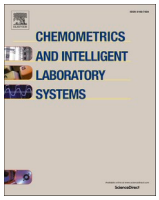
Chemometrics and Intelligent Laboratory Systems 250 (2024) 105145

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics



Identifying 124 new *anti*-HIV drug candidates in a 37 billion-compound database: An integrated approach of machine learning (QSAR), molecular docking, and molecular dynamics simulation

Alexandre de Fátima Cobre [a], Anderson Ara [b], Alexessander Couto Alves [c], Moisés Maia Neto [d], Mariana Millan Fachi [e], Laize Sílvia dos Anjos Botas Beca [f], Fernanda Stumpf Tonin [g], Roberto Pontarolo [h,*]

[a] *Post-graduate Program in Data Science & Big Data, Department of Statistics, Universidade Federal Do Paraná, Curitiba, Brazil*
[b] *Department of Statistics, Universidade Federal Do Paraná, Curitiba, Brazil*
[c] *School of Biosciences and Medicine, Faculty of Health and Medical Sciences, University of Surrey, United Kingdom*
[d] *Post-graduate Program in Pharmaceutical Sciences, Department of Pharmacy, Universidade Federal Do Paraná, Curitiba, Brazil*
[e] *Post-graduate Program in Pharmaceutical Sciences, Universidade Federal Do Paraná, Curitiba, Brazil*
[f] *Department of Pharmacy, Lúrio University, Mozambique*
[g] *H&TRC - Health & Technology Research Centre, ESTeSL, Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa, Lisbon, Portugal*
[h] *Department of Pharmacy, Universidade Federal Do Paraná, Curitiba, Brazil*

ARTICLE INFO

Keywords:
HIV
CCR5
Drug discovery
Machine learning
Molecular docking
Molecular dynamics

ABSTRACT

Recent data from the World Health Organization reveals that in 2023, 38.8 million people were living with HIV. Within this population, there were 1.5 million new cases and 650 thousand deaths attributed to the disease. This study employs an integrated approach involving QSAR-based machine learning models, molecular docking, and molecular dynamics simulations to identify potential compounds for inhibiting the bioactivity of the CC chemokine receptor type 5 (CCR5) protein, a key entry point for the HIV virus. Using non-redundant experimental data from the CHEMBL database, 40 different machine learning algorithms were trained and the top four models (XGBoost, Histogram based gradient Boosting, Light Gradient Boosted Machine, and Extra Trees Regression) were utilized to predict *anti*-HIV bioactivity for 37 billion compounds in the ZINC-22 database. The screening resulted in the identification of 124 new *anti*-HIV drug candidates, confirmed through molecular docking and dynamics simulations. The study underscores the therapeutic potential of these compounds, paving the way for further in vitro and in vivo investigations. The convergence of machine learning and experimental findings presents a promising avenue for significant advancements in pharmaceutical research, particularly in the treatment of viral diseases such as HIV. To guarantee the reproducibility of our study, we have made the Python code (google colab) and the associated database available on GitHub. You can access them through the following link: GitHub Link: https://github.com/AlexandreCOBRE/code.

1. Introduction

The HIV/AIDS pandemic continues to be a highly lethal public health problem, despite the great advances achieved in the field of drug therapy, diagnosis, and social programs. Recent data from the World Health Organization (WHO) show that in 2023, 38.8 million people were living with HIV, 1.5 million new cases and of these 650,000 fatalities from the disease [1].

The CCR5 protein is a chemokine receptor found on human immune system cells, including $CD4^+$ T lymphocytes. The human immunodeficiency virus (HIV) uses the CCR5 receptor as a co-receptor to enter host cell [2]. People who are genetically deficient in the production of the CCR5 receptor have a greater resistance to HIV infection. This discovery has led to the development of antiretroviral drugs that aim to block CCR5 and prevent HIV from entering host cells. An example of an antiretroviral drug that acts as a CCR5 antagonist is maraviroc, which was

* Corresponding author.
E-mail addresses: alexandrecobre@gmail.com (A.F. Cobre), ara@ufpr.br (A. Ara), a.coutoalves@surrey.ac.uk (A.C. Alves), moimaian@gmail.com (M. Maia Neto), marianamfachi@gmail.com (M.M. Fachi), laizebeca@unilurio.ac.mz (L.S.A.B. Beca), stumpf.tonin@ufpr.br (F.S. Tonin), pontarolo@ufpr.br (R. Pontarolo).

https://doi.org/10.1016/j.chemolab.2024.105145
Received 6 February 2024; Received in revised form 9 May 2024; Accepted 12 May 2024
Available online 15 May 2024

approved for clinical use in 2007. Maraviroc is often used in combination with other antiretroviral drugs as part of the treatment of HIV infection in adult patients [3]. The use of drugs that target CCR5 is a promising approach to treating and preventing HIV infection, but drug resistance can occur. Also, it is important to note that these drugs are not a cure for HIV infection and must be used in combination with other antiretroviral drugs [4].

Machine learning plays a key role in the discovery of new drugs, as it allows analyzing large data sets efficiently and identifying patterns that would be impossible to detect manually. Furthermore, machine learning can help predict the effectiveness of new candidate molecules, as well as their safety and toxicity [5,6]. To this end, machine learning can be applied in drug discovery in several ways. One of them is molecular modeling by QSAR (Quantitative Structure-Activity Relationship), which is a technique used in drug discovery that seeks to establish a mathematical relationship between the chemical structure of a molecule and its biological activity. QSAR uses a series of molecular descriptors to quantify the chemical structure of the molecule and then employs statistical techniques to model the relationship between these descriptors and biological activity [7,8]. There are several drugs that have been discovered using QSAR and machine learning approaches, for example Raltegravir - an integrase inhibitor used in the treatment of HIV, Cetuximab - a monoclonal antibody used in the treatment of colon and head and neck cancer, Valsartan - an angiotensin converting enzyme inhibitor used in the treatment of hypertension and heart failure, Sitagliptin - a dipeptidyl peptidase-4 inhibitor used in the treatment of type 2 diabetes and Ibrutinib - a tyrosine inhibitor Ina kinase used in the treatment of chronic lymphocytic leukemia cancer [9,10]. The integration of in silico models such as QSAR, structural similarity, molecular docking, and molecular dynamics has also been used to improve the accuracy of drug candidates for the treatment of various diseases [11,12].

Based on this, this research utilizes a comprehensive strategy that combines QSAR-based machine learning models, molecular docking, and molecular dynamics simulations to pinpoint potential compounds capable of inhibiting the bioactivity of the CCR5 protein—an essential gateway for the HIV virus.

2. Material and methods

Fig. 1 shows the summary of the execution flow of this study. In short, machine learning models based on SHAP values and QSAR approaches were developed, aiming to predict and analyze bioactive compounds that inhibit the human CCR5 protein, for the potential treatment of HIV/AIDS. The study was carried out using the Organization for Economic Cooperation and Development (OECD) guide, which included the following steps: (i) a dataset with an end point; (ii) an exploratory analysis of these data; (iii) using a set of different supervised machine learning algorithms; (iv) use of metrics to evaluate the performance of machine learning models and (v) mechanistic interpretation of machine learning models (analysis of important features) [13]. To guarantee the reproducibility of our study, we have made the Python code (google colab) and the associated database available on GitHub. You can access them through the following link: GitHub Link.

2.1. Description of database used: CHEMBL database

The dataset used in this study for the development of machine learning models for predicting compounds with bioactivity against the human CCR5 protein for the treatment of HIV were collected from the ChEMBL database (https://www.ebi.ac.uk/chembl/). ChEMBL database is a British public domain database containing bioactivity data on over 2.4 million compounds with drug-like properties, distributed across 215 datasets. CHEMBL database, gathers chemical information of compounds, bioactivity data, genomic data, in order to help in the translation of genomic information into new drugs. The CHEMBL database also contains around 86,000 scientific publications, 1.6 million assays, 15,000 therapeutic targets, 6700 drug action mechanisms, 2000 cells, 45,000 drug indications and 782 biological tissues. It is important to highlight that all this information is manually selected and curated by specialists in the field.

2.2. Dataset used for the development of the study

In the CHEMBL database, bioactive compounds that inhibit HIV CCR5 protein (Target ID CHEMBL274) were compiled, which was composed of a total of 5581 bioactivity data corresponding to 3357 compounds. All compound SMILES notations were cured using ChemAxon's Standardizer as described by Simeon (2016) [14]. The initial dataset contained several bioactivity parameters such as EC50, MIC, percentage of activity, Ki, percentage of inhibition, IC50, etc. The IC50 (measured on the nanomolar scale) was selected for further

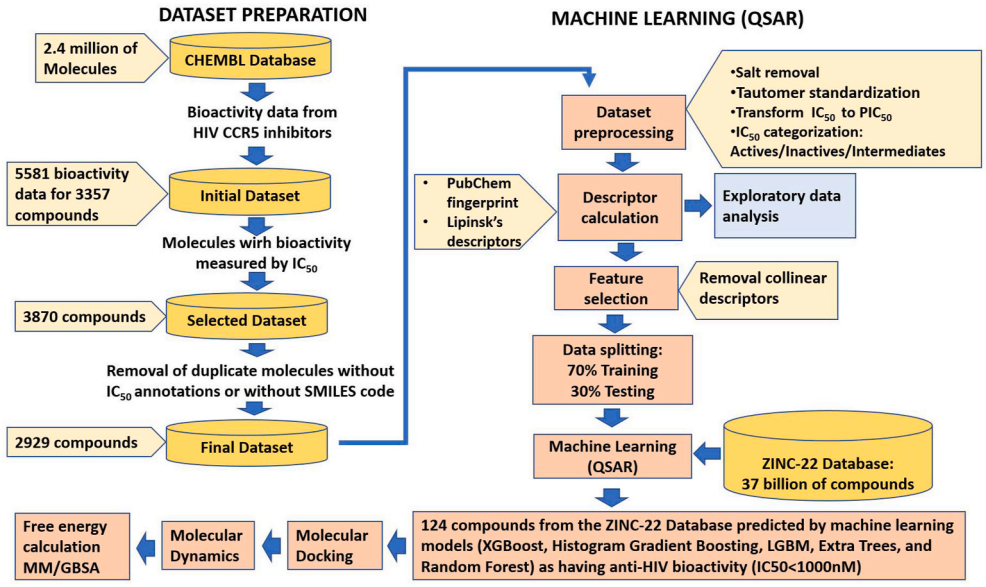


Fig. 1. Flowchart used for data collection and cleaning, and training and validation of machine learning models for predicting the bioactivity of compounds against the CCR5 protein to discover new drug candidates for the treatment of HIV.

investigation as it is the bioactivity parameter that was available in most compounds (n = 3870 compounds). Subsequently, duplicate compounds, compounds without IC50 annotations or compounds that did not have a SMILES code were removed, leaving only 2929 compounds (final dataset) that were used for the analyzes of the next step, machine learning.

. In this analysis, no compound was removed, as all presented complete data. Thus, the final dataset was formed by 2929 bioactive compounds, which were used for the next stage of the study: the pre-processing of the data.

2.3. Dataset pre-processing

The pre-processing of the data of all 2929 bioactive compounds that inhibit the CCR5 protein for the treatment of HIV included in the study, was done using the Padelpy library of the Python language (https://pypi.org/project/padelpy/). This process consisted of removing salts, standardizing tautomer's, converting IC50 to pIC50 (the logarithmic scale). The categorization of ICC50 values was also performed according to the study by Simeon (2016), and the bioactivity was categorized into three groups: active compounds: IC50 < 100 nM; compounds with intermediate activity: IC50 between 100 and 1000 nM; inactive compounds, IC50 > 1000 nM [13]. It's important to note that although this categorization of compound bioactivity may result in unbalanced bioactivity categories, this does not affect the results since the machine learning models used in the studies are based on regression. In other words, bioactivity was not used as a categorical variable, but rather as a quantitative variable, where the dependent variable was the actual bioactivity value [(pIC50 - normalized scale of IC50, i.e., pIC50 = -log (IC50)]). Ligand/target pairs that had different bioactivity values were retained in the database with only one of them being kept.

After this step, the calculation of the fingerprint descriptors (using Pubchem descriptors) of the molecules was carried out. They consist of a set of binary codes that describe the chemical space (2D, 3D, 4D, 5D or even 6D) of molecules. There are 12 different types of fingerprint descriptors (e.g. Pubchem, CDK, substructure, etc) [15]. After a theoretical survey of the literature, we observed that most authors using machine learning-based QSAR models tend to employ PubChem Fingerprint. As our study is a machine learning-based QSAR study, we chose PubChem Fingerprint based on the literature [16–19] The Pubchem descriptor is a binary representation of substructures defined by PubChem. For each molecule, there are 881 Pubchem features [20]. It is also important to highlight that PubChem Fingerprint presents four unique characteristics that distinguish it from other descriptors: (i) Based on extensive public information: PubChem Fingerprint is derived from the public database of PubChem, which is one of the largest and most comprehensive collections of chemical compound information available. This means that PubChem Fingerprint captures a wide and diversified range of molecular characteristics; (ii) Wide coverage of molecular features: PubChem Fingerprint captures a variety of molecular features, including molecular structure, atomic connectivity, functional groups, substructure patterns, among others. This allows for a detailed and comprehensive representation of molecules, which can lead to better QSAR models; (iii) Robustness and generality: Due to its foundation in a vast collection of data, PubChem Fingerprint tends to be robust across a variety of contexts and types of molecules. It is not as prone to issues of overspecificity or inadequate generalization, which can be challenges with some more specialized fingerprint descriptors; (iv) Ease of access and use: Being based on a public database, PubChem Fingerprint is readily available and accessible to researchers. This facilitates study replication, result comparison, and collaboration among different research groups; (v) Support for large-scale modeling: Given its broad nature and data availability, PubChem Fingerprint is suitable for large-scale QSAR studies, where dealing with large datasets and a variety of molecules is necessary [21,22].

2.4. Exploration of multivariate chemical space through Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was employed to conduct the analysis of the multivariate chemical space, aiming to uncover underlying patterns and relationships among the compounds identified as promising candidates for *anti*-HIV treatment by machine learning models.

2.5. Feature selection

Multicollinearity is a serious problem in regression models and is defined as the intercorrelation between pairs of descriptors (features), which causes increased bias and complexity of the developed machine learning model. In order to solve this problem, a variance filter, where the descriptors (features) with variance variability (variance <0.1) were removed from the dataset, aiming to obtain a reduced subset of PubChem descriptors [23]. This entire process was programmed in python language (Fig. 1).

2.6. Data splitting

The Holdout method was carefully employed to partition the data in our QSAR study [24]., ensuring a robust approach in the construction and evaluation of machine learning models. This technique is essential to mitigate any bias in the selection of training, testing, and validation sets, pivotal for the generalization of the obtained results. In the implementation of Holdout, the data were divided into two distinct stages: approximately 70 % of the 2929 bioactive compounds were allocated to the training set, while the remaining 30 % were reserved for the testing set. This strategic division allowed the model to be trained on a substantial amount of diverse data while still maintaining a substantial portion to evaluate its generalization capability on unseen data. By employing the Holdout method, we not only ensured the representativeness of the training and testing data but also guaranteed a reliable assessment of the model's performance in realistic scenarios. This approach was fundamental in ensuring the validity and robustness of our results, providing valuable insights into the relationship between the chemical structure of compounds and their biological activities [24].

2.7. Machine learning

At this stage of the study, all statistical analyzes were performed in Python language. Regression models were implemented, aiming to predict the biological activity (expressed in the form of pIC50) from the chemical structure of the bioactive compounds (expressed in the form of PubChem descriptors). In this analysis, the response variable was the pIC50, while the predictor variables were the PubChem fingerprint descriptors.

The selection of machine learning algorithms was conducted in two stages. Firstly, a screening of various algorithms (n = 40) was conducted using the Lazy Predict library in Python, considering high values of coefficient of determination (R^2) and low values of Root Mean Squared Error (RMSE) as selection criteria [25,26]. Subsequently, the top four best algorithms identified were trained and tested using the Python Sci-kit-Learn library, without hyperparameter optimization, prioritizing those with the highest performance in R^2 [25,26]. The top four best machine learning models were selected to predict new compounds analogous to maraviroc. Additional details about the selection process can be found in the Python source code provided at the end of the abstract of this manuscript, where the screening steps of the algorithms and the criteria used for selecting the best models are documented.

2.8. Important features that contribute to anti-HIV bioactivity of compounds using the SHAP value approach

The interpretation of the important features of the machine learning models was done using the SHAP (SHapley Additive exPlanations) values method. The SHAP values are an interpretability technique in machine learning models that assist in understanding how input features (in this case, PubChem fingerprint descriptors) contribute to the model predictions [27,28]. Specifically, in regression models for predicting the *anti*-HIV bioactivity of compounds, SHAP values can provide valuable insights into which molecular features are most influential in the *anti*--HIV activity of the compounds. The central idea behind SHAP values is to calculate how much each descriptor (variable) contributes to the change in the model prediction relative to its expected contribution. This allows understanding how much each predictor contributes, positively or negatively, to the model outcome. For example, in our study, regression models predict *anti*-HIV bioactivity based on various molecular descriptors derived from PubChem fingerprint descriptors. By calculating SHAP values for these features, it can help uncover that certain fingerprint patterns are associated with higher *anti*-HIV activity, while others are associated with lower activity. This can provide important information for the development of new compounds with greater efficacy against HIV [27,28].

To run *SHAP values*, the "explainer" object was initially created, which helps to analyze an isolated compound or a set of bioactive compounds. Then, the next step was to calculate the average values of each feature using the SoftMax function. The global effect of PubChem fingerprint descriptors (features) was analyzed using the *SHAP plot*, Beeswarm plot, summary plot, and violin plot. On the other hand, to investigate which parts of a specific molecule (features) are important (or not) for biological activity (pIC50) the Bar plot and Waterfall plot were constructed [27,28].

2.9. Use of machine learning models in predicting the anti-HIV bioactivity of compounds analogous to maraviroc

To predict the bioactivity of *anti*-HIV compounds, the following steps were carried out: (i) initially, a screening of 35 billion compounds from the ZINC-22 database (https://cartblanche.docking.org/) was carried out, with the aim of identifying molecules with a structure like that of the drug Maraviroc [29]. (ii) Next, the *anti*-HIV bioactivity of compounds with high structural similarity to Maraviroc was evaluated using the four best selected machine learning algorithms. Compounds with structural similarity scores to maraviroc greater than 50 % were considered significant. Significant bioactivity was those compounds with predicted IC50 values lower than 10000 nM. The assessment of consistency (agreement) of the predicted bioactivity results from the models of the top four machine learning algorithms was performed through the calculation of the percent coefficient of variation, where mean predicted bioactivity values with coefficients of variation less than 5 % were considered consistent.

2.10. Validation of machine learning results via molecular docking and molecular dynamics simulations

The machine learning results allowed the selection of compounds identified as promising, as they possess *anti*-HIV bioactivity via inhibition of the CCR5 protein, and which were subjected to new analyses, using molecular docking and molecular dynamics methods, aiming to carry out a new virtual screening and classification of ligands and obtain information about their binding affinities to the CCR5 protein.

Molecular docking analyses were performed using the Autodock Vina (vina) software according to protocols already described in the literature [30]. Vina was chosen because it is an open-source software considered a first-rate tool for conducting molecular docking, thanks to its speed and accuracy in producing docking results [31–33]. In our analyses, the following parameters were adopted: (i) exhaustiveness of 32; (ii) vina scoring function (iii) optimization of grid box center coordinates for x = 151.223, y = 108.563 and z = 21.821; (iv) optimization of grid box dimensions for x = 40 Å, y = 40 Å, z = 40 Å. The 10 ligands with the highest binding affinity in rigid docking were selected for the new flexible docking process. The cutoff criterion for affinity energy was determined based on literature references [30,34]. To validate the methodology, redocking was carried out using the structure obtained from the protein data bank with PDB ID 4mbs [35,36]. We used the drug maraviroc as a reference standard for defining the flexible residues of the binding site. The molecular docking results were organized based on their binding energies and root mean square deviation (RMSD) values of the 9 poses obtained between our ligands and the CCR5 protein.

The ligand with the best result in flexible molecular docking, more negative binding energy value and lower RMSD value for best pose, was subjected to molecular dynamics (MD) simulations with the CCR5 protein, still using the drug maraviroc as a reference. The CCR5-ligand complex was inserted into a phospholipid bilayer. This membrane was generated on the Charmm-GUI platform respecting the average composition of the human cell membrane [37,38]. The membrane is composed of cholesterol (CHL1, 76 molecules), sphingo lipids (PSM, 36 molecules), phosphatidylinositol (POPI, 17 molecules), phosphatidylethanolamine (POPE, 25 molecules), phosphatidylcholine (POPC, 54 molecules) and phosphatidylserine (POPS, 18 molecules) distributed between the upperleaflet and lowerleaflet faces of the membrane according to the percentages proposed by Escribá and collaborators [37]. The Gromacs 2022.1 software was used in MD calculations with Chemistry at Harvard Macromolecular Mechanics force field (CHARMM36 2021 version) [39,40]. The integration time was 2 fs (fs) and the total simulation time was 100 ns (ns), totaling 50 million steps. The water model used was TIP3P and the physiological ion concentration was 0.15 mM [33]. The optimization of the geometries involved sequential steps with minimization by the Steepest Descents algorithm up to negative Potential Energy of the order of 10^5 to 10^6 and maximum force not exceeding 1000 kJ/mol [40]. The long-range electrostatic cut-off was 1.2 nm, with coulombtype Particle Mesh Ewald (PME). The dynamic equilibrium stages were conducted in two periods of 100 ps (ps), varying between constant temperature (NVT) at 310 K and constant pressure (NPT) with Berendsen barostat [41]. The LINCS algorithm was used for the molecular constraints of holomonic hydrogen bond vibrations. The metrics evaluated RMSD, root mean square fluctuations (RMSF), solvent accessible surface area (SASA) of protein in nm^2, center of mass (COM) of protein-ligand complexes in addition to the calculation of MM/GBSA binding free energy, with graphical representation of the results in Graph software Prism and NCSS.

3. Results and discussion

3.1. Machine learning & QSAR

A total of thirty-two different machine learning algorithms were trained and validated (Figs. S1–S3, in supplementary material), aiming to predict biological activity (pIC50) from PubChem molecular descriptors. It is important to highlight that out of the 881 PubChem descriptors calculated [42], only 204 of them were selected for the training of machine learning models. This selection was carried out using the constant variance method, in which variables (descriptors) with variance less than 0.1 were considered insignificant and therefore excluded. This approach allowed us to minimize the problems of multicollinearity in the trained machine learning models [43,44]. From the thirty-two algorithms trained and tested (using default hyperparameters), the top four best models with high performance in predicting *anti*-HIV bioactivity via inhibition of the CCR5 protein were: Extra Trees regression (R^2 = 0.9276), Extreme Gradient Boosting (R^2 = 0.9202), Light Gradient Boosted Machine (R^2 = 0.8526), and Histogram Based Gradient

Boosting ($R^2 = 0.8520$). Fig. 2 shows the linearity graph of experimental *anti*-HIV bioactivity values and those predicted by machine learning models.

A point worth highlighting is that all of the top four machine learning models that had the greatest capacity in detecting drugs with bioactivity against HIV are based on decision trees (for example, XGBoost, histogram-based boosting, Light Gradient Boosted Machine, extra trees regression), and there are many scientific publications showing the high performance of these algorithms, not only in the area of drug discovery [45–47], but also in other areas, such as omics sciences [48,49]. There are several reasons that explain the high predictive performance shown by these decision tree-based algorithms, and we list some of them: (i) Firstly, these models are versatile and robust, being able to deal with different types of data, including categorical and numerical, without the need for extensive pre-processing, in addition to being resistant to outliers and noisy data. (ii) The inherent ability of decision trees to capture non-linear relationships in data is crucial, especially in situations where the relationships between predictor variables and the response variable are complex and cannot be easily modeled by linear approaches. (ii) Automatic feature selection, assigning importance to each input variable, is another advantage provided by tree-based models. This is valuable for identifying which characteristics have the greatest influence on predictions. (iv) Machine learning models may be susceptible to overfitting. However, techniques such as Gradient Boosting are designed to mitigate this problem by improving model generalization. (v) The ability to naturally capture interactions between variables is another advantage, allowing the model to learn complex relationships between different characteristics. (vi) Finally, the ease of hyperparameter tuning offered by these models provides additional flexibility, allowing you to optimize model performance according to the specific characteristics of the dataset. These characteristics make tree-based models a popular choice in a variety of machine learning scenarios [50,51].

3.2. Interpretation of machine learning models via SHAP values

Figs. S4–S9 of the supplementary material show the most important features in predicting the *anti*-HIV bioactivity of compounds via inhibition of the CCR5 protein. The interpretability of the top four best machine learning models (Xgboost, Extra trees regression, histogram-based boosting and LGBM) in QSAR analysis for drug discovery for HIV treatment was performed using the SHAP values approach [52–55].

In all four machine learning models, the PubchemFP338 descriptor

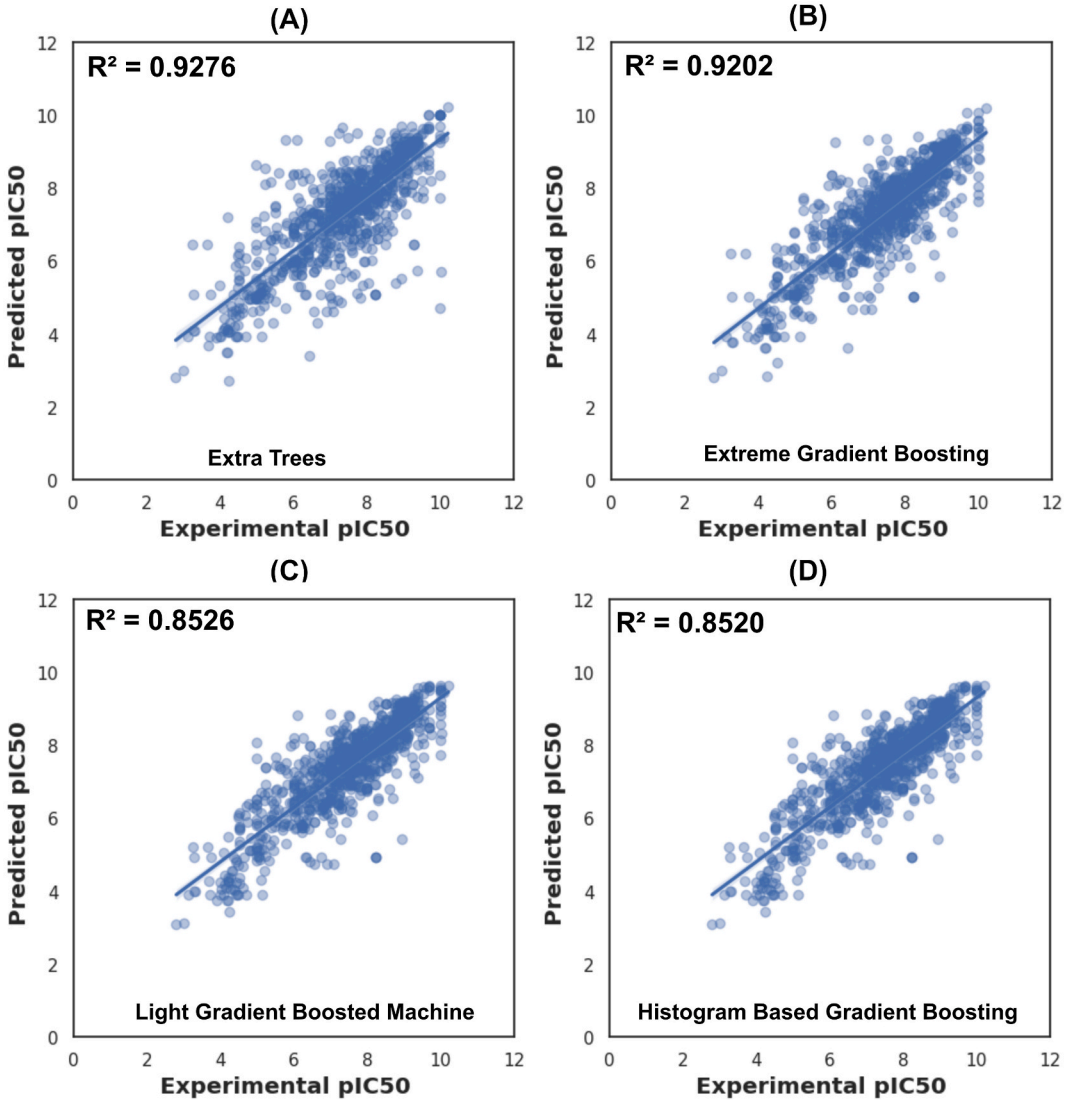


Fig. 2. Curve of experimental values (training and test data) and predictions (validation data) of the bioactivity of compounds against the CCR5 protein for the treatment of HIV. Extra Trees regression, Extreme Gradient Boosting, Light Gradient Boosted Machine, and Histogram Based Gradient Boosting are shown in (A), (B), (C), and (D), respectively.

was the first most important feature in predicting bioactivity against HIV. The presence of the PubchemFP338 feature in the compounds increased their bioactivity (pIC50) against the CCR5 protein for the treatment of HIV (Figs. S3–S7). In scientific literature, PubchemFP338 represents the amino functional group attached to the isopropyl (*N*-isopropylyl amine) group [52–55]. Thus, chemical compounds whose chemical structure presents the functional group, in its structure, increase the inhibitory biological activity against the CCR5 protein for the treatment of HIV. There are QSAR studies also showing that the *N*-isopropylyl substricture (PubchemFP338) is important in inhibiting the protein meprin β, which is a protein involved in the development of various cancers, hyperkeratosis, neurodegenerative disorders, hyperkeratosis, and inflammatory conditions [56]. Other features that also influenced biological activity were the PubchemFP335, PubchemFP13 and PubchemFP566 descriptors (Figs. S3–S7). The feature PubchemFP335 corresponds to an aliphatic hydrocarbon called *n*-butane (Ch3-CH2-CH2-CH3) [57], while the descriptor PubchemFP13 represents the number of carbon atoms greater than or equal to 32 [58]. Finally, PubchemFP566 corresponds to the methyl-group amino ethane (OH–CH2–CH2–NH2). The presence of these three features in the chemical structure of the compounds increased the bioactivity of these compounds in the inhibition of the CCRR-5 protein for the treatment of the HIV virus, and vice versa.

Another most important feature for predicting the bioactivity of compounds against HIV in our study is PubchemFP451. PubchemFP451 represents the formamide functional group C (-N) (=O) [59,60]. In our all models, the four machine learning models showed that the presence of formamide in the chemical structure of chemical compounds decreased the bioactivity of these compounds, on the other hand, the removal of the formamide group increases the bioactivity of the compounds (Figs. S3–S7). Thus, one of the molecular optimization strategies that could be adopted to potentiate the biological activity of the compounds would be the removal of formamide groups in the structures of candidate compounds for CCR5 inhibitor drugs for the treatment of HIV.

Finally, the PubchemFP566 feature, other features that also impacted the biological activity was the PubchemFP566 descriptor, which corresponds to the methyl-amino ethane group (OH–CH2–CH2–NH2). The presence or absence of these functional groups and decreases the bioactivity of the compounds against HIV, respectively.

3.3. Application of the four best machine learning models in the search for new maraviroc analogue compounds

An external database (ZINC-22 Database) containing 35 billion compounds that were not used in model training and testing was used to

Table 1
Predicted *anti*-HIV bioactivity values, expressed as PIC50 (-log IC50), for some of the 124 promising compounds in inhibiting the CCR5 protein (PIC50 > 6). These values were predicted by the four best machine learning models. It is observed that the percentage variation coefficient in the four algorithms is less than 5 %, demonstrating the consistency of the results.

ZINC ID	Predicted *anti*-HIV bioactivity in the form of pIC50 (pIC50 > 6)					Average	Standard deviation	Coefficient of variation (%)
	Similarity score with maraviroc	Histogram based Gradient Boosting	Extreme Gradient Boosting	Light Gradient Boosted Machine	Extra Trees			
ZINC000299762298	0.69	6.32	6.27	6.32	6.41	6.33	0.05	0.80
ZINC000019698225	0.68	6.96	6.71	6.96	6.44	6.77	0.21	3.18
ZINC000223165184	0.67	7.05	7.07	7.05	6.48	6.91	0.25	3.61
ZINC000019370695	0.65	6.31	6.13	6.31	6.11	6.22	0.10	1.53
ZINC000182250475	0.65	6.56	6.54	6.56	7.16	6.71	0.26	3.92
ZINC000257190043	0.64	6.31	6.12	6.31	6.11	6.21	0.10	1.57
ZINC000220159859	0.63	6.19	6.46	6.19	6.11	6.24	0.13	2.13
ZINC000069778775	0.63	5.13	5.54	5.13	5.64	5.36	0.23	4.34
ZINC000169759195	0.62	6.85	6.46	6.85	6.41	6.64	0.21	3.14
ZINC000012718622	0.62	6.51	6.08	6.51	6.7	6.45	0.23	3.52
ZINC000095953914	0.62	6.19	5.73	6.19	6.14	6.06	0.19	3.18
ZINC000019698219	0.62	6.63	6.48	6.63	6.44	6.55	0.09	1.32
ZINC000299801080	0.62	6.43	6.17	6.43	6.18	6.30	0.13	2.02
ZINC000257232629	0.62	6.89	7.27	6.89	7.31	7.09	0.20	2.83
ZINC000328877322	0.61	7.23	7.19	7.23	7.26	7.23	0.02	0.34
ZINC000095392132	0.61	6.23	6.41	6.23	5.94	6.20	0.17	2.72
ZINC000069490977	0.61	5.45	5.41	5.45	4.89	5.30	0.24	4.48
ZINC000581813965	0.61	6.77	6.53	6.77	7.01	6.77	0.17	2.51
ZINC000065384988	0.61	5.96	5.99	5.96	6.46	6.09	0.21	3.49
ZINC000426707376	0.61	6.02	6.24	6.02	6.51	6.20	0.20	3.25
ZINC000068487857	0.61	5.48	5.25	5.48	5.05	5.32	0.18	3.38
ZINC000069659729	0.60	6.71	6.84	6.71	6.58	6.71	0.09	1.37
ZINC000169792086	0.60	7.03	7.03	7.03	6.54	6.91	0.21	3.07
ZINC000169790418	0.60	6.23	5.79	6.23	6.25	6.13	0.19	3.16
ZINC000426477479	0.60	6.1	6.7	6.1	6.51	6.35	0.26	4.11
ZINC000333483163	0.60	7.24	7.09	7.24	7.9	7.37	0.31	4.25
ZINC000077538625	0.60	5.64	5.44	5.64	5.56	5.57	0.08	1.47
ZINC000065287474	0.60	6.18	6.28	6.18	5.93	6.14	0.13	2.11
ZINC000095364697	0.60	7.18	7.47	7.18	7.3	7.28	0.12	1.63
ZINC000534593330	0.60	6.53	6.13	6.53	6.34	6.38	0.17	2.59
ZINC000257318868	0.60	7.00	6.97	7.00	7.49	7.12	0.22	3.05
ZINC000019564570	0.60	5.17	5.14	5.17	4.65	5.03	0.22	4.39
ZINC000091876546	0.60	6.72	7.21	6.72	7.31	6.99	0.27	3.90
ZINC000299758276	0.59	6.22	5.96	6.22	5.87	6.07	0.16	2.57
ZINC000299759540	0.59	6.59	6.55	6.59	7.19	6.73	0.27	3.95
ZINC000065502333	0.59	6.52	6.54	6.52	6.4	6.50	0.06	0.85
ZINC000019713428	0.59	7.71	8.09	7.71	7.2	7.68	0.32	4.12
ZINC000222338399	0.59	7.2	6.72	7.2	7.39	7.13	0.25	3.48
ZINC000065490678	0.59	5.36	4.97	5.36	4.89	5.15	0.22	4.21
ZINC000299785875	0.59	6.29	6.23	6.29	6.41	6.31	0.07	1.04

identify compounds with structural similarity to the maraviroc structure. Among the 37 billion compounds screened, 633 drug-like compounds were identified, which had a structural similarity to maraviroc, with similarity scores varying between 54 and 78 % (Table S1 in supplementary material). Next, the top four best machine learning models were used to predict the *anti*-HIV bioactivity (inhibition of the CCR5 protein) of these 633 compounds. In these analyses, we identified a total of 124 compounds that showed significant *anti*-HIV activity, with IC50 values < 1000 nM (or pIC50 > 6). The coefficient of variation of the bioactivity predicted by the four machine learning models was less than 5 %, showing a high accuracy of these models in predicting the bioactivity of the 124 compounds. An example of some of these 124 compounds ($n = 40$) with their respective *anti*-HIV bioactivity data (pIC50 $>$ 6.0 or IC50 < 1000 nM) are shown in Table 1. The rest of the compounds are shown in Table S2 of the supplementary material. The comprehensive list of all 124 compounds, along with their respective predicted bioactivity values in the four distinct machine learning models, can be found in Table S1 of the supplementary material. In Fig. S4 of the Materials and Methods section, the results of the PCA model are shown, demonstrating the high similarity among the 124 promising compounds.

3.4. Multivariate analysis of chemical space

A multivariate analysis using the PCA [61] model was employed to identify patterns in the chemical space (PubChem descriptors) of the 124 compounds predicted by machine learning models as having *anti*-HIV bioactivity (pIC50 > 6) and those that did not (pIC50 < 6). Fig. 3 shows the PCA model formed by three latent variables (principal components) where it was possible to discriminate between the class of compounds with *anti*-HIV bioactivity ($n = 124$ compounds) and the class of inactive compounds, corroborating with the results obtained by the four machine learning models.

3.5. Molecular docking, molecular dynamics simulations and calculation of MM/GBSA binding free energies

The molecular docking results are depicted in Fig. 4. In (A), the superposition of the maraviroc structures obtained experimentally [25] and generated by docking as the best pose is displayed. The RMSD between these structures corresponds to 0.474 Å, indicating the method's strong predictive capability. In (B), the interaction site of CCR5 with maraviroc is illustrated, highlighting the residues whose side chains formed significant electrostatic interactions and were chosen for flexible docking. In (C), the graph illustrates the binding free energy values of the 9 poses of the pre-selected ligands via rigid coupling. Noteworthy is the ligand ZINC000019713428, which exhibited the highest binding energy release, and the maraviroc pattern, which displayed less deviation between the pose and median values. In (D), the box plot of the RMSD of these conformations of each ligand is presented. It is observable that all median RMSD values of these ligands are below the maraviroc standard and the median less than 2 Å, indicating a quality parameter and indicative of the precision of the results, suggesting a favorable conformational fit of these ligands with the binding site compared to the standard [30].

To exemplify the molecular dynamics results, we selected the ligand ZINC000019713428 among the top 10 ligands ranked by flexible docking out of 126 promising compounds suggested through ML methods. Molecular dynamics simulations were conducted with the CCR5 protein associated with a phospholipid membrane generated on the Charmm-GUI platform, and the results of the selected ligand were compared with the control drug maraviroc. In Fig. S7 of the supplementary material we can visualize the systems and their interactions at time zero of molecular dynamics. In the analysis of Fig. 5, it is apparent, in (A) and (B), that the calculations of the MM/GBSA binding free energies of the CCR5 complexes with ZINC000019713428 and maraviroc exhibited favorable binding energy, with total $\Delta G < 0$ kcal/mol, with ΔG total $= -25.69$ kcal/mol for the former and $\Delta G = -24.42$ kcal/mol for the latter. This data suggests that the compound ZINC000019713428 may have a greater affinity for the site compared to maraviroc. In (C), the RMSD over the 100 ns (ns) of simulation is provided, revealing a smaller oscillation of the CCR5/ZINC000019713428 complex relative to the initial position, indicating greater stability of this ligand compared to the standard. This stability is corroborated by the lower fluctuation shown in most of the CCR5 residues when associated with ZINC000019713428 compared to the complex with maraviroc (D). The analysis of the increase in the solvent accessible surface area (SASA) in the protein complex with ZINC000019713428 suggests possible conformational changes in the protein structure. This increase may indicate a greater exposure of previously protected regions of the protein surface, possibly associated with rearrangements in the protein's conformation to accommodate or interact with the ZINC000019713428 molecule, potentially influencing its interaction with other ligands or molecular partners. In (F), a 3D scatter plot illustrates the displacement of the center of mass (COM) of each complex throughout the simulation. From the graph, we infer that the complex containing the ligand ZINC000019713428 exhibited a more uniform dispersion for each axis (x, y, and z), presenting a more spherical representation throughout the simulation. In contrast, the complex with maraviroc showed greater

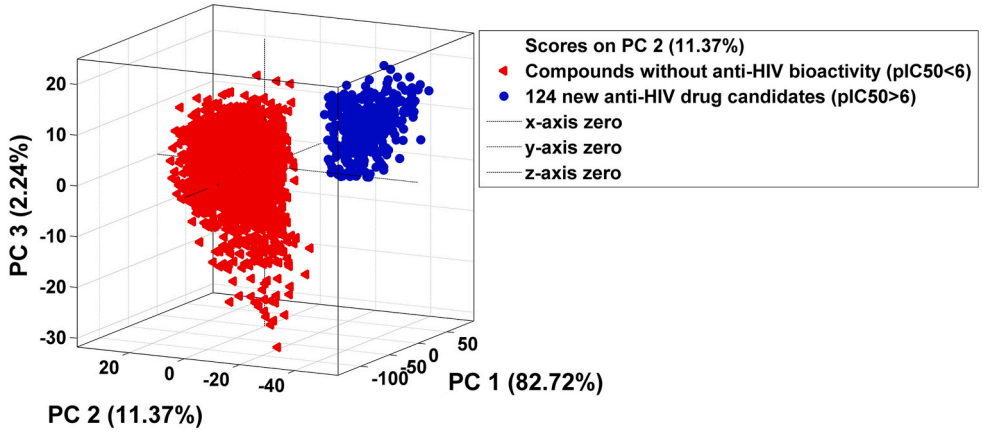


Fig. 3. PCA Model. Discrimination of compounds with *anti*-HIV bioactivity ($n = 124$) versus compounds without bioactivity. Each of the blue circles represents a compound with *anti*-HIV bioactivity (pIC50 > 6) predicted by the machine learning models. The red triangles represent compounds without *anti*-HIV bioactivity (pIC50 < 6) according to the prediction of the machine learning models. It can be observed that there is a discrimination between the class of active and inactive compounds. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

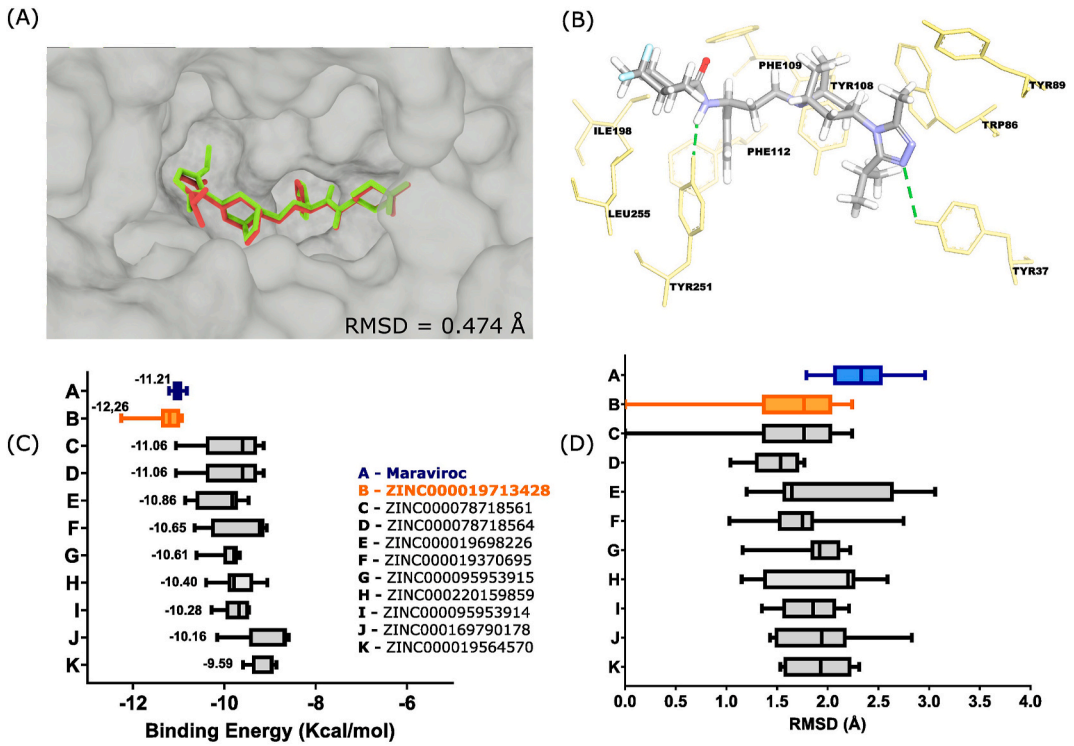


Fig. 4. Results of molecular docking simulations. (A) Shows the result of redocking between the CCR5 protein and the ligand Maraviroc (MRV), with the original structure obtained experimentally represented in green and in red the best pose obtained by docking using Autodock Vina software 1.2.5. In (B), the structure of maraviroc at its binding site with CCR5 is provided. Neighboring residues for which flexible anchoring was used were highlighted in 10 of the 124 promising compounds from rigid docking. In (C) the box plot graph of the binding free energies between the CCR5 protein and the 9 lowest energy poses among the promising flexible coupling compounds is represented. The lowest binding energy values corresponding to the best poses were described on the left. The median RMSD values of all nine conformations generated in each ligand can be seen through the box plot in (D). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

dispersion of COM, assuming a more elongated shape, suggesting greater mobility or conformational changes during the simulation, indicating lower stability compared to the complex with the ligand ZINC000019713428.

This study identified 124 candidates for new *anti*-HIV drugs analogous to maraviroc. However, due to the lack of access to high-performance computing (HPC) equipment, screening was prioritized for compounds with significant similarity to the maraviroc standard, rather than the entire set of billions of compounds in the starting database.

4. Conclusion

In summary, the results of this study highlight the significant impact of machine learning models on the discovery of new drug candidates, especially in the context of HIV treatment through inhibition of the CCR5 protein. The use of advanced algorithms, such as XGBoost, Histogram based Gradient Boosting, LGBM and Extra Trees, demonstrated a remarkable predictive capacity, providing the identification of 124 compounds with inhibitory bioactivity against the CCR5 protein.

These candidates consistently presented coefficients of variation below 5 %, demonstrating the high consistency in predicting the bioactivity of these *anti*-HIV compounds by machine learning models. Furthermore, molecular docking analyses and molecular dynamics simulations corroborated the predictions, revealing a remarkable binding affinity and stability of the identified compounds with the target protein.

In this context, it is crucial to highlight that all 124 compounds shared structural similarities with maraviroc, an FDA-approved *anti*-HIV drug. These results strongly suggest that these compounds emerge as promising drug candidates for the treatment of HIV, offering a solid basis for future in vitro and in vivo studies that further evaluate the therapeutic potential of these substances.

Thus, the intersection between the innovation provided by machine learning models and real experimental discoveries reveals a promising path for significant advances in pharmaceutical research, bringing new hope for the treatment of viral diseases such as HIV.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

CRediT authorship contribution statement

Alexandre de Fátima Cobre: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Anderson Ara:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Alexessander Couto Alves:** Writing – original draft, Methodology, Formal analysis, Conceptualization. **Moisés Maia Neto:** Writing – review & editing, Methodology, Formal analysis. **Mariana Millan Fachi:** Writing – review & editing, Data curation. **Laize Sílvia dos Anjos Botas Beca:** Writing – review & editing, Writing – original draft, Methodology. **Fernanda Stumpf Tonin:** Writing – review & editing, Writing – original draft, Formal analysis, Data curation. **Roberto Pontarolo:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

Not applied.

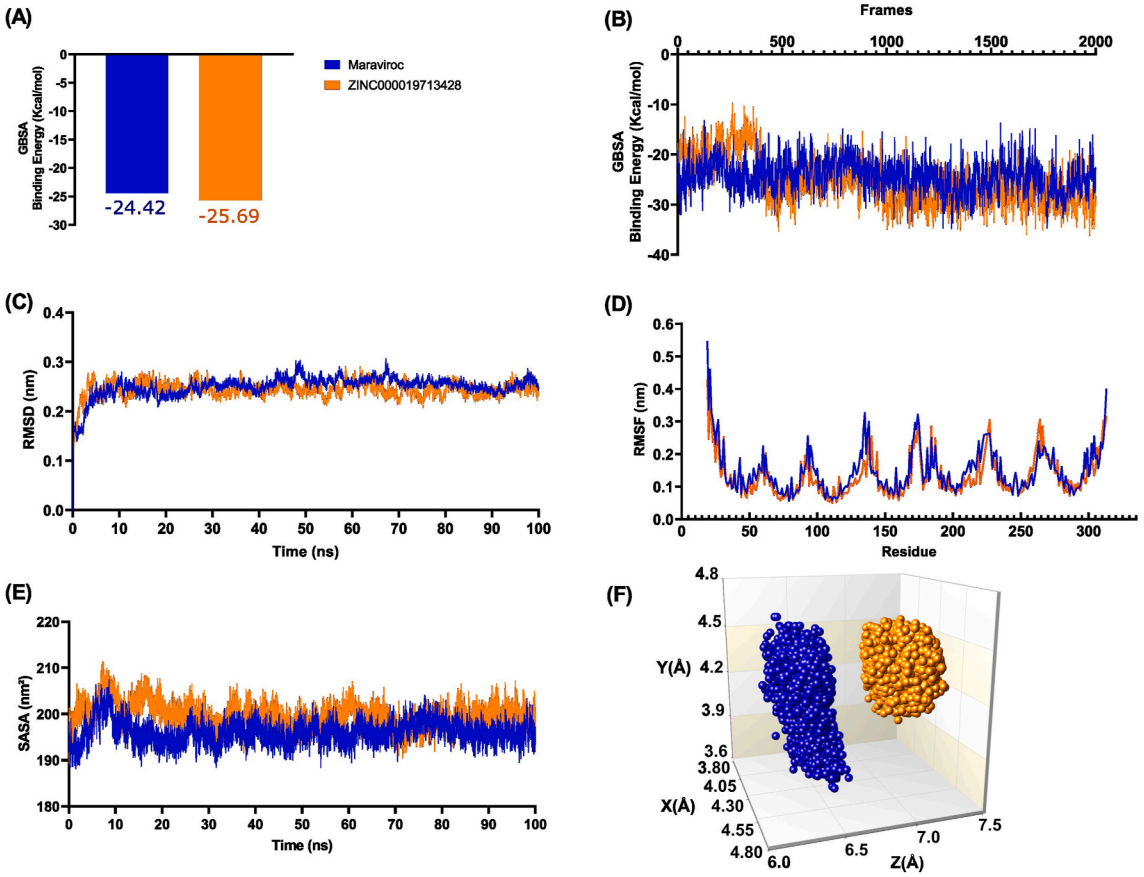


Fig. 5. Results of molecular dynamics (MD) simulations. In (A) we have the total value of the binding free energy (ΔG) obtained by calculating the Molecular Mechanics Generalized Born Surface Area (MM/GBSA) of the ligand with the best docking performance (ZINC000019713428), always represented in orange, and the Maraviroc standard (MRV), always represented in blue. In (B) we have the GBSA representation of both throughout the simulation and we can see that ZINC000019713428 remains below the Maroviroc energy values during most of the simulation. In (C) we have the temporal variation of the RMSD of the protein-ligand complexes of both and the similar stability of this complex is evident with reduced variations in the simulated period. Graph (D) provides the map of fluctuations of CCR5 protein residues complexed with each of the ligands described and we can observe that, in general, the fluctuation of these residues will be smaller for ZINC000019713428. In (E) we have the solvent accessible surface area (SASA) of the CCR5 protein when complexed with the two ligands, and it is possible to observe a higher SASA value in the case of ZINC000019713428. In (F) we have the representation of the displacement of the center of mass of the protein-ligand complex over time through the 3D scatter plot. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors express their gratitude to the Brazilian National Council of Technological and Scientific Development (CNPq) and CAPES (Brazilian Federal Agency for Support and Evaluation of Graduate Education within the Ministry of Education of Brazil) for research funding – Finance Code 001.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2024.105145.

References

[1] World Health Organization, No Title, 2023. https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/hiv/strategic-information/hiv-data-and-statistics. (Accessed 21 March 2023).
[2] P. Lusso, New EMBO Member's Review - HIV and the chemokine system: 10 years later, EMBO J. 25 (2006) 447–456, https://doi.org/10.1038/sj.emboj.7600947.
[3] R.M. Gulick, J. Lalezari, J. Goodrich, N. Clumeck, E. DeJesus, A. Horban, J. Nadler, B. Clotet, A. Karlsson, M. Wohlfeiler, J.B. Montana, M. McHale, J. Sullivan, C. Ridgway, S. Felstead, M.W. Dunne, E. van der Ryst, H. Mayer, Maraviroc for previously treated patients with R5 HIV-1 infection, N. Engl. J. Med. 359 (2008) 1429–1441, https://doi.org/10.1056/NEJMoa0803152.
[4] M.M. Lederman, A. Penn-Nicholson, M. Cho, D. Mosier, Biology of CCR5 and its role in HIV infection and treatment, JAMA 296 (2006) 815–826, https://doi.org/10.1001/jama.296.7.815.
[5] P. Schneider, W.P. Walters, A.T. Plowright, N. Sieroka, J. Listgarten, R.A. J. Goodnow, J. Fisher, J.M. Jansen, J.S. Duca, T.S. Rush, M. Zentgraf, J.E. Hill, E. Krutoholow, M. Kohler, J. Blaney, K. Funatsu, C. Luebkemann, G. Schneider, Rethinking drug design in the artificial intelligence era, Nat. Rev. Drug Discov. 19 (2020) 353–364, https://doi.org/10.1038/s41573-019-0050-3.
[6] T. Ching, D.S. Himmelstein, B.K. Beaulieu-Jones, A.A. Kalinin, B.T. Do, G.P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M.M. Hoffman, W. Xie, G.L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A.E. Carpenter, A. Shrikumar, J. Xu, E.M. Cofer, C.A. Lavender, S.C. Turaga, A.M. Alexandari, Z. Lu, D.J. Harris, D. DeCaprio, Y. Qi, A. Kundaje, Y. Peng, L.K. Wiley, M.H.S. Segler, S.M. Boca, S. J. Swamidass, A. Huang, A. Gitter, C.S. Greene, Opportunities and obstacles for deep learning in biology and medicine, J R Soc Interface 15 (2018), https://doi.org/10.1098/rsif.2017.0387.

[7] Y.-C. Lo, S.E. Rensi, W. Torng, R.B. Altman, Machine learning in chemoinformatics and drug discovery, Drug Discov. Today 23 (2018) 1538–1546, https://doi.org/10.1016/j.drudis.2018.05.010.
[8] L. Zhang, J. Tan, D. Han, H. Zhu, From machine learning to deep learning: progress in machine intelligence for rational drug discovery, Drug Discov. Today 22 (2017) 1680–1685, https://doi.org/10.1016/j.drudis.2017.08.010.
[9] Y. Tian, D. Rai, P. Zhan, C. Pannecouque, J. Balzarini, E. De Clercq, H. Liu, X. Liu, Design, synthesis, and biological evaluation of novel 3,5-disubstituted-1,2,6-thiadiazine-1,1-dione derivatives as HIV-1 NNRTIs, Chem. Biol. Drug Des. 82 (2013) 384–393, https://doi.org/10.1111/cbdd.12160.
[10] A. Akinleye, Y. Chen, N. Mukhi, Y. Song, D. Liu, Ibrutinib and novel BTK inhibitors in clinical development, J. Hematol. Oncol. 6 (2013) 59, https://doi.org/10.1186/1756-8722-6-59.
[11] G.F. Mangiatordi, D. Trisciuzzi, D. Alberga, N. Denora, R.M. Iacobazzi, D. Gadaleta, M. Catto, O. Nicolotti, Novel chemotypes targeting tubulin at the colchicine binding site and unbiasing P-glycoprotein, Eur. J. Med. Chem. 139 (2017) 792–803, https://doi.org/10.1016/j.ejmech.2017.07.037.
[12] O. Nicolotti, M. Catto, I. Giangreco, M. Barletta, F. Leonetti, A. Stefanachi, L. Pisani, S. Cellamare, P. Tortorella, F. Loiodice, A. Carotti, Design, synthesis and biological evaluation of 5-hydroxy, 5-substituted-pyrimidine-2,4,6-triones as potent inhibitors of gelatinases MMP-2 and MMP-9, Eur. J. Med. Chem. 58 (2012) 368–376, https://doi.org/10.1016/j.ejmech.2012.09.036.
[13] S. Simeon, N. Anuwongcharoen, W. Shoombuatong, A.A. Malik, V. Prachayasittikul, J.E.S. Wikberg, C. Nantasenamat, Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking, PeerJ 2016 (2016), https://doi.org/10.7717/PEERJ.2322.
[14] S. Simeon, R. Möller, D. Almgren, H. Li, C. Phanus-umporn, V. Prachayasittikul, L. Bülow, C. Nantasenamat, Unraveling the origin of splice switching activity of hemoglobin β-globin gene modulators via QSAR modeling, Chemometr. Intell. Lab. Syst. 151 (2016) 51–60, https://doi.org/10.1016/j.chemolab.2015.12.002.
[15] C.W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (2011) 1466–1474, https://doi.org/10.1002/jcc.21707.
[16] S. Simeon, N. Anuwongcharoen, W. Shoombuatong, A.A. Malik, V. Prachayasittikul, J.E.S. Wikberg, C. Nantasenamat, Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking, PeerJ 2016 (2016), https://doi.org/10.7717/PEERJ.2322.
[17] T. Srisongkram, P. Khamtang, N. Weerapreeyakul, Prediction of KRASG12C inhibitors using conjoint fingerprint and machine learning-based QSAR models, J. Mol. Graph. Model. 122 (2023), https://doi.org/10.1016/j.jmgm.2023.108466.
[18] G. Sun, T. Fan, X. Sun, Y. Hao, X. Cui, L. Zhao, T. Ren, Y. Zhou, R. Zhong, Y. Peng, In silico prediction of O6-methylguanine-DNA methyltransferase inhibitory potency of base analogs with QSAR and machine learning methods, Molecules 23 (2018), https://doi.org/10.3390/molecules23112892.
[19] S. Kumar, R. Bhowmik, J.M. Oh, M.A. Abdelgawad, M.M. Ghoneim, R.H. Al-Serwi, H. Kim, B. Mathew, Machine learning driven web-based app platform for the discovery of monoamine oxidase B inhibitors, Sci. Rep. 14 (2024), https://doi.org/10.1038/s41598-024-55628-y.
[20] E.E. Bolton, J. Chen, S. Kim, L. Han, S. He, W. Shi, V. Simonyan, Y. Sun, P. A. Thiessen, J. Wang, B. Yu, J. Zhang, S.H. Bryant, PubChem3D: a new resource for scientists, J. Cheminf. 3 (2011) 32, https://doi.org/10.1186/1758-2946-3-32.
[21] National Library of Medicine(PubMed), PubChem Substructure Fingerprint, n.d. http://pubchem.
[22] E.E. Bolton, J. Chen, S. Kim, L. Han, S. He, W. Shi, V. Simonyan, Y. Sun, P. A. Thiessen, J. Wang, B. Yu, J. Zhang, S.H. Bryant, PubChem3D: a new resource for scientists, J. Cheminf. 3 (2011) 32, https://doi.org/10.1186/1758-2946-3-32.
[23] Z.M. Hira, D.F. Gillies, A review of feature selection and feature Extraction methods applied on Microarray data, Adv Bioinformatics 2015 (2015) 198363, https://doi.org/10.1155/2015/198363.
[24] R.J. May, H.R. Maier, G.C. Dandy, Data splitting for artificial neural networks using SOM-based stratified sampling, Neural Network. 23 (2010) 283–294, https://doi.org/10.1016/j.neunet.2009.11.009.
[25] D. Chicco, M.J. Warrens, G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, PeerJ Comput Sci 7 (2021) e623, https://doi.org/10.7717/peerj-cs.623.
[26] W. Wei, J. Jiang, H. Liang, L. Gao, B. Liang, J. Huang, N. Zang, Y. Liao, J. Yu, J. Lai, F. Qin, J. Su, L. Ye, H. Chen, Application of a combined model with Autoregressive integrated Moving average (ARIMA) and generalized regression neural Network (GRNN) in Forecasting hepatitis Incidence in heng county, China, PLoS One 11 (2016) e0156768, https://doi.org/10.1371/journal.pone.0156768.
[27] K. Futagami, Y. Fukazawa, N. Kapoor, T. Kito, Pairwise acquisition prediction with SHAP value interpretation, The Journal of Finance and Data Science 7 (2021) 22–44, https://doi.org/10.1016/j.jfds.2021.02.001.
[28] Y. Meng, N. Yang, Z. Qian, G. Zhang, What makes an online review more helpful: an interpretation framework using xgboost and shap values, Journal of Theoretical and Applied Electronic Commerce Research 16 (2021) 466–490, https://doi.org/10.3390/jtaer16030029.
[29] B.I. Tingle, K.G. Tang, M. Castanon, J.J. Gutierrez, M. Khurelbaatar, C. Dandarchuluun, Y.S. Moroz, J.J. Irwin, ZINC-22—A free multi-billion-scale database of tangible compounds for ligand discovery, J. Chem. Inf. Model. 63 (2023) 1166–1176, https://doi.org/10.1021/acs.jcim.2c01253.
[30] S. Forli, R. Huey, M.E. Pique, M.F. Sanner, D.S. Goodsell, A.J. Olson, Computational protein-ligand docking and virtual drug screening with the AutoDock suite, Nat. Protoc. 11 (2016) 905–919, https://doi.org/10.1038/nprot.2016.051.
[31] D.S. Goodsell, M.F. Sanner, A.J. Olson, S. Forli, The AutoDock suite at 30, Protein Sci. 30 (2021) 31–43, https://doi.org/10.1002/pro.3934.
[32] M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li, R. Wang, Comparative assessment of scoring functions: the CASF-2016 update, J. Chem. Inf. Model. 59 (2019) 895–913, https://doi.org/10.1021/acs.jcim.8b00545.
[33] J. Eberhardt, D. Santos-Martins, A.F. Tillack, S. Forli, AutoDock vina 1.2.0: new docking methods, expanded force field, and Python bindings, J. Chem. Inf. Model. 61 (2021) 3891–3898, https://doi.org/10.1021/acs.jcim.1c00203.
[34] J. Eberhardt, D. Santos-Martins, A.F. Tillack, S. Forli, AutoDock vina 1.2.0: new docking methods, expanded force field, and Python bindings, J. Chem. Inf. Model. 61 (2021) 3891–3898, https://doi.org/10.1021/acs.jcim.1c00203.
[35] Y. Liang, Y. Cao, K. Tanaka, S. Thibivilliers, J. Wan, J. Choi, C.H. Kang, J. Qiu, G. Stacey, Nonlegumes respond to rhizobial nod factors by suppressing the innate immune response, Science 341 (1979) (2013) 1384–1387, https://doi.org/10.1126/science.1242736.
[36] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I. N. Shindyalov, P.E. Bourne, The protein data bank. http://www.rcsb.org/pdb/status.html, 2000.
[37] P.V. Escribá, X. Busquets, J.I. Inokuchi, G. Balogh, Z. Török, I. Horváth, J. L. Harwood, L. Vígh, Membrane lipid therapy: modulation of the cell membrane composition and structure as a molecular base for drug discovery and new disease treatment, Prog. Lipid Res. 59 (2015) 38–53, https://doi.org/10.1016/j.plipres.2015.04.003.
[38] S. Jo, T. Kim, V.G. Iyer, W. Im, CHARMM-GUI: a web-based graphical user interface for CHARMM, J. Comput. Chem. 29 (2008) 1859–1865, https://doi.org/10.1002/jcc.20945.
[39] S. Páll, A. Zhmurov, P. Bauer, M. Abraham, M. Lundborg, A. Gray, B. Hess, E. Lindahl, Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS, J. Chem. Phys. 153 (2020), https://doi.org/10.1063/5.0018516.
[40] J. Lemkul, From proteins to perturbed Hamiltonians: a suite of tutorials for the GROMACS-2018 molecular simulation package [article v1.0], living J, Comput Mol Sci 1 (2019), https://doi.org/10.33011/livecoms.1.1.5068.
[41] A. de F. Cobre, M. Maia Neto, E.B. de Melo, M.M. Fachi, L.M. Ferreira, F.S. Tonin, R. Pontarolo, Naringenin-4'-glucuronide as a new drug candidate against the COVID-19 Omicron variant: a study based on molecular docking, molecular dynamics, MM/PBSA and MM/GBSA, J. Biomol. Struct. Dyn. (2023) 1–14, https://doi.org/10.1080/07391102.2023.2229446, 0.
[42] C.W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (2011) 1466–1474, https://doi.org/10.1002/jcc.21707.
[43] T. Lindner, J. Puck, A. Verbeke, Beyond addressing multicollinearity: robust quantitative analysis and machine learning in international business research, J. Int. Bus. Stud. 53 (2022) 1307–1314, https://doi.org/10.1057/s41267-022-00549-z.
[44] A. Senawi, H.L. Wei, S.A. Billings, A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking, Pattern Recogn. 67 (2017) 47–61, https://doi.org/10.1016/j.patcog.2017.01.026.
[45] S. Kwon, H. Bae, J. Jo, S. Yoon, Comprehensive ensemble in QSAR prediction for drug discovery, BMC Bioinf. 20 (2019) 521, https://doi.org/10.1186/s12859-019-3135-4.
[46] Y. Zhang, Y. Wang, W. Zhou, Y. Fan, J. Zhao, L. Zhu, S. Lu, T. Lu, Y. Chen, H. Liu, A combined drug discovery strategy based on machine learning and molecular docking, Chem. Biol. Drug Des. 93 (2019) 685–699, https://doi.org/10.1111/cbdd.13494.
[47] Y. Zhou, S. Li, Y. Zhao, M. Guo, Y. Liu, M. Li, Z. Wen, Quantitative structure-activity relationship (QSAR) model for the severity prediction of drug-induced rhabdomyolysis by using random forest, Chem. Res. Toxicol. 34 (2021) 514–521, https://doi.org/10.1021/acs.chemrestox.0c00347.
[48] B. Ma, F. Meng, G. Yan, H. Yan, B. Chai, F. Song, Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data, Comput. Biol. Med. 121 (2020) 103761, https://doi.org/10.1016/j.compbiomed.2020.103761.
[49] F. Alghanim, I. Al-Hurani, H. Qattous, A. Al-Refai, O. Batiha, A. Alkhateeb, S. Ikki, Machine learning model for multiomics biomarkers identification for menopause status in breast cancer, Algorithms 17 (2023) 13, https://doi.org/10.3390/a17010013.
[50] D. Che, Q. Liu, K. Rasheed, X. Tao, Decision tree and ensemble learning algorithms with their applications in bioinformatics, Adv. Exp. Med. Biol. 696 (2011) 191–199, https://doi.org/10.1007/978-1-4419-7046-6_19.
[51] M. Hosni, I. Abnane, A. Idri, J.M. Carrillo de Gea, J.L. Fernández Alemán, Reviewing ensemble classification methods in breast cancer, Comput. Methods Progr. Biomed. 177 (2019) 89–112, https://doi.org/10.1016/j.cmpb.2019.05.019.
[52] Y. Ding, M. Chen, C. Guo, P. Zhang, J. Wang, Molecular fingerprint-based machine learning assisted QSAR model development for prediction of ionic liquid properties, J. Mol. Liq. 326 (2021), https://doi.org/10.1016/j.molliq.2020.115212.
[53] A. Wojtuch, R. Jankowski, S. Podlewska, How can SHAP values help to shape metabolic stability of chemical compounds? J. Cheminf. 13 (2021) https://doi.org/10.1186/s13321-021-00542-y.
[54] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (2019) 206–215, https://doi.org/10.1038/s42256-019-0048-x.
[55] S.M. Lundberg, B. Nair, M.S. Vavilala, M. Horibe, M.J. Eisses, T. Adams, D. E. Liston, D.K.W. Low, S.F. Newman, J. Kim, S.I. Lee, Explainable machine-learning

predictions for the prevention of hypoxaemia during surgery, Nat. Biomed. Eng. 2 (2018) 749–760, https://doi.org/10.1038/s41551-018-0304-0.
[56] S. Banerjee, Quantitative structural assessments of potential meprin β inhibitors by non-linear QSAR approaches and validation by binding mode of interaction analysis, New J. Chem. 47 (2023) 7051–7069, https://doi.org/10.1039/D2NJ04753E.
[57] A.A. Malik, C. Phanus-Umporn, N. Schaduangrat, W. Shoombuatong, C. Isarankura-Na-Ayudhya, C. Nantasenamat, HCVpred: a web server for predicting the bioactivity of hepatitis C virus NS5B inhibitors, J. Comput. Chem. 41 (2020) 1820–1834, https://doi.org/10.1002/jcc.26223.
[58] M. Yang, J. Chen, X. Shi, L. Xu, Z. Xi, L. You, R. An, X. Wang, Development of in silico models for predicting P-glycoprotein inhibitors based on a two-step approach for feature selection and its application to Chinese herbal medicine screening, Mol. Pharm. 12 (2015) 3691–3713, https://doi.org/10.1021/acs.molpharmaceut.5b00465.
[59] N. Schaduangrat, A.A. Malik, C. Nantasenamat, ERpred: a web server for the prediction of subtype-specific estrogen receptor antagonists, PeerJ 9 (2021), https://doi.org/10.7717/peerj.11716.
[60] H. Zhang, S. Rodriguez, L. Wang, S. Wang, H. Serezani, R. Kapur, A.A. Cardoso, N. Carlesso, Stem cell reports, Stem Cell Rep. 6 (2016) 940–956.
[61] C.K. Yoo, M. Shahlaei, The applications of PCA in QSAR studies: a case study on CCR5 antagonists, Chem. Biol. Drug Des. 91 (2018) 137–152, https://doi.org/10.1111/cbdd.13064.