©Almeida et al.

LARGE-SCALE PREDICTION OF BIOLOGICAL ACTIVITIES WITH ACTIVE-IT SYSTEM

V.L. Almeida[1,2], O.D.H. dos Santos[3], J.C.D. Lopes[1]*

[1]Chemoinformatics Group — NEQUIM, Departamento de Química, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais (UFMG), av. Pres. Antônio Carlos, 6627, Pampulha, 31.270-901, Belo Horizonte, MG, Brazil; *e-mail: jlopes.ufmg@gmail.com
[2]Serviço de Fitoquímica e Prospecção Farmacêutica, Fundação Ezequiel Dias (FUNED), Belo Horizonte, MG, Brazil
[3]Departamento de Farmácia, Escola de Farmácia, Universidade Federal de Ouro Preto (UFOP), Brazil

Traditional testing methods in pharmaceutical development can be time-consuming and costly, but *in silico* evaluation tools can offer a solution. Our *in-house* Active-IT system, a Ligand-Based Virtual Screening (LBVS) tool, was developed to predict the biological and pharmacological activities of small organic molecules. It includes four independent modules for generating molecular descriptors (3D-Pharma), machine learning modeling (ExCVBA), a database of bioactivity models, and a prediction module. Activity data collected from the PubChem BioAssay database was used for modelling SVM and Naïve Bayes machine learning methods. Models have been constructed using a recursive stratified partition method and validated through an activity randomization (Y-random) process. Over 3500 bioassays were modeled, each comprising 30 SVM and 30 Naïve Bayes models and 60 randomized models. Bioassays with low performance or discrimination between regular and randomized were discarded. Using the Active-IT system we have evaluated three bioactive compounds of Ayahuasca tea. The predictions were thoroughly validated using known targets described in several public databases. The external validation results are noteworthy, with 16 of 33 (48.5% with p-value<0.0001) known targets correctly predicted. This level of accuracy in large-scale virtual screening methods is very significant and demonstrates the effectiveness of the Active-IT methodology in predicting the potential biological activities of small organic molecules.

Key words: ligand-based virtual screening; bioactivity prediction; machine learning modeling; recursive stratified random partition; pharmacophore fingerprint; 3D molecular structures

DOI: 10.18097/PBMC20247006435

INTRODUCTION

Although traditional testing methods in pharmaceutical development can be both time-consuming and costly, the advent of *in silico* evaluation tools offers a promising alternative. The Active-IT system is a cutting-edge platform that combines advanced molecular descriptor generation, machine learning, predictive models, and a dedicated prediction module [1]. This system has proven to be worthy in the field of drug discovery and development [1–5].

Active-IT uses multi-conformational binary pharmacophore fingerprints as molecular descriptors. This innovative approach combines 3D structures and conformational dynamics, moving away from the conventional 2D calculations typically employed in virtual screening [6]. What sets Active-IT apart is its rejection of a single "best model" in favor of a recursive partition approach for developing and validating multiple models. These varied models are also used during the prediction phase, resulting in a more robust and diverse set of predictions.

The primary goal of this study is to present the Active-IT methodology and its validation by investigating the bioactivity of the alkaloids from the stem of *Banisteriopsis caapi* and leaves of *Psychotria viridis*. These two plants are used to make Ayahuasca tea (used in Indigenous religious rituals), known for its psychoactive properties [7]. Our *in-house* Active-IT system targeted alkaloids **1** to **3** (Fig. 1) from these species for predicting potential biological and pharmacological activities [8, 9].

METHODS

The NEQUIM Active-IT system is an innovative ligand-based bioactivity prediction platform with four essential components (Fig. 2) that integrate the creation of descriptors for substances (3D-Pharma), advanced machine learning techniques (ExCVBA), a comprehensive database of activity models, and a robust prediction module (Active-IT). Figure 2 shows a simplified depiction of the Active-IT modules and a detailed modeling process.

The 3D-Pharma module utilizes 3D pharmacophore fingerprints to encode molecular structures from a set of conformations [10]. These fingerprints can help to identify the key features responsible for the biological activity of compounds. Each 3D structure is carefully analyzed, with all non-hydrogen atoms categorized into six groups: positive, negative, hydrogen donor,

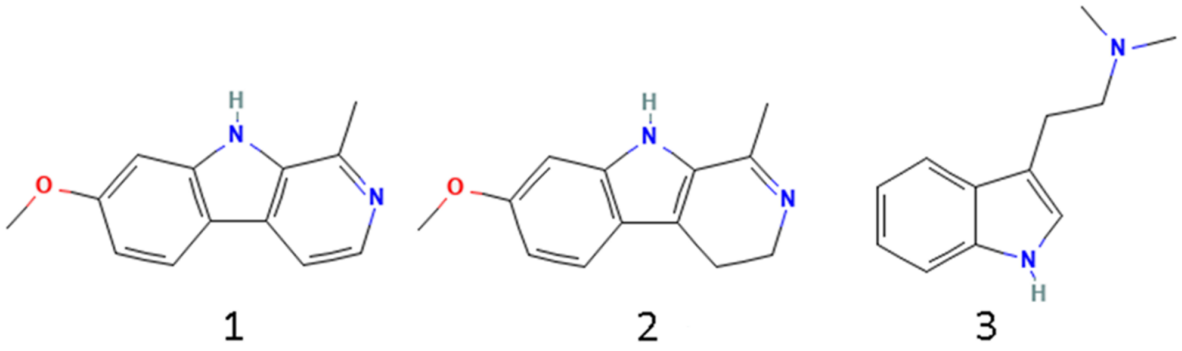

Figure 1. Chemical structure of 7-methoxy-1-methyl-9H-pyrido[3,4-b]indole (harmine **1**) and 7-methoxy-1-methyl-4,9-dihydro-3H-pyrido[3,4-b]indole (harmaline **2**) from stems of *B. caapi*, and 2-(1H-indol-3-yl)-N,N-dimethylethanamine (N,N-dimethyltryptamine **3**) (DMT) from leaves of *P. viridis*.

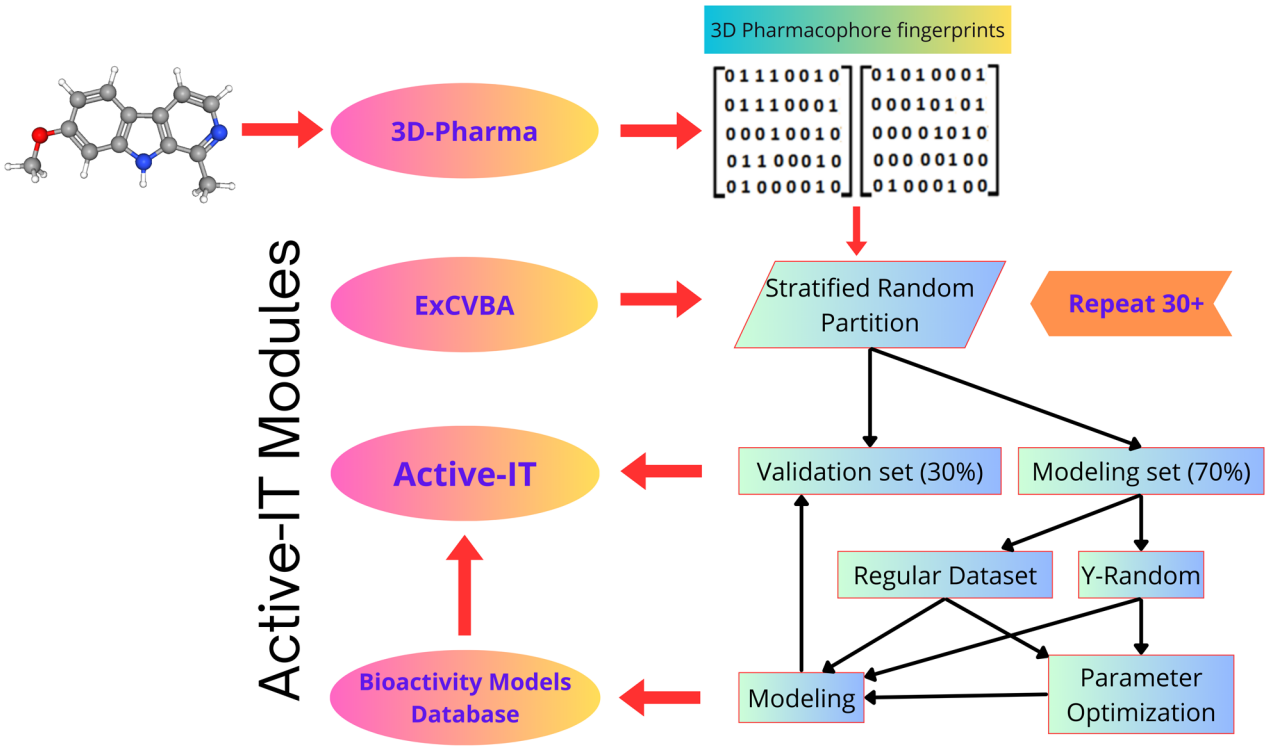

Figure 2. Flow chart of the Active-IT system with its four modules and the modeling protocol using a recursive stratified random partition method controlled by the ExCVBA module is also shown. The color version of the figure is available in the electronic version of the article.

hydrogen acceptor, hydrophobic, and aromatic [11]. Additionally, the 3D-Pharma module calculates all possible combinations of three pharmacophores by considering their distances (discretized in ten distance bins). Each triplet is represented by a 6-character string (3 characters for the feature on vertices and 3 for the distance bins on the edges) that identifies it unequivocally [12]. Each conformation of the compound is associated with an unambiguous vector of 3-point potential pharmacophores (3PPP). To have one representative vector of the compound, all uniconformational vectors of the compound are combined into a unique multiconformational binary modal vector [13]. This single vector contains valuable information about structural details and dynamic behavior of compounds in four dimensions [14].

The Active-IT system relies on extensive data collected from the PubChem BioAssay database [15, 16]. This systematic data collection ensures that our models accurately represent real-world situations. Each compound classified as active or inactive within the *PubChem Outcome* field in each bioassay had its structure (up to 10 conformations) downloaded from the PubChem Compound database [17] and processed as discussed above. We meticulously collected over 3,500 bioassays, assembling them with the ExCVBA module [18] and utilizing the capabilities of supervised machine learning methods, such as Support Vector Machine (implemented through LibSVM by Chang and Lin, [19]) and Naïve Bayes (available in the CPAN repository and developed by Williams [20]) in our custom Perl program.

The construction of the bioactivity models employs a rigorous random recursive stratified partition method. This method ensured that the distribution of active and inactive compounds was consistent in both the test (or validation) and modeling groups. Our approach involved randomly dividing each set of compounds into two subsets, with a 30% and 70% split, without replacement, corresponding to test and modeling groups. This recursive process was repeated at least of 30 times (as illustrated in Fig. 2) for each bioassay, resulting in 30 distinct models for each machine-learning method. To validate the effectiveness of our models, we subjected each dataset to an activity randomization process (known as Y-random). This process generated randomized groups, allowing us to assess the validity of the models against the overfitting risk [21]. As a result, 120 models — 60 regular and 60 randomized — were developed for each bioassay analyzed.

Figure 3 presents the distribution of the area under the receiver operating characteristic curve (AUC-ROC) values for PubChem Bioassay AID 1194 (DSSTox Salmonella Mutagenicity). It shows all four model sets (SVM and Naïve Bayes, both regular and randomized) and provides a compelling visualization of the randomization process. When the models generated through randomizing activities were found to have the same predictive accuracy as conventional models (verified by superposition of AUC distribution curves), they were quickly rejected as unsuitable for accurately predicting potential biological activities. Additionally, the models created for each bioassay were assessed, using only the validation groups, in order to determine the overall AUC for each method. Consequently, we rejected the low-performing modeled bioassay, using the overall AUC>0.6 as an accuracy threshold. About 22% of SVM and 39% of Naïve Bayes modeled datasets were rejected based on these two filters.

The support Vector Machine (SVM) method with linear kernel requires a fine-tuning of cost parameter that we made through a 5-fold cross-validation on the modeling group, using the Power Metric (PM) as the optimization objective [22, 23]. To predict the potential biological activities of new compounds, their structures are obtained from the PubChem Compound database and subjected to the mentioned processing steps. In cases where structures are unavailable, they are constructed with ChemAxon tools (https://www.chemaxon.com), while conformations are obtained using the OpenEye OMEGA program [24].

When beginning to predict a new compound, its multiconformational modal fingerprint is entered into all the models designed for each bioassay. Each model produces a raw score, which is compared with the raw score distribution of active and inactive compounds when they appear in the validation group. Through this comparison, two new values are derived, providing the compound with a likelihood of being either active (Pa) or inactive (Pi) (as illustrated in Fig. 4) [25]. In our previous research work [5], we have set threshold values of 0.5 and 0.8

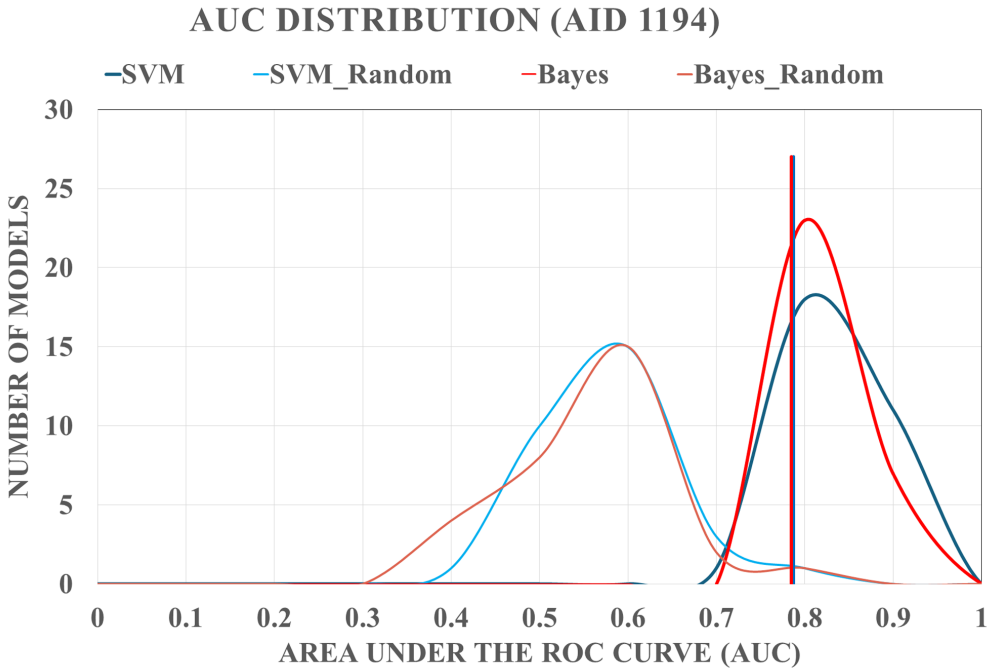

Figure 3. Distributions of the area under the ROC curve (AUC) values for the models generated for each method (SVM and Naïve Bayes) and activity randomization process (Y-random) for the PubChem Bioassay AID 1194 (DSSTox Salmonella Mutagenicity). The vertical lines represent the value of the overall AUC calculated over the full set of models using the validation set only. In the case of this bioassay, the overall AUC exceeds the cutoff value of 0.6 and is considered satisfactory. The color version of the figure is available in the electronic version of the article.

for the Pa – Pi values in the SVM and Naïve Bayes methods, respectively, to be considered promising activity for the compound under study.

The limiting lines in Figure 4 are estimated from the variance of Pa – Pi (σ^2_{Pa-Pi}, analytically calculated as discussed before by Rocha et al. [2] according to Equation (1), where Na and Ni are the number of active and inactive compounds.

$$\sigma^2_{Pa-Pi} = \frac{Pa \times (1 - Pa)}{Na} + \frac{Pi \times (1 - Pi)}{Ni} \quad (1)$$

The confidence interval of Pa – Pi was calculated from the variance and Student *t*-value for a 95% confidence level (Equation 2) [26].

$$(Pa - Pi)_{estimate} = (Pa - Pi)_{mean} \pm t_{stat} \times \sqrt{\sigma^2_{Pa-Pi}} \quad (2)$$

RESULTS AND DISCUSSION

The potential biological activities of Ayahuasca components **1** to **3** were predicted through calculations with the Active-IT system. Their pharmacophore fingerprints were submitted to the Active-IT system to forecast their potential activities using SVM with 2,782 bioassays and Naïve Bayes with 2,176 bioassays. Considering only the bioassays associated with specific biological targets, these numbers are 1,550 for SVM and 1,111 for Naïve Bayes.

For external validation of the predictions, the targets found in the PubChem Compound webpage ("Chemical-Target Interactions") of the compounds harmine (**1**) (77 targets), harmaline (2) (30 targets), and N,N-dimethyltryptamine (**3**) (13 targets) were used. This information was obtained from several databases, such as DrugBank, IUPHAR/BPS, Therapeutic Target Database (TTD), and Comparative Toxicogenomics Database (CTD).

Considering only targets available in the Active-IT system, 33 predictable targets remained. Analyzing only the cases in which the predicted Pa – Pi values exceed the aforementioned cutoff values, the SVM and Naïve Bayes methods correctly predicted 16 targets with high complementarity (Table 1). Thus, around 48% of the targets were correctly predicted.

To estimate the statistical significance of our results, we use the *p*-value, which indicates the probability of obtaining a result equal to or more extreme than the observed due to chance. The *p*-value of the prediction using the Active-IT system was estimated using a hypergeometric distribution (Equation 3), where *N* is the number of modeled bioassays used in the prediction, *k* is the number of known targets, *M* is the number of bioassays selected, and *n* is the number of known targets predicted among the bioassays selected.

$$p\text{-value} = \left(\frac{k}{n}\right)\left(\frac{N-k}{M-n}\right) \quad (3)$$

Table 2 shows the number of bioassays selected for each compound in both prediction methods used to estimate the *p*-values of the predictions. The number of 16 targets predicted using both methods corresponds to a *p*-value<0.0001, demonstrating the high significance of our predictions.

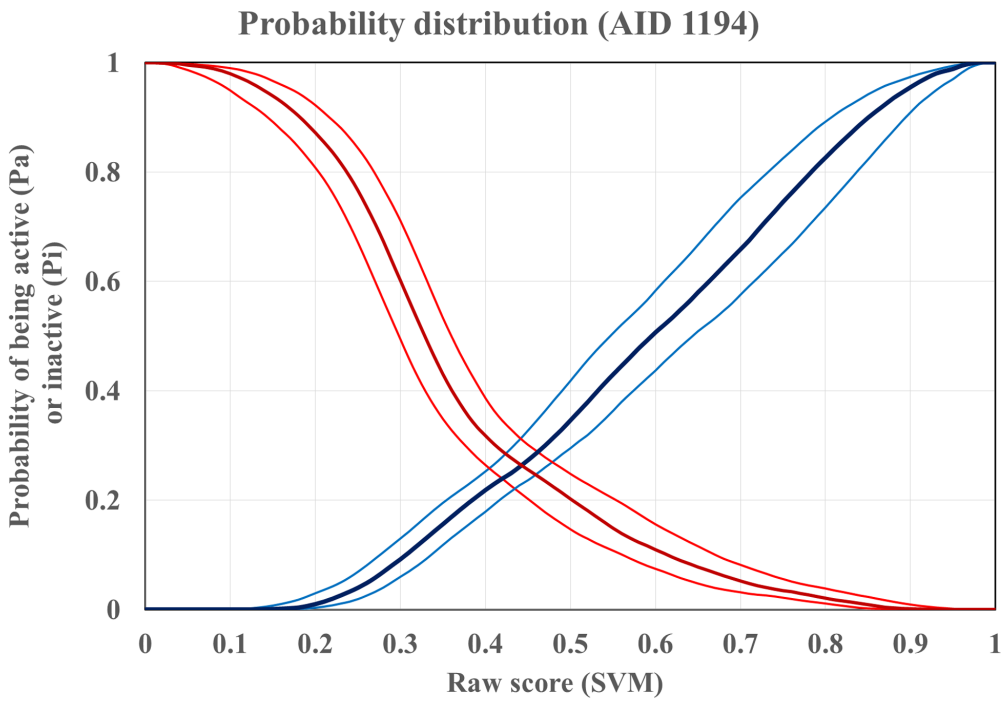

Figure 4. Distribution of the probabilities of a compound being active (Pa), ascending blue lines, and inactive (Pi), descending red lines, as a function of the SVM (or Naive Bayes) raw score for the PubChem Bioassay AID 1194 (DSSTox Salmonella Mutagenicity). The color version of the figure is available in the electronic version of the article.

Table 1. The number of known targets for the analyzed compounds and the number of targets correctly predicted by the Active-IT system

Compound	Targets Known	PubChem Bioassays		
		SVM Selected	Naïve Bayes Selected	SVM or Naïve Bayes Selected
Harmine (**1**)	19	5 (26%)	5 (26%)	8 (42%)
Harmaline (**2**)	11	4 (36%)	4 (36%)	6 (55%)
N,N-Dimethyltryptamine (**3**)	3	1 (33%)	1 (33%)	2 (67%)
Total	**33**	**10 (30%)**	**10 (30%)**	**16 (48%)**

Table 2. The number of modeled PubChem Bioassays associated with biological targets used in the predictions and the number of bioassays selected by each method. The last column shows the *p*-values estimated for the prediction with both methods (SVM or Naïve Bayes)

Compound	Targets Known	PubChem Bioassays			*p*-value
		SVM Selected	Naïve Bayes Selected	SVM or Naïve Bayes Selected	
Harmine (**1**)	19	137	78	193	<0.00005
Harmaline (**2**)	11	125	49	160	<0.00002
N,N-Dimethyltryptamine (**3**)	3	109	32	132	<0.01000
All targets selected	33	371	159	485	<0.00010
All targets predicted	—	1550	1111	2661	—

CONCLUSION

This study focuses on developing and assessing the Active-IT system, a cutting-edge ligand-based virtual screening tool for predicting biological and pharmacological activities of small organic molecules. The Active-IT approach encompasses multi-conformational binary pharmacophore fingerprints for molecular descriptor generation, recursive stratified random dataset partition for machine learning model development, and a robust prediction module for dependable bioactivity predictions.

The Active-IT predictive accuracy was confirmed by evaluating three bioactive compounds from Ayahuasca tea. A remarkable 48.5% (p-value<0.0001) of known targets were accurately predicted. This high level of accuracy in large-scale virtual screening is noteworthy. These external validation results show that the Active-IT system is effective in bioactivity prediction and can contribute significantly to computational drug discovery and development.

ACKNOWLEDGMENTS

The authors thank Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) and Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (CNPq) for financial support.

FUNDING

This study was supported by the Brazilian Science Without Border program (CNPq fellowships 202407/2014-4 to JCDL and 249299/2013-5 to VLA) and FAPEMIG fellowship BIP-00213-24 to VLA.

COMPLIANCE WITH ETHICAL STANDARDS

This article contains no research involving humans or using animals as experimental objects.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

1. *Rocha M.P., Campana P.R.V., Scoaris D.O., Almeida V.L., Lopes J.C.D., Shaw J.M.H., Silva C.G.* (2018) Combined *in vitro* studies and *in silico* target fishing for the evaluation of the biological activities of *Diphylleia cymosa* and *Podophyllum hexandrum*. Molecules (Basel), **23**(12), 3303. DOI: 10.3390/molecules23123303
2. *Rocha M.P., Campana P.R.V., Scoaris D.O., Almeida V.L., Lopes J.C.D., Silva F.A., Pieters L., Silva G.C.* (2018) Biological activities of extracts from *Aspidosperma subincanum* Mart. and *in silico* prediction for inhibition of acetylcholinesterase. Phytother. Res., **32**(10), 2021–2033. DOI: 10.1002/ptr.6133
3. *Briñez-Ortega E., Almeida V.L., Lopes J.C.D., Burgos A.E.* (2020) Partial inclusion of bis(1,10-phenanthroline)silver(I) salicylate in β-cyclodextrin: Spectroscopic characterization, *in vitro* and *in silico* antimicrobial evaluation. Anais da Academia Brasileira de Ciências, **92**(3), e20181323. DOI: 10.1590/0001-3765202020181323
4. *da Silva R.G., Almeida T.C., Reis A.C.C., Filho S.A.V., Brandão G.C., da Silva G.N., de Sousa H.C., de Almeida V.L., Lopes J.C.D., de Souza G.H.B.* (2021) *In silico* pharmacological prediction and cytotoxicity of flavonoids glycosides identified by UPLC-DAD-ESI-MS/MS in extracts of *Humulus lupulus* leaves cultivated in Brazil. Nat. Prod. Res., **35**(24), 5918–5923. DOI: 10.1080/14786419.2020.1803308

5. *Sudan C.R.C., Pereira L.C., Silva A.F., Moreira C.P.S., de Oliveira D.S., Faria G., dos Santos J.S.C., Leclercq S.Y., Caldas S., Silva C.G., Lopes J.C.D., de Almeida V.L.* (2021) Biological activities of extracts from *Ageratum fastigiatum*: Phytochemical study and *in silico* target fishing approach. Planta Medica, **87**(12–13), 1045–1060. DOI: 10.1055/a-1576-4080
6. *Axen S.D., Huang X.P., Cáceres E.L., Gendelev L., Roth B.L., Keiser M.J.* (2017) A simple representation of three-dimensional molecular structure. J. Med. Chem., **60**(17), 7393–7409. DOI: 10.1021/acs.jmedchem.7b00696
7. *Gonçalves J., Luís Â., Gallardo E., Duarte A.P.* (2023) A systematic review on the therapeutic effects of Ayahuasca. Plants, **12**(13), 2573. DOI: 10.3390/plants12132573
8. *Pires A.P., de Oliveira C.D., Moura S., Dörr F.A., Silva W.A., Yonamine M.* (2009) Gas chromatographic analysis of dimethyltryptamine and beta-carboline alkaloids in Ayahuasca, an Amazonian psychoactive plant beverage. Phytochem. Anal., **20**(2), 149–153. DOI: 10.1002/pca.1110
9. *Callaway J.C., McKenna D.J., Grob C.S., Brito G.S., Raymon L.P., Poland R.E., Andrade E.N., Andrade E.O., Mash D.C.* (1999) Pharmacokinetics of Hoasca alkaloids in healthy humans. J. Ethnopharmacology, **65**(3), 243–256. DOI: 10.1016/s0378-8741(98)00168-8
10. *Domingues B.F., Martins-José A., Lopes J.C.D.* (2024) 3D-Pharma, a ligand-based virtual screening tool using 3D pharmacophore fingerprints. ChemRxiv (Preprint), **2024**, DOI: 10.26434/chemrxiv-2024-dkvf8
11. *Sud M.* (2016) Mayachemtools: An open source package for computational drug discovery. J. Chem. Inf. Model., **56**(12), 2292–2297. DOI: 10.1021/acs.jcim.6b00505
12. *Abrahamian E., Fox P.C., Naerum L., Christensen I.T., Thøgersen H., Clark R.D.* (2003) Efficient generation, storage, and manipulation of fully flexible pharmacophore multiplets and their use in 3-D similarity searching. J. Chem. Inf. Comput. Sci., **43**(2), 458–468. DOI: 10.1021/ci025595r
13. *Shemetulskis N.E., Weininger D., Blankley C.J., Yang J.J., Humblet C.* (1996) Stigmata: An algorithm to determine structural commonalities in diverse datasets. J. Chem. Inf. Comput. Sci., **36**(4), 862–871. DOI: 10.1021/ci950169
14. *Domingues B.F., Lopes J.C.D.* (2012) 3D-Pharma: Uma Ferramenta para Triagem Virtual Baseada em Fingerprints de Farmacyforos. [Doctoral dissertation, Universidade Federal de Minas Gerais]. UFMG Institutional Repository. (in Portuguese) Retrieved September 29, 2024 from: http://hdl.handle.net/1843/BUBD-9DKHDA
15. *Kim S., Chen J., Cheng T., Gindulyte A., He J., He S., Li Q.,Shoemaker B.A., Thiessen P.A., Yu B., Zaslavsky L., Zhang J., Bolton E.E.* (2023) PubChem 2023 update. Nucleic Acids Res., **51**(D1), D1373–D1380. DOI: 10.1093/nar/gkac956
16. *Kim S., Bolton E.E.* (2024) PubChem: A Large-Scale Public Chemical Database For Drug Discovery. In: Open Access Databases and Datasets for Drug Discovery (Daina A., Przewosny M., Zoete V., eds.). pp. 39–66. DOI: 10.1002/9783527830497.ch2
17. *Bolton E.E., Chen J., Kim S., Han L., He S., Shi W., Simonyan V., Sun Y., Thiessen P.A., Wang J., Yu B., Zhang J., Bryant S.H.* (2011) PubChem3D: A new resource for scientists. J. Cheminformatics, **3**(1), 32. DOI: 10.1186/1758-2946-3-32
18. *Santos F.M., de Winter H., Augustyns K., Lopes J.C.D.* (2015) Use of extensive cross-validation and bootstrap application (ExCVBA) for molecular modeling of some pharmacokinetics properties. Poster presented at OPENTOX EURO 2015 — OpenTox InterAction Meeting — Innovation in Predictive Toxicology, Dublin, Ireland. DOI: 10.13140/RG.2.1.2274.8888
19. *Chang C., Lin C.* (2011) LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol., **2**(3), 27. DOI: 10.1145/1961189.196119
20. *Williams K.* (2004) Naïve Bayes algorithm at comprehensive perl archive network. Retrieved September 29, 2024 from: https://metacpan.org/pod/Algorithm::NaiveBayes
21. *Tropsha A.* (2010) Best practices for QSAR model development, validation, and exploitation. Mol. Inform., **29**(6–7), 476–488. DOI: 10.1002/minf.201000061
22. *Lopes J.C.D., dos Santos F.M., Martins-José A., Augustyns K., de Winter H.* (2017) The power metric: A new statistically robust enrichment-type metric for virtual screening applications with early recovery capability. J. Cheminformatics, **9**, 7. DOI: 10.1186/s13321-016-0189-4
23. *de Winter H., Lopes J.C.D.* (2018) Reply to the comment made by Šicho, Voršilák and Svozil on "The power metric: A new statistically robust enrichment-type metric for virtual screening applications with early recovery capability". J. Cheminformatics, **10**, 14. DOI: 10.1186/s13321-018-0262-2
24. *Hawkins P.C., Nicholls A.* (2012) Conformer generation with OMEGA: Learning from the data set and the analysis of failures. J. Chem. Inf. Model., **52**(11), 2919–2936. DOI: 10.1021/ci300314k
25. *Filimonov D.A., Lagunin A.A., Gloriozova T.A., Rudik A.V., Druzhilovskii D.S., Pogodin P.V., Poroikov V.V.* (2014) Prediction of the biological activity spectra of organic compounds using the PASS online web resource. Chem. Heterocycl. Compd., **50**(3), 444–457. DOI: 10.1007/s10593-014-1496-1
26. *Nicholls A.* (2014) Confidence limits, error bars and method comparison in molecular modeling. Part 1: The calculation of confidence intervals. J. Comput.-Aided Mol. Des., **28**(9), 887–918. DOI: 10.1007/s10822-014-9753-z

Received: 07. 10. 2024.
Revised: 01. 11. 2024.
Accepted: 03. 11. 2024.

КРУПНОМАСШТАБНОЕ ПРЕДСКАЗАНИЕ БИОЛОГИЧЕСКОЙ АКТИВНОСТИ С ИСПОЛЬЗОВАНИЕМ СИСТЕМЫ ACTIVE-IT

В.Л. Алмейда[1,2], О.Д.Х. дос Сантос[3], Х.С.Д. Лопес[1]*

[1]Chemoinformatics Group — NEQUIM, Departamento de Quimica, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais (UFMG),
av. Pres. Antônio Carlos, 6627, Pampulha, 31.270-901, Belo Horizonte, MG, Brazil; *e-mail: jlopes.ufmg@gmail.com
[2]Servico de Fitoquimica e Prospeccao Farmaceutica, Fundacao Ezequiel Dias (FUNED), Belo Horizonte, MG, Brazil
[3]Departamento de Farmacia, Escola de Farmacia, Universidade Federal de Ouro Preto (UFOP), Brazil

Традиционные методы тестирования при разработке новых фармацевтических препаратов являются трудоёмкими и дорогими, однако инструменты *in silico* оценки могут помочь в решении этой проблемы. Система Active-IT — “инструмент” для проведения виртуального скрининга на основе структуры лигандов, который был разработан нами для предсказания биологической активности малых органических молекул. Она включает в себя четыре независимых модуля: модуль генерации молекулярных дескрипторов (3D-Pharma); модуль машинного обучения (ExCVBA); базу данных о биологических активностях; модуль предсказания. Данные о биологических активностях были получены из базы данных PubChem BioAssay. Для построения моделей машинного обучения использованы метод опорных векторов и наивный байесовский классификатор. Модели были сконструированы с использованием случайного рекурсивного стратифицированного разбиения, их валидацию проводили путём рандомизации данных по активности (Y-random). Были построены модели для 3500 биологических тест-систем, каждая из которых состоит из: (i) 30 моделей, построенных с использованием метода опорных векторов; (ii) 30 моделей, построенных по наивному байесовскому алгоритму; (iii) 60 рандомизированных моделей для валидации. Биологические тест-системы, обладающие низкой производительностью или невысокой дискриминационной способностью, были исключены. С использованием системы Active-IT в данной работе была проведена оценка трёх биоактивных компонентов чая Аяуаска. Прогнозы были проверены с использованием известных мишеней, описанных в нескольких общедоступных базах данных. Результаты внешней валидации показали, что 16 из 33 (48,5%, p<0,0001) известных мишеней были предсказаны верно. Такой уровень точности при крупномасштабном виртуальном скрининге является удовлетворительным, что демонстрирует эффективность методологии Active-IT в прогнозировании биологической активности для малых органических молекул.

Полный текст статьи на русском языке доступен на сайте журнала (http://pbmc.ibmc.msk.ru).

Ключевые слова: виртуальный скрининг на основе структуры лигандов; предсказание биологической активности; машинное обучение; случайное рекурсивное стратифицированное разбиение; фармакофорные фингерпринты; 3D молекулярные структуры

Финансирование. Исследование было поддержано бразильской программой “Наука без границ” (CNPq стипендия 202407/2014-4 to JCDL и 249299/2013-5 to VLA) и FAPEMIG стипендия BIP-00213-24 to VLA.

Поступила в редакцию: 07.10.2024; после доработки: 01.11.2024; принята к печати: 03.11.2024.