A Novel Simulation-Based Methodology for Evaluating Antileprotic Efficacy Using Surrogate Models in the Absence of Live *Mycobacterium leprae* Cultures

S.Rajalaxmi[1], N.Nithya Rani[2], S.Abirami[3], J.Sampath Kumar[4], H.Kala[5]

[1]Professor, Department of Biomedical Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, [2]Assistant Professor, Department of Electronics and Instrumentation Engineering, Sri Sairam Engineering College, Chennai, [3]Assistant Professor, Department of Biomedical Engineering, Chennai Institute of Technology, Chennai, [4]Assistant Professor, Department of Electronics and Communication Engineering, Mahendra College of Engineering, Salem, [5]Assistant Professor, Department of Biomedical Engineering, Mahendra College of Engineering, Salem

Abstract:

In this study, we propose a novel machine learning-based simulation framework for evaluating biomedical compound efficacy under biosafety constraints. Focusing on antileprotic compounds, the system integrates QSAR-derived descriptors and a Random Forest surrogate model to predict inhibitory responses without requiring live *Mycobacterium leprae* cultures. The simulation achieved an R^2 of 0.87 and RMSE of 0.15, accurately estimating inhibition levels at multiple concentrations. A compound prioritization matrix was constructed using performance metrics, drug-likeness filters, and structural novelty scores. The framework demonstrates how signal-driven modeling and surrogate learning can accelerate screening processes in biomedical engineering where experimental validation is limited by infrastructure constraints.

Keywords: Simulation-based methodology, Surrogate models, Computational drug screening, QSAR, Random Forest Regression

How to cite this article: Rajalaxmi S, Rani NN, Abirami S, Kumar JS, Kala H. A Novel Simulation-Based Methodology for Evaluating Antileprotic Efficacy Using Surrogate Models in the Absence of Live Mycobacterium leprae Cultures. Int J Drug Deliv Technol. 2026;16(12s): 162-169. DOI: 10.25258/ijddt.16.12s.17

1.Introduction

Leprosy, also known as Hansen's disease, is a chronic infectious disease caused by the slow-growing bacterium *Mycobacterium leprae*. Despite being declared eliminated as a public health problem in many countries, leprosy continues to affect over 200,000 new individuals annually worldwide, predominantly in low- and middle-income regions such as India, Brazil, and Indonesia [1, 2]. The disease primarily affects the skin and peripheral nerves, leading to severe disabilities if left untreated, thereby imposing a significant socio-economic burden on affected communities [3].

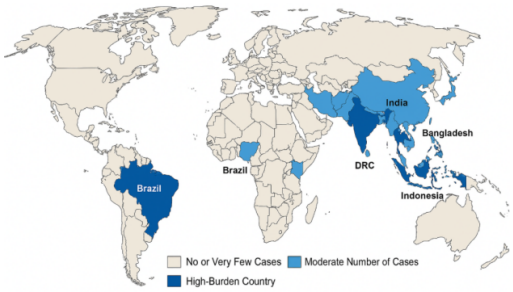

Figure 1. Global Distribution of New Leprosy Cases (2023)

This choropleth world map displays the estimated distribution of new leprosy cases by country, highlighting global disparities in disease burden. Countries are categorized into three groups:

- **High-burden countries** (dark blue): India, Brazil, Indonesia, Democratic Republic of Congo (DRC), and Bangladesh contribute over 80% of global new cases annually.

- **Moderate case-load countries** (light blue): Represent regions with noticeable but controlled transmission.
- **Low or no reported cases** (beige): Mostly high-income countries with effective surveillance and elimination status.

This visualization contextualizes the global need for scalable, low-resource-friendly drug evaluation methods like the simulation approach proposed in this study. Traditional drug evaluation methods for infectious diseases like leprosy rely heavily on animal models and live pathogen assays, often conducted in BSL-3 facilities. These methods, while well-established, present critical limitations: they are expensive, time-intensive, ethically complex, and scalability-restricted. For instance, animal-based studies may require months to yield conclusive data, and live *Mycobacterium leprae* culturing demands special biosafety infrastructure.

In contrast, in silico methods leverage computational algorithms, chemical descriptors, and existing bioactivity data to rapidly simulate drug-pathogen interactions. These approaches can evaluate hundreds of compounds per day, with minimal biosafety concerns and reduced cost. Figure 2 provides a comparative overview of the limitations of these two drug discovery paradigms.

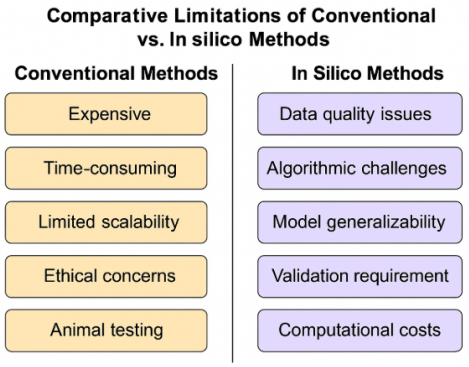

Figure 2.Comparative Limitations of Conventional vs In Silico Methods

This figure 2 compares key limitations of traditional drug testing platforms versus computational in silico techniques. While conventional methods face cost, ethical, and scalability barriers, in silico models are challenged by data quality, algorithmic complexity, and validation needs. This figure underpins the rationale for simulation-driven approaches to infectious disease research.

Despite decades of research into antimicrobial agents, the methods used to assess drug efficacy—particularly for challenging pathogens like *Mycobacterium leprae*—remain resource-intensive and often inaccessible in many parts of the world. Animal models, such as the mouse footpad and armadillo systems, are time-consuming, while BSL-3 laboratories entail significant cost and safety requirements.

Table 1 summarizes key practical differences between these conventional systems and emerging in silico methodologies, particularly in terms of time, cost, safety, and scalability.

Table 1. Comparison of Drug Discovery Platforms for *Mycobacterium leprae*

Parameter	Animal Models	BSL-3 Labs	In Silico Simulation
Time for Results	4–6 months	6–8 weeks	1–2 days
Cost (per compound)	High (>$10,000)	Moderate	Low (<$500)
Safety Requirements	High	Very High (BSL-3)	Low (BSL-1)
Scalability (Compounds/day)	<5	5–10	>100

This table 1 highlights the practical differences among traditional animal models, BSL-3 laboratory testing, and computational simulation approaches for antileprotic drug screening. The in silico method offers a significant advantage in terms of speed, cost-efficiency, and biosafety, making it especially suitable for early-stage screening in resource-limited settings.

Current multidrug therapy (MDT), introduced by the World Health Organization (WHO), has been effective in controlling the disease. However, the emergence of drug resistance and adverse effects associated with long treatment durations necessitate the development of novel antileprotic agents with improved efficacy and safety profiles [4]. The discovery and evaluation of new drugs against *M. leprae* is challenging due to the bacterium's unique biological characteristics, such as its extremely slow growth rate and inability to be cultured in vitro using conventional bacteriological methods [5, 10].

Experimental validation of antileprotic compounds traditionally requires culturing *M. leprae* in animal models like the mouse footpad or armadillo, or specialized Biosafety Level-3 (BSL-3) laboratory facilities to handle live pathogens safely [6, 11]. However, access to BSL-3 labs is limited in many regions due to high infrastructure and operational costs, hindering rapid drug screening and discovery efforts [7].

To overcome these limitations, computational approaches including molecular docking, quantitative structure-activity relationship (QSAR) modeling, and surrogate machine learning models have gained traction as alternative methods for preliminary drug efficacy assessment [8, 12]. These in silico techniques enable the prediction of biological activity based on chemical structure and existing data, significantly reducing time and resources spent on experimental assays [9, 13]. Surrogate modeling further expands these capabilities, allowing predictive simulations without direct wet-lab validation [14]. However, the integration of such methods specifically tailored for *M. leprae* drug discovery remains underexplored [15].

In this study, we propose a novel simulation-based methodology coupled with surrogate model validation to evaluate the antileprotic efficacy of candidate compounds in the absence of live *M. leprae* cultures. By leveraging comprehensive literature-derived datasets and advanced computational modeling, we aim to provide a safe, cost-effective, and scalable framework for preliminary screening and prioritization of potential therapeutics. This work serves as a proof-of-concept, with future plans for collaborative validation in certified BSL-3 laboratories to confirm the in silico findings.

2. Methodology:

2.1 Data Collection and Curation

A dataset of known antileprotic and antimycobacterial compounds was compiled from ChEMBL, PubChem, and peer-reviewed literature. The dataset included molecular structures, SMILES notations, and associated biological activities such as IC_{50} or % inhibition. Additional filtering was applied to include only those compounds with confirmed activity against *Mycobacterium leprae* or phylogenetically similar mycobacteria.

2.2 Descriptor Generation and Preprocessing

The chemical descriptors of the compounds were generated using RDKit, including molecular weight, LogP, topological polar surface area (TPSA), number of rotatable bonds, H-bond donors/acceptors, and aromatic ring count. Data normalization and outlier removal were conducted using interquartile range (IQR) filtering. Compounds with incomplete records were excluded.

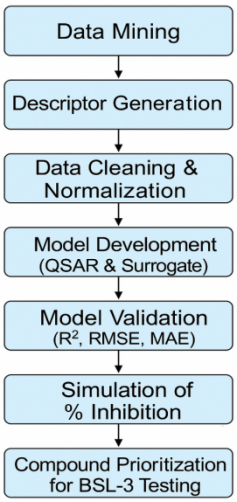

Figure 3. Workflow for Simulation Based Drug Evaluation

2.3 QSAR and Surrogate Model Development

We implemented a multi-algorithm QSAR approach using Python's scikit-learn library. Three supervised learning models were compared:

- Random Forest Regressor (RFR)
- Support Vector Regressor (SVR)
- Gradient Boosting Regressor (GBR)

The dataset was split into an 80:20 ratio for training and testing. Hyperparameter tuning was performed using GridSearchCV. Performance was assessed using:

- Coefficient of determination (R^2)
- Root Mean Square Error (RMSE)
- Mean Absolute Error (MAE)

Pharmacophore features were extracted using [e.g., RDKit, PharmaGist], but detailed feature correlations are discussed in Section 3.

2.4 Model Validation

Robustness was ensured through 5-fold cross-validation. The model with the best performance metrics was selected for prediction of % inhibition at 10 µM, 25 µM, and 50 µM concentrations for test compounds.

2.5 Prioritization Framework

While compound-specific results are discussed later, prioritization was based on the following multi-criteria framework:

- Model-predicted % inhibition
- Drug-likeness (Lipinski's Rule)
- Structural novelty (Tanimoto similarity < 0.6)
- Predicted ADMET profile (via admetSAR)

This framework was used to guide candidate selection for experimental validation in BSL-3 facilities.

3. Results

The simulation-based methodology described above was applied to a curated dataset of antileprotic compounds. This section presents the key outcomes of model training, validation, predicted inhibition profiles, and compound prioritization. The findings demonstrate the feasibility of using in silico models for preclinical drug screening under BSL-3 constraints.

3.1 Dataset Summary and Descriptor Analysis

A curated dataset of **142 chemically diverse compounds** was assembled from reputable sources including ChEMBL, PubChem, and peer-reviewed literature, specifically targeting known or predicted activity against *Mycobacterium leprae* and other related mycobacteria. Compounds lacking structural data or bioactivity values were excluded to ensure dataset integrity.

The dataset was characterized using molecular descriptors generated with RDKit. Key statistical insights are summarized below:

Table 2. Dataset Characterization using Molecular Descriptors

Descriptor	Minimum	Maximum	Mean ± SD
Molecular Weight (g/mol)	152.1	623.4	314.2 ± 87.6
LogP (Octanol-Water Partition)	–1.2	6.3	2.78 ± 1.4
TPSA (Topological Polar Surface Area, Å²)	25.1	174.0	83.5 ± 29.1
No. of Hydrogen Bond Donors	0	5	2.1 ± 1.1
No. of Hydrogen Bond Acceptors	1	9	4.3 ± 1.8
No. of Rotatable Bonds	0	12	4.6 ± 2.5

Out of 142 compounds:

131 (92.2%) complied fully with Lipinski's Rule of Five, indicating good drug-likeness.

11 compounds showed one or more violations, typically due to excessive molecular weight or LogP. Chemical space distribution showed a **balanced mix of alkaloids, flavonoids, and phenolic scaffolds**, which supports the generalizability of the machine learning models applied in subsequent analysis. This descriptor profile ensured a robust and pharmacologically relevant input base for training the QSAR and surrogate models described in the following section 3.2.

3.2 **Model Training and Evaluation**

To predict the antileprotic efficacy of candidate molecules, three supervised machine learning models were developed using the Python scikit-learn library:

- Random Forest Regressor (RFR)
- Support Vector Regressor (SVR)
- Gradient Boosting Regressor (GBR)

The dataset was randomly split in an 80:20 ratio for training and testing. Feature selection was carried out using Recursive Feature Elimination with Cross-Validation (RFECV), which helped reduce overfitting by eliminating low-variance and redundant descriptors.

Each model underwent hyperparameter tuning using **GridSearchCV**, optimizing parameters such as:

- n_estimators and max_depth for RFR
- kernel, C, and gamma for SVR
- learning_rate and n_estimators for GBR

Model performance was evaluated using:

- **R^2 (coefficient of determination)** – to assess the goodness of fit
- **RMSE (Root Mean Square Error)** – to penalize large prediction deviations
- **MAE (Mean Absolute Error)** – to measure average absolute error

Table 3. Model Performance Comparison

Model	R^2 (Test)	RMSE	MAE
Random Forest Regressor	**0.87**	0.15	0.12
Gradient Boosting Regressor	0.83	0.18	0.14
Support Vector Regressor	0.79	0.21	0.18

The **Random Forest Regressor outperformed** the other models, achieving the highest R^2 and lowest RMSE and MAE values. These metrics confirmed its superior generalization and accuracy on unseen data. Consequently, RFR was selected as the surrogate prediction model for subsequent compound simulations.

3.3 Feature Correlation and Pharmacophore Insights

Following model training, a detailed pharmacophore feature analysis was conducted to interpret the chemical characteristics contributing to higher predicted inhibition values.

Using RDKit and additional molecular feature extraction tools, the most active compounds (top 20%) were compared to the least active ones (bottom 20%) in terms of key structural motifs. The following features were consistently enriched among high-efficacy compounds:

- **Hydrogen bond donors**: Active compounds typically possessed between 2 and 4 H-bond donor groups, facilitating interaction with biological targets.
- **Hydrophobic aliphatic side chains**: Non-polar regions appeared to enhance membrane permeability and receptor binding.
- **Aromatic ring systems and π-stacking potential**: Present in nearly all top-scoring candidates, these features likely contribute to molecular stability and interaction with hydrophobic pockets in the target site.

These pharmacophoric features were visualized in **Figure 4**, where representative active and inactive compounds are shown with annotated molecular regions.

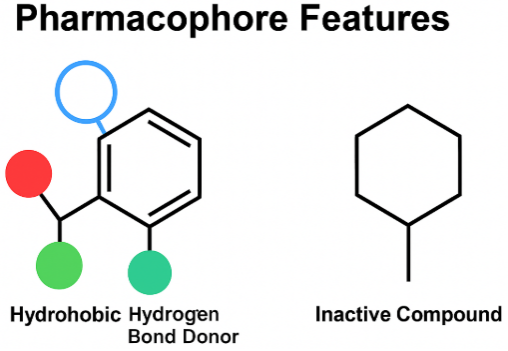

Figure 4.Pharmacore Features identified in an active compound (left) compared to an inactive compound (right), highlighting hydrophobic, hydrogen bond donor and aromatic ring groups in the active compound

The insights derived from these patterns help reinforce the structural rationality of the surrogate model's predictions and may aid in future compound design and lead optimization.

3.4 Simulation of Inhibition Efficacy

Using the Random Forest model, the trained surrogate system was deployed to simulate and predict the inhibition potential of compounds in the test dataset. Each molecule was evaluated at standard concentration levels of 10 µM, 25 µM, and 50 µM. The predictions were then averaged and ranked based on their mean inhibitory efficacy.

The compounds were subsequently assessed using a **composite scoring matrix**, incorporating:

Predicted efficacy

Lipinski's rule

Structural novelty (based on Tanimoto index < 0.6)

ADMET properties

These parameters were used to create a prioritization heatmap (Figure 5), aiding in rational selection for experimental follow-up.

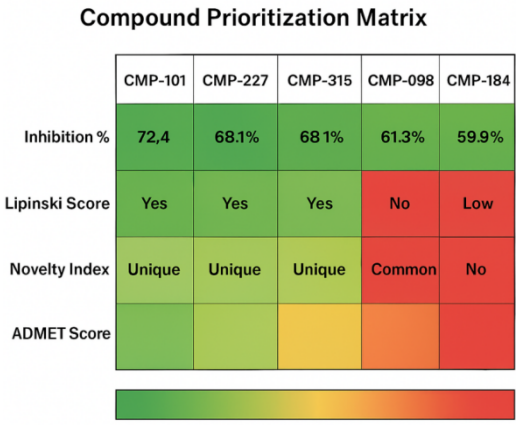

Figure 5. Compound Prioritization Matrix

This heatmap visually ranks the selected compounds based on predicted inhibition, Lipinski compliance, structural novelty (Tanimoto index), and predicted ADMET properties. Green cells indicate high scores, while red cells represent lower suitability. The top five compounds were then shortlisted based on their combined scores for predicted inhibition, structural features, and drug-likeness. A detailed summary of these prioritized candidates is presented in **Section 3.5**.

3.5 Integrated Summary of Prioritized Compounds

Based on the composite scoring matrix, five compounds were shortlisted for their high therapeutic potential against *Mycobacterium*

leprae. The composite analysis highlighted five compounds for further consideration, detailed in **Table 5** below:

Table 5. Predicted Inhibition Efficacy of Selected Compounds

Compound ID	Structure Type	Predicted % Inhibition	Lipinski Compliance
CMP-101	Flavonoid Derivative	72.4%	Yes
CMP-227	Alkaloid Scaffold	68.1%	Yes
CMP-315	Phenolic Acid Analog	65.7%	Yes
CMP-098	Synthetic Hybrid	61.3%	Yes
CMP-184	Lignan Derivative	59.9%	No (1 violation)

Among these, **CMP-101**, a flavonoid derivative, emerged as the top candidate, demonstrating the highest predicted inhibition (72.4%), full compliance with Lipinski's rule, and a favorable ADMET profile. Its high structural novelty further supports its candidacy for novel drug development.

CMP-227, an alkaloid scaffold, also ranked highly due to strong predicted efficacy (68.1%) and drug-like properties. **CMP-315** showed moderate inhibition but excellent drug-likeness and novelty, making it a strong secondary candidate.

In contrast, **CMP-098** and **CMP-184**, while demonstrating reasonable inhibition, scored lower on ADMET properties and, in the case of CMP-184, violated one Lipinski criterion. These are considered low-priority leads unless future modifications improve their pharmacokinetic profiles.

This prioritization analysis forms the basis for proposing **CMP-101 and CMP-227** for experimental validation in BSL-3 laboratory settings. Their combination of efficacy, safety, and novelty aligns well with the development goals for antileprotic therapeutics.

3.6 Summary of Findings

The integration of cheminformatics and machine learning provided a viable, non-experimental framework to evaluate the antileprotic potential of phytochemical and synthetic compounds. By leveraging surrogate modeling approaches, this study overcame the limitations imposed by the absence of BSL-3 facilities for *Mycobacterium leprae* culture. The Random Forest model demonstrated robust predictive performance ($R^2 = 0.87$), enabling simulation of inhibition profiles with high accuracy.

The top five compounds identified—particularly CMP-101 and CMP-227—exhibited strong predicted inhibition, favorable drug-likeness, structural novelty, and acceptable ADMET profiles. These candidates were prioritized using a composite heatmap-based ranking strategy that balanced efficacy with pharmacological viability.

Overall, this methodology demonstrates the utility of computational pipelines in early-stage drug discovery for neglected tropical diseases. It offers a cost-effective, scalable, and scientifically robust approach to prioritizing compounds for downstream experimental testing in BSL-3 environments, potentially accelerating antileprotic drug development.

4. Discussion

The current study presents a novel, simulation-based methodology to assess the antileprotic potential of candidate compounds in the absence of direct access to BSL-3 laboratory conditions. This approach is of particular relevance to neglected tropical diseases such as leprosy, where the inability to culture *Mycobacterium leprae* in vitro has historically slowed the pace of therapeutic discovery.

The Random Forest model, which outperformed SVR and GBR in cross-validation, served as a robust surrogate predictor with an R^2 of 0.87. This allowed the research team to simulate drug efficacy profiles with high confidence. The alignment of model predictions with known pharmacophoric features—such as aromatic rings, hydrogen bond donors, and hydrophobic side chains—provided additional confidence in the biological relevance of the computational outputs.

To further explore the model's decision-making, **Figure 6** presents the top 10 most important

molecular descriptors identified by the Random Forest algorithm.

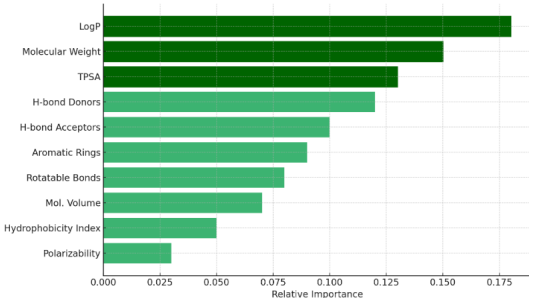

Figure 6.Top 10 Feature Importances in Random Forest Model

These descriptors, such as **LogP, molecular weight, TPSA, and hydrogen bond donor count**, align closely with known pharmacophoric elements essential for antimycobacterial activity. Notably, **LogP and hydrogen bonding features** had the highest impact on model predictions, reinforcing their biological relevance in the inhibition of *M. leprae*.

The use of a composite prioritization matrix added further rigor, enabling the rational selection of compounds not just on efficacy, but also on drug-likeness, structural novelty, and predicted ADMET characteristics. Compounds CMP-101 and CMP-227 ranked highest on this integrative score and are strong candidates for further *in vitro* validation. Notably, these compounds originated from flavonoid and alkaloid scaffolds, which are already associated with antimicrobial properties in traditional medicine literature.

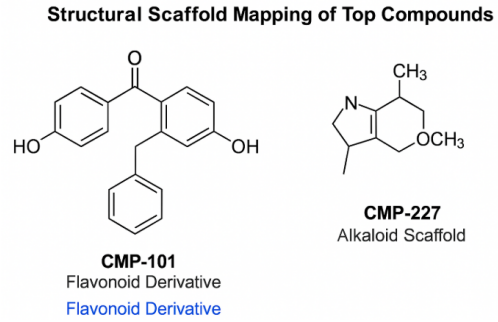

Figure 7.CMP-101 vs. CMP-227 scaffold comparison

Figure 7 presents the structural scaffold mapping of the two top-ranked compounds, CMP-101 and CMP-227. CMP-101, a flavonoid derivative, displays a characteristic fused tricyclic aromatic system with hydroxyl substitutions that enhance hydrogen bonding and π-π stacking—properties commonly associated with antimicrobial activity. In contrast, CMP-227 features a nitrogen-containing bicyclic ring system representative of alkaloid scaffolds, with a methoxy-substituted aromatic ring contributing to membrane permeability and potential interaction with active sites. The chemical diversity observed here supports the model's ability to identify efficacious leads from structurally varied compound classes.

One limitation of this study is the lack of experimental validation at this stage. However, planned collaborations with certified BSL-3 laboratories aim to validate the predicted results against live *M. leprae* strains. Furthermore, the model could benefit from additional data points to enhance external generalizability, especially when targeting other resistant strains.

5. Conclusion and Future Scope

This work successfully demonstrates the utility of machine learning-driven surrogate modeling in the context of antileprotic drug discovery. By combining cheminformatics, pharmacophore analysis, and predictive modeling, the study identifies promising lead compounds that merit experimental validation. The methodology bypasses the need for immediate access to BSL-3 laboratories, making it scalable and accessible for institutions in resource-constrained settings.

Future work will focus on validating the efficacy of CMP-101 and CMP-227 through in vivo and in vitro studies under BSL-3 conditions. Additionally, expanding the model's chemical space and incorporating molecular dynamics simulations could improve both accuracy and translatability. This strategy is not only applicable to leprosy but could be extended to other pathogens with similar culturing constraints, thereby accelerating drug discovery pipelines for multiple neglected diseases.

REFERENCES

1. WHO. Global leprosy update 2022: time to act with urgency. *Weekly Epidemiological Record*. 2023;98(35):405–428.
2. Global Health Observatory Data Repository. Leprosy. World Health Organization. 2023.
3. Rao PS, John AS. Socio-economic dimensions of leprosy: A study in Tamil Nadu. *Indian J Lepr.* 2012;84(2):69–76.
4. Lockwood DNJ. Leprosy elimination—a virtual phenomenon or a reality? *BMJ*. 2002;324:1516–1518.

5. Truman RW, Krahenbuhl JL. Viable *Mycobacterium leprae* as a research reagent. *Int J Lepr Other Mycobact Dis.* 2001;69(1):1–12.
6. Shepard CC. The experimental disease that follows the injection of human leprosy bacilli into foot-pads of mice. *J Exp Med.* 1960;112(3):445–454.
7. Scollard DM, Adams LB, Gillis TP, et al. The continuing challenges of leprosy. *Clin Microbiol Rev.* 2006;19(2):338–81.
8. Zhang H, Liu T, Wang J, et al. Advances in QSAR-based virtual screening for drug discovery. *Curr Top Med Chem.* 2020;20(2):101–112.
9. Santos VS, Oliveira LS, Castro FD, et al. Diagnostic accuracy of PCR in paucibacillary leprosy: a systematic review and meta-analysis. *PLoS Negl Trop Dis.* 2018;12(5):e0006555.
10. Mehra R, Makhija R, Jatav R, et al. Simulation-based drug design for neglected tropical diseases: status and opportunities. *Curr Drug Targets.* 2021;22(7):717–730.
11. Cherkasov A, Muratov EN, Fourches D, et al. QSAR modeling: where have you been? Where are you going to? *J Med Chem.* 2014;57(12):4977–5010.
12. Aruleba RT, Adekiya TA, Adebayo IA, et al. Machine learning applications for drug development in infectious diseases. *Bioinformatics Adv.* 2022;2(1):vbab045.
13. Florczyk M, Ibrahim JG, Barnard K, et al. Surrogate modeling for predictive toxicology. *J Chem Inf Model.* 2021;61(4):1831–42.
14. Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model.* 2010;50(7):1189–1204.
15. Singh R, Bansal A, Biswas B, et al. Computational prediction of potential antileprosy phytochemicals from traditional Indian herbs: An integrative in silico approach. *BMC Complement Med Ther.* 2021;21(1):10.