JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DEVELOPMENT OF A COMPUTATIONAL QSAR MODEL FOR PREDICTING ANTI-TUBERCULAR ACTIVITY OF CHEMICAL COMPOUNDS

Dr.M.Vijaya Bhargavi[1]*, Chencharam Tulasi[2], M.Varalaxmi[3], M.Shivani[4], M.Deekshitha[5], S.Lahari[6]

*[1]*Associate Professor, [2]Assistant Professor, [3-6]Research Scholar*

[1-6] Department of Pharmaceutical Chemistry*

RBVRR Women's College of Pharmacy, Hyderabad, India

ABSTRACT

Tuberculosis (TB), caused by *Mycobacterium tuberculosis*, remains a global health challenge due to the emergence of drug-resistant strains and the limitations of existing therapeutic agents. In the search for novel anti-tubercular compounds, computational approaches such as Quantitative Structure-Activity Relationship (QSAR) modelling have emerged as powerful tools. This study focuses on the development of computational QSAR models to predict the antitubercular activity of diverse chemical compounds. A curated data-set of compounds with known activity was analyzed, and molecular descriptors were calculated. Various statistical and machine learning techniques, including Multiple Linear Regression (MLR), Partial Least Squares (PLS), kNN (k-Nearest Neighbor), SVM(Support Vector Machine)/SVR(Support Vector Regression) were employed to generate predictive models. The models were validated using internal and external validation techniques, including cross-validation and external test sets. The results demonstrate that the developed QSAR models can reliably predict the biological activity of novel compounds and assist in the rational design of potent anti-tubercular agents. This computational strategy offers a cost-effective and time-saving approach to support anti-TB drug discovery and accelerate lead identification.

KEYWORDS: Anti-tubercular, Computational QSAR, PubChem, Chem master, IC50, PIC5O.

INTRODUCTION

Tuberculosis (TB) continues to be a major global health threat, claiming millions of lives annually and posing significant challenges due to the rise of drug-resistant strains of *Mycobacterium tuberculosis*. Despite ongoing efforts, the development of new anti-tubercular agents has not kept pace with the urgent clinical need. Traditional drug discovery methods are often time-consuming, costly, and resource-intensive. As a result, there is a growing interest in computational approaches that can accelerate the identification and optimization of novel therapeutic candidates.

QSAR (Quantitative Structure–Activity Relationship)

Quantitative structure-activity relationship (QSAR) is a computational modeling method for revealing relationships between structural properties of chemical compounds and biological activities. QSAR modeling is essential for drug discovery, but it has many constraints. Ensemblebased machine learning approaches have been used to overcome constraints and obtain reliable predictions. Ensemble learning builds a set of diversified models and combines them. However, the most prevalent approach random forest and other ensemble approaches in QSAR prediction limit their model diversity to a single subject.

Types of QSAR Modelling 1. Linear QSAR Modelling

Linear QSAR models identify a direct (linear) relationship between molecular descriptors and a compound's biological or physicochemical activity.

Examples: Simple Linear Regression (SLR), Multiple Linear Regression (MLR)

MLR involves using multiple descriptors to predict a property or activity, with the equation:

$Y = a + bx$

Advantages

Easy to interpret

Suitable for initial analysis

Disadvantages

Cannot model complex, non-linear relationships Risk of over-fitting

with too many descriptors.

2. Non-Linear QSAR Modelling

Non-linear QSAR uses advanced computational methods to model complex, non-linear relationships between descriptors and activity.

Examples: Artificial Neural Networks (ANNs), Support Vector Machines (SVMs)

Advantages

Better accuracy and predictive power

Effective for complex chemical data **Disadvantages**

Less transparent ("black-box" models)

Requires more data and computational resources.

DIMENSIONS OF QSAR:

DIMENSIONS	**DESCRIPTORS, DATA**
1D-QSAR	Affinity correlates with global molecular properties of ligands ($pK_{a,}$ Log P, PSA etc.)
2D-QSAR	Affinity correlates with structural patterns (connectivity, 2-D pharmacophore etc.) without consideration of an explicit 3D representation of these properties.

3D-QSAR	Affinity correlates with the 3-Dimensional structure of the ligands.
4D-QSAR	Ligands are represented as am ensemble of potential binding modes (different conformations, orientations, protonation states, tautomer, and stereoisomers.)
5D-QSAR	4D-QSAR + Explicit representation of different induced-fit models.
6D-QSAR	5D-QSAR + Different solvation scenarios.

SIGNIFICANCE OF QSAR

1. Drug Discovery and Development: Prediction of Biological Activity, Cost and Time Efficiency, Optimization of Leads.

2. Environmental and Toxicological Assessment: Predicts Toxicity, Regulatory Use.

3. Mechanistic Insight: Understanding Structure-Activity Relationships, Design of Safer Chemicals.

4. Integration with AI and Machine Learning. QSAR models are increasingly enhanced by AI/ML techniques, improving their accuracy and allowing for large-scale virtual screening of compound libraries.

5. Economical and Ethical Benefits.

   - ➢ Reduces the reliance on animal testing.
   - ➢ Minimizes the cost and resources needed for lab experiments.
   - ➢ Supports ethical guidelines and sustainability in chemical research.

DATABASES :

PubChem

PubChem is a free, public chemical database maintained by the National Center for Biotechnology Information (NCBI), which is part of the U.S. National Institutes of Health (NIH). It provides information on the chemical structures, properties, biological activities, and more for millions of chemical substances.

SIGNIFICANCE

□ PubChem is a cornerstone resource for QSAR modelling, providing reliable chemical and biological data to build, validate, and test predictive models—fuelling advancements in drug discovery, toxicology, and cheminformatics·

CHEM MASTER

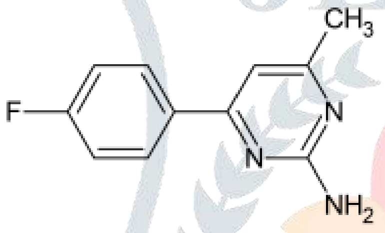
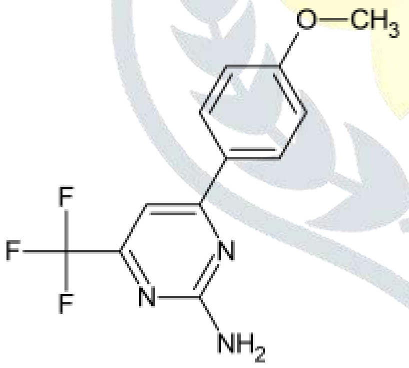
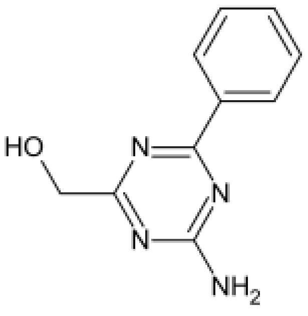
CHEMMASTER refers to a specialized software or platform (depending on context, it may also be a custom tool or acronym) used for chemical informatics, including Quantitative Structure– Activity Relationship (QSAR) modelling. QSAR is a computational method that correlates chemical structure with biological activity using statistical tools and machine learning algorithms**.**

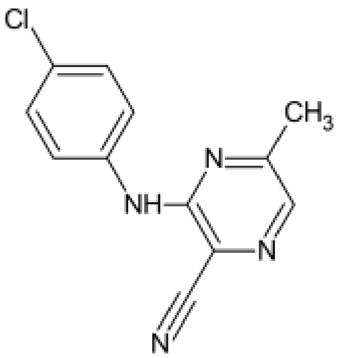
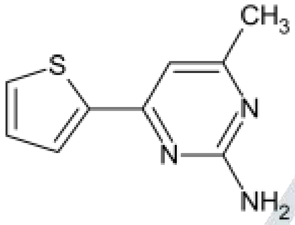
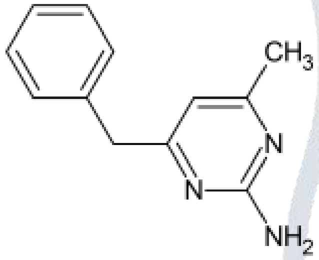
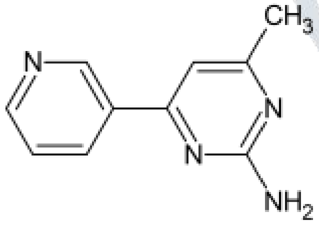
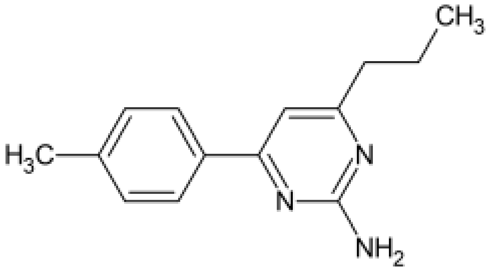
SIGNIFICANCE

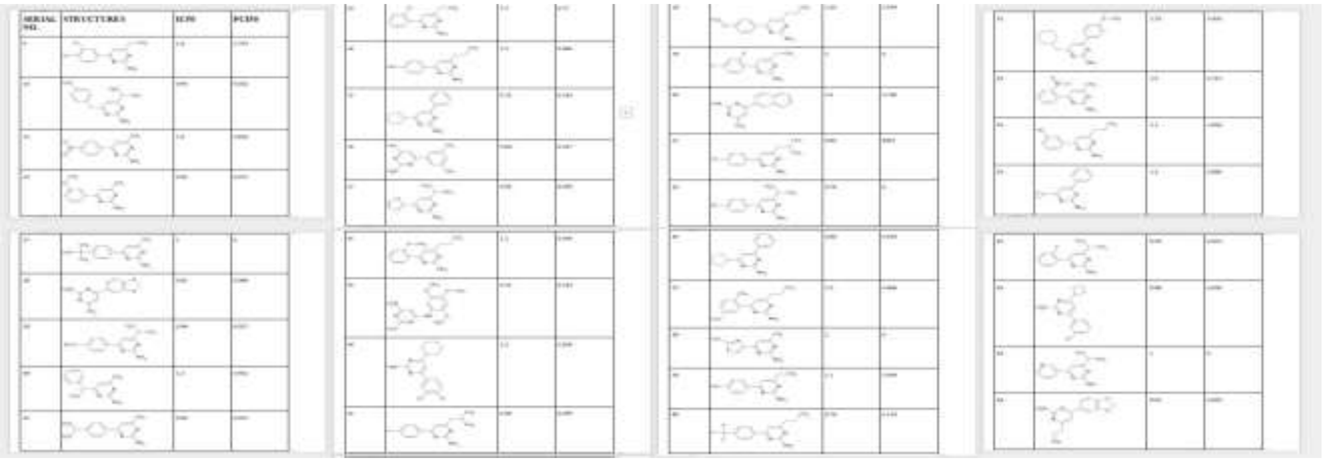
 Chem Master is a powerful cheminformatics and QSAR modelling tool that plays a significant role in streamlining the drug discovery and molecular modelling process. It allows researchers to efficiently perform tasks such as molecular descriptor calculation, substructure searching, fingerprint generation, and conformer creation within a single platform.The software supports both regression and classification models, incorporating advanced methods such as Hologram-QSAR and Auto-QSAR. It enables users to develop accurate predictive models for biological activity or chemical properties with ease. A key advantage is its ability to streamline QSAR model development by automating essential steps like feature selection, model training, and validation. This automation boosts efficiency and ensures consistent, reproducible modeling.

EXPERIMENTALMETHODOLOGY

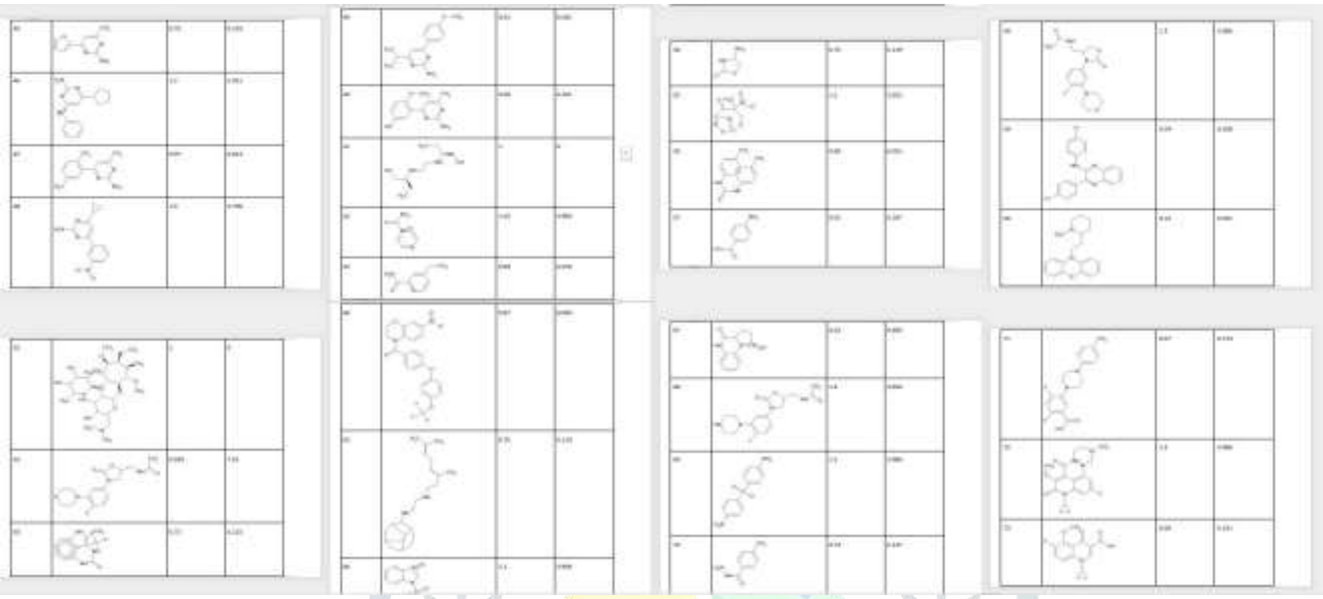
Structures, IC50 and PIC50 values of the selected compounds from the various articles:

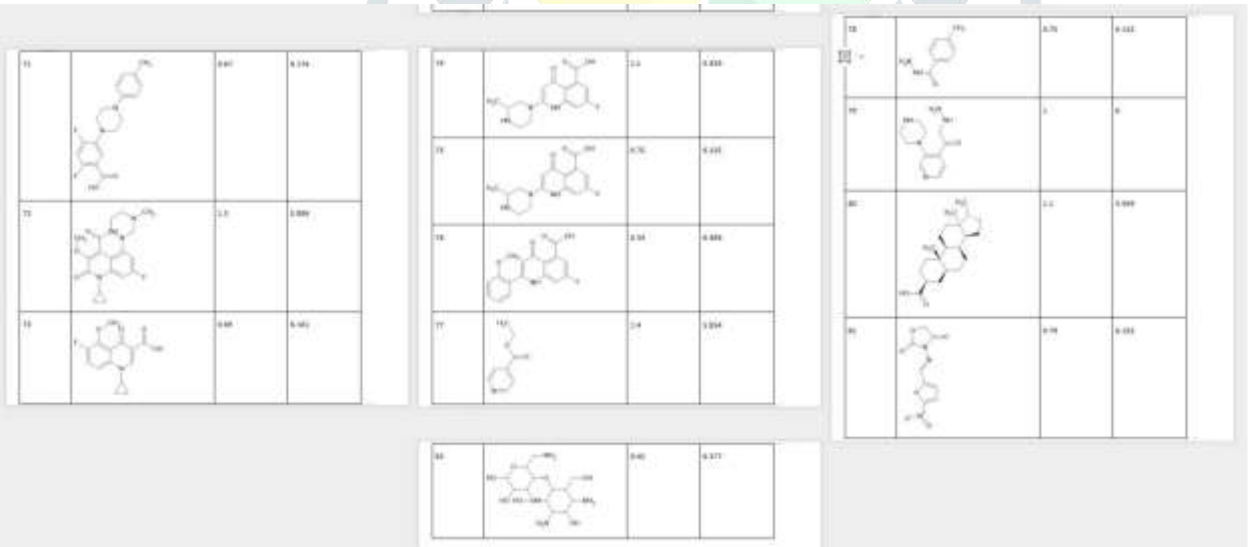
SERIAL NO.	**STRUCTURES**	**IC50**	**PCI50**
1		1.2	5.9
2		0.9	6.046
3		2.3	5.683

		1	6
		0.6	6.222
		1.5	5.824
		0.75	6.125
		1.1	5.959



SERIAL NO.	STRUCTURES	IC50	PIC50





VARIOUS STEPS INVOLVED IN QSAR MODELLING

Step 1:- Using keywords or PubChem ID’S of the compounds which act as Anti-tubercular agents are retrieved from the PubChem Database.

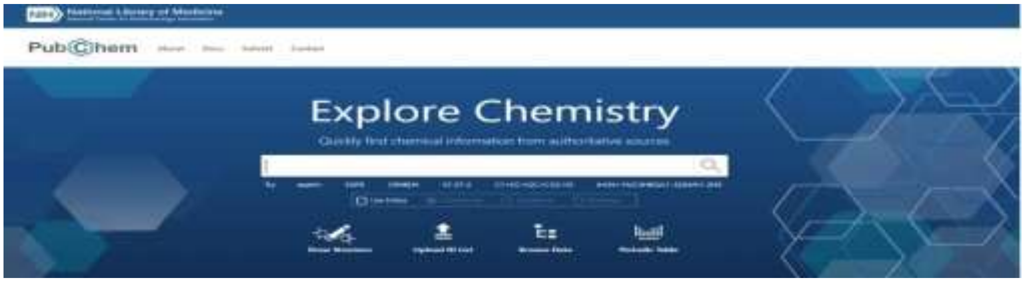


Fig no. 1 **PubChem database search bar.**

Step 2:- Explore and search.

Step 3:- Download the active substance database in the SDF format (Excel) from the data table.

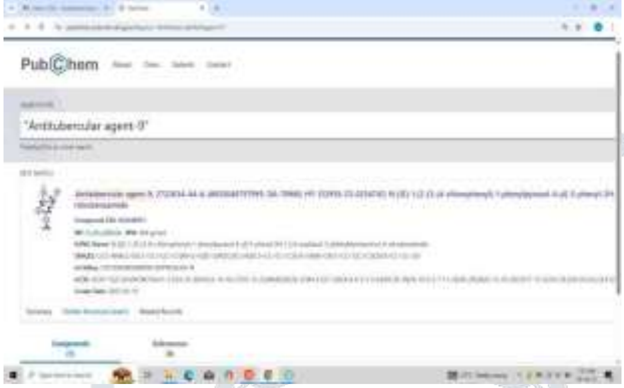


Fig no. 2 **Pubchem database search for Anti-tubercular drugs with IC50 Value.**

Fig no. 3 **Download of the data in SDF format.**

Step 4:-Prepare a datasheet containing only SMILES Notation and IC50 Values.

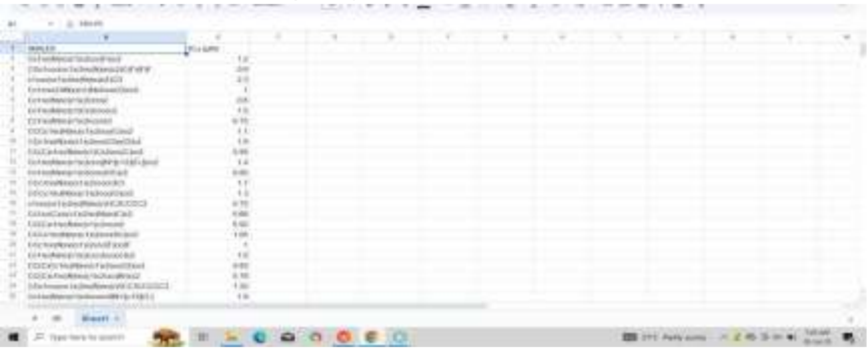


Fig no. 4 **A excel data-sheet of SMILES and IC50 values.**

Step 5:-Convert IC50 Values into PIC50 Values using the converter.

IC50 VALUE :- Half-maximal inhibitory concentration (IC50) is the most widely used and informative measure of a drug's efficacy. It indicates how much drug is needed to inhibit a biological process by half, thus providing a measure of potency of an antagonist drug in pharmacological research.

PIC50 VALUE :- PIC50 is the negative logarithm of IC50 in molar concentration.

PIC50 = 6-LOG(IC50)

Step 6:- In Chem Master by using downloaded SMILES generate structures one after the other

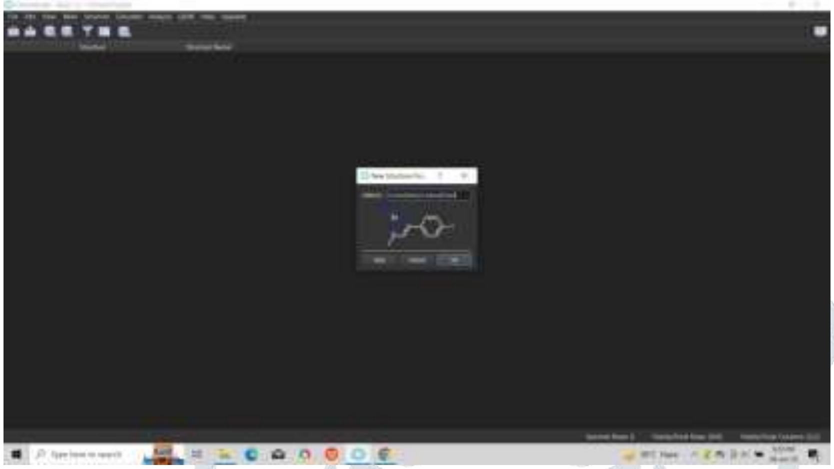


Fig no. 5 **Generating structures from SMILES in Chem Master.**

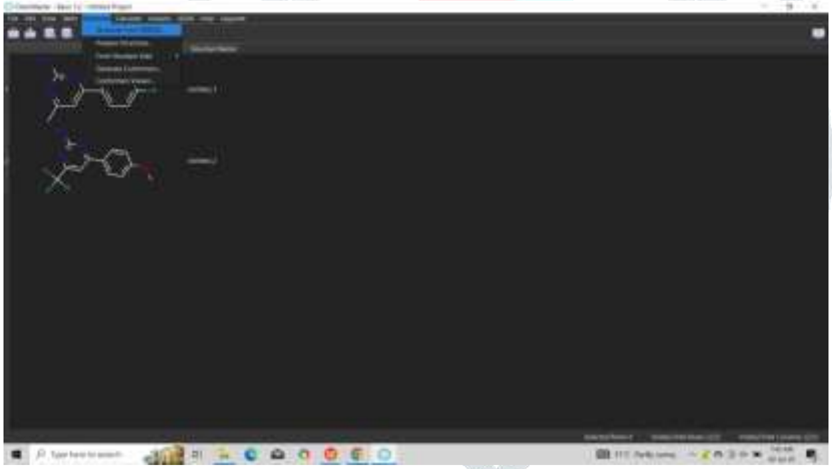


Fig no. 6 **Generating Structures one after another from SMILES.**

Step 7:- Import IC50 and PIC50 values.

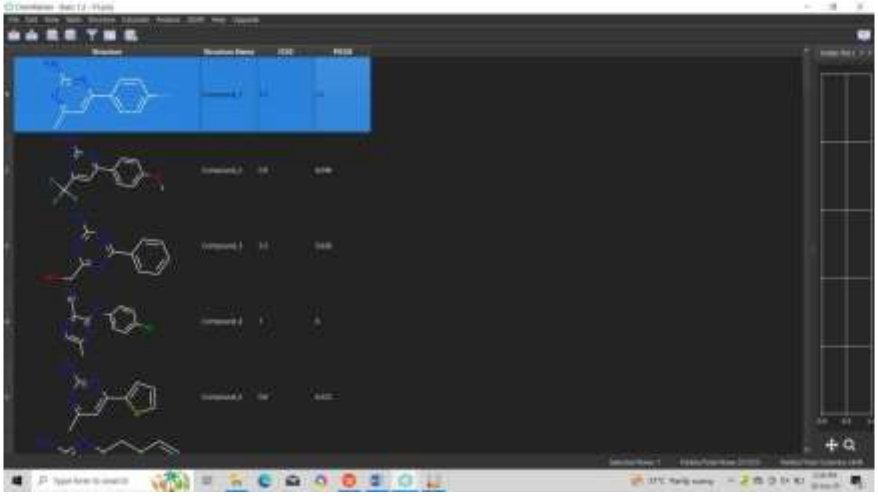


Fig no. 7 **Importing IC50 and PIC50 values**

Step 8:- Molecular descriptors were calculated using drug-like properties such as Molecular weight, Log P, MR by selecting calculate molecular descriptors and shifting all the properties to right side.

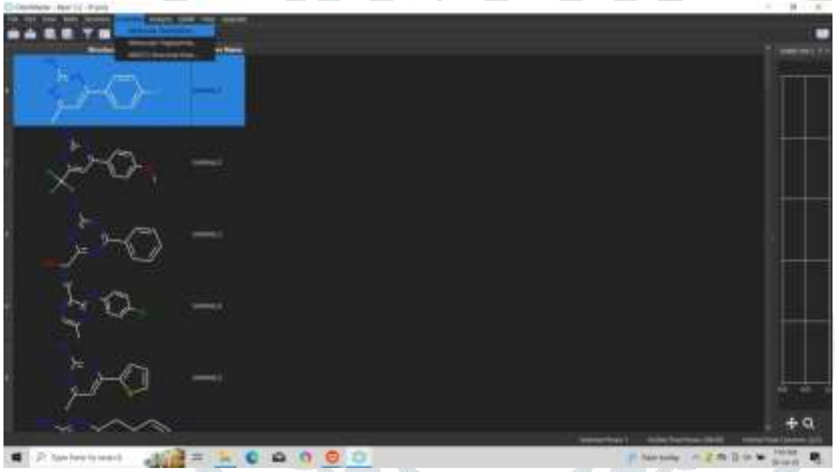


Fig no. 8 **Selecting calculate molecular descriptors.**

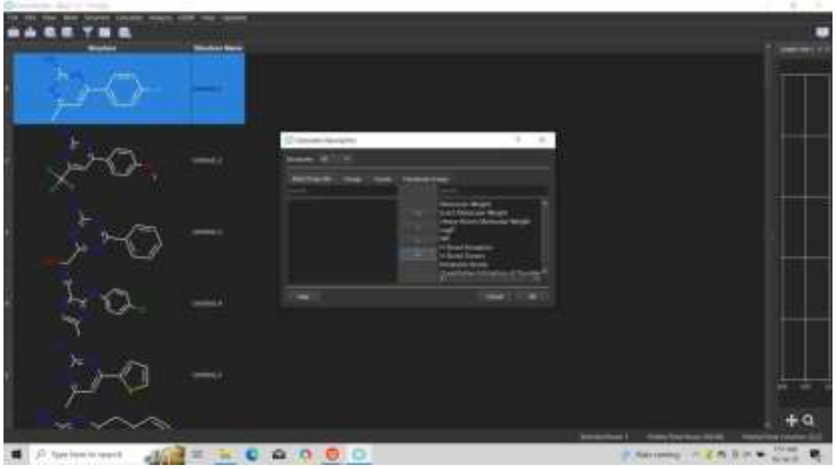


Fig no. 9 **Choosing all the Descriptors to the right side and selecting OK.**

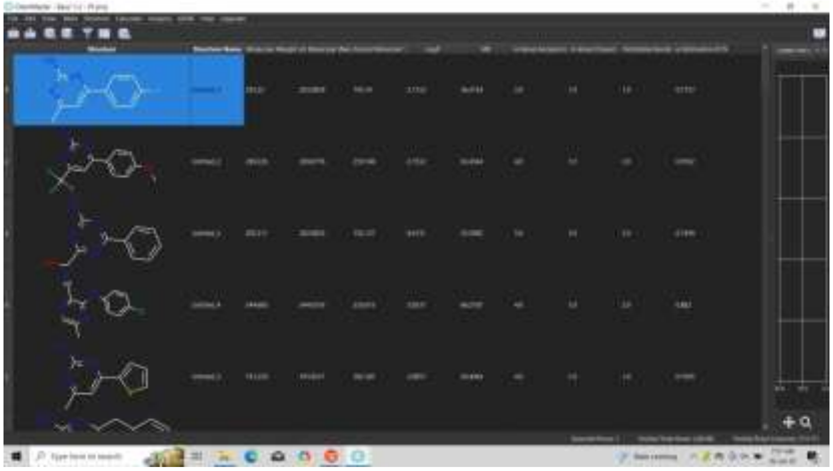


Fig no. 10 **Molecular descriptors calculated.**

Step 9:- Select QSAR and build the model by taking PIC$_{50}$ as the Y-variable and shift all molecular descriptors to the right side(X-variables) before proceeding with further steps.

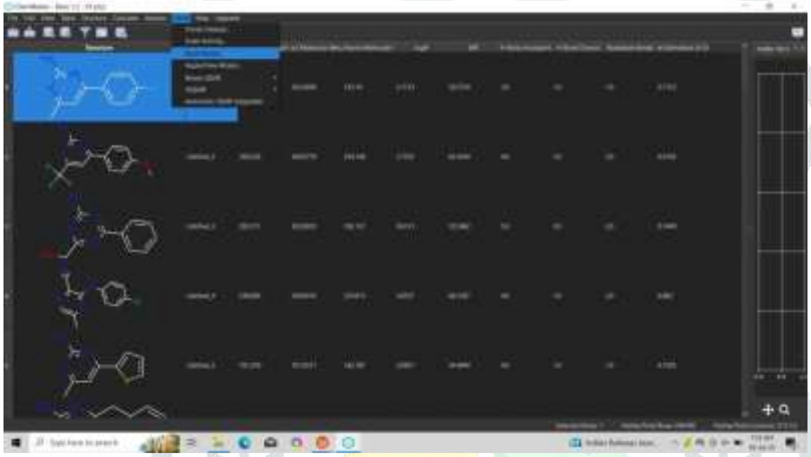


Fig no. 11 **Selecting QSAR > Build Model.**

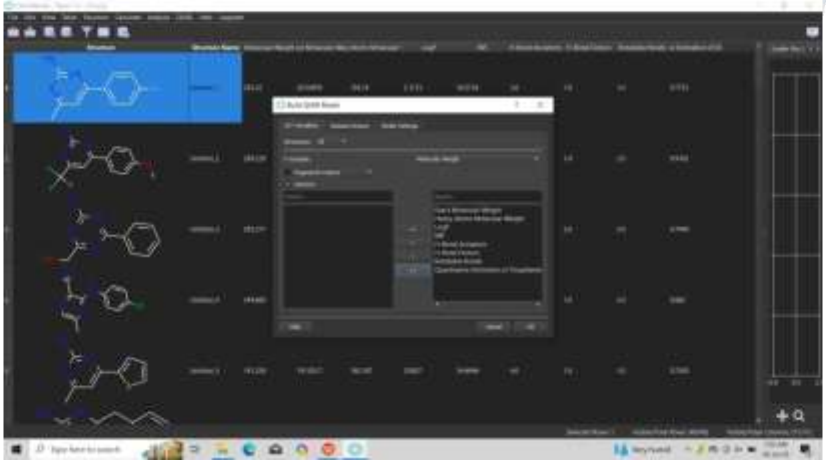


Fig no. 12 **Choosing all the parameters before proceeding.**

Step 10:- Select MLR (Multiple Linear Regression) method, PLS (Partial Least Squares), kNN (K Nearest Neighbor), SVR/SVM(Support Vector Machine) select compute Q^2 (CV LOO) for Leaveone-out cross-validation, and proceed by clicking OK.

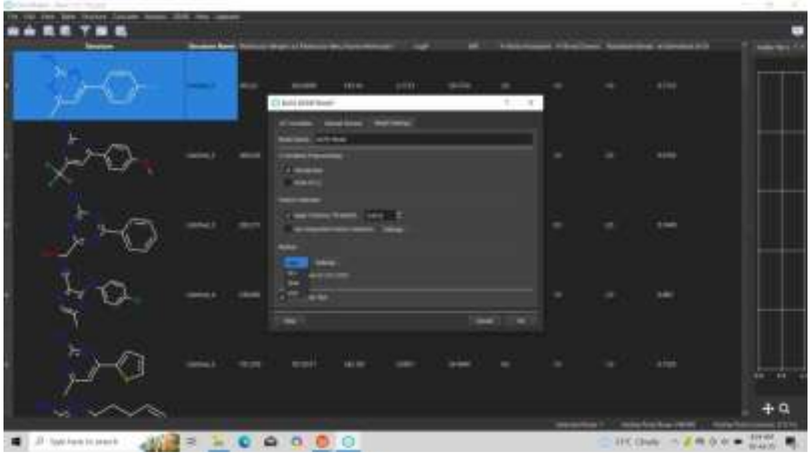


Fig no. 13 **Selecting MLR method and Compute Q^2.**

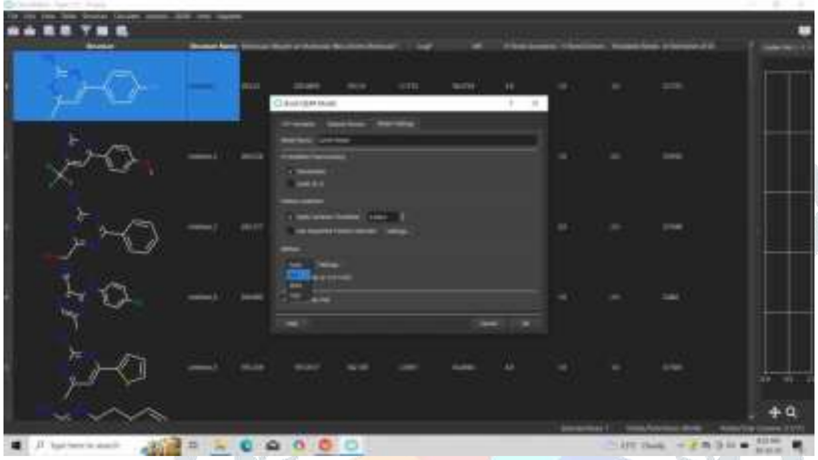


Fig no. 14 **Selecting PLS Method and Compute Q^2.**

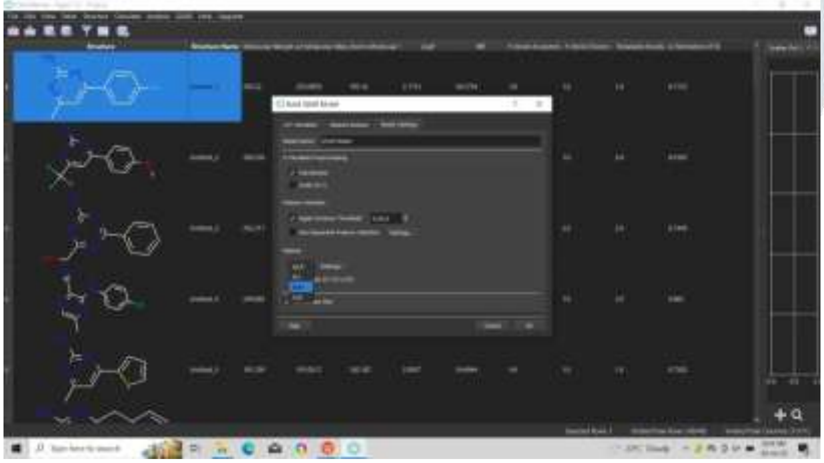


Fig no. 15 **Selecting kNN Method and Compute Q^2**

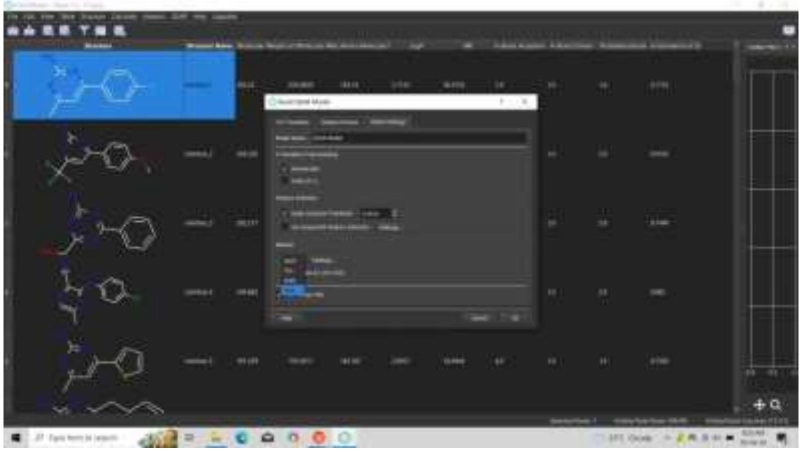


Fig no. 16 **Selecting SVR/SVM Method and Compute Q^2.**

RESULTS AND DISCUSSION

The results for the QSAR Model built for Training and Test set were obtained.

On graph X-axis will be Predicted Y (QSAR MODEL) and Y-axis will be Molecular weight.

Through this modelling we can predict the R^2, RMSE, MAE, Q^2.

MLR METHOD:

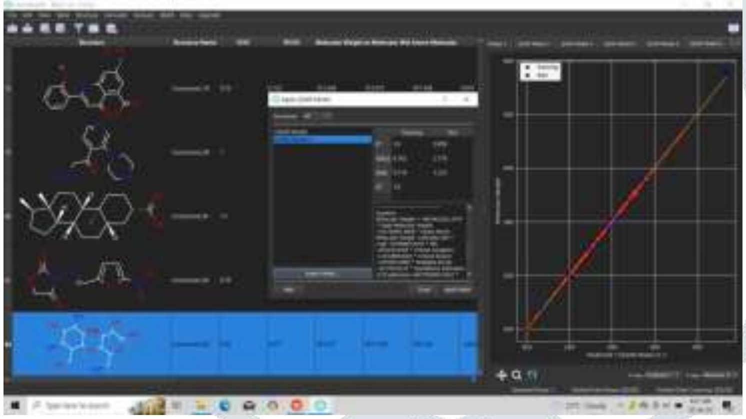


Fig no. 17 **Results of QSAR MODEL by MLR Method.**

PARAMETERS	TRAINING	TEST
R2	1.0	1.0
RMSE	0.162	2.778
MAE	0.114	1.332
Q2	1.0	

PLS METHOD:

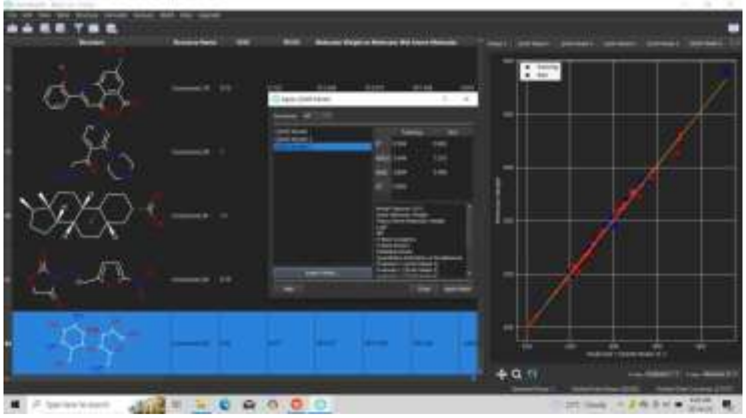


Fig no. 18 **Results of QSAR MODEL by PLS Method.**

PARAMETERS	TRAINING	TEST
R2	0.994	0.993
RMSE	5.649	7.353
MAE	3.884	5.406
Q2	0.985	

kNN METHOD:

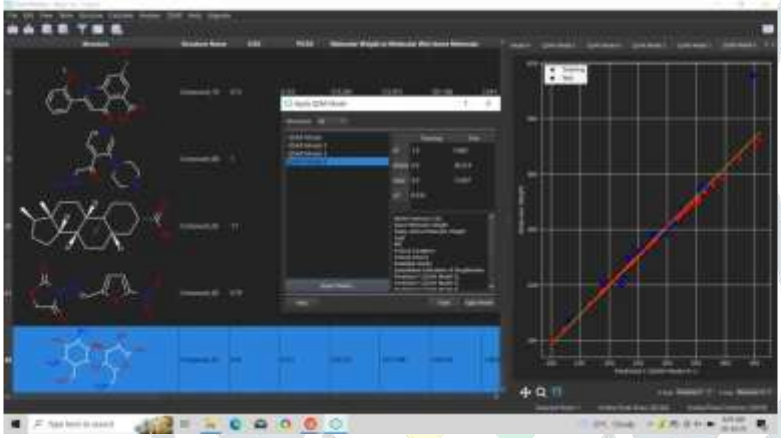


Fig no. 19 **Results of QSAR MODEL by kNN Method.**

PARAMETERS	TRAINING	TEST
R2	1.0	0.885
RMSE	0.0	30.214
MAE	0.0	13.607
Q2	0.932	

SVR/SVM METHOD:

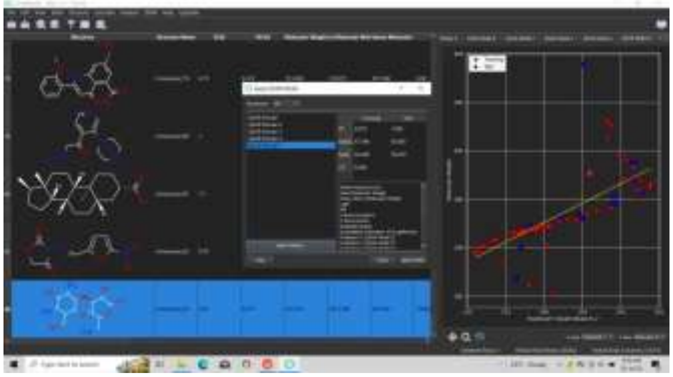


Fig no. 20 **Results of QSAR MODEL by SVR/SVM Method.**

PARAMETERS	TRAINING	TEST
R2	0.075	-0.08
RMSE	67.396	92.645
MAE	42.496	59.419
Q2	0.046	

1. R² (Coefficient of Determination)

R² is a crucial statistical metric that indicates goodness of fit of a QSAR model. It helps determine how well the model can predict the activity of new molecules based on their chemical structure Range: 0 to 1 (sometimes negative if forced through origin).

High R² (e.g. > 0.7–0.8) → your model captures most of the trends in the training data.

Low R² (e.g. < 0.5) → poor fit; descriptors may not correlate well with activity.

R² always increases when you add descriptors; does not penalize for overfitting.

2. RMSE (Root-Mean-Square Error)

RMS is measured by taking the square root of the average of the squared difference(error) between the prediction and the actual value

Units: Same as your activity (e.g., PIC_{50} units).

Lower RMS → predictions are on average closer to the true values.

Sensitive to large outliers (squares amplify large deviations).

Formula

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where:

- n= number of data points (compounds)
- yi= observed biological activity (experimental value) of the i-th compound
- y^i= predicted biological activity by the QSAR model for the i-thcompound

3. MAE (Mean Absolute Error)

It measures the average magnitude of the errors between predicted and actual values in a QSAR model.

Units: Same as your activity.

Lower MAE → on average, your predictions deviate less from actual values.

Less sensitive to outliers than RMS.

Formula

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Where:

- n= number of data points (compounds)
- yi= observed (experimental) biological activity of thei-th compound
- y^i = predicted biological activity from the QSAR model for thei-th compound

4. Q^2 (Cross-validated R^2)

The R^2 obtained from cross-validation (often leave-one-out or k-fold).

Range: Can be negative if predictive performance is worse than the mean model.

High Q^2 (e.g. > 0.5) → the model generalizes reasonably to unseen data.

Q^2 much lower than R^2 → overfitting: model fits training data well but fails to predict new compounds.

Different cross-validation schemes (LOO vs. 5-fold) give different Q^2 values; always specify which you used.

Putting it all together

Train/Test Split: You typically build the model on a training set, compute R^2/RMS/MAE there, then apply it to a validation set and compute Q^2 (and often R^2val, RMS val, MAE val).

Overfitting Check:

If $R^2 \gg Q^2$ (e.g. $R^2 = 0.9$ vs. $Q^2 = 0.3$), suspect overfitting.

PARAMETRS	ACCEPTABLE RANGE	IDEAL RANGE
R^2 (TAINING SET)	0.60-0.79	≥0.80
Q^2(CV LOO)	0.50-0.59	≥0.60
R^2(TEST SET)	0.50-0.59	≥0.60
RMSE(TRAIN/TEST)	0.20-1.00	≥0.20
MAE(TRAIN/TEST)	0.10-0.50	≥0.10

By combining these statistical indicators, we ensure the QSAR models are both accurate on the known data and robust enough to predict new compounds.

MLR METHOD Multiple Linear Regression (MLR) is a statistical method used to model the relationship between a dependent variable and multiple independent variables.

PLS METHOD Partial Least Squares (PLS) is a multivariate statistical method that relates two data tables (or blocks) by finding latent variables that maximize covariance between them.

kNN METHOD The k-Nearest Neighbors (kNN) algorithm is a supervised machine learning method used for both classification and regression tasks.

SVM METHOD A Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for both classification and regression tasks.

CONCLUSION

The development of a computational QSAR model for predicting anti-tubercular activity presents a time-efficient strategy to accelerate the discovery of effective treatments for tuberculosis. In this study, the Q^2 value was evaluated using various statistical methods, including Multiple Linear Regression (MLR), Partial Least Squares (PLS), k-Nearest Neighbors (kNN), and Support Vector Machine/Support Vector Regression (SVM/SVR). Among these methods, the MLR model yielded a Q^2 value of 1.0, indicating a statistically significant and highly predictive model.By integrating molecular descriptors with biological activity data, the constructed model demonstrates a statistically significant correlation between the chemical structures of compounds and their inhibitory activity against *Mycobacterium tuberculosis* targets. Utilizing platforms such as PubChem for data collection and Chem Master for model building, this study underscores the value of computational techniques in streamlining drug discovery, minimizing experimental efforts and guiding the rational design of novel antitubercular agents

REFERENCES

1. Domenech, P., Reed, M. B., Barry, C. E. (2005). Contribution of the Mycobacterium tuberculosis MmpL protein family to virulence and drug resistance. Infect. Immun. 73, 34923501. doi: 10.1128/Iai.73.6.3492-3501.2005
2. Dorman, S. E. et al. Four-month rifapentine regimens with or without moxifloxacin for tuberculosis. N. Engl. J. Med. 384, 1705–1718 (2021).
3. Perveen, S., Kumari, D., Singh, K. & Sharma, R. Tuberculosis drug discovery: progression and future interventions in the wake of emerging resistance. Eur. J. Med. Chem. 229, 114066 (2022).
4. World Health Organization. Global tuberculosis report 2021 (WHO, 2021).
5. World Health Organization. Global tuberculosis report 2020 (WHO, 2020).
6. Cadena, A. M., Fortune, S. M. & Flynn, J. L. Heterogeneity in tuberculosis. Nat. Rev. Immunol. 17, 691–702 (2017).
7. Sterling, T. R. et al. Guidelines for the treatment of latent tuberculosis infection: recommendations from the National Tuberculosis Controllers Association and CDC, 2020. MMWR Recomm. Rep. 69, 1–11 (2020).
8. Shah, M. & Dorman, S. E. Latent tuberculosis infection. N. Engl. J. Med. 385, 2271–2280 (2021).
9. Franke, M. F. et al. Culture conversion in patients treated with bedaquiline and/or delamanid. A prospective multicountry study. Am. J. Respir. Crit. Care Med. 203, 111–119 (2021).

10. Andries, K., Verhasselt, P., Guillemont, J., et al. (2005). A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis. Science, 307(5707), 223-227. https://doi.org/10.1126/science.1106753

11. Cole, S. T., Riccardi, G. (2011). New tuberculosis drugs on the horizon. Current Opinion in Microbiology, 14(5), 570-576. https://doi.org/10.1016/j.mib.2011.07.021

12. World Health Organization (WHO). (2021). WHO Consolidated Guidelines on Tuberculosis: Module 4: Treatment - Drug-Resistant Tuberculosis Treatment. Geneva: World Health Organization. https://www.who.int/publications/i/item/9789240028173

13. Koul, A., Dendouga, N., Vergauwen, K., et al. (2007). Diarylquinolines target subunit c of mycobacterial ATP synthase. Nature Chemical Biology, 3(6), 323-324. https://doi.org/10.1038/nchembio883

14. Balabanov, S., Gille, L., Pashkova, N., et al. (2019). Mechanisms of action and resistance to Bedaquiline in Mycobacterium tuberculosis. Antimicrobial Agents and Chemotherapy, 63(4), e00511-19. https://doi.org/10.1128/AAC.00511-19

15. Lounis, N., Veziris, N., Chauffour, A., et al. (2011). Bactericidal and sterilizing activities of Bedaquiline (TMC207) in murine tuberculosis. Antimicrobial Agents and Chemotherapy, 55(11), 5287-5290. https://doi.org/10.1128/AAC.00652-11

16. Wang, F., Langley, R., Gulten, G., et al. (2010). Mechanism of action of diarylquinolines: Inhibition of mycobacterial ATP synthase by Bedaquiline. Journal of Biological Chemistry, 285(34), 25273-25279. https://doi.org/10.1074/jbc.M110.139246

17. Pontali, E., Tiberi, S., D'Ambrosio, L., et al. (2017). Bedaquiline and multidrug-resistant tuberculosis: A systematic and critical analysis of the evidence. European Respiratory Journal, 49(4), 1700462. https://doi.org/10.1183/13993003.00462-2017

18. A review of recent advances in anti-tubercular drug development Indian Journal of Tuberculosis 2020Théoneste Umumararungu a, Marie Jeanne Mukazayire a, Matabishi Mpenda a, Marie Françoise Mukanyangezi a, Jean Bosco Nkuranga b, Janvier Mukiza c, Emmanuel Oladayo Olawode d.

19. Anti-tuberculosis treatment strategies and drug development: challenges and priorities Nature review microbiology.2022 Veronique A Dartois, Eric J Rubin.