



Integrating traditional QSAR and read-across-based regression models for predicting potential anti-leishmanial azole compounds

Rajat Nandi¹ · Anupama Sharma² · Ananya Priya² · Diwakar Kumar¹

Received: 22 October 2024 / Accepted: 25 November 2024
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2024

Abstract

Leishmaniasis, a neglected tropical disease caused by various *Leishmania* species, poses a significant global health challenge, especially in resource-limited regions. Visceral Leishmaniasis (VL) stands out among its severe manifestations, and current drug therapies have limitations, necessitating the exploration of new, cost-effective treatments. This study utilized a comprehensive computational workflow, integrating traditional 2D-QSAR, q-RASAR, and molecular docking to identify novel anti-leishmanial compounds, with a focus on *Glycyl-tRNA Synthetase (LdGlyRS)* as a promising drug target. A feature selection process combining Genetic Function Approximation (GFA)-Lasso with Multiple Linear Regression (MLR) was used to characterize 99 azole compounds across ten structural classes. The baseline MLR model (MOD1), containing seven simple and interpretable 2D features, exhibited robust predictive capabilities, achieving an R^2_{train} value of 0.82 and an R^2_{test} value of 0.87. To further enhance prediction accuracy, three qualified single models (two MLR and one q-RASAR) were used to construct three consensus models (CMs), with CM2 ($\text{MAE}_{\text{test}} = 0.127$) demonstrating significantly higher prediction accuracy for test compounds than the MOD1. Subsequently, Support Vector Regression (SVR) and Boosting yielded 0.88 (R^2_{train}), 0.86 (R^2_{test}), 0.92 (R^2_{train}), and 0.82 (R^2_{test}), respectively. Molecular docking highlighted interactions of potent azoles within the QSAR dataset with critical residues in the *LdGlyRS* active site (Arg226 and Glu350), emphasizing their inhibitory potential. Furthermore, the pIC50 values of an accurate external set of 2000 azole compounds from the ZINC20 database were simultaneously predicted by CM2 + SVR + Boosting models and docked against the *LdGlyRS*, which identified Bazedoxifene, Talmecatin, Pyrinium, Enzastaurin as leading FDA candidates, whereas three novel compounds with the database code ZINC000001153734, ZINC000011934652, and ZINC000009942262 displayed stable docked interactions and favourable ADMET assessments. Subsequently, Molecular Dynamics (MD) simulations for 100 ns were conducted to validate the findings further, offering enhanced insights into the stability and dynamic behaviour of the ligand–protein complexes. The integrated approach of this study underscores the efficacy of 2D-QSAR modelling. It identifies *LdGlyRS* as a promising leishmaniasis target, offering a robust strategy for discovering and optimizing anti-leishmanial compounds to address the critical need for improved treatments.

Keywords Leishmaniasis · Azole · q-RASAR · Consensus modelling · Machine learning · Virtual screening

Introduction

Leishmaniasis, an overlooked tropical disease transmitted by vectors, is caused by various *Leishmania* parasite species and poses a substantial global health challenge, particularly

in regions with limited resources and inadequate healthcare infrastructure [1]. Among the diverse forms of leishmaniasis, with around 1.3 million new cases per year, Visceral leishmaniasis (VL) stands out as one of the most severe and potentially fatal manifestations, with around 50,000–90,000 new cases occurring worldwide annually. *Leishmania donovani*, a subspecies of the parasite transmitted through the bites of infected female sandflies, is the primary causative agent of VL in East Asia, the Indian Subcontinent, parts of South America, and the Middle East [2].

The need for new, potent, and affordable drugs against *L. donovani* is paramount due to the lack of an efficacious

✉ Diwakar Kumar
diwakar11@gmail.com

¹ Department of Microbiology, Assam University,
Silchar 788011, Assam, India

² IHub-Data, International Institute of Information Technology,
Hyderabad 500032, India

vaccine and the persistent global burden posed by the VL. Existing therapies like Amphotericin B (in its Liposomal formulation), Miltefosine, and paromomycin have limitations. Miltefosine, initially an anti-tumour agent, was the first oral VL drug but had a long half-life, severe gastrointestinal toxicity, and contraindications in pregnancy. Paromomycin-resistant *Leishmania* strains pose a risk of clinical resistance at the same time. However, partially effective liposomal Amphotericin B is costly, may lead to resistant parasites, and might be unsuitable in certain regions like East Africa [3]. With these treatment limitations, there is an urgent need for novel, effective, and affordable drug candidates.

In the last decade, many extensively researched novel anti-leishmanial compounds shared a common substructure: nitro-heterocycles. These compounds, including fexinidazole sulphone, DNDI-VL-2098, CGI-17341, DNDI-0690, and others, are currently undergoing clinical trials [4, 5]. Moreover, FDA-approved azole compounds like Ketoconazole, fluconazole, and itraconazole have proven effective in inhibiting the growth and proliferation of the parasites [6]. Azoles, including isoxazole, are crucial in many drugs and bioactive compounds, showing potent activity. For instance, Acivicin disrupts the parasite's pyrimidine pathway, while isoxazole-based hetero-retinoids and derivatives like 3-nitro and 3-amino isoxazoles exhibit superior inhibition compared to the standard drug Miltefosine [7–9]. Other compounds like bis-Arylimidamides, bis-AIAs, and AIA-azole hybrids demonstrate inhibitory effects in the single-digit micromolar range [10–12]. Recently developed Furanyl and thiophenyl azoles have shown comparable potency [13]. Other azole variants, such as aryloxy alkyl and aryloxy aryl alkyl imidazoles, Aryloxy cyclohexyl imidazoles, benzocycloalkyl azole oximino ethers, tetrahydronaphthyl azoles, and related cyclohexyl azoles, displayed significant inhibitory effects [14–17]. The availability of such a diverse array of azoles targeting the intracellular amastigote stage offers a valuable opportunity for modelling anti-leishmanial activity using QSAR approaches.

Quantitative structure–activity relationship (QSAR) modelling has been pivotal in drug discovery, streamlining research efforts and reducing costs and timelines for academic laboratories and pharmaceutical companies. QSAR investigation has been utilized to elucidate fundamental guidelines for a particular target, connecting observed inhibitory properties with ligands' molecular descriptors to establish correlations [18]. In recent decades, the scientific literature has documented numerous studies employing QSAR techniques against *Leishmania* spp. [19]. Bernal and Schmidt developed Multiple Linear Regression (MLR)-based QSAR models with genetic algorithm-driven variable selection to predict the activity of cinnamate esters against *L. donovani*, achieving high predictive accuracy ($Q^2 > 0.90$).

Although limited by a dataset of only 34 compounds, their work lays a strong foundation for studies with larger datasets. Goodarzi et al. integrated QSAR and docking approaches to evaluate diarylsulphides and sulphonamides targeting *L. donovani* α,β -tubulin. Using MLR and SVM models with DRAGON descriptors, the SVM model showed superior predictability ($Q^2 = 0.90$). They also identified four promising compounds with high predicted IC₅₀ values and docking affinity, although experimental validation was lacking. Similarly, Lorenzo et al. utilized a Random Forest-based QSAR model combined with docking to study aporphynic and aza-phenanthrene alkaloids, achieving 82% accuracy in classifying a 1,397-compound dataset and identifying multitarget candidates. Docking and RF analysis of external alkaloid sets revealed 13 potentially active compounds, with one as a promising multitarget candidate. However, a limitation was the focus on promastigote-stage assays rather than the clinically relevant amastigote stage. Ugbe et al. conducted a combined 2D and 3D QSAR study on 36 arylimidamide–azole hybrids, supported by molecular docking and pharmacokinetic analysis. Using genetic function approximation (GFA) and uninformative variable elimination-partial least square (UVE-PLS), they built robust QSAR models with high validation metrics (e.g. $R^2 = 0.9614$ for 2D-QSAR). Docking studies identified potent compounds with strong binding affinity against pyridoxal kinase and superior predicted activity compared to the reference drug pentamidine. While the study shows promise, expanding datasets and refining descriptor selection could further enhance predictive accuracy and applicability. Recently, Casanova-Alvarez et al. introduced a KNIME-automated workflow to model a highly imbalanced dataset of anti-leishmanial compounds, addressing data bias with balanced training sets and a decision tree-based consensus model. This approach achieved 71–76% accuracy across the full chemical space and up to 92% for reduced chemical spaces with higher consensus levels, offering a reproducible framework for future QSAR studies [20–24].

Despite promising experimental research on azole compounds, computational studies, including QSAR and molecular docking, have mainly focused on broad molecular classes, often needing more specificity for *L. donovani* targets and diverse azole scaffolds. In pursuit of this objective, we have developed a comprehensive workflow employing QSAR as a predictive tool to establish correlations between the molecular descriptors of a diverse set of azole compounds with their respective IC₅₀ values against intracellular *L. donovani* amastigote. Multiple Linear Regression (MLR) was employed to construct the baseline QSAR model through a representative data splitting and Genetic Algorithm (GFA)-Lasso integrated feature selection processes. The model's predictive prowess was confirmed through extensive validation metrics. Additionally, we validated

the reliability of selected descriptors using a read-across approach. We developed a q-RASAR model by combining read-across descriptors with original physiochemical descriptors, leading to the development of consensus models. Furthermore, we expanded our approach by integrating other machine learning models, such as Support Vector Machine (SVM) and Boosting, to assess the baseline model's robustness and generalization capabilities. This study is the first to combine a read-across strategy with QSAR for azole derivatives, aiming to improve anti-leishmanial activity assessments' predictive power and relevance.

Concurrently, this investigation also employed a molecular docking strategy to identify novel compounds with anti-leishmanial properties, targeting the *Glycyl-tRNA synthetase* protein. This enzyme, crucial for translation processes, holds considerable significance for the growth and survival of *L. donovani* [25], and like numerous secretory proteins, the *GlyRS* protein serves varied functions beyond protein synthesis, including transcriptional regulation and modulation of inflammatory responses [26, 27]. As a class II aminoacyl-tRNA synthetases (aaRSs) member, *L. donovani*

Glycyl-tRNA Synthetase (LdGlyRS) remains a relatively unexplored drug target compared to other aaRS family members. aaRSs are established drug targets for antibacterial and antifungal activities, featuring inhibitors like Mupirocin targeting *IleRS* in bacteria and Benzoxaborole AN2690 inhibiting *LeuRS* in methicillin-resistant *Staphylococcus aureus* and fungal-infective onychomycosis [28, 29]. Natural compounds such as febrifugine, synthetic halofuginone (HF), and borrelidin (BF) have displayed potent antimalarial properties by inhibiting *ProRS* and *ThrRS* [29]. Borrelidin, for instance, exhibited a strong affinity for *L. donovani* *ThrRS* and inhibited the promastigote stage of the parasite. Given the precedent of aaRSs as a successful and established drug target, *GlyRS*, a secretory protein of *L. donovani*, was chosen as a viable drug target.

This investigation amalgamated 2D-QSAR, virtual screening, molecular docking, ADMET profiling, and Molecular Dynamic Simulation (Fig. 1). Briefly, the most potent anti-leishmanial azole from the QSAR training dataset was docked into the *LdGlyRS* active sites for ligand–receptor interaction study. Concurrently, the most

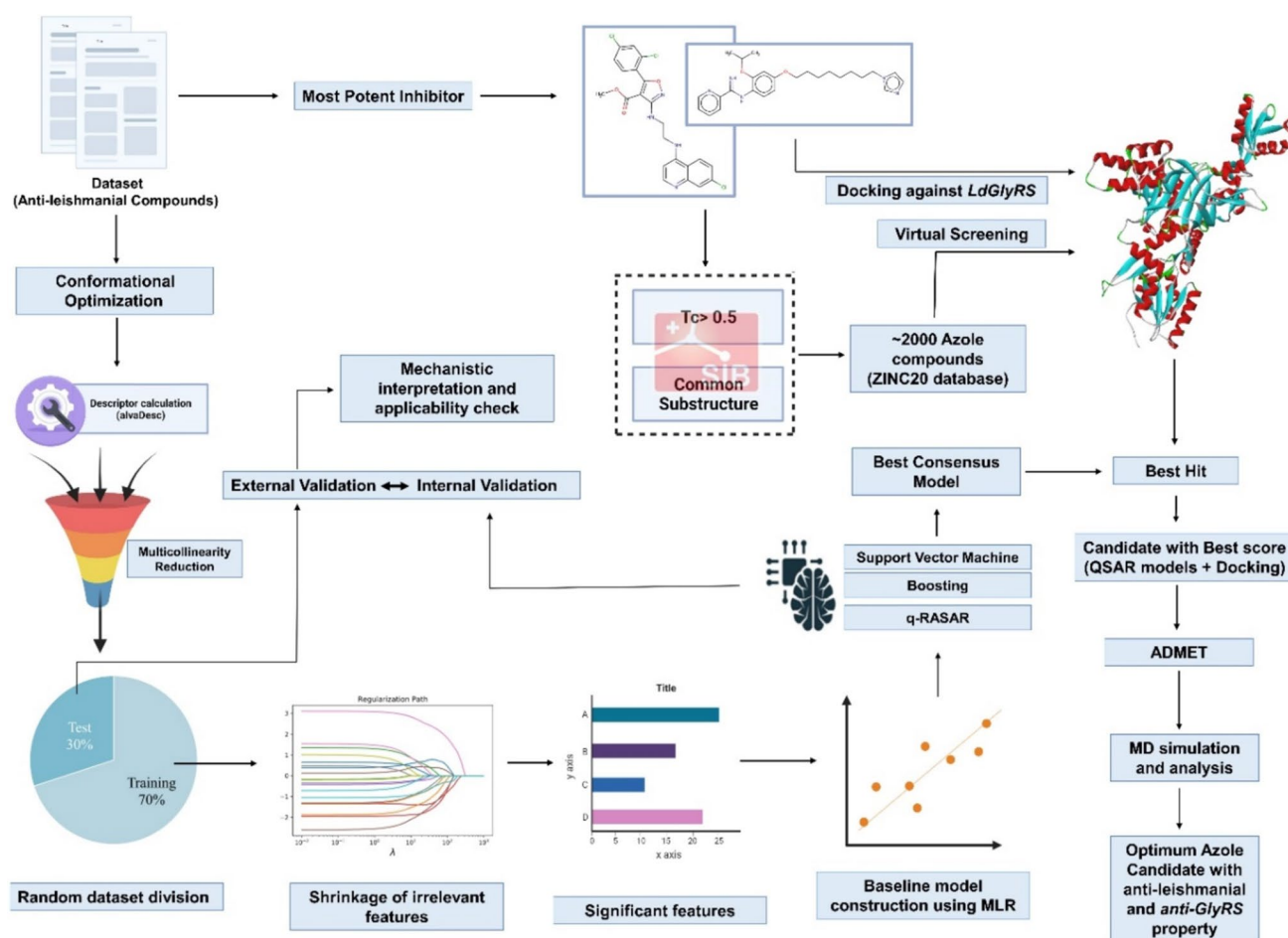


Fig. 1 General workflow of our study

potent compounds from the QSAR dataset were selected as a template for substructure search in SwissSimilarity, with a Tanimoto coefficient threshold of 0.5, and a compound library of relevant substructures was developed. This library was subjected to docking studies against the *LdGlyRS*, and their anti-leishmanial potency was predicted simultaneously using the developed consensus model (CM2 + Support Vector Regression (SVR) + Boosting). We further assessed standard computational pharmacokinetics parameters (ADMET) and drug-likeness criteria to identify the most optimal candidate with potential as both an anti-*GlyRS* and an anti-leishmanial inhibitor. Subsequent MD simulations provided insights into the dynamic behaviour of the ligand–protein interactions over time (100 ns), enhancing our understanding of binding stability and conformational changes.

Methods and methodology

Dataset selection and chemical curation

The foundation of a robust QSAR model hinges on obtaining accurate and reliable endpoint data. Thus, meticulous and vigilant data curation is paramount [18]. In this study, we meticulously retrieved a dataset of 108 azole compounds targeting *L. donovani* from the ChEMBL database with the ID: 365, in .csv format, comprising essential information such as the Molecule ChEMBL ID, Compound Key (Substance Identifier), and the Standard IC₅₀ value. The IC₅₀ values for ten diverse groups of azole compounds in this study were sourced from seven carefully selected research papers, all of which assessed the compounds' efficacy against the intracellular amastigote stage of the parasite, using the J-774A.1 mouse macrophage cell line as a host model [9, 12–17]. Notably, Amphotericin B, the control drug, consistently maintained an IC₅₀ value of 0.05 ± 0.005 μM across these studies, serving as a reliable benchmark for comparison. All the IC₅₀ values were normalized to a molar unit (μM) and subsequently transformed into a negative logarithmic scale, denoted here as pIC₅₀. Unlike conventional IC₅₀, where lower values indicate higher potency, in this context, the pIC₅₀ values directly correlate with the potency of the selected drugs.

All the drugs were drawn with the aid of Marvin 17.21.0, Chemaxon (<https://www.chemaxon.com>). Subsequently, a rigorous curation process was initiated, wherein unconnected, redundant compounds were discerningly excluded with the aid of “Connectivity” RDKit nodes extensions of KNIME version 4.3.2 (<https://www.knime.com/download>) [30, 31]. Further, the stabilization of certain aromatic and nitro groups was carried out by “Standardizer” Indigo nodes of KNIME [31]. To further ensure the uniqueness of the

dataset, a Tanimoto coefficient cutoff of > 0.9 was applied to check for similarity and remove duplicate compounds. The final curated dataset, containing 99 unique compounds, was saved in Smiles and .mol2 formats, aligning with the recommended input formats for the subsequent descriptor calculations and molecular docking process.

Descriptor calculation

Sophisticated molecular structures are condensed into molecular descriptors, encapsulating an extensive array of chemical information [32]. These descriptors comprehensively cover various physicochemical properties, ranging from topological to electrostatic attributes and intricately derived conformational details.

In the current study, computations were carried out on a specific category of descriptors [2D] that exhibit well-defined physicochemical meaning features. Constitutional descriptors, Topological indices, Functional group counts, Connectivity indices, 2D Atom pairs, Atom-centred fragments, Pharmacophore and Ring descriptors, and atom-type E-state indices, among others, were calculated using “AlvaDesc v.2.0.14” [33]. The rationale for opting for 2D descriptors was to reduce the conformational complexity and unpredictability of including 3D descriptors.

Further, redundant features, inadequate data, and descriptors where over 80% of compounds shared identical values were eliminated, resulting in a refined set of 580 descriptors for each ligand [34]. Subsequently, the Pearson correlation coefficients (ρ) were used to evaluate the intercorrelation among each descriptor. Since elevated intercorrelation ($\rho > 0.95$ or $\rho < -0.95$), referred to as multicollinearity, can lead to insignificance in regression coefficients, features exhibiting correlations with multiple features were systematically removed individually, based on the maximum occurrence of multicollinearity terms. Within the remaining features, we refined the dataset by eliminating pairwise intercorrelation terms with the minimum correlation with the response variable. The above procedure resulted in a final set of 225 refined features, which were taken as input features for QSAR modelling.

Dataset partitioning: training and test set division

A pivotal consideration in QSAR modelling with diverse scaffolds involves assessing the consistency of ligand-induced energy landscapes, ensuring that incremental shifts in ligand properties align seamlessly with the target activity function [35]. To mitigate these concerns, QSAR has been explicitly employed on molecules sharing a common scaffold group, i.e. azoles. However, for further normal distribution of the datasets, it is vital to appropriately divide the training and test sets to create a cohesive landscape wherein

the model can increase their local minima and predict a wide range of subsets of the selected compounds [36]. It is essential to highlight that an equitable distribution (70:30) between training and test set molecules must accurately represent the complete sample.

For this aim, principal component analysis (PCA) was first applied to reduce the descriptor's dimensionality (Fig. 2a), followed by dataset splitting with the aid of two different methods, i.e. the Kennard-Stone method and Euclidean distance-based method. Since using such a deterministic method to split a dataset can be bias-prone, a third

method, i.e. Random splitting, was also carried out with multiple iterations. As predicted, statistically better models were obtained through the random splitting-data division method (only the results for random splitting are presented in this manuscript). The distribution of response variables (pIC50 values) for both the training and test sets was visualized using a box plot (Fig. 2b). The random splitting ensured that a uniform representation of the analogue space was carried out so that the resulting model had more confidence in predicting test set molecules from the same training set distribution (Fig. 2c). The final models were based on 73

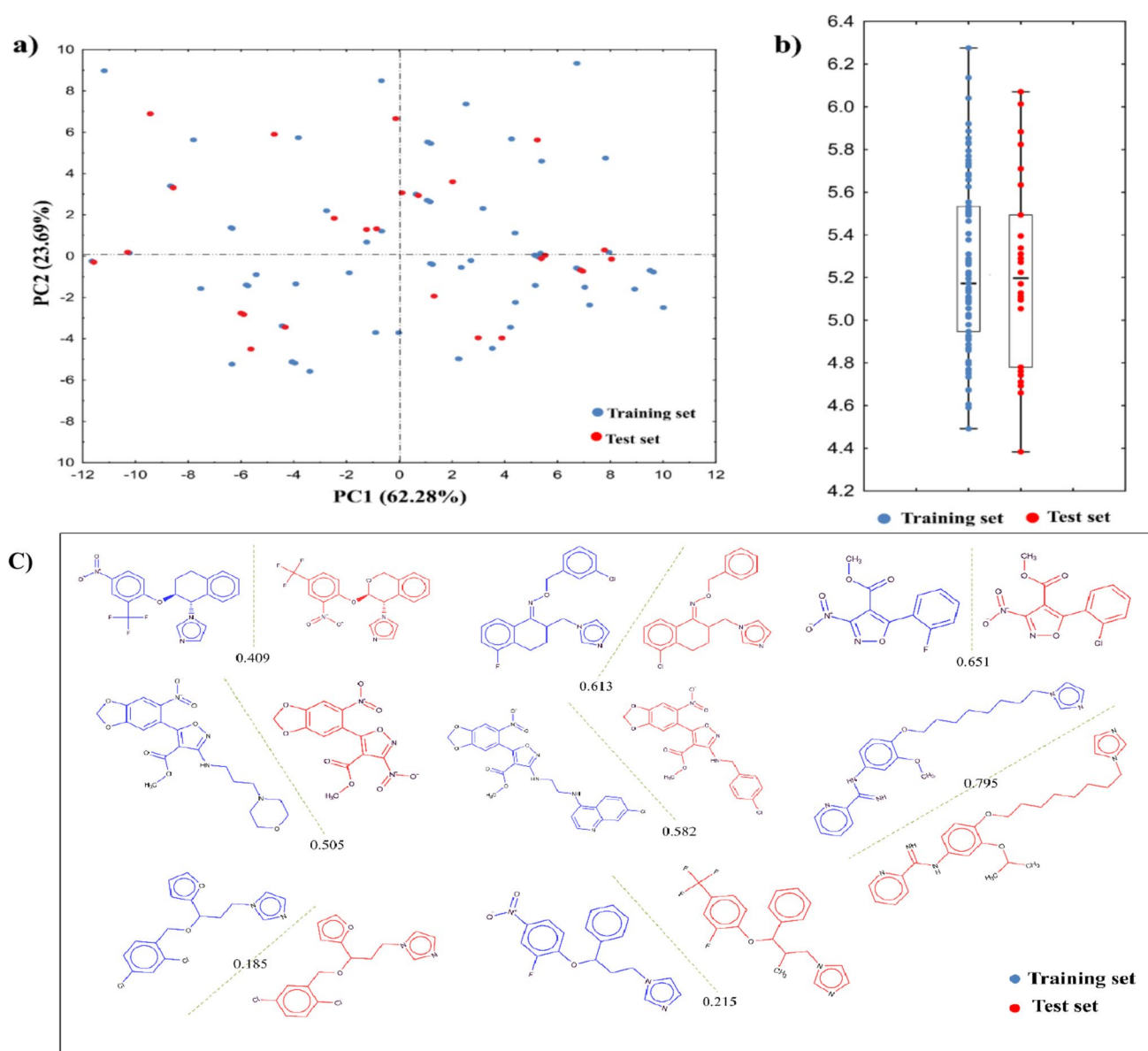


Fig. 2 **a** PCA plot depicting the location of training and test sets' molecule in a 2D chemical space, based on two principal components (62.28 and 23.69%, respectively). **b** Distribution of the response variables for the training and test sets, as visualized in the box plot.

c Chemical structure of the closest neighbouring training (blue) and test sets' (red) molecules, as selected by Random splitting. The molecules are separated by a similarity line (green), calculated by the Tanimoto coefficient

compounds in the training sets and 26 compounds in the test set.

Feature selection and QSAR model development

Our primary objective was to develop a robust, predictive, and interpretable baseline QSAR model. As the preliminary stride, we implemented Multiple Linear Regression (MLR), leveraging its stature as a benchmark standard for multivariate data analysis [37]. However, given the significantly high number of descriptors compared to the limited observations, we deftly integrated feature selection methodologies. The Genetic feature approximation (GFA) was initially employed to sift through the pool of descriptors and identify the most prevalent ones from the diverse array constituting the model's initial population. By navigating the vast expanse of potential feature combinations, the algorithm, with the aid of selection, crossover (combining features), and mutation (introducing random changes), relentlessly seeks the optimal or near-optimal solution by defined stopping criteria, be it a pre-determined maximum generation threshold (1000 number of generation) or the convergence of fitness scores (MAE-based criteria) [38]. The initial pool of 225 was reduced to 58, considering the recurrent presence of descriptors in the initial population of genetic algorithm models.

The subsequently selected descriptors were further subjected to lasso feature selection analysis. The Lasso (Least Absolute Shrinkage and Selection Operator) [39], extensively utilized in QSAR studies, is employed to control model complexity and improve performance by imposing a penalty constraint on the loss function during the modelling process [39]. A hyperparameter λ , tuneable to control the model's complexity, dictates the extent of descriptor shrinkage, where a higher value of λ results in more significant descriptor shrinkage.

The final set of 16 descriptors selected by Lasso was subjected to exhaustive combinational analysis by the MLR. MLR posits a linear association between the independent and dependent variables, defining the model as a linear equation. The equation can be expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n * X_n \quad (1)$$

where Y is the dependent variable, X_1, X_2, \dots, X_n are the independent variables, and $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ are the regression coefficients representing the weights assigned to each independent variable.

The maximum number of descriptors targeted for MLR QSAR modelling was set to 7, based on the Topliss rule, according to which incorporating an additional descriptor in the model necessitated a minimum of five compounds within the training set [40]. This extensive investigation resulted

in the generation of multiple models, further tested on the test sets, along with various statistical validation metrics described below.

q-RASAR model development

The final feature set, consisting of seven variables, was further employed for predictive modelling using the Read-Across-v4.3 software [41]. Briefly, this software assigns appropriate weights to the endpoint of each selected training compound. It uses these weights to predict the endpoint of a test or query molecule based on three distance/similarity metrics (Euclidean distance, Gaussian kernel, and Laplacian kernel). This process involves adjusting parameters of the similarity kernel function and applying various distance thresholds (ranging from 1 to 0) and similarity thresholds (ranging from 0 to 1), as well as considering different numbers of the most similar training compounds (ranging from 2 to 10) to optimize prediction quality. These RA prediction results generated a new set of descriptors comprising 18 features derived from the Gaussian Kernel function using RASAR-Desc-Calc-v3.0.1 software [42]. These additional descriptors were integrated with the initial 2D descriptors and subjected to further screening using a genetic algorithm (GA) to identify the optimal feature combination for constructing the q-RASAR-based MLR equation.

Consensus modelling

In this study, concurrent deployment of multiple individual models is utilized to predict the response endpoint of the query molecule through consensus modelling, aiming to extend the applicability domain and enhance predictive accuracy. This approach was facilitated by the "Intelligent Consensus Predictor" software [43], where the following criterion was used to generate the three consensus models.

1. CM1: The predicted value for any unknown compound was determined by averaging the predictions from n -qualified models.
2. CM2: The Weighted Average Prediction (WAP) from n -qualified models for unknown compounds was determined using specific model weights, as detailed in Roy et al. (2018b).
3. CM3: Best selection of predictions (per compound) from 'qualified' individual models.

Examining the baseline model with other machine learning models

The MOD1 (MLR baseline model) was further evaluated with the models constructed by SVR and Boosting.

Support vector regression (SVR): As a machine learning algorithm, it is designed to tackle complex non-linear regression problems using the kernel trick. The kernel function, encompassing the linear, polynomial, radial basis function (RBF) kernel (the go-to kernel with just two parameters to tweak) and sigmoid kernel, facilitates the transformation of input data into a higher-dimensional feature space. A linear hyperplane is crafted to aptly capture complex connections between input features and the target variable within this feature space. Identifying these hyperplanes relies on support vectors corresponding to data points near the hyperplane [44, 45]. The function equation for SVR typically takes the form:

$$f(x) = \sum_{i=1}^n (\alpha_i \cdot K(x_i, x)) + b, \quad (2)$$

where $f(x)$ represents the prediction for the input x , α_i are the Lagrange multipliers obtained during training, x_i is the support vectors, $K(x_i, x)$ represents the kernel function evaluated between the support vectors x_i and the input x , and b is the bias term. The SVR model's effectiveness relies heavily on adequately optimized hyperparameters. These pivotal parameters include the Regularization parameter (C), which governs the trade-off between training error and model complexity, which is crucial for avoiding overfitting. Tuning ' C ' allows for adjustments between model simplicity (with smaller values) and increased complexity (with larger ones). The Kernel Coefficient (Gamma) in the RBF kernel shapes the kernel and significantly affects model complexity. Adjusting 'Gamma' fine-tunes the kernel width, where smaller 'gamma' values result in wider kernels, while larger values yield narrower kernels. The Epsilon parameter sets the margin for classification errors, determining the width of the epsilon-insensitive tube within the SVR model. Finally, the degree parameter in polynomial kernels allows for the capture of complex data relationships, with its adjustment serving to strike a balance between model intricacy and overfitting prevention. Precise hyperparameter tuning is vital for achieving a well-fitted and reliable SVR model [43, 45].

Boosting, a machine learning technique, refers to a sequential ensemble learning method where a series of weak predictive models are combined to develop a predictive model that is both resilient and accurate. These weak models, typically decision trees, also known as weak learners or base learners, are trained on subsets of data, focusing on misclassified instances from previous models. The output of each weak learner is combined through a weighted voting or averaging scheme to form the final prediction of the boosted model. The weights assigned to each weak learner are determined based on their performance during training. The overall idea is to leverage the

collective knowledge of multiple weak models to improve the accuracy and predictive power of the QSAR model [46].

The hyperparameters optimized for model generation were Learning Rate (LR), indicating the step size at which the boosting algorithm updates its model. A lower learning rate often leads to a more robust and accurate model but may require more iterations. Additive Terms reflect the number of weak learners (typically decision trees or stumps) sequentially combined to form a robust predictive model. Random Data Proportion is a percentage of data randomly selected for each boosting iteration. This random subsampling can introduce variability into the model and help prevent overfitting, as well as the Subsample Proportion, where each boosting iteration uses a subsample of the available data. This technique, known as stochastic gradient boosting, further enhances model generalization by introducing randomness [46].

All the statistical analysis, including feature selection (Lasso) and Model building (MLR, SVR, Boosting), was carried out in the Statistica Version 14.0.0.15 (<https://www.tibco.com/products/tibco-statistica>) [47].

Model validation and applicability domain check

The final selected MOD1 model was validated rigorously using various statistical methods like leave-one-out cross-validation R^2_{loocv} (Iteratively excluding one sample from the training set to construct models predicting the excluded sample's target value for validation) and correlation coefficient (Q^2) [48, 49], explained as

$$Q^2 = 1 - \frac{\sum (e_i - p_i)^2}{\sum (e_i - P_{\text{mean}})^2}, \quad (3)$$

where e_i and p_i are the experimental and predicted values of sample i , respectively.

Further, Y-randomization was conducted to validate the robustness of the baseline QSAR model. Briefly, the response variable (pIC_{50}) was randomly shuffled multiple times (50 iterations), and new models were built using the same algorithm and parameters as the original model. Performance metrics (R^2 and Q^2) were calculated for each randomized model and compared to the original model. This process ensured that the predictive power of the QSAR model was not due to chance correlations [50].

External prediction regression coefficient (R^2_{Test}), Standard deviation of residuals (S_{Train} & S_{Test}), variance ratio (F), p -value, Standardized Prediction Error Sum of Squares (S-Press), and Standard Deviation of Prediction Errors in the Sum of Squares (SDEP) were also used to assess the model's robustness [49].

For the SVR model, tenfold cross-validation was employed to explore the ideal kernel function and its associated parameters. Subsequently, leave-one-out cross-validation (LOOCV) was used to evaluate the model's overall reliability. The selection of the kernel function, capacity parameter C , insensitive loss function ε , and their respective gamma g was based on assessing the lowest root mean square error (RMSE) and highest correlation coefficient R for evaluation criteria [45].

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (p_i - e_i)^2}{N}} \quad (4)$$

Here, N is the number of total samples, and e_i and p_i are the experimental and predicted values of sample i , respectively. The RMSE quantifies the overall magnitude of the errors in the model's predictions. A lower RMSE value indicates that the model's predictions closely match the observed values, suggesting higher accuracy and precision.

Applicability domain check

As the QSAR model is constructed using a constrained dataset of available anti-leishmanial azole compounds, this model's applicability is limited to representing the chemical diversity within the training set. Thus, a proper range of small molecules must be defined, ensuring their relevance within the model's applicable scope (Applicability domain) [51]. This study utilized the William plot to define the applicability domain, enclosing a square area spanning 3 standard deviation units. A leverage threshold, denoted as h^* , was determined using the equation: $h^* = 3(k + 1)/n$, where k represents the number of selected descriptors and n is the dataset size. William plot was generated to identify Y-outliers (data points displaying considerable deviations in response values, i.e. std. residual value of $> \pm 3$) and influential compounds characterized by their substantial impact on regression, often having extreme descriptor values or leverage values.

Virtual screening for a novel anti-leishmanial, anti-GlyRS compound

Virtual screening aimed at identifying potential inhibitors targeting anti-leishmanial *LdGlyRS* was executed via the similarity and standard substructure search tool in the SwissSimilarity web [52] to further screen the ZINC20 database [53]. The approach involved utilizing the potent inhibitors 44, 59, 68, 82, and 90 from the QSAR dataset as templates for similarity searches against the 10 million compounds of the ZINC drug-like library using the 2D FP2 fingerprints screening method. Subsequently, a filter of the Tanimoto coefficient ($T_c \geq 0.5$) was applied to prune the

unrelated compounds. Additionally, a substructure search that identified compounds containing isoxazole, imidazole, and oxazole rings was used, resulting in a final list of ~2000 compounds. All the compounds with their respective ZINC ID and SMILES were downloaded in a.csv file. The smiles were then used to calculate the molecular descriptor from the AlvaDesc, as described before. In the following step, the most optimal consensus model (CM2 + SVR + Boosting) was employed to predict the biological activity of the potential anti-leishmanial inhibitors. Concurrently, the whole compound sets were acquired in Mol2 format and transformed into.pdbqt format via Open Babel [54] for subsequent molecular docking evaluations against *LdGlyRS*.

Molecular docking

The amino acid primary sequence of *L. donovani Glycyl-tRNA Synthetase (LdGlyRS)* was retrieved from the kegg database (LDBPK_364030). The sequence similarity search for *LdGlyRS* within the protein data bank (PDB) was performed with Blastp (<https://blast.ncbi.nlm.nih.gov/>) [55]. The crystal structure of *Thermus thermophilus GlyRS* (PDB ID: 1ggm), with sequence similarity of 27% and a good query cover of 89%, was selected as the template structure to build a three-dimensional model of *LdGlyRS*, with the aid of Robetta software (<http://robetta.bakerlab.org>) [56]. The quality assessment of the model was assessed using the Ramachandran plot, Verify3D, and ERRAT plot, sourced from the PROCHECK webserver (<https://saves.mbi.ucla.edu/>) (Supple. Figure S1). The homolog structure was energy-minimized and superimposed to the template backbone (PDB ID: 1ggm) with the help of Pymol software [57], resulting in a superimposed RMSD value of 1.506 Å (Supple. Figure S2).

The active site exploited in the study (Supple. Figure S2) was defined by the Multiple sequence alignment and PDB template: 1ggm, referenced Glycyl-Adenosine-5'-Phosphate bound sites of *Glycyl-tRNA Synthetase* from *Thermus thermophilus* (Supple. Fig. S2). The docking study was performed using the AutoDock Vina software [58], where the centre grid box was configured with dimensions of $x = 5.629$ Å, $y = 21.280$ Å, and $z = -12.099$ Å. The docked protein–ligand complexes were visualized in PyMol software [57], and hydrogen bond interaction (2D) was studied using Discovery Studio Visualizer v21.1 [59].

Molecular dynamics simulations

Molecular dynamics (MD) simulations were performed to examine the dynamic behaviour and stability of protein–ligand complexes using GROMACS 2021.4 with PLUMED 2.8.0 [60]. Ligand structures were initially optimized with GAUSSIAN09, and the output files were

converted to fch format for RESP atomic charge calculations using Multiwfn 3.8 (dev) [61]. These charges were then integrated into the ligand topology files, which were generated using AmberTools18. The General Amber Force Field (GAFF) was applied for the MD simulations to parameterize the ligands, while the protein was modelled using the AMBER99SB-ILDN force field. The systems were solvated in a cubic water box with the TIP3P model, maintaining a 2-nm buffer between the protein and the box edges to ensure adequate hydration. Ions were added to neutralize the system and replicate physiological conditions. Energy minimization was done using the steepest descent algorithm until the force dropped below 1000 kJ/mol. Electrostatic interactions were treated with the Particle Mesh Ewald (PME) method, while van der Waals interactions were computed using a cutoff range of 10–12 Å. The Linear Constraint Solver (LINCS) maintained hydrogen bond constraints throughout the simulation. The system was able to be equilibrated in two stages. First, an NVT ensemble was run for 100 ps at constant volume, followed by an NPT ensemble for another 100 ps at 1 bar pressure using the Parrinello–Rahman barostat. The temperature was controlled at 300 K during both equilibration stages using the Berendsen thermostat. Production MD simulations were subsequently conducted for 100 ns with a time step of 100 ps, where the centre of mass translation and rotation of the protein–ligand complexes were restrained to ensure stable interactions.

Post-simulation analysis was carried out using GROMACS utilities and Visual Molecular Dynamics (VMD) software [62] to investigate the structural and

dynamic properties of the protein–ligand complexes. Several metrics were analysed, including the Root Mean Square Deviation (RMSD) of backbone atoms, Root Mean Square Fluctuation (RMSF) of individual residues, Radius of Gyration (Rg), Solvent-Accessible Surface Area (SASA), and hydrogen bond formation. Principal Component Analysis (PCA) assessed the conformational space explored by the apoprotein and ligand-bound complexes. Binding free energy (ΔG) calculations were performed using gmX_MMPBSA [63], incorporating the contributions from van der Waals interactions (ΔV_{DWAALS}), electrostatic energy (ΔE_{EL}), polar solvation energy (ΔE_{PB}), and non-polar solvation energy (ΔE_{NPOLAR}). The total binding free energy (ΔG_{TOTAL}) is the sum of the gas-phase free energy (ΔG_{GAS}) and solvation free energy (ΔG_{SOLV}).

Results and discussion

Baseline QSAR model

In pursuit of a robust baseline model, we employed lasso selection to iteratively shrink the descriptor set by gradually increasing the hyperparameter λ (Fig. 3a). The optimal value of λ was found to be 0.02 through tenfold cross-validation. The optimal λ yielded a set of 16 descriptors, forming the basis for an exhaustive model exploration that spanned from straightforward two-feature models to more intricate seven-feature linear models. The model candidates underwent thorough scrutiny of their performance on the training

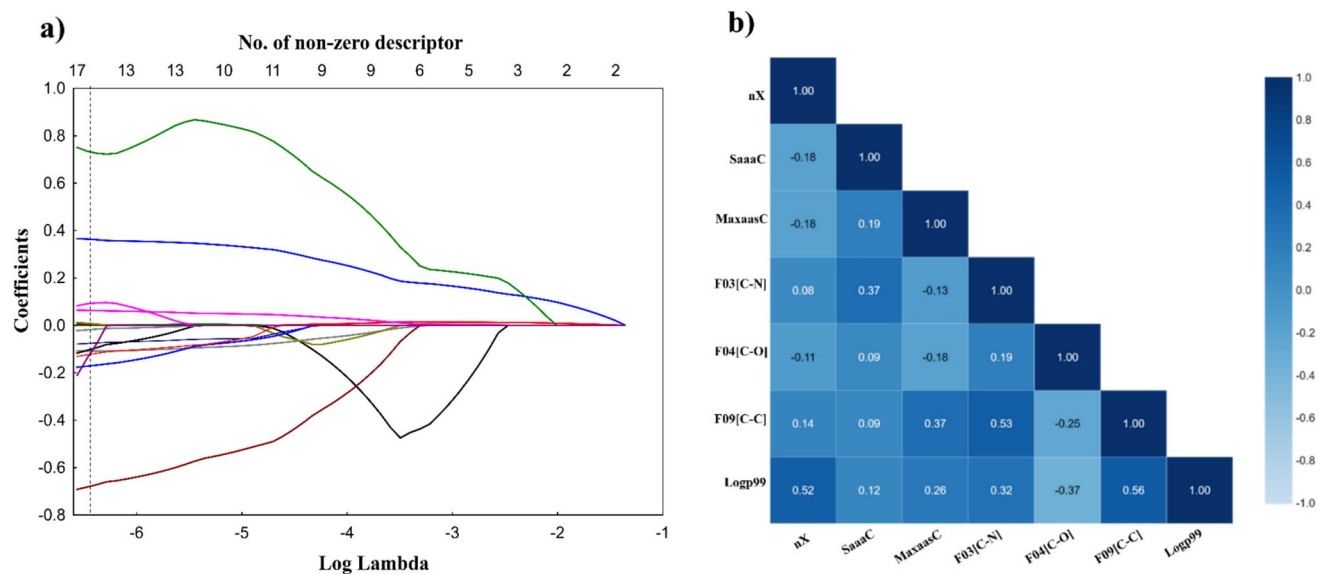


Fig. 3 **a** Using the lasso regression, the coefficients of the descriptors were shrunk as λ (lambda) increases. Each curve, distinguished by a different colour, represents the descriptor coefficient shrinkage. The upper x-axis denoted the count of descriptors retaining non-zero

coefficients at specific λ values. The optimal λ value (0.02) was determined through tenfold cross-validation. **b** Pearson correlation coefficient matrix heat map of seven selected features for MOD1

set with selection criteria $R^2 > 0.75$, $R^2_{\text{loocv}} > 0.75$, followed by statistical significance parameter p -value < 0.05 , which resulted in the selection of seven final descriptors, with small correlation and good orthogonality (Fig. 3b).

The final baseline model (Fig. 4) for pIC₅₀ against intracellular amastigote of *L. donovani* (Eq. 5) was highly robust and reliable: it effectively explained 82% of the variance in the training set (with a LOO variance of 77%) and 87% of the variance in the test set (R^2_{test}).

$$\begin{aligned} \text{pIC}_{50} = & -0.1623 \text{ nX} + 0.4182 \text{ SaaaC} \\ & - 0.6393 \text{ MaxaasC} - 0.0966 \text{ F03 [C-N]} \\ & + 0.0238 \text{ F04 [C-O]} + 0.0576 \text{ F09 [C-C]} \\ & + 0.3568 \text{ Logp99} + 4.1627 \end{aligned} \quad (5)$$

$$\text{nTrain} = 73; R^2 = 0.821; R^2_{\text{adj}} = 0.8017; R^2_{\text{loocv}} = 0.776;$$

$$S_{\text{Train}} = 0.182; F = 42.588; p < 0.0001;$$

$$S_{\text{Press}} = 0.204; \text{SDEP} = 0.194$$

$$\text{nTest} = 26 : R^2_{\text{test}} = 0.872; R^2_{\text{test adj}} = 0.8223; S_{\text{Test}} = 0.188$$

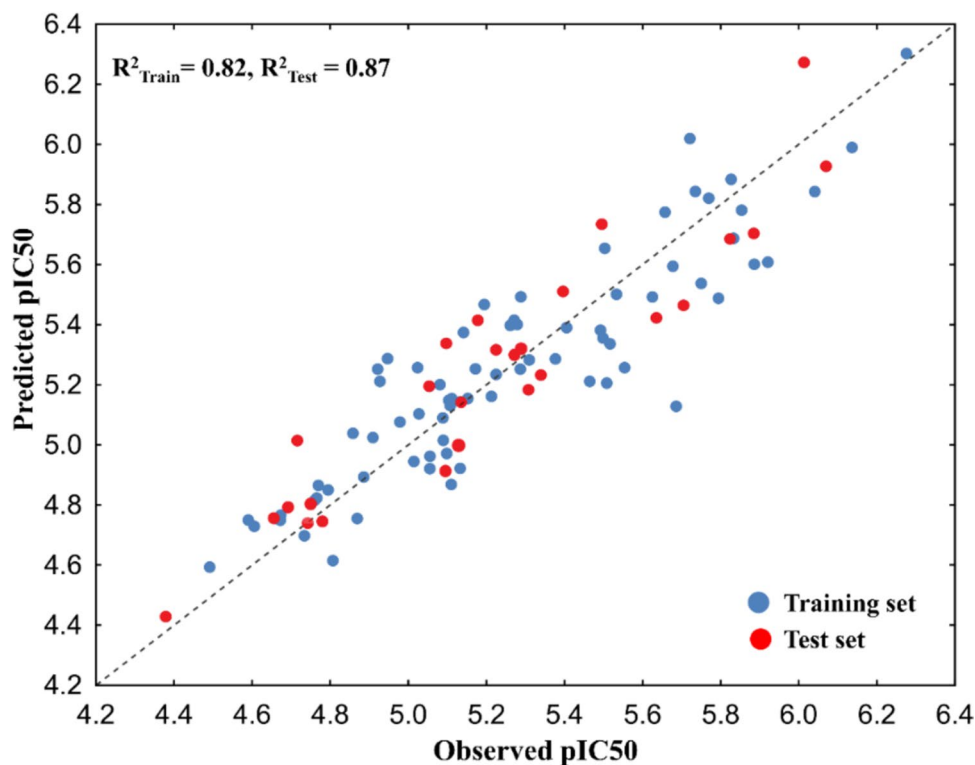
Here, nX represents the number of halogen atoms in the molecule; SaaaC and MaxaasC are descriptors related to carbon connectivity in aromatic systems; F03 [C-N], F04 [C-O], and F09 [C-C] indicate fragment-based contributions of specific atomic pairs; and LogP99 refers to

the Wildman–Crippen octanol–water partition coefficient (LogP), which quantifies molecular hydrophobicity. The constants in the equation account for the baseline effect.

The model's statistical validity was reinforced by calculating various validation metrics. A high and statistically significant $R^2_{\text{train adj}}$ (0.801) and $R^2_{\text{test adj}}$ (0.8223) values indicate the model's goodness-of-fit and generalization capabilities. Additionally, the model demonstrated robustness, as evidenced by a low standard deviation of residuals ($S_{\text{Train}} = 0.182$, $S_{\text{Test}} = 0.188$), a high F-value (42.58), low values of Prediction Error Sum of Squares ($S_{\text{Press}} = 0.204$) and Standard Deviation of Prediction Errors in the Sum of Squares ($\text{SDEP} = 0.194$), and a significant p -value. Furthermore, the model's reliability was checked using the "Xternal Validation Plus" software [64], which indicated the absence of systematic error and classified the prediction quality as "Good" for the test set based on MAE-based criteria ($\text{MAE}_{95\% \text{ test}} < 0.1 \text{ TSR}$ (training set range) and $\text{MAE}_{95\% \text{ test}} + 3\text{SD}_{95\% \text{ test}} < 0.2 \times \text{TSR}$) (Supple. Table S3).

The Y-randomization test further confirmed the robustness of the QSAR model. Across y, the models built on the randomized datasets consistently exhibited significantly lower predictive performance than the original model. The average R^2 and Q^2 values for the randomized models were approximately 0.091 and -0.157 (Supple. Table S4). These results demonstrate that the QSAR model's predictive ability arises from genuine correlations between molecular descriptors and biological activity rather than chance.

Fig. 4 The baseline model depicting observed pIC₅₀ vs. predicted pIC₅₀ was developed by multiple linear regression (MOD1)



Moreover, an additional eligible GFA-Lasso MLR model (MOD2), depicted in Eq. 6, was employed for “consensus modelling” in the subsequent section.

$$\begin{aligned} \text{pIC}_{50} = & -0.0762 X\% + 0.4410 \text{SaaaC} \\ & - 0.6206 \text{MaxaasC} - 0.1050 \text{F03}[\text{C} - \text{N}] \\ & + 0.0249 \text{F04}[\text{C} - \text{O}] + 0.0561 \text{F09}[\text{C} - \text{C}] \\ & + 0.3515 \text{Logp99} + 4.2499 \end{aligned} \quad (6)$$

In a comparative analysis, 3D descriptors from AlvaDesc, including 3D autocorrelation, RDF descriptors, and 3D-MoRSE descriptors, were employed with the same training and test sets to create QSAR models. As expected, the standalone baseline model (MOD1) demonstrated better performance; however, the consensus model, which combined set A (MOD1 + MOD2) and set B (RDF descriptors + 3D-MoRSE descriptors), yielded improved results with an R^2 value of 0.86 for training and 0.87 for testing (Supple. Table S5). The consensus model was developed using OCHEM (<https://ochem.eu/>) [65].

q-RASAR model development and consensus modelling

The seven descriptors from the MOD1 model were utilized as inputs for Read-Across predictions with default parameters: a sigma value (σ) of 1 for the Gaussian kernel function, a gamma value (γ) of 1 for the Laplacian kernel function, and distance and similarity threshold values set at 0.5 and 0, respectively. Among the three kernels, the Gaussian Kernel function yielded the most favourable results, with Q^2_{F1} (0.767) and MAE_{pr} (0.155) (Supple. Table S6). These parameters were subsequently employed to generate 18 new descriptors, which were used to develop the q-RASAR model. This model was developed by integrating two RASAR descriptors and four 2D original descriptors from the baseline model (Eq. 7).

$$\begin{aligned} \text{pIC}_{50} = & 0.9038 \text{RA functions}[\text{gk}] + 0.2095 \text{gm_class}[\text{gk}] \\ & - 0.0822 \text{MaxaasC} - 0.0176 \text{F03}[\text{C} - \text{N}] \\ & + 0.0114 \text{F04}[\text{C} - \text{O}] + 0.0080 \text{F09}[\text{C} - \text{C}] \\ & + 0.4333 \end{aligned} \quad (7)$$

The conclusive statistical parameters of the q-RASAR model, as depicted in Table 1, displayed significantly enhanced internal robustness (R^2 of 0.87, R^2_{loocv} of 0.84) compared to the baseline model. However, the external predictive quality capability ($R^2_{\text{test}} = 0.762\text{--}0.795$, $\text{MAE}_{\text{test}} = 0.161$) experienced a relative reduction. Consequently, the baseline model was retained for subsequent testing with machine learning models.

Consensus modelling was implemented to enhance the external predictive accuracy and broaden the model's applicability domain by combining the two qualified MLR models with the q-RASAR model. Table 1 comprehensively compares the three models and their corresponding consensus models (CMs). According to the statistical parameters for the test sets, CM2 emerged as the superior model, exhibiting an R^2_{test} range of 0.854–0.874 and an MAE_{test} of 0.129. The CMs were later used to screen the untested true external prediction set from the ZINC20 database.

Analysis of modelled descriptors of the MOD1

Logp99, F09 [C–C], F03 [C–N], F04 [C–O], MaxaasC, SaaaC, and nX represent the seven physicochemical descriptors included in the baseline model (Table 2). According to coefficient normalization, MaxaasC, SaaaC, Logp99, and nX possess the most significant makeshift for the modelled response against *L. donovani*. Conversely, all other variables exhibit equal or decreasing significance in the equation, as reflected by their standard coefficients.

MaxaasC emerges as the most significant descriptor in this model. Belonging to the class of Atom-type E-state indices descriptors, it provides insights into the electronic and topological state of carbon atoms within a molecule.

Table 1 The validation metrics (internal and external) of the two MLR models, the q-RASAR model, and the external validation metrics of the three consensus models (CMs)

Scheme	R^2	R^2_{adj}	R^2_{loocv}	SDEP_{tr}	MAE_{test}	Q^2_{Fn}	CCC_{test}	r^2_m	Δr^2_m
MOD1 (baseline)	0.8210	0.8017	0.7760	0.1939	0.14155	0.836–0.858	0.92535	0.79211	0.11454
MOD2	0.8199	0.8005	0.7688	0.1970	0.14545	0.833–0.856	0.92395	0.7924	0.11676
q-RASAR	0.8723	0.8607	0.8454	0.1611	0.16117	0.762–0.795	0.8802	0.64046	0.17904
CM1	–	–	–	–	0.12927	0.855–0.875	0.93021	0.75809	0.11165
CM2	–	–	–	–	0.1272	0.854–0.874	0.93034	0.76487	0.11093
CM3	–	–	–	–	0.14967	0.798–0.826	0.90502	0.72969	0.14225

The bold entries in highlight the best-performing models in the study. Specifically, they represent the best model among the MLR approaches (MOD1) and the best consensus model (CM2) based on the presented validation metrics

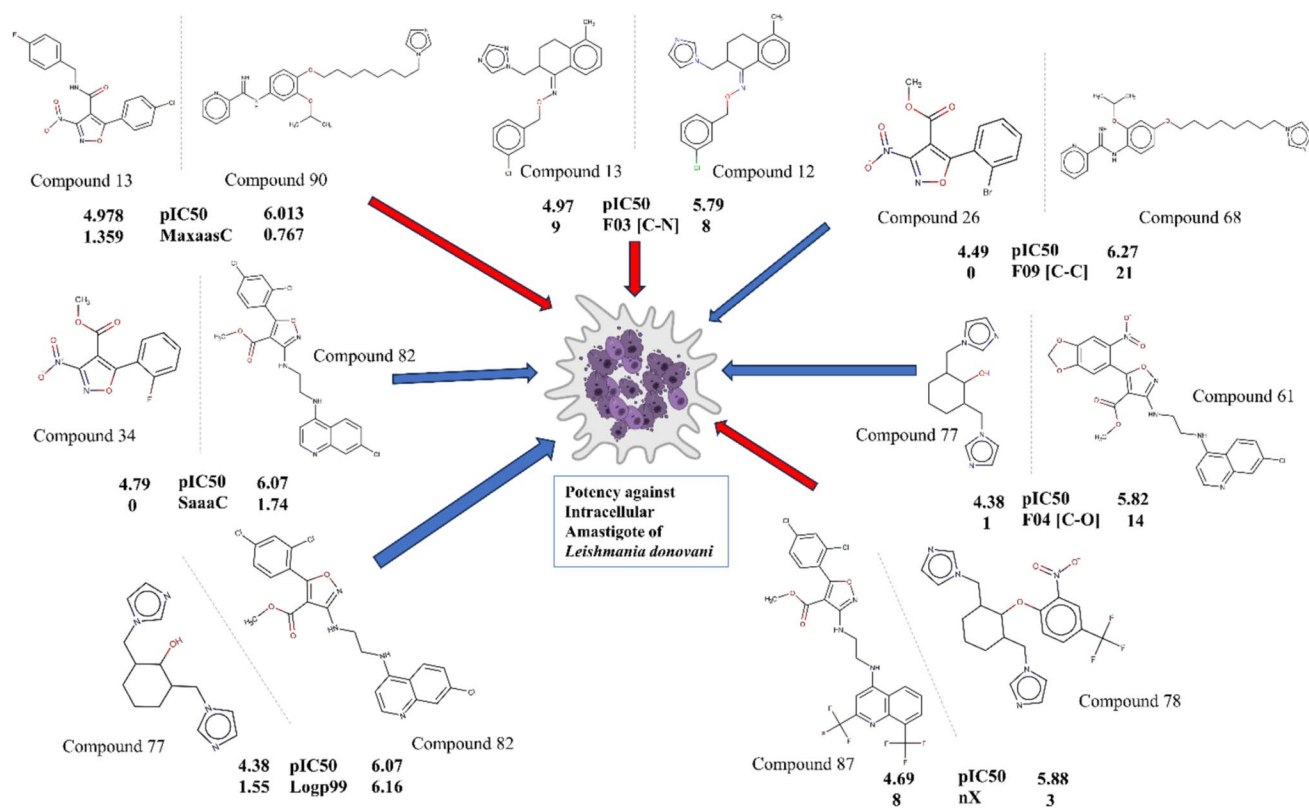
Table 2 Molecular descriptors are involved in three models (MOD1 + SVR + Boosting) and their meanings

Feature	Meaning	Class	Contribution to potency
nX	Number of halogen atoms	Constitutional indices	Negative
SaaaC	Sum of aaaC E-states	Atom-type E-state indices	Positive
MaxaasC	Maximum aasC	Atom-type E-state indices	Negative
F03(C–N)	Frequency of C–N atomic pair at topological distance 3	2D Atom Pairs	Negative
F04(C–O)	Frequency of C–O atomic pair at topological distance 4	2D Atom Pairs	Positive
F09(C–C)	Frequency of C–C atomic pair at topological distance 9	2D Atom Pairs	Positive
Logp99	Wildman–Crippen octanol–water partition coefficient (LogP)	Molecular properties	Positive

"E-state" refers to the Electrotopological State, which integrates information about an atom's electronic environment and its connectivity within the molecular structure. Specifically, MaxaasC represents the maximum electron accessibility for carbons with two aromatic bonds and one single bond. Our study revealed a negative correlation between this descriptor and the modelled response. Figure 5 illustrates this relationship, where the pIC50 value decreases from 6.013 (Compound 90) to 4.978 (Compound 13) as MaxaasC increases from 0.767 to 1.359, respectively. This correlation underscores the significance of MaxaasC, i.e. aromatic

bonds, in influencing the bioactivity of the compounds under investigation.

SaaaC, also an Atom-type E-state indices class of descriptors, holds the second-highest significance among the descriptors. Its presence within the compound's structure demonstrates a favourable association with increased bioactivity, as depicted in Fig. 5. Notably, a distinct trend emerges in Compounds 82 and 34: as the SaaaC value decreases from 1.74 to 0, there is a corresponding decrease in the pIC50 value, which declines from 6.07 to 4.79, respectively. SaaaC is derived by summing the E-states of all carbon atoms

**Fig. 5** Positive (blue) and Negative (red) contribution of the selected descriptors towards the pIC50 of the corresponding compounds

within the molecule, featuring three aromatic bonds. This descriptor offers valuable insights into the reactivity of such atom types, as it increases proportionally with the number of carbon atoms possessing three aromatic bonds and their corresponding reactivity.

The third most significant feature in the model, as indicated by its standardized coefficient, was LogP99. This molecular property descriptor, calculated using the Wildman–Crippen method, estimates LogP based on predefined atomic contributions. The “99” in LogP99 refers to its specific implementation or variation within the descriptor set generated by the AlvaDesc software. LogP99 positively influences the pIC50 against *L. donovani* as a measure of lipophilicity. This study observed that the compounds containing chlorine or other halogen components exhibit increased molecular size and London dispersion forces, indirectly elevating lipophilicity [66]. Consequently, azole compounds exhibiting greater lipophilicity demonstrated enhanced penetration through the cell wall of *L. donovani*, resulting in increased potency.

nX represents the total halogen group count and is a class of constitutional indices. In this study, nX exhibits a negative correlation with the modelled response. Figure 5 shows that less potent compounds (pIC50 = 4.69) are associated with a higher nX value of 8, while highly potent compounds (pIC50 = 6.27) exhibit lower nX values of zero. Further exploration reveals that highly potent anti-leishmanial compounds, bearing at least 2–3 halogen groups such as chlorine (Cl) and fluorine (F), glean substantial benefits in terms of compound potency, as seen in compounds 87 and 68. However, caution is warranted as excess halogen groups (> 4) diminish the compound's potency.

F03 [C–N], a 2D atom pair descriptor, exhibits an inverse correlation with the modelled endpoint in this study. This descriptor specifically encapsulates the frequency of Carbon–Nitrogen atomic pairs at a topological distance of 3. The unique attributes of this descriptor are observed in rows 18 and 19; notably, both the compounds share a common substructure, except for an additional Nitrogen group added to the ortho-position of the benzene ring in the structure of row 19. This addition led to a minor increase of a single unit in F03 [C–N] value, shifting from 8 to 9, respectively; it also corresponded to a significant reduction in the pIC50 value, transitioning from 5.79 (Compound 12) to 4.97 (Compound 13).

F09 [C–C] is another descriptor in the model that falls under the class of 2d atom pair descriptor and signifies the Frequency of the Carbon–Carbon atomic pair at topological distance 9. According to the standardized coefficient of this model, F09 [C–C] exhibits a positive relationship with the modelled endpoint. As the F09 [C–C] value increases from zero (Compound 26) to 21 (Compound 68), the pIC50 value increases from 4.49 to 6.27, respectively (Fig. 5).

The least important descriptor in this baseline model is F04 [C–O]; once more a 2D atom pairs descriptor, it delves into the Frequency of Carbon–Oxygen atomic pair at topological distance 4. The model suggests that the presence of this descriptor helped to improve (positively) the pIC50 of the compounds. The positive attributes of this descriptor were seen in Compound 77 and Row 61. As the F04 [C–O] increases from 1 to 14, the pIC50 value increases from 4.38 to 5.82, justifying its proportional relationship with the pIC50.

Examining the baseline model with SVR and Boosting

Throughout the SVR modelling process, optimal model parameters were identified via a grid search approach, where the primary goal was to uncover the parameters that would yield the lowest RMSE values across three distinct kernel functions: LKF, RBF, and PKF.

RMSE values were calculated using different parameters through tenfold cross-validation in the grid search. For LKF and PKF, this included the capacity parameter (C) ranging from 1 to 250 (step = 10) and the ϵ -insensitive loss function parameter (ϵ) from 0.01 to 0.1 (step = 0.01). Meanwhile, the RBF kernel function involved C (ranging from 1 to 500, step = 10), ϵ (ranging from 0.01 to 0.1, step = 0.02), and Gamma (g , ranging from 0.5 to 1.5, step = 0.1). This process unveiled minimum RMSE values of 0.41, 0.45, and 0.50 for the RBF, PKF, and LKF kernel functions, respectively (Supple. Figure S3). Consequently, the optimal SVR model attained was with the RBF kernel function, characterized by the parameters $C = 70$, $\epsilon = 0.07$, and $g = 0.6$, which yielded a correlation coefficient value of 0.940 (Fig. 6a).

The LOOCV method was implemented to internally validate the QSAR model's predictive capability. The RMSE value in the LOOCV_{test} using SVR for the training set was 0.263 (Table 3), thus signifying the model's ability to generalize effectively.

The baseline model was further evaluated with the model constructed by the ensemble tree method, i.e. Boosting. The boosting algorithm was optimized with the following key parameters: Learning Rate (LR): 0.1200, number of additive terms: 400, Random Data Proportion: 30%, Subsample Proportion: 70%, Stopping parameters: Maximum Number of Cases: 07, Maximum Number in Child Nodes: 01, Maximum Number of Levels: 10, and Maximum Number of Nodes: 03. Figure 6b depicts the final boosting model trained using 524 trees. The precision of this model's predictive capacity is evident through its notable correlation coefficient, R^2 value of 0.927, and Q^2 of 0.921.

Results of independent SVR test: Predictions were extended to an independent dataset using models developed from the training dataset to demonstrate the generated

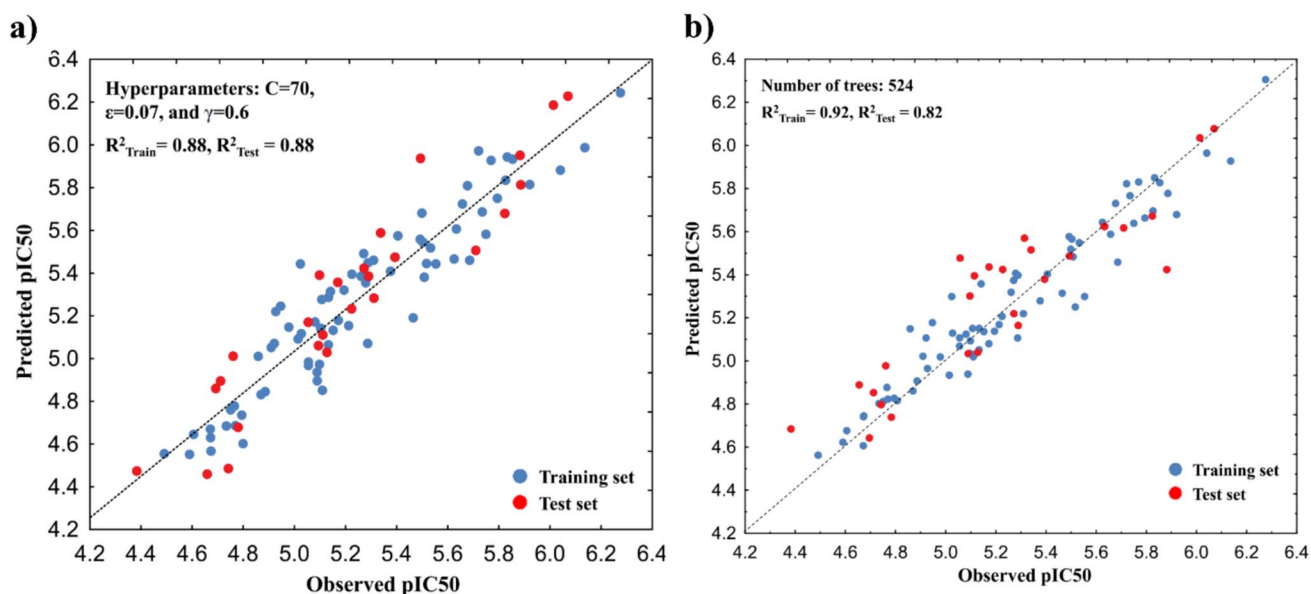


Fig. 6 **a** Optimized Support Vector Regression (SVR) model, built with the following hyperparameters settings: $C=70$, $\varepsilon=0.07$, and $\gamma=0.6$. **b** The booster model was built with 524 trees

Table 3 RMSE was calculated through LOOCV for the SVR model and Performance evaluation of three models (SVR, MOD1, and Boosting)

	SVR	MOD1 (baseline)	BOOSTING
<i>LOOCV cross-validation for SVR models</i>			
CV-RMSE	0.263		
<i>Model performance</i>			
	<i>Training</i>		
R^2	0.884	0.821	0.927
Q^2	0.879	0.820	0.921
RMSE	0.141	0.170	0.191
	<i>Test</i>		
R^2	0.882	0.872	0.821
Q^2	0.836	0.871	0.800
RMSE	0.176	0.156	0.196

model's practical utility. Table 3 details all the models (MOD1, SVR, Boosting) with their respective R^2_{Test} and $\text{RMSE}_{\text{Test}}$. Though all the models performed well on the independent set, the SVR model showcased superior generalization prowess (training and test predictability) over the other models.

Evaluation and applicability domain check of the model

To substantiate the fundamental regression presumption of normality in MOD1's standardized residuals, we employed quantile–quantile (Q–Q) plots. These plots are a

conventional tool for juxtaposing the spread or arrangement of two datasets. Here, the x -axis denotes standard quantiles from a normal distribution, while the y -axis represents MOD1-derived standardized residuals, enabling a comparative assessment. Notably, the Q–Q plot (Fig. 7a) revealed an alignment of the linear regression residuals closely aligning with the 45-degree reference line, thus validating the normality assumption.

The William plot has been generated to define an applicability domain, as shown in Fig. 7b. As indicated by the Williams plot results, the data points of all compounds within both the training and test sets, from the perspective of the response variable, were situated below the cautionary leverage threshold ($h^*=0.32$). However, one compound no. 62 exhibited substantial fitting residuals, yet it fell within the shared descriptor space due to its low leverage value. Excluding this compound from the training set would have substantially raised the R^2 value from 0.82 to 0.85.

Further, it was also observed that the compound nos. 38, 77, 86, 87 were plotted near the designated warning leverage value. While removing these outliers significantly improved the R^2 value, the decision was made to retain these compounds to maintain the model's broader applicability.

Molecular docking

The homologous structure generated with Robetta showed 99.6% of residues in favoured regions (Ramachandran plot), 91.88% in acceptable spatial arrangement (Verify 3D plot), and a 93.126 overall quality score (ERRAT plot), confirming the model's reliability for further analysis (Supple.

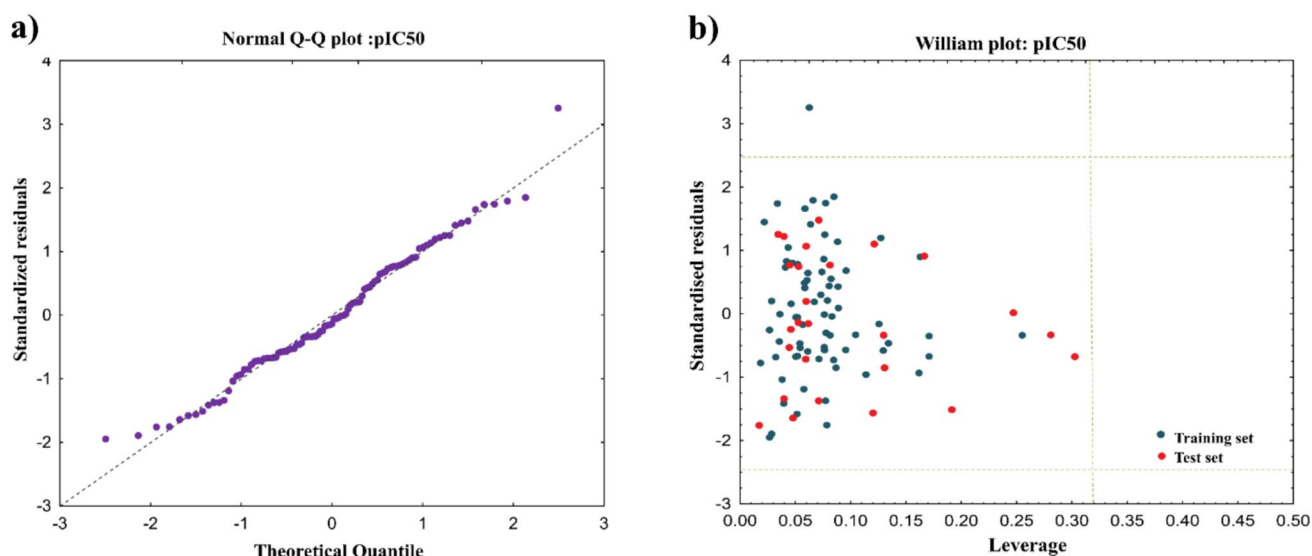


Fig. 7 **a** Normal quantile–quantile plot. **b** William plot, defining the applicability domain of the developed model

Figure S1). The active sites were defined using MSA and the template, Glycyl-Adenosine-5'-Phosphate bound *GlyRS* of *Thermus thermophilus*. This analysis selected critical amino acid residues, specifically Pro 215–Phe 249, Gly 347–Gln 362, and Pro 466–Ile 478. These chosen residues collectively form a contiguous beta sheet arrangement, which holds the potential to act as plausible binding sites for ligands (Supple. Figure S2).

The molecular docking analysis was first carried forward with the most potent inhibitor in the training dataset to gather insights into the primary forms of interaction established with the active site residues of the *LdGlyRS*. The interaction modes obtained from compounds 82 and 87 are displayed in Fig. 8.

Within the *LdGlyRS* active site, molecular docking of Compound 82 ($\text{pIC}_{50} = 6.07$) revealed three hydrogen bonds, with a docking score of -8.9 kcal/mol. These interactions involved nitrogen atoms and the amino acids Glu350 and Asp339, exhibiting bond distances of 3.40 Å, 3.28 Å, and 3.37 Å, respectively. An alkyl–halogen bonding interaction was observed involving a chlorine atom and amino acids Pro192 (4.25 Å) and Leu174 (4.32 Å), in addition to the significant active site residue Arg226 (4.35 Å). Additionally, theazole motif displayed a π – π T-shaped interaction with the amino acid Phe93 (4.60 Å), a feature also noted in the model's template–ligand interactions (1ggm). This π – π T-shaped interaction, along with the hydrophobic interactions, namely, Ala331, His325, Gly353, Arg357, Ser472, Arg477, and Glu329, contributes to the overall three-dimensional arrangement of the complexes and plays an essential role in determining binding modes and affinities. Glu 193

formed a carbon–hydrogen bond, indicating the ligand's robust affinity for the target protein.

The least bioactive ligand, Compound 87, was also docked within the active site of the *LdGlyRS*. Moreover, the compound bound firmly with a score of -10.7 kcal/mol, with conventional hydrogen bonds forming between residues Arg357, Ser472, and Gly353 and the Fluorine with distances of 3.26 Å, 4.29 Å, and 3.86 Å, respectively. Though the compound formed multiple π – π T-shaped halogen bonding and salt bridge interactions with the residues of the *LdGlyRS*, this compound also proved to be a vital compound in defining the limitation of the halogen group in constructing a drug compound. As described in our baseline QSAR model, more than 3–4 halogen groups can be detrimental; where halogen substitution can often enhance binding affinity, it can also lead to changes in the compound's shape and properties that affect its biological activity against the whole organism (pIC_{50} of 4.69). Fluorine atoms are highly electronegative and can have electron-withdrawing effects on neighbouring atoms and functional groups. As the anti-leishmanial activity is primarily attributed to the nitrogen atom within the azoles, multiple fluorine groups might "mask" the reactivity or interactions of neighbouring functional groups, including nitrogen atoms, resulting in decreased bioactivity. In addition, the critical Arg226 residue was also seen to form an unfavourable positive–positive bond with theazole group of the compound.

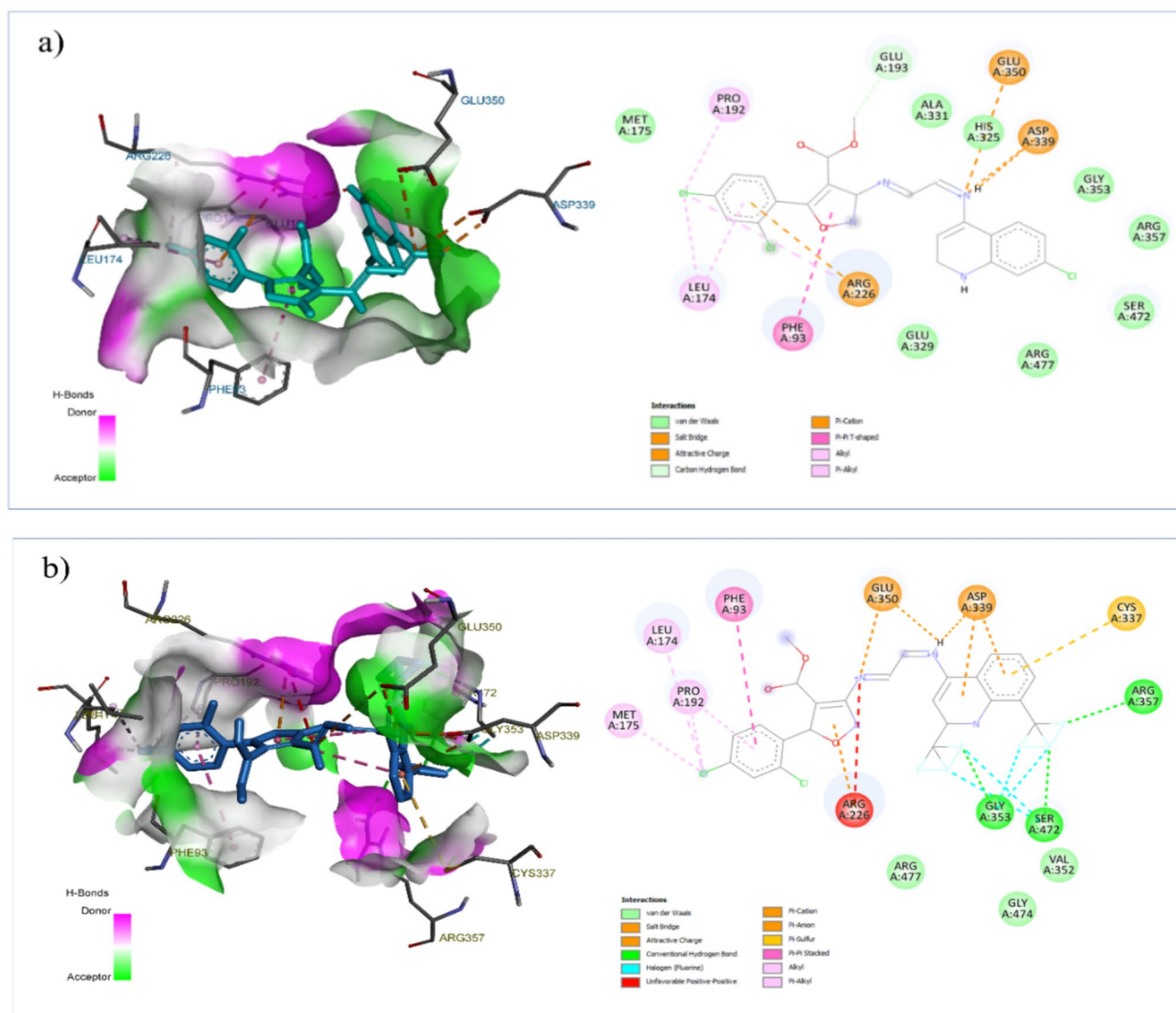


Fig. 8 **a, b** 2D & 3D docking poses showing interactions of compounds 82 (**a**) & 87 (**b**) in the active site residues of *LdGlyRS*

Table 4 The predicted pIC50 for the top FDA-approved and experimental drug Candidates with their respective Docking Scores

Compounds	Prediction pIC50					
	MLR		SVR	Boosting	Docking Score	AD
	MOD1	CM2				
Bazedoxifene	6.91	6.58	7.01	5.93	−9.4	In
Delamanid	5.62	5.61	5.37	5.51	−9.2	In
Talmetacin	7.52	7.03	6.84	6.32	−10.0	Out
Pipendoxifene	6.75	6.48	6.88	5.76	−8.9	In
Enzastaurin	6.24	7.13	5.96	6.04	−10.6	Out
Pyrvinium	6.63	6.34	6.53	6.32	−8.8	In
ZK-806711	6.33	6.13	5.28	5.66	−9.7	In
Niometacin	5.10	5.01	5.32	5.68	−7.6	In
Ketoconazole (control)	5.75	5.76	5.54	5.31	−8.6	In

Virtual screening for potential anti-leishmanial and *LdGlyRS* compounds

The QSAR models were employed to screen a selection of FDA-approved and experimental drugs with azole moieties from the ZINC drug-like library (Table 4 and Supple. Table S7). Bazedoxifene, Talmecatin, Pyrvinium, and Enzastaurin emerged as the leading candidates with notable

bioactivity predictions and docking scores. As indicated in Table 4, both Talmecatin and Enzastaurin were identified as falling outside the model's applicability domain. However, their retention in this study was based on their high docking scores and stable interaction with *LdGlyRS*. Figure 9 illustrates the chemical scaffold of these selected compounds alongside the reference ketoconazole, as well as their respective docked poses against *LdGlyRS*.

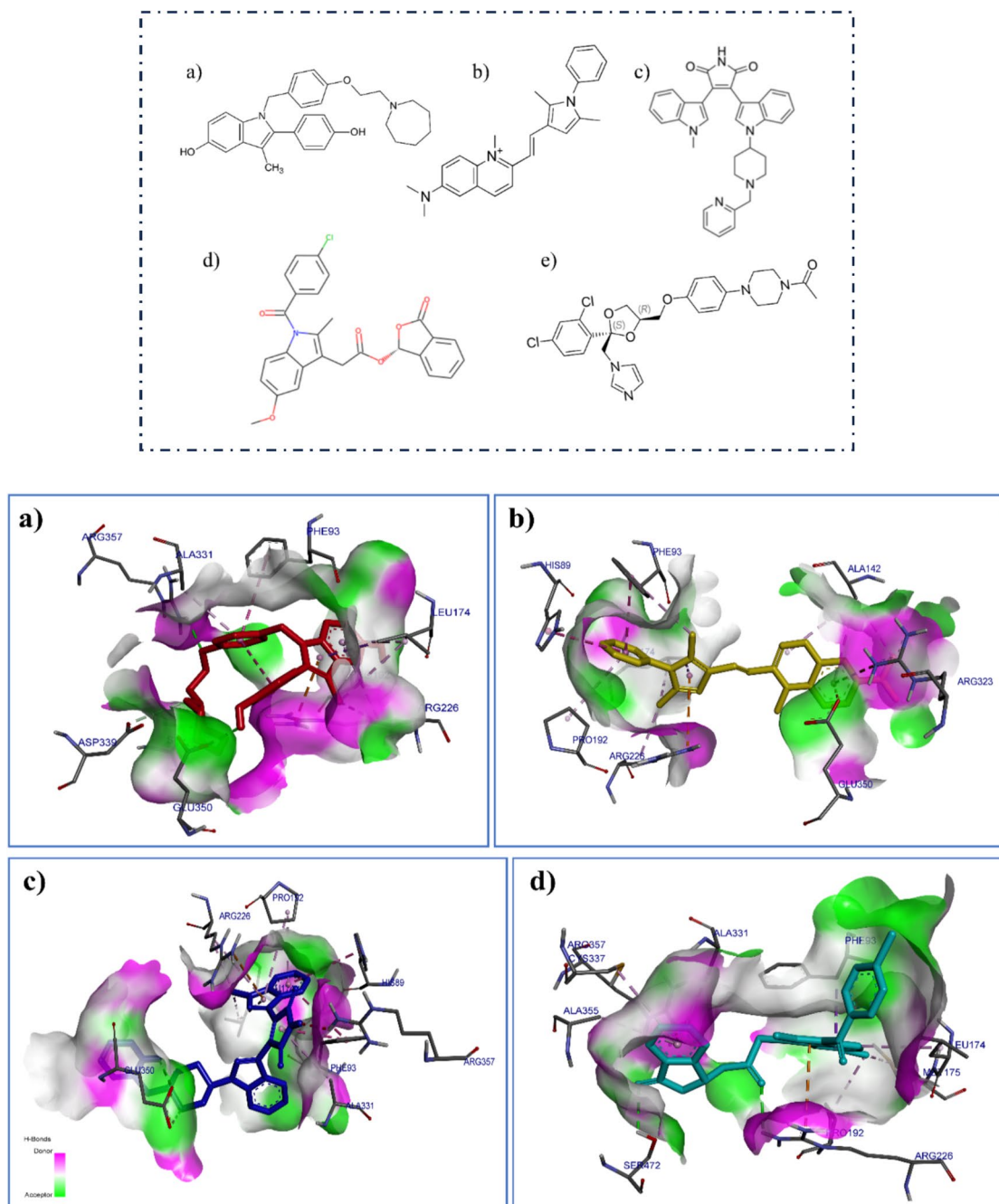


Fig. 9 Chemical scaffold of **a** Bazedoxifene, **b** Pyrvinium, **c** Enzastaurin, **d** Talmecatin, and **e** Ketoconazole & Docking poses of (A) Bazedoxifene (red), (B) Pyrvinium (yellow), (C) Enzastaurin (Blue), and (D) Talmecatin (Cyan) in the active site residues of *LdGlyRS*

Bazedoxifene (−9.4 kcal/mol), an FDA-approved selective oestrogen receptor modulator (SERM) with strong antimalarial properties [67], displayed a conventional hydrogen bond of 2.46 Å between its central amide bond and Arg357 and another H-bond of 2.47 Å between its hydroxyl (-OH) group and Glu350. Furthermore, a π -cation interaction was observed between the ligand's indole group and the critical residue Arg226. The ligand also participates in a T-shaped π - π stacking interaction with Phe93, involving aromatic ring interactions.

Talmetacin (−10.0 kcal/mol), an analgesic and antipyretic compound, forms conventional hydrogen bonds with Arg226 and Ser472. A π -cation interaction is also present between Arg226 and one of the ligand's aromatic rings. Additionally, the ligand participates in a π -sigma interaction with Phe93, where one of its aromatic rings stacks against the phenylalanine residue. It also establishes alkyl and π -alkyl interactions with residues like Leu174, Pro192, and Met175. Similarly, Pyrvinium (−8.8 kcal/mol), an FDA-approved drug for pinworm treatment with antitubercular and antibacterial bioactivity [68, 69], displayed favourable predictive bioactivity and formed π -cation interactions with crucial active site residues Glu350, Arg226, and Arg323. Likewise, the ligand also formed a T-shaped π - π stacking interaction with Phe93, with its aromatic moiety. Enzastaurin (−10.6 kcal/mol), a synthetic bisindolylmaleimide with potential anticancer properties, formed a hydrogen bond with the critical Glu350, a π -cation interaction with Arg226, and other significant interactions, including a π - π T-shaped interaction with Phe93 and a π -alkyl interaction with Pro192, involving its indole ring. All selected drug candidates notably demonstrated interactions with crucial residues Arg226 and Glu350.

Moreover, it was noted that the compound's potency in the QSAR dataset against the *L. donovani* amastigote was independent in its capacity to establish strong bonds with the *LdGlyRS* protein. However, to identify novel anti-leishmanial inhibitors that could also inhibit the *LdGlyRS* protein, only the dataset's most potent inhibitors were considered templates for a new compound search. The substructure in the similarity search contained isoxazole, imidazole, and oxazole ring, along with a Tanimoto coefficient threshold set at $T_c \geq 0.5$, resulting in a final list of ~2000 compounds. These compounds were concurrently screened against the *LdGlyRS*, and their predicted pIC50 values were calculated through a consensus model (CM2 + SVR + Boosting), with a voting score conducted. Table 5 depicts the optimum azole candidate with their predictive bioactivity against *L. donovani* amastigote and binding affinity score to *LdGlyRS*. The 3D binding pose, along with the 2D interaction, is given in the Supple. Figure S4.

ADMET prediction

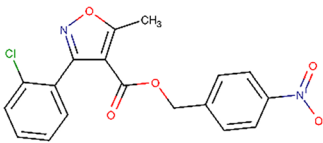
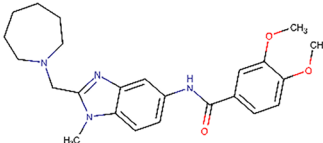
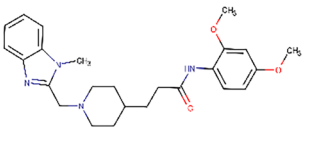
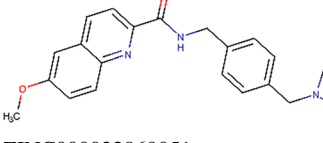
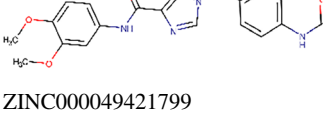
The selected optimal azole candidates were subjected to ADMET and drug-likeness assessments, with compounds 82 and 87 serving as reference points. The calculations used SwissADME, AdmetSAR 2.0, and ADMETlab 2.0 tools [70–72]. Intercorrelated properties were chosen based on the most accurate aggregated results (Table 6), and the attributes related to drug-likeness were predicted through the SwissADME online tool (Table 7).

All the selected compounds have an excellent absorption value of more than 90%, indicating good human intestinal absorption. The value of the human oral bioavailability depicted here (Table 6) is the per cent probability of a compound to be either +ve or -ve orally bioavailable to enter the systemic circulation and be available to exert its intended pharmacological effects. Here, the maximum of the compounds exerts good bioavailability, except for three compounds that are less probable in -ve bioavailability. Moreover, ZINC000009942262, ZINC000011934652, ZINC000006665032, and ZINC000032869051 displayed considerable potential to cross the Blood-Brain Barrier.

Metabolic metabolism refers to the drug's biochemical transformation process within the body. As a result of the metabolism, each possesses distinct physical and pharmacological characteristics [73]. The drug metabolism process, notably phase I metabolism (oxidation), aligns with the primary focus of our investigation and is prominently facilitated by the cytochrome P450 (CYP450) enzyme system. Human CYP450 comprises 17 identified families, of which CYP1-4 chiefly participates in drug metabolism. Notably, CYP enzymes such as CYP1A2, CYP2C9, CYP2C19, CYP2D6, and CYP3A4 metabolize more than 90% of drugs undergoing phase I metabolism [74]. Moreover, the inhibition of cytochrome CYP3A4 was particularly important in this study [74]. Each of the selected compounds was identified as a substrate for CYP3A4, and except ZINC000001153734, all were inhibitors of CYP3A4.

The notion of clearance illustrates the relationship between a drug's concentration within the body and its elimination rate. In this context, we calculated clearance using predefined thresholds (High: > 15 mL/min/kg; Moderate: 5–15 mL/min/kg; < 5 mL/min/kg) [72]. Notably, all the compounds displayed moderate clearance values, implying a moderate level of drug persistence within the body. Additionally, the compounds were examined for Human-hepatotoxicity and AMES toxicity. Most of the compounds in the QSAR dataset were hepatotoxic, and this pattern was also seen in ZINC000001153734 and ZINC000049421799, whereas all other selected compounds (Table 6) were non-hepatotoxic.

Table 5 The predicted pIC50 for the novel sets by three QSAR models and their respective Docking Results

Compounds	Predicted pIC50				
	MLR		SVR	Boosting	Docking score
	MOD1	CM2			
	5.78	5.67	6.06	6.01	− 8.6
ZINC000001153734					
	5.93	5.94	6.03	5.88	− 8.3
ZINC000009942262					
	6.14	6.07	6.26	6.09	− 8.1
ZINC000011934652					
	5.13	5.52	5.52	5.44	− 8.5
ZINC000032869051					
	5.2	5.21	5.11	5.19	− 7.9
ZINC000049421799					

MLR multiple linear regression, SVM support vector machine, MOD1 Baseline model, CM2 consensus model number 2

The compounds acquired from the ZINC Database underwent further evaluation regarding their synthetic feasibility, rated on a scale from 1 (indicating very easy synthesis) to 10 (suggesting highly complex synthesis). The calculated ease of synthesis for the selected compounds is represented by a low value 3 (Table 7). Moreover, all these compounds are readily available for purchase from various vendors, accessible through the ZINC20 database. It is worth noting that these compounds adhere to all drug-likeness criteria, except for Compound 87, which, indeed, ranks as the least bioactive among the QSAR dataset.

Molecular dynamics (MD) simulations

MD simulation was conducted over 100 ns to analyse five protein–ligand complexes' structural and dynamic behaviour and assess critical properties such as stability, flexibility, compactness, and interaction strength.

The RMSD values for the protein–ligand complexes (Fig. 10a) were used to assess the overall stability of the complexes during the MD simulation. The apoprotein (black) showed significant fluctuations with RMSD values reaching up to 15 Å, reflecting a higher degree of conformational flexibility. The Enzastaurin (yellow), Bazedoxifene (blue) and ZINC000009942262 (brown) complexes showed lower RMSD values, with Enzastaurin stabilizing around 4–6 Å, indicating more stable binding and reduced deviation from the initial structures. ZINC000009942262, however, exhibited early increased fluctuation but stabilized after 20 ns, suggesting late-stage stability.

Figure 10b illustrates the RMSF values for individual amino acid residues, revealing the protein's flexible and rigid regions. Elevated RMSF values were observed in the 350–400 residue region, particularly in the apoprotein (black), the Pyrvinium (red), and ZINC000011934652 (cyan) complexes, suggesting greater flexibility in this area.

Table 6 ADMET properties of the selected compounds

Compound name	Absorbance		Distribution		Metabolism		Inhibitor				Excretion		Toxicity	
	Human Intestinal Absorption ^a	Blood Brain Barrier	Human oral bioavailability ^b	Substrate	3A4	2D6	3A4	2C9	2C19	2D6	1A2	Clearance (CL) ^c	H-HT ^d	AMES Toxicity
Comp 82	0.9908	No	0.8286 (+)	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	4.243	Yes	No
Com 87	0.9919	No	0.6571 (+)	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	5.020	Yes	No
ZINC000001153734	0.9794	No	0.5286(+)	Yes	No	No	No	Yes	Yes	No	Yes	6.852	Yes	Yes
ZINC000049421799	0.8929	No	0.6857(+)	Yes	No	Yes	Yes	Yes	Yes	Yes	No	9.753	Yes	No
ZINC000009942262	0.9953	Yes	0.5429 (−)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	9.007	No	No
ZINC000011934652	1.0000	Yes	0.5143 (−)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	10.696	No	No
ZINC000032869051	0.9908	Yes	0.6714(+)	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	9.201	No	No

^aNumeric % absorbed^bNumeric % probability in either Positive (+) & Negative (−)^cHigh: > 15 mL/min/kg; moderate: 5–15 mL/min/kg low: < 5 mL/min/kg^dH-HT: Human-hepatotoxicity

Conversely, ligands such as Enzastaurin (yellow) and Bazedoxifene (blue) significantly reduced fluctuations in crucial active site residues (Gly 347–Gln 362 and Pro 466–Ile 478), indicating more robust and more stable interactions. Meanwhile, core residues within the active site (Pro 215–Phe 249) maintained a stable conformation, underlining the importance of these residues in preserving complex stability.

The radius of gyration (Rg) plots, shown in Fig. 10c, demonstrates the compactness of the protein–ligand complexes throughout the 100-ns simulation. The apoprotein displayed higher Rg values, fluctuating around 32–35 Å, indicating more structural flexibility without a ligand. Among the ligand-bound complexes, Enzastaurin (yellow), Talmetacin (green), and ZINC000009942262 (brown) showed relatively stable Rg values of approximately 30 Å, suggesting greater compactness and stability in their interactions with the protein. Conversely, ZINC000011934652 (cyan) exhibited higher Rg values, indicating a less compact protein structure.

Figure 10d illustrates the hydrogen bond (H-bond) count for the protein–ligand complexes over time. Notably, the Bazedoxifene (blue) complex exhibited the highest number of H-bonds, peaking at around six H-bonds during the 50–80-ns range, indicating strong and sustained interactions with the target protein. Subsequently, ZINC000011934652 (cyan) displayed 2–4 hydrogen bonds over time, maintaining these bonds relatively consistently from 20 to 70 ns. In contrast, the other ligands, such as Pyrvinium and Talmetacin, formed fewer H-bonds, suggesting comparatively weaker interaction dynamics with the protein.

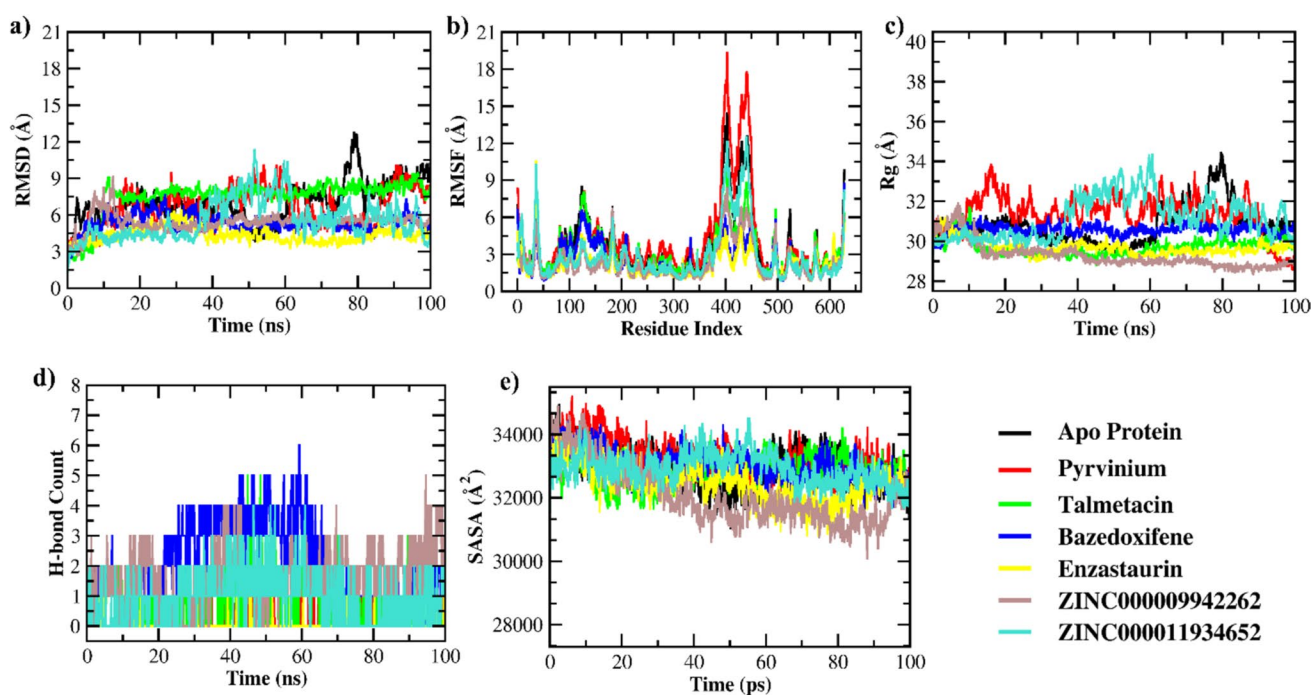
The solvent-accessible surface area (SASA) analysis (Fig. 10e) provides insights into the exposure of the protein–ligand complexes to the solvent environment. The apoprotein (black) had consistently lower SASA values (~30,000–32,000 Å²), indicating limited solvent exposure due to its more compact form. In contrast, the ligand-bound complexes, particularly ZINC000009942262 (brown) and Enzastaurin (yellow), displayed significantly higher SASA values, reflecting increased exposure to the solvent, which may correlate with conformational changes induced by ligand binding.

Principal component analysis (PCA)

To further explore the conformational space sampled by the protein–ligand complexes during the simulations, Principal Component Analysis (PCA) was performed. The PCA results revealed significant differences in the conformational dynamics between the apoprotein and ligand-bound complexes, particularly in the first two principal components

Table 7 Drug-likeness prediction of compounds 82, 87 and the selected compounds and their synthetic accessibility

	Lipinski	Ghose	Veber	Egan	Muegge	Synthetic accessibility ^a
Com 82	Yes	No	Yes	Yes	No	3.69
Com 87	No	No	Yes	No	No	3.94
ZINC000001153734	Yes	Yes	Yes	Yes	Yes	3.40
ZINC000049421799	Yes	Yes	Yes	Yes	Yes	3.13
ZINC000009942262	Yes	Yes	Yes	Yes	Yes	3.1
ZINC000011934652	Yes	No	Yes	Yes	Yes	3.23
ZINC000032869051	Yes	Yes	Yes	Yes	Yes	2.35

^aSynthetic accessibility score: from 1 (very easy) to 10 (very difficult)**Fig. 10** MD Simulation. **a** Root Mean Square Deviation (RMSD) of C α atoms over time, indicating the conformational stability of the protein. **b** Root Mean Square Fluctuation (RMSF) values reflecting the flexibility of individual residues throughout the simulation. Peaks in the graph correspond to residues with higher mobility. **c** Radius of Gyration (Rg) illustrating the compactness of the protein structure over time. **d** Total number of hydrogen bonds (H-bonds) formed throughout the simulation, providing insights into the stability and interactions within the protein structure. **e** Solvent-Accessible Surface Area (SASA) demonstrating the exposure of the protein surface to the solvent, with changes indicating dynamic interactions with the environment

(PC1 and PC2). Ligands such as Bazedoxifene, Enzastaurin, and ZINC000009942262 induced more compact and restricted motions in the protein, consistent with their favourable binding energies and strong interactions. Conversely, ZINC000011934652 showed broader conformational fluctuations, likely contributing to its less favourable binding profile. Detailed PCA plots illustrating the clustering of different conformations along PC1 and PC2 are provided in Fig. 11.

MM-GBSA binding free energy calculation

Among the studied complexes, Pyrvinium showed the most favourable binding free energy ($\Delta G_{\text{TOTAL}} = -50.00$ kcal/mol), reflecting strong binding affinity, primarily driven by significant van der Waals interactions ($\Delta \text{VDWAALS} = -51.52$ kcal/mol). Similarly, Bazedoxifene and Talmetacin exhibited strong binding free energies of -48.78 and -46.17 kcal/mol, respectively, benefiting from dominant van der Waals forces. Enzastaurin also demonstrated a relatively favourable binding energy (-43.34 kcal/mol), while ZINC000009942262 showed moderate

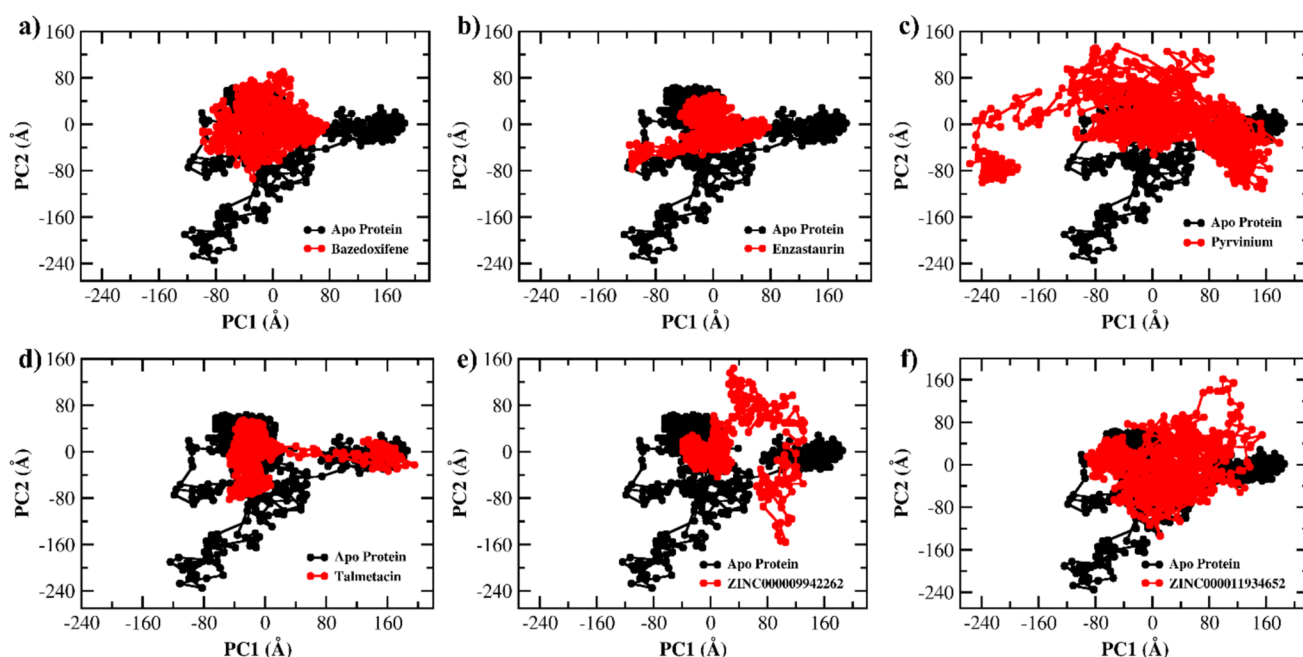


Fig. 11 Principal Component Analysis (PCA) plots comparing apo-protein (black) to various ligands (red). Each plot shows the distribution of the apo-protein and ligand-bound forms based on the first two principal components (PC1 and PC2), which capture the major variations in structural dynamics. **a** Apoprotein vs. Bazedoxifene. **b** Apoprotein vs. Enzastaurin. **c** Apoprotein vs. Pyrvinium. **d** Apopro-

tein vs. Talmecatin. **e** Apoprotein vs. ZINC000009942262. **f** Apoprotein vs. ZINC000011934652. PC1 and PC2 axes are measured in Å, representing structural variations. The red clusters indicate how the protein structure changes upon ligand binding compared to the apo form (black)

Table 8 Binding free energy decomposition of ligand–protein complexes (kcal/mol)

Complexes	Energy (kcal/mol)						
	Δ VDWAALS	Δ EEL	Δ EPB	Δ ENPOLAR	ΔG_{GAS}	ΔG_{SOLV}	Δ TOTAL
Bazedoxifene	−49.27	−11.89	17.79	−5.41	−61.16	12.39	−48.78
Enzastaurin	−46.88	−3.08	11.99	−5.36	−49.96	6.63	−43.34
Pyrvinium	−51.52	−2.29	8.93	−5.12	−53.81	3.81	−50.0
Talmecatin	−48.53	−6.98	14.41	−5.08	−55.50	9.33	−46.17
ZINC000009942262	−36.77	−8.17	11.55	−3.48	−44.94	8.06	−36.87
ZINC000011934652	494.27	−4.33	8.96	−4.90	489.94	4.06	494.00

interactions with a binding free energy of −36.87 kcal/mol. Interestingly, ZINC000011934652, despite forming a substantial number of hydrogen bonds, displayed a high positive binding free energy ($\Delta G_{\text{TOTAL}} = 494.00$ kcal/mol), influenced mainly by less favourable van der Waals interactions (Δ VDWAALS = 494.27 kcal/mol) and gas-phase energy ($\Delta G_{\text{GAS}} = 489.94$ kcal/mol) (Table 8).

These results suggest that while ZINC000011934652 forms numerous hydrogen bonds, its weak van der Waals forces undermine the complex's stability, leading to an energetically unfavourable profile in the MM-PBSA analysis. In contrast, ligands such as Bazedoxifene and Enzastaurin, which exhibit strong hydrogen bonding and highly negative van der Waals and total binding energies, establish more stable and favourable interactions with the protein.

Conclusion

The quest for discovering new anti-leishmanial compounds has surged due to the limitations of existing drugs. Azole compounds have emerged as potential candidates with promising anti-leishmanial activity. To comprehensively evaluate their efficacy in terms of structural characteristics, a systematic workflow utilizing a 2D-QSAR analysis has been developed, showcasing the application of this approach in evaluating a diverse range of azole compounds. This is the first time a traditional 2D-QSAR and q-RASAR have been used to predict parameters with diverse azole scaffolds against *L. donovani*.

By utilizing a representative data division, we constructed models from 73 small molecules encompassing 10 distinct structural classes as the basis of our understanding of azole-based anti-leishmanial chemical space. The trained model effectively explained the bioactivity of the drugs, and subsequently, the prediction of 26 test compounds displayed good precision within the models' applicable domain. Mechanistic interpretation revealed that the structural features MaxaasC and F04 [C–O] exerted the most significant positive and negative influence on pIC₅₀ against intracellular *L. donovani*. Comparing model performance with q-RASAR and consensus modelling using three qualified SMs, all consensus models displayed superior statistical performance compared to SMs and q-RASAR models. CM2 emerged as the best overall model with an impressive MAE_{test} of 0.127, outperforming the baseline model (MAE_{test} = 0.14155) and q-RASAR model (MAE_{test} = 0.161).

Furthermore, the well-established Support Vector Regression (SVR) outperformed the Multiple Linear Regression (MLR) and the Boosting method in terms of quality and generalization capabilities. However, due to inherent limitations in SVR modelling, such as parameter optimization and kernel function selection, the simpler MLR emerged as a more effective choice for baseline model generation. This highlights the potential of MLR models to guide compound design.

Simultaneously, the docking analysis unveiled how the most potent azole compound within the training set effectively inhibits *L. donovani* Glycyl-tRNA synthetase by interacting with crucial residues within the active site, notably arginine 226 and glutamic acid 350, which have a significant impact on aiding the binding of the Glycine-AMP substrate (1ggm). Through the integration of the 2D-QSAR models and docking analysis, a virtual screening of the ZINC database using the SwissSimilarity tool was executed, which led to the identification of FDA-approved drugs and new azole compounds exhibiting notable anti-GlyRS and anti-leishmanial properties. Notably, Bazedoxifene, Talmetacin, Pyrvinium, and Enzastaurin, among the marketed drugs, demonstrated effective binding to *LdGlyRS* with good binding energy and significant predicted bioactivity against *L. donovani*, as determined by the consensus models. On the other hand, the newly identified hits, ZINC000011934652 and ZINC000009942262, displayed promising predictive bioactivity and favourable outcomes in in-silico ADMET assessments and drug-likeness evaluations.

Molecular Dynamics (MD) simulations were conducted to validate the findings further, offering enhanced insights into the stability and dynamic behaviour of the ligand–protein complexes over a 100-ns period. The RMSD and Rg analyses confirmed stable binding for compounds such as Enzastaurin and ZINC000009942262. At the same time, RMSF plots revealed reduced flexibility in critical active

site regions for Bazedoxifene and Enzastaurin, signifying stronger interactions. Hydrogen bond analysis underscored the strong and sustained interactions of Bazedoxifene with the *LdGlyRS*, and SASA analysis revealed significant solvent exposure for complexes such as ZINC000009942262 and Enzastaurin, correlating with the conformational changes upon ligand binding. These novel hits, exhibiting variations from their parent compounds, offer promising opportunities for further optimization. The workflow and insights gained from these potential drug candidates represent an effective strategy for hit identification and lead optimization against specific anti-leishmanial targets.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11030-024-11070-w>.

Acknowledgements The lab is supported by the Department of Biotechnology (India) (BT/PR24504/NER/95/746/2017) to Diwakar Kumar. Rajat Nandi recognizes the financial support from the ICMR-SRF (Letter No. 45/06/2022-DDI/BMS, dated 17/05/2022).

Author contribution RN, AS, AP, and DK carried out the experiment. RN and DK wrote the manuscript. RN and DK contributed to the analysis of the results. DK supervised the project and conceived the original idea.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that no conflict of interest exists.

References

- Burza S, Croft SL, Boelaert M (2018) Leishmaniasis. Lancet 392:951–970. [https://doi.org/10.1016/S0140-6736\(18\)31204-2](https://doi.org/10.1016/S0140-6736(18)31204-2)
- WHO. WHO fact sheet on Leishmaniasis (Internet). 2023 Jan 12 [cited 2023 Sep 1]. Available from: <https://www.who.int/en/news-room/fact-sheets/detail/leishmaniasis>.
- Sangshetti JN, Khan FAK, Kulkarni AA, Arote R, Patil RH (2015) Antileishmanial drug discovery: comprehensive review of the last 10 years. RSC Adv 5:32376–32415. <https://doi.org/10.1039/C5RA02669E>
- Patterson S, Fairlamb AH (2019) Current and future prospects of nitro-compounds as drugs for trypanosomiasis and leishmaniasis. Curr Med Chem 26:4454–4475. <https://doi.org/10.2174/0929867325666180426164352>
- Petri e Silva SC, Palace-Berl F, Tavares LC, Soares SR, Lindoso JA (2016) Effects of nitro-heterocyclic derivatives against *Leishmania (Leishmania) infantum* promastigotes and intracellular amastigotes. Exp Parasitol 163:68–75. <https://doi.org/10.1016/j.exppara.2016.01.007>
- Croft SL, Yardley V (2002) Chemotherapy of leishmaniasis. Curr Pharm Des 8:319–342. <https://doi.org/10.2174/1381612023396258>
- Mukherjee T, Roy K, Bhaduri A (1990) Acivicin: a highly active potential chemotherapeutic agent against visceral leishmaniasis.

- Biochem Biophys Res Commun 170:426–432. [https://doi.org/10.1016/0006-291x\(90\)92109-d](https://doi.org/10.1016/0006-291x(90)92109-d)
8. Suryawanshi SN, Tiwari A, Chandra N, Ramesh GS (2012) Chemotherapy of leishmaniasis Part XI: synthesis and bioevaluation of novel isoxazole containing heteroretinoid and its amide derivatives. *Bioorg Med Chem Lett* 22:6559–6562. <https://doi.org/10.1016/j.bmcl.2012.09.024>
 9. Mukhopadhyay S, Barak DS, Karthik R, Verma SK, Bhatta RS, Goyal N et al (2020) Antileishmanial assessment of isoxazole derivatives against *L. donovani*. *RSC Med Chem* 11:1053–1062. <https://doi.org/10.1039/d0md00083c>
 10. Stephens CE, Brun R, Salem MM, Werbovetz KA, Tanious F, Wilson WD et al (2003) The activity of diguanidino and “reversed” diamidino 2,5-diarylfurans versus *Trypanosoma cruzi* and *Leishmania donovani*. *Bioorg Med Chem Lett* 13:2065–2069. [https://doi.org/10.1016/s0960-894x\(03\)00319-6](https://doi.org/10.1016/s0960-894x(03)00319-6)
 11. Reid CS, Farahat AA, Zhu X, Pandharkar T, Boykin DW, Werbovetz KA (2012) Antileishmanial bis-arylimidamides: DB766 analogs modified in the linker region and bis-arylimidamide structure-activity relationships. *Bioorg Med Chem Lett* 22:6806–6810. <https://doi.org/10.1016/j.bmcl.2012.06.037>
 12. Abdelhameed A, Feng M, Joice AC, Zywoit EM, Jin Y, La Rosa C et al (2021) Synthesis and antileishmanial evaluation of arylimidamide-azole hybrids containing a phenoxyalkyl linker. *ACS Infect Dis* 7:1901–1922. <https://doi.org/10.1021/acscinfecdis.0c00855>
 13. Marrapu VK, Mittal M, Shivhare R, Gupta S, Bhandari K (2011) Synthesis and evaluation of new furanyl and thiophenyl azoles as antileishmanial agents. *Eur J Med Chem* 46:1694–1700. <https://doi.org/10.1016/j.ejmech.2011.02.021>
 14. Bhandari K, Srinivas N, Marrapu VK, Verma A, Srivastava S, Gupta S (2010) Synthesis of substituted aryloxy alkyl and aryloxy aryl alkyl imidazoles as antileishmanial agents. *Bioorg Med Chem Lett* 20:291–293. <https://doi.org/10.1016/j.bmcl.2009.10.117>
 15. Srinivas N, Palne S, Nishi GS, Bhandari K (2009) Aryloxy cyclohexyl imidazoles: a novel class of antileishmanial agents. *Bioorg Med Chem Lett* 19:324–327. <https://doi.org/10.1016/j.bmcl.2008.11.094>
 16. Verma A, Srivastava S, Sane SA, Marrapu VK, Srinivas N, Yadav M et al (2011) Antileishmanial activity of benzocycloalkyl azole oximino ethers: the conformationally constraint analogues of oxiconazole. *Acta Trop* 117:157–160. <https://doi.org/10.1016/j.actatropica.2010.10.011>
 17. Marrapu VK, Srinivas N, Mittal M, Shakyia N, Gupta S, Bhandari K (2011) Design and synthesis of novel tetrahydronaphthyl azoles and related cyclohexyl azoles as antileishmanial agents. *Bioorg Med Chem Lett* 21:1407–1410. <https://doi.org/10.1016/j.bmcl.2011.01.026>
 18. Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Porokov V et al (2020) QSAR without borders. *Chem Soc Rev* 49:3525–3564. <https://doi.org/10.1039/d0cs00098a>
 19. Castillo-Garit JA, Abad C, Rodríguez-Borges JE, Marrero-Ponce Y, Torrens F (2012) A review of QSAR studies to discover new drug-like compounds active against leishmaniasis and trypanosomiasis. *Curr Top Med Chem* 12:852–865. <https://doi.org/10.2174/156802612800166756>
 20. Bernal FA, Schmidt TJ (2019) A comprehensive QSAR study on antileishmanial and antitrypanosomal cinnamate ester analogues. *Molecules* 24(23):4358. <https://doi.org/10.3390/molecules24234358>
 21. Goodarzi M, da Cunha EF, Freitas MP, Ramalho TC (2010) QSAR and docking studies of novel antileishmanial diaryl sulfides and sulfonamides. *Eur J Med Chem* 45:4879–4889. <https://doi.org/10.1016/j.ejmech.2010.07.060>
 22. Lorenzo VP, Lúcio AS, Scotti L, Tavares JF, Filho JM, Lima TK, Rocha JD, Scotti MT (2016) Structure- and ligand-based approaches to evaluate aporphynic alkaloids from Annonaceae as multi-target agents against *Leishmania donovani*. *Curr Pharm Des* 22:5196–5203. <https://doi.org/10.2174/1381612822666160513144853>
 23. Ugbe FA, Shallangwa GA, Uzairu A, Abdulkadir I (2022) A combined 2-D and 3-D QSAR modeling, molecular docking study, design, and pharmacokinetic profiling of some arylimidamide-azole hybrids as superior *L. donovani* inhibitors. *Bull Natl Res Centre* 46:189. <https://doi.org/10.1186/s42269-022-00874-1>
 24. Casanova-Alvarez O, Morales-Helguera A, Cabrera-Pérez MA, Molina-Ruiz R, Molina C (2021) A novel automated framework for QSAR modeling of highly imbalanced *Leishmania* high-throughput screening data. *J Chem Inf Model* 61:3213–3231. <https://doi.org/10.1021/acs.jcim.0c01439>
 25. Freist W, Logan DT, Gauss DH (1996) Glycyl-tRNA synthetase. *Biol Chem Hoppe Seyler* 377:343–356
 26. Guo RT, Chong YE, Guo M, Yang XL (2009) Crystal structures and biochemical analyses suggest a unique mechanism and role for human glycyl-tRNA synthetase in Ap4A homeostasis. *J Biol Chem* 284:28968–28976. <https://doi.org/10.1074/jbc.M109.030692>
 27. Park MC, Kang T, Jin D, Han JM, Kim SB, Park YJ, Cho K, Park YW, Guo M, He W, Yang XL, Schimmel P, Kim S (2012) Secreted human glycyl-tRNA synthetase implicated in defense against ERK-activated tumorigenesis. *Proc Natl Acad Sci USA* 109:E640–E647. <https://doi.org/10.1073/pnas.1200194109>
 28. Francklyn CS, Mullen P (2019) Progress and challenges in aminoacyl-tRNA synthetase-based therapeutics. *J Biol Chem* 294:5365–5385. <https://doi.org/10.1074/jbc.REV118.002956>
 29. Gill J, Sharma A (2023) Exploration of aminoacyl-tRNA synthetases from eukaryotic parasites for drug development. *J Biol Chem* 299:102860. <https://doi.org/10.1016/j.jbc.2022.102860>
 30. KNIME Analytics Platform. Available from: <http://update.knime.com/analytics-platform/4.0>. Accessed 26 Nov 2020
 31. KNIME Trusted Community Contributions. Available from: <http://updateknime.com/community-contributions/trusted/4.0>. Accessed 26 Nov 2020
 32. Todeschini R, Consonni V (2010) Molecular descriptors. In: Recent advances in QSAR studies, pp 29–102
 33. Alvascience. AlvaDesc (Software for Molecular Descriptors Calculation). Version 1.0.18, 2020.
 34. Stumpfe D, Bajorath J (2012) Exploring activity cliffs in medicinal chemistry. *J Med Chem* 55:2932–2942. <https://doi.org/10.1021/jm201706b>
 35. Bajorath J, Peltason L, Wawer M, Guha R, Lajiness MS, Van Drie JH (2009) Navigating structure-activity landscapes. *Drug Discov Today* 14(13–14):698–705. <https://doi.org/10.1016/j.drudis.2009.04.003>
 36. Chukhrova N, Johannssen A (2019) Fuzzy regression analysis: systematic review and bibliography. *Appl Soft Comput* 84:105708. <https://doi.org/10.1016/j.asoc.2019.105708>
 37. Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S (2013) QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. *J Comput Chem* 34:2121–2132. <https://doi.org/10.1002/jcc.23361>
 38. Rogers D (1999) Genetic function approximation: evolutionary construction of novel, interpretable, nonlinear models of experimental data. Rational drug design. Springer, New York
 39. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B (Methodological)* 58:267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
 40. Topliss JG, Edwards RP (1979) Chance factors in studies of quantitative structure-activity relationships. *J Med Chem* 22:1238–1244. <https://doi.org/10.1021/jm00196a017>

41. Chatterjee M, Banerjee A, De P, Gajewicz-Skretna A, Roy K (2022) A novel quantitative read-across tool designed purposefully to fill the existing gaps in nanosafety data. *Environ Sci Nano* 9:189–203. <https://doi.org/10.1039/D1EN00725D>
42. Banerjee A, Roy K (2022) First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability. *Mol Divers* 26:2847–2862
43. Roy K, Ambure P, Kar S, Ojha PK (2018) Is it possible to improve the quality of predictions from an “intelligent” use of multiple QSAR/QSPR/QSTR models? *J Chemometrics* 32:e2992. <https://doi.org/10.1002/cem.2992>
44. Niu B, Lu WC, Yang SS, Cai YD, Li GZ (2007) Support vector machine for SAR/QSAR of phenethyl-amines. *Acta Pharm Sinica* 28:1075–1086. <https://doi.org/10.1111/j.1745-7254.2007.00573.x>
45. Gunn SR (1998) Support vector machines for classification and regression. Department of Electronics and Computer Science, University of Southampton. <https://doi.org/10.1039/b918972f>
46. Schapire RE (2003) The boosting approach to machine learning: an overview. *Nonlinear estimation and classification*. Springer, Berlin
47. TIBCO Statistica (2017) Version 13.3.0. TIBCO Software Inc, Palo Alto, CA, USA. Available from: <https://www.tibco.com/products/tibco-statistica>
48. Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ Health Perspect* 111:1361–1375. <https://doi.org/10.1289/ehp.5758>
49. Gramatica P (2007) Principles of QSAR models validation: internal and external. *QSAR Comb Sci* 26:694–701. <https://doi.org/10.1002/qsar.200610151>
50. Rücker C, Rücker G, Meringer M (2007) y-Randomization and its variants in QSPR/QSAR. *J Chem Inf Model* 47:2345–2357. <https://doi.org/10.1021/ci700157b>
51. Roy K, Kar S, Ambure P (2015) On a simple approach for determining applicability domain of QSAR models. *Chemom Intell Lab Syst* 145:22–29. <https://doi.org/10.1021/acsomega.8b01647>
52. Zoete V, Daina A, Bovigny C, Michielin O (2016) SwissSimilarity: a web tool for low to ultra high throughput ligand-based virtual screening. *J Chem Inf Model* 56:1399–1404. <https://doi.org/10.1021/acs.jcim.6b00174>
53. Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, Moroz YS, Mayfield J, Sayle RA (2020) ZINC20-A free ultralarge-scale chemical database for ligand discovery. *J Chem Inf Model* 60:6065–6073. <https://doi.org/10.1021/acs.jcim.0c00675>
54. O’Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. *J Cheminform* 3:33. <https://doi.org/10.1186/1758-2946-3-33>
55. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
56. Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* 32:W526–W531. <https://doi.org/10.1093/nar/gkh468>
57. Lill MA, Danielson ML (2011) Computer-aided drug design platform using PyMOL. *J Comput Aided Mol Des* 25:13–19. <https://doi.org/10.1007/s10822-010-9395-8>
58. Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 31:455–461. <https://doi.org/10.1002/jcc.21334>
59. Dassault Systèmes (2021) Discovery studio visualizer. Version 21.1.0.20298. Dassault Systèmes, San Diego
60. Spoel VD (2020) GROMACS 2020.6 Source code. Zenodo <https://doi.org/10.5281/zenodo.4576055>
61. Lu T, Chen F (2012) Multiwfn: a multifunctional wavefunction analyzer. *J Comput Chem* 33:580–592. <https://doi.org/10.1002/jcc.22885>
62. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph* 14:33–38. [https://doi.org/10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5)
63. Miller BR III, McGee TD Jr, Swails JM, Homeyer N, Gohlke H, Roitberg AE (2012) MMPBSA.py: an efficient program for end-state free energy calculations. *J Chem Theory Comput* 8:3314–3321. <https://doi.org/10.1021/ct300418h>
64. Roy K, Das RN, Ambure P, Aher RB (2016) Be aware of error measures: further studies on validation of predictive QSAR models. *Chemom Intell Lab Syst* 152:18–33. <https://doi.org/10.1016/j.chemolab.2016.01>
65. Sushko I, Novotarskyi S, Körner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY, Todeschini R (2011) Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J Comput Aided Mol Des* 25:533–554. <https://doi.org/10.1007/s10822-011-9440-2>
66. Testa B, van de Waterbeemd H (1996) Lipophilicity in drug action and toxicology. *J Med Chem*. <https://doi.org/10.1021/jm960775b>
67. Sudhakar R, Adhikari N, Pamnani S, Panda A, Bhattacharjee M, Rizvi Z, Shehzad S, Gupta D, Sijwali PS (2022) Bazedoxifene, a postmenopausal drug, acts as an antimalarial and inhibits hemozoin formation. *Microbiol Spectrum* 10:e02781-e2821. <https://doi.org/10.1128/spectrum.02781-21>
68. Loughheed KE, Taylor DL, Osborne SA, Bryans JS, Buxton RS (2009) New anti-tuberculosis agents amongst known drugs. *Tuberculosis* 89:364–370. <https://doi.org/10.1016/j.tube.2009.07.002>
69. Torres NS, Abercrombie JJ, Srinivasan A, Lopez-Ribot JL, Ramasubramanian AK, Leung KP (2016) Screening a commercial library of pharmacologically active small molecules against *Staphylococcus aureus* biofilms. *Antimicrob Agents Chemother* 60:5663–5672. <https://doi.org/10.1128/aac.00377-16>
70. Daina A, Michielin O, Zoete V (2017) SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci Rep* 7:42717. <https://doi.org/10.1038/srep42717>
71. Yang H, Lou C, Sun L, Li J, Cai Y, Wang Z, Li W, Liu G, Tang Y (2019) admetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties. *Bioinformatics (Oxford)* 35(6):1067–1069. <https://doi.org/10.1093/bioinformatics/bty707>
72. Xiong G et al (2021) ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Res* 49:W5–W14. <https://doi.org/10.1093/nar/gkab255>
73. Kok-Yong S, Lawrence L (2015) Basic pharmacokinetic concepts and some clinical applications. InTech
74. Otyepka M et al (2012) Is there a relationship between the substrate preferences and structural flexibility of cytochromes P450? *Curr Drug Metab* 13:130–142. <https://doi.org/10.2174/138920012798918372>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.