


Applied Machine Learning Toward Drug Discovery Enhancement: Leishmaniases as a Case Study

Emna Harigua-Souiai¹ , Rafeh Oualha¹, Oussama Souiai² , Ines Abdeljaoued-Tej^{2,3} , and Ikram Guizani¹

¹Laboratory of Molecular Epidemiology and Experimental Pathology-LR16IPT04, Institut Pasteur de Tunis, Université de Tunis El Manar, Tunis, Tunisia. ²Laboratory of Bioinformatics, BioMathematics and BioStatistics LR20IPT09, Institut Pasteur de Tunis, Université de Tunis El Manar, Tunis, Tunisia. ³Engineering School of Statistics and Information Analysis, University of Carthage, Ariana, Tunisia.

Bioinformatics and Biology Insights
Volume 16: 1–10
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11779322221090349


ABSTRACT: Drug discovery (DD) research is a complex field with a high attrition rate. Machine learning (ML) approaches combined to cheminformatics are of valuable input to this field. We, herein, focused on implementing multiple ML algorithms that shall learn from different molecular fingerprints (FPs) of 65 057 molecules that have been identified as active or inactive against *Leishmania major* promastigotes. We sought to build a classifier able to predict whether a given molecule has the potential of being anti-leishmanial or not. Using the RDKit library, we calculated 5 molecular FPs of the molecules. Then, we implemented 4 ML algorithms that we trained and tested for their ability to classify the molecules into active/inactive classes based on their chemical structure, encoded by the molecular FPs. Best performers were random forest (RF) and support vector machine (SVM), while atom-pair and topology torsion FPs were the best embedding functions. Both models were further assessed on different stratification levels of the dataset and showed stable performances. At last, we used them to predict the potential of molecules within the Food and Drug Administration (FDA)-approved drugs collection to present anti-*Leishmania* effects. We ranked these drugs according to their anti-Leishmanial probability and obtained in total seven anti-*Leishmania* agents, previously described in the literature, within the top 10 of each model. This validates the robustness of the approach, the algorithms, and FPs choices as well as the importance of the dataset size and content. We further engaged these molecules into reverse docking experiments on 3D crystal structures of seven well-studied Leishmania drug targets and could predict the molecular targets for 4 drugs. The results bring novel insights into anti-Leishmania compounds.

KEYWORDS: Machine learning, molecular fingerprints, *Leishmania*, classification, drug repurposing, drug target

RECEIVED: July 12, 2021. **ACCEPTED:** March 4, 2022.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: E.H.-S. is a recipient of a NAS-USAID grant within the PEER Women Mentoring Programme; Grant Award Number AID-OAA-A-11-00012. The study also received support from the ministry of Higher Education and Research, Tunisia (LR16IPT04 to I.G.).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Emna Harigua-Souiai, Laboratory of Molecular Epidemiology and Experimental Pathology—LR16IPT04, Institut Pasteur de Tunis, Université de Tunis El Manar, Tunis 1002, Tunisia. Email: harigua.emna@gmail.com; emna.harigua@pasteur.utm.tn

Background

Drug discovery (DD) and lead optimization are becoming very challenging tasks to researchers from academia and the private sector. As the number of approved drugs is decreasing, the development costs of a novel valid therapeutic molecule exceed US \$1.2 billion for a discovery process lasting over more than 12 years. In this context, computational approaches appeared as promising avenues in lowering the cost, duration, and attrition rate of the DD process. Multiple groups have focused on developing machine learning (ML) and deep learning (DL) algorithms for drug discovery and development.^{1–3} Unfortunately, drug discovery and lead optimization are still low-data domains with too few molecules reaching the market. Because of this low-data challenge, adapted ML/DL approaches were proposed, such as one-shot learning methods based on structure–activity relationships for activity predictions Altae-Tran et al.⁴ Compared to more classical approaches, they demonstrated higher predictive power using small positives in their training sets. However, they showed poor capability of generalization to distinct datasets. Thus, sufficient data are still a cornerstone in the domain.

The low-data issue is even worse for neglected tropical diseases (NTDs), which mainly affect the poorest and most vulnerable populations. For such diseases, available data mostly describe early stages of drug discovery, such as extracellular assays or primary high-throughput screenings (HTSs). It is not likely to access large datasets of molecules validated at different levels of experimental validation (enzymatic, in vitro, in cellulo, ex vivo, in vivo, etc). There is an urgent need to implement alternative cost-effective approaches for drug discovery and repurposing against such diseases using early-stage drug discovery data. We herein focus on Leishmaniases, a group of largely distributed vector-borne parasitic diseases affecting 700 000–1 million cases/year and causing 26 000–65 000 deaths annually. The lack of low-cost and non-toxic treatments for these diseases brings the urge to develop novel therapeutic strategies. However, drugs Research & Development (R&D) faces many challenges that range from increasing costs and pressure on pricing to the high attrition rate. This puts NTDs—including Leishmaniases—in the least prioritized diseases for therapeutics development because of their poor financial potential in the market.



The present work was designed to demonstrate the power of common ML algorithms in classifying molecules as active or inactive against *Leishmania* parasites based on their chemical structures. We used data originating from a bioassay targeting the *Leishmania major* promastigote growth and viability, retrieved from the PubChem database. We also implemented a pipeline for data preprocessing and encoding toward their use as input to ML algorithms. Multiple encoding systems of chemical structures, called molecular fingerprints (FPs), were used. We then compared the performances of different ML algorithms on stratified subsets of the data. The best combinations of (ML, FP, and subset) were further validated on unseen data from a second bioassay, prior to their use for drug repurposing predictions against Leishmaniasis, and identification of potential targets for the molecules selected.

Methods

Datasets

Data were retrieved from 2 phenotypic bioassays targeting the *Leishmania* promastigotes and deposited in the PubChem database under the references AID 1063 and AID 1258. Data corresponded to parasite growth and viability inhibition experiments with a binary activity outcome (active and inactive). The experiments measured *Leishmania* spp promastigotes drug susceptibility through Alamar blue-based assay. The number of viable promastigotes correlated with a colorimetric reading, which indicated the anti-Leishmanial effect of the active molecules. The first bioassay (AID 1063) is a preliminary HTS of 65 057 molecules whose results were used for the training of multiple ML algorithms and their optimization toward selecting the most performing one(s). The second bioassay (AID 1258) is a confirmatory screening performed on 1 122 molecules derived from the first bioassay that were tested at a lower concentration of 1 μ M. We used it as an external validation set that has not been seen by the best performer(s). For simplicity's sake, molecules with "Inconclusive" activity outcome were discarded from the AID (1258) dataset.

The Food and Drug Administration (FDA)-approved drugs collection used herein was downloaded on April 13, 2021, from the ZINC database.⁵

Data encoding

Chemical structures of the molecules were encoded using the Simplified Molecular Input Line Entry System (SMILES), retrieved directly from the PubChem database, and merged with the bioassay data frame according to the PubChem identifier of each compound (CID).

Prior to molecular structure encoding into numerical vectors, we performed a series of data splitting to obtain multiple datasets with different sizes. We first performed a data equilibration of the dataset using the random over sampling (ROS) algorithm, previously described as one of the most performing methods in drug discovery.³ Then, we performed a random

undersampling (RUS) through a reduction of the elements within the largest class inactive. Finally, we performed 2 subsampling of the RUS dataset up to 10% and 1%. Thus, we disposed of 5 datasets, namely, the original dataset, the ROS dataset, the RUS dataset, the 10% sub-sample of the RUS dataset, and the 1% sub-sample of the RUS dataset. Molecule structures were then encoded into intelligible format for the ML algorithms used herein. Out of the SMILES within each dataset, we generated molecule objects using the RDkit⁶ library under Python (<http://www.rdkit.org>). We then calculated different types of molecular FPs that consist of binary vectors. All datasets were split into training (80%), validation (10%), and test (10%) sets. The validation set will be used to fine-tune the ML models' hyperparameters prior to evaluating its performances on the test set.

Five molecular FPs were calculated using RDkit for each compound within the datasets: (1) the RDkit molecular fingerprints (RDFPs), (2) the atom-pair FPs (APFPs), (3) the topology torsion FPs (TTFPs), and (4) the extended-connectivity FPs with a radius of 2 atoms (ECFP4) and the extended-connectivity FPs with a radius of 3 atoms (ECFP6).

Machine learning

Four ML algorithms were implemented under Sci-kit learn, an open-source Python library.⁷ The molecular FP vectors were used as input to perform a binary classification of the molecules into active and inactive classes. The algorithms are: linear regressor (LR), gradient boosting (GB), random forest (RF), and support vector machine (SVM). We performed a 5-fold cross-validation tuning to optimize the ML models based on their accuracy when trained on all 5 datasets.

For most performant models, the receiver operating characteristic (ROC) and the precision-recall (PR) curves were generated and their area under the curve (AUC) scores were calculated. Based on the confusion matrix elements: the true positives (TP), the false positives (FP), the true negatives (TN), and the false negatives (FN), we calculated different metrics to assess models' performances, namely, sensitivity (also called recall), specificity, precision, balanced accuracy, and the *F1*-score.

The best classifier(s) were then used to identify potential anti-*Leishmania* effectors within the FDA-approved drugs collection for which suitable FPs were generated. For each molecule, probabilities of being classified as active and inactive are calculated.

All simulations were run on one machine with the following specifications: Hardware (CPU—Intel i7-9750H @ 5.00 Ghz, RAM-32GB DDR4) and Software (Ubuntu 20.04 LTS, Python-3.6, RDkit 2017.09.1, Scikit-Learn 0.23.1, Matplotlib 3.2.2, Numpy 1.19.0).

Reverse docking. Crystal structures of *Leishmania* proteins that are considered as potential drug targets for anti-*Leishmania* therapeutics were retrieved from the Protein Data Bank (PDB).

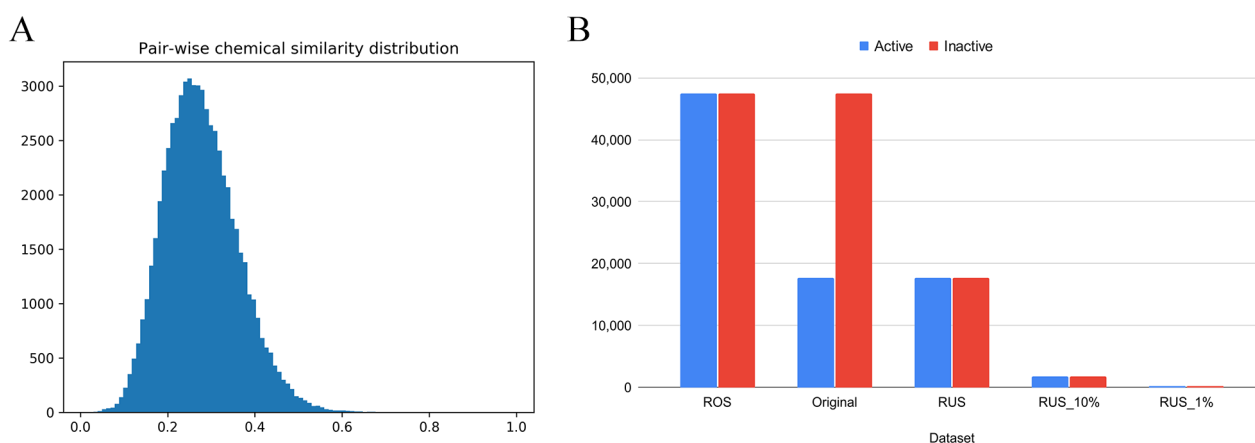


Figure 1. Datasets descriptive statistics. (A) Histogram of the chemical similarity distribution calculated pairwise between molecules constituting the dataset. (B) Bar plots of the proportions of active and inactive molecules within each dataset. RUS indicates random undersampling.

These were the pteridine reductase 1 (PTR1; PDBid: 2BFM), the trypanothione reductase (TR; PDBid: 2KJ6), the dihydroorotate dehydrogenase (DHODH; PDBid: 3MJY), the mitogen-activated protein kinase 10 (MAPK10; PDBid: 3UIB), the arginase (PDBid: 4IU0), the UDP-glucose pyrophosphorylase (UDP-GP; PDBid: 5NZL) and the N-myristoyl-transferase (NMT; PDBid: 6QD9).

All structures were retrieved from the PDB using their IDs and prepared for docking simulations using the Open Babel software.⁸ Water molecules and co-crystallized ligands were removed. Explicit hydrogens were added and atomic charges were calculated using the Gasteiger model. Prepared structures will be here referred to as receptors. On each receptor, we defined the docking space as a large virtual box that included the catalytic site, based on the 3D coordinates of the co-crystal ligand atoms. Drug molecules were also prepared for docking simulations through adding hydrogen atoms and Gasteiger atomic charges, using Open Babel. They will be here referred to as ligands. Docking of all ligands were performed against all receptors using the AutoDock Vina program.⁹

The docking outputs were analyzed and scores from all simulations were retrieved. The minimum, the maximum, and the mean value of all docking scores were calculated to assess their distribution and interval of variation.

In addition, a contact analysis was performed to assess for each receptor the list of residues within 3.5 Å of both the co-crystal ligand and the docked drug considering the best scored pose. For co-crystal ligands and docked drugs, we calculated a contact rate (CR) as the fraction between the total number of contacts established between the target and the ligand divided by the total number of the ligand atoms. The software PyMol, Schrödinger & DeLano¹⁰ was used for visualization of the docking outputs and figures generation.

Results

Datasets presented a satisfactory chemical diversity

The first dataset (AID1063) contained 47 427 inactive compounds and 17 630 active compounds; indicating a non-equilibrated

state, with 72.9% of inactive versus 27.1% of active molecules. We further explored the content of this dataset to estimate its chemical diversity prior to its use for the training of the ML models. We calculated pairwise chemical similarity of the molecules using the RDkit FPs as embedding function and the Tanimoto coefficient as a distance metric. The pairwise distance histogram showed that the dataset presents a satisfactory chemical diversity with an average value of the Tanimoto coefficient equal to 0.28 \pm 0.08 and too few values higher than 0.6 (Figure 1).

Through 2 equilibration simulations based on oversampling and undersampling methods, we generated the ROS and RUS datasets, respectively. Then, we randomly subsampled 10% and 1% of the RUS dataset with respect to class equilibration to obtain the RUS_10% and RUS_1% datasets, respectively. The 4 equilibrated datasets presented equivalent class sizes of 47 427, 17 630, 1736, and 177 for ROS, RUS, RUS_10%, and RUS_1%, respectively (Figure 1).

RF and SVM were identified as the best performing models

For the 5 datasets derived from the AID1063 bioassay, we generated molecular FPs based on the molecules' SMILES. Five molecular FPs were used to encode all molecules, namely: ECFP4, ECFP6, RDFP, APFP, and TTFP. Each encoded dataset was split into a training, a validation and a test set with 80/10/10 proportions. The training sets were used to train 4 ML algorithms and the validation sets were used for hyperparameters optimization. We performed cross-validation with 5-fold stratification and estimated the models' accuracy. Figure 2 shows the obtained results for each algorithm trained on all datasets encoded through the different molecular FPs. Accuracy distributions obtained by each ML model revealed similar distributions for all types of FPs suggesting little impact on the models' performances (Figure 2A; Supplementary Figure S1A). On the other hand, the size and composition of the datasets appeared to be the most influencing condition on accuracy distributions (Figure 2B;

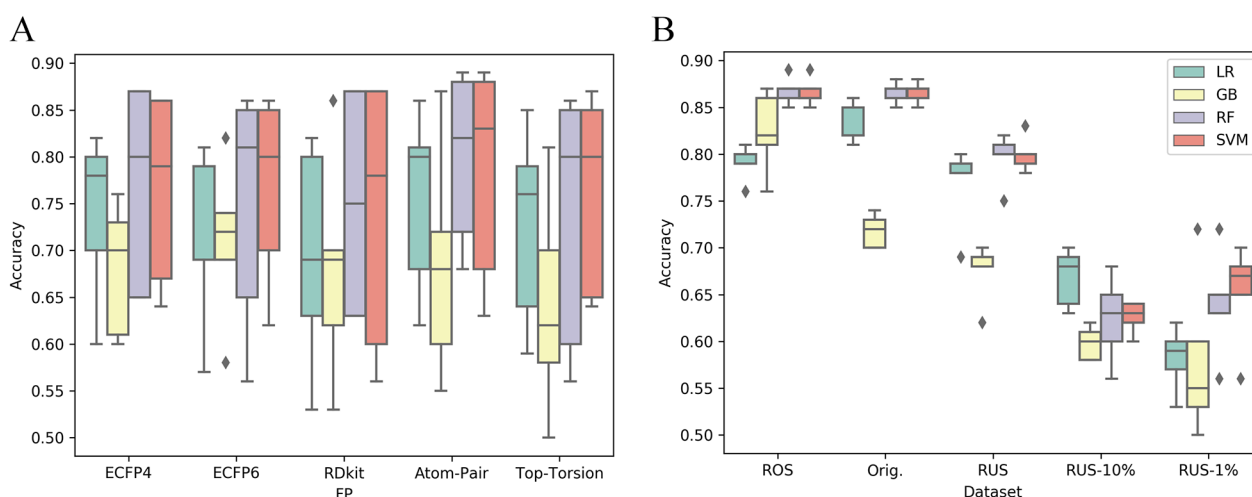


Figure 2. Algorithms' accuracy obtained with the different ML models on the 5 datasets encoded through 5 molecular fingerprints. All simulations included hyperparameters' tuning through a five-fold cross-validation. (A) ML models' accuracy distribution for each type of fingerprints. (B) ML models' accuracy distribution for each dataset. GB indicates gradient boosting; LR, linear regressor; ML, machine learning; RF, random forest; ROS, random over sampling; SVM, support vector machine.

Supplementary Figure S1B). When comparing the models' performances, GB presented the least accuracy values, followed by LR while RF and SVM exhibited the highest accuracy values overall (Figure 2; Supplementary Figure S1C). Noticeably, the ROS dataset encoded with the APFP demonstrated the highest accuracy value of 0.89 obtained with both RF and SVM, while the 10% and 1% subset of the RUS dataset led to values lower than 0.70 (Figure 2B; Supplementary Table S1). Overall, the original dataset presented a comparable accuracy distribution with the ROS dataset, while the RUS dataset led to slightly lower values with the highest obtained with RF and SVM. We chose to focus on these 2 models trained on the original, the ROS and the RUS datasets for further validation.

We aimed at further assessing the impact of the size and the imbalance of the data on models' performances. We generated the ROC and the precision–recall (PR) curves for each model trained on the original, the ROS and the RUS datasets and calculated their AUC scores. All simulations induced comparable performances in terms of AUC_ROC and AUC_PR scores (Figure 3). The highest AUC scores were obtained with the ROS dataset for both models. Nonetheless, simulations on the ROS dataset revealed an overfitting of the models with AUC scores of 0.99 and 0.98 on the training sets with RF and SVM, respectively. Training the models on the original and RUS datasets induced no overfitting with AUC_ROC scores on the training sets ranging from 0.79 to 0.81 and AUC_PR scores ranging from 0.80 and 0.85 for both models. Thus, we retained models trained on the original dataset for further validation.

External validation revealed satisfactory predictive power

The original dataset contained activity information on 65 057 molecules obtained through a preliminary screening (AID 1063).

A subset of 1122 molecules was subject to a confirmatory screening (AID 1258). The results led to confirming 146 active molecules, 963 inactive molecules and 13 molecules with inconclusive activity outcome. We randomly extracted 73 out of the confirmed actives, and 482 out of the inactive molecules from the original dataset to constitute an external validation set. Then, we retrained the RF and SVM models with their optimal parameters on the truncated original dataset using the AID 1063 activity outcome data. RF and SVM achieved TP counts of 59 and 62 out of 73 and FP counts of 193 and 174, respectively (Table 1). As the data are imbalanced, we calculated the balanced accuracy to assess whether the models are correctly classifying both active and inactive molecules. RF and SVM both achieved scores of 0.72. We also calculated the *F1*-score that balances precision and recall on the positive class and obtained scores of 0.37 and 0.39 for RF and SVM, respectively. Both models achieved high FP counts, which is common in cases where class imbalance is important. Nonetheless, we consider the predictive power of both models satisfying with regards to the fraction of active molecules equal to 13% (73/555) of the total test set.

Application of the selected models to FDA-approved collection suggested repurposing some drugs against *Leishmaniasis*

At last, we used the optimized SVM and RF model previously trained on the original dataset to predict which of 1 065 FDA-approved drugs can be considered as anti-*Leishmania* effectors. We sorted the molecules predicted as potentially *Leishmanicidal* according to their probability of being active and selected the top 10 candidates for each model (Table 1). For RF, literature review demonstrated that 5 out of the top 10 molecules were indeed described as anti-*Leishmania* agents, namely, albendazole,¹¹ pyrazinamide (Mendez et al., 2009)¹⁹, domperidone,¹²⁻¹⁴

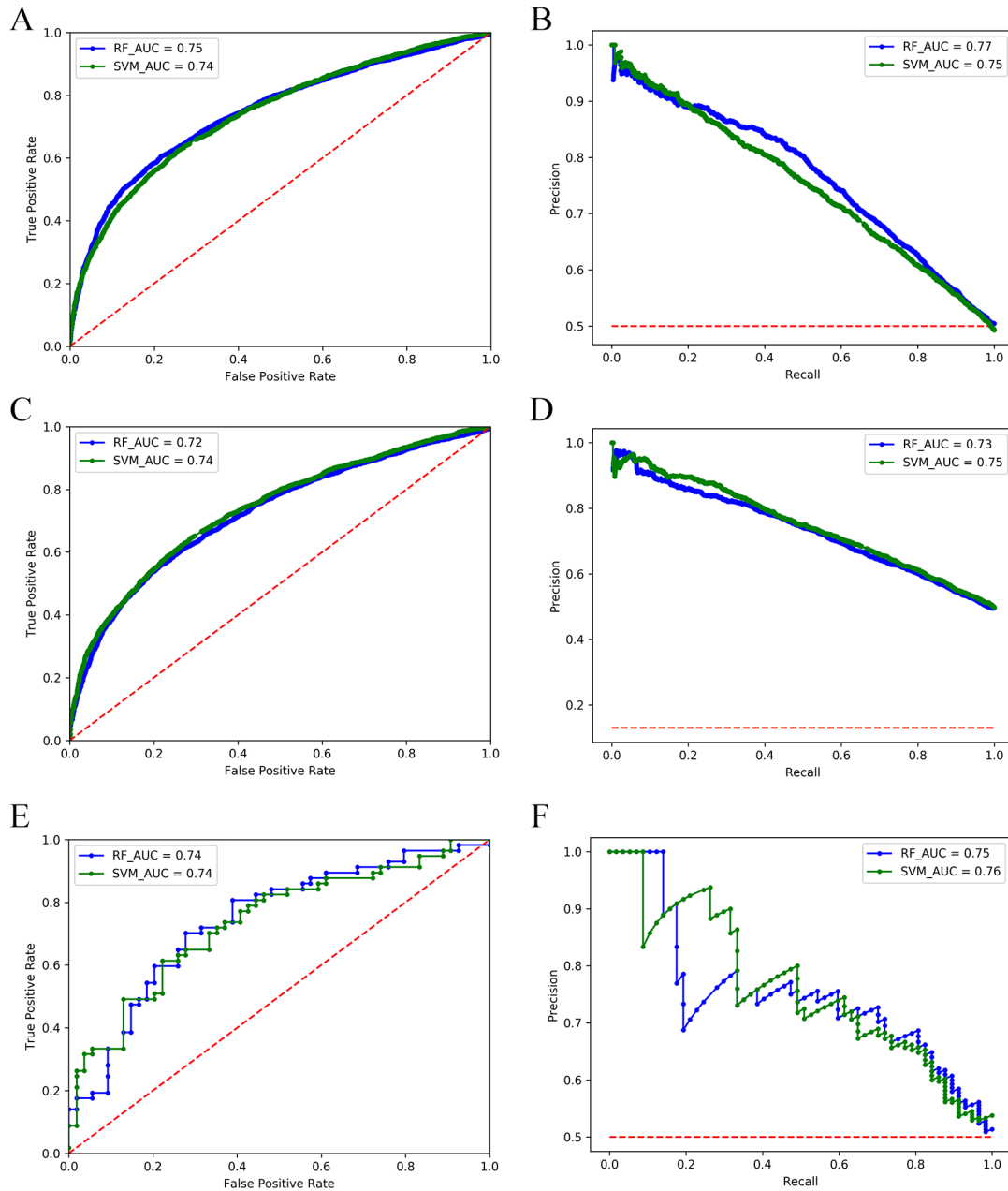


Figure 3. ROC and PR curves of the optimized RF and SVM optimized on the original dataset. (A) ROC curves for models trained on the ROS dataset. (B) PR curves for models trained on the ROS dataset. (C) ROC curves for models trained on the original dataset. (D) PR curves for models trained on the original dataset. (E) ROC curves for models trained on the RUS dataset. (F) PR curves for models trained on the RUS dataset. AUC indicates area under the curve; PR, precision–recall; RF, random forest; ROC, receiver operating characteristic; ROS, random over sampling; SVM, support vector machine.

Table 1. RF and SVM models' performances when trained on the original dataset. For each model TP, FP, TN, and FN counts were reported. Percentages of the TP and TN (shown between parentheses) were calculated as the fraction: (predicted count/real count) × 100.

MODEL	TP (/73)	FP	TN (/482)	FN	SENSITIVITY	SPECIFICITY	BALANCED ACCURACY	F1-SCORE
RF	59 (81%)	193	292 (61%)	11	0.84	0.60	0.72	0.37
SVM	62 (85%)	174	303 (63%)	16	0.79	0.64	0.72	0.39

Abbreviations: FN, false negatives; FPs, fingerprints; RF, random forest; SVM, support vector machine; TN, true negatives, TP: true positives.

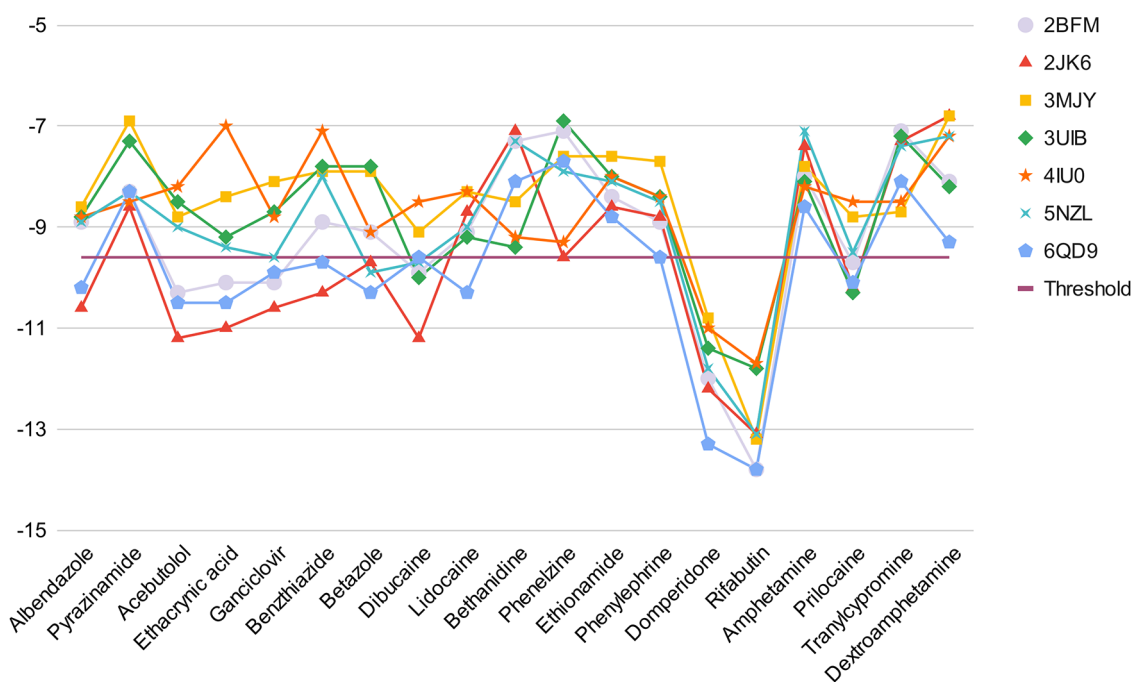


Figure 4. Docking scores for each (receptor, ligand) pair. Scores from the best poses (lowest scores) were retrieved. The threshold value $T = -9.6$ kcal/mol is also indicated for reference.

dibucaine¹⁵, and lidocaine.^{16,17} For SVM, three molecules out of the 10 were described as anti-*Leishmania* effectors, namely, phenelzine,¹⁸ domperidone,¹²⁻¹⁴ and rifabutin.¹⁹ Noticeably, domperidone has been selected by both models. Its potential to treat Leishmaniasis was investigated up to clinical trial phases.¹²⁻¹⁴ Albendazole was explored for its anti-*Leishmania* effects until it proved of limited efficacy in hamster models.¹¹ Also pyrazinamide, an anti-tuberculosis drug showed promising results in early stages of drug discovery investigations.²⁰ On the other hand, investigations of dibucaine's effects had stopped at preclinical stages in hamster models.¹⁵ Lidocaine was described as enhancing the outcome of first line anti-*Leishmania* treatment antimonials through clinical trials.^{16,17} Rifabutin was validated for inhibiting intracellular parasite growth and had promising pharmacokinetic properties.¹⁹ Finally, phenelzine is an antidepressant molecule with validated anti-*Leishmania* effects in vitro and in vivo.¹⁸ To conclude, 7 out of 19 FDA-approved drugs predicted by RF and SVM were previously described for their anti-*Leishmania* effect. This further validated our approach toward the identification of novel anti-*Leishmania* drug candidates either within the FDA-approved drugs collection or within other chemical collections for lead discovery in future studies.

Reverse docking predicted targets among selected Leishmania proteins. The approach used herein to identify novel anti-*Leishmania* effectors is a top-down approach, also called ligand-based drug discovery. It permits the prediction of molecules activity with no prior knowledge on their molecular targets within the cell. Through this method, we were able to

identify, although with no prior knowledge on the targets, 19 drugs among which a set of 7 were previously described in the literature as being anti-*Leishmania* effectors. This constituted a solid confirmation of the proposed approach.

In addition to these confirmed anti-*Leishmania* molecules, the second set of 12 drugs thus corresponds to novel potential effectors. Drugs from both sets have unknown molecular targets. Thus, we performed additional analyses using in silico reverse docking simulations of these 19 potential anti-*Leishmania* drugs against a series of *Leishmania* enzymes identified as potential drug targets, having available crystal structures in complex with ligands in the PDB database. The score values for all molecules docked on all targets followed a Gaussian-like distribution (Supplementary Figure S2), with a maximal value of -6.8 kcal/mol, a minimal value of -13.9 kcal/mol, and a mean value of -9.6 kcal/mol, that we set as a threshold (T) to define relevant docking poses as those with scores inferior to T .

For each (target and drug) pair, we identified the best docking pose as the one exhibiting the lowest docking score (Figure 4). Noticeably, rifabutin and domperidone exhibited the lowest scores on all receptors. Targets 3MJY, 4IU0, and 5NZN had docking scores mostly higher or equal to the threshold T , except with rifabutin and domperidone (Figure 4). On the other hand, 2JK6 followed by 2BFM and 6QD9 presented the lowest scores for all docked ligands, overall.

To identify the most relevant (target and drug) pairs based on the docking results, we calculated the CRs of each docked drug. We then compared them to those observed with the co-crystal ligands of each receptor. Drugs presenting a CR with a given target that is higher or equal to 50% of the CR of the

Table 2. List of the top 10 FDA-approved drugs with the highest probabilities of being anti-Leishmania effectors that were predicted by RF and SVM trained on the original dataset encoded using the atom-pair fingerprints. The compounds already described as anti-Leishmania appear in bold. Four drugs matched with Leishmania targets in our docking experiments. Their docking scores with the identified potential targets are indicated along with the PDB structure used for the reverse docking simulations.

MODEL	DRUG NAME (CHEMICAL CLASS)	PROBABILITY OF BEING ANTI-LEISHMANIAL	REFERENCE IF PREVIOUSLY DESCRIBED AS ANTI-LEISHMANIAL	PREDICTED TARGET REFERENCE	BEST DOCKING SCORE (kcal/mol)
RF	Albendazole	0.97	11	6QD9	-10.2
	Pyrazinamide	0.97	20		
	Acebutolol	0.94	—		
	Ethacrynic acid	0.94	—		
	Ganciclovir	0.94	—	2BFM	-10.1
	Domperidone	0.93	12-14	2JK6	-12.2
	Benzthiazide	0.92	—		
	Betazole	0.92	—		
	Dibucaine	0.92	15		
	Lidocaine	0.92	16		
SVM	Bethanidine	0.84	—		
	Phenelzine	0.83	18		
	Ethionamide	0.82	—		
	Phenylephrine	0.82	—		
	Domperidone	0.821	12-14	2JK6	-12.2
	Rifabutin	0.81	19		
	Amphetamine	0.81	—		
	Prilocaine	0.80	—	3UIB	-10.3
	Tranlycypromine	0.80	—		
	Dextroamphetamine	0.80	—		

Abbreviations: FDA: Food and Drug Administration; PDB, Protein Data Bank; RF, random forest; SVM, support vector machine.

corresponding co-crystal ligand, were retained, which led to identify 4 potential (target and drug) pairs, namely: (2BFM/PTR1, ganciclovir), (2JK6/TR, domperidone), (3UIB/MAPK10, prilocaine), and (6QD9/NMT, albendazole). Corresponding docking scores were, respectively, -10.1, -10.2, -10.3, and -12.2 kcal/mol for ganciclovir, albendazole, prilocaine, and domperidone, indicating high affinity between the drugs and their predicted targets (Table 2). Visual examination of the docking poses from the above-mentioned (target and drug) pairs revealed superimposition of common chemical substructures with the co-crystal ligand (Figure 5), suggesting similar receptor-ligand interactions between the co-crystal ligands and the predicted drugs.

Discussion

DD is a lengthy and costly process that also suffers from a high attrition rate. Classical approaches to drug discovery refer to

screening chemical libraries using either phenotypic or enzymatic assays according to different ligand- or structure-based pipelines.²¹ To reduce time and cost, computational and artificial intelligence approaches are increasingly involved in the drug development process.²² The most common applications of such approaches are (1) novel hit compound identification through docking simulations on targets and (2) predictive analytics of molecules' biological activity based on their chemical structures. The first approach, structure-based, is heavily focused on known targets and aims at identifying potential inhibitors of its biochemical and/or biological activity. The second approach, called ligand-based, focuses on ligands, and uses their chemical structure/properties toward assessing their potential as novel hits with no prior knowledge of their target. For these ligand-based approaches, combined chemoinformatics and ML are now rapidly evolving toward successful drug discovery research outcomes.²³⁻²⁵ Many groups developed

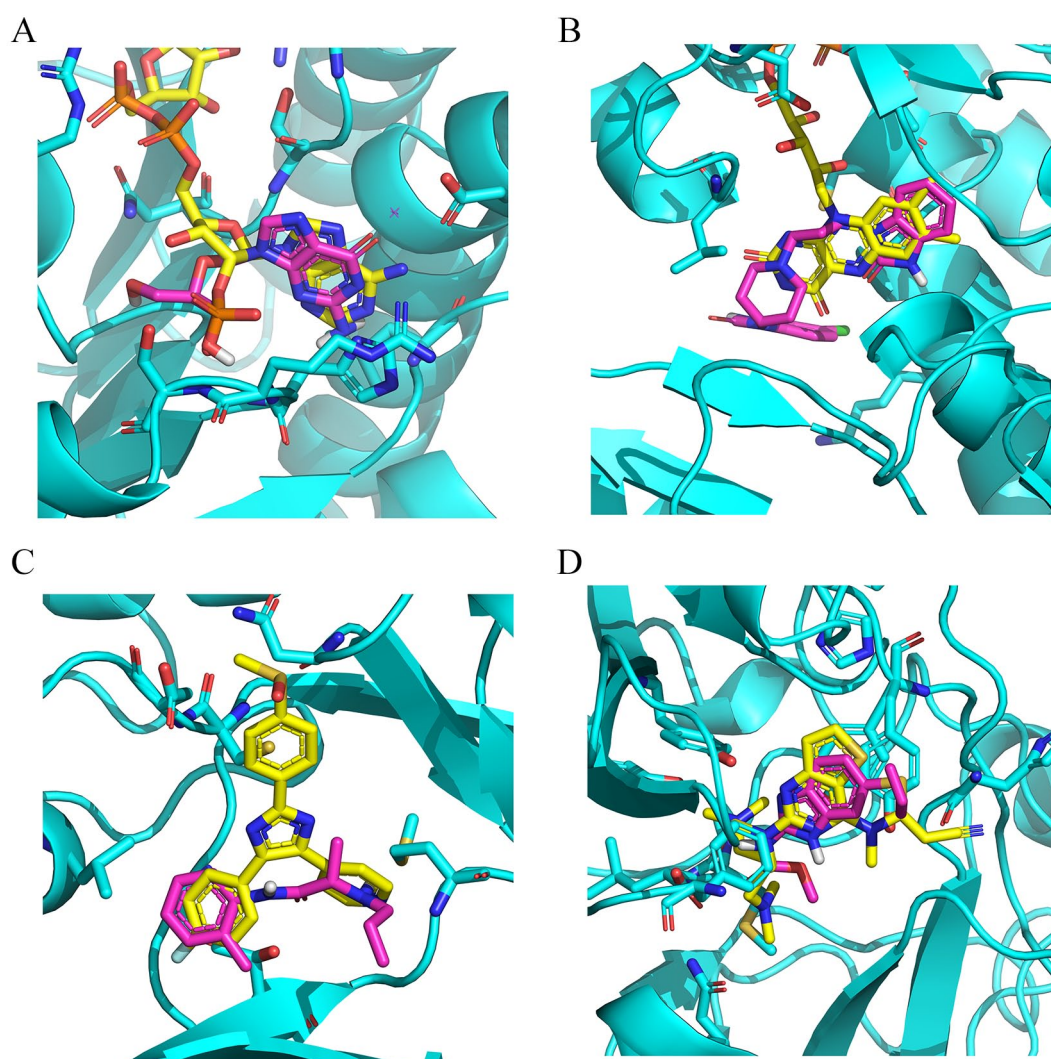


Figure 5. Docking poses of the predicted target–drug pairs using reverse docking simulations. All figures show the protein targets in cyan cartoon representation, with superimposed co-crystal ligands in yellow licorice and docked drugs in magenta. (A) Docking of ganciclovir on the pteridine reductase 1 (PDBid: 2BFM). (B) Docking of domperidone on the trypanothione reductase (PDBid: 2JK6). (C) Docking of prilocaine on the MAPK10 (PDBid: 3UIB). (D) Docking of albendazole on the N-myristoyltransferase (PDBid: 6QD9). MAPK10 indicates mitogen-activated protein kinase 10.

sophisticated deep learning architectures that shall outperform classical ML algorithms. Besides the algorithm performances, the molecular descriptors and/or FPs that are used to encode the chemical structures of the molecules are of utmost importance within the process.^{26–28} In fact, the hypothetical basis of such approaches considers that molecules with similar chemical structures and properties are likely to have similar biological activities through similar molecular interactions with key proteins (targets).

In this context, we implemented a ligand-based approach for the identification of novel drugs and treatments against Leishmaniasis, a group of infectious diseases for which this aim continues to be a challenge and a priority.^{29–31} We presented a proof of concept that classic ML algorithms can achieve satisfying results. In fact, we have put together a pipeline to compare and optimize the combination of the molecular FPs and an ML algorithm to achieve the highest accuracy. In our specific case, RF and SVM demonstrated the highest

accuracy scores on data encoded using the APFPs. Moreover, we performed a series of data stratification to assess the impact of data size and content on algorithms. Results supported the importance of data size toward high accuracy of all ML algorithms tested in the present work. Conversely, data encoding systems (FPs) had no impact on models' performances. Surprisingly, data imbalance also had no significant effect on the performances of the best models, RF and SVM. In fact, most equilibration techniques add novel entries to the dataset through duplication-based algorithms, and are less likely to add new information.

Best performing algorithm, SVM, demonstrated an accuracy of 0.89 and AUC_ROC score of 0.82. Other groups used similar approaches to train and validate ML algorithms RF, SVM, Bayesian models, etc using the CDD TB database Collaborative Drug Discovery Inc toward antimalarial compounds identification.²⁵ Using 4 different datasets of sizes between 1248 and 2273, the authors obtained comparable

AUC_ROC scores for 4 ML algorithms varying between 0.57 and 0.86. These authors have found that no specific model could be considered as outperforming the others and that the chemical diversity and content of the training dataset was the most important parameter. This is in line with our hypothesis and findings as we were able to validate the robustness of our models across the different datasets derived through stratification and balancing techniques. Through an external validation, we were able to provide reliable proof on the models' performances in classifying drugs into active vs inactive against *Leishmania*. We also assessed the models' performances through the balanced accuracy and the *F1*-score. Although highly dependent on the number of TP, the *F1*-score is weighted by the number of FN and FP. Moderate *F1*-score values are often obtained with unbalanced datasets. Korotcov et al³² obtained *F1*-scores for a series of ML models trained on multiple unbalanced datasets that varied between 0.2 and 0.8. Lowest *F1*-scores 0.2–0.4 correlated with high imbalance rates of the data. In the specific case studied herein, where the active class represents only 13% of the test set, low to medium *F1*-scores 0.37–0.39 appeared acceptable. These figures were mainly due to the high number of FP, a recurrent issue with imbalanced sets-based classification in drug discovery,^{28,33,34} even when random oversampling techniques are applied toward equilibration.³ We used the optimized RF and SVM models to predict potential anti-*Leishmania* molecules within the FDA-approved drugs collection. The top 10 predicted molecules by each of the RF and SVM models, respectively, contained 5 and 3 molecules with confirmed anti-*Leishmania* effects.

Our approach has the advantage to overcome the need for a preconceived knowledge about the molecular targets. It uses data from phenotypic screens to train the algorithms and identify active molecules within collections of compounds. We herein used the FDA-approved drugs collection to predict novel anti-*Leishmania* effectors. Nonetheless, the models could be tested on natural products databases or other large chemical collections. This highlights the potential of bioinformatics and artificial intelligence in drug discovery and design. Furthermore, for neglected diseases, such as Leishmaniases, little knowledge on drug targets is available. Despite the recent progress on *Leishmania* species genomes elucidation,³⁵ too few *Leishmania* proteins have resolved crystal structures or are validated as drug targets. Through our AI-based approach (ie, based on a phenotypic drug screening), we were able to overcome this knowledge shortage and predict novel potential anti-*Leishmania* effectors.

Molecules

We complemented these findings with target predictions using reverse docking by referring to the crystal structure of well-defined *Leishmania* drug targets, expecting that some of these could constitute targets for the set of selected molecules.

Despite the current limitations, it was possible to identify some of the selected molecules as potential binders to 4 out of 7 known *Leishmania* targets. Four target and drug pairs were predicted, namely, TPR1, ganciclovir, TR, domperidone, MAPK10, prilocaine, and NTM, albendazole. The docking highlighted alignment of substructures of the co-crystal ligands and the docked drugs. This brings novel insights about the predicted compounds and their potential targets. It also confirms the relevance of using chemical similarity of known active molecules to predict novel bioactive entities. Owing to literature validation of the selected molecules as potentially relevant to DD pipeline to fight Leishmaniases, the developed approach appears accurate. The other novel drugs herein predicted as potential anti-*Leishmania* will need to be investigated in the future. This work supports the suggestion of drug repurposing as a new strategy for discovering anti-*Leishmania* candidates³⁶ or optimizing them. It combined both ligand- and structure-based approaches to validate ML models, identify novel potential anti-*Leishmania* drugs and provide exploratory data on their potential targets in *Leishmania*.

Conclusions

Machine learning has proved its interest and power in the field of drug discovery and development. With the rise of data curation and centralization efforts and the democratization of computational power, it is evolving toward more effectiveness. We herein demonstrated the usefulness of state of the art ML algorithms combined with suitable molecular embedding functions to efficiently predict biological activity of molecules based on their chemical structure. Data availability, size, and content remain a cornerstone in this field. Further development can be made toward quantitative activity predictions, for which dedicated datasets are a prerequisite.


Author Contributions

EH-S conceived the research. EH-S, OS, and IA-T designed the pipeline. EH-S implemented the code and tested the ML models' performances. EH-S and RO performed the docking simulations. EH-S drafted the original manuscript. EH-S, OS, RO, and IG reviewed and edited the manuscript. All authors read and approved the final manuscript.

ORCID iDs

Emna Harigua-Souiai  <https://orcid.org/0000-0003-2974-9157>

Oussama Souiai  <https://orcid.org/0000-0003-2443-114X>

Ines Abdeljaoued-Tej  <https://orcid.org/0000-0002-1796-7897>

Availability of Data and Materials

The datasets used and the code generated during the current study are available in the public repository Github https://github.com/Harigua/ML_DD-applications.

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

- Lo YC, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. *Drug Discov Today*. 2018;23:1538-1546.
- Varnek A, Baskin I. Machine learning methods for property prediction in chemoinformatics: Quo Vadis? *J Chem Inf Model*. 2012;52:1413-1437.
- Korkmaz S. Deep learning-based imbalanced data classification for drug discovery. *J Chem Inf Model*. 2020;60:4180-4190.
- Altae-Tran H, Ramsundar B, Pappu AS, Pande V. Low data drug discovery with one-shot learning. *ACS Cent Sci*. 2017;3:283-293.
- Irwin JJ, Shoichet BK. ZINC—a free database of commercially available compounds for virtual screening. *J Chem Inf Model*. 2005;45:177-182.
- RDKit. Open-source cheminformatics. <http://www.rdkit.org>.
- Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
- O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: an open chemical toolbox. *J Cheminform*. 2011;3:1-14.
- Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov*. 2019;18:463-477. doi:10.1038/s41573-019-0024-5.
- Schrödinger L, DeLano W. PyMOL. <http://www.pymol.org/pymol>. Updated 2020.
- Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. 2010;31:455-461.
- Gómez-Ochoa P, Castillo JA, Gascón M, Zarate JJ, Alvarez F, Couto CG. Use of domperidone in the treatment of canine visceral leishmaniasis: a clinical trial. *Vet J*. 2009;179:259-263.
- Passos SR, de Azevedo Rodrigues T, Madureira AP, Giunchetti RC, Zanini MS. Clinical treatment of cutaneous leishmaniasis in dogs with furazolidone and domperidone. *Int J Antimicrob Agents*. 2014;44:463-465.
- Travi B, Osorio Y. Failure of albendazole as an alternative treatment of cutaneous leishmaniasis in the hamster model. *Mem Inst Oswaldo Cruz*. 1998;93:515-515.
- Perfetti DJC, Hurtado JYY, Reverol NA, de Mirt AS, Acosta MEQ. Anti-Leishmania effect of intralésional procaine and dibucaine in hamsters. *Revista Científica*. 2004; XIV:291-296.
- Weng HB, Chen HX, Wang MW. Innovation in neglected tropical disease drug discovery and development. *Infect Dis Poverty*. 2018;7:67.
- Añez N, Rojas A, Scorza-Dagert JV, Morales C. Successful treatment against American cutaneous leishmaniasis by intralésional infiltration of a generic antimonial compound-lidocaine combination. A follow up study. *Acta Trop*. 2018;185:261-266.
- Evans AT, Croft SL, Peters W, Neal RA. Hydrazide antidepressants possess novel antileishmanial activity in vitro and in vivo. *Ann Trop Med Parasitol*. 1989;83:19-24.
- Bustamante C, Ochoa R, Asela C, Muskus C. Repurposing of known drugs for leishmaniasis treatment using bioinformatic predictions, in vitro validations and pharmacokinetic simulations. *J Comput Aided Mol Des*. 2019;33:845-854.
- Mendez S, Traslavina R, Hinchman M, et al. The antituberculosis drug pyrazinamide affects the course of cutaneous leishmaniasis in vivo and increases activation of macrophages and dendritic cells. *Antimicrob Agents Chemother*. 2009;53:5114-5121.
- Munir A, Vedithi SC, Chaplin AK, Blundell TL. Genomics, computational biology and drug discovery for mycobacterial infections: fighting the emergence of resistance. *Front Genet*. 2020;11:965.
- Zhang H, Pan J, Wu X, Zuo AR, Wei Y, Ji ZL. Large-scale target identification of herbal medicine using a reverse docking approach. *ACS Omega*. 2019;4:9710-9719.
- Yépez J, Cazorla D, de Mirt AS, Añez N, De Yarbuh A. Effect of intralésional treatment with lidocaine and Glucantime® in hamsters infected with Leishmania (Viannia) Braziliensis. *Bol Malarial Salud Ambient*. 1999;39:10-20.
- Gupta MK, Gupta S, Rawal RK. Impact of artificial neural networks in QSAR and computational modeling. In: Puri M, Pathak Y, Sutariya VK, et al, eds. *Artificial Neural Network for Drug Design, Delivery and Disposition*. Academic Press, Elsevier; 2016:153-179.
- Ekins S, Freundlich JS, Reynolds RC. Fusing dual-event data sets for Mycobacterium tuberculosis machine learning models and their evaluation. *J Chem Inf Model*. 2013;53:3054-3063.
- Fan T, Sun G, Zhao L, Cui X, Zhong R. QSAR and classification study on prediction of acute oral toxicity of N-nitroso compounds. *Int J Mol Sci*. 2018;19:3015.
- Mahmoud RS, Yousef AH. Using molecular fingerprints as descriptors in toxicity prediction: a survey. Paper presented at: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); November 18-21, 2019:2649-2654; San Diego, CA. New York: IEEE.
- Harigua-Souiai E, Heinhane MM, Abdelkrim YZ, Souiai O, Abdeljaoued-Tej I, Guizani I. Deep learning algorithms achieved satisfactory predictions when trained on a novel collection of anti-coronaviruses molecules. *Front Genet*. 2021;12:744170.
- Winkler DA. Use of artificial intelligence and machine learning for discovery of drugs for neglected tropical diseases. *Front Chem*. 2021;9:614073.
- Njogu PM, Guantai EM, Pavadai E, Chibale K. Computer-aided drug discovery approaches against the tropical infectious diseases malaria, tuberculosis, trypanosomiasis, and leishmaniasis. *ACS Infect Dis*. 2016;2:8-31.
- Alcântara LM, Ferreira TCS, Gadelha FR, Miguel DC. Challenges in drug discovery targeting TriTryp diseases with an emphasis on leishmaniasis. *Int J Parasitol Drugs Drug Resist*. 2018;8:430-439.
- Korotcov A, Tkachenko V, Russo DP, Ekins S. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. *Mol Pharm*. 2017;14:4462-4475.
- Zheng Y, Peng H, Zhang X, Zhao Z, Gao X, Li J. Old drug repositioning and new drug discovery through similarity learning from drug-target joint feature spaces. *BMC Bioinformatics*. 2019;20:605.
- Trott O, Olson A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2), 455-461.
- Jones NG, Catta-Preta CM, Lima APC, Mottram JC. Genetically validated drug targets in Leishmania: current knowledge and future prospects. *ACS Infect Dis*. 2018;4:467-477.
- Charlton RL, Rossi-Bergmann B, Denny PW, Steel PG. Repurposing as a strategy for the discovery of new anti-leishmanials: the-state-of-the-art. *Parasitology*. 2018;145:219-236.