MOL2NET, International Conference Series on Multidisciplinary Sciences
http://sciforum.net/conference/mol2net-03

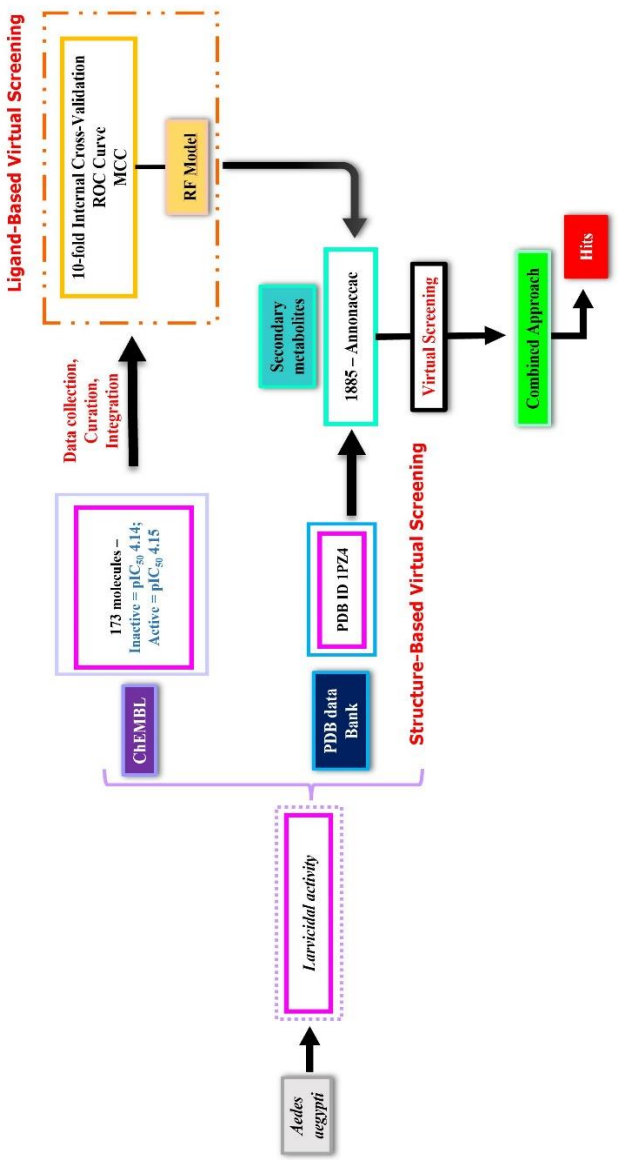
Ligand-Based and Structure-based virtual screening for the discovery of natural larvicidal against *Aedes aegypti*

<Renata Priscila Barros de Menezes> (renatabarros@ltf.ufpb.br)[a], <Chonny Herrera Acevedo> (chonny622@gmail.com)[a], ,<Luciana Scotti> (luciana.scotti@gmail.com)[a] <Marcus Tullius Scotti> (mtscotti@gmail.com)[a]

[a] < *Post-Graduate Program in Natural Synthetic Bioactive Products, Federal university of Paraiba* >

.
.
.

Graphical Abstract



Abstract. The *Aedes aegypti* mosquito belongs to the order Diptera and is one of the main vectors of transmission of etiological agents that cause several diseases. This mosquito can transmit diseases such as dengue, yellow fever, Zika, chikungunya, among others. The aim of this study was combining structure-based and ligand-based virtual screening (VS) techniques to select potentially larvicidal active molecules against *Ae. aegypti* from in-house secondary metabolite dataset (SistematX). From the ChEMBL database, we selected a set of 161 chemical structures with larvicidal activity against *Ae. aegypti* to create random forest models with an accuracy value higher than 82% for cross-validation and test sets. Afterward, the ligand-based virtual screen selected 38 secondary metabolites. In addition, a structure-based virtual screening was also performed for the 38 molecules selected. Finally, using consensus analyzes approach combining ligand-based and structure-based VS, five molecules were selected as potential larvicidal against *Ae. aegypti*.

Keywords: Virtual Screening; Secondary Metabolites; Annonaceae; Larvicidal activity; Aedes aegypti

Introduction

The *Aedes aegypti* mosquito belongs to the order Diptera and is one of the main vectors of transmission of etiological agents that cause several diseases [1]. According to the World Health Organization (WHO), the diseases transmitted by this insect are classified as neglected tropical diseases, as they mainly affect vulnerable socioeconomic populations where there is little investment in their control as well as in the development of treatments [1–3]. This mosquito can transmit diseases such as dengue, yellow fever, Zika, chikungunya, among others [1–3].

The life cycle of *Ae. aegypti* starts in the egg, from which larvae emerge. After going through four stages, the larvae turn into pulps and then into adult mosquitoes [1,4]. The eggs of this mosquito can remain viable for more than a year even without the presence of water, which represents a great threat to the control of *Ae. aegypti* [1,4]. The main method for the prevention and spread of these diseases is vector control, especially during the larval and adult stages [3,5].

Chemical pesticides, despite being effective, can cause several unwanted effects for both man and the environment. In addition to already having reports in the literature of resistance by *Ae. aegypti* to various chemical pesticides [1,2,6–8]. Thus, the search for alternatives to combat the vector is extremely important. Secondary metabolites from plants can be an excellent alternative to search for new insecticides.

In this perspective, a combination of ligand-based and virtual structure-based screening techniques was performed on secondary metabolite Annonaceae dataset to select the best larvicidal active molecules against *Ae. aegypti*.

Materials and Methods

1. Dataset

From the ChEMBL database, was selected a dataset of 161 chemical structures with larvicidal activity against *Ae. aegypti* for construction of predictive models. The compounds were classified as active (85) ($pIC_{50} > 4.15$) or inactive (76) ($pIC_{50} < 4.15$). After a literature search, 11 flavonoids and palmitic acid were added. They had known activity against A. aegypti larvae, with six being identified as active and six as inactive, based on the cuto_ point (pIC_{50}). The compounds were classified using values of – $\log IC_{50}$ (mol/L) = pIC_{50}. In this case, IC_{50} represented the concentration required for 50% inhibition of *Ae. aegypti*.

A dataset of secondary metabolites composed by 1885 structures from Annonaceae were extracted from our in-house databank SistematX available at http://sistematx.ufpb.br [9,10]. This database was used for virtual screening to select the molecules with the highest values of probability to inhibit the *Ae. aegypti*. For all structures, SMILES codes were used as input data in Marvin 19.27.0, 2019, ChemAxon (http://www.chemaxon.com) [11] and Standardizer software [JChem 19.27.0, 2019; ChemAxon (http://www.chemaxon.com)] [12] to canonize structures, add hydrogens, perform aromatic form conversions, clean the molecular graph in three dimensions and save compounds in sdf format.

Molecular descriptors are used to calculate the physicochemical properties of the molecules of each set of molecules. To obtain the molecular descriptors, the DRAGON 7.0 program34 was used [13].

The DRAGON 7.0 software can calculate 5270 molecular descriptors, covering several approaches. These molecular descriptors are arranged in 30 logic blocks [13]. This calculus was realized for all sets of chemical structures.

1.2 Predict Model

The Knime 4.4.2 software (Knime 4.4.2 the Konstanz Information Miner Copyright, 2003–2021, www.knime.org) [14] was used to perform the analyses and to generate the *in silico* model. Datasets of molecules, along with their calculated descriptors and class variables were imported from the Dragon 7.0 software. The dataset was divided using the “Partitioning” tool, with the “stratified sample” option, to create a training set and an external test set, which represented 80% and 20% of the compounds,

respectively. Although the compounds were selected randomly, the same proportion of active and inactive samples was maintained in both sets. Was used the Random Forest algorithm for created the predict model was used 50 Trees and 1 seed for Random generator.

For internal validation, we employed cross-validation using 10 randomly selected, stratified groups, and the distributions according to activity class variables were found to be maintained in all validation groups and in the training set. Descriptors were selected, and a model was generated using the training set and the Random Forest algorithm (RF), using the WEKA nodes [15,16].

The internal and external performances of the selected models were analyzed for sensitivity (true positive rate, i.e., active rate), specificity (true negative rate, i.e., inactive rate) and accuracy (overall predictability). In addition, the sensitivity and specificity of the Receiver Operating Characteristic (ROC) curve were found to describe true performance with more clarity than accuracy. Using Knime nodes the most important descriptors in the generation of prediction model was evaluated.

The model was also analyzed by the Matthews Correlation Coefficient (MCC), a way to evaluate the model globally from the results obtained from the confusion matrix. The MCC is a correlation coefficient between observed and predictive binary classifications. It results in a value between -1 and +1, where a coefficient of +1 represents a perfect forecast, 0 is nothing more than a random forecast, and -1 indicates total disagreement between forecast and observation [17].

The Matthews correlation coefficient can be calculated from the following formula (Equation 1):

$$MCC = \frac{VP\ x\ VN - FP\ x\ FN}{\sqrt{(VP+FP)(VP+FN)(VN+FP)(VN+FN)}} \quad (1)$$

where VP is the value of true positive, VN is the value of true negative, FP is the value of false positives and FN of false negatives.

The domain of applicability (APD) was used to analyze the compounds of the test sets evaluating whether or not their predictions were reliable. The APD is based on Euclidean distances and similarity measures between the descriptors of the training set are used to define the applicability domain, so if a test set compound has distances and similarity beyond this limit, its prediction is not reliable. The APD calculation is performed behind the formula (Equation 2):

$$APD = d + Z\sigma \quad (2)$$

where d and σ are the Euclidean distances and the standard mean deviation, respectively, of the compounds in the training set. Z is an empirical cut-off value, and in this work the Z value was used as 0.5 [18,19].

1.3 Molecular Docking

The target protein of *Aedes aegypti* 1PZ4 [20], with their respective inhibitor ligands were downloaded from Protein Data Bank (http://www.rcsb.org/pdb/home/home.do). All water molecules were deleted from the enzyme structure, and the enzyme and compound structures were prepared using the same default parameter settings in the same software package (score function: MolDock Score; ligand evaluation: internal ES, internal H-bond, sp2–sp2 torsions, all checked; number of runs: 10 runs; algorithm: MolDock SE; maximum interactions: 1500; max. population size: 50; max. steps: 300; neighbor distance factor: 1.00; max. number of conformations returned: 5). The docking procedure was performed using a GRID with a radius of 15 Å and a resolution of 0.30 Å to cover the ligand-binding site in the structures of the four enzymes.

Results and Discussion

Analyzing the *Ae. aegypti* model, you can see that the internal cross validation and the external test demonstrated similar statistical performance, with accuracy higher than 81%, showing to be a model with great performance. The Table 1 summarizes the statistical rates of the RF model.

Table 1: Summary of parameters corresponding to the results obtained in model.

	Specificity	Sensitivity	Accuracy	PPV	NPV
External Test	0.83	0.82	0.82	0.83	0.81
Internal Cross Validation	0.87	0.81	0.85	0.84	0.84

Two parameters were used to evaluate the quality of these binary models: The Receiver Operating Characteristic (ROC) curve and Matthews correlation coefficient (MCC) [17]. In the model, the area under the curve was greater than 94% for the cross-validation sets, and greater than 91% for the test sets, revealing that the models can perform a good classification and prediction rate. Figure 1 shows the ROC curves of the test and cross-validation for the model.

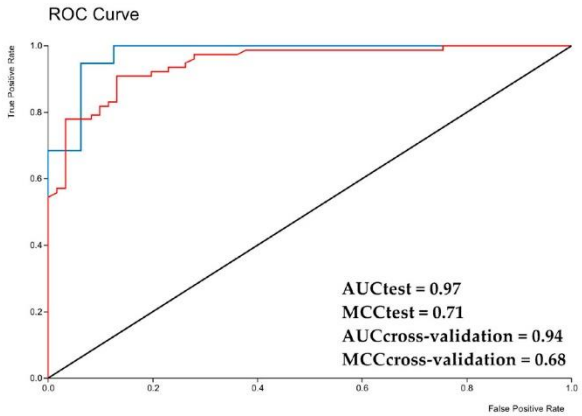


Figure 1: . ROC chart with area under a curve for the Aedes aegypti model test set obtained with Random Forest. AUC – area under the curve; red line – Internal cross validation; blue line – External test.

Of the 1885 secondary metabolites analyzed, the RF model was able to select 1300 chemical structures that are within the applicability domain and its predictions are reliable. These 1300 molecules obtained a prediction between 50 and 91%.

38 molecules had a prediction equal to or greater than 80%, these molecules were then selected to undergo structure-based virtual screening. The chosen protein was Sterol Carrier Protein-2 (PDB ID 1PZ4) which is a protein present in the intestine of the Ae. aegypti. A relevant target since we are analyzing the larvicidal potential of secondary metabolites of Annonaceae.

The molecular docking, Structure-based virtual screening, was first validated by redocking of the original ligand for the 1PZ4 protein. The MolDock scores are listed in Table 2 along with their respective RMSD values and the energies from the PDB.

Table 2: The docking energy (kJ/mol) of the ligand PDB for the 1PZ4 enzyme the *Ae. aegypti*. Ligand energy of the MolDock score and the RMSD values obtained from the redocking procedure.

Proteins	PDB Ligand	Energy PDB (KJ/mol)	Energy Moldock (KJ/mol)	Redocking RMSD
1PZ4	Palmitic acid	-106.543	-113.687	0.29

Therefore, the molecular docking was performed for the 38 molecules with the best's prediction (higher than 80%) in the RF prediction model. Based on the binding energy values, all tested molecules were ranked using the following probability calculation (Equation 3):

$$ps = \frac{E_{TM}}{E_M}, IF\ E_{TM} < E_L \quad (3)$$

where *p*s = structure-based probability, E_{TM} = docking energy of molecule test and TM ranges from 1 to 38 (secondary metabolites dataset); E_M = lowest energy value obtained from tested molecules; E_L = the ligand energy from protein crystallography.

This equation aims to normalize the scores obtained from molecular docking (structure-based virtual screening) so that the values can be compared with the active probability values from the ligand-

based virtual screening [21–23]. In addition, a principle of selection is that the structures must have an energy lower than the value obtained for the ligand in the crystallography study. The secondary metabolites were classified as active if the structure-based probability values are greater than or equal to 0.5. The numbers of molecules with probability values greater than 0.5 and binding energy values less than the ligand was 53, just five molecules don't was predict like active in the molecular docking.

An approach combining structure-based and ligand-based virtual screening was realized to verify potentially active molecules as well as their possible mechanism of action. This approach also seeks to minimize the probability of selecting false positive molecules because it considers the scores of both virtual tracking techniques and correlates them with the true negative rate [21–23]. The calculation is done with the following equation (Equation 4):

$$P_c = \frac{p_s+(1+Esp)\ x\ p}{2+Esp} \quad (4)$$

where P_c is combined probability, p_s is the structure-based probability, Esp is the specificity rate the cross validation and p is the ligand-based probability. In this equation, the ligand-based score is conditioned to a decrease in the false positive rate with the increment of Esp. Thus, the probability of selecting inactive molecules as active molecules is minimized.

Table 3 summarizes the results for the best-ranked molecules obtained using the combined approach, and Figure 2 shows the best-rated structures.

Table 5: Summary of the best-ranked structures obtained using an approach combining ligand-based and structure-based virtual screening; p = active probability value in ligand-based VS; p_s = active probability value in structure-based VS. P_c = combined probability value.

Protein	Molecule	*p*	p_s	p_c
1PZ4	Cherimoline	0.91	0.95	0.92
	Oropheolide	0.85	0.99	0.90
	ester 21-(2-furyl)heneicosa-14,16-diyne-19-(2-furyl)nonadeca-5,7-diynoate	0.90	0.86	0.88
	N-Palmitoyltryptamine	0.89	0.87	0.88
	9,10-Dihydrooropheolide	0.83	0.94	0.88

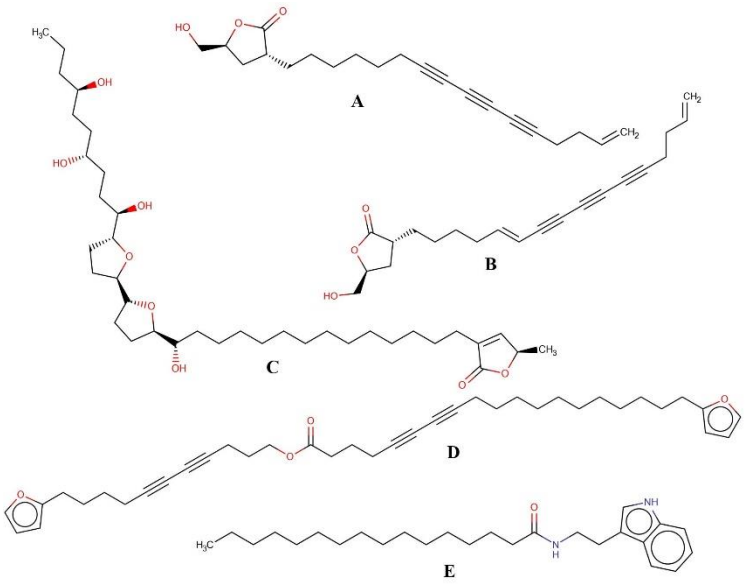
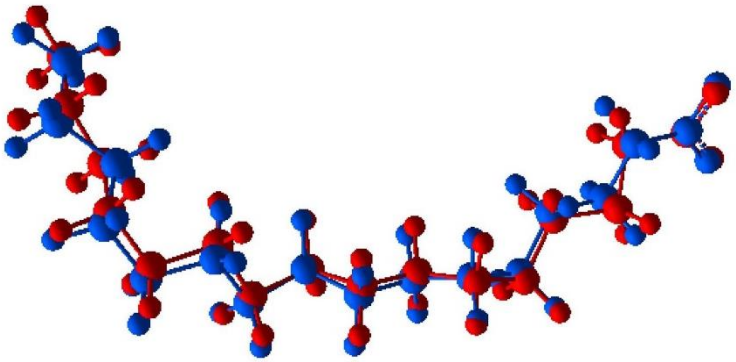


Figure 2: The best rated structures in approach combining structure-based and ligand-based virtual screening; A- Oropheolide; B- 9,10-Dihydrooropheolide; C- Cherimoline; D - ester 21-(2-furyl)heneicosa-14,16-diyne-19-(2-furyl)nonadeca-5,7-diynoate; E – N-Palmitoyltryptamine.

Molecular docking was validated by the redocking and RMSD. The redocking compare the assumed conformation of the binder in redocking with the conformation of the crystallographic ligand. In this analysis we observed that the assumed conformation by the ligand in the redocking and the ligand cristallized was very similar, validating the docking for this enzyme. Figure 3 shows the conformation of the inhibitory linker the enzyme 1PZ4 assumed in the redocking superimposed with the conformation of the inhibitory linker assumed in the X-ray crystallography of the enzyme.



RMSD = 0.29

Figure 3: Redocking of the target protein 1PZ4 against the *Ae. aegypti*. The blue conformation is the conformation of the ligand in X-ray crystallography, and the red conformation is assumed by the redocking.

As observed in the best molecules selected by the approach combining structure-based and ligand-based virtual screening present same common characteristics with the ligand PDB, the palmitic acid. This leads us to believe that these characteristics are important to give the active potential of these molecules.

Conclusions

In this study, we selected five secondary metabolites as potential larvicidal against *Ae. aegypti* through rapid approaches using ligand-based and structure-based VS of 1885 secondary metabolites from Annonaceae, obtained from an in-house database. The compounds selected have structural similarities with other secondary metabolites related in the literature as antiviral compounds. The selected structures are a start point to further studies in order to develop new insecticidal compounds based on natural products.

References

1. Marques, D.M.; Rocha, J. de F.; de Almeida, T.S.; Mota, E.F. Essential Oils Of Caatinga Plants With Deletary Action For Aedes Aegypti: A Review. *South African J. Bot.* 2021, *143*, 69–78.
2. World Health Organization (WHO) *Integrating neglected tropical diseases into global health and development: fourth WHO report on neglected tropical diseases.*; 2017;
3. Fernandes, D.A.; Barros, R.P.C.; Teles, Y.C.F.; Oliveira, L.H.G.; Lima, J.B.; Scotti, M.T.; Nunes, F.C.; Conceição, A.S.; Vanderlei de Souza, M.D.F. Larvicidal compounds extracted from helicteres velutina K. Schum (Sterculiaceae) evaluated against aedes aegypti L. *Molecules* **2019**, *24*.
4. Neves, D.P.; Melo, A.L.; Linardi, P.M.; Vitor, R.W.A. *Parasitologia Humana*; 13th ed.; Atheneu, 2016; ISBN 9788538807155.
5. Nunes, F.C.; Leite, J.A.; Oliveira, L.H.G.; Sousa, P.A.P.S.; Menezes, M.C.; Moraes, J.P.S.; Mascarenhas, S.R.; Braga, V.A. The larvicidal activity of Agave sisalana against L4 larvae of Aedes aegypti is mediated by internal necrosis and inhibition of nitric oxide production. *Parasitol. Res.* **2015**, *114*, 543–549.
6. Brogdon, W.G.; McAllister, J.C. Insecticide resistance and vector control. *Emerg. Infect. Dis.* **1998**, *4*, 605–613.
7. Braga, I.A.; Valle, D. Aedes aegypti: inseticidas, mecanismos de ação e resistência. *Epidemiol. e Serviços Saúde* **2007**, *16*, 279–293.
8. Carson, R. *Silent Spring 40th Anniversary Edition*; Carson Rachel, Ed.; 40th Anniversary Ed….; Mariner Books, 2002;
9. Scotti, M.T.; Herrera-Acevedo, C.; Oliveira, T.B.; Costa, R.P.O.; Santos, S.Y.K. de O.; Rodrigues, R.P.; Scotti, L.;

Da-Costa, F.B. SistematX, an Online Web-Based Cheminformatics Tool for Data Management of Secondary Metabolites. *Molecules* **2018**, *23*, 103.

10. Costa, R.P.O.; Lucena, L.F.; Silva, L.M.A.; Zocolo, G.J.; Herrera-Acevedo, C.; Scotti, L.; Da-Costa, F.B.; Ionov, N.; Poroikov, V.; Muratov, E.N.; et al. The SistematX Web Portal of Natural Products: An Update. *J. Chem. Inf. Model.* **2021**, *61*, 2516–2522.
11. ChemAxon Marvin Copyright © 1998-2021, ChemAxon Ltd. All rights re.
12. ChemAxon Standardizer software Copyright © 1998-2021, ChemAxon Ltd. All rights re.
13. Talete srl Dragon - Software for Molecular Descriptor Calculation) Version 7 Available online: http://www.talete.mi.it/.
14. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD Explor. Newsl.* **2009**, *11*, 26–31.
15. Salzberg, S.; Quinlan, R. Book Review: C4. 5: Programs for machine learning by J. *Ross Quinlan* **1994**, 1–6.
16. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. *The WEKA Data Mining Software: An Update*;
17. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BBA - Protein Struct.* **1975**, *405*, 442–451.
18. Scotti, M.T.; Scotti, L.; Ishiki, H.M.; Peron, L.M.; de Rezende, L.; do Amaral, A.T. Variable-selection approaches to generate QSAR models for a set of antichagasic semicarbazones and analogues. *Chemom. Intell. Lab. Syst.* **2016**, *154*, 137–149.
19. Aptula, A.O.; Roberts, D.W. Mechanistic applicability domains for nonanimal-based prediction of toxicological end points: general principles and application to reactive toxicity. *Chem. Res. Toxicol.* **2006**, *19*, 1097–1105.
20. Dyer, D.H.; Lovell, S.; Thoden, J.B.; Holden, H.M.; Rayment, I.; Lan, Q. The Structural Determination of an Insect Sterol Carrier Protein-2 with a Ligand-bound C16 Fatty Acid at 1.35-Å Resolution *. *J. Biol. Chem.* **2003**, *278*, 39085–39091.
21. Barros, R.P.C.; Scotti, L.; Scotti, M.T. Exploring secondary metabolites database of apocynaceae, menispermaceae, and annonaceae to select potential anti-HCV compounds. *Curr. Top. Med. Chem.* **2019**, *19*.
22. Acevedo, C.H.; Scotti, L.; Scotti, M.T. In Silico Studies Designed to Select Sesquiterpene Lactones with Potential Antichagasic Activity from an In-House Asteraceae Database. *ChemMedChem* **2018**, *13*, 634–645.
23. Lorenzo, V.P.; Lúcio, A.S.S.C.; Scotti, L.; Tavares, J.F.; Filho, J.M.B.; Lima, T.K. de S.; Rocha, J. da C.; Scotti, M.T. Structure- and Ligand-Based Approaches to Evaluate Aporphynic Alkaloids from Annonaceae as Multi-Target Agent Against Leishmania donovani. *Curr. Pharm. Des.* **2016**, *22*, 5196–5203.