Biological Forum – An International Journal 15(1): 505-513(2023)

ISSN No. (Print): 0975-1130
ISSN No. (Online): 2249-3239

Development and Validation of an Easily Interpretable QSAR Model for Inhibitory Activity Prediction against Dihydrofolate Reductase from *Candida albicans*

Sharav Desai[1*], ***Vijay K. Patel***[2], ***Ankita S. Patel***[3] ***and Jaini Patel***[2]
[1]Department of Pharmaceutics,
Sanjivani College of Pharmaceutical Education and Research, Kopargaon (Maharashtra) India.
[2]Department of Pharmaceutics,
Sharda School of Pharmacy, Pethapur, Gandhinagar (Gujarat), India.
[3]Department of Pharmaceutical Chemistry,
Sharda School of Pharmacy, Pethapur, Gandhinagar (Gujarat), India.

(Corresponding author: Sharav Desai)*

(Received: 09 December 2022; Revised: 11 January 2023; Accepted: 15 January 2023; Published: 21 January 2023)
(Published by Research Trend)

ABSTRACT: Candidiasis is a devastating infection caused by the fungi *Candida albicans* species of the genus Candida. The current treatment available for candidiasis is affected by drug-resistant strains. The primary objective of the study was to developa robust and accurate regression-based QSAR model to virtually predict the inhibitory activity of the dihydrofolate reductase enzyme present in *Candida albicans*. We have collected 281 chemical compounds with known inhibitory activity from the ChEMBL webserver. We initially used manual curation to remove blank and false entries from the downloaded databases. All the structures were converted into sdf format using OpenBabel software. We calculated more than 2400 structural descriptors for each class of chemical compound using Alvadesc software. The main challenge encountered during the study was handling such massive data produced after calculating descriptors. Several feature selection techniques are used to reduce the number of insignificant descriptors. A total of four machine learning algorithms named MLR, SVR, RF, and RT were used to build the QSAR model on the training dataset. We used R2, MAE, Y-randomization, applicability domain, and prediction reliability indicators as statistical tools to find out the robustness, stability, and predictability of the model. The model showed satisfactory results in all the calculated parameters under the acceptable range. The developed can be used to screen inhibitors against *Candida albicans*.

Keywords: QSAR, *Candida albicans*, Dihydrofolate reductase, Y-randomization, Applicability domain, prediction reliability indicator.

INTRODUCTION

Fungal infections are becoming common nowadays and among them, the main infection is invasive candidiasis which is typically caused by *Candida albicans*, *Candida tropicalis*, *Candida glabrata,* and *Candida parapsilosis* belonging to the genus Candida. Candidiasis is an opportunistic infection and it can infect the oral cavity, vaginal, penis, and other moist locations of the body. It is also recorded that untreated candidiasis infections resulted in systemic infection with the involvement of other internal organs. Furthermore, this infection is common to occur in immunocompromised individuals like in patients with leukemia, or lymphoma as they consumed corticosteroids or cytotoxic drugs which compromised their immunity. Antibiotic usage, diabetes, pregnancy, use of oral contraceptives, HIV, TB, and hypoparathyroidism are other conditions where candidiasis is found infect commonly. Patients with Xerostomia in which they are in absence of antifungal proteins like histatin and calprotectin are liable to get candidiasis more commonly than other individuals. Candidiasis is treated with antifungal agents but only a few classes of drugs are available to treat systemic or mucosal infections. The most common antifungal agents recommended for candidiasis are derivatives of azoles, Echinocandins, polyenes, nucleoside analogs like flucytosine and allylamines, and thiocarbamates are also known to have antifungal activities('Candidiasis - StatPearls - NCBI Bookshelf', n.d.; Schuster and Fisher 2022; Vanani *et al.*, 2019).

Today current pharmaceutical industries are facing two major challenges to deal with in the area of novel drug discovery and development. One of them is drug resistance which is developed by pathogenic fungi after the prolonged use of the antifungal drug. Drug resistance is becoming a significant threat not just to developing countries but also around the world. Infection with drug-resistant fungi can result in prolonged illness, the cost of secondary drugs, and also an increase in overall healthcare costs. According to the center for disease control and prevention antimicrobial resistance adds a 20-billion-dollar surplus in the united states only (Aslam *et al.*, 2018). Several factors are governing increase in the drug resistance. Misuse and overuse of drugs through self-medication, use of antibiotics with the belief to get relief in viral infections like a common cold, sale of poor-quality antibiotics over the counter, and administering of antibiotics by physicians when they are not required (Chaw *et al.,* 2018; Fletcher-Lartey *et al.,* 2016). Biological factors like mutations, gene transfer, plasmid transfer, and the use of drugs on plants and animals are also some of the factors that contribute the drug resistance (Sun *et al.,* 2019). Infection with drug resistance also affects the morbidity and mortality of the patients. Infection with resistance strains is known to cause serious health issues along with greatly increasing the chance of death. Currently, a total of 700,000 individuals loses their lives because of drug-resistant strains. Companies are spending millions of dollars within a long-time frame along with the involvement of hundreds of staff. If the drug developed through this process and resistance developed then will create a great loss to the company along with sentimental disappointment to the associated staff. It is great for today's pharmaceutical industries to find an alternate way that can reduce the time, effort, and cost associated with the drug discovery and development process (Zaman *et al.,* 2017).

DHFR is the key enzyme for folate metabolism. It catalyzes essential reactions for de novo glycine and purine synthesis and also for purine synthesis. It catalyzes NADPH-dependent reduction of 7,8 dihydrofolate to 5,6,7,8 tetrahydrofolate reductase a cofactor required for the metabolism of some amino acids, purines, and thymidine. The choice of antifungal protein as a target is always a complicated task. DHFR is present in both mammals and fungi, but there are major differences present in the binding site for the human and candida species. It makes a base for novel and advanced antifungal agents (Otzen *et al*., 2004). Virtual screening today has emerged in drug discovery as a powerful tool that uses computational techniques to screen a large pool of data for new hits which can later be tested experimentally. Virtual screening is not an alternative to *in vitro* and *in vivo* experiments but it helps in speeding up the discovery process. It helps in reducing the number of candidates to be tested experimentally along with it also helps in rationalizing the choice. Among all the methods available in virtual screening, QSAR is becoming more popular due to its speed and good hit rate. The method was developed almost fifty years ago (Hansch and Fujita 1964). Until now the method remains the first choice to build a mathematical model. Such a model attempts to create a correlation between structural features present within the active compound and biological or toxicological activity associated with it (Cherkasov *et al.,* 2014). Because of the high throughput screening techniques, a large pool of data is available for QSAR model development. Hundreds of free databases and web servers are available those are providing chemical data with proven experimental biological activity (Mueller *et al.,* 2012). Initially, the QSAR model development was limited to very few numbers of structural properties and their relation to the final activity. But sudden advancements in IT technology and cloud computing made it possible to count thousands of such properties associated with the chemical compounds. Software like Alvadesc has made it possible to count more than 5000 descriptors for a single compound within a fraction of the time. Alvadesc also provides to reduce the number of insignificant properties present in the generated dataset (Mauri, 2020). Programming languages like Python and R can handle such huge databases generated for the development of the model (Andresen, 2021; Millman and Aivazis 2011). These programming languages provide ready-to-use techniques to reduce the number of features or variables generated by descriptor calculation software. It is also possible to develop a QSAR model for many algorithms like Multiple linear regression, Support vector regression, Random Forest, Random Tree, and another categorical model at the same time for the same data. This makes it possible to choose a suitable and significant algorithm from all available. In the present investigation, we have developed a QSAR model based on the supervised machine learning approach. We have tried to maintain and consider all the OECD guidelines given to achieve the regulatory acceptance of the QSAR model (Laub, 1999). According to that, our model is associated with the defined endpoint, it is developed through four unambitious algorithms, and it is having a defined domain of applicability, along with goodness of fit, robustness, and predictivity. We also have tried to incorporate the mechanistic interpretation of the developed model with descriptors with their contribution to the model. For all the techniques performed with the generated dataset, we first carried out very stringent manual data curation to remove insignificant descriptors present (Neves *et al.,* 2018).

MATERIAL AND METHODS

Database collection. In the present work total of 281 compounds with their known IC50 values were collected from the ChEMBL web server (ID: CHEMBL2329). The database was obtained by searching the target section for *Candida albicans* as an organism. From all the listed targets, dihydrofolate reductase with their IC50 values was downloaded to the local computer. The structure database was downloaded into SMILE format. The SMILES were converted in to 2D sdf structure format using OpenBabel software. The compounds downloaded were having experimental activity in IC50 (nM). We have converted IC50 values in pIC 50 values (-log of IC 50) for model development. Before the development of the 2D-QSAR model, all the structures were carefully checked for significant model development. The complete workflow used in the work is represented in Fig. 1.

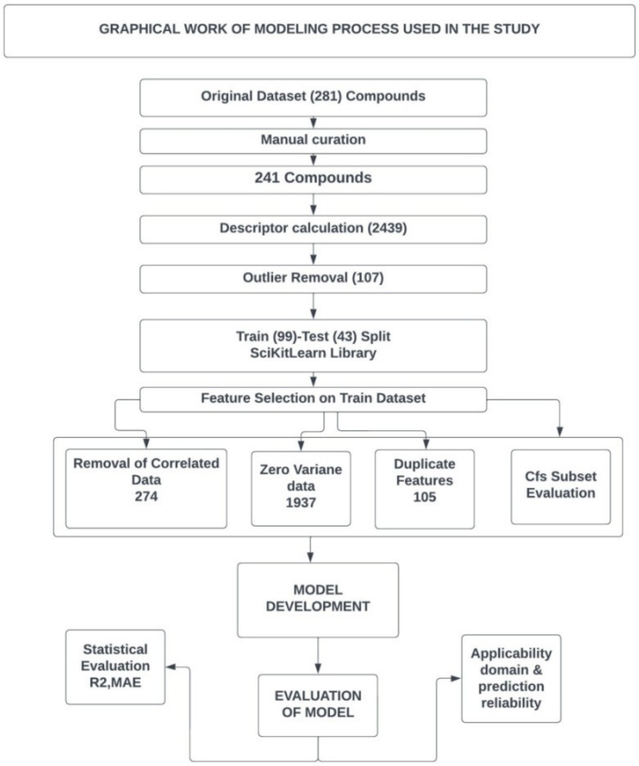

GRAPHICAL WORK OF MODELING PROCESS USED IN THE STUDY
Original Dataset (281) Compounds
Manual curation
241 Compounds
Descriptor calculation (2439)
Outlier Removal (107)
Train (99)-Test (43) Split
SciKitLearn Library
Feature Selection on Train Dataset
Removal of Correlated Data 274
Zero Variane data 1937
Duplicate Features 105
Cfs Subset Evaluation
MODEL DEVELOPMENT
Statistical Evaluation R2,MAE
EVALUATION OF MODEL
Applicability domain & prediction reliability

Fig. 1. Graphical representation of the Modelling process used in the study.

Descriptor calculation. In the present work, Alvadesc was used to count the descriptors for the selected compounds. Alvadesc can count more than 5000 types of descriptors including 2D and 3D types. Only 2D Descriptors of Constitutional, ring, functional, 2D atom pairs, and molecular properties. Topological indices, atom-type e-states, and atom-centered fragments types were considered for calculation. The database was thoroughly checked for "na" values manually using Python programming language and such entries were removed from progressing database (Mauri, 2020).

Outlier removal. Every QSAR model prediction power relies on the presence of outlier values. Outliers are values that are extremes from the average values of the database. In the present work, we have used the IQR method to remove the structural and functional outliers present (Tropsha, 2010).

Dataset division. It is always recommended to check the prediction power of the developed model on the dataset which is unknown to the model or the model has not seen it before. To, achieve this entire dataset was divided into training and testing datasets using the train test split function present in the scikit learn module of the python programming language. We have used a 70 % and 30 % ratio split, where 70 % dataset was used in the training set while 30 % was used for a testing set (Kumar and Roy 2020; Pedregosa *et al*., 2011).

Feature selection. A feature or variable is a column heading with its numerical value present in the dataset. In the present work, we have calculated 2439 descriptors for each of the compounds present in the dataset. To, train the optimal model, we should use only optimal features. If we have too many insignificant features, the model can learn from noise. The prediction power of any model depends on the number and overall significance of each descriptor or feature contributing to the final model (Barcelos *et al.,* 2021; Hira and Gillies 2015).

Constant columns. Columns that do have values that are nearly constant or constant are known as constant columns. The output or dependent variable won't be significantly impacted by such columns or variables. Sklearn has a function called variance threshold that may be used to eliminate these constant columns. To exclude the zero variance columns from the dataset, we used a default setting.

Duplicate Features. Duplicate features or variables from the developing database were removed to reduce the number of descriptors for the QSAR model. I used the sklearn duplicate and convert function to get rid of these redundant features.

Correlation. The output or target features should be connected to the independent features, but independent features shouldn't be correlated to one another. This is called as multicollinearity of the features. Multicollinearity is a phenomenon that significantly impairs the performance of the model. A high correlation between independent features is linearly dependent and hence affects the target in the same way. Therefore, if there is a strong correlation between two features, one of them should be discarded. To do this, a cut of 0.9 was set, and features with correlations higher than 0.9 were eliminated from the selection. Next, we used CfsSubsetEval together with locally predictive attributes for significant descriptor selection. CfsSubsetEval produces subsets of the features that are highly correlated to the class/activity at the same time they have low intercorrelation. Locally predictive attribute classifies locally predictive attributes and iteratively add attributes with the highest correlation with the class as long as there is not already an attribute in the subset that has a higher correlation with the class (Hall *et al.*, 2009).

QSAR model development. We have used supervised machine learning techniques to develop a QSAR model based on the final set of features selected in the previous steps. Total of four machine learning algorithms namely Multiple linear regression, Support vector regression, Random Forest regression, and Random tree were used to produce the final QSAR prediction model (Asha Kiranmai and Jaya Laxmi 2018; Breiman, 2001; Pandis, 2016a; Shevade *et al.,* 1999).

Multiple linear regression. It is the regression model that establishesa relationship between a dependent variable and one or more independent variables. In the present work, the M5 method for regression was adopted, which removes attributes based on the smallest standardized coefficient until no improvement is observed in the estimate of error given by the Akaike information criterion (Pandis, 2016b; Uyanık and Güler 2013).

Support vector regression. To, predict the continuous-discrete values, we employed a support vector regression approach. SVR and SVMs both operate on the same principles. The PolyKernel kernel was used to create the best-fit line in SVR, termed the hyperplane (Smola and Schoelkopf 1998).

Random forest. Another supervised machine learning algorithm Random Forest was used to finding out the prediction power of the QSAR model. Random forests are learning methods for classification and regression tasks. It works by constructing multiple decision trees. For the regression task in the RF, the mean or average prediction of individual trees was used(Breiman, 2001; Trinh *et al.,* 2022).

Random tree. Another supervised machine learning tree-based algorithm called random tree for examining the prediction capability of the model was used. Recursive portioning to split the data and finding the best split using a reduction in impurity index was used to evaluate the model performance (Asha Kiranmai and Jaya Laxmi 2018).

Evaluation of model. The QSAR model developed by four different algorithms was checked for its robustness and predictive power using several statistical validation criteria. For the present investigation, we have used both internal and external validation tools to find out the significance of the model. The models were evaluated for Pearson's correlation coefficient, coefficient of determination, mean absolute error, and root mean squared error. The model's performance was also evaluated on the external dataset or test dataset to find out the prediction power of the model. Furthermore, we also have applied the Y-randomization or Y-Scrambling method to check the robustness of the model (Christopher *et al.,* 2007; Rücker *et al.,* 2007a). This test is used to check whether the model is not produced by chance of prediction or not. It was performed by training the model first on the original dataset and the performance of the model was noted. In the next step, we shuffled the target or activity column to replace the correct target-feature pair with the incorrect one, and the performance of the model was noted. This scrambling was done 50 times and 50 QSAR models with their performance were noted and evaluated.

Applicability Domain. It is the space of a mathematical model where we can apply our model with confidence. For new compounds like test compounds to be inside the applicability domain and they have to be sufficiently similar to training dataset compounds used to develop the model. we have determined the applicability domain using MLR plus validation tool available at (https://dtclab.webs.com/software-tools) (Roy *et al.*, 2015).

Prediction Reliability. It always needs to predict the experimental value of the unknown compounds without performing any wet lab experiments. It is a significant application of the QSAR Model to predict the target value with the help of structural descriptors calculated for the unknown dataset. One needs to have computational measures to check the reliability of the prediction made by the model. We have used the prediction reliability indicator tool available at https://dtclab.webs.com/software-tools. The tool was used to describe the quality of prediction in three scores "good", "bad" and "moderate" based on the criterion explained earlier (De *et al.,* 2022).

RESULT AND DISCUSSION

We have developed the QSAR model to predict the inhibitory activity of the compounds against the dihydrofolate reductase enzyme present in the *Candida albicans* fungus. We have used Chembl database web server to find out the compounds with proven inhibitory activity against the selected target. A total of 281 compounds were downloaded to the local computer for their use in the development of the QSAR model (Supplementary_1). We have manually curated the database for removing false IDs, missing values, or zero values (Supplementary_2). Structures were present in the SMILE format and were converted into a structure data format (sdf) (Supplementary_3 and 4). We have calculated around 2439 descriptors of the described types in the materials and method sections. Descriptors are the chemical characteristics of the compounds in numerical form (Supplementary_5). The progressing database was then subjected to remove outliers. Outlier is defined as a data point which is having inconsistent values in comparison to other data point values. The presence of an outlier in the dataset can significantly influence the power of the model. Detection of an outlier was performed using the interquartile range value. This method is widely accepted to remove outliers from a huge dataset. We have divided the dataset into four equal segments or quartiles and the distance between the quartiles was used to determine the IQR. We have defined the outliers as the values that fall outside of 1.5* IQR below Q1 and 1.5* IQR above Q3. A total of 107 structural and functional outliers were found and were removed from the progressing database (Supplementary_6). The remaining 142 compounds were split in to training and testing databases with a random split of 70 % and 30 %. This was done to test the developed model on the unseen data with known activity values. We have used feature selection or sometimes called dimensionality reduction techniques to improve the estimator accuracy scores and to boost their performance on a high dimensional database. We have used the variance threshold approach from the sci-kit-learn 1.1.2 library to remove the descriptors with zero variance present within. As these cannot impart significant values to the final model prediction power. A total of 1937 constant columns were found and were removed to reduce the number of descriptors (Supplementary_7). Similarly, 106 duplicate columns were removed using the duplicate transform function available in the python sci-kit-learn 1.1.2 library (Supplementary_8). We have manually defined a function to remove the features with multicollinearity. Features with high correlation are linearly dependent and have the same effect on the target. To remove such features, we kept a cut-off of 0.9 *i.e.,* features with 90 % similarity were removed from the dataset. In this work, we have found 274 highly correlated features that were removed from the progressing database (Supplementary_9). Final feature selection from the existing database was done using CfsSubsetEval along with locally predictive attributes. CfsSubsetEval is a correlation-based feature selection technique. It was used through the select attribute function in the Weka machine learning program. Using CfsSubsetEval we have identified a total of 54 features that can be used to develop a QSAR model for the prediction of inhibitory activity against a selected target of *Candida albicans* fungus. We have used a supervised machine-learning technique to build the model. This technique takes the advantage of existing data present in the training dataset along with known activity values. A total of four different algorithms MLR, SVR, RF, and RT were used to build the model. All the statistical parameters used to validate the model performance are given in (Table 1).

Table 1: Statistical Values for QSAR models.

Algorithm	Statistical Parameters	
MLR	**R^2**	**MAE**
Training set	94.44	0.2533
Test Set	0.61	0.66
SVR		
Training set	0.93	0.21
Test Set	0.84	0.5
RF		
Training set	0.95	0.21
Test Set	0.86	0.39
RT		
Training set	0.95	0.15
Test Set	0.78	0.5

From the values represented in the table, it is observed that R^2 values for all the models are above 90%, and maximum R^2 is observed for RF and RT algorithms. The model was developed with 54 significant descriptors chosen from a large pool of 2439 descriptors calculated initially for compounds. The equation produced by the model can be used to predict the inhibitory activity of the compounds. Furthermore, the scatter plot produced between observed and predicted values also confirms the closeness and indicates the robustness and stability of the model (Figs. 2-5).

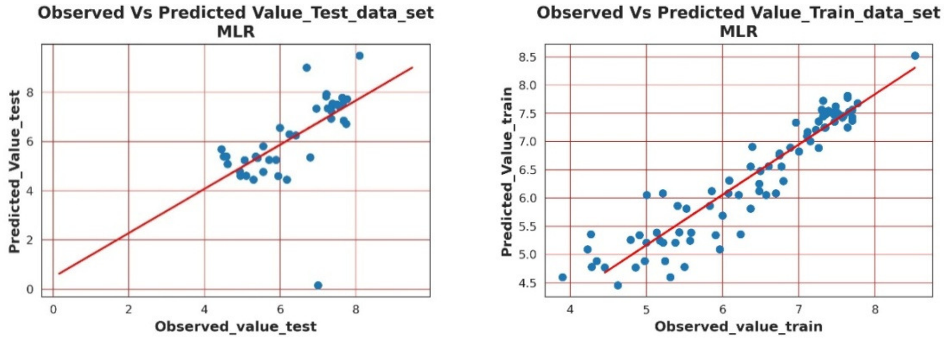


Fig. 2. Scatter plot of observed and predicted values for training and testing dataset from the developed Multiple linear regression model.

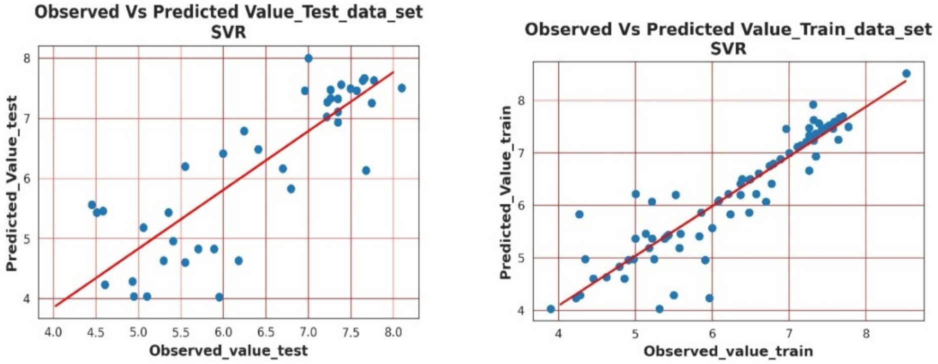


Fig. 3. Scatter plot of observed and predicted values for training and testing dataset from the developed Support Vector regression model.

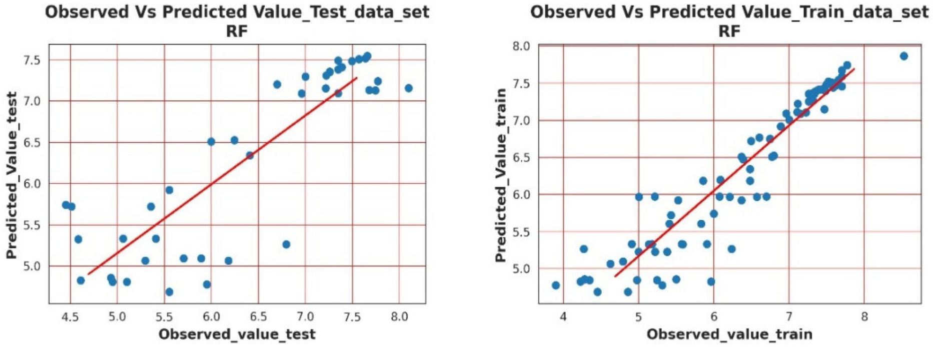


Fig. 4. Scatter plot of observed and predicted values for training and testing dataset from the developed Random Forest model.

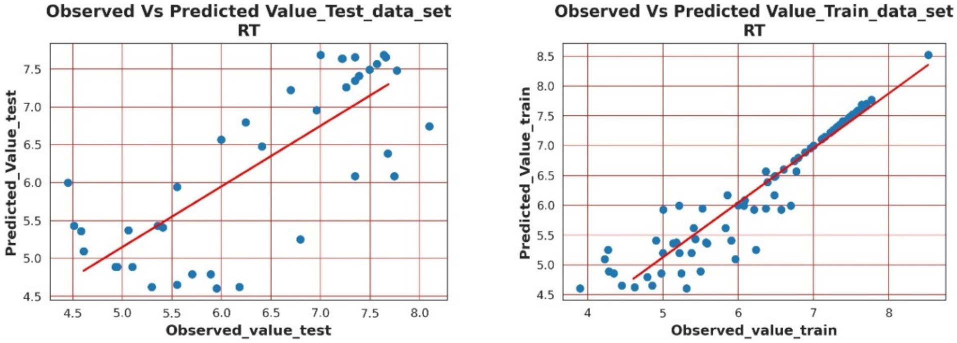


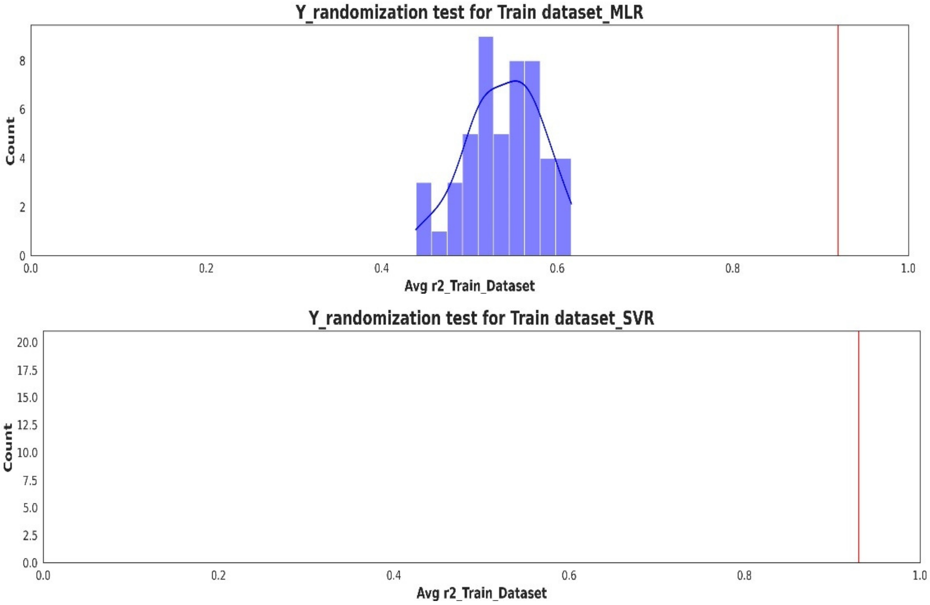
Fig. 5. Scatter plot of observed and predicted values for training and testing dataset from the developed Random trees model.

The MAE for the prediction of both train and test datasets is also not very high and it strongly suggests the applicability of the build model for this class of compounds.

We also carried out the Y-randomization test to check whether the model is produced by chance or not. The randomization of the activity values was carried out 50 times and for each model produced R^2 values were calculated. The values are represented in the table and from the table, it is clear that the model performance for each algorithm is worse than before randomly shuffling the activity values (Rücker *et al.*, 2007b). The results indicate that underlying relationship between structural features and activity or target variable (Table 2).

Table 2: R^2 values for the actual and randomized model.

Algorithm	Actual R^2	Y-randomized R^2
MLR	0.92	0.61
SVR	0.93	-0.06
RF	0.95	0.62
RT	0.95	0.68



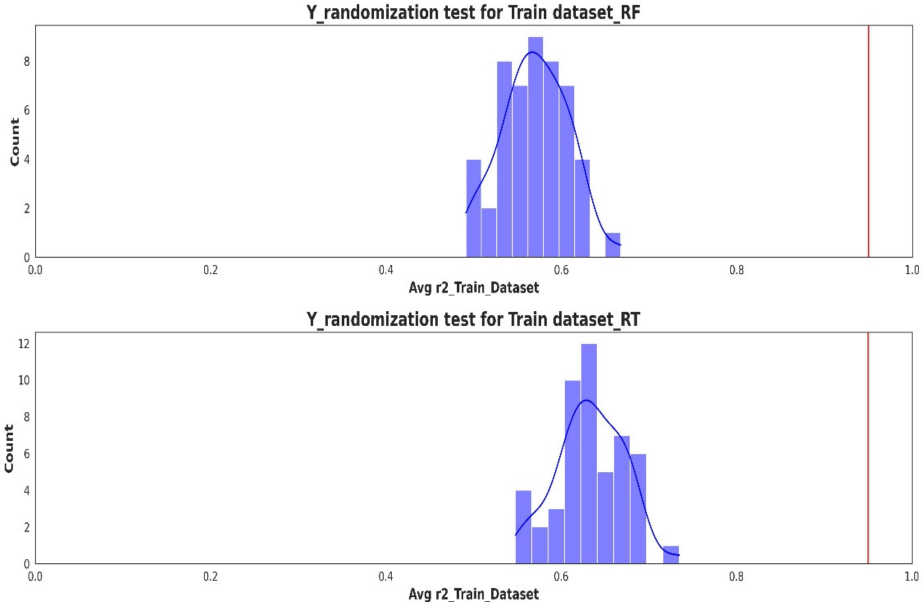


Fig. 6. The performance of all four QSAR models built with randomized data is inferior to that of the QSAR model developed with the non-permuted data set. (a) Y-randomization test for train dataset for MLR; (b) Y-randomization test for train dataset for SVR; (c) Y-randomization test for train dataset for RF; (d) Y-randomization test for train dataset for RT.

We performed external validation on the 43 compounds that were kept separate before the development of the model. In all the models as shown in table No:1 test dataset R^2 values and their respective MAEs were found to be under the acceptable range. In addition, we performed an applicability domain test to define the hypothetical domain in the chemical space for the QSAR model. From the analysis conducted for the test compound, we found that there are no compounds in the test dataset available outside the applicability domain area (Supplementary_10).

The developed model was used to predict the activity of the test dataset, kept hidden during the process of model development. The model was able to predict the activity of the target in the test dataset with confidence as was suggested by the prediction reliability indicator (Supplementary_11). The equation developed for MLR, SVR, and RF regression are attached as supplementary files. The descriptors present in the equation are representing negative and positive values. These indicate the contributions made by them to the final activity of the compounds. Positive contribution indicates descriptors impart positivity to the activity and negative impart negative to the activity. The supplementary file attached clearly describes the type of descriptor along with its contribution to the final QSAR model.

CONCLUSION

To find structural significant properties for a compound to show inhibitory activity against dihydrofolate reductase enzyme, the QSAR model was developed. We used 281 compounds with known and proven IC50 values for the model development. Feature selection and supervised machine learning techniques were used to develop the model. The statistical validation results obtained showed good predictive ability based on both internal and external parameters for validation. The model developed can be used to screen a large pool of compounds for their inhibitory activity and features obtained through studies can be used to design novel inhibitors. The list of the inhibitors screened from the virtual screening can be used for the further development of novel inhibitors.

FUTURE SCOPE

The developed QSAR model can come out as time saver and cost-effective tool to screen the inhibitors against dihydrofolate reductase enzyme from *Candida albicans* fungi.

Conflict of Interest. None.

List of Abbreviations

ML-Machine learning
MLR-Multiple Linear Regression
SVR-Support Vector Regression
RF-Random Forest
RT-Random Tree
MAE-Mean Absolute error
QSAR-Quantitative Structure-Activity Relationship
DHFR-Dihydro Folate Reductase
IQR-Inter Quartile Range
OECD- Organization for Economic Co-Operation and Development

REFERENCES

Andresen, M. A. (2021). R (Statistical Software). In *The Encyclopedia of Research Methods in Criminology and Criminal Justice: Volume II: Parts 5-8*. https://doi.org/10.1002/9781119111931.ch167

Asha Kiranmai, S. and Jaya Laxmi, A. (2018). Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy. *Protection and Control of Modern Power Systems*, *3*(1).

Aslam, B., Wang, W., Arshad, M. I., Khurshid, M., Muzammil, S., Rasool, M. H. and Baloch, Z. (2018). Antibiotic resistance: a rundown of a global crisis. *Infection and Drug Resistance*, *11*, 1645–1658. https://doi.org/10.2147/IDR.S173867

Barcelos, H. C., Mendoza, M. R., & Moreira, V. P. (2021). *Identifying and Fusing Duplicate Features for Data Mining*. https://doi.org/10.5753/sbbd.2020.13631

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Candidiasis – Stat Pearls - NCBI Bookshelf. (n.d.). Retrieved 27 September 2022, from https://www.ncbi.nlm.nih.gov/books/NBK560624/

Chaw, P. S., Höpner, J., & Mikolajczyk, R. (2018). The knowledge, attitude and practice of health practitioners towards antibiotic prescribing and resistance in developing countries—A systematic review. *Journal of Clinical Pharmacy and Therapeutics*, *43*(5), 606–613. https://doi.org/10.1111/JCPT.12730

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M. and Tropsha, A. (2014). QSAR modeling: Where have you been? Where are you going to? *Journal of Medicinal Chemistry*, Vol. 57. https://doi.org/10.1021/jm4004285

Christopher Rucker and Gerta Rucker, M. M. (2007). Y-Randomization-A useful tool in QSAR Validation, or Folklore? *Journal of Chemical Information and Modeling*, *47*(6).

De, P., Kar, S., Ambure, P., & Roy, K. (2022). Prediction reliability of QSAR models: an overview of various validation tools. *Archives of Toxicology*, Vol. 96. https://doi.org/10.1007/s00204-022-03252-y

Fletcher-Lartey, S., Yee, M., Gaarslev, C. and Khan, R. (2016). Why do general practitioners prescribe antibiotics for upper respiratory tract infections to meet patient expectations: A mixed methods study. *BMJ Open*, *6*(10). https://doi.org/10.1136/BMJOPEN-2016-012244

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, *11*(1). https://doi.org/10.1145/1656274.1656278

Hansch, C. and Fujita, T. (1964). ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *Journal of the American Chemical Society*, *86*(8), 1616–1626. https://doi.org/10.1021/JA01062A035/ASSET/JA01062A035.FP.PNG_V03

Hira, Z. M. and Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, *2015*. https://doi.org/10.1155/2015/198363

Kumar, V. and Roy, K. (2020). Development of a simple, interpretable and easily transferable QSAR model for quick screening antiviral databases in search of novel 3C-like protease (3CLpro) enzyme inhibitors against SARS-CoV diseases. *SAR and QSAR in Environmental Research*, *31*(7). https://doi.org/10.1080/1062936X.2020.1776388

Laub, J. A. (1999). Assessing the servant organization; Development of the Organizational Leadership Assessment (OLA) model. Dissertation Abstracts International. *Procedia - Social and Behavioral Sciences*, *1*(2).

Mauri, A. (2020). alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. In *Methods in Pharmacology and Toxicology* (pp. 801–820). Humana Press Inc. https://doi.org/10.1007/978-1-0716-0150-1_32

Millman, K. J. and Aivazis, M. (2011). Python for scientists and engineers. *Computing in Science and Engineering*, Vol. 13. https://doi.org/10.1109/MCSE.2011.36

Mueller, R., Dawson, E. S., Meiler, J., Rodriguez, A. L., Chauder, B. A., Bates, B. S. and Lindsley, C. W. (2012). Discovery of 2-(2-Benzoxazoyl amino)-4-Aryl-5-Cyanopyrimidine as Negative Allosteric Modulators (NAMs) of Metabotropic Glutamate Receptor5 (mGlu 5): From an Artificial Neural Network Virtual Screen to an In Vivo Tool Compound. *ChemMedChem*, *7*(3). https://doi.org/10.1002/cmdc.201100510

Neves, B. J., Braga, R. C., Melo-Filho, C. C., Moreira-Filho, J. T., Muratov, E. N. and Andrade, C. H. (2018). QSAR-based virtual screening: Advances and applications in drug discovery. *Frontiers in Pharmacology*, *9*(NOV), 1275. https://doi.org/10.3389/FPHAR.2018.01275/BIBTEX

Otzen, T., Wempe, E. G., Kunz, B., Bartels, R., Lehwark-Yvetot, G., Hänsel, W. and Seydel, J. K. (2004). Folate-Synthesizing Enzyme System as Target for Development of Inhibitors and Inhibitor Combinations against *Candida albicans* - Synthesis and Biological Activity of New 2,4-Diaminopyrimidines and 4′-Substituted 4-Aminodiphenyl Sulfones. *Journal of Medicinal Chemistry*, *47*(1). https://doi.org/10.1021/jm030931w

Pandis, N. (2016a). Multiple linear regression analysis. *American Journal of Orthodontics and Dentofacial Orthopedics*, Vol. 149. https://doi.org/10.1016/j.ajodo.2016.01.012

Pandis, N. (2016b). Multiple linear regression analysis. *American Journal of Orthodontics and Dentofacial Orthopedics*, Vol. 149. https://doi.org/10.1016/j.ajodo.2016.01.012

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. and Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*.

Roy, K., Kar, S., &Ambure, P. (2015). On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems*, *145*. https://doi.org/10.1016/j.chemolab.2015.04.013

Rücker, C., Rücker, G. and Meringer, M. (2007a). Y-randomization and its variants in QSPR/QSAR. *Journal of Chemical Information and Modeling*, *47*(6). https://doi.org/10.1021/ci700157b

Rücker, C., Rücker, G. and Meringer, M. (2007b). Y-randomization and its variants in QSPR/QSAR. *Journal of Chemical Information and Modeling*, *47*(6). https://doi.org/10.1021/ci700157b

Schuster, J. E. and Fisher, B. T. (2022). Candidiasis. *Pediatric Transplant and Oncology Infectious Diseases*, 195-205.e3. https://doi.org/10.1016/B978-0-323-64198-2.00035-X

Shevade, S. K., Keerthi, S. S., Bhattacharyya, C. and Murthy, K. R. K. (1999). Improvements to the SMO Algorithm for SVM Regression. *IEEE Transactions on Neural Networks*.

Smola, A. J. and Schoelkopf, B. (1998). *A tutorial on support vector regression*.

Sun, D., Jeannot, K., Xiao, Y. and Knapp, C. W. (2019). Editorial: Horizontal gene transfer mediated bacterial antibiotic resistance. *Frontiers in Microbiology*, *10*(AUG). https://doi.org/10.3389/FMICB.2019.01933

Trinh, T. X., Seo, M., Yoon, T. H. and Kim, J. (2022). Developing random forest based QSAR models for predicting the mixture toxicity of TiO_2 based nano-mixtures to Daphnia magna. *NanoImpact*, *25*. https://doi.org/10.1016/j.impact.2022.100383

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*, Vol. 29. https://doi.org/10.1002/minf.201000061

Uyanık, G. K. and Güler, N. (2013). A Study on Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences*, *106*. https://doi.org/10.1016/j.sbspro.2013.12.027

Vanani, A. R., Mahdavinia, M., Kalantari, H., Khoshnood, S. and Shirani, M. (2019). Antifungal effect of the effect of *Securigera securidaca* L. Vaginal gel on *Candida* species. *Current Medical Mycology*, *5*(3). https://doi.org/10.18502/cmm.5.3.1744

Zaman, S. bin, Hussain, M. A., Nye, R., Mehta, V., Mamun, K. T. and Hossain, N. (2017). A Review on Antibiotic Resistance: Alarm Bells are Ringing. *Cureus*. https://doi.org/10.7759/CUREUS.1403

How to cite this article: Sharav Desai, Vijay K. Patel, Ankita S. Patel and Jaini Patel (2023). Development and Validation of an Easily Interpretable QSAR Model for Inhibitory Activity Prediction against Dihydrofolate Reductase from *Candida albicans*. *Biological Forum – An International Journal, 15*(1): 505-513.