically select 2D and 3D molecular descriptors. The artificial neural network model (QSPR$_{\text{ANN}}$), derived from selected molecular descriptors, incorporating both 2D and 3D aspects.

4. Conclusion

We have established a database encompassing both 2D and 3D molecular descriptors. The 2D-QSAR models facilitate the interpretation and rapid prediction of SAR-CoV-2 inhibition for a derivative through a linear regression model, namely QSARMLR. This 2D-QSAR model has demonstrated success in predicting and designing derivatives of Pyridines and Thiophenes that inhibit SARS-CoV. The findings of this study have successfully unveiled a comprehensive set of molecular descriptors, incorporating both 2D and 3D aspects.
descriptors of the multivariable linear regression model (QSPRMLR), demonstrated improved predictive capabilities for logS values in both the validation and prediction groups.

References


Employing artificial neural networks for the assessment of the aqueous solubility of drug-like substances


