Contents lists available at GrowingScience

Current Chemistry Letters

homepage: www.GrowingScience.com

The use of combined machine learning and in-silico molecular approaches for the study and the prediction of anti-HIV activity

Mohamed Ouabane[a,b], Zouhir Dichane[c], Marwa Alaqarbeh[d], Radwan Alnajjar[e,f], Chakib Sekkate[b], Tahar Lakhlifi[a] and Mohammed Bouachrine[a*]

[a]*Molecular Chemistry and Natural Substances Laboratory, Department of Chemistry, Faculty of Science, Moulay Ismail University, BP 11201, Meknes, Morocco*
[b]*Chemistry-Biology Applied to the Environment URL CNRT 13, Department of Chemistry, Faculty of Science, Moulay Ismail University, BP 11201, Meknes, Morocco*
[c]*Water Sciences and Environmental Engineering Team, Department of Geology, Faculty of Sciences, Moulay Ismail University, BP 11201, Meknes, Morocco*
[d]*Basic Science Department, Prince Al Hussein Bin Abdullah II Academy for Civil Protection, Al-Balqa Applied University, Al-Salt 19117, Jordan*
[e]*Department of Chemistry, Faculty of Science, University of Benghazi, Benghazi, Libya*
[f]*PharmD, Faculty of Pharmacy, Libyan International Medical University, Benghazi, Libya*

C H R O N I C L E

Article history:
Received January 25, 2024
Received in revised form
March 30, 2024
Accepted June 27, 2024
Available online
June 27, 2024

Keywords:
Anti-HIV
Machine Learning
QSAR
Docking
MD Simulation

A B S T R A C T

While the number of AIDS-related deaths continues to rise, efforts have been made to transform the disease into a manageable chronic condition. HIV protease inhibitors have become central to combination therapy. As a result, these inhibitors have become a major focus of anti-HIV drug development. This research takes a data-driven approach to drug development through the use of quantitative structure-activity relationship (QSAR) analysis. A dataset of 450 anti-HIV drugs was used to construct and validate models. Using extensive validation methods and various machine learning algorithms, the results clearly showed that the "ET" regression outperformed the other models ("XGB", "LGBM", "DT", "RF", "GB", "Bag", and "HGB") in terms of goodness of fit, predictivity, generalizability, and model robustness. Promising compounds were subjected to molecular docking and molecular dynamics simulation, resulting in drugs with favourable pharmacokinetic and pharmacodynamic properties that consistently interact with the therapeutic target.

1. Introduction

Acquired immunodeficiency syndrome (AIDS) continues to be a lethal illness that progresses inexorably[1]. There is currently no complete and efficient chemotherapy[2]. The Human Immunodeficiency Virus (HIV) is the retrovirus responsible for the illness[3]. This virus destroys the body's natural ability to defend itself, weakening the immune system and fostering the growth of several infectious illnesses (opportunistic infections) with a higher risk of fatality[4]. Different neoplasms are thought to develop through the same process[5]. HIV also targets the nerve system, which can have serious psychological and neurological effects[6]. The probability of containing the disease decreases after the acceleration of viral replication[7].

HIV infection development is associated with both infectious complications (sometimes known as opportunistic infections) and non-infectious complexity[8]. The advent of Highly Active Antiretroviral Therapy (HAART) has underlined the relevance of non-infectious difficulties encountered by people living with HIV[8], even while proper prophylaxis has improved the management of opportunistic infections[9].

* Corresponding author
E-mail address m.bouachrine@umi.ac.ma (M. Bouachrine)

doi: 10.5267/j.ccl.2024.6.004

The illness brought on by HIV is no longer referred to as “AIDS”; instead, the phrase “AIDS stage” is used to denote a crucial stage of immune system degeneration in the patient. Based on certain opportunistic infections known as “AIDS-defining illnesses” (such as Kaposi's sarcoma and oropharyngeal candidiasis), an AIDS-stage diagnosis can be made[9].

As of 2021, no experimental vaccine was shown to be successful despite enormous efforts in the development of an HIV vaccine[10]. The management of comorbidities associated with HIV infection, particularly neurocognitive disorders and cardiovascular diseases (CVD), which now accompany this chronic condition, is one of the emerging challenges due to the virus' continued persistence in human populations and the rising life expectancy of HIV-positive patients[11]. The comprehension of HIV virus and its molecular repercussions, as well as the structure-activity connections among diverse types of viral inhibitors, are both highly influenced by these studies[12].

Quantitative Structure-Activity Relationship (QSAR) tries to identify connections between chemical structure and biological or other activities by creating a QSAR model[13]. With this method, it is feasible to forecast the functions of recently proposed compounds before deciding whether to synthesize and test them[14]. The determination of theoretical parameters for the target compounds is the first step in building a QSAR model[15].

When developing a model, experimental data connected to biological characteristics are regarded as dependent variables[16]. In the QSAR study, many descriptors may be produced, but only a portion of them are statistically significant in terms of their association with the target biological activity for the analysis at hand[17]. The difficult part is deciding which selection of descriptors best captures the essential structural and physicochemical characteristics connected to anti-HIV inhibitory action[18]. The QSAR modeling procedure includes the effective selection of descriptors or variables[19]. The quality of biological data, descriptor choice, and statistical techniques are only a few of the variables that affect the development of a high-quality QSAR model[17]. There are many different approaches to choosing variables; the most well-liked ones are stepwise regression, neural networks, fuzzy logic, and evolutionary algorithms[20]. Decision Trees (DT) were suggested as a variable selection method in this study[21]. Researchers in a variety of domains have become interested in DT algorithms as a unique computational technique[22].

Machine learning-based modeling, which excels at resolving prediction and data analysis problems, is vital for developing a trustworthy QSAR model[23]. XGBoost, Random Forest, Gradient Boosting, Bagging, Extra Trees, and Histogram-Based are important algorithms among the many others that are available[24]. While Random Forest delivers great predicted accuracy and durability against overfitting[24], XGBoost stands out for its speed and skill in managing enormous volumes of complex data[25]. For its capacity to simulate intricate interactions between variables, gradient boosting is highly regarded. To lower variance and improve model stability, bagging is used, and Extra Trees works well for high-dimensional datasets. As a strong alternative to the Gradient Boosting framework, Histogram-Based Gradient Boosting uses histograms to describe continuous information. The use of these many methods enables machine learning professionals to investigate and choose the best solution for their unique issue while taking performance, interpretability, and the features of the available data into account.

2. Materials and Methods

2.1. Dataset

Numerical data on the activity of HIV-1 protease inhibitors expressed as pIC_{50} values, were obtained from the literature[26]. The total number of compounds investigated in this study is 450. It should be noted that the SMILES strings were converted into 3D structures using RDKit[25].

To perform a lazy regression[24] on the chemical data, the following steps were taken:

1. Import the necessary libraries: pandas, rdkit, Mordred, and sklearn.
2. Use the pd.read_csv function to read the CSV file and create a pandas Data Frame.
3. Specify Data Frame column names.
4. Convert SMILES strings into 3D structures.
5. Add hydrogens to molecules using RDKit's Chem. AddHs function.
6. Generate multiple conformers for each molecule using RDKit's AllChem. Embed Multiple Confs function.
7. Calculate molecular descriptors using Mordred and RDKit. Obtain the list of available descriptors, initialize a dictionary to store descriptor values, and calculate descriptors for each compound using a for loop and the getter function.
8. Merge the descriptor results with the original Data Frame.
9. Filter descriptors using principal component analysis.
10. Divide the dataset into training and test sets.
11. Apply automatic model predictions (“XGB”, “LGBM”, “DT”, “RF”, “GB”, “Bag”, “ET”, and “HGB”).
12. Evaluate model performance using appropriate performance measures.
13. Examine results.

14. Use the predictive model to make predictions on new data.

It should be emphasized that lazy regression gives a rough estimate of how well various models perform without calling for fine-tuning or a particular model choice[24]. However, additional model investigation, hyperparameter modification, and more thorough assessment may be required to get higher results.

2.2. Machine Learning Models

2.2.1. XGBoost (XGB)

Extreme Gradient Boosting, or XGB, is a potent machine learning technique applied in various domains, including QSAR modeling. A more regularized model is used in this boosting method to reduce overfitting and promote generalization. The great performance and efficiency of XGB make it a preferred option for huge datasets.

In QSAR modeling, XGB has proven its capacity to precisely predict the activity of different chemicals. Studies have demonstrated, for instance, that XGB works better than other machine learning algorithms like Random Forests and Support Vector Machines in predicting the inhibitory action of certain substances[27].

Carlos Guestrin and Tianqi Chen first introduced XGB in 2011[27]. It is a framework for machine learning that uses boosting tree models. As opposed to conventional Boosting tree models, XGB uses a second-order Taylor expansion on the loss function, enabling it to take into account more complex data during tree training[28].

The numerous strategies used by XGB help to reduce overfitting. For instance, it prevents learning of certain situations if all sample weights in the leaf nodes are below a predetermined level. Additionally, random characteristics are chosen from samples throughout the creation of each tree, strengthening the generalizability of the model and enhancing its performance in real-world applications.

To maximize the performance of the model while using XGB, a few parameters must be modified[28]. These parameters include the depth of the tree (max_depth), the number of iterations (n_estimators), the sum of the weights of the smallest leaf nodes (min_child_weight), the subsampling rate of training samples (subsample), the subsampling rate of features during the construction of each tree (colsample_bytree), and the learning rate.

In finality, XGB is a well-liked and effective machine learning technique, especially in QSAR modeling. To improve generality and avoid overfitting, it uses regularized Boosting tree models. XGB offers great prediction performance across a range of applications because of its sophisticated methodologies and careful parameter selection[29].

2.2.2. LightGBM (LGBM)

Microsoft created LightGBM (LGBM), a gradient-boosting framework that makes use of a tree-based learning method. This model can handle huge datasets with millions of samples and thousands of features since it is quick, effective, and scalable[30].

The capacity of LGBM to handle high-dimensional data and its quick learning time makes it a popular choice for QSAR modeling. LGBM was used, for example, to develop a QSAR model to predict the binding affinity of tiny molecules to human serum albumin. According to the researchers, LGBM outperformed other models regarding prediction efficiency and accuracy, including random forests and support vector machines[31].

XGB and pGBRT are two effective implementations of the well-known machine learning technique gradient-boosting decision tree (GBDT)[30]. Although these solutions incorporate several technological refinements, efficiency, and scalability are still inadequate when dealing with high-dimensional features and huge datasets. One major factor is that it might take much time to traverse all data instances for each characteristic to assess the information benefit of every potential split point. Two innovative methods are suggested as a solution to this issue: gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB)[32].

In GOSS, a sizeable portion of data instances with minor gradients are disregarded, and the information gain is only estimated for the remaining cases. GOSS delivers accurate information gain estimates with a significantly lower amount of data because cases with greater gradients are more crucial to the computation of information gain. EFB, on the other hand, reduces the number of features by grouping mutually exclusive features (those that seldom take non-zero values concurrently)[33]. It has been demonstrated that determining the best way to combine exclusive features is an NP-hard issue, but a greedy approach can offer a good approximation ratio, allowing for a reduction in the number of features while maintaining split point precision[33].

The combination of both methods is known as LightGBM, which outperforms the standard GBDT learning process by a factor of more than 20 while retaining essentially the same accuracy. Experiments on various open datasets are used to validate this[33].

2.2.3. Decision Tree (DT)

Because the number of pathways inside DTs is often substantially fewer than the total number of features, decision trees (DTs) are frequently viewed as interpretable machine learning models[34]. This study shows that, in some situations, it might be difficult to accept DTs as interpretable since their pathways can be arbitrarily longer than those of a Partial Instance (PI) explanation, which is a set of feature values that can be used to make a prediction with the fewest possible features. In order to compute PI explanations of DTs, this paper offers a novel model that enables PI explanations to be computed in polynomial time[35].

In addition, it is demonstrated that counting PI explanations may be simplified to counting minimum hitting sets[35]. Utilizing well-known DT learning technologies, experimental results have been attained over a wide variety of publicly accessible datasets. These findings demonstrate that DTs typically contain supersets of PI explanations that are appropriate supersets.

This study looks at the boundaries of interpretability in Decision Trees (DTs), demonstrating that, even for irreducible DTs, certain routes inside a DT may include many literals that are unimportant to an explanation. The paper also presents a linear-time test to see if a path in a DT contains unnecessary literals and uses this test to create a polynomial-time method for constructing a PI explanation of a DT. The study also shows that counting minimum hitting sets and counting PI explanations of DTs are related. Induced routes inside DTs can certainly contain irrelevant literals, even when the DT is irreducible, according to experimental results using publicly accessible datasets and cutting-edge DT learning methods. The suggested algorithms' runtime is low or on par with the time needed for tree learning[36].

Due to its simplicity, interpretability, and ability to handle non-linear relationships between input and output variables, the Decision Tree (DT) is a widely used machine learning algorithm in QSAR modeling. The DT works by recursively partitioning data into subsets based on input feature values until a stopping criterion is met and then assigning an output value to each subset[34]. DTs have been applied to a wide range of QSAR applications, including the prediction of compound toxicity, solubility, and bioactivity. Several studies have demonstrated the successful use of DTs in QSAR modeling. For example, DTs were employed to predict the oral acute toxicity of pesticides with a robust accuracy of 86%[36].

2.2.4. Random Forest (RF)

A modeling strategy that makes use of decision tree ensembles is called random forests. Each tree in the forest is separately built using values from a randomly selected vector, and they all have the same distribution. As the number of trees in the forest increases, the generalization error of Random Forests converges to a limit[37]. The strength of each individual tree and how they are correlated with one another determines the generalization error of a forest. Each node of the trees is divided randomly to improve performance, resulting in error rates that are equivalent to or even better than those attained with the AdaBoost method while being more noise resistant. The impact of increasing the number of features utilized in the splits is evaluated using internal estimates to track inaccuracy, strength, and correlation. Additionally, these internal estimates enable assessing the significance of variables employed in the modeling procedure[38].

In the area of QSAR modeling, the Random Forest (RF) technique is used as an excellent example. In this application, RF is used to forecast a molecule's activity based on its structural characteristics. Multiple decision trees are built by RF utilizing randomly chosen subsets of data, and their predictions are then combined to get a conclusion. Numerous research studies have demonstrated how well RF performs in QSAR simulation[38].

As an illustration, RF was used to predict the inhibitory action of quinoline derivatives against human acetylcholinesterase in their study[39]. The RF model beat conventional machine learning methods, including support vector machines and artificial neural networks, according to the authors. RF in a different study to forecast the toxicity of ionic liquids based on their chemical characteristics. The authors discovered that RF demonstrated high prediction accuracy and, using this method, determined the most significant toxicity characteristics. As a result, RF seems to be a useful technique for QSAR modeling and other related activities[40].

2.2.5. Gradient Boosting (GB)

In QSAR modeling for drug discovery and material design, Gradient Boosting (GB) is a popular machine-learning approach. The approach creates a few decision trees to forecast the target variable by iteratively fitting new models to the residuals of earlier models. In this way, the algorithm improves upon past errors and seeks to lower prediction errors with each iteration[41].

Gradient Boosting (GB) is a preferred QSAR method due to its computational efficiency and precision. A gradient-boosting model consists of a series of decision trees. The sum of the predictions of the decision trees on a molecule is the model prediction. The trees are added to the model through iterations so that the current model's errors are used to build a new tree, requiring the new tree to learn information that the current model has not extracted. As a result, we expect that the trees learned earlier in the sequence contribute more to overall prediction than those learned later in the sequence[42].

2.2.6. Bagging (Bag)

A machine learning ensemble approach called bagging (Bootstrap Aggregating) combines many models to improve prediction accuracy and minimize variance[43]. By bootstrapping the original dataset, the technique creates many training sets, and each set is used to train a different model. After then, the combined forecasts from all the models are used to establish the final projection. Bagging is frequently employed in QSAR modeling to increase forecast stability and accuracy[44].

In a work, bagging was used to build a QSAR model for predicting the acute toxicity of organic chemicals[44]. To improve prediction accuracy over the use of each method alone, the authors integrated bagging, random forest, and support vector machines. In research, bagging was also used to create a QSAR model for predicting the carcinogenicity of chemical[43]. By combining bagging with partial least squares regression, the authors improved prediction accuracy and resilience.

2.2.7. Extra Trees (ET)

A particular ensemble learning approach for decision trees is called Extra Trees (ET). Although similar to random forests, the way the trees are built differs significantly. To make the final forecast, Extra Trees creates a lot of random decision trees and combines their guesses. Due to its capacity for handling highly dimensional data and resilience against noise and overfitting, ET is frequently employed in QSAR modeling. Studies have successfully used ET in QSAR modeling in a number of cases[45]. In 2013 research by Lusci et al.[46], ET was used to predict the activity of drugs against acetylcholinesterase, and it performed better than other machine learning techniques, including support vector machines and random forests. The authors asserted that ET had superior predictive performance compared to other machine learning techniques like random forests and support vector machines. A QSAR model for predicting the activity of anti-hepatitis peptides was created using the Extra Trees regressor as one of the machine learning techniques[47].

As a sophisticated machine learning approach for QSAR modeling, Extra Trees can handle high-dimensional data and is comparatively resistant to overfitting. It is a popular option for creating QSAR models since several research have shown it successful.

2.2.8. Histogram-Based Gradient Boosting (HGB)

In order to forecast the biological activities of substances, QSAR modeling employs the machine learning algorithm HGB. It is a development of the widely used regression and classification technique known as the gradient boosting decision trees (GBDT). Numerous research has demonstrated the usefulness of HGB for QSAR modeling. For instance, predicted the binding affinity of substances to the alpha estrogen receptor using HGB[48].

The results show that HGB outperforms various machine learning techniques, such as support vector regression and random forests. HGB in a different investigation to forecast the inhibitory efficacy of drugs against cholinesterase, a target for treating Alzheimer's disease Regarding prediction accuracy and computational effectiveness, the scientists discovered that HGB performed better than a number of machine-learning techniques[49].

For QSAR modeling, HGB is a promising machine learning approach, especially when working with huge datasets and high-dimensional feature spaces. It is a desirable choice for researchers in the field because of its capacity to manage a large number of different feature values while preserving computing effectiveness LightGBM[50].

2.3. Validation Methodologies

The validation of derived models is a critical stage in molecular modeling research in order to evaluate the importance and prediction capacities of QSAR models for the activities of newly suggested compounds with a target of anti-HIV activity. Since these models are the consequence of predictive studies, it is crucial to evaluate and use them specifically in the context of reviewing these results. Any usage outside of this modeling context needs careful thought, and the more one veers from the framework, the more unclear it gets[51].

It is critical to define the limits of the model in detail in order to prevent mistakes during model validation and use. The stability and dependability of the model must be confirmed, together with its capability for internal-external prediction and the chemical space in which it may be used. Various statistical factors have been used in this work to validate prediction models, including:

2.3.1. Benchmark measures

The Root Mean Squared Error (RMSE), defined as the square root of the average of the residual sum of squares (RSS), is the standard number used to measure model quality in regression analysis:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n_{TR}}(y_i - \hat{y}_i)^2}{n_{TR}}} = \sqrt{\frac{RSS}{n_{TR}}}$$

where n_{TR} is the number of items in the training set (used to calibrate the regression model), y_i and $\hat{y}_i$ are the experimental and calculated responses for the i-th object, respectively[52]. Because the RMSE is scale-dependent, the standard measure of model quality that is independent of the response scale is the coefficient of determination (R^2), which represents the variance explained by the model as follows:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n_{TR}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{TR}}(y_i - \bar{y}_{TR})^2}$$

where $\bar{y}_{TR}$ is the average response of the training objects, while i y_i and $\hat{y}_i$ are the experimental and the calculated response of the i-th object, respectively[53].

2.3.2. Adjusted determination coefficient R^2_{adj}

The adjusted determination coefficient R^2_{adj} This coefficient is used in multiple regression because it takes the degree of freedom into account:

$$R^2_{adj} = \sqrt{\frac{R^2(n-1) - p}{n - p - 1}}$$

With: n is the number of observations (the molecules); p is the number of independent variables (the descriptors); and R^2 is the model's coefficient of determination[54].

2.3.3. Indicator Q^2 metrics

The analyzed Q^2 metrics are introduced briefly in this line. The first three metrics (Q^2_{F1} , Q^2_{F2} and Q^2_{F3}) were investigated and addressed in two of our prior articles, and they are as follows:

$$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n_{OUT}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{OUT}}(y_i - \bar{y}_{TR})^2} \qquad Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n_{OUT}}(y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^{n_{OUT}}(y_i - \bar{y}_{OUT})^2} \qquad Q^2_{F3} = 1 - \frac{\sum_{i=1}^{n_{OUT}}(y_i - \hat{y}_{i/i})^2/n_{OUT}}{\sum_{i=1}^{n_{TR}}(y_i - \bar{y}_{i/i})^2/n_{TR}}$$

where y_i is the experimental response of the i-th object, $\hat{y}_{i/i}$ is the predicted response when the i-th object is not in the training set, n_{TR} and n_{OUT} are the number of training and test objects, respectively; $\bar{y}_{TR}$ is the average value of the training set experimental responses, and $\bar{y}_{OUT}$ is the average value of experimental responses of external test objects. It should also be noted that, using conventional cross-validation approaches (e.g., leave-one-out), where each training object is utilized as a test object only once, $Q^2_{F1} = Q^2_{F2} = Q^2_{F3}$, where $n_{OUT} = n_{TR}$ and $\bar{y}_{TR} = \bar{y}_{OUT}$. It is worth noting that the measure Q^2_{F1} is the most commonly employed in QSAR modeling[55].

2.3.4. Fisher's F index

Fisher's F index, also known as Fisher's F test, is used to calculate the statistical significance of a model at a specified confidence level, which is commonly indicated as "x%" (the standard level being 95%). This evaluates the model's parameter selection quality. It is crucial to remember, however, that the conclusion does not indicate that the association has a “x%” chance of being true. Rather, it indicates that the association is confirmed for “x%” of the reference compounds used, while others are eliminated[56]. The observed F is the value to be estimated in this test, which is derived using the formula:

$$F = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \hat{y}_i)^2} \frac{n - p - 1}{n}$$

where F is Fisher’s index; y_i and y_i denote the observed and calculated values of the dependent variable, respectively; y_i is the mean of the predicted values; n denotes the number of observations (molecules); and p denotes the number of independent variables (descriptors). Following calculation of Fisher (observed), it is compared to theoretical Fisher obtained

from conventional statistical test (F test). If the observed F is larger than the theoretical F, the null hypothesis H0 is rejected, suggesting that the variances of the samples are too dissimilar to be deemed homogenous. If the observed *F* is less than the theoretical *F*, the null hypothesis H1 is accepted, which means that the two variances have values that are near enough for us to accept the assumption that they are homogeneous[57].

2.4. Pharmacodynamic and Pharmacokinetic Properties

In silico research aiming at predicting and assessing the pharmacokinetic characteristics and safety of prospective chemical compounds, ADME-Tox attributes (Absorption, Distribution, Metabolism, Excretion, and Toxicity) play a critical role. Based on the information presented previously about the SwissADME Web tool, it is clear that these filters are key components for quickly identifying compounds that may offer potential concerns during drug development[58].

ADME-Tox filters are used as pre-selection criteria to remove drugs with undesirable features such as limited bioavailability, poor absorption, excessive toxicity, or bad interactions with target proteins. ADME-Tox filters quickly identify compounds with poor drug development potential by employing prediction models and indicators generated using technologies such as SwissADME. For example, by examining factors like lipophilicity, solubility, skin permeability, and others, these filters serve to reject compounds that may have difficulty being absorbed into the body or penetrating the brain, potentially affecting their therapeutic efficacy[59].

Predicting interactions with proteins such as transporters and hepatic enzymes also allows for the prediction of possible dangers associated with metabolism and excretion. The importance of ADME-Tox filters stems from their capacity to quickly reject doubtful therapeutic candidates while concentrating resources on the most promising chemicals[60]. This saves drug development costs and time by removing low-potential compounds early on. As a result, using ADME-Tox filters in *in silico* research can considerably improve drug development efficiency by finding the most promising candidates while avoiding risks associated with toxicity and other pharmacokinetic concerns[61].

2.5. Molecular Docking

Ligand preparation was conducted using Chem3D, applying the MM2 method to minimize steric energy. The 3D structure of the target protein was downloaded from the Protein Data Bank (PDB) database, with a specific focus on the 3D structure of chain A (ID: 3OYA), carried out using Discovery Studio 2021[62]. The docking method was carried out utilizing the MOE 2014 (Molecular Operating Environment) program[63]. The energy of the investigated substances was reduced using the MMFF94x force field. The chain A structure was corrected by adding extra hydrogen atoms and allocating partial charges with Amber12: EHT. Following that, additional minimization was attempted using the same force field and an RMSD of 0.01 Å. To achieve a crystalline structure, water molecules, and co-crystallized inhibitors were removed, and hydrogen atoms were added. The MMFF94x force Field was used to regulate charges[64]. The final energies, generated poses, and scores were evaluated using the Rescoring1 procedure: London dG with a retention factor of 50 and Rescoring 2: GBVI/WSA dG with a retention factor of 1 by selecting the lowest affinity score[65].

2.6. Molecular Dynamics simulations

MD simulations were performed using the Desmond simulation package developed by Schrödinger LLC. The simulations used the NPT package, with a temperature of 300 K and a pressure of 1 bar applied consistently throughout the cycles. Each simulation lasted 150 ns with a relaxation time of 1 ps[66]. The OPLS3 force field parameters were used universally in all simulations to ensure accurate representation. Coulomb interactions were considered up to a cutoff radius of 20Å. The boundaries of the orthorhombic periodic box were placed at 10Å from the protein atoms. The three-point transferable intermolecular potential (TIP3P) model was used to describe the water molecules. In addition, a salt concentration of 0.15 M NaCl(aq) was incorporated into the system, which was created using Desmond's System Builder utility[67].

The Martyna-Tuckerman-Klein chain coupling scheme was used for pressure control with a coupling constant of 2.0 ps[68].Temperature control was accomplished using the Nose-Hoover chain coupling scheme. The RESPA integrator was used for unbound force calculations, with short-range forces updated at each step and long-range forces updated every three steps[69]. Trajectories were recorded at 20 ns intervals to facilitate later analysis. The Simulation Interaction Diagram tool in the Desmond MD software package was used to examine the behavior and interactions between the ligands and the protein. The stability of the MD simulations was checked by evaluating the root mean square deviation (RMSD) of ligand and protein atom positions over time[70].

3. Discussion and Results

3.1. Evaluating the performance of several ML algorithms

We calculated 208 widely used descriptors from the literature. The 208 descriptors are then filtered using a Principal Component Analysis (PCA) approach to remove non-informative descriptors and avoid information redundancy. Finally,

we had 63 descriptors with correlations greater than 0.4. The following **Table 1** shows the statistical validation settings for both internal and external validation.

Table 1. Performance of ensemble machine learning model.

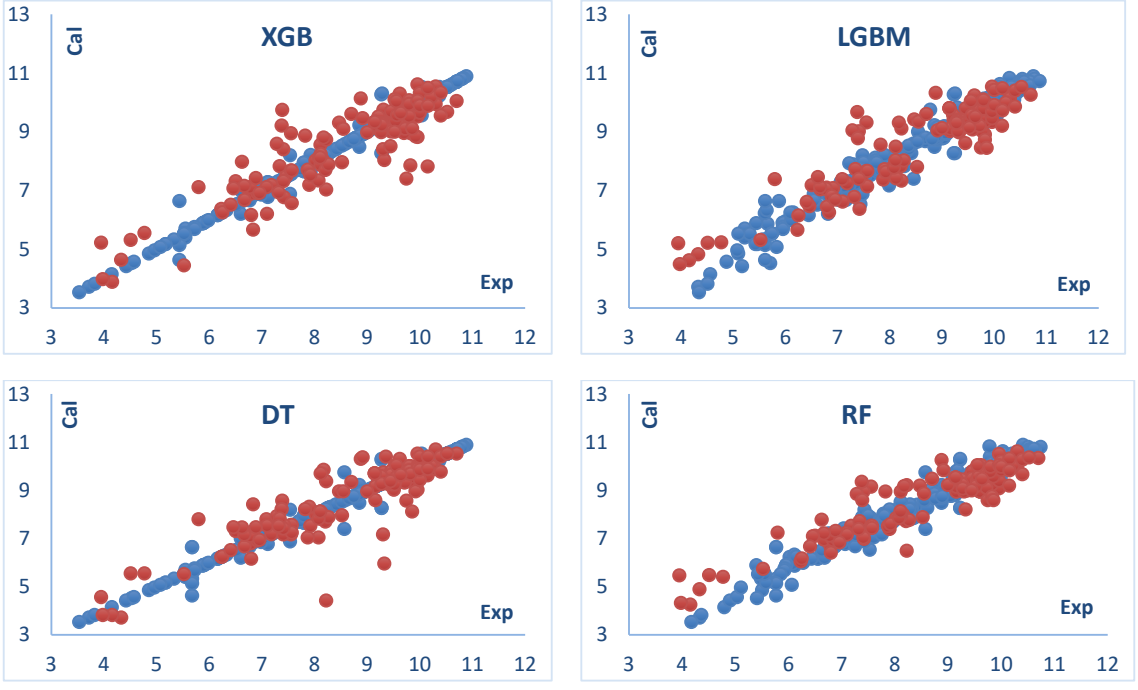
	Models	R^2	Q^2	R^2_{adj}	RMSE	F	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	k	k'	P-value
Training set	XGB	0.990	0.990	0.987	0.167	30640.152	0.985	0.99	0.995	0.995	0.99	$1.1.10^{-16}$
	LGBM	0.969	0.969	0.961	0.294	9938.251	0.968	0.964	0.974	0.985	0.969	$1.1.10^{-16}$
	DT	0.985	0.985	0.982	0.201	21011.978	0.981	0.986	0.989	0.993	0.985	$1.1.10^{-16}$
	RF	0.964	0.964	0.955	0.316	8905.184	0.962	0.962	0.966	0.983	0.964	$1.1.10^{-16}$
	GB	0.971	0.971	0.963	0.285	10761.708	0.969	0.971	0.971	0.986	0.971	$1.1.10^{-16}$
	Bag	0.952	0.952	0.94	0.364	6499.651	0.939	0.956	0.959	0.977	0.952	$1.1.10^{-16}$
	ET	0.985	0.985	0.982	0.201	21011.978	0.981	0.986	0.989	0.993	0.985	$1.1.10^{-16}$
	HGB	0.971	0.971	0.964	0.28	10929.46	0.969	0.968	0.977	0.986	0.971	$1.1.10^{-16}$
Test set	XGB	0.788	0.601	0.788	0.697	508.730	0.802	0.781	0.754	0.890	0.789	$1.1.10^{-16}$
	LGBM	0.836	0.690	0.836	0.614	683.240	0.797	0.852	0.821	0.915	0.837	$1.1.10^{-16}$
	DT	0.783	0.590	0.783	0.706	533.892	0.664	0.823	0.797	0.895	0.785	$1.1.10^{-16}$
	RF	0.849	0.716	0.849	0.588	755.202	0.793	0.858	0.864	0.922	0.850	$1.1.10^{-16}$
	GB	0.852	0.720	0.852	0.584	768.751	0.802	0.872	0.843	0.923	0.852	$1.1.10^{-16}$
	Bag	0.830	0.679	0.830	0.625	658.417	0.809	0.809	0.850	0.912	0.832	$1.1.10^{-16}$
	ET	0.859	0.733	0.859	0.569	822.527	0.854	0.853	0.849	0.928	0.860	$1.1.10^{-16}$
	HGB	0.835	0.689	0.835	0.615	682.053	0.797	0.850	0.822	0.915	0.836	$1.1.10^{-16}$

The results highlight the performance of the different models used for prediction, evaluated with measures such as R^2, Q^2, R^2_{adj}, RMSE, F, Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, k, k', and P-value for the training and test datasets.

For the training dataset, the ET model shows high for R^2, Q^2, and R^2_{adj}, suggesting excellent prediction of the data. The low RMSE value indicates that the model's predictions are accurate. The high values for F, Q^2_{F1}, Q^2_{F2}, and Q^2_{F3}, show that the ET model fits the data well and captures subtle variations in the data.

On the test dataset, the ET model maintains its performance with a high R^2, suggesting a significant generalization capability. The Q^2, R^2_{adj}, values confirm that the ET model maintains its ability to predict accurately even with additional data. The low value of RMSE indicates that the model's predictions remain close to the actual values.

If we compare the reliability of models such as XGB, LGBM, DT, RF, GB, Bag and HGB, with the ET model consistently outperforms in terms of R^2, Q^2, R^2_{adj}, and RMSE, both on the training set and on the test set. These results indicate that the ET model is the most reliable for predicting specific events of anti-HIV activity (**Fig. 1**).



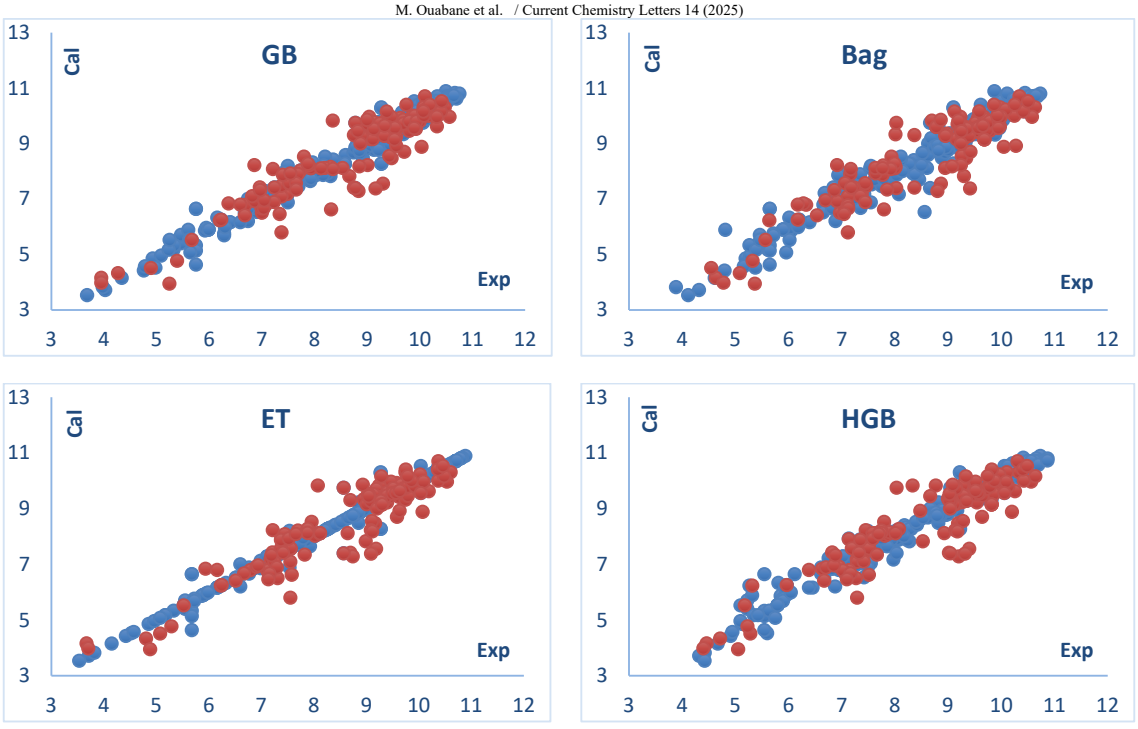
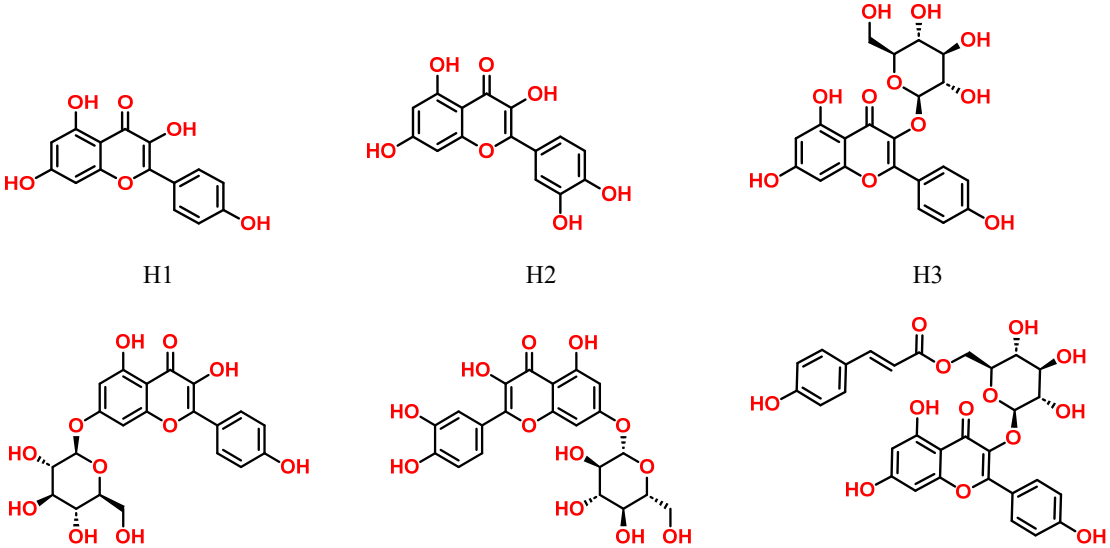


Fig. 1. Experimental and calculated pIC_{50} values for each prediction model

3.2. New inhibitors proposed

Naheed Mahmood et al.[71] investigated the anti-HIV activity of *Rosa damascena* flower extracts. These extracts have been shown to have mild antiviral activity against the virus. Among these compounds, Kaempferol (H1) has been shown to be effective in preventing the development of infectious viral progeny by selectively inhibiting viral protease. However, quercetin (H2) and two kaempferol derivatives were shown to prevent HIV infection by inhibiting the binding of gp120 to CD4. 2-phenylethanol-O-(6-O-galloyl)-b-D-glucopyranoside (compound H8), coupled irreversibly with gp120 and destroyed the virus's infectivity. Our findings show the anti-HIV efficacy of *Rosa damascena* extracts and isolated components. These compounds' various modes of action contribute to their overall efficacy against HIV infection, giving insights into prospective antiviral therapies based on natural products. For these reasons, we used prediction models to estimate the pIC_{50} values (**Table 2**) in order to assess the inhibitory concentration, ADME-Tox profile, interaction modes, and stability of these ligands as HIV inhibitors (**Fig. 2**).



H1 H2 H3

H4 H5 H6

H7 H8

19

Fig. 2. Modelling the structures of new anti-HIV inhibitors.

Table 2. Prediction of pIC50 values of new anti-HIV inhibitors using machine learning models

Ligands	XGB	LGBM	DT	RF	GB	Bag	ET	HGB
H1	6.224	5.365	6.250	5.227	6.035	5.038	6.519	5.444
H2	6.289	5.356	6.251	5.232	6.151	5.038	6.725	5.424
H3	6.945	6.902	7.409	7.385	7.602	6.515	7.858	6.714
H4	6.945	6.902	7.409	7.384	7.602	6.761	7.855	6.714
H5	8.208	7.449	7.409	7.788	8.100	7.126	8.101	7.062
H6	8.512	8.891	9.620	9.164	8.828	9.507	8.606	8.667
H7	5.076	5.230	6.250	5.413	5.631	5.001	6.142	4.991
H8	6.021	7.327	3.979	6.359	6.417	6.322	5.234	6.976
19	**10.221**	**10.171**	**10.222**	**10.036**	**10.100**	**9.882**	**10.222**	**10.137**

3.3. Validation of crystalline protein structure

Multiple criteria were used to assess the quality and validate the protein structure (3OYA.pdb)[53]. The average worldwide quality factor, 93.1689, represents the level of confidence in rejecting locations that exceed the error value. High-resolution structures often produce results of 95% or higher. However, the quality factor is around 91% for lower-resolution structures (between 2.5 and 3). An investigation of 118 structures with a resolution of 2 or higher and an R factor of less than 20% found that a good model should have more than 90% in the most favorable regions (**Fig. 3**).

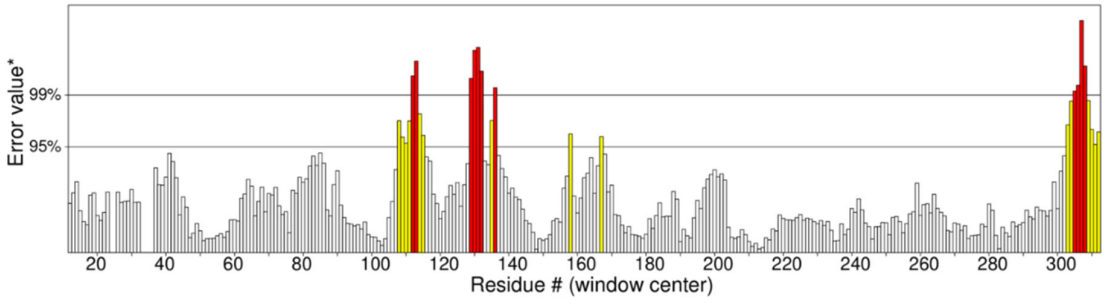


Fig. 3. Evaluation of the Quality and Structural Validation of the Protein (3OYA.pdb).

PROCHECK was used to do a structural evaluation of the molecule from the file "/var/www/SAVES/Jobs/1417031/saves.pdb," which had 588 residues. According to the Ramachandran plot (**Fig. 4**),

89.7% of the residues are in the “core” area, suggesting their preferred shape. Furthermore, 9.5% of the residues fall into the “allowed” category, 0.8% fall into the “generously allowed” category, and none fall into the “disallowed” category.

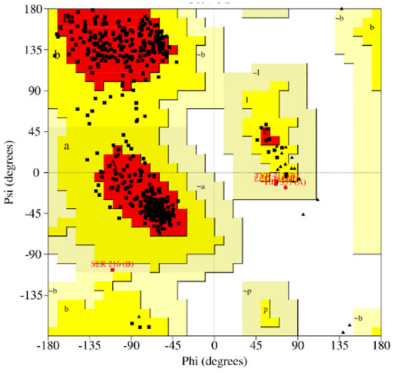


Fig. 4. PROCHECK Ramachandran diagrams showing various amino acid regions of the protein shaded according to the degree of favorability of high-resolution structures.

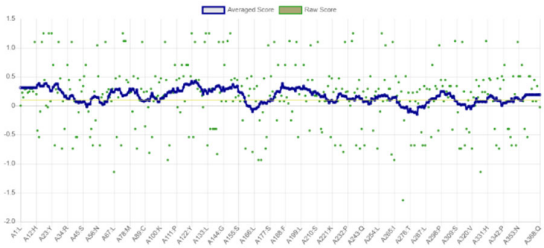


Fig. 5. Evaluation of the structural quality of the molecular entity and 3D-1D alignment

Detailed analysis revealed that only 9 of 548 residues have non-ideal Ramachandran conformations. Similarly, out of 313 residues, only 6 have atypical chi1-chi2 dihedral angle combinations. In terms of side-chain parameters, 5 residues have improved properties, while none are labeled as “inside” or “outside”. The largest outlier departure in terms of residue characteristics is 5.1, and 33 poor interactions have been discovered in the structure. Bond lengths and angles are all in the 3.2 range, and the structure is classified as 1-2-2. In the structure, three cis-peptides have been found.

The G factors, which indicate the dependability of energy terms, have been determined. The G factor of dihedrals is 0.28, the G factor of covalent terms is 0.45, and the total G factor is 0.34, suggesting a relatively well-optimized structure. Planar group analysis reveals that 100% of these groups are within permissible limits, with no observed aberrations.

In summary, the PROCHECK-based thorough analysis gave useful structural insights into the analyzed molecular entity (**Fig. 5**). This assessment emphasizes the protein's structural features as well as its overall quality. Another test, called VERIFY3D, was run to assess the quality of the 3D-1D alignment. However, only 70.29% of the residues had a 3D-1D score of 0.1 or higher, suggesting a failure since less than 80% of the amino acids received a 3D/1D value of 0.1 or higher.

3.4. Molecular Docking results

The analysis of the data shown in **Table 3** offers a complete overview of the scores and quality indicators awarded to distinct ligands marked from H1 to H8, as well as ligand 19. Each column relates to criteria that are required to assess the potential efficiency of these ligands in their interaction with the target. The "Ligands" section contains the IDs of the ligands studied. The "Energy Score" (S) is an important indicator of the ligand's capacity to bind to the target, with lower values indicating more affinity.

Table 3. Molecular docking results by MOE.2014.

Ligands	Affinity	Rmsd-refine	E-Conf	E-Place	E-Score1	E-Refine	E-Score2
H1	-5.235	3.714	34.469	-40.545	-7.233	-14.389	-5.235
H2	-4.940	4.253	35.078	-59.887	-7.780	-16.242	-4.940
H3	-5.993	2.834	123.710	-87.252	-10.297	-19.389	-5.993
H4	-6.805	3.091	139.455	-93.456	-11.251	-25.753	-6.805
H5	-7.261	4.202	137.792	-88.665	-12.498	-26.952	-7.261
H6	-6.915	5.274	139.113	-72.249	-10.077	-16.238	-6.915
H7	-4.894	3.258	109.558	-47.590	-7.002	-14.806	-4.894
H8	-6.542	4.448	118.724	-94.477	-13.979	-20.878	-6.542
19	-7.868	3.140	129.609	-79.429	-7.220	-25.596	-7.868

The “Root Mean Square Deviation after Refinement” (Rmsd-refine) metric assesses the spatial agreement between the refined ligand's atomic locations and the reference, indicating the ligand's correctness after corrections. The Conformation Energy (E-Conf) measures the structural stability of the ligand in its bound state, whereas the Placement Energy (E-Place) evaluates the initial ligand placement within the binding site. The “E-Score1” and “E-Score2” parameters offer overall indicators of the ligand's quality after different calculation stages, whilst the Energy after Refinement (E-Refine) parameter gauges its post-refinement stability. An in-depth examination of the data indicates that ligand H5 has the largest negative energy score (S), indicating a potentially significant interaction with the target. Similarly, ligand H3 has the lowest RMSD-refined, indicating a tight overlap with the revised reference conformation. Despite its negative energy score, ligand 19 has a rather high E-Place, indicating difficulties in its initial location. Overall, ligands with significant negative E-Score1 and E-Score2 values appear to have more stable interactions with the target. However, in order to reach meaningful conclusions

about the intrinsic efficiency of the ligands under research, it is critical to evaluate acceptable quality levels for each metric, as well as other characteristics.

3.5. Analysis of complex interactions

The molecular interactions of H1_3OYA (**Fig. 6**) involve several interactions that shape the complex world of H1 ligands. Hydrophobic interactions, characterized by their aversion to water, reveal their influence in specific cases. By indexing these interactions, we find residue 222A (ARG) involved in a hydrophobic bond at distances of 3.99 and 3.73 Å, further highlighting its significant role in molecular arrangements. Hydrogen bonds, known for their role in molecular cohesion, emerge as essential linkers in this analysis. Indexed accordingly, residue 114A (ARG) forms a hydrogen bond with a hydrogen acceptor distance of 2.65 Å. Similarly, residue 117A (ARG) demonstrates its interactional potential with hydrogen bond distances of 3.29 and 2.2 Å in different cases. Residue 223A (Lysine) contributes to the molecular interactions by forming a hydrogen bond at a distance of 3.19 Å, while residue 226A (Aspartic acid) forms a hydrogen bond at a distance of 2.43 Å. These hydrogen bonds collectively facilitate dynamic interactions within the molecular framework. Additionally, π-cation interactions play a significant role. Residue 219A (Lysine) exemplifies this interaction, establishing a binding force at 3.9 Å. Through these mechanisms, molecules navigate complex pathways and form connections, underscoring the intricacies of molecular dynamics.

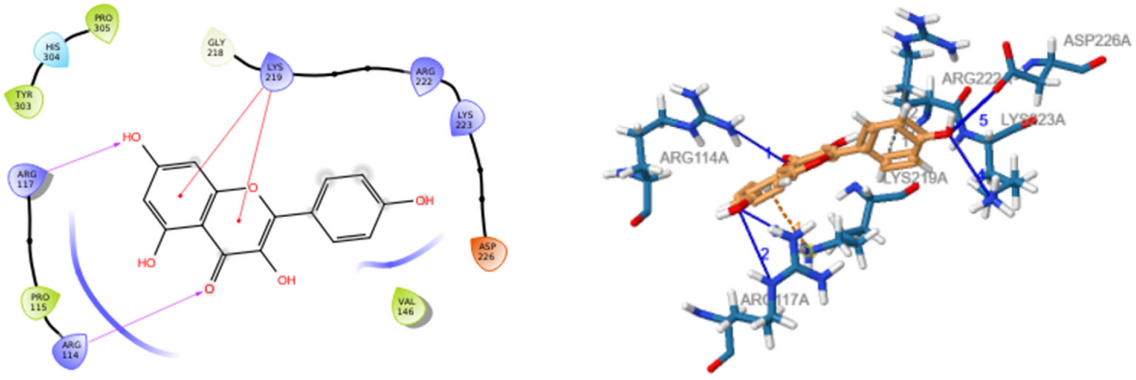


Fig. 6. Visualization of H1_3OYA complex interactions

In H2_3OYA complex molecular interactions (**Fig. 7**), distinct forces come into play to shape the complex H2 ligand world, with hydrophobic forces playing an essential role. In this context, a significant case emerges where residue 215A (GLN) engages in a hydrophobic interaction, maintaining 3.94 Å. This emphasizes the significance of hydrophobic interactions in the creation of complex landscapes in molecular configurations where certain residues of amino acids play a role. Hydrogen bonds, known for their ability to link molecules, reveal their presence in distinct cases. Residue 114A (ARG) is involved in two hydrogen bonds, each characterized by unique distances. In the first case, the distance between the hydrogen atom and the acceptor is 3.14 Å, and that between the donor and acceptor atoms is 4.1 Å. In the second case, hydrogen bonding is manifested by distances of 2.89 Å and 3.66 Å between hydrogen and acceptor and between donor and acceptor, respectively.

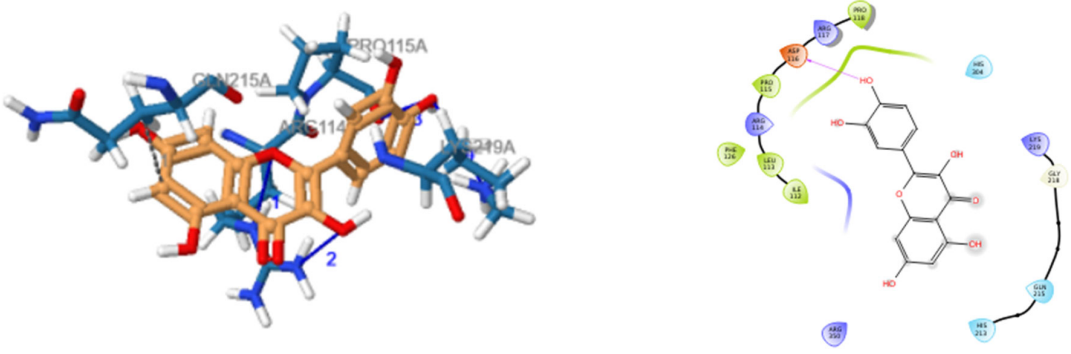


Fig. 7. Visualization of H2_3OYA complex interactions

In addition, residue 115A (PRO) contributes to the hydrogen bonding network, forming an interaction with 2.69 Å between hydrogen and acceptor atoms and 3.43 Å between donor and acceptor atoms. In addition, residue 219A (LYS) is also engaged in a hydrogen bonding interaction, displaying distances of 3.42 Å and 3.98 Å for the hydrogen-acceptor and donor-acceptor spans, respectively.

These hydrogen bonds illustrate the complex nature of molecular interactions, highlighting the diverse ways in which amino acid residues establish connections to maintain the structural and functional integrity of complex systems. In H3_3OYA complex molecular interactions (**Fig. 8**), distinct forces come into play to shape the complex world of ligand H3; molecular interactions are the basis of complex systems, revealing their subtleties through distinct forces. In the context of hydrophobic interactions, a notable example emerges where residue 112A (ILE) orchestrates a dynamic interaction. This interaction is characterized by 3.67 Å, demonstrating the role of hydrophobic forces in molecular arrangements. Hydrogen bonds, crucial for molecular cohesion, appear repeatedly in this analysis. Residue 114A (ARG) engages in two distinct hydrogen bonds. The first bond is defined by a hydrogen-acceptor distance of 2.39 Å and a donor-acceptor span of 3.04 Å, forming at a donor angle of 117.58°. The second interaction features hydrogen and donor-acceptor span distances of 2.7 Å and 3.28 Å, respectively. In addition, residue 219A (LYS) contributes to the hydrogen bonding network, participating in an interaction with a hydrogen-acceptor distance of 1.89 Å and a donor-acceptor span of 2.95 Å the donor angle measures 166.95°, further delineating the geometric complexities of the bond. The influence of residue 350A (ARG) is also evident through hydrogen bonding. It establishes two distinct interactions with distances of 2.03 Å and 2.66 Å for hydrogen-acceptor spans, and 3.01 Å and 3.46 Å for donor-acceptor spans. The respective donor angles are measured at 150.19° and 131.59°, underscoring the diversity of these interactions. In addition to hydrogen bonds, salt bridges are also a significant feature. Residue 114A (Arginine) contributes to this by forming a salt bridge at 4.68 Å. This interaction is enhanced by the presence of a positive charge on the protein and is linked to the carboxylate group of ligands. Together, these interactions underscore the intricate forces within molecular systems, highlighting their essential role in shaping the structure and function of complex biological entities.

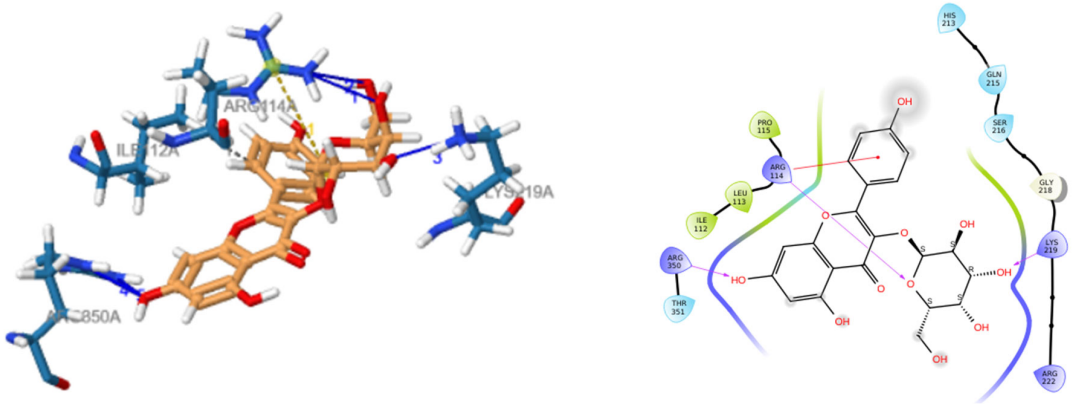


Fig. 8. Visualization of H3_3OYA complex interactions

In H4_3OYA complex molecular interactions (**Fig. 9**), distinct forces come into play to shape the complex world of H4 ligand, with the forces involved in weaving a tapestry of connections. Hydrophobic interactions, which play an essential role in the formation of these arrangements, are illustrated by a striking example involving residue 351A (THR). This interaction occurs at 3.75 Å, linking ligand atom 5950 to protein atom 5545. Here, hydrophobic forces play an important role in maintaining molecular architecture. Hydrogen bonds, essential for molecular cohesion, appear repeatedly in this analysis. Residue 114A (ARG) is involved in the formation of hydrogen bonds, each characterized by specific distances and angles. The first interaction has a hydrogen-acceptor distance of 2.1 Å and a donor-acceptor span of 2.88 Å, while the second interaction involves residue 216A (SER) with distances of 3.07 Å and 4.09 Å, respectively.

Fig. 9. Visualization of H4_3OYA complex interactions

Residue 219A (LYS) reinforces the hydrogen bond network, with distances of 1.85 Å and 2.92 Å for the hydrogen and donor-acceptor spans. Residue 350A (ARG) contributes twice, with distances of 2.13 Å and 2.12 Å, demonstrating the diversity of these interactions. In addition, residue 351A (THR) participates in hydrogen bonding, with distances of 2.94 Å

and 2.82 Å, highlighting the complexity of these connections. Salt bridges, characterized by their role in molecular stability, are an important feature in this context. Residue 114A (ARG) contributes to this phenomenon by forming a salt bridge at 4.81 Å. This interaction is enriched by the presence of a positive entity for the protein and involves the "Carboxylate" ligand group. Collectively, these interactions reflect the multifaceted nature of molecular systems, where diverse forces converge to influence the structural and functional properties of biological entities. In H5_3OYA complex molecular interactions (Fig. 10), distinct forces come into play to shape the complex H5 ligand world. Hydrophobic interactions, essential in molecular arrangements, present a notable example involving residue 115A (PRO). This interaction is characterized by 3.73 Å, highlighting the role of hydrophobic forces in molecular stabilization. Hydrogen bonds, known for their cohesive role, play an important role in this analysis. Residue 114A (ARG) contributes to multiple hydrogen bonds, each marked by specific distances. The first interaction involves distances of 3.06 Å for the hydrogen-acceptor span and 3.71 Å for the donor-acceptor span. The second interaction features distances of 1.97 Å and 2.82 Å for the hydrogen and donor-acceptor spans, respectively. In addition, residue 207A (GLU) exhibits a hydrogen bonding interaction with distances of 2.43 Å and 3.22 Å. Residue 208A (PHE) participates with distances of 3.21 Å and 3.95 Å, while residue 216A (SER) contributes twice, with distances of 2.87 Å and 2.32 Å, respectively. Residue 218A (GLY) is also involved, with distances of 3.35 Å and 4.09 Å, while residue 219A (LYS) completes the network with distances of 2.06 Å and 3.13 Å. The hydrogen bonding network is further enriched by residue 350A (ARG) on two occasions, with distances of 2.1 Å and 2.28 Å. Finally, residue 351A (THR) solidifies this network with a hydrogen bond distance of 2.66 Å. Intriguingly, the phenomenon of π-cation interactions surfaces. Residue 183A (HIS) illustrates this interaction, forming a bond at 5.16 Å with an offset of 1.11 Å, highlighting the fascinating geometric complexities of these interactions. Collectively, these interactions reveal the complex dynamics of molecular forces, showing how they collaborate harmoniously to shape the structure and function of biological entities.

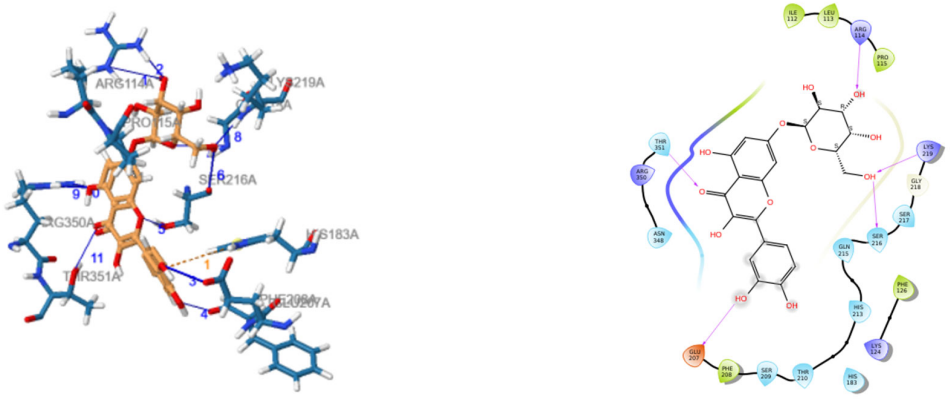


Fig. 10. Visualization of H5_3OYA complex interactions

Several interactions are involved in the formation of the H6 complex in the complicated molecular interactions of H6_3OYA (**Fig. 11**). Hydrophobic interactions, which are critical in the maintenance of molecular structures, express themselves in spectacular circumstances. Residue 115A (PRO) has a hydrophobic interaction with 3.9 Å, whereas residue 350A (ARG) has a distance of 3.8 Å. Hydrogen bonds, renowned for their cohesive role, appear several times in this study. Residue 114A (ARG) participates in two distinct hydrogen bonds, each marked by specific distances. The first interaction reveals a hydrogen-acceptor distance of 3.08 Å and a donor-acceptor span of 3.95 Å, while the second features hydrogen and donor-acceptor span distances of 3.43 Å and 3.95 Å, respectively. Residue 124A (LYS) contributes to this network via a hydrogen bond involving distances of 1.68 Å and 2.67 Å, highlighting the complex geometry of the interaction. Residue 210A (THR) reinforces the network by forming two hydrogen bonds, each with specific distances. The first bond has distances of 2.72 Å and 3.38 Å, while the second interaction has distances of 2.79 Å and 3.38 Å. Residue 216A (SER) also engages twice, with distances of 3.11 Å and 2.73 Å for one and 2.73 Å and 3.63 Å for the other. To add to the complexity, residue 348A (ASN) establishes hydrogen bonding at distances of 2.31 Å and 3.13 Å, while residue 351A (THR) contributes at distances of 2.99 Å and 3.55 Å. In addition, residue 365A (SER) engages in hydrogen bonding at distances of 2 Å and 2.88 Å, while residue 366A (ILE) forms an interaction at distances of 3.33 Å and 4.1 Å. Salt bridges, characterized by their role in molecular stability, are an integral part of this molecular landscape. Residue 183A (HIS) participates in this phenomenon by forming a salt bridge at 4.99 Å, enriched by the presence of a positively charged protein entity. Collectively, these interactions highlight the rich diversity of forces within molecular systems, shedding light on their role in shaping the complex structure and function of biological entities. In H7_3OYA complex molecular interactions (Fig. 12), distinct forces come into play to shape the H7 ligand complex world. Hydrogen bonds, pivotal to molecular cohesion, are highlighted in this study through distinct cases. Residue 115A (PRO) participates in the dynamics of hydrogen bonds, revealing two interactions with specific distances. The first bond has a hydrogen-acceptor distance of 2.49 Å and a donor-acceptor span of 3.32 Å, while the second interaction has hydrogen and donor-acceptor span distances of 2.36 Å and 3.02 Å, respectively. Residue 219A (LYS) further enriches the network by contributing to a hydrogen bond with distances of 2.67 Å and 3.36 Å.

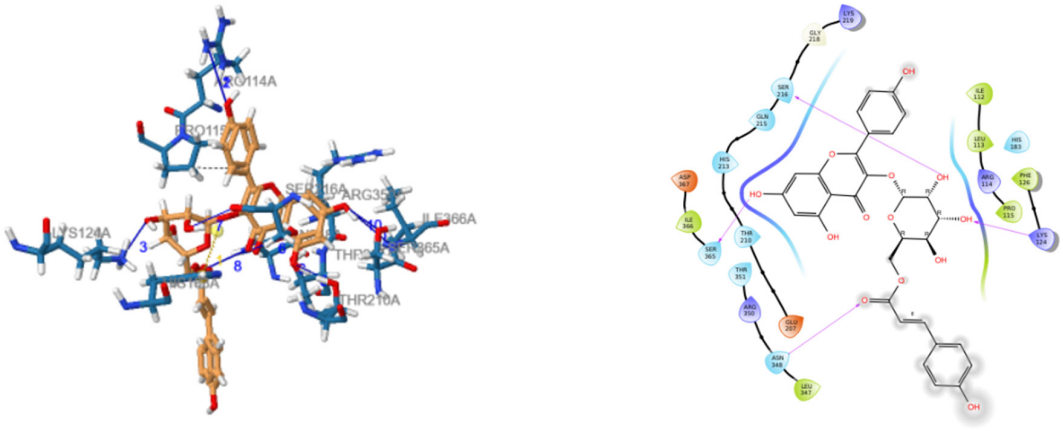


Fig. 11. Visualization of H6_3OYA complex interactions

Intriguingly, the concept of π-cation interactions surfaces. Residue 114A (ARG) illustrates this phenomenon by forming a dynamic interaction at 4.57 Å with an offset of 0.47 Å, highlighting the fascinating geometry underlying these interactions. Furthermore, salt bridges, essential for molecular stability, appear to be an integral factor in this molecular landscape. Residues 114A (ARG) and 219A (LYS) participate in this phenomenon, forming salt bridges at distances of 4.99 Å and 3.95 Å, respectively. Enhanced by the presence of positively charged protein entities, these interactions highlight the role of charge-based forces in the formation of molecular arrangements. Together, these interactions provide a glimpse into the complex tapestry of molecular forces, highlighting their role in shaping the structure, dynamics, and function of biological entities.

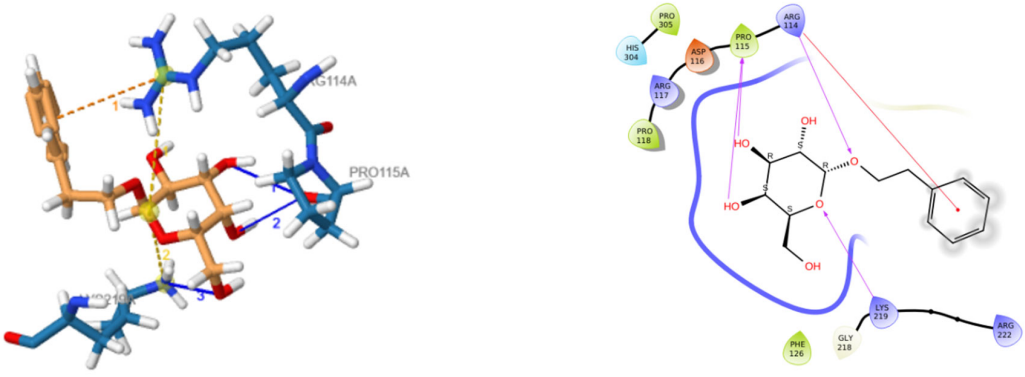


Fig. 12. Visualization of H7_3OYA complex interactions

In H8_3OYA complex molecular interactions (**Fig. 13**), distinct forces come into play to shape the H8 ligand complex world. Hydrogen bonds, which play an essential role in molecular cohesion, occupy a central place in this study, as illustrated by several examples. Residue 113A (LEU) contributes to the hydrogen bonding network, forming an interaction with a hydrogen-acceptor distance of 3.01 Å and a donor-acceptor span of 3.88 Å. Residue 208A (PHE) continues this story with hydrogen and donor-acceptor distances of 2.07 Å and 2.84 Å, respectively. Residue 216A (SER) is also part of this network, with hydrogen bond distances of 2.75 Å and 3.42 Å, while residue 350A (ARG) participates with distances of 2.58 Å and 2.06 Å in separate cases. This set of hydrogen bonds highlights the complex network of forces governing molecular interactions. Intriguingly, the concept of π-cation interactions also comes to the fore. Residue 114A (ARG) illustrates this phenomenon by forming a π-cation interaction at 5.01 Å with an offset of 1.78 Å. This dynamic interaction highlights the geometric complexities involved in these types of bonds. Collectively, these interactions provide insight into the delicate balance of forces that govern molecular arrangements, shaping the structure and function of biological entities. In the molecular interactions of complex 19_3OYA (Fig. 14), distinct forces come into play to shape the complex world of ligand 19. Hydrophobic interactions, which play a key role in molecular arrangements, are highlighted by residue 222A (ARG) and residue 260A (VAL). The interaction involving residue 222A (ARG) shows 3.5 Å, while residue 260A (VAL) engages in a similar interaction with a distance of 3.83 Å. These examples underline the role of hydrophobic forces in the formation of molecular architectures.

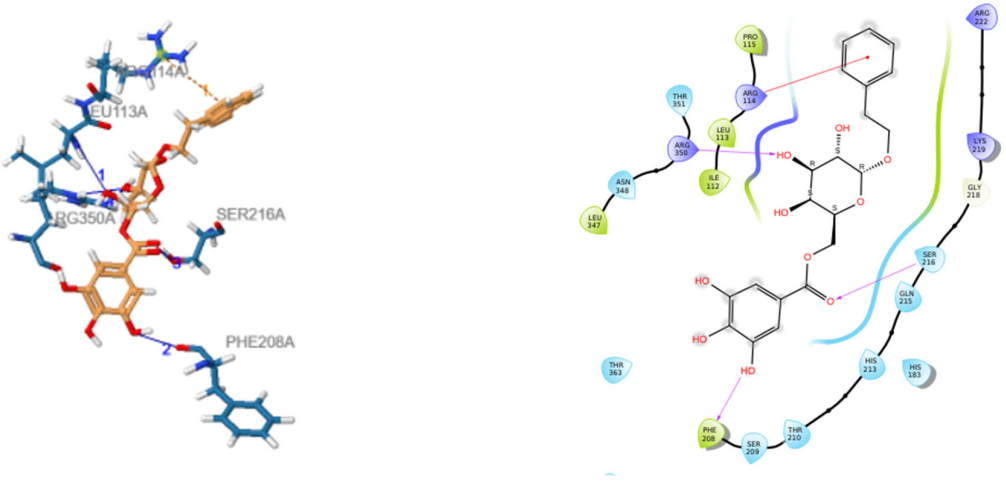


Fig. 13. Visualization of H8_3OYA complex interactions

Hydrogen bonds, known for their cohesive role, are illustrated by distinct interactions in this analysis. Residue 114A (ARG) contributes to the hydrogen bonding network by participating in two distinct interactions. The first interaction reveals a hydrogen-acceptor distance of 2.17 Å and a donor-acceptor span of 3.23 Å. The second shows distances of 3.14 Å and 4.02 Å for the hydrogen and donor-acceptor spans, respectively. In addition, residue 219A (LYS) reinforces the network by forming a hydrogen bond with distances of 3.18 Å and 3.78 Å. Intriguingly, the concept of π-cation interactions surfaces. Residue 219A (LYS) illustrates this phenomenon by forming a dynamic interaction at 3.56 Å with an offset of 1.58 Å, highlighting the fascinating geometry underlying these interactions. Collectively, these interactions provide insight into the complex interplay of forces that shape the intricate structure and function of biological entities.

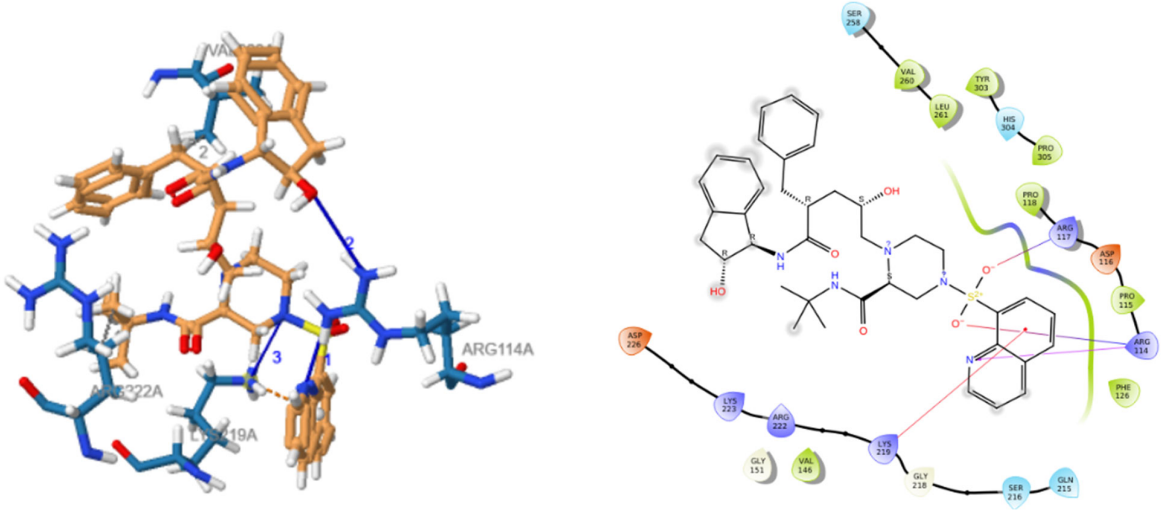


Fig. 14. Visualization of 19_3OYA complex interactions

3.6. ADME-Tox results

❖ Ligand 19

The physicochemical properties of the compound under study provide essential clues to its molecular characteristics. The chemical formula is $C_{39}H_{47}N_5O_6S$, corresponding to a molecular weight of 713.89 g/mol. With 51 heavy atoms, including 22 aromatic heavy atoms, the compound is structurally complex. The fraction of sp3-hybridized carbon atoms (Csp3) is 0.41, meaning a balanced distribution between sp2 and sp3 carbons. The presence of 14 rotating bonds highlights molecular flexibility, while 9 hydrogen bond acceptors and 4 hydrogen bond donors reflect its potential for specific interactions. The molar refractivity is 203.45, and the topological polar surface area (TPSA) spans 160.55 Å², indicating a large surface area for potential interactions. Regarding lipophilicity, several Log Po/w values give an indication of the compound's partition coefficient. The iLOGP value is 4.26, the XLOGP3 value is 3.70, the WLOGP value is 3.20, the MLOGP value is 1.36, and the SILICOS-IT value is 3.46. The consensus Log Po/w is 3.20, reflecting its moderate lipophilicity. Water solubility plays a key role in the compound's behavior. The Log S (ESOL) value of -5.99 corresponds

to a moderately soluble class. Solubility is 7.27e-04 mg/ml or 1.02e-06 mol/l. In contrast, the Log S (Ali) value is -6.76, indicating low solubility, with a solubility value of 1.23e-04 mg/ml or 1.73e-07 mol/l. The Log S (SILICOS-IT) value of -9.49 classifies it as poorly soluble, with a solubility of 2.32e-07 mg/ml or 3.25e-10 mol/l.

Pharmacokinetic attributes reveal important information about the compound's behavior in the body. It exhibits low gastrointestinal (GI) absorption and is not considered permeable to the blood-brain barrier (BBB). It acts as a substrate for P-glycoprotein (P-gp) but does not inhibit CYP1A2 or CYP2C9. However, it does inhibit CYP2C19 and CYP3A4, key enzymes in drug metabolism. The compound's skin permeation potential is reflected by its Log Kp value of -8.03 cm/s.

Several filters are applied to assess pharmacokinetics. Lipinski's rule of five identifies two violations - molecular weight (MW) greater than 500 and number of oxygen and nitrogen atoms (NorO) greater than 10. The Ghose filter identifies three violations: molecular weight greater than 480, molecular refractive index (MR) greater than 130, and number of atoms greater than 70. Veber's rule identifies two violations: rotating bonds exceeding 10 and topological polar surface area (TPSA) exceeding 140. Egan's rule identifies a violation when TPSA exceeds 131.6. Similarly, Muegge's rule reveals two violations: a soft mass greater than 600 and a topological polar area greater than 150. The bioavailability score is 0.17, indicating a potential for bioavailability.

In the medicinal chemistry evaluation, the compound shows no alerts according to the PAINS or Brenk filters. However, it does not meet the parentage criteria, with three violations: a wavelength greater than 350, rotating bonds greater than 7, and an XLOGP3 value greater than 3.5. Finally, synthetic accessibility is rated at 6.24, indicating a moderate level of ease in synthesizing the compound.

Overall, this comprehensive range of physicochemical properties provides an understanding of the compound's molecular attributes, solubility, pharmacokinetics, and medicinal chemistry potential, factors that are essential for assessing its suitability for drug development.

- Ligand H1

The physicochemical properties provided give a comprehensive overview of the compound's molecular attributes. With a chemical formula of $C_{15}H_{10}O_6$ and a molecular weight of 286.24 g/mol, the compound contains 21 heavy atoms, 16 of which are aromatic. The compound is notably devoid of sp3-hybridized carbon atoms (Csp3), indicating a highly aromatic structure. The presence of a single rotatable bond suggests limited molecular flexibility. The compound exhibits 6 hydrogen bond acceptors and 4 hydrogen bond donors, indicating its potential for specific interactions. Its molar refractive index is 76.01, and its topological polar surface area (TPSA) spans 111.13 Å², suggesting a moderate surface area for interactions.

With regard to lipophilicity, the compound's Log Po/w values give an indication of its partition coefficient. The iLOGP value is 1.70, the XLOGP3 value is 1.90, the WLOGP value is 2.28, the MLOGP value is -0.03, and the SILICOS-IT value is 2.03. The consensus Log Po/w is 1.58, underlining its moderate to low lipophilicity.

Water solubility is a crucial factor influencing compound behavior. The Log S (ESOL) value is -3.31, indicating its soluble nature. Its solubility is 1.40e-01 mg/ml or 4.90e-04 mol/l, which classifies it as a soluble substance. The Log S (Ali) value of -3.86 also suggests solubility, with a reported solubility of 3.98e-02 mg/ml or 1.39e-04 mol/l. The Log S (SILICOS-IT) value of -3.82 aligns with its soluble classification, supported by a solubility of 4.29e-02 mg/ml or 1.50e-04 mol/l.

Exploration of pharmacokinetics enables us to elucidate the compound's behavior in the body. It has a high gastrointestinal absorption and is not considered permeable to the blood-brain barrier (BBB). It is not a P-glycoprotein (P-gp) substrate, indicating limited efflux potential. While it inhibits CYP1A2 and CYP2D6, it does not inhibit CYP2C19 or CYP2C9. In addition, it acts as an inhibitor of CYP3A4, a pivotal enzyme in drug metabolism. The compound's skin permeation potential is reflected by its Log Kp value of -6.70 cm/s.

In pharmacokinetic evaluation, the compound performs well against several filters. Lipinski's Rule of Five identifies no violations, indicating that the compound exhibits drug-like characteristics. Similarly, the Ghose, Veber, Egan, and Muegge filters give positive results, indicating their potential as a drug candidate. The bioavailability score is 0.55, indicating a moderate probability of bioavailability. In the evaluation of medicinal chemistry attributes, the compound presents no alerts according to the PAINS or Brenk filters. In addition, it meets the lead likeness criteria, further enhancing its potential for drug development. Finally, the compound's synthetic accessibility is rated at 3.14, suggesting a reasonable level of ease of synthesis. Collectively, these physicochemical properties provide insight into the compound's molecular attributes, solubility, pharmacokinetics, and medicinal chemistry potential. These attributes play a crucial role in assessing its suitability for drug development and further exploration.

- Ligand H2

The physicochemical properties provided give an overview of the compound's molecular characteristics. With a chemical formula of $C_{15}H_{10}O_7$ and a molecular weight of 302.24 g/mol, the compound comprises 22 heavy atoms, 16 of

which are aromatic. The absence of sp3-hybridized carbon atoms (Csp3) indicates a highly aromatic structure. It has a single rotatable bond, which means limited molecular flexibility. The compound has 7 hydrogen bond acceptors and 5 hydrogen bond donors, suggesting its potential for specific interactions. Its molar refractive index is 78.04, and its topological polar surface area (TPSA) spans 131.36 Å², indicating a moderate surface area available for interactions.

The compound's lipophilicity is elucidated by various Log Po/w values. The iLOGP value is 1.63, the XLOGP3 value is 1.54, the WLOGP value is 1.99, the MLOGP value is -0.56, and the SILICOS-IT value is 1.54. The consensus Log Po/w is 1.23, indicating moderate lipophilicity.

Water solubility is a key determinant of compound behavior. A Log S (ESOL) value of -3.16 indicates its soluble nature. Its solubility is 2.11e-01 mg/ml or 6.98e-04 mol/l, which classifies it as a soluble substance. The Log S (Ali) value of -3.91 also suggests solubility, with a reported solubility of 3.74e-02 mg/ml or 1.24e-04 mol/l. The Log S (SILICOS-IT) value of -3.24 corresponds to its soluble classification, supported by a solubility of 1.73e-01 mg/ml or 5.73e-04 mol/l.

Pharmacokinetic attributes provide an overview of the compound's behavior in the body. It exhibits high gastrointestinal absorption and is not permeable to the BBB. It does not act as a P-glycoprotein (P-gp) substrate, indicating limited efflux potential. It inhibits CYP1A2 and CYP2D6 but not CYP2C19 or CYP2C9. In addition, it inhibits CYP3A4, an enzyme important in drug metabolism. Its skin permeation potential is indicated by its Log Kp value of -7.05 cm/s.

In terms of pharmacokinetics, the compound passes through several filters. Lipinski's Rule of Five reveals no violations, underlining its adherence to drug characteristics. The Ghose, Veber, Egan, and Muegge filters also give positive results, indicating its potential as a drug candidate. The bioavailability score is 0.55, suggesting a moderate probability of bioavailability.

Assessing medicinal chemistry attributes, the compound presents a single alert according to the PAINS filter for the presence of a catechol group. Similarly, the Brenk filter identifies one catechol alert. Nevertheless, it meets the lead likeness criteria, reinforcing its potential for drug development. Its synthetic accessibility is rated at 3.23, indicating a reasonable level of ease in synthesizing the compound.

These comprehensive physicochemical properties provide valuable insights into the compound's molecular characteristics, solubility, pharmacokinetics, and medicinal chemistry potential. These attributes play a crucial role in assessing its suitability for drug development and further exploration.

- Ligand H3

The compound's physicochemical properties shed light on its molecular attributes and potential suitability for drug development. With a chemical formula of $C_{21}H_{20}O_{11}$ and a molecular weight of 448.38 g/mol, it contains 32 heavy atoms, 16 of which are aromatic. The fractional Csp^3 value of 0.29 indicates a moderate aliphatic character, while the presence of 4 rotatable bonds suggests a certain molecular flexibility. The compound has 11 hydrogen bond acceptors and 7 hydrogen bond donors, indicating its potential for specific interactions. Its molar refractive index is 108.13, and its topological polar surface area (TPSA) covers 190.28 Å², suggesting a substantial surface area available for interactions.

Lipophilicity is assessed using various Log Po/w values. The iLOGP value is 0.53, the XLOGP3 value is 0.72, the WLOGP value is -0.24, the MLOGP value is -2.10, and the SILICOS-IT value is -0.12. The consensus Log Po/w value is -0.25, indicating a general tendency towards hydrophilicity.

Water solubility is a key factor in determining a compound's behavior. The Log S (ESOL) value of -3.18 implies its soluble nature. It has a solubility of 2.97e-01 mg/ml or 6.61e-04 mol/l, which classifies it as soluble. The Log S (Ali) value of -4.29 indicates moderate solubility, with a reported solubility of 2.28e-02 mg/ml or 5.08e-05 mol/l. The Log S (SILICOS-IT) value of -2.10 corresponds to soluble classification, supported by a solubility of 3.55e+00 mg/ml or 7.91e-03 mol/l.

Pharmacokinetic characteristics provide an overview of the compound's behavior in the body. It has low gastrointestinal (GI) absorption and is not permeable to the BBB. It does not act as a substrate for P-glycoprotein (P-gp), indicating limited efflux potential. Furthermore, it does not inhibit any of the cytochrome P450 enzymes tested (CYP1A2, CYP2C19, CYP2C9, CYP2D6, CYP3A4). Its skin permeation potential is indicated by its Log Kp value of -8.52 cm/s.

With regard to pharmacokinetics, the compound's adherence to various filters was evaluated. It violates Lipinski's rule of five, with two violations linked to the presence of nitrogen heteroatoms (NH or OH > 5) and a high number of oxygen-hydrogen groups (N or O > 10). It passes the Ghose filter, suggesting that it is suitable for oral absorption. Although it fails the Veber filter due to a TPSA>140, it adheres to the Egan filter criteria but fails the Muegge filter due to a high TPSA, a high number of hydrogen bond acceptors (H-acc>10) and hydrogen bond donors (H-don>5). The bioavailability score is 0.17, indicating a moderate probability of bioavailability.

In terms of medicinal chemistry, the compound passes both the PAINS and Brenk filters, suggesting that interference compounds common to all assays can be avoided. It does not meet the reliability criteria due to its molecular weight (MW>350). Its synthetic accessibility is evaluated at 5.29, indicating a reasonable level of ease of synthesizing the compound.

These comprehensive physicochemical properties provide valuable insights into the compound's molecular characteristics, solubility, pharmacokinetics, and medicinal chemistry potential. These attributes play an essential role in assessing its suitability for drug development and further exploration.

❖ Ligand H4

The compound's physicochemical properties provide valuable insights into its molecular characteristics and drug development potential. With a chemical formula of $C_{21}H_{20}O_{11}$ and a molecular weight of 448.38 g/mol, the compound contains 32 heavy atoms, 16 of which are aromatic. The Csp3 fraction value of 0.29 indicates a moderate aliphatic character. It has 4 rotatable bonds, suggesting molecular flexibility, and possesses 11 hydrogen bond acceptors and 7 hydrogen bond donors, highlighting its potential for specific interactions. The compound's molar refractive index is 108.13, and its topological polar surface area (TPSA) covers 190.28 Å², suggesting a substantial surface area available for interactions.

Several Log Po/w values were used to assess lipophilicity: iLOGP (1.55), XLOGP3 (0.72), WLOGP (-0.24), MLOGP (-2.10) and SILICOS-IT (-0.12). The consensus value for Log Po/w is -0.04, indicating a tendency towards hydrophilicity.

The compound's water solubility is a determining factor in its behavior. The Log S (ESOL) value of -3.18 indicates its solubility. It has a solubility of 2.97e-01 mg/ml or 6.61e-04 mol/l, which classifies it as soluble. The Log S (Ali) value of -4.29 indicates moderate solubility, with a reported solubility of 2.28e-02 mg/ml or 5.08e-05 mol/l. The Log S (SILICOS-IT) value of -2.10 corresponds to soluble classification, supported by a solubility of 3.55e+00 mg/ml or 7.91e-03 mol/l.

The pharmacokinetic profile highlights the compound's behavior in the body. It has low gastrointestinal (GI) absorption and is not permeable across the blood-brain barrier (BBB). It is not a P-glycoprotein (P-gp) substrate, indicating limited efflux potential. Furthermore, it does not inhibit any of the cytochrome P450 enzymes tested (CYP1A2, CYP2C19, CYP2C9, CYP2D6, CYP3A4). Its skin permeation potential is indicated by its Log Kp value of -8.52 cm/s.

In terms of pharmacokinetics, the compound violates Lipinski's rule of five with two violations: the presence of nitrogen heteroatoms (NH or OH > 5) and a high number of oxygen-hydrogen groups (N or O > 10). It adheres to the Ghose filter, suggesting that it can be absorbed orally. Although it violates the Veber filter due to a TPSA>140, it adheres to the Egan filter but violates the Muegge filter due to a high TPSA, a high number of hydrogen bond acceptors (H-acc>10) and hydrogen bond donors (H-don>5). The bioavailability score is 0.17, indicating a moderate probability of bioavailability.

In terms of medicinal chemistry, the compound passes both the PAINS and Brenk filters, suggesting that interference compounds common to all assays can be avoided. It does not meet the reliability criteria due to its molecular weight (MW>350). Its synthetic accessibility is evaluated at 5.24, suggesting a reasonable level of ease in synthesizing the compound.

These comprehensive physicochemical properties provide valuable insights into the compound's molecular characteristics, solubility, pharmacokinetics, and medicinal chemistry potential. These attributes play a crucial role in assessing its suitability for drug development and further exploration.

❖ Ligand H5

The physicochemical properties of the compound $C_{21}H_{20}O_{12}$ provide information on its molecular characteristics and its potential for drug development. With a molecular mass of 464.38 g/mol, the compound contains 33 heavy atoms, 16 of them aromatic, and has a Csp3 fraction of 0.29, indicating a moderate aliphatic character. It exhibits 4 rotatable bonds, suggesting molecular flexibility, and has 12 hydrogen bond acceptors and 8 hydrogen bond donors, indicating its potential for specific interactions. The compound's molar refractive index is 110.16, and its topological polar surface area (TPSA) covers 210.51 Å², indicating a substantial surface area available for interactions.

Lipophilicity evaluation involved several Log Po/w values: iLOGP (1.54), XLOGP3 (0.36), WLOGP (-0.54), MLOGP (-2.59) and SILICOS-IT (-0.59). The Log Po/w consensus value is -0.37, suggesting a trend towards hydrophilicity.

The compound's solubility in water is an important determinant of its behavior. The Log S (ESOL) value of -3.04 indicates its solubility, with a reported solubility of 4.23e-01 mg/ml or 9.10e-04 mol/l, classifying it as soluble. The Log S (Ali) value of -4.35 indicates moderate solubility, with a reported solubility of 2.10e-02 mg/ml or 4.51e-05 mol/l. The Log S (SILICOS-IT) value of -1.51 suggests high solubility, with a reported solubility of 1.43e+01 mg/ml or 3.08e-02 mol/l.

The pharmacokinetic profile provides information on the compound's behavior in the body. It shows low gastrointestinal (GI) absorption, is not permeable to the blood-brain barrier (BBB), and is not a P-glycoprotein (P-gp) substrate, indicating limited efflux potential. It does not inhibit any of the cytochrome P450 enzymes tested (CYP1A2, CYP2C19, CYP2C9, CYP2D6, CYP3A4). Its skin permeation potential is indicated by its Log Kp value of -8.88 cm/s.

With regard to drug similarity, the compound violates Lipinski's Rule of Five with 2 violations: the presence of nitrogen atoms (NH or OH > 5) and a high number of oxygen-hydrogen groups (N or O > 10). It also violates the Ghose filter due to its low Log Po/w value (WLOGP<-0.4). It violates Veber's filter due to TPSA>140, and it adheres to Egan's filter but violates Muegge's filter due to high TPSA, a high number of hydrogen bond acceptors (H-acc>10) and hydrogen bond donors (H-don>5). The bioavailability score is 0.17, indicating a moderate probability of bioavailability.

In terms of medicinal chemistry, the compound passes both the PAINS and Brenk filters, suggesting its potential to avoid common interfering compounds in testing. It does not violate drug similarity criteria, except for its molecular weight (MW>350). Its synthetic accessibility is rated at 5.31, indicating a reasonable level of ease of compound synthesis.

These detailed physicochemical properties provide valuable information on the compound's molecular characteristics, solubility, pharmacokinetics, drug similarity, and medicinal chemistry potential. These attributes are essential for assessing its relevance to drug development and for further exploration.

- Ligand H6

The molecule's physicochemical properties present a set of essential attributes. Its chemical formula is $C_{30}H_{26}O_{13}$, giving it a molecular weight of 594.52 g/mol. The structure contains 43 heavy atoms, including 22 aromatic atoms, with a Csp^3 fraction of 0.20, indicating some presence of single bonds in its structure. It is composed of 8 rotatable bonds and has 13 hydrogen bond acceptors and 7 hydrogen bond donors. Its molar refractive index is 149.51, and its total polar surface area (TPSA) is 216.58 Å². With regard to lipophilicity, several indicators were measured, including Log Po/w (iLOGP) of 2.99, Log Po/w (XLOGP3) of 2.47 and Log Po/w (WLOGP) of 1.62. However, it also shows negative values for Log Po/w (MLOGP) of -1.04 and Log Po/w (SILICOS-IT) of 1.56, with a consensus of 1.52. Its water-solubility properties show some variability: Log S (ESOL) is -4.93, indicating moderate solubility, while Log S (Ali) is -6.66, suggesting low solubility, and Log S (SILICOS-IT) is -4.23, highlighting moderate solubility. Pharmacokinetic parameters reveal that the molecule has low gastrointestinal absorption and does not permeate the blood-brain barrier (BBB). It is not a P-glycoprotein (P-gp) substrate, nor is it an inhibitor of CYP1A2, CYP2C19, CYP2C9, CYP2D6, or CYP3A4 enzymes. Skin permeation (Log Kp) was measured at -8.17 cm/s. The characteristics of similarity with drugs show that the molecule does not comply with all the established rules. For example, it violates the Lipinski rule with 3 violations: a molecular weight greater than 500 and the presence of more than 10 NH and OH groups combined. It also violates the Ghose rule and the Veber rule. However, it does not trigger an alert according to the PAINS and Brenk filters, and its medicinal chemistry shows only one violation concerning the criterion of molecular weight (MW) greater than 350. Finally, its synthetic accessibility is evaluated at 5.96, indicating a degree of feasibility in its synthesis.

- Ligand H7

The physicochemical properties of the molecule in question provide a comprehensive overview of its attributes. Its chemical formula is $C_{14}H_{20}O_6$, giving it a molecular weight of 284.31 g/mol. The molecule contains 20 heavy atoms, 6 of which are aromatic, with a Csp3 fraction of 0.57, indicating a predominance of single bonds in its structure. It has 5 rotating bonds, 6 hydrogen bond acceptors, and 4 hydrogen bond donors. Its molar refractive index is 69.76, and its total polar surface area (TPSA) is 99.38 Å². With regard to lipophilicity, several indicators were evaluated. Log Po/w (iLOGP) is 1.49, while Log Po/w (XLOGP3) is -0.69, and Log Po/w (WLOGP) is -0.95. Log Po/w (MLOGP) and Log Po/w (SILICOS-IT) show similar values at -0.69 and -0.01 respectively. The consensus Log Po/w is -0.17. Water solubility properties reveal the molecule to be highly soluble, with Log S (ESOL), Log S (Ali), and Log S (SILICOS-IT) values of -1.06, -0.92, and -1.02, respectively. Its pharmacokinetic properties indicate high gastrointestinal absorption, but it cannot cross the blood-brain barrier (BBB). It is not a P-glycoprotein (P-gp) substrate nor an inhibitor of CYP1A2, CYP2C19, CYP2C9, CYP2D6, or CYP3A4 enzymes. Skin permeation (Log Kp) is measured at -8.52 cm/s. Drug similarity characteristics indicate that the molecule respects Lipinski's rule, triggering no violations. It does not comply with Ghose's rule due to a negative value for Log Po/w (WLOGP) but does comply with Veber's, Egan's, and Muegge's rules. The bioavailability score is 0.55, and in terms of medicinal chemistry, it meets the criteria for lead likeness (similarity to lead) but has a low degree of synthetic accessibility, assessed at 4.36.

- Ligand H8

The molecule's physicochemical properties reveal its intrinsic characteristics. Its chemical formula is $C_{21}H_{24}O_{10}$, giving it a molecular weight of 436.41 g/mol. The molecule contains 31 heavy atoms, 12 of which are aromatic, and has a Csp3 fraction of 0.38, indicating a notable presence of single bonds in its structure. It has 8 rotatable bonds, 10 hydrogen bond acceptors, and 6 hydrogen bond donors. Its molar refractive index is 105.47, and its total polar surface area (TPSA) is 166.14

Å². In terms of lipophilicity, various indicators were calculated. Log Po/w (iLOGP) is 1.72, while Log Po/w (XLOGP3) is 0.52. Log Po/w (WLOGP) is 0.03, and Log Po/w (MLOGP) has a value of -0.61. Log Po/w (SILICOS-IT) is 0.15, and consensus Log Po/w is 0.36. Water solubility properties indicate that the molecule is soluble. Log S (ESOL), Log S (Ali), and Log S (SILICOS-IT) values are -2.63, -3.58, and -1.97, respectively. Pharmacokinetically, the compound has low gastrointestinal absorption and cannot cross the blood-brain barrier (BBB). It is a P-glycoprotein (P-gp) substrate but does not inhibit CYP1A2, CYP2C19, CYP2C9, CYP2D6, or CYP3A4 enzymes. Skin permeation (Log Kp) is estimated at -8.59 cm/s. With regard to its suitability as a drug candidate, the molecule complies with Lipinski's rule, with the exception of a violation linked to the presence of hydrogen bond donors (NH or OH > 5). It complies with Ghose's rule but violates Veber's criteria due to a high total polar surface area value (TPSA>140). About other drug similarity criteria, it satisfies Egan and Muegge's rules while displaying a bioavailability score of 0.55. In medicinal chemistry, it displays PAINS (chemical substances with reactive motifs) alerts linked to the presence of catechol groups. The molecule is also alerted by Brenk for the same reason. Although it does not meet certain lead likeness criteria (similarity with lead compounds) due to its high molecular weight and the presence of rotating bonds, it has a synthetic feasibility rating of 4.84.

3.7. MD simulation results

The graph above shows the evolution of the Root Mean Square Deviation (RMSD) of the alpha carbon (Cα) atoms of the protein, protein backbone atoms, protein side chains, protein heavy atoms, complexes, and the ligand (**Fig. 15**). All protein structures are first aligned to the protein backbone reference structure, and then the RMSD is calculated based on atom selection. RMSD analysis can provide insight into the structural conformation throughout the simulation and indicate whether the simulation has equilibrated - its fluctuations towards the end of the simulation are around a thermal mean structure. Changes in the order of 1 to 3 Å are perfectly acceptable for small, globular proteins. Higher RMSD changes for ligands H6, H7, and the most active 19 compared to H5 suggest that these complexes undergo conformational changes during the simulation time. The H5_3OYA complex, with an RMSD value close to 7 Å, converges compared to the others, confirming the stability of the H5 ligand structure with a value of less than 1 Å.

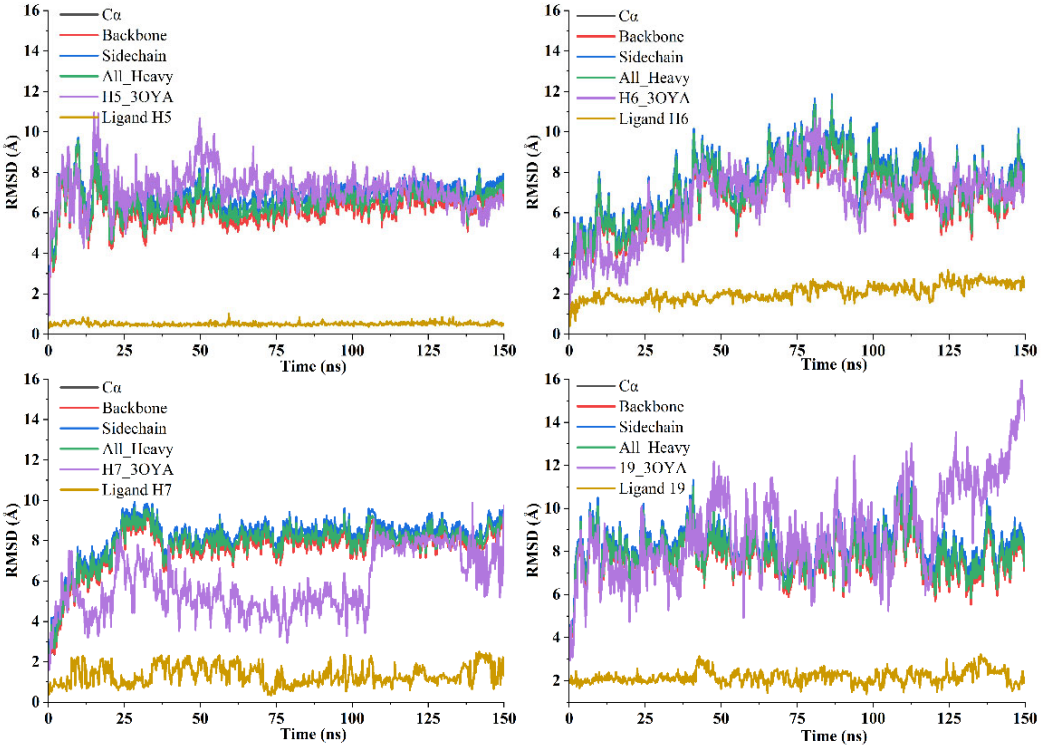


Fig. 15. Root Mean Square Deviation (RMSD) of ligands, proteins, and complexes.

RMSF provides a quantitative measure of atomic fluctuations for alpha carbon (Cα) atoms of the protein, protein backbone, protein side chains, and protein heavy atoms (Fig. 16). RMSF values over a period of 150 ns can help reveal their dynamic behavior, thereby characterizing the molecular flexibility of amino acids. For example, in the H5_3OYA complex, the RMSF values in the range of residues 100 to 275 are less than 3 Å, indicating the stability of these residues during bond formation. Conversely, other residues and complexes generally exhibit values above 3 Å, with limits of 7 Å for H5_3OYA and H6_3OYA complexes and 8 Å for H7_3OYA and 19_3OYA complexes. A higher RMSF value for a particular residue or atom implies greater positional fluctuations during the simulation, indicating increased flexibility in that region.

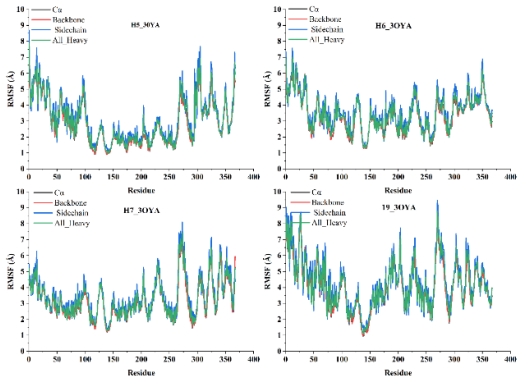


Fig. 16. Mean Square Deviation of Fluctuations (RMSF) of the Atomic Positions of Protein Constituents.

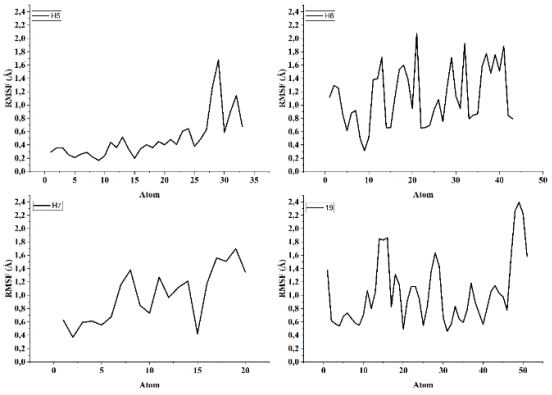


Fig. 17. Mean Square Deviation of Fluctuations of the Atomic Positions of Ligand Constituents.

A number of connections between these groups and neighboring amino acids may have formed as a result of the significant RMSF values for some ligand atoms found when compared to the initial position. Such events could potentially indicate a relaxation of the binding or a conformational change of the ligand (**Fig. 17**). For example, atom number 29 (OH) in ligand H5 has a remarkably high RMSF value compared to other atoms due to its freely rotatable σ bond, which allows it to form a polar interaction with the amino acid GNL:215. Analysis of the RMSF of ligand H5, excluding the hydroxyl group (No. 29), can reveal which atoms or residues of the ligand tend to show fluctuation compared to other ligands. Notable fluctuations in these regions could indicate significant structural changes, possibly aimed at maximizing interactions or sometimes triggering shifts in the mechanism of action. Comparative analysis of the molecular properties of ligands H5, H6, H7, and 19 reveals significant variation among these compounds. The ranges of variation of properties such as radius of gyration (Rg), total molecular surface area (MolSA), solvent accessible surface area (SASA), and total polar surface area (PSA) vary considerably from ligand to ligand. The H5 ligand has an Rg between 4.75 and 5.15 Å, a MolSA of 360 to 380 Å^2, a SASA of 100 to 150 Å^2 and a PSA of 390 to 410 Å^2. In comparison, ligand H6 has similar Rg values (4.74 to 5.6) Å but is characterized by a higher MolSA (425 to 500) Å^2, a SASA of 200 to 400 Å^2 and a PSA between 350 and 400 Å^2. Strikingly, ligand H7 is characterized by a lower Rg (3 to 4) Å, a reduced MolSA (250 to 275) Å^2, a SASA varying from 20 to 320 Å^2 and a PSA of 190 to 210 Å^2. Finally, the most active ligand, ligand 19, has an Rg of 5.4 to 6 Å, a significantly higher MolSA (575 to 625) Å^2, a SASA ranging from 270 to 460 Å^2, and a lower PSA (160 to 180) Å^2. These marked differences in structural and surface properties between the ligands suggest substantial variation in their structures and possible interactions with the target protein. For example, ligands H7 and 19 exhibit particularly diverse SASA and PSA values (**Fig. 18**), which could influence their solubility, solvent accessibility, and potential interaction with protein residues. These observations underscore the importance of these molecular properties in understanding the binding mechanisms and potential biological effects of these ligands.

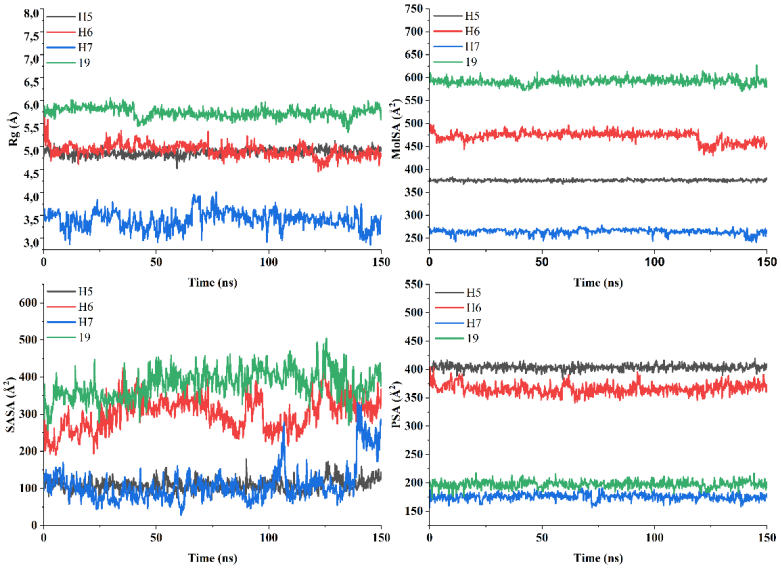


Fig. 18. molecular properties of ligands

The ligand H5 formed five hydrogen bonds (H-bonds) with the amino acids LEU: 113, GLU: 207, PHE: 208, GNL: 215, THR: 351, and SER: 365. These types of bonds play a significant and critical role in drug design due to their strong influence on drug specificity, metabolism, and absorption. This confirms the stability of the H5_3OYA complex. It is closely followed by the H6_3OYA complex, which forms four hydrogen bonds with the amino acids LEU: 113, LYS: 124, GLU: 207, and ASP: 367, with very substantial proportions. On the other hand, the other ligands, H7 and 19, formed multiple hydrogen bonds but with lower intensity.

One ionic interaction (**Fig. 19**) is observed between ligand H6 and amino acid LYS: 219, and two interactions between ligands H7 and 19 with amino acids (ARG: 114, LYS: 219) and (ARG: 114, ARG: 117), respectively. Other hydrophobic interactions and more significant water bridges are formed in all complexes. These results underscore the complex interplay of different types of binding and further emphasize the intricate nature of ligand-protein interactions within these complexes.

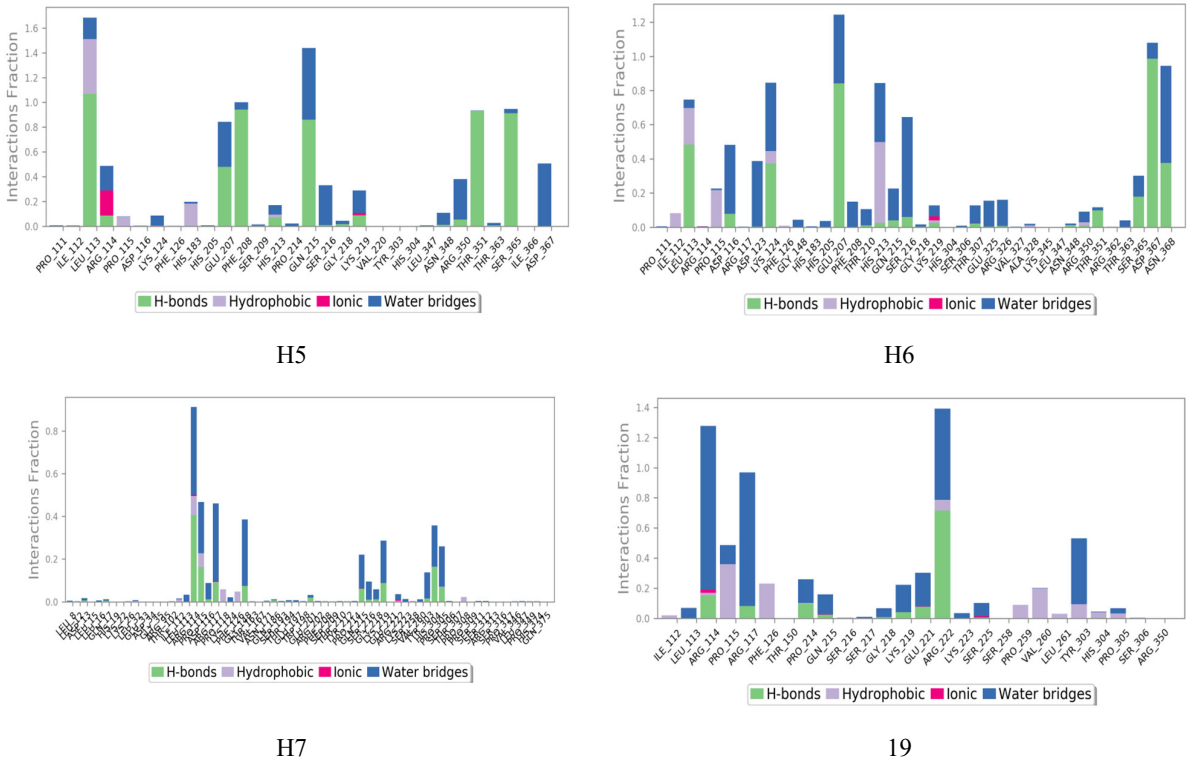


Fig. 19. Bonding Types Include H-Bond Contacts, Hydrophobic Contacts, Ionic Contacts, and Water Bridges

3.8. Comparison of results obtained

The experimental results highlight the diverse activities of compounds H1 to H9 against HIV infection, with kaempferol (H1) and its 3-O-β-D glucopyranoside derivatives (H3 and H6) standing out as the most effective, displaying EC50 values of 4 and 8 respectively on C8166 cells. Quercetin (H2) also showed significant activity with an EC50 of 20, while compound H4 showed no effect even at a high concentration of 250 mg/ml. Compound H8 showed moderate anti-HIV activity. Compounds 1 and 2 are particularly promising, reducing HIV-1IIIB infectivity by more than 99% with EC50s of 0.8 and 10 respectively[71]. The mechanisms of action of the compounds show that kaempferol acts primarily as a viral protease inhibitor, while its derivatives substituted in position 3 modulate the gp120/CD4 interaction without significantly influencing protease activity. This diversity in mechanisms of action highlights the crucial importance of the structure-activity relationship (SAR) in optimizing these compounds as therapeutic agents against HIV. In parallel, the application of machine learning models to predict the pIC50 values of new anti-HIV inhibitors offers a complementary theoretical perspective. The results obtained by different models, such as XGB, LGBM, DT, RF, GB, Bag, ET, and HGB, show a significant correlation with experimental results, validating the predictive efficacy of these approaches. For example, the ET model predicts high pIC50 values for compounds H1, H2 and H3, in line with their high experimental activity. These theoretical predictions reinforce the validity of machine learning models in predicting the biological activity of compounds, providing valuable cross-validation with experimental data. The analysis of molecular interactions by docking enriches our understanding of the mechanisms of action at a structural level. For example, in the H1_3OYA complex, kaempferol (H1) interacts mainly via hydrophobic interactions with the ARG:222 residue, highlighting its central role in the molecular configuration. Hydrogen bonds with ARG residues 114, 117, 223 and 226 also help to stabilize this ligand-protein interaction, enhancing its inhibitory efficacy. Similarly, the H2_3OYA complex shows that quercetin (H2) interacts via hydrophobic forces with residue GLN:215, in addition to multiple hydrogen bonds with other residues, illustrating the

diversity of molecular interactions involved in its mechanism of action. For the H3_3OYA complex, kaempferol 3-O-β-D glucopyranoside (H3) shows specific interactions with residue ILE:112 through hydrophobic interactions and hydrogen bonds with residues ARG:114, 219 and 350, confirming its inhibitory efficacy through distinct mechanisms. By integrating these experimental and theoretical results, this hybrid approach enriches our understanding of the molecular interactions underlying the anti-HIV activities observed. It demonstrates the importance of combining molecular docking data with experimental results to elucidate SAR, thereby guiding the development of new, more effective and targeted antiviral therapies.

4. Conclusion

In this study, eight prediction models (XGB, LGBM, DT, RF, GB, Bag, ET, and HGB) were used to develop and predict anti-HIV activity by relating the activity to the molecular structure of each molecule using machine learning approaches. This was accomplished utilizing a large dataset encompassing 450 experimentally synthesized compounds. 208 descriptors were successfully computed using an automated technique. To develop high-performance models, descriptors with low correlation to activity were excluded using a principal component analysis. The ET model outperformed the others based on statistical characteristics of internal and external validation. Furthermore, blind molecular docking was used to determine the type of amino acids responsible for the biological action of these chemical molecules. The proposed chemicals predicted physical closeness and interaction with mutant HIV-1 IN residues, supporting the coevolution of mutations in drug-resistant viral strains. According to the results of this approach, the affinity values of the proposed ligands are about similar to the most active No. 19. In this investigation, pharmacokinetic and pharmacodynamic features were used to investigate the ADME-Tox profile of Compound No. 19, which had multiple violations. In contrast, suggested molecules H1, H2, and H8 demonstrated favorable pharmacokinetic and pharmacodynamic characteristics. Understanding these qualities is a critical step before moving on to clinical and pre-clinical phases.

The results of dynamic simulation of complexes favored by molecular docking show that H5 and H7 have lower RMSD fluctuations than H6 and Compound No. 19. This shows that the structure of H5 and H7 remains reasonably constant during the 150ns dynamic simulation. Furthermore, hydrogen bond interactions of ligands H5 and H6 are more significant than H7 and No. 19. Overall, these findings add to our understanding of ligand-protein interactions and anti-HIV processes.

Acknowledgments

We are grateful to the ‘Association Marocaine des Chimistes Théoriciens’ (AMCT) for its pertinent help concerning the programmes.

References

[1] Halder AK., Jha T. **(2010)** Validated predictive QSAR modeling of N-aryl-oxazolidinone-5-carboxamides for anti-HIV protease activity. *Bioorg Med Chem Lett.,* 20 (20) 6082–7.
[2] Awi NJ, Teow S. **(2018)** Antibody-Mediated Therapy against HIV/AIDS: Where Are We Standing Now?. *J Pathog.,* (20)1–9.
[3] Su Q, Xu X, Zhou L. **(2008)** QSAR model of triterpene derivatives as potent anti-HIV agents. *Mol Simul.,* 34 (7) 651–9.
[4] Sierawska O., Małkowska P., Taskin C., Hrynkiewicz R., Mertowska P., Grywalska E., Korzeniowski T., Torres K., Surowiecka A., Niedźwiedzka-Rystwej P., Strużyna J. **(2022)** Innate Immune System Response to Burn Damage—Focus on Cytokine Alteration. *Int J Mol Sci*., 23 (2) 117.
[5] Vang R., Shih I. M., Kurman R. J. (**2013**) Fallopian tube precursors of ovarian low- and high-grade serous neoplasms. *Histopathology*., 62 (1) 44–58.
[6] Mcarthur J. C., Steiner J., Sacktor N., Nath A. (**2010**) Human Immunodeficiency Virus- Associated Neurocognitive Disorders Mind the Gap., *Ann. Neurol.* 67 (6) 699–714.
[7] Watkins C. C., Treisman G. J. (**2015**) Cognitive impairment in patients with Aids – Prevalence and severity. *Dove Med Press.,* 7 35–47.
[8] Karkhur S, Hasanreisoglu M, Vigil E, Halim MS, Hassan M, Plaza C., Nguyen N. V., Afridi R., Tran A. T., Do D. V., Sepah Y. J., Nguyen Q. D. (**2019**) Interleukin-6 inhibition in the management of non-infectious uveitis and beyond. *J. Ophthal. Inflamm. Infect.* 9 (1) 17.
[9] Garcia J. C. (**2015**) Review of Radiologic Infectious and Non-infectious Pulmonary Complications in Human Immunodeficiency Virus Patients. *J. Pulm. Respir Med*. 5 (3) 1000260.
[10] Kumar S., Singh S., Luthra K. (**2023**) An Overview of Human Anti-HIV-1 Neutralizing Antibodies against Diverse Epitopes of HIV-1. *ACS Omega*. 8 (8) 7252–61.
[11] Bloch M., John M., Smith D., Rasmussen T.A. (**2020**) Wright E. Managing HIV-associated in fl ammation and ageing in the era of modern ART. *HIV Med*. 21 2–16.
[12] Ensoli B, Moretti S, Borsetti A, Teresa M, Stefano M, Picconi O, Tripiciano A., Sgadar C., Monini P., Cafaro A. (**2021**) New insights into pathogenesis point to HIV - 1 Tat as a key vaccine target. *Arch Virol.* 166 2955–74.
[13] Zhang F., Wang Z., Peijnenburg W. J. G. M., Vijver M. G. (**2023**) Machine learning-driven QSAR models for predicting the mixture toxicity of nanoparticles. *Environ Int*. 177 108025.
[14] Durrant J. D., Amaro R. E. (**2015**) Machine-learning techniques applied to antibacterial drug discovery. *Chem Biol Drug Des.* 85 14–21.
[15] Chen M., Yang X., Lai X., Gao Y. (**2015**) 2D and 3D QSAR models for identifying diphenylpyridylethanamine based inhibitors against cholesteryl ester transfer protein. Bioorg Med Chem Lett. 25 4487–95.

[16] Maltarollo V. G., Gertrudes J.C., Oliveira P. R., Honorio K. M. (**2015**) Applying machine learning techniques for ADME-Tox prediction: A review. *Expert Opin Drug Metab Toxicol*. 11 259–71.
[17] Nongonierma A. B., Fitzgerald R. J. (**2016**) Learnings from quantitative structure-activity relationship (QSAR) studies with respect to food protein-derived bioactive peptides: A review. RSC Adv. 6 75400–13.
[18] Sarfaraz k. N., Zamara M. (**2023**) Recent Advances in Machine-Learning-Based Chemoinformatics : A Comprehensive Review. *Int. J. Mol. Sci.* 24(14) 11488
[19] Ghafourian T., Cronin M. T. D. The impact of variable selection on the modelling of oestrogenicity. *SAR QSAR Environ. Res.* 16 171–90.
[20] Shahlaei M., Madadkar-Sobhani A., Saghaie L., Fassihi A. (**2012**) Application of an expert system based on Genetic Algorithm-Adaptive Neuro-Fuzzy Inference System (GA-ANFIS) in QSAR of cathepsin K inhibitors. *Expert. Syst. Appl.* 39 (6) 6182-6191.
[21] Tirelli T., Pessani D. (**2011**) Importance of feature selection in decision-tree and artificial-neural-network ecological applications. Alburnus alburnus alborella: A practical example. *Ecol. Inform.* 6 (5) 309–315.
[22] Izza Y., Ignatiev A., Silva J. M. (**2021**) On Explaining Decision Trees To cite this version: HAL Id: hal-03312480 pp:0–21.
[23] Pandya, P. N., Kumar, S. P., Bhadresha, K., Patel, C. N., Patel, S. K., Rawal, R. M., & Mankad, A. U. (**2020**) Identification of promising compounds from curry tree with cyclooxygenase inhibitory potential using a combination of machine learning, molecular docking, dynamics simulations and binding free energy calculations. *Mol. Simul.* 46 (11) 812–822.
[24] Qi A., Zhao D., Yu F., Asghar A., Wu Z., Cai Z., Alenezi F., Mansour R. F., Chen H., Chen M. (**2022**) Directional mutation and crossover boosted ant colony optimization with application to COVID-19 X-ray image segmentation. *Comput. Biol. Med.* 148 105810.
[25] Yadav A. K., Singh T. R. (**2021**) Novel inhibitors design through structural investigations and simulation studies for human PKMTs (SMYD2) involved in cancer. *Mol. Simul.* 47 (14) 1149–1158.
[26] Toropov A. A., Toropova A. P., Carnesecchi E., Benfenati E., Dorne J. L. (**2020**) The index of ideality of correlation and the variety of molecular rings as a base to improve model of HIV-1 protease inhibitors activity. *Struct. Chem.* 31 1441–8.
[27] Liu Y., Liu Y., Wang S., Zhu X. (**2023**) LBCE-XGB: A XGBoost Model for Predicting Linear B-Cell Epitopes Based on BERT Embeddings. *Interdiscip. Sci.* 15 293–305.
[28] Chen T., Guestrin C. (**2016**) XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13 (17) 785–94.
[29] Sikander R., Ghulam A., Ali F. (**2022**) XGB-DrugPred: computational prediction of druggable proteins using extreme gradient boosting and optimized features set. *Sci. Rep.* 12 5505.
[30] Huang B., Wang C. (**2024**) Retraction Note: Research on Data Analysis of Efficient Innovation and Entrepreneurship Practice Teaching Based on LightGBM Classification Algorithm. *Int. J. Comput. Intell. Syst*. 17 (52).
[31] Ju Y., Sun G., Chen Q., Zhang M., Zhu H., Rehman M. U. (**2019**) A model combining convolutional neural network and lightgbm algorithm for ultra-short-term wind power forecasting. *IEEE Access*. 7 28309 – 28318.
[32] Zhang D., Gong Y. (**2020**) The Comparison of LightGBM and XGBoost Coupling Factor Analysis and Prediagnosis of Acute Liver Failure. *IEEE Access*. 8 220990 - 221003.
[33] Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu. Y. (**2023**) LightGBM : A Highly Efficient Gradient Boosting Decision Tree To cite this version : *HAL Id* : hal-03953007: 0–9.
[34] Mukherjee K., Colón Y. J. **(2021**) Machine learning and descriptor selection for the computational discovery of metal-organic frameworks. *Mol. Simul.* 47(10–11) 857–877.
[35] AlKheder S., AlOmair A. (**2022**) Urban traffic prediction using metrological data with fuzzy logic, long short-term memory (LSTM), and decision trees (DTs). *Nat. Hazards.* 111 1685–1719.
[36] Aher R. B., Khan K., Roy K. (**2020**) A brief introduction to quantitative structure-activity relationships as useful tools in predictive ecotoxicology. *SAR QSAR Environ Res.* 27–53.
[37] Calle L., Marrero-Ponce Y., Mora J. R. (**2021**) Molecular simulation of the (GPx)-like antioxidant activity of ebselen derivatives through machine learning techniques. *Mol. Simul.* 47 1402–10.
[38] Johnston A., Johnston B. F., Kennedy A. R., Florence A. J. (**2008**) Targeted crystallisation of novel carbamazepine solvates based on a retrospective Random Forest classification. *Cryst. Eng. Comm.* 10 23–5.
[39] Deepak T., Mohd Anul H., Gazi R., Prashant B., Joydip D. (**2019**) Comparison of Performance of Artificial Neural Network (ANN) and Random Forest (RF) in the Classification of Land Cover Zones of Urban Slum Region. *LNCE*. 51 225–36.
[40] Ferreira Neto, D. C., Alencar Lima, J., Sobreiro Francisco Diz de Almeida, J., Costa França, T. C., Jorge do Nascimento, C., & Figueroa Villar, J. D. (**2018**) New semicarbazones as gorge-spanning ligands of acetylcholinesterase and potential new drugs against Alzheimer's disease: Synthesis, molecular modeling, NMR, and biological evaluation. *J. Biomol. Struct. Dyn.* 36 (15) 4099–4113.
[41] Danush S., Dutta A. (**2023**) Machine learning-based framework for predicting toxicity of ionic liquids. *Mater Today Proc*. 72 (1) 75–80.
[42] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (**2017**). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Neural Information Processing Systems.
[43] Xu Y., Liaw A., Sheridan R. P., Svetnik V. (**2023**) Development and Evaluation of Conformal Prediction Methods for QSAR. q-bio.BM. https://doi.org/10.48550/arXiv.2304.00970.
[44] Vishwakarma G., Sonpal A., Hachmann J. (**2021**) Metrics for Benchmarking and Uncertainty Quantification: Quality, Applicability, and Best Practices for Machine Learning in Chemistry. *Trends Chem*. 3 146–56.
[45] Shahhosseini M., Hu G., Pham H. (**2022**) Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. *Machine Learning with Applications* 7 100251.
[46] Noviandy T. R., Maulana A., Emran T. B., Idroes G. M., Idroes R. (**2023**) QSAR Classification of Beta-Secretase 1 Inhibitor Activity in Alzheimer's Disease Using Ensemble Machine Learning Algorithms. *HJAS*. 1.
[47] Lusci A., Pollastri G., Baldi P. (**2013**) Deep architectures and deep learning in chemoinformatics: The prediction of aqueous solubility for drug-like molecules. *J. Chem. Inf. Model*. 53 1563–75.
[48] Rao, H., Zeng, X., Wang, Y., He, H., Zhu, F., Li, Z., & Chen, Y. (**2012**). Identification of DNA adduct formation of small molecules by molecular descriptors and machine learning methods. *Mol. Simul.* 38 (4) 259–273.

[49] Liu X, Zhu B, Dai XW, Xu ZA, Li R, Qian Y, Lu Y. P. Zhang Y., Liu Y., Zheng J. (**2023**) GBDT_KgluSite: An improved computational prediction model for lysine glutarylation sites based on feature fusion and GBDT classifier. *BMC Genomics* 24.
[50] Li N., Chen K., Bai J., Geng Z., Tang Y., Hou Y., Fan F., Ai X., Hu Y., Meng X., Wang X., Zhang Y. (**2021**) Tibetan medicine Duoxuekang ameliorates hypobaric hypoxia-induced brain injury in mice by restoration of cerebrovascular function. *J. Ethnopharmacol.* 270 113629.
[51] Hamada M., Zaqoot H. A., Sethar W. A. (**2023**) Using Supervised Learning Machine Approach to Predict Water Quality at Gaza's Wastewater Treatment Plant. *Environmental Science: Advances*. https://doi.org/10.1039/d3va00170a.
[52] Binuya M. A. E, Engelhardt E. G., Schats W., Schmidt M. K., Steyerberg E. W. (**2022**) Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. *BMC Med. Res. Methodol.* 22 316.
[53] Ouabane M., Tabti K., Hajji H., Elbouhi M., Khaldan A., Elkamel K., Sbai A., Ajana M. A., Sekkate C. Bouachrine M. Lakhlifi T. (**2023**) Structure-odor relationship in pyrazines and derivatives: a physicochemical study using 3D-QSPR, HQSPR, Monte Carlo, Molecular Docking, ADME-Tox and. Arabian Journal of Chemistry 16 (11) 105207.
[54] Ouabane M., Hajji H, Belhassan A. Koubi Y., Elbouhi M., Badaoui H., Sekkate C., Lakhlifi T. (**2022**) RHAZES: Green and Applied Chemistry molecules in pectin gels of different concentration. *RHAZES: Green and Applied Chemistry* 14 15-35.
[55] Luo X., Yang X., Qiao X., Wang Y., Chen J., Wei X., Willie J. G. M. P. (**2017**) Development of a QSAR model for predicting aqueous reaction rate constants of organic chemicals with hydroxyl radicals. *Environ Sci Process Impacts* 19 350–6.
[56] Todeschini R., Ballabio D., Grisoni F. (**2016**) Beware of Unreliable Q2! A Comparative Study of Regression Metrics for Predictivity Assessment of QSAR Models. *J. Chem. Inf. Model.* 56 (10) 1905-1913.
[57] Zou G. Y. (**2007**) Toward Using Confidence Intervals to Compare Correlations. *Psychol Methods* 12 (4) 399-413.
[58] Mulaik S. A., Raju N. S., Harshman R. A. (**2016**) There is a time and a place for significance testing. What If There Were No Significance Tests. Classic Edition 2016: 61–106.
[59] Hajji H, En-nahli F, Aanouz I, Zaki H., Lakhlifi T., Ajana M. A., Bouachrine M. (**2021**) Catastrophic Collision Between Obesity and COVID-19 Have Evoked the Computational Chemistry for Research in Silico Design of New CaMKKII Inhibitors Against Obesity by Using 3D-QSAR, Molecular Docking, and ADMET. *Orbital* 13 (4) 316-327.
[60] Zhao D., Zhong S. (**2021**) Binding mechanisms of varic acid inhibitors on protein tyrosine phosphatase 1B and in silico design of the novel derivatives. *Mol. Simul*. 47 (9) 771–784.
[61] Wichur T., Pasieka A., Godyń J., Panek D., Góral I., Latacz G., et al. Discovery of 1-(phenylsulfonyl)-1H-indole-based multifunctional ligands targeting cholinesterases and 5-HT6 receptor with anti-aggregation properties against amyloid-beta and tau. *Eur. J. Med. Chem.* 225 113783.
[62] Belghalia E., Ouabane M., El Bahi S., Muzzammel H., Sbai A., Lakhlifi T., Bouachrine M. (**2023**) In silico research on new sulfonamide derivatives as BRD4 inhibitors targeting acute myeloid leukemia using various computational techniques including 3D-QSAR, HQSAR, molecular docking, ADME /Tox, and molecular dynamics. *J. Biomol. Struct. Dyn.* 0 1–19.
[63] Tabti K., Elmchichi L., Sbai A., Maghat H., Bouachrine M., Lakhlifi T. (**2022**) Molecular modelling of antiproliferative inhibitors based on SMILES descriptors using Monte-Carlo method, docking, MD simulations and ADME/Tox studies. *Mol. Simul*. 48 (17) 1575–1591.
[64] Ouabane M., Zaki K., Tabti K., Alaqarbeh M., Sbai A., Sekkate C., Bouachrine M., Lakhlifi T. (**2024**) Molecular toxicity of nitrobenzene derivatives to Tetrahymena pyriformis based on SMILES descriptors using Monte Carlo, Docking, and MD simulations. *Comput. Biol. Med.*: 169 107880.
[65] Calugi L., Sautariello G., Lenci E., Mattei M. L., Coppa C., Cini N., Contini A., Trabouchi A. (**2023**) Identification of a short ACE2-derived stapled peptide targeting the SARS-CoV-2 spike protein. *Eur. J. Med. Chem.* 249 (2023) 115118.
[66] Berenger F., Kumar A., Zhang K. Y. J., Yamanishi Y. (**2021**) Lean-Docking: Exploiting Ligands' Predicted Docking Scores to Accelerate Molecular Docking. *J. Chem. Inf. Model.* 61 2341–52.
[67] Alnajjar R., Mostafa A., Kandeil A., Al-Karmalawy A. A. (**2020**) Molecular docking, molecular dynamics, and in vitro studies reveal the potential of angiotensin II receptor blockers to inhibit the COVID-19 main protease. *Heliyon* 6 e05641.
[68] Alnajjar R., Mohamed N., Kawafi N. (**2021**) Bicyclo[1.1.1]Pentane as Phenyl Substituent in Atorvastatin Drug to improve Physicochemical Properties: Drug-likeness, DFT, Pharmacokinetics, Docking, and Molecular Dynamic Simulation. *J. Mol. Struct.* 1230 129628.
[69] Martyna G. J., Klein M. L., Tuckerman M. (**1992**) Nosé-Hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.* 97 2635–43.
[70] El-Masry R. M., Al-Karmalawy A. A., Alnajjar R., Mahmoud S. H., Mostafa A., Kadry H. H, Abou-Seri S. M., Taher A. T. (**2022**) Newly synthesized series of oxoindole–oxadiazole conjugates as potential anti-SARS-CoV-2 agents: in silico and in vitro studies. *New J. Chem*. 46 5078-5090.
[71] Mahmoud N., Piacente S., Pizza C., Burke A., Khan A. I., Hay A. J. (**1996**) The Anti-HIV Activity and Mechanisms of Action of Pure Compounds Isolated from *Rosa damascene. Biochem. Biophys. Res. Commun.* 229, 73-79.