Machine Learning Approaches to Predict Biological Effects of Organic Compounds

Aniket Sanjay Shitole
Institute of Chemical Technology
Jalna, India
aniket.shitole@ieee.org

Shrikant Mete
Institute of Chemical Technology
Jalna, India
ss.mete@staffmarj.ictmumbai.edu.in

Navnath Hatvate
Institute of Chemical Technology
Jalna, India
nt.hatvate@marj.ictmumbai.edu.in

***Abstract*—Tuberculosis, caused by Mycobacterium tuberculosis, remains a persistent global health problem, affecting more than two billion people and posing a significant risk of morbidity and mortality, especially in developing regions. Individuals with immunosuppressive conditions, such as HIV/AIDS, Cancer, and Lupus, are particularly vulnerable. Urgent efforts are needed to discover new and effective drugs to combat this disease. Machine learning offers a transformative approach by predicting the biological activity of organic compounds, thereby expediting drug discovery processes. This paper focuses on integrating machine learning models with drug discovery data to enhance identifying potential drug candidates efficiency and success rate. By accurately predicting the biological activity of structurally similar organic compounds, the drug discovery process can be accelerated, developing more effective drugs. Specifically, this research aims to predict Minimum Inhibitory Concentration (MIC) values for analogous compounds, focusing on Q203, an antituberculosis drug that inhibits Mycobacterium tuberculosis growth. Physicochemical properties, including partition coefficient (LogP), Molecular refractivity (MR), calculated Molecular refractivity (CMR), calculated partition coefficient (CLogP), Henry's law, topological polar surface area (tPSA), and the logarithm of solubility (LogS) are computed using ChemDraw software. Machine learning models trained on MIC data extracted from literature sources and Delta learning rule was used to predict MIC values for similar compounds. Implementing the delta learning rule to predict the MIC values of structurally identical compounds builds on previous successful applications of machine learning algorithms in drug development. Machine learning in drug discovery holds promise to accelerate the identification of novel drugs to combat tuberculosis and other infectious diseases.**

I. Introduction

The persistent spread of tuberculosis among populations is a long-standing problem. Tuberculosis, caused by Mycobacterium tuberculosis, affects more than two billion people worldwide and is a significant cause of morbidity and mortality, particularly in developing countries. Individuals with immunosuppressive conditions such as HIV/AIDS, cancer, and Lupus are at higher risk of contracting the disease. This underscores the urgency of discovering new and efficacious drugs to combat the disease. Machine learning emerges as a key tool in predicting the biological activity of organic compounds, thereby expediting drug discovery processes. In drug discovery and development, prediction of biological activity is of critical importance [1]. By integrating machine learning models with drug discovery data, the efficiency and success rate of identifying potential drug candidates can be substantially enhanced. Accurate prediction of the biological activity of structurally similar organic compounds accelerates the drug discovery process, leading to the development of more effective drugs [1]. Machine learning has emerged as a powerful tool in the field of drug discovery. There has been a growing interest in utilising machine learning algorithms and techniques to streamline various tasks in the drug discovery process [2]. Some of these are polypharmacology, which involves the prediction of multiple targets for a given drug molecule, virtual screening to identify potential drug candidates from large databases, prediction of drug toxicity and ADME properties, optimisation of lead compounds, and even designing clinical trials. Also, deep learning is used to screen potential drug candidates more efficiently by analysing large-scale data. Machine learning can provide valuable insight into drug mechanisms, establish biomarkers, optimise drug candidates, and even repurpose existing drugs. Apart from targeting the drug's biological activity, AI/ML is being used as Natural language processing to extract and analyse valuable information from scientific publications, clinical trials, and other sources [3]. The predictive power of machine learning can be easily harnessed to get the best possible results in the drug discovery process, leading to more effective and safer medicines for patients. One such approach has been suggested and mentioned in this article for the prediction of the biological activity of drug molecules based on their physicochemical properties.

This article focuses on utilizing various physicochemical properties to train machine learning models for predicting Minimum Inhibitory Concentration (MIC) values for analogous compounds. Q203, an antituberculosis drug that inhibits Mycobacterium tuberculosis growth, serves as the study drug compound [4]. Telacebec, previously known as Q203, is a pioneering drug that targets the cytochrome bc1 complex of Mycobacterium tuberculosis, crucial for its survival and growth.

This research aims to determine the MIC values for substances similar to the specific drug through machine learning. MIC represents the minimum amount of substance required to inhibit microbial growth effectively. Determining MIC involves methods such as solid agar plates or liquid broth dilution. Physicochemical properties, including the partition

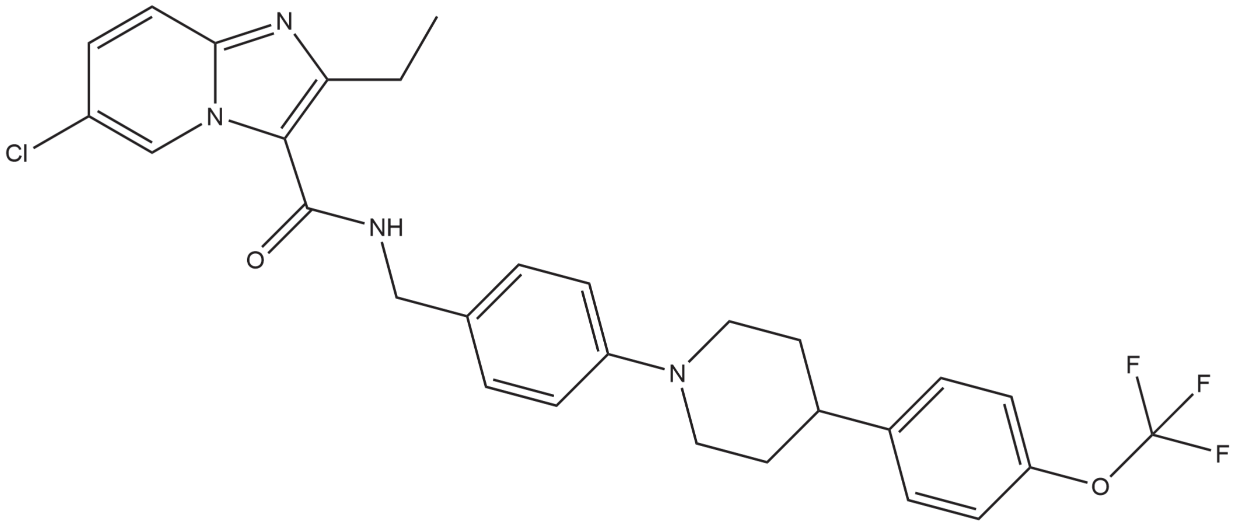

Fig. 1. Chemical structure of the Telacebec (Q203) drug molecule. Q203 is a promising anti-tuberculosis agent targeting the mycobacterial cytochrome bc1 complex. The molecule is characterised by its unique quinoline core and hydrophobic tail, which are crucial for its selective inhibition of bacterial respiration.

coefficient (LogP), Molecular refractivity (MR), calculated Molecular refractivity (CMR), calculated partition coefficient (CLog P), Henry's Law, Topological Polar Surface Area (tPSA), and Logarithm of Solubility (LogS), are calculated using ChemDraw software. MIC 50, 80, and 90 data are extracted from pertinent sources in the literature. Machine learning models are trained to predict MIC values for similar compounds and compared with a 2D Quantitative Structure-Activity Relationship (QSAR) model to ensure precision in predicting the MIC of new compounds. Delta learning rules update the weights and biases of the model to minimize the error between predicted and actual MIC values. A portion of the data serves as the training and testing dataset to evaluate the machine learning models' performance. This sets the groundwork for implementing feedforward neural networks to predict MIC values of structurally similar compounds.

Previous research demonstrates the successful application of machine learning algorithms in various stages of drug development, including prediction of target structure, biological activity, and optimization of hits [5]. Furthermore, [6] predicts the pMIC values of the antituberculosis drug 3-phenyl-1,3-benzoxazine-2,4 (3) derivatives using artificial neural networks (ANN), using lipophilicity and solubility indices as derivatives for prediction.

II. Methodology

The study utilized experimental methods to calculate the physicochemical properties of analogues of the drug molecule Q203. This involved employing ChemDraw to identify structurally similar compounds and compute their physiochemical properties using computational tools for LogP, pKa, MR cm3/mol, CMR, CLog P, Gibbs energy, Henry's law constant, tPSA, and LogS. Furthermore, the MIC 50, 80 and 90 data were extracted from relevant research articles.

A. Artificial Neural Network

Machine learning models such as Random Forest, Support Vector, Linear Regression, and others were considered. However, artificial neural networks (ANN) emerged as the primary model for testing the hypothesis that machine learning algorithms accurately predict the biological activity of structurally similar organic compounds. Inspired by the brain's neural network, which comprises approximately 100 billion interconnected neurons to process and transmit information, ANN mimics this architecture [7].

B. Learning Rules

Various learning models are used to train ANN models and improve their performance. Examples include Hebbian Learning, Perceptron Learning Rule (PLR), and Delta Learning Rule (DLR) [7]. This study focusses mainly on the Delta learning rules. The delta learning rule is employed in feedforward neural network, such as backpropagation algorithm and which can be used for further prediction purposes.

C. Data Collection

The study requires two types of data sets: MIC values extracted from research articles and the physiochemical properties of Q203 drug analogues, calculated using ChemDraw software. These properties include LogP, pKa, MR, CMR, CLog P, Gibbs energy, Henry's law, tPSA, and LogS. Seven of these properties (LogPMR, Henry's Law, tPSA, CLog P, CMR, and LogS) were selected for training the ANN model, based on their relevance to Q203 drug analogue performance and their availability in the data set. MIC values (MIC 50, 80, and 90) were obtained from relevant research articles [8] [9] [10].

D. Training the Model

MIC 80 and 90 and the corresponding attributes as shown in Table I are used to train the DLR model, with 4 vector arrays

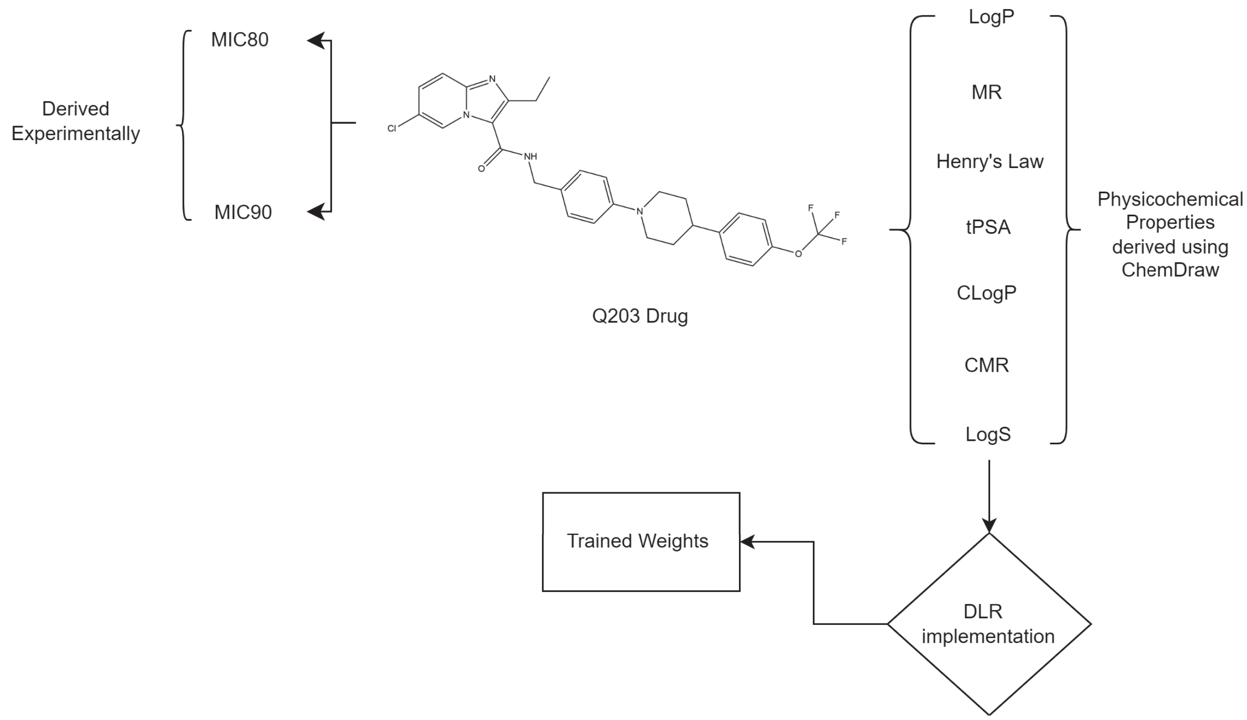

Fig. 2. The flow diagram represents a systematic approach that combines experimental data, chemical insights, and Delta learning rule to enhance our understanding of the Q203 drug's physicochemical characteristics.

each of input, respectively. The classification of the data sets involved using exactly 50 for the training of the models and the remaining 22 for the testing.

III. Results and Discussion

The input parameters included LogP, pKa, MR, CMR, CLogP, Gibbs energy, Henry's Law, tPSA and LogS. These were normalized before being used to train the artificial neural network model using the Delta Learning Rule model. During the testing phase, data that are not used in training is used to get an idea of the performance of this model for prediction of MIC values. DLR was applied on the same data set against the MIC80 and MIC90 values of the drug molecule analogues. The results indicated that the Delta Learning Rule model can be trained effcetively and can be used for prediction of MIC values. The DLR calculates the error between the predicted and target values then adjusts weights until desired outputs are achieved. Thus, we successfully trained DLR which effectively handles non-binary output data by updating weights based on prediction values while avoiding infinite loops.

In delta learning rule model, the least squared error between d_i and o_i is calculated using the mathematical model given in Eq. 1.

$$E = \frac{1}{2}(d_i - o_i)^2 \tag{1}$$

Normalizing the data brings all input parameters to a similar scale, allowing better comparison and analysis [11]. This was achieved by dividing each input value by the highest value in the respective parameter. The normalization process ensured that all predictor variables had the same influence on the predictive model, effectively improving its performance and accuracy.

This model works by calculating the square of the difference between the desired value and the output value for each adjusted weight unless the error does not equal or is less than 0.1. Hence, the error keeps decreasing over time as the adjusted weights push the overall output value close to the desired value.

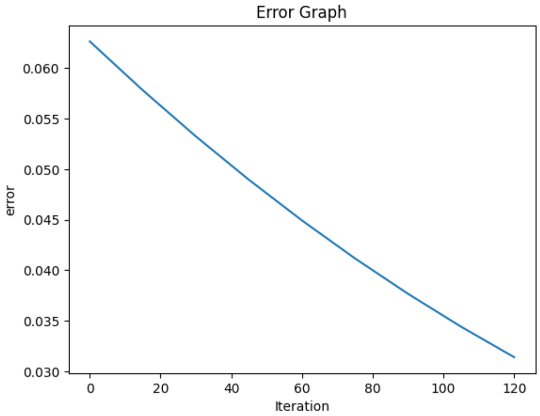

Fig. 3. The decreasing error in each iteration of the DLR can be seen. After each iteration the weights are adjusted to new set of weights which can be again fed in the loop eventually minimising error.

In the future, we wish to implement multilayer feedforward neural networks on the existing collected extensive dataset of

TABLE I. TRAINED WEIGHTS OBTAINED AFTER TRAINING THE INPUT ATTRIBUTES WITH 7 VALUES IN EACH ARRAY. ONLY 3 SUCH INPUTS ARE SHOWN HERE FOR REPRESENTATION PURPOSE DUE TO SPACE CONSTRAINTS.

Learning Rule	Desired Parameter	Input Parameters (Attributes)	Trained Weights
		[0,1,0.1,0.5,0.1,0.1,0.1]	
	MIC80	[0,1,0.2,0.7,0,0.1,0]	[0.3, 0.31, 0.66, 0.08, 0.49, 0.48, 0.31]
Delta Learning Rule		[0,1,0.2,0.6,0,0.1,-0.1]	
		[0,1,0.1,0.6,0,0.1,0]	
	MIC90	[0.0,1.0,0.1,0.5,0.0,0.1,0]	[0.3 0.3 0.67 0.10 0.5 0.48 0.3]
		[0.0,1.0,0.2,0.6,0.0,0.1,0]	

around 200 compounds for accurate prediction of the MIC values of the drug molecule. The learning rule implemented in this, such as DLR, will establish a foundation for achieving good accuracy in predicting the MIC values of the drug molecule based on the extensive dataset. Therefore, we can expect to achieve accurate predictions of MIC values for the drug molecule by utilizing the generalized Delta Learning Rule and error backpropagation training to implement multilayer feedforward neural networks with the existing dataset.

IV. CONCLUSION

The widespread challenge of tuberculosis persists globally, affecting millions and necessitating urgent drug discovery efforts. Machine learning stands out as a vital tool that accelerates the prediction of compound activity. This article uses various physicochemical properties to predict Minimum Inhibitory Concentration (MIC) values for analogous compounds, explicitly targeting Mycobacterium tuberculosis growth inhibition. By training machine learning models and integrating data from various sources, our objective is to enhance the efficiency of drug discovery processes. The successful application of these methods promises to accelerate the development of more effective drugs to combat tuberculosis and other infectious diseases.

ACKNOWLEDGMENT

I wish to acknowledge the PhD scholars and Master students at my institute who accompanied me in running my programs smoothly along with the help in data collection process in a timely manner, as well as my friends and family who helped me by supporting me through the entire process.

REFERENCES

[1] W. Duch, K. Swaminathan, and J. Meller, “Artificial intelligence approaches for rational drug design and dis-covery,” *Current pharmaceutical design*, vol. 13, no. 14, pp. 1497–1508, 2007.

[2] N. Noorbakhsh-Sabet, R. Zand, Y. Zhang, and V. Abedi, “Artificial intelligence transforms the future of health care,” *The American journal of medicine*, vol. 132, no. 7, pp. 795–801, 2019.

[3] S. Kolluri, J. Lin, R. Liu, Y. Zhang, and W. Zhang, “Machine learning and artificial intelligence in pharma-ceutical research and development: a review,” *The AAPS journal*, vol. 24, pp. 1–10, 2022.

[4] R. Lahiri, L. B. Adams, S. S. Thomas, and K. Pethe, “Sensitivity of mycobacterium leprae to telacebec,” *Emerging Infectious Diseases*, vol. 28, no. 3, p. 749, 2022.

[5] A. N. Lima, E. A. Philot, G. H. G. Trossini, L. P. B. Scott, V. G. Maltarollo, and K. M. Honorio, “Use of machine learning approaches for novel drug discovery,” *Expert opinion on drug discovery*, vol. 11, no. 3, pp. 225–239, 2016.

[6] P. Nemeˇcek, J. Moc´ak, J. Lehotay, and K. Waisser, “Prediction of anti-tuberculosis activity of 3-phenyl-2 h- 1, 3-benzoxazine-2, 4 (3 h)-dione derivatives,” *Chemical Papers*, vol. 67, pp. 305–312, 2013.

[7] J. Zurada, *Introduction to artificial neural systems*. West Publishing Co., 1992.

[8] S. Kang, Y. M. Kim, R. Y. Kim, M. J. Seo, Z. No, K. Nam, S. Kim, and J. Kim, “Synthesis and structure-activity studies of side chain analogues of the anti-tubercular agent, q203,” *European Journal of Medicinal Chemistry*, vol. 125, pp. 807–815, 2017.

[9] S. Kang, Y. M. Kim, H. Jeon, S. Park, M. J. Seo, S. Lee, D. Park, J. Nam, S. Lee, K. Nam *et al.*, “Synthesis and structure-activity relationships of novel fused ring analogues of q203 as antitubercular agents,” *European Journal of Medicinal Chemistry*, vol. 136, pp. 420–427, 2017.

[10] K. D. Farrell, Y. Gao, D. A. Hughes, R. Henches, Z. Tu, M. V. Perkins, T. Zhang, and C. L. Francis, “3-methoxy-2-phenylimidazo [1, 2-b] pyridazines highly active against mycobacterium tuberculosis and mycobac-terium marinum,” *European Journal of Medicinal Chem-istry*, vol. 259, p. 115637, 2023.

[11] M. Bowles, *Machine learning in Python: essential tech-niques for predictive analysis*. John Wiley & Sons, 2015.