

# QSAR study of novel indole derivatives in hepatitis treatment by stepwise- multiple linear regression and support vector machine

Parastou Fattahi Sadr<sup>1</sup>, Mahmoud Ebrahimi<sup>1\*</sup>, Mehdi Nekoei<sup>2</sup>, Behzad Chahkandi<sup>2</sup>

<sup>1</sup> Department of Chemistry, Mashhad Branch, Islamic Azad University, Mashhad, Iran. <sup>2</sup>Department of Chemistry, Shahrood Branch, Islamic Azad University, Shahrood, Iran.

## Abstract

The quantitative structure – activity relationship (QSAR) of the novel indole derivatives for prediction of the half maximal inhibitory concentration (IC<sub>50</sub>), in hepatitis treatment was studied. After calculating the 1481 molecular descriptors, the stepwise (SW) was used as variable selection method to select the most appropriate molecular descriptors. The selected descriptors are Mp, MATS6e, GATS8e, Mor22v, R7v+ and MLOGP, which are of the Constitutional, 2D autocorrelations, 3D-MoRSE, GETAWAY, Molecular properties groups. Modeling was then performed using multiple linear regression (MLR) and support vector machine (SVM). The robustness and the predictive performance of the developed models was tested using both the internal and external statistical validation (test set) of ten compounds, randomly chosen out of 48 compounds. The SVM model with optimal parameters C of 11,  $\gamma$  of 1 and  $\epsilon$  of 0.07 has the R<sup>2</sup> (0.993, 0.844) and RMS errors (0.068, 0.269) for the training and test sets, respectively, which are better than MLR method (R<sup>2</sup>=0.886, 0.583 and RMS error=0.272, 0.441). Based on the information derived from the model, some key features for increasing the activity of compounds have been identified and can be utilized to designing new indole derivatives in hepatitis treatment.

**Keywords:** Indole derivatives, QSAR, MLR, SVM, hepatitis.

## INTRODUCTION

Hepatitis means inflammation in the liver parenchyma and can occur for a variety of reasons, some of which are contagious and some are not contagious. Factors that causes hepatitis include alcohol over-effects of certain medications as well as viruses [1]. Viral hepatitis can lead to liver infection [2, 3]. Hepatitis factor primarily symptoms like colds symptoms such as fatigue severe muscle pain and nausea. But advanced cases include abdominal swelling and organs, jaundice, gastrointestinal bleeding, and liver cirrhosis or liver cancer [4]. Hepatitis is one of the major health challenges and has over 325 million people worldwide. Hepatitis B and C are the root causes of liver cancer that kill more than one million people each year [1, 5]. Hepatitis B and C are chronic infections that do not show symptoms for a long time and may even persist for years or decades. At least 60% of liver cancer cases are due to delayed diagnosis and treatment of hepatitis B and C [6-9]. Because of these problems, it is necessary to find effective drugs to treat this disease. The process that led to the discovery and development of new drugs in the past was through trial and error, a time-consuming and costly method. Another problem that bothers scientists is their lack of knowledge of the pharmaceutical activity of the compounds prior to their synthesis and experimental investigation and that's why one of the most important goals of drug researchers

is to predict the activity of drug compounds before they are synthesized [10]. Therefore, it is necessary to use theoretical and computational methods that can predict the properties or activity of pharmaceutical compounds [11]. The advent of chemometrics science has largely eliminated these problems. One of the most important areas of application of chemometric methods is the study of the relationship between the activity of compounds and their structural properties [12]. These types of studies, called quantitative structure-activity relationships (QSAR), examine the relationship between activity of molecules with their structural and intrinsic

**Address for correspondence:** Mahmoud Ebrahimi, Department of Chemistry, Mashhad Branch, Islamic Azad University, Mashhad, Iran.  
E-mail: m.ebrahimi@mshdiau.ac.ir

This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License, which allows others to remix, tweak, and build upon the work noncommercially, as long as the author is credited and the new creations are licensed under the identical terms.

**How to cite this article:** Fattahi Sadr, P., Ebrahimi, M., Nekoei, M., Chahkandi, B. QSAR study of novel indole derivatives in hepatitis treatment by stepwise- multiple linear regression and support vector machine. Arch Pharma Pract 2020;11(S1):27-37.

properties<sup>[13]</sup>. In recent years, the increasing trend of papers published in science-based sources by QSAR implies the unique position of this perspective in deepening scientists' insights into understanding and justifying the mechanisms involved in the operation of a wide range of compounds<sup>[14-17]</sup>. Linear methods such as multiple linear regression (MLR) and nonlinear methods such as artificial neural networks (ANN) and support vector machines (SVM) are among the methods used in QSAR studies to model and predict drug activity<sup>[18-24]</sup>.

Indoles are the strongest compounds found in plants and vegetables that are widely effective in cancer, diabetic and hepatitis and other disorders. Indoles also have antioxidant and anti-atherogenic properties<sup>[25]</sup>. The indole nucleus is found in important drug compounds, such as sumatriptan, used in the treatment of migraine and ondansetron, which is used in the treatment of vomiting caused by cancer chemotherapy<sup>[26]</sup>. The figure 1 shows an indole structure. The main purpose of this study was to predict the pharmacological activity of some indole derivatives in the treatment of hepatitis C using MLR and SVM methods.

## MATERIALS AND METHODS

### Data Set

In this work, data series including drug activity of 48, indole derivatives as inhibitors for hepatitis C treatment were collected from the literature<sup>[27]</sup>. The inhibitory potency of these compounds is as the half maximal inhibitory concentration (IC<sub>50</sub>), which the minimum concentration of a drug compound that causes a 50% inhibitory effect on the disease. The IC<sub>50</sub> values were converted to the logarithmic scale pIC<sub>50</sub> [-log IC<sub>50</sub> (M)] and then used as the response variables for QSAR model. The chemical structures and activity data for the complete set of compounds are presented in Table 1.

### Descriptors Calculation and Reduction

Descriptors are numerical values that express the different properties of a molecule. In this work, Dragon software was used to calculate descriptors. 1481 descriptors, belonging to 18 different types of the theoretical descriptors, such as (the constitutional descriptors, topological descriptors, Galves topological charge indices, charge descriptors, 2D autocorrelations, geometrical descriptors, aromaticity indices, 3D-MoRSE descriptors, WHIM descriptors, radial distribution function (RDF) descriptors, functional groups, GETAWAY descriptors and other descriptors) can be calculated by DRAGON for each molecule. For this purpose, molecular structures were plotted and optimized by Hypercom software. Then the optimized structures were transferred to Dragon software and 1481 of descriptors was calculated for each compound. Due to the large number of descriptors, they should be reduced. Descriptors that have constant or near constant values (more than 90% of their data constant) are omitted. This will remove 303 descriptors. Then descriptors with a correlation higher than 0.9 were examined.

And among them, the descriptor less correlated with the independent variable was eliminated. Therefore, the number of 761 descriptors is eliminated, and finally the number of 417 descriptors remains.

### Variable Selection Method

One of the important steps in QSAR methods is to select the most appropriate descriptors for better prediction. Stepwise-Multiple Linear Regression (SW-MLR) was used to select the best descriptors. In this way the variables are entered one by one into the model. At first, the variable is selected that has the highest correlation with the dependent variable. The second variable that enters into the regression analysis is the variable that causes the most increase in the amount of R<sup>2</sup> after dividing the first variable. This will continue until the meaningful variable reaches %95 i.e., the error level is %5. Six descriptors were selected by SW-MLR from the 417 descriptors as the most appropriate. The descriptors selected by this method, with a brief description of them, are listed in Table 2.

## RESULT AND DISCUSSION

### Regression Models

Firstly, The SW-MLR was employed to model the quantitative structure-activity relationships with a different set of descriptors. In order to build and test the model, a data set consisting of 48 compounds was randomly divided into a training set of 38 compounds (80%), which were applied to build the model and a test set of 10 compounds (20%), which were used to test the QSAR model.

For the 48 compounds, the best correlation equation involving six descriptors (the closest to the ratio of five training molecules for each descriptor) with prediction ability for the test set was obtained. It is described by the following equation:

$$\begin{aligned} \text{pIC}_{50} = & 13.586(\pm 3.630) - 17.335(\pm 5.245) (\text{Mp}) \\ & - 4.294(\pm 0.940) (\text{MATS6e}) \\ & + 1.902(\pm 0.327) (\text{GATS8e}) \\ & - 1.047(\pm 0.408) (\text{Mor22v}) \\ & + 28.892(\pm 5.887) (\text{R7v+}) \\ & + 0.093(\pm 0.070) (\text{MLOGP}) \quad (1) \end{aligned}$$

$$\begin{aligned} N_{\text{train}} = 38; R^2_{\text{train}} = 0.886; \text{RMSE}_{\text{train}} = 0.272; Q^2_{\text{LOO}} = 0.835; \\ Q^2_{\text{LGO}} = 0.686; F_{\text{train}} = 40.552 \quad N_{\text{test}} = 10; R^2_{\text{test}} = 0.583; \\ \text{RMSE}_{\text{test}} = 0.441; F_{\text{test}} = 0.826 \end{aligned}$$

In this equation, N is the number of compounds, R<sup>2</sup> is the squared correlation coefficient, RMSE is the root mean square error, Q<sup>2</sup>LOO and Q<sup>2</sup>LGO are the squared cross-validation coefficients for leave one out and leave group out respectively, and F is the Fisher F statistic. Characteristics of the molecular descriptors in eq. (1) are shown in Table 3. As can be seen in Table 3, the variance inflation factor (VIF=1/(1-R<sub>j</sub><sup>2</sup>)) value for each descriptor is much lower than 10, indicating that the descriptors are poorly correlated with

each other. Therefore, it can be said that these descriptors can independently reflect the amounts of drug activity.

The experimental and predicted values based on SW-MLR model are listed in Table 1 and shown in Fig. 2. The squared correlation coefficient ( $R^2$ ) and RMS error are 0.583 and 0.441, respectively, for the test set. Although this model yields good results for the training set, but the results are not desirable for test set ( $R^2_{test} = 0.583$ ). Therefore, SVM was used for modeling and obtain better results. SVM is one of the supervised learning methods used for classification and regression. The non-linear SVM algorithm was invented by Vapnik in 1995. The theory of SVM is fully explained in our previous articles [18, 23]. The advantages of SVM are as follows: 1) The training is relatively simple. 2) Unlike artificial neural networks, it does not get stuck in the local maximums. 3) It works well for high dimensional data and 4) If the kernel function and parameters are selected well the error of prediction will be low.

The SVM parameters must be optimized first to obtain good results. The regularization constant  $C$  and the Gaussian function parameter  $\gamma$  and  $\epsilon$ -insensitive loss function were optimized as following: the search range parameter  $C$  is [1, 300]; and the search range of parameter  $\gamma$  is [0.1, 2]; and the search range of  $\epsilon$  is [0.01, 0.1]. The optimization results from the training set show that the optimal SVM model based on six molecular descriptors possesses parameters  $C$  of 11,  $\gamma$  of 1 and  $\epsilon$  of 0.07 (Figures 3-5). Table 1 shows the predicted values based on optimized SVM model. The plot of predicted pIC50 versus experimental values, obtained by the SVM modeling, are shown in Figure 6, too. This Figure shows a good agreement between the predicted and experimental values. The squared correlation coefficient ( $R^2$ ) for the training and test sets are 0.993 and 0.844, respectively. The RMS errors for two sets are 0.068 and 0.269, respectively, which are less than MLR method (Eq. (1)). The table 4 presents the statistical parameters calculated for SW-MLR and SVM methods.

Therefore, there is a nonlinear relationship between the inhibitory activity of indole derivatives and the six molecular descriptors.

### Interpretation of the descriptors

In this study, six descriptors selected by SW to obtain the best equation for prediction of pIC50. These molecular descriptors constitute a field for discovery a new drug for treatment of cancer.

The first descriptor in the model is  $M_p$  (Mean atomic polarizability (scaled on Carbon atom)). Atomic polarizability is the polarization caused by the deformation of the electric charge distribution between different atoms in the molecules. This descriptor is one of the constitutional descriptors [28].  $M_p$  displays a negative sign, which indicates that the pIC50 value is indirectly related to this descriptor.

The second and third descriptors are MATS6e and GATS8e. These descriptors are among 2D-autocorrelation (2DA) descriptors. 2DA descriptors generate histograms of atom pair distances within a molecule up to a cutoff distance. The major difference between these descriptors that designates their dimensionality is in their representation of interatomic distance. For 2DA, distances are measured in terms of the number of bonds between two connected atoms [28].

The 2DA descriptors are molecular descriptors calculated from molecular graph by summing the products of atom weights of the terminal atoms of all the paths of the considered path length (the lag). In 2DA class MATS and GATS are Moran autocorrelation and Geary Autocorrelation of Topological Structure. The physico-chemical property in MATS6e and GATS8e is an atomic sanderson electronegativities [29]. MATS6e displays a negative sign, which indicates that the pIC50 value is indirectly related to this descriptor while GATS8e displays a positive sign, which indicates that the pIC50 value is directly related to this descriptor.

Another important descriptor is Mor22v (3D-MoRSE - Signal 22/ weighted by van der Waals volume), which is among 3D-MoRSE descriptors. 3D MoRSE descriptors (3D Molecule Representation of Structures based on Electron diffraction) are derived from Infrared spectra simulation using a generalized scattering function. At a typical MoRSE descriptor  $m$  is weighted by mass  $v$  is weighted by van der Waals volume  $e$  is weighted by electronegativity  $p$  is weighted by polarizability [28]. Mor22v display a negative sign, which indicates that the pIC50 value is indirectly related to this descriptor.

The fifth descriptor is R7v+ (R maximal autocorrelation of lag 7 / weighted by van der Waals volume) which is one of the GETAWAY (Geometry, Topology, and Atom-Weights Assembly) descriptors [30]. GETAWAY descriptors are calculated from the leverage matrix obtained by the centered atomic coordinates and usually used for regression diagnostics are presented [31]. R7e+ display a positive sign, which indicates that the pIC50 value is directly related to this descriptor.

The final descriptor is MLOGP (Moriguchi octanol-water partition coeff. (logP)) which is one of the molecular properties descriptors [28]. MLOGP display a positive sign, which indicates that the pIC50 value is directly related to this descriptor.

Summarizing, from the above discussion it is concluded that the atomic polarizability, atomic sanderson electronegativities, van der Waals volume and octanol-water partition coefficient play a main role in the prediction of the half maximal inhibitory concentration (IC50) of the indole derivatives in hepatitis treatment.

### CONCLUSION

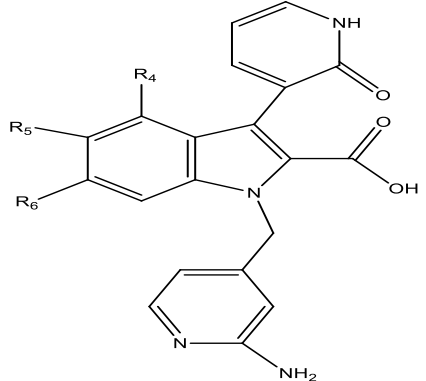
In this paper, QSAR models for prediction of the half maximal inhibitory concentration (IC<sub>50</sub>) of the novel indole derivatives in hepatitis treatment were successfully developed by means of MLR and SVM methods. The predictive ability and robustness of the models were evaluated by various statistical parameters such as the RMSEP, AARD and etc. The MLR model provided a clear output with RMSE of 0.441 and AARD of 5.392% for the test set. Whereas the SVM model shows more accurate predictions than the MLR model (RMSE of 0.269 and AARD of 2.677%). Therefore, the results of this research show the SVM method, together with appropriate descriptors, can be used to predict the activity of new indole derivatives in the treatment of hepatitis.

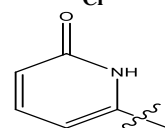

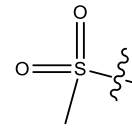
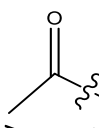
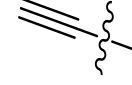
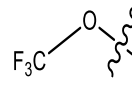
## REFERENCES

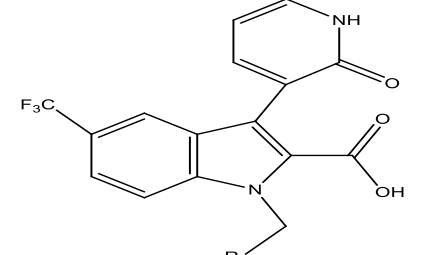
1. Manns MP, Buti M, Gane E, Pawlowsky JM, Razavi H, Terrault N. and Younossi Z. Hepatitis C virus infection. *Nature reviews Disease primers*, 2017; 3(1), pp.1-19.
2. Dienstag JL. Hepatitis B virus infection. *New England Journal of Medicine*, 2008; 359(14), pp.1486-1500.
3. Desmet VJ, Gerber M, Hoofnagle JH, Manns M. and Scheuer PJ. Classification of chronic hepatitis: diagnosis, grading and staging. *Hepatology*, 1994; 19(6), pp.1513-1520.
4. Islam N, Krajden M, Shoveller J, Gustafson P, Gilbert M, Buxton JA, Wong J, Tyndall MW, Janjua NZ. and Cohort BCHT. Incidence, risk factors, and prevention of hepatitis C reinfection: a population-based cohort study. *The lancet Gastroenterology & hepatology*, 2017; 2(3), pp.200-210.
5. Rossi C, Butt ZA, Wong S, Buxton JA, Islam N, Yu A, Darvishian M, Gilbert M, Wong J, Chapinal N. and Binka M. Hepatitis C virus reinfection after successful treatment with direct-acting antiviral therapy in a large population-based cohort. *Journal of hepatology*, 2018; 69(5), pp.1007-1014.
6. Wang C, Ji D, Chen J, Shao Q, Li B, Liu J, Wu V, Wong A, Wang Y, Zhang X. and Lu L. Hepatitis due to reactivation of hepatitis B virus in endemic areas among patients with hepatitis C treated with direct-acting antiviral agents. *Clinical Gastroenterology and Hepatology*, 2017; 15(1), pp.132-136.
7. Lok, A.S., Zoulim, F., Dusheiko, G. and Ghany, M.G., 2017. Hepatitis B cure: from discovery to regulatory approval. *Journal of hepatology*, 67(4), pp.847-861.
8. Ringehan M, McKeating JA and Protzer U. Viral hepatitis and liver cancer. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2017; 372(1732), p.20160274.
9. Pawlowsky JM. Hepatitis C virus resistance to direct-acting antiviral drugs in interferon-free regimens. *Gastroenterology*, 2016; 151(1), pp.70-86.
10. Kitchen DB, Decornez H, Furr JR. and Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*, 2004; 3(11), pp.935-949.
11. Sliwoski G, Kothiwale S, Meiler J. and Lowe EW. Computational methods in drug discovery. *Pharmacological reviews*, 2014; 66(1), pp.334-395.
12. Favia AD. Theoretical and computational approaches to ligand-based drug discovery. *Front Biosci*, 2011; 16, pp.1276-90.
13. Liu H. and Gramatica P. QSAR study of selective ligands for the thyroid hormone receptor  $\beta$ . *Bioorganic & medicinal chemistry*, 2007; 15(15), pp.5251-5261.
14. Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C. and Prachayasittikul V. QSAR study of H1N1 neuraminidase inhibitors from influenza A virus. *Letters in Drug Design & Discovery*, 2014; 11(4), pp.420-427.
15. Davood A. and Iman M. Molecular docking and QSAR study on imidazole derivatives as 14 $\alpha$ -demethylase inhibitors. *Turkish Journal of Chemistry*, 2013; 37(1), pp.119-133.
16. Ha H, Park K, Kang G. and Lee S, QSAR study using acute toxicity of *Daphnia magna* and *Hyalella azteca* through exposure to polycyclic aromatic hydrocarbons (PAHs). *Ecotoxicology*, 2019; 28(3), pp.333-342.
17. Pourbasheer E, Aalizadeh R, Ganjali MR. and Norouzi P. QSAR study of IKK $\beta$  inhibitors by the genetic algorithm: multiple linear regressions. *Medicinal Chemistry Research*, 2014; 23(1), pp.57-66.
18. Nekoei M, Mohammadhosseini M. and Pourbasheer E. QSAR study of VEGFR-2 inhibitors by using genetic algorithm-multiple linear regressions (GA-MLR) and genetic algorithm-support vector machine (GA-SVM): a comparative approach. *Medicinal Chemistry Research*, 2015; 24(7), pp.3037-3046.
19. Nekoei M, Salimi M, Dolatabadi M. and Mohammadhosseini M. Prediction of antileukemia activity of berbamine derivatives by genetic algorithm-multiple linear regression. *Monatshefte für Chemie-Chemical Monthly*, 2011; 142(9), p.943.
20. Beheshti A, Pourbasheer E, Nekoei M. and Vahdani S. QSAR modeling of antimalarial activity of urea derivatives using genetic algorithm-multiple linear regressions. *Journal of Saudi Chemical Society*, 2016; 20(3), pp.282-290.
21. Pourbasheer E, Ahmadvpour S, Zare-Dorabei R. and Nekoei M. Quantitative structure activity relationship study of p38 $\alpha$  MAP kinase inhibitors. *Arabian Journal of Chemistry*, 2017; 10(1), pp.33-40.
22. Nazari M, Tabatabai SA. and Rezaee E. Quantitative Structure Activity Relationships Study of Soluble Epoxide Hydrolase Inhibitors Using MLR, ANN, CoMFA and CoMSIA Methods. *ChemistrySelect*, 2019; 4(20), pp.6348-6353.
23. Pourbasheer E, Aalizadeh R, Ganjali MR. and Norouzi P. QSAR study of  $\alpha$ 1 $\beta$ 4 integrin inhibitors by GA-MLR and GA-SVM methods. *Structural Chemistry*, 2014; 25(1), pp.355-370.
24. Leechaisit R, Pingaew R, Prachayasittikul V, Worachartcheewan A, Prachayasittikul S, Ruchirawat S. and Prachayasittikul V. Synthesis, molecular docking, and QSAR study of bis-sulfonamide derivatives as potential aromatase inhibitors. *Bioorganic & medicinal chemistry*, 2019; 27(19), p.115040.
25. Almagro L, Fernández-Pérez F. and Pedreño MA. Indole alkaloids from *Catharanthus roseus*: bioproduction and their effect on human health. *Molecules*, 2015; 20(2), pp.2973-3000.
26. Demurtas M, Baldisserotto A, Lampronti I, Moi D, Balboni G, Pacifico S, Vertuani S, Manfredini S. and Onnis V. Indole derivatives as multifunctional drugs: Synthesis and evaluation of antioxidant, photoprotective and antiproliferative activity of indole hydrazones. *Bioorganic chemistry*, 2019; 85, pp.568-576.
27. Chen KX, Vibulbhan B, Yang W, Sannigrahi M, Velazquez F, Chan TY, Venkatraman S, Anilkumar GN, Zeng Q, Bennet F. and Jiang Y. Structure-activity relationship (SAR) Development and discovery of potent indole-based inhibitors of the Hepatitis C Virus (HCV) NS5B Polymerase. *Journal of medicinal chemistry*, 2012; 55(2), pp.754-765.
28. Todeschini R, Consonni V. *Handbook of molecular descriptors*. Wiley-VCH, Weinheim, 2000.
29. Gupta MK. and Prabhakar YS. Topological descriptors in modeling the antimalarial activity of 4-(3', 5'-disubstituted anilino) quinolines. *Journal of chemical information and modeling*, 2006; 46(1), pp.93-102.
30. Consonni V, Todeschini R. and Pavan M. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *Journal of Chemical Information and Computer Sciences*, 2002; 42(3), pp.682-692.
31. Khan AKR, Sahu VK, Singh RK. and Khan SA. Comparative QSTR study of saturated alcohols based on topological, constitutional, geometrical, and getaway descriptors. *Medicinal chemistry research*, 2009; 18(9), p.770.



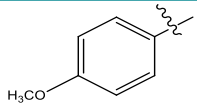
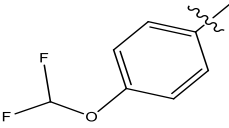
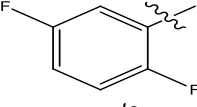
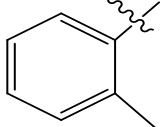
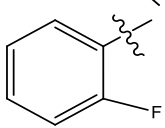
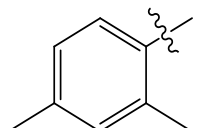
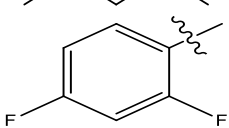
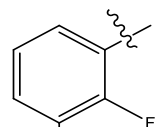
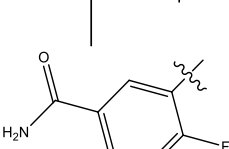
**Table 1:** Chemical structures and experimental and predicted activities (pIC<sub>50</sub>) for indole derivatives by SW-MLR and SVM.

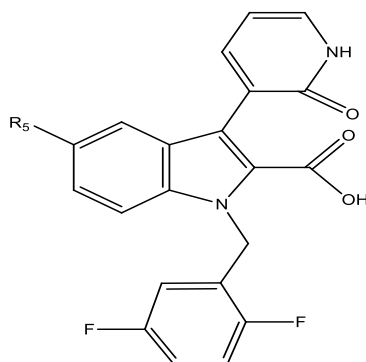


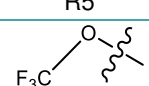
| Comp. | R <sub>4</sub> | R <sub>5</sub>  | R <sub>6</sub> | IC <sub>50</sub> | pIC <sub>50</sub> | SW-MLR       | SW-SVM       |
|-------|----------------|---|----------------|------------------|-------------------|--------------|--------------|
| 1     | Br             | H   | H              | 4.0              | 5.398             | 5.397        | 5.328        |
| 2*    | H              | Cl  | H              | <b>0.053</b>     | <b>7.276</b>      | <b>6.916</b> | <b>7.283</b> |
| 3     | H              | Cl  | Cl             | 0.047            | 7.328             | 7.138        | 7.258        |
| 4     | H              | Br  | H              | 0.044            | 7.357             | 7.113        | 7.287        |
| 5*    | Cl             | Cl  | H              | <b>0.75</b>      | <b>6.125</b>      | <b>6.509</b> | <b>6.043</b> |
| 6     | H              |    | H              | 7.0              | 5.155             | 5.462        | 5.225        |
| 7     | H              |   | H              | 0.065            | 8.301             | 6.900        | 7.117        |
| 8     | H              |  | H              | 1.4              | 5.854             | 6.396        | 5.924        |
| 9     | H              |  | H              | 0.33             | 6.481             | 6.295        | 6.411        |
| 10    | H              |  | H              | 0.048            | 7.319             | 7.464        | 7.249        |
| 11    | H              |  | H              | 0.016            | 7.796             | 7.704        | 7.726        |
| 12    | H              | CF <sub>3</sub>   | H              | 0.017            | 7.770             | 7.534        | 7.700        |

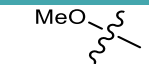

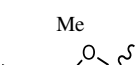
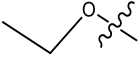
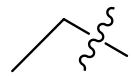
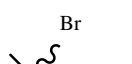
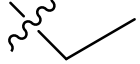
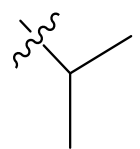
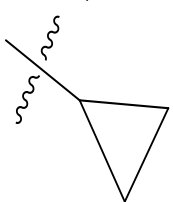


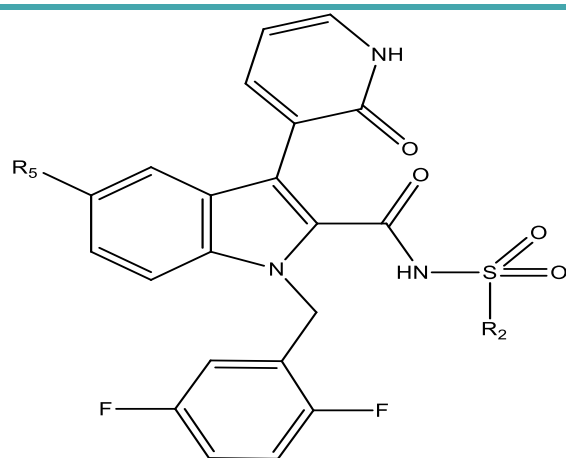
| Comp. | R <sub>1</sub> | IC <sub>50</sub> | pIC <sub>50</sub> | SW-MLR | SW-SVM |
|-------|----------------|------------------|-------------------|--------|--------|
|       |                |                  |                   |        |        |


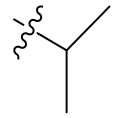
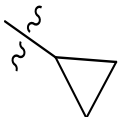
|     |   |              |              |              |              |
|-----|---|--------------|--------------|--------------|--------------|
| 13  |    | 0.025        | 7.602        | 7.768        | 7.672        |
| 14  |    | 0.067        | 7.174        | 7.728        | 7.244        |
| 15* |    | <b>0.004</b> | <b>8.398</b> | <b>8.104</b> | <b>8.270</b> |
| 16  |    | 0.009        | 8.046        | 7.920        | 7.994        |
| 17  |    | 0.009        | 8.046        | 8.001        | 7.976        |
| 18  |    | 0.016        | 7.796        | 7.896        | 7.726        |
| 19  |   | 0.025        | 7.602        | 7.049        | 7.672        |
| 20* |  | <b>0.005</b> | <b>8.301</b> | <b>7.590</b> | <b>7.942</b> |
| 21  |  | 0.007        | 8.155        | 7.524        | 8.085        |



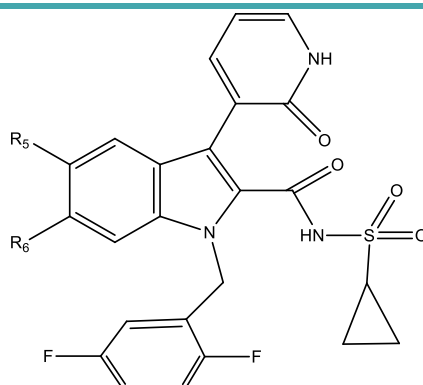
| Comp. | R5  | IC50         | pIC50        | SW-MLR       | SW-SVM       |
|-------|---|--------------|--------------|--------------|--------------|
| 22*   |  | <b>0.024</b> | <b>7.620</b> | <b>8.039</b> | <b>7.913</b> |

|            |   |              |              |              |              |
|------------|---|--------------|--------------|--------------|--------------|
| 23         |  | 0.015        | 7.824        | 7.888        | 7.894        |
| 24         |  | 0.15         | 7.225        | 7.223        | 8.894        |
| 25         |  | 0.008        | 8.097        | 8.444        | 8.167        |
| <b>26*</b> |  | <b>0.079</b> | <b>7.294</b> | <b>7.757</b> | <b>7.700</b> |
| 27         |  | 0.004        | 8.398        | 8.186        | 8.328        |
| 28         |  | 0.003        | 8.523        | 8.809        | 8.453        |
| 29         |  | 0.008        | 8.097        | 7.931        | 8.027        |
| 30         |  | 0.008        | 8.097        | 7.808        | 8.027        |
| <b>31*</b> |  | <b>0.009</b> | <b>8.046</b> | <b>7.874</b> | <b>7.746</b> |



| Comp.      | R5        | R2  | IC50         | pIC50        | Pred.        | Res.         |
|------------|-----------|---|--------------|--------------|--------------|--------------|
| 32         | Me        | Me  | 0.007        | 8.155        | 8.179        | 8.223        |
| 33         | Me        |  | 0.010        | 8.000        | 8.063        | 8.070        |
| 34         | Me        |  | 0.007        | 8.155        | 8.345        | 8.102        |
| 35         | Me        |  | 0.006        | 8.222        | 8.296        | 8.292        |
| <b>36*</b> | <b>Et</b> | <b>Me</b>   | <b>0.006</b> | <b>8.222</b> | <b>7.871</b> | <b>8.116</b> |

|    |    |  |       |       |       |       |
|----|----|--|-------|-------|-------|-------|
| 37 | Et |  | 0.006 | 8.222 | 7.921 | 8.152 |
| 38 | Et |  | 0.005 | 8.301 | 8.004 | 8.231 |
| 39 | Et |  | 0.003 | 8.523 | 8.611 | 8.453 |



| Comp. | R5               | R6              | IC50         | pIC50        | SW-MLR       | SW-SVM       |
|-------|------------------|-----------------|--------------|--------------|--------------|--------------|
| 40    | CF <sub>3</sub>  | H               | 0.003        | 8.523        | 8.019        | 8.453        |
| 41    |                  | H               | 0.006        | 8.222        | 8.040        | 8.152        |
| 42*   | F <sub>3</sub> C | H               | <b>0.007</b> | <b>8.155</b> | <b>7.882</b> | <b>7.984</b> |
| 43    |                  | H               | 0.006        | 8.222        | 8.351        | 8.292        |
| 44    | Me               | Cl              | 0.005        | 8.301        | 8.231        | 8.231        |
| 45*   | Me               | F               | <b>0.008</b> | <b>8.097</b> | <b>8.596</b> | <b>8.106</b> |
| 46    | Me               | CF <sub>3</sub> | 0.007        | 8.155        | 7.936        | 8.085        |
| 47    | CF <sub>3</sub>  | F               | 0.007        | 8.155        | 8.324        | 8.225        |
| 48    |                  | F               | 0.005        | 8.301        | 8.488        | 8.285        |

\* Used as test set

**Table 2:** The name, meaning and type of the descriptors selected by the SW-MLR.

| Descriptor symbol | Meaning   | Descriptor Type         |
|-------------------|---|-------------------------|
| <b>Mp</b>         | Mean atomic polarizability (scaled on Carbon atom).                       | Constitutional indices. |
| <b>MATS6e</b>     | Moran autocorrelation of lag 6 / weighted by Sanderson electronegativity. | 2D autocorrelations.    |
| <b>GATS8e</b>     | Geary autocorrelation of lag 8 / weighted by Sanderson electronegativity. | 2D autocorrelations.    |
| <b>Mor22v</b>     | Signal 22/ weighted by van der Waals volume.                              | 3D-MoRSE                |
| <b>R7v+</b>       | R maximal autocorrelation of lag 7 / weighted by van der Waals volume.    | GETAWAY                 |
| <b>MLOGP</b>      | Moriguchi octanol-water partition coeff. (logP).                          | Molecular properties.   |



**Table 3:** Characteristics of the descriptors selected in the Eq. (1).

| Descriptor | Coefficient | Std. Error | Mean effect | t-Value | VIF   |
|------------|-------------|------------|-------------|---------|-------|
| Mp         | -17.335     | 5.245      | 2.042       | -3.304  | 3.373 |
| MATS6e     | -4.294      | 0.940      | 0.002       | -4.563  | 2.061 |
| GATS8e     | 1.902       | 0.327      | -0.674      | 5.808   | 2.005 |
| Mor22v     | -1.047      | 0.408      | -0.057      | -2.565  | 1.440 |
| R7v+       | 28.892      | 5.887      | -0.264      | 4.907   | 2.789 |
| MLOGP      | 0.093       | 0.070      | -0.048      | 1.333   | 2.915 |

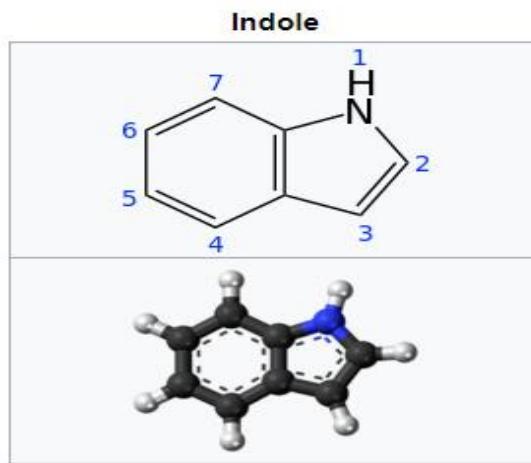
**Table 4:** Statistical parameters calculated for SW-MLR and SVM methods

| Parameter                           | Set      | Modeling approach |        |
|-------------------------------------|----------|-------------------|--------|
|                                     |          | SW-MLR            | SW-SVM |
| Determination coefficient ( $R^2$ ) | Test set | 0.583             | 0.844  |
| REP <sup>a</sup>                    | Test set | 5.709             | 3.483  |
| RMSEP <sup>b</sup>                  | Test set | 0.441             | 0.269  |
| AARD <sup>c</sup>                   | Test set | 5.392             | 2.677  |
| F-statistical                       | Test set | 0.826             | 2.602  |

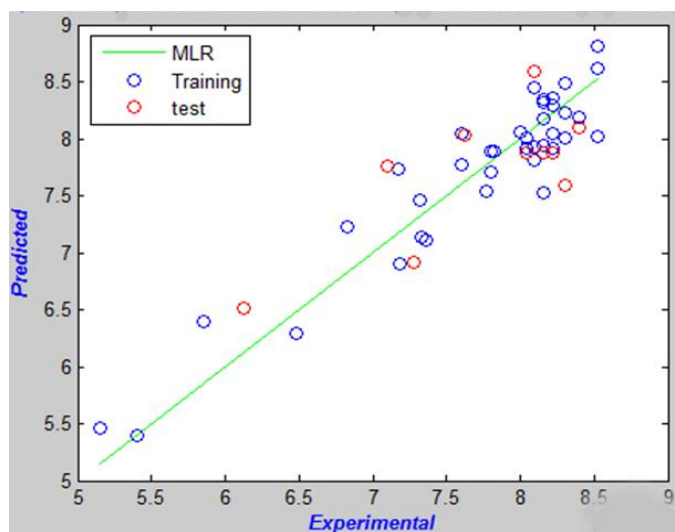
<sup>a</sup> Relative error of prediction

<sup>b</sup> Root mean square error of prediction

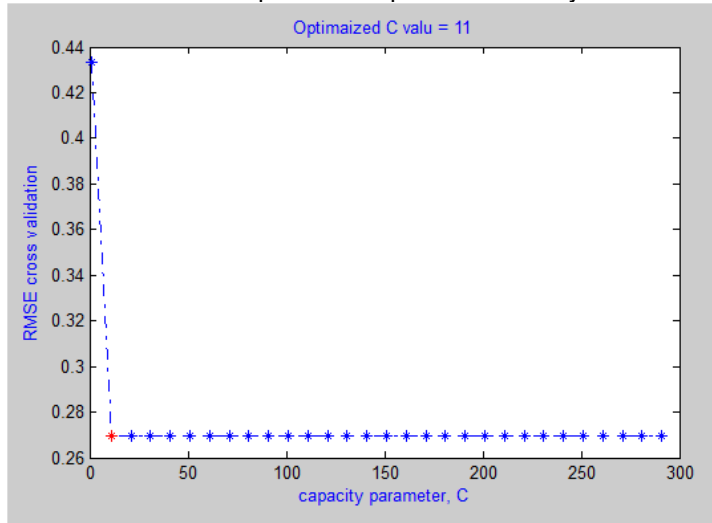
<sup>c</sup> Absolute average relative deviation



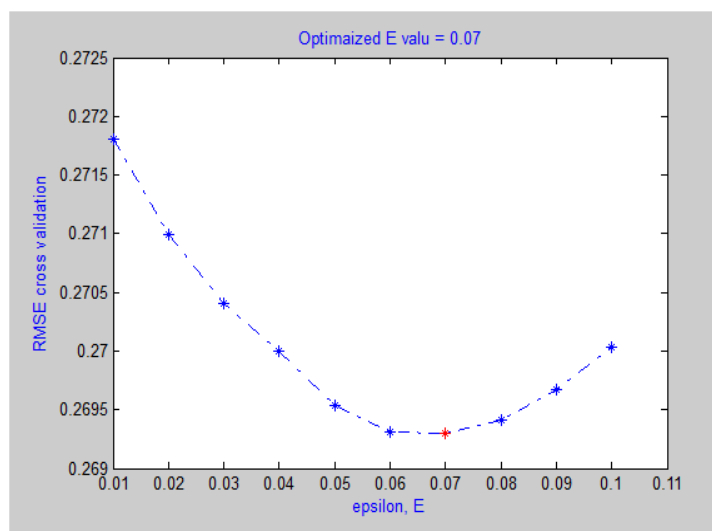
**Figure 1:** The indole structure (C<sub>8</sub>H<sub>7</sub>N: an aromatic heterocyclic organic compound)



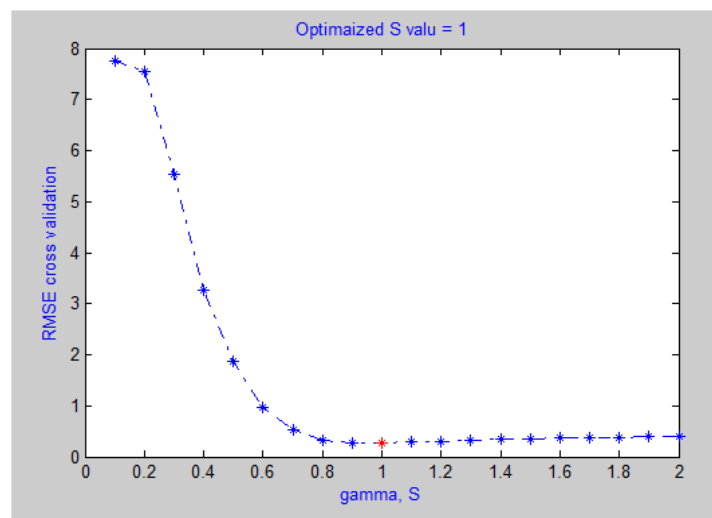
**Figure 2:** The predicted versus the experimental pIC50 values by the SW-MLR modeling.



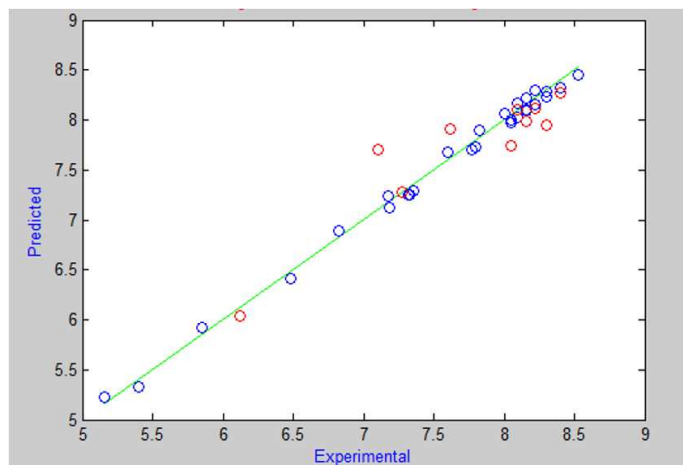
**Figure 3.** The RMSE versus capacity parameter (C) for the training set.



**Figure 4:** The RMSE versus epsilon (E) for the training set.



**Figure 5:** The RMSE versus gamma for the training set.



**Figure 6:** The predicted versus the experimental pIC50 values by the SW-SVM modeling.