



Research Article

Machine learning algorithm used to build a QSAR model for pyrazoline scaffold as anti-tubercular agentT Prabha*¹, C Selvinthanuja¹, S Hemalatha², S Sengottuvelu², J Senthil³

Nandha College of Pharmacy, Erode, Tamilnadu, India

ABSTRACT

Machine learning has become an essential tool for drug research to generate pertinent structural information to design drugs with higher biological activities. In this paper, we used python program language on pyrazoline scaffold, which is collected from diverse literature for the inhibition of *Mycobacterium tuberculosis*. Pyrazoline, a small molecule scaffold could block the biosynthesis of mycolic acids, resulting in mycobacteria death and leading to anti-tubercular drug discovery. The generated QSAR model afforded the ordinary least squares (OLS) regression as $R^2 = 0.380$, $F=4.909$, and $Q^2 = 0.303$, reg. coef_ developed were of 0.00651593 (molecular weight), -0.0069445 (hydrogen bond acceptor), -0.07576775 (hydrogen bond donor), -0.239021 (LogP) and reg. intercept of 3.10331589018553 developed through statsmodels.formula module. The support vector machine of the sklearn module generated the model score of 0.6294242262068762, the developed model was cross-validated by using the test set compounds and plotting the linear curve between the predicted and actual $pMIC_{50}$ value. We have found that the values obtained using this script correlated well and may be useful in the design of a similar group of pyrazoline analogs as anti-tubercular agents.

Keywords: Machine Learning, QSAR, Python, *Mycobacterium tuberculosis*, Pyrazoline scaffold.

Received - 11/12/2021, Reviewed - 20/12/2021, Revised/ Accepted- 29/12/2021

Correspondence: T Prabha* ✉ drtpappa@yahoo.com

Nandha College of Pharmacy, Tamilnadu, India

INTRODUCTION

Mycobacterium tuberculosis (MTB) remains the leading cause of worldwide death among infectious diseases. The MTB enoyl acyl carrier protein reductase (InhA) plays a vital role in the MTB fatty acid synthesis pathway as essential in the mycolic acid biosynthesis. Therefore, inhibition of InhA could block the biosynthesis of mycolic acids, resulting in mycobacteria death^[1, 2]. A series of pyrazoline conjugates have been collected from various reported literature and identified as promising molecules against *Mycobacterium tuberculosis* (MTB).

The drug discovery requires the use of hybrid technologies for the discovery of new chemical substances that could be the hopeful novel candidates for treating *Mycobacterium tuberculosis*. Quantitative structure-activity relationship research (QSAR) has been considered an important tool in drug discovery to design newer candidates for several therapeutic areas^[3]. It provides useful insights into the structural features which are responsible for the biological activity and helps to generate a mathematical model that can predict the activity of untested compounds quantitatively. With the growth of chemical data from combinatorial chemistry and virtual screening,

machine learning has become an essential tool for drug research to

generate pertinent structural information to design drugs with higher biological activities^[4]. Nowadays, sophisticated machine learning tools could be used to establish a QSAR model, also the development of machine learning models could be used to predict the potential impact of changes in chemical structures on biological activity^[5]. Currently, there is an increasing interest in the use of Artificial Neural Networks (ANN) and Support Vector Machines (SVM) to the QSAR field. Many of these QSAR models published in the literature are not utilized for designing new drugs. The main reason is use of costly commercial software which prevents many researchers from testing and adopting published models. If computational models were generated by open source software and then it can be more easily shared in scientific community. Here, we report a new python script developed for doing the statistical analysis and validation of this script using data of a series of pyrazoline derivatives, which are collected from various reported literature and identified as possible lead compounds against *Mycobacterium tuberculosis* (MTB).

From literature survey, in recent years pyrazolines have attracted considerable interest because of their therapeutics and the pharmacological properties such as antibacterial, antifungal^[6], antiviral^[7], antitubercular^[8], ant amoebic^[9], anticancer^[10], anti-

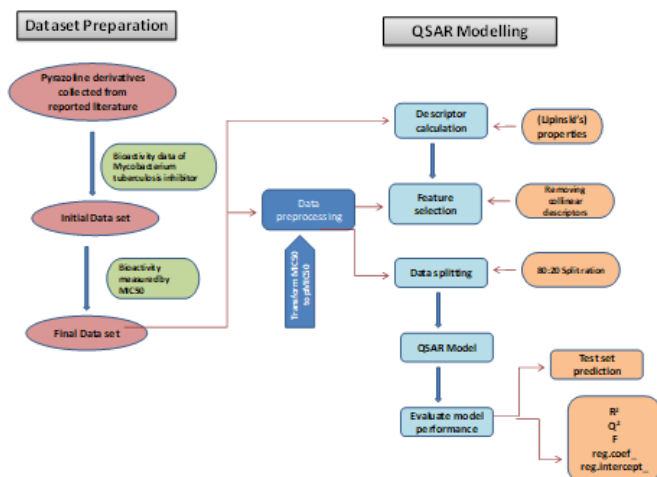
inflammatory^[11], analgesic^[12], antidepressant^[13], and Anti-convulsant activity^[14]. This study aimed to develop a QSAR models based on the machine learning for pyrazoline derivatives as antitubercular agents. Lipinski's molecular descriptors such as, logp, molecular weight, hydrogen bond donor and hydrogen bond acceptor were calculated and used after appropriate pre-processing in building QSAR models. Machine learning techniques like multiple Linear regression (MLR) and support vector machine (SVM) were employed to identify the correlation between the structures of pyrazoline (i.e., as described by the molecular descriptors) and their respective bioactivity (i.e., the MIC₅₀ values). The number of nodes used in the input layer was equal to the number of the descriptors presented in data set (i.e., four descriptors) while one node was used in the output layer corresponding to the MIC₅₀ value.

MATERIAL AND METHOD

Programming Language used

Workflow of the software Using Python programming language, software was written to perform regression analysis. Python (Python Programming Language—Official Website, python.org) has several advantages like open source, cross platform, object-oriented programming, dynamic typing features, simple and easy to learn and understand a program, rich set of supporting libraries for mathematics, statistics, and visualization.

Figure 1. Workflow of QSAR model via Machine learning



We used python modules like NumPy (Scientific Computing Tools for Python—Numpy, numpy.scipy.org), Scipy (Open Source Library of Scientific Tools, <http://www.scipy.org>), Python-Sklearn^[15], and matplotlib for linear algebra calculations, statistical values, machine/statistical learning, data mining, and plotting, respectively. A flow chart illustrating the work of QSAR model is given in Figure 1. The input data for the software is a comma-separated values file with each descriptor calculated from Molinspiration online tools (molinspiration.com/cgi-bin/properties) that calculates Lipinski's molecular descriptors,

descriptors (X=independent variables), as a column and biological activity data (Y=dependent variable) in first column.

Data collection

The antitubercular effects of diverse pyrazoline derivatives on *Mycobacterium tuberculosis* strain H37Rv, which is collected from reported literature^[16-18]. Such data are derived from different laboratories, have been generated at different times, most likely with different reagents and laboratory equipment.

Dataset preparation

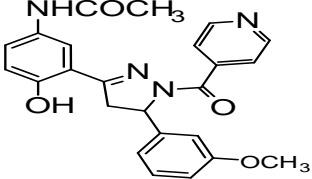
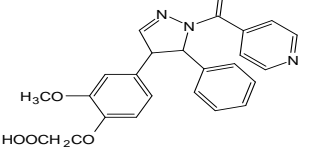
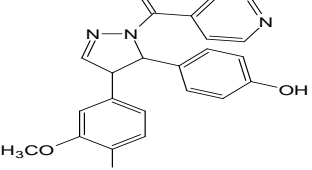
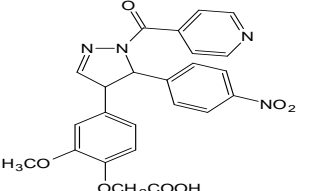
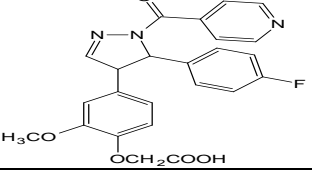
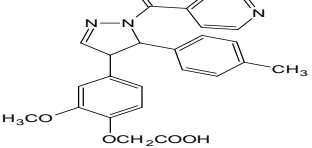
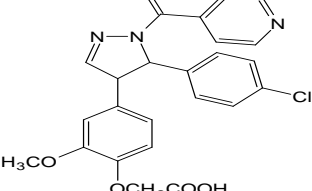
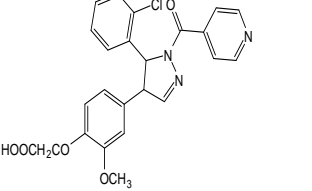
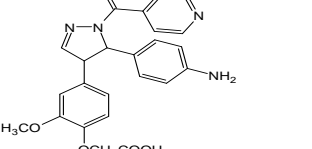
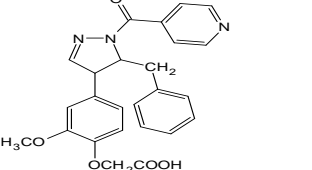
A dataset of 37 compounds with their antitubercular activity values (MIC₅₀ in μM/ml) on *Mycobacterium tuberculosis* strain H37Rv were collected from the reported literature^[16-18]. Then we converted the biological activity data into logarithmic scale [-log(IC₅₀)] based on data and shuffled for doing the machine learning model. From this, 80 % of the data were considered as training set (29 observations) and 20% were test set (8 observations). The structures of pyrazoline derivatives are listed in Table 1.

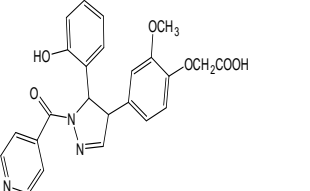
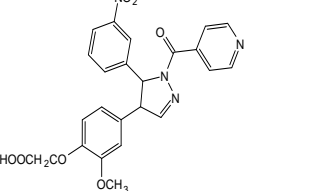
Table 1: The structures of pyrazoline derivatives

CODE	STRUCTURE	MIC ₅₀ (μM/ml)	pMIC ₅₀ (μM/ml)
3a		135.72	3.8674
3b		16.34	4.7868
3c		64.7	4.1891
3d		24.82	4.6052
3e		111.79	3.9516
4a		65.37	4.1846

4b		25.22	4.5983
4c		62.43	4.2046
4d		15.01	4.8236
4e		108.38	3.9651
3a		67.34	4.1717
3b		16.21	4.7902
3c		30.8	4.5115
3d		15.4	4.8125
3e		64.52	4.1903
3f		30.01	4.5227

3g		15	4.8239
3h		31.15	4.5065
4a		124.87	3.9035
4b		57.49	4.2404
4c		7.18	5.1439
4d		6.65	5.1772
4e		30.02	4.5226
4f		14.03	4.8529
4g		7.01	5.1543

4h		14.52	4.838
1.		12.17	4.9147
2.		0.045	7.3468
3.		6.82	5.1662
4.		0.18	6.7447
5.		11.05	4.9566
6.		1.94	5.7122
7.		2.28	5.6421
8.		4.82	5.317
9.		9.54	5.0205

10.		6.44	5.1911
11.		12.91	4.8891

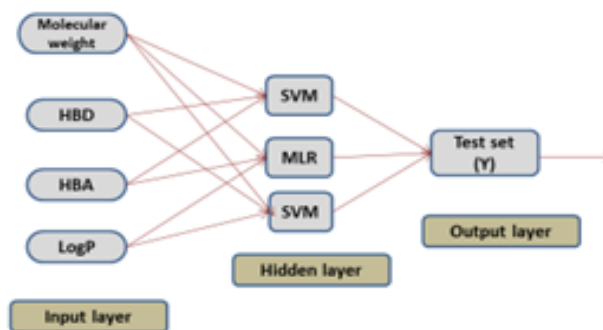
Pre-processing and feature selection

We generated distinct QSAR models with four descriptors and pre-processed the data using python [19]. In the pre-processing step the independent and dependent variables are clearly stated through defining X and Y, respectively. We used four Lipinski's independent variables, which have good correlation with dependent variable and a Pearson correlation matrix is plotted as a heat map. After that the dataset was split into training and test set from *sklearn*. *Preprocessing* module of python program.

Classifier Used

We made use of two machine learning algorithms such as, MLR and SVM to build the models, which could able to predict with reasonable accuracy the effect of substances against mycobacterium tuberculosis. The neural network used in our study presented in Fig. 2

Figure 2. The neural network architecture



A predictive mathematical model is built using selected descriptor(s), all statistical terms associated with the model like Multiple Linear Regression r^2 , adjusted r^2 , F statistics, t value, and p value are calculated. Further model is validated both by internal and external test set, and corresponding predicted r^2 is calculated. Residual and regression plots are saved as image files for quick analysis. Based on summary of the regression equation, one has to select right model for predicting activity of compounds whose activity are unknown.

The support vector machine (SVM) is a supervised machine learning technique for classification and regression tasks,

makes use of a hyperplane separating the data from the variable space into classes that is based on the statistical learning approach [20]. Variables are first mapped in a high-dimensional space through a variety of kernel functions, then the algorithm identifies in this high-dimensional space the maximal margin hyperplane, thus separating the compounds in classes. The built model was saved by using *pickle* module of Python.

Results and Discussion

QSAR investigations focused on machine learning strategies, such as multiple linear regression analysis (MLR) followed by support vector machine (SVM) algorithm [21]. Machine learning technique has been caught the attention of academic community with enormously successful performance in building QSAR models. It can automatically select features directly from raw, high dimension, and heterogeneous chemical data and as well from data curated from the literatures [22].

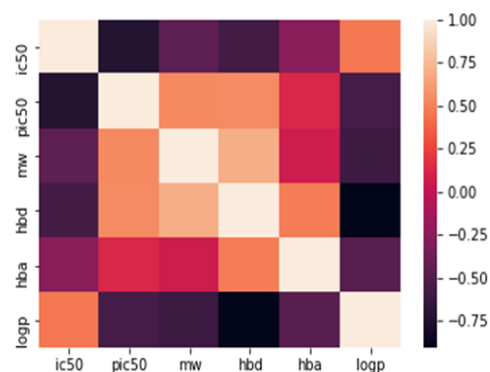
The generated QSAR model afforded the ordinary least squares (OLS) regression as $R^2 = 0.380$, $F=4.909$, and $Q^2 = 0.303$, reg. coef_ developed were of 0.00651593 (molecular weight), -0.0069445 (hydrogen bond acceptor), -0.07576775 (hydrogen bond donor), -0.239021 (Log P), and reg. intercept_ of 3.10331589018553 developed through *stats models .formula* module. The support vector machine of *sklearn* module generated the model score of 0.6294242262068762, the developed model was cross validated via internal and external validation by using the test set compounds and plotted the linear curve between the predicted and actual pMIC₅₀ value. As a consequence, the established QSAR models based on machine learning methods could help us to understand the structural requirements necessary to design new compounds with improved biological activity.

Generated QSAR model

$pMIC_{50} = 0.00651593 (MW) - 0.0069445 (HBA) - 0.07576775 (HBD) - 0.239021 (LogP) + 3.10331589018553$

Lipinski's rule-of-five descriptors comprising of MW, LogP, HBD, and HBA. MW represents the molecular size of a compound that is commonly used because of it can be easily interpreted and calculated as well as appropriate size of a compound is important for its passage via lipid membrane. LogP is a widely used parameter for determining the lipophilicity of a compound and used in calculating the membrane penetration and permeability of compounds. HBD and HBA describe the number of hydrogen bond donors and hydrogen bond acceptors, respectively, which is used to measuring hydrogen bonding capacity. Lipinski's independent variables, which have good correlation with dependent variable and a Pearson correlation matrix is plotted as a heat map followed by dataset split into training and test set from *sklearn.preprocessing* module of python program (Fig.3).

Figure 3. Heatmap shows Pearson correlation matrix of Descriptors



QSAR model for predicting Mycobacterium tuberculosis inhibitory activity

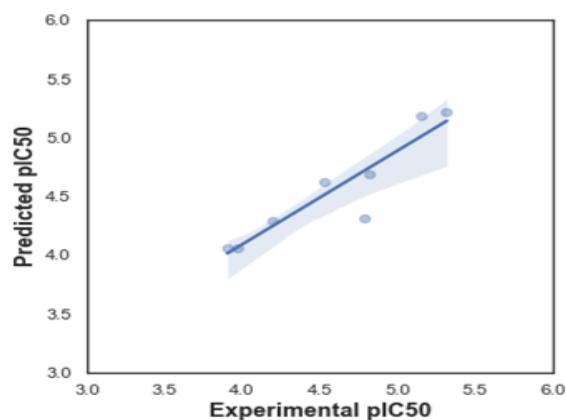
A data set comprising of 37 pyrazoline derivatives were used for construction of QSAR models. Particularly, four sets of fingerprint descriptors were benchmarked in order to find the best performing set. Prior to modelling, data preprocessing was carried out to clean the missing data if any and feature selection was applied to reduce dimensionality between the descriptors. However, in our case we have not found any of these issues. Each of the two models were then built using a data split ratio of 80:20 in which 80% of the data set was used as the internal set (training set) and 20% as the external set (test set). The performance results given in Fig.4. As shown in Fig. 5, it can also be seen that scatter plots of experimental versus predicted pIC₅₀ of panels displayed narrower variance of the data points

Figure 4. Predicted value of test set

```
In [192]: df1['pic50'] = df.pic50
          df1
```

	Y_pred	pic50
0	4.892299	3.86736
1	4.606332	4.78675
2	4.447771	4.18910
3	4.661150	4.60520
4	4.642604	3.95160
5	4.805830	4.18462
6	5.066220	4.59825
7	4.724526	4.20461

Figure 5. Linear regression curve



CONCLUSION

A QSAR study through machine learning algorithm was carried out to know the residual difference between observed and predicted anti-tubercular potency of the 37 selected molecules of pyrazoline derivatives. It was observed that Lipinski's fingerprint afforded good performance for the constructed models indicating that they could capture the feature space of *Mycobacterium tuberculosis* inhibitors. The predicted pMIC₅₀ values of the compounds have an acceptable correlation with the experimental values from the generated multiple linear regression and SVM algorithms. In conclusion, the QSAR model generated from the present study should be useful for designing a similar group with hopeful anti-tubercular agents. It is anticipated that the knowledge gained from this study could be used as general guidelines for the design of novel anti-tubercular agents.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENT

The authors acknowledge our institution, Nandha College of Pharmacy, Erode, for providing the necessary facilities to carry out our research work

REFERENCE

1. T Prabha, P Aishwaryah, E Manickavalli, R Chandru, G Arulbharathi, A Anu, T Sivakumar 2019. A Chalcone Annulated Pyrazoline Conjugates as a Potent Antimycobacterial Agents: Synthesis and in Silico Molecular Modeling Studies. Research J. Pharm. Tech., 12(8), 3857-3865.
2. T Prabha, T Sivakumar 2018. Design, Synthesis, and Docking of Sulfadiazine Schiff Base Scaffold for their Potential Claim as INHA Enoyl-(Acyl-Carrier-Protein) Reductase Inhibitors. Asian J Pharm Clin Res. 11(10), 233-237.
3. M T Chhabria, B M Mahajan, P S Brahmshatriya 2001. QSAR Study of a Series of Acyl Coenzyme A (CoA): Cholesterol Acyltransferase Inhibitors Using Genetic Function Approximation. Med. Chem. Res., 20, 1573-1580.
4. T Katsila, G A Spyroulias, G P Patrinos, M T Matsoukas, 2016. Computational approaches in target identification and drug discovery, Computational and Structural Biotechnology Journal, 14, 177-184.
5. Cherkasov, 2014. QSAR Modeling: Where Have You Been? Where Are You Going To, J. Med. Chem., 57, 4977-5010.
6. S Kini, A M Gandhi 2008. Novel 2 pyrazoline derivatives as potential antibacterial and antifungal agents. Indian J Pharm Sci, 70:1058.
7. O I ElSabbagh, M Baraka, S M Ibrahim, C Pannecouque, G Andrei, R Snoeck, 2009. Synthesis and antiviral activity of new pyrazole and thiazole derivatives. Eur J Med Chem. 44, 3746-53.
8. Khunt, R.C., Khedkar, V.M., Chawda, R.S., Chauhan, N.A., Parikh, A.R., Coutinho, E.C. 2012. Synthesis, antitubercular evaluation and 3D-QSAR study of N-phenyl-3-(4-fluorophenyl)-4-substituted pyrazole derivatives. Bioorg Med Chem Lett 22, 666-78.

9. F Hayat, A Salahuddin, S Umar, A Azam 2010. Synthesis, characterization, antiamebic activity and cytotoxicity of novel series of pyrazoline derivatives bearing quinoline tail. Eur J Med Chem. 45, 4669-75.
10. C V Li, Li QS, Yan L, Sun XG, Wei R, Gong, HB 2012. Synthesis, biological evaluation and 3D-QSAR studies of novel 4,5-dihydro-1H-pyrazole niacinamide derivatives as BRAF inhibitors. Bioorg Med Chem. 20, 3746-55.
11. Shoman ME, Abdel-Aziz M, Aly OM, Farag H, Morsy MA 2009. Synthesis and investigation of anti-inflammatory activity and gastric ulcerogenicity of novel nitric oxide-donating pyrazoline derivatives. Eur J Med Chem. 44, 3068-76.
12. Khode S, Maddi V, Aragade P, Palkar M, Ronad PK, Mamledesai S, 2009. Synthesis and pharmacological evaluation of a novel series of 5-(substituted) aryl-3-(3-coumarinyl)-1-phenyl-2-pyrazolines as novel anti-inflammatory and analgesic agents. Eur J Med Chem. 44, 1682-1688.
13. Rajendra P.Y., Lakshmana R.A., Prasoona L., Murali K., Ravi Kumar P. 2005. Synthesis and antidepressant activity of some 1,3,5-triphenyl-2-pyrazolines and 3-(2''-hydroxy naphthalen-1''-yl)-1,5-diphenyl-2-pyrazolines. Bioorg Med Chem Lett. 15, 5030-5034.
14. Ozdemir Z., Kandilci H.B., Gümüsel B., Caliş U., Bilgin A.A. 2007. Synthesis and studies on antidepressant and anticonvulsant activities of some 3-(2-furyl)-pyrazoline derivatives. Eur J Med Chem. 42, 373-379.
15. P Fabian, V Gaël, G Alexandre, M Vincent, T Bertrand, G Olivier, B Mathieu, Peter Ron W, D Vincent, V Jake, P Alexandre, C David, B Matthieu, P Matthieu, D Édouard 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825-2830.
16. S N Shelke, G R Mhaske, V D Bonifácio, 2012. Green synthesis and anti-infective activities of fluorinated pyrazoline derivatives. Bioorg Med Chem Lett. 22(17), 5727-5730.
17. A Aftab, H Asif, A K Shah, Mujeeb Mohd, B Anil 2016. Synthesis, antimicrobial and antitubercular activities of some novel pyrazoline derivatives. Journal of Saudi Chemical Society. 20(5), 577-584.
18. M A Ali, M S Yar, M Kumar, G S Pandian 2007. Synthesis and antitubercular activity of substituted novel pyrazoline derivatives. Nat Prod Res. 21(7), 575-579.
19. S Kim, K H Cho 2019. PyQSAR: A Fast QSAR Modeling Platform Using Machine Learning and Jupyter Notebook. Bull. Korean Chem. Soc. 40, 39-44.
20. C Cortes, V Vapnik, 2009. Support-vector networks. Chem. Biol. Drug Des. 297, 273-297.
21. T He, H Mao, J Guo, Z Yi 2016. Cell Tracking Using Deep Neural Networks with Multi-task Learning. Image and Vision Computing. 60, 12-14.

How to cite this article

T Prabha, C Selvinthanuja, S Hemalatha, S Sengottuvelu, J Senthil, 2021. Machine learning algorithm used to build a QSAR model for pyrazoline scaffold as anti-tubercular agent. Jour. of Med. P'ceutical & Allied. Sci. V 10 - I 6, 2562, P-4024 - 4029. doi: 10.22270/jmpas.V10I6.2562