ejpmr, 2024, 12(4), 517-529

SJIF Impact Factor 7.065

EUROPEAN JOURNAL OF PHARMACEUTICAL AND MEDICAL RESEARCH

www.ejpmr.com

Research Article

ISSN 2394-3211

EJPMR

EXPLORATION OF QSAR-BASED VIRTUAL SCREENING FOR THE DISCOVERY OF QUINOLONE-BASED ANTIBACTERIAL DRUGS

Abdulhaqq Taiwo Abdulwaheed*[1], Toheeb Ayodeji Sodiq[2], Wahab Abolore Badmus[3], Makuochukwu Collins Mbah[4], Olaitan Ebenezer Oluwadare[5], Miracle Oladoyin Ojedayo[6], Emmanuel Faderin[7], Precious Ifeanyi Ebere[8] and Oluwagbade Joseph Odimayo[9]

[1]Chemistry Department, Sa'adu Zungur University Gadau, Bauchi State, Nigeria.
[2]Pharmacology Department, Sa'adu Zungur University Gadau.
[3]Chemistry, Sam Houston State University.
[4]Radiology Department, BT Health and Diagnostic Centre LASUTH.
[5]Division of Physics, Engineering, Mathematics and Computer Science (PeMaCs) Delaware State University, USA.
[6]Pharmacy Department, University of Jos.
[7]Department of Pharmaceutical sciences, Southern Illinois University, Edwardsville.
[8]Pharmacy Department, Ahmadu Bello University, Zaria.
[9]Independent Researcher.

***Corresponding Author: Abdulhaqq Taiwo Abdulwaheed**

Chemistry Department, Sa'adu Zungur University Gadau, Bauchi State, Nigeria.

Article Received on 20/02/2025 **Article Revised on 11/03/2025** **Article Accepted on 31/03/2025**

ABSTRACT

Drugs discovering is essential for assessing the potential impact on human health. Using 2D autocorrelation descriptors as predictor variables, a binary logistic regression model was developed to identify active antibacterial among quinolone compounds. The classifications made by the model on the training set compounds resulted in an overall accuracy, sensitivity and specificity of 91.80%, 90.62%, 93.10% dataset. The areas under the ROC curves, constructed with the training set data, was found to be 0.933 for the model. Predictions made by the model on the dataset to the test sets correctly classified 93% of test set compounds selected from datasets. The developed models are considered reliable for rapid discovery of drugs.

KEYWORDS: Autocorrelation descriptor, Binary logistic regression, antibacterial, Quantitative structure-activity relationship.

INTRODUCTION

Antibacterial drugs lie in the need for new and effective antibiotics to combat bacterial infections. Quinolones are a class of antibiotics that have been widely used to treat various bacterial infections for many years. However, the emergence of drug-resistant bacteria poses a significant challenge in the field of antibiotic discovery. Traditional methods of drug discovery are time-consuming, costly, and often yield limited success. Therefore, there is a growing interest in utilizing computational techniques such as virtual screening to accelerate the drug discovery process. QSAR (quantitative structure activities relationships), is a computational method that correlates the chemical structure of a molecule with its biological activity. By developing a QSAR model specific to quinolone-based antibacterial drugs, researchers can gain insights into what structural features are important for their activity against bacteria. This information can then be used to screen large databases of chemical compounds and identify potential drug candidates with similar structural characteristics.[1]

The exploration of QSAR-based virtual screening for the discovery of quinolone-based antibacterial drugs is quite extensive. Researchers have conducted studies to develop QSAR models that can predict the activity of quinolone compounds against bacteria. One example of a study in this field is "QSAR models for predicting the antibacterial activity of quinolone derivatives" by Khan et al. (2019). The researchers utilized various computational methods to develop QSAR models that could predict the antibacterial activity of quinolone derivatives. They used molecular descriptors to represent the structural features of the compounds and employed machine learning algorithms to build the models. Another study, "QSAR models for predicting the activity of novel quinolone derivatives against Methicillin-resistant Staphylococcus aureus"

Virtual screening is a computational method used in drug discovery to identify molecules that have the potential to interact with a target protein or enzyme. QSAR, on the other hand, is a technique that correlates the chemical structure of a molecule with its biological activity or property. By combining these two approaches, researchers can predict and prioritize potential quinolone-based compounds that could be effective in fighting bacterial infections. To develop a QSAR model specific to quinolone-based antibacterial drugs, which can accurately predict their activity against bacteria. This model could then be used to virtually screen large databases of chemical compounds and identify potential drug candidates for further exploration and development.

Despite the availability of various quinolone-based antibiotics, the emergence of drug-resistant bacteria poses a significant challenge in the field of antibacterial drug discovery. Traditional experimental methods for screening potential drug candidates are often time-consuming and costly. Therefore, there is a need to explore QSAR-based virtual screening techniques to efficiently identify new quinolone derivatives with improved antibacterial properties. This research aims to develop a reliable QSAR model that can accurately predict the antibacterial activity of quinolone compounds.

The main aim of this study was to discover quinolone-based antibacterial drugs using QSAR-based virtual screening. Specifically, the study set out to achieve the following objectives:

- To develop a classification-based QSAR model that can discriminate between compounds that are active against *Escherichia coli* and those that are inactive against *Escherichia coli*.
- To determine the predictive ability of the developed QSAR model on test set compounds.
- To use the developed QSAR model to screen ChemBL database for possible identification of drug-like compounds that are active against *Escherichia coli*.

The development of a binary logistic regression model for categorizing chemical compounds into antibacterial and non-antibacterial has several significant implications.

Efficient drug discovery: The ability to accurately predict the antibacterial properties of chemical compounds can greatly accelerate the drug discovery process. It allows researchers to prioritize and focus on compounds with higher potential for antibacterial activity, saving time, and resources.

Targeted antimicrobial therapy: The model can help in the development of targeted antimicrobial therapies by identifying compounds that specifically target antibacterial mechanisms. This can lead to more effective and tailored treatments for bacterial infections.

Reduced resistance development: Antibiotic resistance is a growing concern globally. By accurately categorizing compounds, the model can aid in the discovery of new chemical entities that have a lower likelihood of developing resistance. This can help researchers stay ahead of evolving bacterial resistance mechanisms.

Cost-effective screening: Traditional experimental screening of chemical compounds can be expensive and time consuming. The logistic regression model provides a cost-effective alternative.

RESEARCH METHODOLOGY

Dataset and its Sources

The dataset used for developing and validating the binary logistic regression model reported in this research project was obtained from literature.[2] A total of 82 molecules belonging to the quinolone family of antibacterial compounds were selected and split into two groups, 43 compounds with proven antibacterial activity and 39 compounds described as inactive against E. coli.[2]

Calculation and Preprocessing of Molecular Descriptors

Two-dimensional structures of the 82 quinolone-family of antibacterial molecules were drawn using the 2D sketch palette in Spartan '14 software.[3] These 2D structures were converted into 3D structures and then optimized using semi-empirical AM1 model as implemented in Spartan '14 software.[3] These optimized structures were then imported into PaDEL-Descriptor software and a total of 1444 2D molecular descriptors were calculated for each quinolone molecule in the dataset.[4] Highly-correlated descriptors (redundant descriptors) and descriptors with constant or nearly constant values (irrelevant descriptors) were eliminated from the pool of molecular descriptors calculated by PaDEL-Descriptor software.[4] In this project research, highly-correlated descriptors with correlation coefficient exceeding 0.90 and constant-value descriptors with variance lower than 0.0001 were removed using V-WSP algorithm[5] as implemented in V-WSP tool (version 1.2) developed by Ambure et al. (2015). Correlation matrix was constructed to verify the absence of multicollinearity in the final 2D autocorrelation descriptors selected for model building.

Dataset Division

The 43 quinolone molecules listed as active compounds in Table 3.1 were split into training and test sets, with the training set being 75% of the total active compounds and the test set being 25% of the total active compounds. The 39 inactive compounds listed in Table 3.2 were also divided into training and test sets, with the training set being 75% of the entire inactive compounds and the test set being 25% of the entire inactive compounds. The dataset division procedure described above was implemented in Dataset split GUI 1.2 developed by Ambure et al. (2015) using Kennard-Stone.[6] The 32 active compounds and the 29 inactive compounds assigned to the training set were then combined to form 61 training set compounds. These 61 training set compounds was used to develop the binary logistic regression model reported in this project research. Similarly, the 11 active compounds and the 10 inactive compounds assigned to the test set were also combined to form 21 test set compounds. These 21 tests set compounds were reserved for external validation of the developed model. The 61 quinolone molecules assigned to the training set, along with the values of 2D autocorrelation descriptors selected for model building, are shown in the Appendix I. Similarly, the 21 quinolone molecules assigned to the test set, along with the values of 2D autocorrelation descriptors used for model validation are shown in Appendix II.

Development of Quantitative Structure-Activity Relationship Model

The classification-based QSAR models reported for datasets in this research was developed using binary logistic regression as implemented in IBM® SPSS® Statistics (version 26). In this multivariate statistical method, the 2D autocorrelation descriptors for the training set compounds (61 quinolone molecules) was used as input independent variables while the coded values of discrete class labels of compounds in the training set (1 for active quinolone compounds and 0 for inactive quinolone compounds) were used as dependent variable. Feature selection was carried out using forward conditional procedures as implemented in IBM® SPSS® Statistics (version 26). The goodness-of-fit and reliability of the developed binary logistic regression model were evaluated using Wald test, Omnibus test, Hosmer and Lemeshow test, and Nagelkerke R square. The binary logistic regression models generated for datasets was then used to compute the logit values and posterior probabilities of group memberships for all the training set compounds. These predicted posterior probabilities were used to classify the training set compounds into antibacterial and non-antibacterial compounds. Compounds with predicted probability greater than 0.5 were classified as active (antibacterial compounds) while compounds with predicted probability lower than 0.5 were classified as inactive (non-antibacterial compounds).

Evaluation of Model Performance

The classification obtained for the training set compounds in the datasets was organized in a specific table layout known as confusion matrix that allows easy calculation of true positive (TP), true negative (TN), false positive (FP) and false negative (FN). In this project, TP and TN refer to the number of active and inactive antibacterial compounds that was correctly classified by the models as active and inactive antibacterial compounds, while FP and FN refer to the number of inactive and active antibacterial compounds that was misclassified by the models as active and inactive antibacterial compounds respectively. From the values of TP, TN, FP and FN obtained, performance evaluation metrics such as accuracy (ACC), sensitivity or true positive rate (TPR), and specificity or true negative rate (TNR) were calculated using the formulae shown in Equations 3.1–3.3. The performance of the developed models vis-à-vis the classifications made on the training set compounds in datasets was also evaluated graphically using receiver operating characteristic (ROC) curves.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3.1)$$

$$TPR = \frac{TP}{TP+FN} \qquad (3.2)$$

$$TNR = \frac{TN}{TN+FP} \qquad (3.3)$$

External Validation of QSAR Model

The values of the 2D autocorrelation descriptors calculated for the test set compounds (Appendix II) were used to calculate the logit values and the posterior probabilities of group memberships for the test set compounds in datasets. The predicted posterior probability calculated for each test compound was then used to classify the compound into either active or inactive antibacterial. As stated earlier in this project research, compounds with predicted probability greater than 0.5 were classified as active while compounds with predicted probability lower than 0.5 were classified as inactive. The predictive abilities of the classification model developed in this research project was then externally evaluated by calculating the proportion of active and inactive compounds in the test set that was correctly classified by the model.

Virtual Screening of ChemBL Database

A total of 46 quinolone compounds, obtained from ChemBL database, were screened using the classification-based QSAR model developed and validated in the preceding sections. Two-dimensional structure of each of the 46 quinolone molecules was drawn using the 2D sketch palette in Spartan '14 software.[4] These 2D structures were converted into 3D structures and then optimized using semi-empirical AM1 model as implemented in Spartan '14 software.[4] The optimized structures were then imported into PaDEL-Descriptor software and relevant autocorrelation descriptors were calculated.[5] Values of molecular descriptors and predicted group memberships of compounds screened are shown in Appendix V. Thereafter, the classification-based QSAR model described above was applied to screen and identify quinolone molecules that could potentially acts as antibacterial drugs against E. coli among the 46-quinolone obtained from ChemBL database.

RESULTS AND DISCUSSION

Reliability of the Developed QSAR Model

Two 2D autocorrelation descriptors was selected for building the binary logistic regression model reported in this research project. The symbols and definitions of these two 2D autocorrelation descriptors are shown in Table 1.0 The two autocorrelation descriptors listed in Table 1.0 are ATSC4p and AATSC0i, belonging to centered Broto-Moreau autocorrelation and average centered Broto-Moreau autocorrelation respectively. The values of ATSC4p and AATSC0i computed for compounds assigned to the training set and test set in dataset are shown in Appendix I and Appendix II respectively. Using the values of ATSC4p and AATSC0i shown in Appendix I as predictor variables and the coded values of the discrete class labels of the training set compounds as outcome variable (1 for active quinolone compounds and 0 for inactive quinolone compounds), application of binary logistic regression method produced the logistic regression coefficients (B), their standard errors (S.E.), the p-values, the odds ratios (Exp(B)) and the 95% confidence intervals of the odds ratios listed in Table 2.0. From the values of the logistic regression coefficients listed in Table 2.0, the binary logistic regression model displayed in Eq. 1 was constructed. The p-value displayed in Tables 2.0 for the predictor variables indicates that the strength of the relationship between the outcome variable and the predictor variables was statistically significant at $p < 0.05$. In Tables 2.0 the odds ratios of ATSC4p and AATSC0i were found to be greater than one. This indicates that quinolones compounds with higher value of ATSC4p and AATSC0i have higher likelihood of being classified as active antibacterial compounds.

The result of the Omnibus test of model coefficients to assess the goodness-of-fit is presented in Tables 3.0. The result of the Omnibus test of model coefficients shown in Table 3.0 indicate that there was significant improvement in fit ($p < 0.05$) for the QSAR model displayed in Eq. 1 when compared to the null model constructed without any predictor variable. The result of the Hosmer and Lemeshow test to also assess the goodness-of-fit of the QSAR model displayed in Eq 1 is presented in Tables 4.0 As shown in Table 4.0, there was no significant difference between the observed outcome and the outcome predicted by the QSAR model ($\chi^2(8) = 5.461$, $p = 0.707$), indicating that the QSAR model displayed in Eq. 4.1 adequately fit the data in the training set. In Table 5.0, pseudo-R-squared values (Nagelkerke R^2 and Cox and Snell R^2) were presented for the QSAR model displayed in Eq. 1. The Nagelkerke R^2 is an adjusted version of the Cox and Snell R^2. It adjusts the scale of the statistic to cover the full range from 0 to 1. As shown in Table 5.0, the Nagelkerke R^2 value of 0.819 reported for the QSAR model indicates that 81.9% of variation in the outcome variable in the QSAR model can be accounted for by the predictor variables in the model.

$$ln\left(\frac{P}{1-P}\right) = -12.297 + 0.915\ \mathrm{ATSC4p} + 7.049\ \mathrm{AATSC0i} \qquad (1)$$

Table 1.0: Symbols and definitions of molecular descriptors utilized in building models I and II.

Symbol	Definition	Type	Class
ATSC4p	Centered Broto-Moreau autocorrelation–lag 4 / weighted by polarizabilities	2D	Autocorrelation
AATSC0i	Average centered Broto-Moreau autocorrelation–lag 0 / weighted by first ionization potential	2D	Autocorrelation

Table 2.0: Logistic regression coefficients and odds ratios of 2D autocorrelation descriptors utilized in building binary logistic regression model.

	B	S.E.	Wald	df	p-value	Exp(B)	95% C.I. for Exp(B)	
							Lower	Upper
ATSC4p	0.915	0.256	12.775	1	0.000	2.496	1.511	4.121
AATSC0i	7.049	2.410	8.555	1	0.003	1151.964	10.232	129693.503
Constant	-12.297	4.860	6.402	1	0.011	0.000		

Table 3.0: Omnibus test of model coefficients.

	Chi-square	Df	p-value
Step	58.079	2	0.000
Block	58.079	2	0.000
Model	58.079	2	0.000

Table 4.0: Hosmer and Lemeshow test.

Parameter	Value
Chi square	5.461
Df	8
p-value	0.707

Table 5.0: Model summary.

Parameter	**Value**
-2Log likelihood	26.337
Cox & Snell R square	0.614
Nagelkerke R square	0.819

Internal Validation of the Developed QSAR Model

Having established the fitness of the developed QSAR model in the preceding section (Section 4.1.2), the model was then used to calculate the probabilities of allotting the training set compounds to the active or inactive class. Appendix III shows the predicted probabilities of allotting the training set compounds to the active or inactive class. As shown in Appendix III, a compound was classified as active if P(active) > 0.5 but classified as inactive if P(active) < 0.5. The classifications of the training set compounds in Appendix III by the QSAR model displayed in Eq. 4.1 are summarized in the confusion matrix shown in Table 6.0 As shown Table 6.0, exactly 29 out of the 32 active compounds and 27 out of the 29 inactive compounds in the training set were correctly classified by the QSAR model.

Evaluating the performance of the prediction made on the training set compounds by the QSAR model displayed in Eq. 1 using the performance metric defined in Equations 3.1, 3.2 and 3.3 (Chapter Three) resulted in the values of the performance metrics listed in Table 7.0, As shown in Table 7.0, the overall accuracy (ACC), sensitivity or true positive rate (TPR) and specificity or true negative rate (TNR) obtained for the prediction made on the training set compounds are 91.80%, 90.62% and 93.10% respectively. The values of the performance metrics reported in Table 7.0, indicate satisfactory classifications of the training set compounds by the QSAR model displayed in Eq. 1. The performance of the classifications predicted by the QSAR model on the training set compounds was also evaluated graphically using the receiver operating characteristic (ROC) curves shown in Figure 1. The area under this ROC curve (AUC) was 0.972. The high value of the AUC reported in Figure 1 suggests excellent discriminating ability of the QSAR model displayed in Eq. 1.

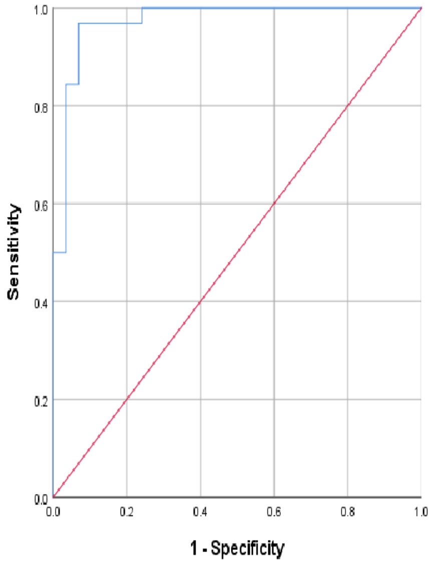
Table 6.0: Confusion matrix for the prediction made on training set compounds*

	Predicted group membership			
Class	Active (1)	Inactive (0)	Total	Correct prediction
Active (1)	29	3	32	90.6%
Inactive (0)	2	27	29	93.1%

* TP = 29, TN = 27, FP = 2, FN = 3

Table 7.0: Values of performance metrics for the prediction made on the training set compounds.

Performance metric		Value (%)
Metric	Symbol	
Accuracy	ACC	91.80
Sensitivity (or true positive rate)	TPR	90.62
Specificity (or true negative rate)	TNR	93.10



AUC = 0.972; std. error = 0.020; p-value = 0.000; 95% confidence interval = 0.933 to 1.000

Figure 1: Receiver operating characteristic (ROC) curve for evaluating the performance of the developed model on training set compounds.

Predictive Ability of Classification-Based QSAR Model

The predictive abilities of the QSAR model displayed in Eq. 4.1 was externally evaluated using quinolone compounds that were not part of the compounds used for model building. To accomplished this task, the binary logistic regression model displayed in Eq. 1 was used to calculate the probabilities of allotting the test set compounds to active or inactive class using the values of the 2D autocorrelation descriptors shown in Appendix II for the test set compounds. The results of the classifications made by the QSAR model on the test set were shown in Appendix IV. The results presented in Appendix IV are depicted graphically in Figure 2. In Figure 2, compounds above the horizontal cut-off line were classified as active while compounds below the horizontal cut-off lines were classified as inactive. As shown in Fig. 2, of the 21 quinolone compounds assigned to the test, only compounds A34 and A38 were misclassified by the QSAR model. The proportions of active and inactive compounds that were correctly classified in Figure 2 by the QSAR model are 82% and 100% respectively. The result of external validation of the QSAR model presented in Figure 2 suggests that the binary logistic regression model developed in this research project has good predictive ability when applied to new quinolone compound that was not part of the compounds used for building the QSAR model displayed in Eq. 1.

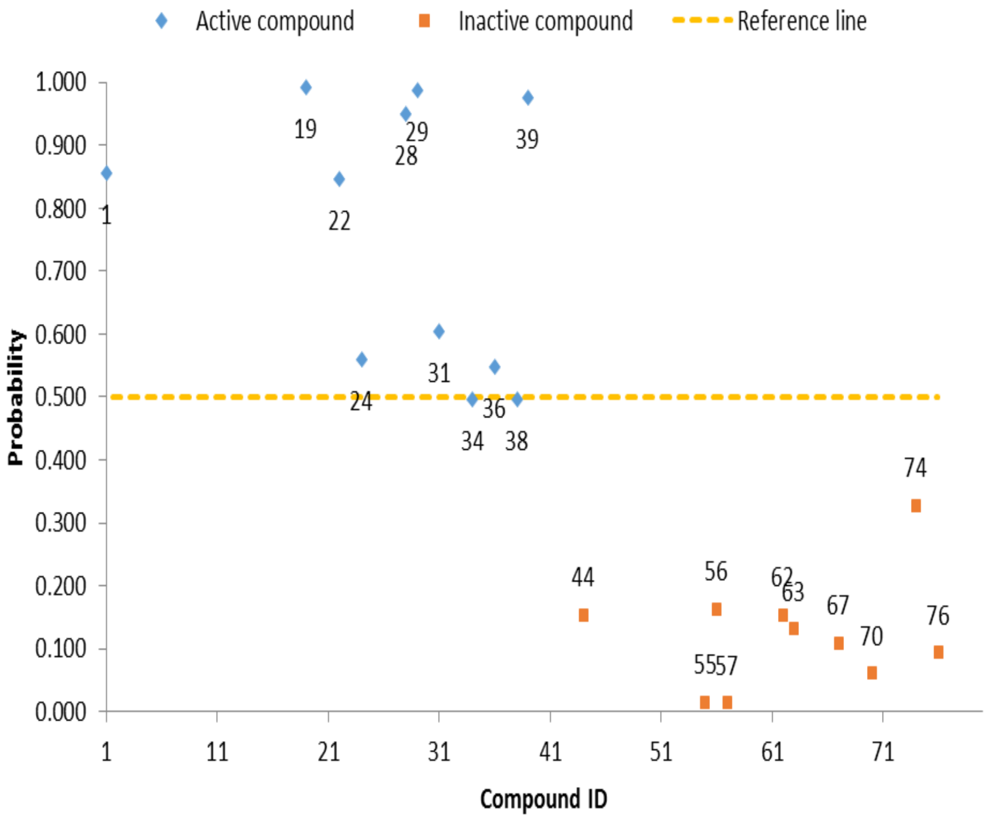
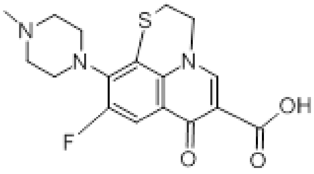


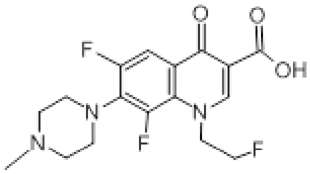
Figure 2: Graphical representation of the prediction made on test set compounds.

Virtual Screening of ChEMBL Database

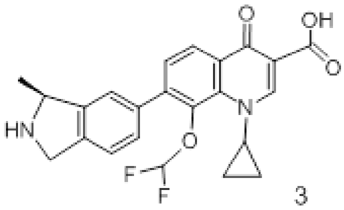
Finally, the 46 quinolone compounds with unknown antibacterial activities in the ChEMBL database were screened using the QSAR model displayed in Eq. 4.1. To accomplish this task, the values of ATSC4p and AATSC0i were computed for each of the 46 quinolone compounds in the ChEMBL database (see Appendix V) and the values of these descriptors were substituted in Eq. 4.1. The probabilities of allotting the screened quinolone compounds in the ChEMBL database to the active or inactive class were then predicted (see Appendix V). The 2D structures of the 15 novel quinolone compounds from ChEMBL database predicted to be active against *Escherichia coli* are shown in Figure 3.0



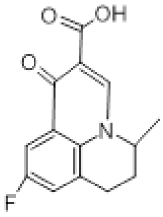
1



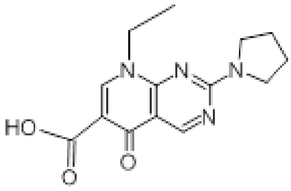
2



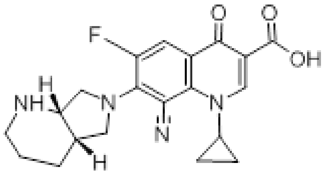
3



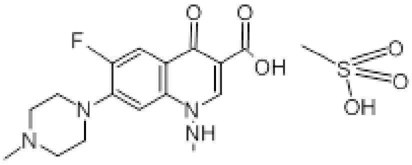
4



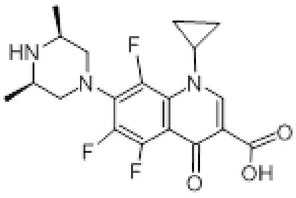
5



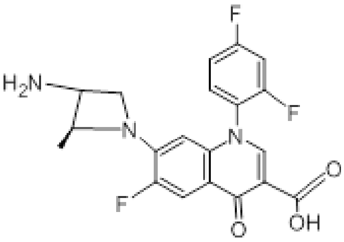
6



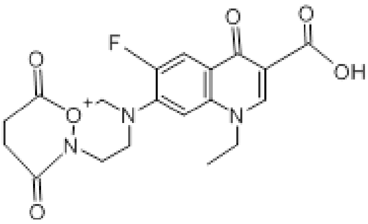
7



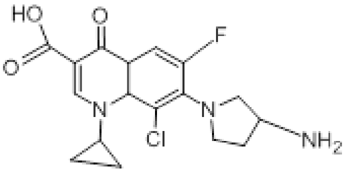
8



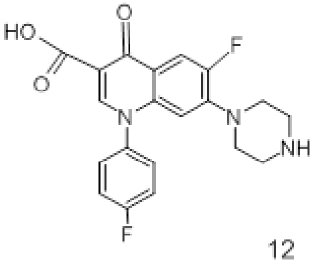
9



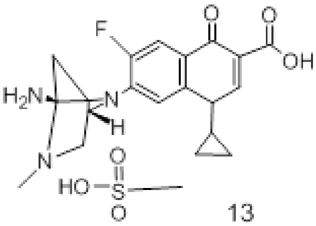
10



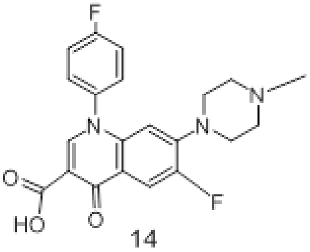
11



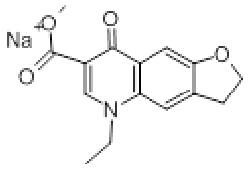
12



13



14



15

Figure 3: New quinolone compounds from ChEMBL database predicted to be active against *Escherichia coli.*

DISCUSSION

Classification of quinolone derivatives into active and inactive antibacterial compounds was accomplished using the classification-based QSAR model developed in this research project. The performance of the classification made on the training and test set compounds by the developed QSAR model was found to be reasonably good. The QSAR model developed in this research project was able to identify 15 novel quinolone-based antibacterial compounds from the ChEMBL database. Classification-based QSAR models, using theoretically-derived molecular descriptors as predictor variables, have been used by several researchers to identify antibacterial compounds from Pubchem, ChemSpinder, and Chematical.[8] The findings reported in this research project is consistent with what were reported in the literature cited above. Some of the main attractions of using binary logistic regression algorithm for building classification-based QSAR models include easy implementation of the algorithm, easy interpretation of the resulting models, no assumption about distribution of classes in feature space is required, the algorithm is less inclined to over-fitting, and it is one of the most efficient algorithms when the different outcomes represented by the dataset are linearly separable.[9] The classification-based QSAR model developed in this research project is therefore considered suitable for rapid identification of quinolone-based antibacterial compounds in chemical databases of.

CONCLUSION

The binary classification models, utilizing 2D descriptors as predictor variables, was developed using binary logistic regression. The performance of the classification model, their abilities to correctly classify into active and inactive antibacterial, was found to be satisfactory. Some of the main attractions of using binary logistic regression algorithm for building classification-based QSAR models include easy implementation of the algorithm, easy interpretation of the resulting models, no assumption about distribution of classes in feature space is required, the algorithm is less inclined to over-fitting, and it is one of the most efficient algorithms when the different outcomes represented by the dataset are linearly separable, The binary logistic regression model developed in this project are therefore considered suitable for rapid identification of quinolone family of antibacterial.

RECOMMENDATION

Further studies on synthesis and experimental validation of the new quinolone compounds identified as potential antibacterial agents in this research project are suggested.

REFRENCES

1. Z. Zhang, C. Jia, Y. Hu, L. Sun, J. Jiao, L. Zhao, D. Zhu, J. Li, Y. Tian, H. Bai, R. Li, J. Hu. The estrogenic potential of salicylate esters and their possible risks in foods and cosmetics. Toxicol. Lett., 2012; 209(2): 146-153.
2. Lukman k.Akinola, Adamu Uzair, Gideon A, shallagwa, Stephie E abechi. In silico prediction of nuclear receptor binding to polychlorinated dibenzofurans and its implication on endocrine disruption in humans and wildlife, 2021 The Authors. Published by Elsevier B.V.
3. Dalisay D. S., Lievens S. L., Saludes J. P., Molinski T. F., Nat. Rev. Drug Discov. 8, 69 (2008). Google Scholar Skropeta D., Nat. Prod. Rep. 2008; 25: 1131. PubMed Abstract | CrossRef Full Text | Google Scholar
4. AlMatar, M., AlMandeal, H., Var, I., Kayar, B., and Köksal, F. New drugs for the treatment of Mycobacterium tuberculosis infection. Biomed. Pharmacother, 2017; 91: 546–558. doi: 10.1016/j.biopha.2017.04.105. PubMed Abstract | CrossRef Full Text | Google Scholar
5. Bajorath, J. Computational chemistry in pharmaceutical research: at the crossroads. J. Comput. Aided. Mol. Des., 2012; 26: 11–12. doi: 10.1007/s10822-011-9488-z. PubMed Abstract | CrossRef Full Text | Google Scholar
6. Ban, F., Dalal, K., Li, H., LeBlanc, E., Rennie, P. S., and Cherkasov, A. Best practices of computer-aided drug discovery: lessons learned from the development of a preclinical candidate for prostate cancer with a new mechanism of action. J. Chem. Inf. Model, 2017; 57: 1018–1028. doi: 10.1021/acs.jcim.7b00137. PubMed Abstract | CrossRef Full Text | Google Scholar
7. Fourches, D., Muratov, E., and Tropsha, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J. Chem. Inf. Model., 2010; 50: 1189–1204. doi: 10.1021/ci100176x. PubMed Abstract | CrossRef Full Text | Google Scholar
8. Kuntz, A. N., Davioud-Charvet, E., Sayed, A. A., Califf, L. L., Dessolin, J., Arnér, E. S. J., et al. Thioredoxin glutathione reductase from Schistosoma mansoni: an essential parasite enzyme and a key drug target. PLoS Med., 2007; 4: e206. doi: 10.1371/journal.pmed.0040206. PubMed Abstract | CrossRef Full Text | Google Scholar
9. Cihlar, T., and Fordyce, M. Current status and prospects of HIV treatment. Curr. Opin. Virol., 2016; 18: 50–56. doi: 10.1016/j.coviro.2016.03.004. PubMed Abstract | CrossRef Full Text | Google Scholar
10. Ekins, S., Lage de Siqueira-Neto, J., McCall, L.-I., Sarker, M., Yadav, M., Ponder, E. L., et al. Machine learning models and pathway genome data base for Trypanosoma cruzi drug discovery. PLoS Negl. Trop. Dis., 2015; 9: e0003878. doi: 10.1371/journal.pntd.0003878

APPENDIX I

Values of molecular descriptors for training set compounds utilized in building the binary logistic regression model developed in the research project

Compound ID	**Actual class**	**Molecular descriptor**	
		ATSC4p	**AATSC0i**
A2	Active	-1.741659	2.298635
A3	Active	-1.712508	2.335964
A4	Active	-0.599320	2.335964
A5	Active	-0.853962	2.052579
A6	Active	-0.020991	2.292271
A7	Active	-1.655569	2.408287
A8	Active	-0.926884	2.335964
A9	Active	-1.308759	2.353237
A10	Active	3.004714	2.326603
A11	Active	3.746369	2.262945
A12	Active	3.620820	2.281289
A13	Active	-0.172052	2.368356
A14	Active	0.937394	2.368356
A15	Active	1.574973	2.301940
A16	Active	1.733726	2.331794
A17	Active	-0.302501	2.443411
A18	Active	0.386172	2.368356
A20	Active	-5.355444	2.718713
A21	Active	-7.303096	2.424958
A23	Active	2.939857	1.892872
A25	Active	4.873740	1.989665
A26	Active	4.006051	3.054220
A27	Active	-0.181951	2.032496
A30	Active	-4.437629	2.734069
A32	Active	1.036849	1.958772
A33	Active	-0.935387	1.905417
A35	Active	-1.291181	2.395300
A37	Active	-2.282827	1.992832
A40	Active	1.113959	2.297586
A41	Active	4.523334	2.992993
A42	Active	-2.483469	2.159541
A43	Active	-2.854034	2.076074
A45	Inactive	-1.997974	1.943574
A46	Inactive	-7.763079	2.110541
A47	Inactive	-8.809261	2.100579
A48	Inactive	-8.452686	2.055480
A49	Inactive	-6.033951	1.795398
A50	Inactive	-8.126481	2.384928
A51	Inactive	-5.472503	2.263200
A52	Inactive	-6.639824	2.048005
A53	Inactive	-4.466652	2.038389
A54	Inactive	-6.383564	2.090817
A58	Inactive	-6.998472	2.003510
A59	Inactive	-4.485569	2.038389
A60	Inactive	-3.761884	1.951891
A61	Inactive	-5.758592	2.337165
A64	Inactive	-5.960265	1.906918
A65	Inactive	-4.335950	1.949659
A66	Inactive	-6.232556	2.338406
A68	Inactive	-4.309184	1.878927
A69	Inactive	-5.657789	1.907038
A71	Inactive	-4.665113	1.874844
A72	Inactive	-1.813098	2.152911

A73	Inactive	-1.910323	2.493591
A75	Inactive	-1.960988	1.828694
A77	Inactive	-5.018936	1.712757
A78	Inactive	-7.452917	1.872520
A79	Inactive	-1.131784	1.480856
A80	Inactive	-6.493638	1.423101
A81	Inactive	-1.508969	1.457516
A82	Inactive	-5.340324	1.951615

APPENDIX II

Values of molecular descriptors for test set compounds utilized in validating the predictive ability of the binary logistic regression model developed in the research project

Compound ID	Actual class	Molecular descriptor	
		ATSC4p	AATSC0i
A1	Active	-1.851426	2.238068
A19	Active	0.239609	2.396187
A22	Active	-1.762280	2.214772
A24	Active	-1.731223	2.004060
A28	Active	1.315396	1.991431
A29	Active	-3.576805	2.835645
A31	Active	-0.573852	1.879510
A34	Active	-1.393283	1.922850
A36	Active	-0.614316	1.851939
A38	Active	-1.393283	1.922850
A39	Active	2.288192	1.970600
A44	Inactive	-3.324084	1.934149
A55	Inactive	-6.563876	2.003510
A56	Inactive	-4.050972	2.038389
A57	Inactive	-6.675153	2.007327
A62	Inactive	-3.625497	1.973774
A63	Inactive	-3.743379	1.961462
A67	Inactive	-3.439449	1.892865
A70	Inactive	-6.607311	2.214843
A74	Inactive	-0.961638	1.767300
A76	Inactive	-3.876221	1.926668

APPENDIX III

Probabilities and group memberships predicted by the binary logistic regression model developed in the research project for training set compounds

Compound ID	Actual group membership		Predicted group membership	
	Class	Class code	Predicted probability	Predicted class
A2	Active	1	0.910	1
A3	Active	1	0.931	1
A4	Active	1	0.974	1
A5	Active	1	0.801	1
A6	Active	1	0.979	1
A7	Active	1	0.959	1
A8	Active	1	0.965	1
A9	Active	1	0.957	1
A10	Active	1	0.999	1
A11	Active	1	0.999	1
A12	Active	1	0.999	1
A13	Active	1	0.986	1
A14	Active	1	0.995	1
A15	Active	1	0.995	1
A16	Active	1	0.997	1
A17	Active	1	0.991	1
A18	Active	1	0.991	1

A20	Active	1	0.878	1
A21	Active	1	0.132	0
A23	Active	1	0.977	1
A25	Active	1	0.998	1
A26	Active	1	1.000	1
A27	Active	1	0.866	1
A30	Active	1	0.949	1
A32	Active	1	0.921	1
A33	Active	1	0.569	1
A35	Active	1	0.968	1
A37	Active	1	0.416	0
A40	Active	1	0.993	1
A41	Active	1	1.000	1
A42	Active	1	0.658	1
A43	Active	1	0.432	0
A45	Inactive	0	0.396	0
A46	Inactive	0	0.011	0
A47	Inactive	0	0.004	0
A48	Inactive	0	0.004	0
A49	Inactive	0	0.006	0
A50	Inactive	0	0.051	0
A51	Inactive	0	0.206	0
A52	Inactive	0	0.019	0
A53	Inactive	0	0.118	0
A54	Inactive	0	0.032	0
A58	Inactive	0	0.010	0
A59	Inactive	0	0.116	0
A60	Inactive	0	0.121	0
A61	Inactive	0	0.252	0
A64	Inactive	0	0.013	0
A65	Inactive	0	0.075	0
A66	Inactive	0	0.180	0
A68	Inactive	0	0.048	0
A69	Inactive	0	0.017	0
A71	Inactive	0	0.034	0
A72	Inactive	0	0.772	1
A73	Inactive	0	0.972	1
A75	Inactive	0	0.231	0
A77	Inactive	0	0.008	0
A78	Inactive	0	0.003	0
A79	Inactive	0	0.052	0
A80	Inactive	0	0.000	0
A81	Inactive	0	0.032	0
A82	Inactive	0	0.032	0

APPENDIX IV

Probabilities and group memberships predicted by the binary logistic regression model developed in the research project for test set compounds

Compound ID	Actual group membership			Predicted group membership	
	Class	Class code		Predicted probability	Predicted class
A1	Active	1		0.856	1
A19	Active	1		0.992	1
A22	Active	1		0.846	1
A24	Active	1		0.561	1
A28	Active	1		0.950	1
A29	Active	1		0.988	1
A31	Active	1		0.605	1
A34	Active	1		0.496	0

A36	Active	1		0.549	1
A38	Active	1		0.496	0
A39	Active	1		0.976	1
A44	Inactive	0		0.154	0
A55	Inactive	0		0.015	0
A56	Inactive	0		0.163	0
A57	Inactive	0		0.014	0
A62	Inactive	0		0.154	0
A63	Inactive	0		0.131	0
A67	Inactive	0		0.109	0
A70	Inactive	0		0.061	0
A74	Inactive	0		0.328	0
A76	Inactive	0		0.094	0

APPENDIX V

Values of molecular descriptors and predicted group memberships of compounds screened from ChemBL database

Compound ID	Molecular descriptor		Predicted group membership	
	ATSC4p	AATSC0i	Predicted probability	Predicted class
T1	4.864649	2.998620	1.000	1
T2	1.188616	2.093948	0.972	1
T3	-4.299364	2.322997	0.536	1
T4	-3.576805	2.835645	0.988	1
T5	-2.329256	1.440312	0.014	0
T6	3.160794	1.551774	0.823	1
T7	-0.918454	2.158122	0.888	1
T8	2.288192	1.970600	0.976	1
T9	-2.841175	1.436609	0.008	0
T10	-4.519961	1.905417	0.047	0
T11	-5.120566	1.627260	0.004	0
T12	0.751619	2.944007	1.000	1
T13	2.625753	2.032496	0.988	1
T14	0.227822	2.255044	0.978	1
T15	-0.612542	2.458745	0.989	1
T16	-3.193319	1.842347	0.097	0
T17	-2.726240	1.444160	0.010	0
T18	-2.606785	1.951451	0.284	0
T19	-1.563665	1.841171	0.321	0
T20	-1.478435	2.149064	0.817	1
T21	1.575399	2.739969	1.000	1
T22	-0.505296	2.409687	0.986	1
T23	-0.729288	2.692404	0.998	1
T24	0.176318	2.489299	0.996	1
T25	-1.291181	2.395300	0.968	1
T26	-1.962787	1.807013	0.205	0
T27	-0.729288	2.692404	0.998	1
T28	-2.646889	2.076074	0.479	0
T29	4.873740	1.989665	0.998	1
T30	1.006401	1.910459	0.890	1
T31	0.176318	2.489299	0.996	1
T32	-3.673842	1.976016	0.151	0
T33	-1.436795	1.960223	0.551	1
T34	-1.906332	1.421120	0.018	0
T35	-2.614264	1.850353	0.162	0
T36	-4.299364	2.322997	0.536	1
T37	-1.611742	1.905417	0.416	0
T38	-2.119647	1.959206	0.395	0
T39	-2.987418	2.093948	0.433	0
T40	3.093427	1.953924	0.987	1

T41	-1.436795	1.960223	0.551	1
T42	-0.505296	2.409687	0.986	1
T43	2.651891	1.443439	0.575	1
T44	-1.815838	1.433995	0.021	0
T45	3.678308	1.939692	0.991	1
T46	-3.673842	1.976016	0.151	0