

INVESTIGATION OF THE ACTIVITY OF 8-METHYLQUINOLONES AGAINST MYCOBACTERIUM TUBERCULOSIS USING THEORETICAL MOLECULAR DESCRIPTORS: A CASE STUDY

Gowal M. Eric

Adamu Uzairu

Paul A.P Mamza

Department of Chemistry, Ahmadu Bello University, Zaria, Nigeria

Abstract

A quantitative structure-activity relationship (QSAR) study on a set of 36 structurally-similar 8-methylquinolones was performed using a large pool of theoretical molecular descriptors. The molecular structures of the compounds were pre-optimized using molecular mechanics (MM2). Full optimization was done with the density functional theory (DFT) using Becke's three-parameter hybrid functional with LYP correlation functional in combination with the standard Pople's basis set 6311G*. HOMO and LUMO energies, dipole moment, total energy and many other properties served as quantum-chemical descriptors. The GA-MLRA technique was used to select the most significant descriptors and to generate a linear model for predicting the biological activity, Minimal Inhibitory Concentration (MIC), treated as negative decade logarithm, ($pMIC$). The best model was obtained with $R^2=0.90323$. The model was tested internally using the leave-one-out (LOO) cross validation procedure on the training set and validated against the external validation set ($Q^2_{LOO} = 0.83115$ and $R^2_{Pred} = 0.78708$). The Y-scrambling/randomization validation also confirmed the statistical significance of the model. Leverage approach was used to define the applicability domain of the model. This validated model could be used to design new potential drug candidates, within the 8-methylquinolone family, with high activity against tuberculosis.

Keywords: Tuberculosis, Mycobacterium Tuberculosis, Molecular descriptors, Quinolones, Genetic algorithm, QSAR

Introduction

Background of the Study

Tuberculosis (TB) is a disease of antiquity which is thought to have evolved sometime between the seventh and sixth millennia BC (Manchester, 1984). This disease has been recognized for thousands of years and the etiological agent has been identified since earliest days of medical biology, however, the global burden of TB has continue to loom as one of the largest among infectious diseases, with an enormous toll in morbidity and mortality. (Leonard, 2009). This disease which is caused by various strains of mycobacterium, usually *mycobacterium tuberculosis* (Luo 2012) has been reported to be the leading cause of death from a single infectious agent, after the human immuno deficiency virus (HIV). (WHO, 2014). Currently one-third of the world is believe to harbor the latent form of mycobacterium tuberculosis, with a lifelong risk of activation and disease development, particularly in people co-infected with HIV. (Anil, 2011) The World Health Organization (2014) reported 1.5 million deaths from tuberculosis globally in 2013. Ironically, available therapeutic agents are highly efficacious in TB, with cure exceeding 95% during clinical trials (Frieden, 2003). The natural question then is; why has the control of TB been so problematic? The answer lies in the indolent clinical nature of the disease, which needs prolong complex therapy and the unusual microbiological properties of the pathogen. (Leonard,2009).

The treatment of TB is quite long ,taking approximately 6-9months (Blumberg *et al.*, 2003). The prolonged duration of the treatment as well as the toxicity and the poor patience compliance are risk factors which frequently lead to selection of drug resistant and very often deadly multi-drug resistant strains. This increasing problem of multi-drug resistant strains is the major challenge for the investigation and designs of novel drug candidates which are not only active against stable drug resistant m.tuberculosis but also shorten the length of therapy.

Quinolones

In the search for new anti- tubercular agents, quinolones are particularly interesting. This is because they potentially offer many of the attributes of an ideal antibiotic, combining high potency, a broad spectrum of activity, good bioavailability, oral and intravenous formulations, high serum levels, a large volume of distribution indicating concentration in tissues and a potentially low incidence of side-effects.(Emami , Shafiee and Foroumadi ,2005)

The first quinolone, nalidixic acid, (1-ethyl-7-methyl-1,4-dihydro-4-oxo-1,8- naphthyridine-3-carboxylic acid , was isolated as a by-product of the synthesis of chloroquine by George Leshner and his coworkers in 1962,it

was found to be effective in the treatment of urinary tract infections (UTIs) (Lescher et al, 1962; Norris S. and Mandell,1988).

Quinolones act by blocking bacterial DNA synthesis through inhibition of bacterial topoisomerase-II (DNA gyrase) and topoisomerase IV. Inhibition of DNA gyrase prevents the relaxation of positively super coiled DNA that is required for normal transcription and replication. Inhibition of topoisomerase-IV interferes with separation of replicated chromosomal DNA into the respective daughter cells during cell division (Alangaden and Lerner ,1997, Flamm et al ,1995). Quinolone agents basically possess a bicyclic aromatic core; this can contain a carbon at the 8-position, yielding a true quinolone, or a nitrogen, which provides a ring system technically termed a naphthyridone (Figure 1).

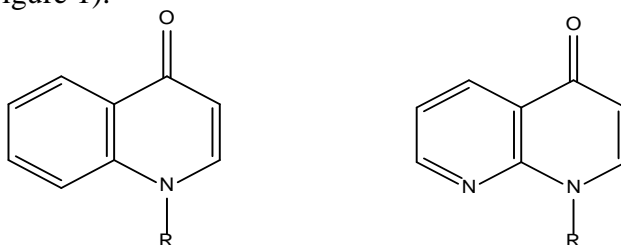


Fig 1. Quinolone and Naphthyridone nuclei; general structural features required for antibacterial activity

It is common practice, however, for both quinolone and naphthyridone structures to be classified as “quinolone antibacterial agents.” The pyridone ring on the right-hand side is an essential necessity for Antibacterial activity. (Bright and Gootzy, 2000)

Experimental procedure

Data set

The data set contains 36 derivatives of 1-(cyclopropyl/2,4-difluorophenyl/tert-butyl)-1,4-dihydro-8-methyl-6-nitro-4-oxo-7-(substituted secondary amino)quinoline-3-carboxylic acids with different substitution on position 2 and 5. The in vitro antimycobacterial activity (MIC) of the molecules against Mycobacterium Tuberculosis strain H37Rv determined by agar dilution method for the determination of MIC in duplicates (NCCLS, 1995), were taken from the work of Senthilkumar *et al.*, (2009). Table 1 lists the chemical structure and the anti-tuberculosis activity (pMIC) of the molecules used in this study.

The inhibitory data (MIC in μM) was converted to negative logarithmic dose in mole (pMIC) because a QSAR is a linear free energy relationship, and from the van't Hoff isotherm, free energy change during a process is proportional to the logarithm of the rate or equilibrium constant of the process ($\Delta G = 2.303 RT \log K$). (Gupta *et al.*, 2007)

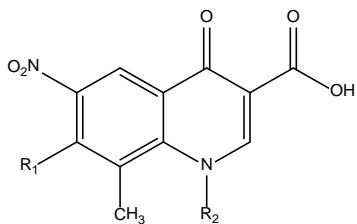
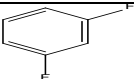
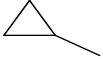
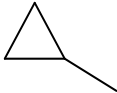
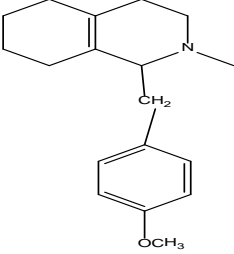
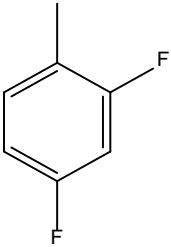
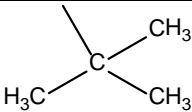
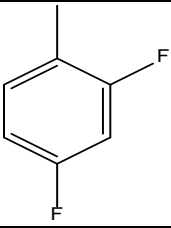
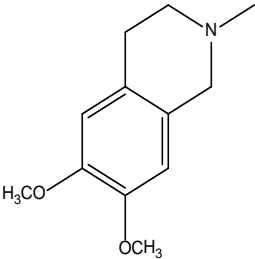
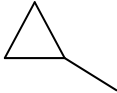


Table 1

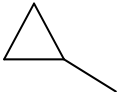
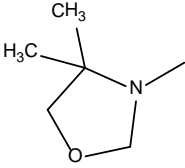
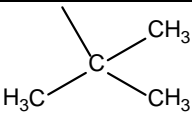
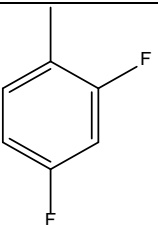
ID	R ₂	R ₁	pMIC
1a			5.20
2a			4.61
1b			5.54
1c			4.59
1d			4.85
2d			5.74

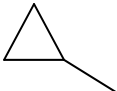
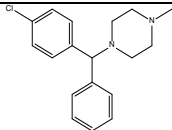
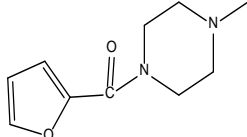
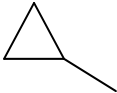
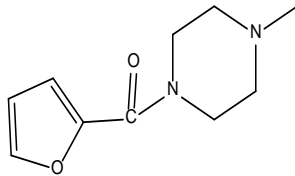
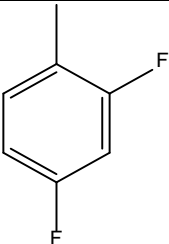
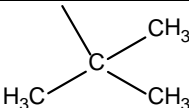
... continued Table 1

ID	R ₂	R ₁	pMIC
1e			4.63
2e			5.82
3e			4.07
1f			4.57

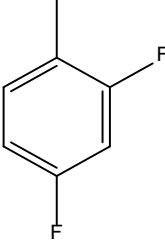
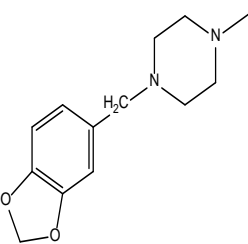
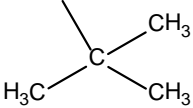
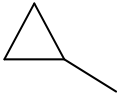
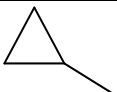
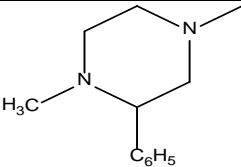
2f			5.22
3f			6.38
1g			5.53
2g			4.69
1h			4.60
2h			4.94
3h			6.41

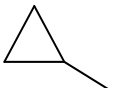
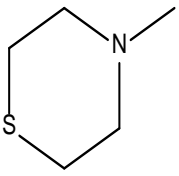
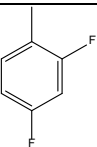
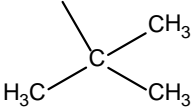
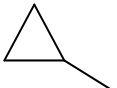
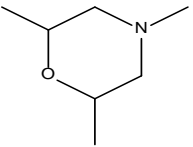
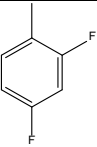
... continued Table 1

ID	R ₂	R ₁	pMIC
1i			5.70
2i			5.16
3i			4.81

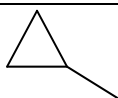
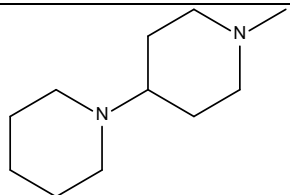
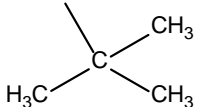
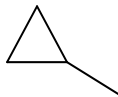
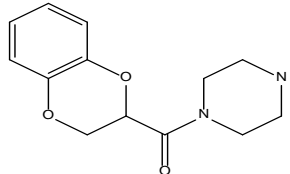
1j		 	6.17
1k			6.08
2k			4.39
3k			4.59

... continued Table 1

ID	R ₂	R ₁	pMIC
1l			4.67
2l			4.62
3l			6.11
1m			6.08

1n			5.70
2n			4.87
3n			4.51
1o			6.01
2o			5.17

... continued Table 1

ID	R ₂	R ₁	pMIC
1p			5.16
2p			4.58
1q			5.53

*These are compounds in the external Validation Set

Geometry optimization and calculation of quantum-chemical descriptors

The molecular structures previously collected were in 2D format (ChemBioOffice 2D sketch (*.mol)). Using the SPARTAN “14 v 1.1.0 software package (2013) each structure was subsequently resketched in a 3D environment, checked by visual inspection in order to ensure that the 3D geometry is correct, and saved as SAPRTAN input file format (*.spartan) for geometry optimization. The molecular geometry of all model structures were subjected to energy minimization using molecular mechanics (MM2) and then re-optimized using the density functional theory (DFT) with Becke’s

three-parameter hybrid functional (Becke,1988) using LYP correlation functional (Lee , Yang and Parr 1988) . The standard Pople's 6-311G* basis set was used. All calculations were carried out using the SPARTAN 14" v 1.1.0 (2013) suite of programs. Several descriptors evaluated by quantum- chemical calculations as described above, were used in this study. These descriptors include the Highest occupied molecular orbital, HOMO; Lowest unoccupied molecular orbital, LUMO; Energy gap ,Total energy, total dipole moment, polarizability, ovality Ionization energy (IP) , electron affinity(EA), Global hardness (η) ,Chemical potential (μ) ,global electrophilicity index (ω) and total softness S

Calculation of 0D,1D,2D and 3D descriptors

The 0D, 1D,2D, 3D molecular descriptors for each compound of the training/validation were calculated using PADEL (Yap, 2001) software package. The final pool of the calculated 0D, 1D, 2D and 3D theoretical molecular descriptors, 1537 in total can be separated into sixty-one classes:

Acidic group count, ALOGP, APol , Aromatic atoms count, Aromatic bonds count, Atom count, Autocorrelation. Barysz matrix, Basic group count, BCUT, Bond count, BPol ,Burden modified eigenvalues, Carbontypes, Chi chain, Chi cluster, Chi path cluster, Chi path, Constitutional, Crippen logP and MR, Detour matrix, Eccentric connectivity index, Atom type electrotopological state, Extended topochemical atom, FMFDescriptor, Fragment complexity, Hbond acceptor count, Hbond donor count, Hybridization ratio, Information content, Kappa shape indices, Largest chain, Largest Pi system, Longest aliphatic chain, MannholdLogP, McGowan volume, Molecular distance edge, Molecular linear free energy relation, Path counts, Petitjean number, Ringcount, Rotatable bonds count, Rule of five, Topological, Topological charge,Topological distance matrix, Topological polar surface area, Van der Waals volume, Vertex adjacency information (magnitude),Walk counts, Weight, Weighted path, Wiener numbers, XLogP, Zagreb index,3D autocorrelation, Charged partial surface area, Gravitational index, Moment of inertia, Petitjean shape index, RDF.

The preprocessing of the independent variables (i.e., descriptors) was done by removing invariable (constant column) and cross-correlated descriptors (with $R^2 = 0.75$), this left a total of 316 descriptors which were then individually normalized to the [0, 1] range, and used for QSAR analysis.

Development of a Quantitative Structure- Activity Relationships Model

Initially the dataset was divided into training and test sets using the random division approach. This division procedure resulted in 28 compounds in the training set and 8 compounds in the test set (see Table 1). The

selection of significant descriptors, which constructs a relationship between the biological activity of the data and its molecular structures, is an important step in QSAR modeling. For this purpose the variable selection genetic algorithm (GA) and Multiple Linear Regression Analysis (MLRA) methods were used.

The GA has been applied widely in many areas, and specifically, in recent studies as a powerful tool to address many problems in cheminformatics. GA is an optimization algorithm based on the evolutionary mechanisms as proposed by Darwin, it uses random mutation, crossover and selection procedures to produce better models or solutions from an originally random starting population or sample. (Davis, 1991; Devillers, 1996)

The MLR, also known as linear free-energy relationship method is an extension of simple linear regression analysis to more than one dimension (Berk, 2003). MLR generates QSAR equations by performing standard multivariable regression calculations to identify the dependence of a drug property on any or all of the descriptors under investigation. (Thangapandian *et al.*, 2012)

The combination of the GA-MLRA technique was employed for preliminary model selection, as implemented in the BuildQSAR (De Oliveira and Gaudio, 2001) software.

The GA began with a population of 100 random models, 1000 iterations to evolution, and the mutation probability of 35%. Initially several models which contain between 1–5 variables were selected. A final set of three models consisting of two 5-parameter and one 4-parameter models was selected for further analysis, based on their high squared correlation coefficient R^2 , low standard error s , high Fisher coefficient F . A detailed statistical analysis of the selected models was carried out using Material Studio Software 7.0

Model Validation

Internal Validation - The three models were internally validated using the cross-validation leave-one-out procedure. In the leave-one-out (LOO) method of cross validation, the process of removing a molecule, and creating and validating the model against the individual molecules is performed for the entire training set. Once complete, the mean is taken of all the Q^2 values and reported (Ravichandran *et al.*, 2011)

The Q^2_{LOO} was calculated using the equation (10) (Consonni, Ballabio, and Todeschini, 2009)

$$Q^2_{tr} = 1 - \frac{\sum_{i=1}^{N_{tr}} (y_{i,exp} - y_{i,pred})^2}{\sum_{i=1}^{N_{tr}} (y_{i,exp} - \bar{y}_{i,exp})^2} \dots\dots\dots (10)$$

Where N_{tr} is the total number of training set objects; $y_{i,exp}$ and $y_{i,pred}$ are the experimental and predicted values, respectively; $\bar{y}_{i,exp}$ is the average response value of the training set.

External Validation

The models were externally validated by testing the previously excluded compounds which form the test set. The value of R^2_{Pred} which gives an indication of the predictive power of a model was calculated using equation (11) (Schüürmann *et al.*, 2008)

$$R^2_{Pred} = 1 - \frac{\sum_{i=1}^{N_{ext}} (y_{i,exp} - y_{i,pred})^2}{\sum_{i=1}^{N_{ext}} (y_{i,exp} - \bar{y}_{i,pred})^2} \dots\dots\dots (11)$$

where N_{ext} is the total number of external validation set objects; $y_{i,exp}$ and $y_{i,pred}$ are the experimental and predicted values, respectively; $\bar{y}_{i,pred}$ is the average response value of the external validation set.

The selection of robust and well predictive QSAR models on the basis of only R^2 , Q^2 and R^2_{pred} might mislead the search for the ideal predictive model, so additional statistical parameters have been suggested by Tropsha (2010) ; a model that must be regarded as truly predictive must satisfy the following conditions

1. $Q^2 > 0.5$
2. $R^2_{pred} > 0.6$
3. $|R_0^2 - R_0'^2| < 0.3$
4. $\left[\frac{|R_0^2 - R_0'^2|}{R^2} < 0.1 \right]$ and $0.85 < k < 1.15$ or $[(R^2 - R_0'^2)/R^2] < 0.1$ and $0.85 < k' < 1.15$

Where R is the correlation coefficient between the predicted and observed activities; R_0^2 is coefficients of determination (predicted versus observed activities, and observed versus predicted activities $R_0'^2$ for regressions through the origin); (iii) slopes k and k' of regression lines through the origin. These additional statistical parameters were calculated for the selected models and the result summarized in table (4). The model with the highest R^2_{Pred} value was chosen as the final model.

Robustness of QSAR Models

The final Model was validated by the Y-randomization (randomization of response). This is approach widely use to validate the robustness of QSAR models (Wold and Ericksson, 1995). It consists of rebuilding models using randomized activities of the training set and subsequent assessment of the model statistics. It is expected that models obtained for the training set with randomized activities should have significantly lower values of pMIC for the training set than the models built

using training set with real activities. If this condition is not satisfied, models built for this training set with real activities are not reliable and should be discarded.

In addition, the parameter R_p^2 (Roy ,2007) which penalize the model R^2 for the difference between squared mean correlation coefficient (R_r^2) of the randomized models and squared correlation coefficient (R^2) of the non-randomized model given (Roy ,2007) by eq (12) was also calculated for the model.

$$R_p^2 = R^2 * \left(1 - \sqrt{|R^2 - R_r^2|} \right) \dots\dots\dots (12)$$

A value of R_p^2 greater than 0.5 may be taken as an indicator of model acceptability

Applicability domain

The applicability domain of a (Q)SAR model is a theoretical region in chemical space, defined by the model descriptors and modeled response, and thus by the nature of the chemicals in the training set, as represented in each model by specific molecular descriptors. Thus only the predictions for chemicals falling within this domain can be considered reliable and not model extrapolations (Gramatica, 2007). Without the restriction placed by the applicability domain, QSAR models can predict the activity of any compound even if such a compound is structurally different from those included in the training set. This would lead to unreasonable extrapolation of the model in chemistry space and therefore heighten the chances of inaccurate predictions. Thus for a QSAR model to give reliable outcome, its applicability domain must de be define.

In this study, the applicability domain of the final model was calculated by the leverage approach (Eriksson et al, 2003),The leverage value of every compound was calculated and plotted against the standardized residuals (William plot). This method offers a graphical assessment of the leverage values (h_{ii}) as a function of the standardized cross validated residuals and it is suitable not only for detection of the structurally-influential outliers, but also for determination of the response outliers (Gramatica, 2007). The leverage is defined as a compound's distance from the centroid of X. Mathematically, the leverage (h_{ii}) of a given compound in the multidimensional descriptor space, can be calculated as (Eq. 13)

$$h_{ii} = x_i^T (X^T X)^{-1} x_i \dots\dots\dots (13)$$

Where x_i the descriptor row matrix of the compound under consideration and X is the multidimensional matrix carrying the structural information (calculated molecular descriptors) for each training set compound.

The model predictions should be referred as unreliable for those compounds for which h_{ii} diagonal elements are greater than the cut-off leverage value (h^*). These compounds are located far from the structural centroid of the model, and therefore could be referred as structurally-influential outliers. The cut-off leverage value (h^*) is usually defined by

Eq. (14)

$$h^* = \frac{3(P+1)}{n} \dots\dots\dots(14)$$

where p is the total number of descriptors used for developing of the QSAR model, while n is the total number of the training set compounds. Moreover, the compounds for which the calculated standardized residual values are greater than three standard deviation units ($>\pm 3\sigma$) could be considered as response outliers (Gramatica, 2007).

Result and Discussion

This study was focused on developing a valid model that capable of predicting the activity of derivatives of 1-(cyclopropyl/2,4-difluorophenyl/tert-butyl)-1,4-dihydro-8-methyl-6-nitro-4-oxo-7-(substituted secondary amino)quinoline-3-carboxylic acids which could subsequently be used to design new potential drug candidates, with possible better activity. For this purpose a QSAR approach was utilized. One of the most important steps in QSAR modeling is to define the number of independent variables in the model equation obtained. In this way the over- parameterization of the mathematical model as well as the chance correlation between the molecular descriptors is avoided (Topliss and Edwards, 1979). Since the modeling procedure started with a pool of several theoretical molecular descriptors, a possibility exists to encounter a chance correlation in a case where the number of examined variables is higher than the number of observations. The GA was applied to select from the pool of calculated descriptors, only the best combinations of those most relevant for obtaining models with the highest predictive power for activity. Eventually the GA-MLRA to build different QSAR models.

Initially several 1-5 parameter models were generated, a set of three models, consisting of two 5-parameter models and one 4-parameter model was selected for further analysis based on their high squared correlation coefficient R^2 , low standard error s and high Fisher coefficient F . The three models were validated internally using the Leave One Out technique (LOO) where their various Q^2_{LOO} were calculated. Also the models were externally validated by testing their predictive performances using the external validation set that were not used for model development. See table 2 for the statistical parameters as well as the internal and external validation

parameters, (Q^2_{LOO}), (R^2_{Pred}) respectively of the three selected models. Table 3 compares the Observed versus Predicted activities of the selected models.

Selected Models

pMIC = 5.19(+/-0.05) -0.37(+/-0.06) MATS4c -0.23(+/-0.06) VE3_Dzp +0.63(+/-0.06) SCH-3 -0.22(+/-0.05) PetitjeanNumber -0.15(+/-0.04) RDF95u.....(1)

pMIC = 5.20(+/-0.04) -0.23(+/-0.06) AATSC5p -0.43(+/-0.05) MATS4c -0.35(+/-0.06)VE3_Dzp +0.62(+/-0.05)SCH-3 -0.26(+/-0.05) PetitjeanNumber.....(2)

pMIC = 5.26(+/-0.05) -0.30(+/-0.06) MATS4c +0.69(+/-0.07) SCH-3 +0.26(+/-0.07) ZMIC3 -0.30(+/-0.06) RDF10u.....(3)

Table 2 Fitting Parameters For The Selected models

S/N	R ²	R ² _{adj}	F	s	Q ² _{LOO}	R ² _{Pred}	SPress	SDEP	n	k
1.	0.88	0.85	31.62	0.26	0.82	0.77	2.17	0.28	28	5
2	0.90	0.88	41.07	0.23	0.83	0.78	2.03	0.27	28	5
3	0.85	0.82	32.38	0.28	0.78	0.57	2.66	0.31	28	4

Since the selection of a robust and well predictive QSA R model on the basis of only R^2 , Q^2 and R^2_{pred} might be misleading, the selected models were subjected to further statistical analysis as suggested by Golbraikh and Tropsha, the result is summarized in table 4. From table 4 it can be seen that models 1 and model 2 both fulfilled all the Golbraikh and Tropsha criteria for a truly predictive model, but model 3 failed, as its R^2_{pred} of 0.57 is lower than the accepted value of 0.6. Finally model (2) was selected as the final model since it performed best at predicting the activities of the external validation set as indicated by the value of R^2_{pred} .

Table 4 Golbraikh and Tropsha acceptable model criteria For The Selected Models

MODEL	Q ²	R ² _{Pred}	R _o ² - R' _o ²	[(R ² -R _o ²)/R ²]	[(R ² -R _o ²)/R ²]	k	k'
1	0.82	0.77	0.08	0.00	0.10	0.99	1.01
2	0.83	0.78	0.01	0.01	0.01	0.98	1.02
3	0.78	0.57	0.02	0.09	0.12	0.96	1.04

The mathematical equation of model (1) is as follows:

pMIC = 5.20(+/-0.04) -0.25(+/-0.06) AATSC5p -0.43(+/-0.05) MATS4c -0.35(+/-0.06) VE3_Dzp +0.62(+/-0.05) SCH3 -0.26(+/-0.05) PetitjeanNumber.....(2)

R² = 0.90323 R²_{adj} = 0.88123 F = 41.0676 s = 0.2301 Q²_{LOO} = 0.83115 R²_{Pred} = 0.78708 SPress = 2.03242 SDEP = 0.26942 n = 28

The high correlation coefficient R (0.91) indicates the susceptibility of descriptors (AATSC5p, MATS4c, VE3_Dzp, SCH-3 and SCH-3) to form the above model (2). Squared correlation coefficient (R^2) of 0.903 explains 90% variance in biological activity of the tested compounds. It also indicates the statistical significance >99.9% with F-values (41.0676). Cross-validated square correlation coefficient (Q^2) by LOO technique was 0.83115 which showed a good internal predictive ability of the model. The model was also validated by applying the Y-randomization test. Several random shuffles of the Y vector were performed and the obtained results are in good agreement with the suggested limits (Eriksson et al., 2003).

The low R^2 and Q^2_{LOO} values of the random models shown in Table 5 and the value of $R^2_p = 0.808101$ ($R^2_p \geq 0.5$) indicates that there is no chance of correlation or structural dependency in the proposed model. Consequently model (2) can be considered as a perfect model with both high statistical significant and excellent predictive ability.

Table 5 Y-Randomization Result

Model	R	R^2	Q^2
Original	0.95	0.90	0.83
Random 1	0.41	0.17	-0.42
Random 2	0.53	0.28	-0.35
Random 3	0.48	0.23	-0.32
Random 4	0.46	0.21	-0.35
Random 5	0.51	0.26	-0.67
Random 6	0.32	0.10	-1.03
Random 7	0.24	0.04	-1.25
Random 8	0.34	0.11	-0.54
Random 9	0.64	0.41	-0.08
Random 10	0.30	0.09	-0.94

Random Models Parameters

Average R : 0.42

Average R^2 : 0.19 Average Q^2 : -0.59 $R^2_p = 0.81$

It was observed that, AATSC5p, MATS4c , VE3_Dzp, SCH-3 and PetitjeanNumber are the best descriptors in the establishment of the QSAR model for the class of compounds in this study. Table 6 shows the experimental and calculated activities of the training set while Fig 4 shows the corresponding plot of the experimental activities with the calculated ones.

Table 7 also shows the experimental and calculated activities of the external validation set and figure 5 captures the graphic correlation between them.

The applicability domain (AD) of the five-descriptor linear model previously selected (Williams plot) was assessed utilizing the well known leverage approach (Fig. 3). Training set objects (28 compounds with experimental activity values) used in the model development are presented as solid dots, whereas the external validation set objects (9 compounds) as solid rectangles labeled with the corresponding number (ID signature). The analysis of AD for the training set objects shows that only one compound labeled with (ID20) signature can be identified as a typical

X-outlier ($h > h^* = 0.643$).

However the pMIC prediction of this compound was fairly good as can be seen in Table 6

None of the external validation set is a typical X-outliers with $h > h^*$. Also no compound within the training as well as the external validation set is a typical Y-outlier since the cut off value for standard deviation is $\pm 3.0\sigma$

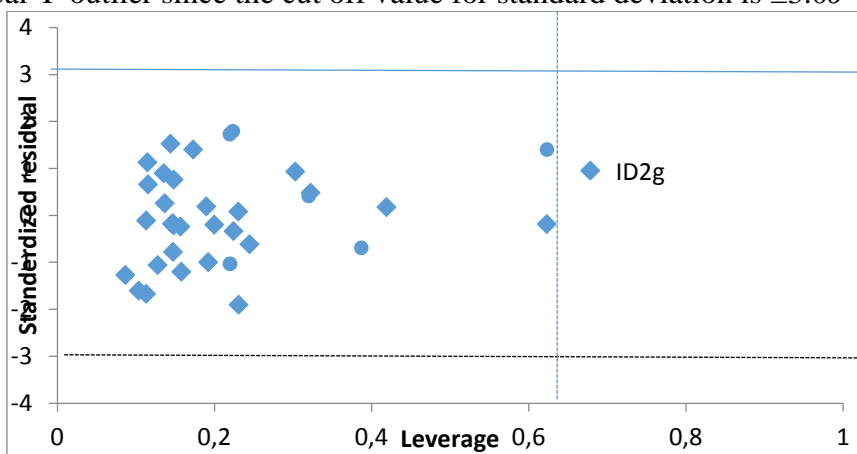


Fig 3 Graphical representation (Williams plot) of the five- descriptor MLR model's applicability domain (AD) together with external validation set object

Table6 Observed Activity versus Predicted Activity OF Model 2

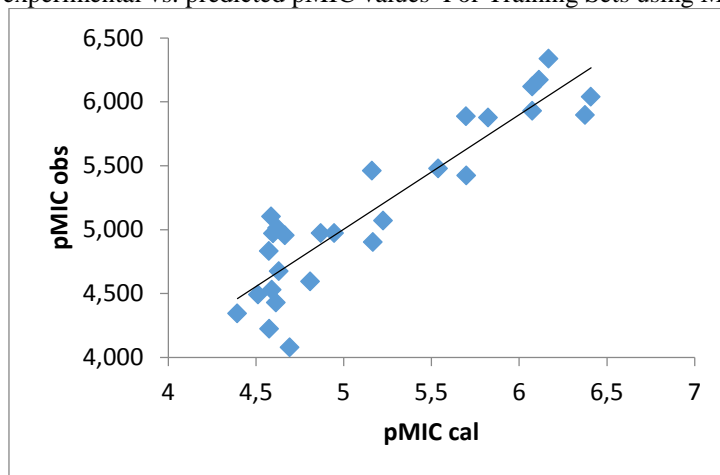
COMP. NO	pMIC OBSERVED	pMIC CALCULATED	RESIDUAL	STANDARD DEV.
3	4.61	4.46	0.16	0.69
4	5.54	5.49	0.05	0.23
7	4.59	4.56	0.04	0.15
14	4.63	4.67	-0.04	-0.16
15	5.82	5.87	-0.04	-0.19
16	4.58	4.78	-0.21	-0.90
17	5.23	5.09	0.14	0.59
18	6.38	6.00	0.37	1.62
20	4.69	4.50	0.20	0.86
22	4.60	4.93	-0.33	-1.45
23	4.95	4.97	-0.02	-0.10
24	6.41	6.09	0.32	1.37

25	5.70	5.86	-0.16	-0.71
26	5.17	4.93	0.23	1.02
27	4.81	4.62	0.19	0.81
28	6.17	6.30	-0.13	-0.56
31	6.08	6.11	-0.04	-0.16
32	4.39	4.36	0.04	0.17
33	4.59	4.98	-0.30	-1.72
34	4.67	4.93	-0.26	-1.15
35	4.62	4.97	-0.35	-1.51
36	6.11	6.16	-0.05	-0.22
39	6.08	5.98	0.10	0.43
43	5.70	5.51	0.19	0.84
44	4.87	4.91	-0.04	-0.17
45	4.51	4.50	0.02	0.07
49	5.16	5.41	-0.25	-1.09
51	4.58	4.29	0.29	1.26

Table 6 Observed Activity versus Predicted Activity For Test Sets

Name	Y_{obs}	Y_{pred}	$(Residual)^2$	$(Y_{obs} - Y_{bar})^2$
1	5.20	5.60	0.16	0.00
11	4.85	5.24	0.15	0.11
12	5.74	5.38	0.13	0.31
19	5.53	5.52	0.00	0.12
47	5.18	4.91	0.07	0.00
48	6.01	6.30	0.09	0.68
53	5.53	5.70	0.03	0.12
52	4.07	4.27	0.04	1.24

Fig 4. The experimental vs. predicted pMIC values For Training Sets using MLR method.



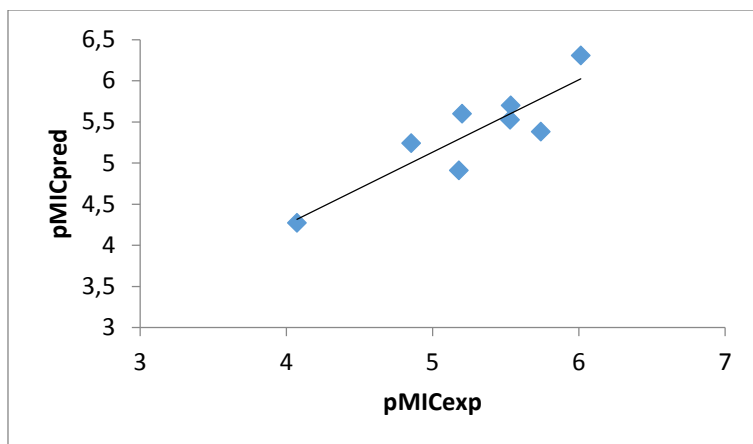


Fig 5 The experimental vs. predicted pMIC values For Test Sets using MLR method

The correlation matrix for the descriptors present in all three models and pMIC used in the present study is shown in Table 8. The sign of the correlation tells us whether the two variables are positively (more X means more Y) or negatively (more X means less Y) related. pMIC and SCH-3 had a good positive correlation ($r=0.7533$) and strongly associated with . However, the correlations for this descriptors considered as a single descriptor in the model was not sufficient to be considered significant in predicting anti-tuberculosis activity

Table 8 Correlation matrix among the descriptors

	<i>pMIC</i>	<i>AATSC</i> <i>5p</i>	<i>MATS</i> <i>4c</i>	<i>VE3_D</i> <i>zp</i>	<i>SCH-3</i>	<i>Petitjea</i> <i>n</i> <i>Number</i>	<i>RDF9</i> <i>5u</i>	<i>RDF1</i> <i>0u</i>	<i>ZMI</i> <i>C3</i>
pMIC	1								
AATSC5p	-0.37	1							
MATS4c	-0.09	-0.15	1						
VE3_Dzp	0.04	-0.42	-0.35	1					
SCH-3	0.75	-0.38	0.32	0.11	1				
Petitjean Number	-0.09	-0.07	-0.17	-0.07	0.11	1			
RDF95u	-0.38	0.64	-0.06	-0.08	-0.21	-0.02	1		
RDF10u	-0.30	0.35	-0.11	0.17	-0.13	-0.02	0.81	1	
ZMIC3	-0.26	0.12	-0.01	-0.13	-0.44	-0.34	0.33	0.53	1

Description of the various descriptors in the Final Model.

AATSC5p :Average centered Broto-Moreau autocorrelation - lag 5 / weighted by polarizabilities

MATS4c: is the Moran autocorrelation - lag 4 / weighted by charges

VE3_Dzp: is the Logarithmic coefficient sum of the last eigenvector from Barysz matrix / weighted by polarizabilities

SCH-3: Valence cluster, order 3

PetitjeanNumber. is the petitjean Number descriptor

Conclusion

A quantitative structure-activity relationships (QSAR) study on a set of 36 structurally-similar 8-methylquinolone analogs was performed using a comprehensive set of theoretical molecular descriptors. The GA-MLR method was employed for the construction of a robust model for prediction of the inhibitory activity (pMIC) against *M. tuberculosis*. The best regression equation was obtained on the following descriptors AATSC, MATS4c ,VE3_Dzp ,SCH-3 ,PetitjeanNumber. The robustness and the predictive ability of the model were verified using a method for internal validation (cross-validation leave-one-out) and Y-Randomization. The predictive power of the model was tested through the extrapolation of the model over the external previously excluded validation data set. The result obtained in this study ($R^2_{Pred} = 0.7393$) suggests that the QSAR model can be used to predict the pMIC of novel 8-methylquinolone analogs which fall within the applicability domain of the model, before synthesizing them.

References:

- Alangaden GJ, Lerner SA.(1997) The clinical use of fluoroquinolones for the treatment of mycobacterial diseases. *Clin Infect Dis.*, 25, 1213
- Anil ,K., Eric A., Nacer L., Jerome G and Koen A. (2011). The challenge of new drug discovery for tuberculosis. *Nature*, 469(483).
- Alexander Tropsha (2010) Best Practices for QSAR Model Development, Validation, and Exploitation *Mol. Inf.*, 29, 476 – 488
- Becke A.D. (1988) Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys Rev A*;38:3098–3100
- Blumberg, HM., Burman, WJ., Chaisson, RE., Daley, CL., Etkind, SC., Friedman, LN.,Fujiwara, P., Grzemska, M., Hopewell, PC., Iseman, MD., Jasmer, R Koppaka V., Menzies, RI., O'Brien, RJ., Reves, RR., Reichman, LB., Simone, PM. Starke, JR., Vernon, AA. (2003). Treatment of tuberculosis. *AM J Crit Care Med* 5 :167(4): 603-62
- Berk, R.A (2003) The formalities of multiple regression, in Regression Analysis: A Constructive Critique, R.A. Berk, ed., London: SAGE Publications Ltd, pp. 103–110.
- Davis, L., (1991). Handbook of Genetic Algorithms. Van Nostrand R, New York
- Deoliveira, D.B ;Gaudio , A. C(2001) ; BuildQsar : A New Computer Program for QSAR Analysis. *Quant. Struc-Activ. Relat.*, 19(6) : 599 – 601,
- Devillers, J.,(1996) . *Genetic Algorithms in Molecular Modeling*. Academic Press Ltd London. J. Devillers, Ed . Academic Press, London, pp. 35-66

- Emami S, Shafiee A and Foroumadi A (2005) .Quinolones: Recent Structural and Clinical Developments *IJPR* 3: 123-136
- Eriksson L , Jaworska J, Worth A.P, Cronin M.T.D, . McDowell R.M and Gramatica P. (2003). Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs *Environmental Health Perspectives* 111 (10): 361.
- Flamm RK, Vojtko C, Chu DT, Li Q, Beyer J, Hensey D, Ramer N, Clement JJ, Tanaka SK (1995) . In vitro evaluation of ABT-719, a novel DNA gyrase inhibitor. *Antimicrob Agents Chemother*, 39: 964-970
- Frieden Fr, Sterling TR, Munsiff SS, Watt CJ, Dye C. (2003). Tuberculosis. *Lancet* 362:887-899.
- Gootz T.D., Brighty K.E(1998).; Chemistry and mechanism of action of the quinolone antibacterials in: *The Quinolones*. 2nd ed. Academic Press, San Diego, pp29.
- Gramatica P. (2007) Principles of QSAR models validation: internal and external. *QSAR Comb Sci* . 26,(5,) 694 – 701.
- Gupta R.A, Gupta A.K, Soni L.K and Kaskhedikar S.G (2007) Rationalization of physicochemical characters of oxazolyl thiosemicarbazone analogs towards multi-drug resistant tuberculosis: A QSAR approach *European Journal of Medicinal Chemistry* 421109 -1116
- Lee C., Yang W., Parr R.G. (1988) Development of the Colle-Salvetti correlation energy formula into a functional of the electron density. *Phys Rev*; 37:785–789
- Leonard V.S and Racheal E B. (2009) Challenges, successes and hopes in the development of novel TB therapeutics. *Future Med. Chem.* 1(4)749-756
- Lescher G.Y., Froelich E.J., Gruett M.D., Bailey J.H. and Brundage R.P.; *J.Med. Pharm. Chem.* 5, 1962, 1063
- Manchester K. (1984) Tuberculosis and leprosy in antiquity: an interpretation. *Medical History*,28:162-173
- National Committee for Clinical Laboratory Standards. Antimycobacterial susceptibility testing for Mycobacterium tuberculosis(1995). Proposed standard M24–T. National Committee for Clinical Laboratory Standards, Villanova, Pa Norris S. and Mandell G.L (1988).; The quinolones: History and Overview, The quinolones, San Diego, Academic Press Inc, pp 1
- Ravichandran V, Harish, Abhishek J, Shalini , Christopher P. V and Ram K A
2011 *International Journal of Drug Design and Discovery*
2(3) :511- 519). Validation of QSAR Models - Strategies and Importance
Roy, K. *Expert Opin. Drug Discov.* (2007), 2: 1567-1577

- Schüürmann G, Ralf-Uwe, E.J. Wang C.and Kühne B .J. (2008) *Chem. Inf. Model.* 48 (11): 2140
- Siram D., Palaniappan S., Murugesan D.,Yogesh C., and Perumal Y.(2009).
Synthesis and in-vitro Antimycobacterial Evaluation of 1-(Cyclopropyl/2,4-difluorophenyl/tert-butyl)-1,4-dihydro-8-methyl-6-nitro-4-oxo-7-(substituted secondary amino)quinoline-3-carboxylic acids *Arch Pharm Chem. Life sci*, 342,100112
- Thangapandian S, John S and Lee K. W .(2012) Classical and 3D QSAR studies on inverse agonists of human histamine H1 receptor 38, *Molecular Simulation* (13):1143–115
- Topliss J. G and Edwards R .P (1979) Chance factors in studies of quantitative structure- activity relationships *J. Med. Chem.* 22 (10): 1238-1244
- Warren Hehre and Sean Ohlinger The *Spartan'14* Tutorial User's Guide (www.wavefun.com)
- Wold, S.and Ericksson, L. statistical validation of QSAR Results in : Van de Waterbeemed H.,(Ed.), Chemometric methods in Molecular design,VCH, Wienheim (Germany) .1995, pp. 309 – 318
- World Health Organization. Tuberculosis Fact sheet. Geneva: World Health Organization; 2014
- Yap CW (2011). PaDel –Descriptor : An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry.* 32 (7):1466-1474