# Novel machine learning approach toward classification model of HIV-1 integrase inhibitors

Tieu-Long Phan[a, b] , Hoang-Son Lai Le[c], Gia-Bao Truong[c], The-Chuong Trinh[d], Van-Thinh To[c], Phuoc-Chung Van Nguyen[c], Thanh-An Phamc, Tuyen Ngoc Truong[*c]

HIV-1 (Human immunodeficiency virus-1) has been causing severe pandemics by attacking the immune system of its host. Left untreated, it can lead to AIDS (acquired immunodeficiency syndrome), where death is inevitable due to opportunistic diseases. Therefore, discovering new antiviral drugs against HIV-1 is crucial. This study aimed to explore a novel machine learning approach to classify compounds that inhibit HIV-1 integrase and screen the dataset of repurposing compounds. The present study had two main stages: selecting the best type of fingerprint or molecular descriptor using the Wilcoxon signed-rank test and building a computational model based on machine learning. In the first stage, we calculated 16 different types of fingerprint or molecular descriptors from the dataset and used each of them as input features for 10 machine-learning models, which were evaluated through cross-validation. Then, a meta-analysis was performed with the Wilcoxon signed-rank test to select the optimal fingerprint or molecular descriptor types. In the second stage, we constructed a model based on the optimal fingerprint or molecular descriptor type. This data followed the machine learning procedure, including data preprocessing, outlier handling, normalization, feature selection, model selection, external validation, and model optimization. In the end, an XGBoost model and RDK7 fingerprint were identified to be the most suitable. The model achieved promising results, with an average precision of 0.928 ± 0.027 and an F1-score of 0.848 ± 0.041 in cross-validation. The model achieved an average precision of 0.921 and an F1-score of 0.889 in external validation. Molecular docking was performed and validated by redocking for docking power and retrospective control for screening power, with the AUC metrics being 0.876 and the threshold being identified at –9.71 kcal/mol. Finally, 44 compounds from DrugBank repurposing data were selected from the QSAR model, then three candidates were identified as potential compounds from molecular docking, and PSI-697 was detected as the most promising molecule, with in vitro experiment being not performed (docking score: -17.14 kcal/mol, HIV integrase inhibitory probability: 69.81%)

## 1. Introduction

According to the UNAIDS 2021 statistics (United Nations Joint Programme on HIV/AIDS)[1], there were more than 38.4 million people worldwide living with HIV. Since HIV was first discovered in 1980s, the disease has caused about 34.7 million deaths. However, there are no specific drugs or vaccines, so individuals living with HIV can only be treated with antiviral therapy like antiretroviral drug (ARV), suppressing symptoms and slowing down the process leading to AIDS. Following several HIV-1 treatment regimens, clinical therapy should incorporate multiple ARV drugs to ensure the antiviral effect and reduce the risk of drug resistance. Therefore, many ARV drugs have been studied and developed, including reverse transcriptase inhibitors comprising both nucleoside[2] and non-nucleoside inhibitors[3], protease inhibitors[4], integrase inhibitors[5] and fusion inhibitors[6]. Regarding protein targets, the integrase (IN) enzyme stands as a prominent target for medicinal chemistry researchers[7]. This enzyme is produced by a virus with reverse transcription, a process in which viral nucleic acids are catalyzed to form covalent bonds between its genetic information with the DNA (deoxyribonucleic acid) of the host's infected cells[8]. Thus, inhibition of integrase during strand transfer can prevent viral proliferation, and therefore, prolonging the host's lifetime. These inhibitory compounds are called integrase strand transfer inhibitors (INSTIs), and are often in combination IN inhibitors with other HIV medicines to mitigate drug resistance[9].

Machine learning has revolutionized many fields, including drug discovery. In this field, AI (artificial intelligence) has been used to create predictive models for ADMET, drug response[10], toxicity[11], and anticancer activity[12]. These models allow virtual screening and prediction of compound activity[13]. Several studies have been conducted on building models based on machine learning to predict the activity of compounds against IN, including those by Kurczyk A et al (2015), Li Y et al (2017), and Lucas A et al (2022)[14-16]. In contrast to traditional QSAR models[17], which solely focused on linear equations to correlate the molecular descriptor with biological activities, the machine learning approach enables the implementation of non-linear models for QSAR.

This study aimed to discover promising candidates for organic synthesis plans by building computational models and used those models to screen the dataset of repurposing compounds from DrugBank database for potential IN inhibitors. In this study, a novel approach was taken to select fingerprint or descriptor using the Wilcoxon signed-rank test[18]. Likewise, the model selection process for determining the optimal model was conducted differently from the approach used by Li Y et al. (2017)[14]. The study of Li Y et al. utilized ECFP_4 fingerprint as the model input along with Support Vector Machine, Decision

[a.] *Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, Universität Leipzig, Härtelstraße 16-18, 04107, Leipzig, Germany.*
[b.] *Department of Mathematics and Computer Science, University of Southern Denmark, Odense M DK-5230, Denmark.*
[c.] *University of Medicine and Pharmacy, Ho Chi Minh city, Falcuty of Pharmacy, Ho Chi Minh city, 700000, Vietnam.*
[d.] *Faculty of Pharmacy, Grenoble Alpes University, La Tronche, 38700, France.*
[*c]Corresponding Author.
E-mail address: truongtuyen@ump.edu.vn

Tree, Function Tree, and Random Forest for machine learning model.

## 2. Methods

This study was carried out using Python 3.8 with AMD Ryzen 9 3900X CPU core consisting of 12 processors, 3.79 GHz processor speed, 500 GB memory, 96.0 GB RAM operating on Linux 22.04. Molecular descriptor and fingerprint were generated via open-source packages, including Padel 2.21 Descriptor[14], RDKit 2020.3.1[19], Mordred 1.2.0[20], Map4 1.0[21], and MHFP[22]. The machine learning model was completed using Scikit-learn 1.1.1 library[23] with the steps described in the diagram below (Fig 1). All stages of the study were conducted with the same random state (value = 42) to ensure reproducibility. The source code, all datasets, and the results of this study are available at: https://github.com/Medicine-Artificial-Intelligence/HIV_IN_Classification

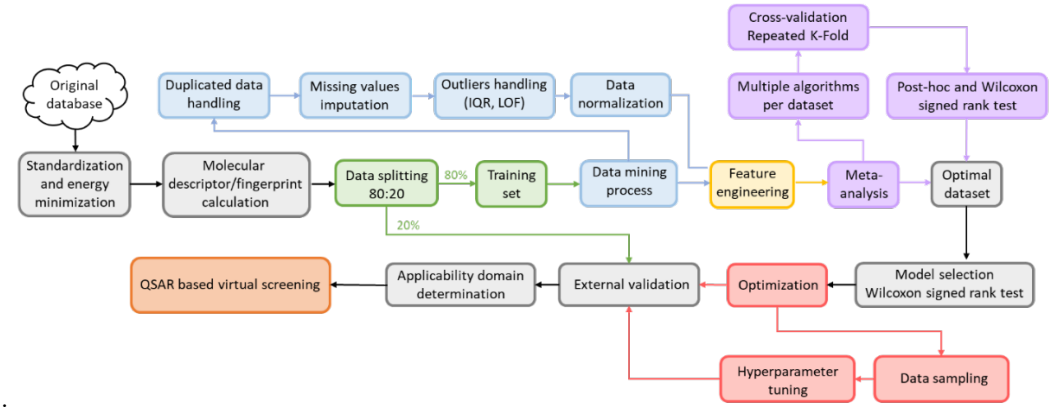


**Figure 1.** Model development pipeline includes two stages: dataset selection and model establishment.

### 2.1. Dataset

8979 molecules inhibiting HIV-1 integrase were collected from the ChEMBL 33 database. Biological activity standardization was performed, including target organism being "Human immunodeficiency virus 1", assay type being "B type", and "IC50" value in measurement unit columns, followed by canonicalizing SMILES structures. Upon completion of these stages, 2834 compounds remained in the dataset for building the model. Besides, 15235 structures collected from the DrugBank database were prepared for virtual screening, with repurposing and repositioning strategies.

### 2.2. Optimal dataset selection

SMILES structures were converted into molecules to calculate 16 types of fingerprints or descriptors, called molecular features, including 3D-Mordred, RDKit descriptor, Mol2vec, MACCS, PubChem, Avalon, ECFP2, ECFP4, ECFP6, RDK5, RDK6, RDK7, Cats2D, 2D-Pharmacophore Gobbi (Ph4), MAP4, and SECFP (raw_data_features). The data preparation process was performed for all molecular features set, including target normalization with the threshold of pIC50 being 7 (meaning active or class 1 if pIC50 is equal or above 7, and inactive or class 0 for the counterpart), dataset division (80:20) with stratification principle resulted in 1995 compounds in the training set and 499 compounds in the external validation set with the imbalance ratio of 0.404 between the active and inactive classes.

Then, the data mining process was conducted on the training set and applied similar methods for the external validation set. First, 1995 compounds with molecular features underwent data removal to eliminate duplicate rows and columns, followed by missing values handling utilizing KNNImputer from the Scikit-learn library (only for the 3D-Mordred dataset) before ending up with low variance removal using a threshold being 0.05. Next, Local Outlier Factor (LOF) was employed with a parameter setting of "n_neighbors = 20" to remove outliers in the training set and novelty in the external validation set. The final step in the data mining process is data normalization using a rescaling method, in which MinMaxScaler was applied to map all data into the range [0,1].

In the next step, feature extraction was conducted with feature importance algorithms derived from the Random Forest algorithm. Then, 10 different machine learning models were applied for all molecular features, including logistic regression (Logic), k-nearest neighbor (KNN), support vector machine (SVM), random forest (RF), extra tree (ExT), Adaboost (ADA), Gradient Boosting (Grad), XGBoost (XGB), CatBoost (Catbst), and Multilayer Perceptron (MLP). The Wilcoxon signed-rank test conducted a meta-analysis to select the optimal type of feature set. This was based on 10-fold internal cross-validation repeated three times (3x10 RepeatedStratifiedKFold) with the evaluation metric being F1-score. The outcome yielded the optimal feature set that best suited the model.

### 2.3. Machine learning model development

The most effective molecular features set selected above underwent similar data processing and mining steps but experienced a slight difference in the feature extraction stage. Instead of just using Random Forest to select essential features, eight different methods, consisting of Chi-squared (Chi2), Mutual information (Mutual), Random Forest (RF), Extra Tree (ExT), Adaboost (ADA), Gradient Boosting (Grad), XGBoost (XGB), and Logistic Regression (Logic), were performed to compare the performance of these models with baseline model,

which was not performed feature extraction. The optimal algorithm was then used to reduce the dimension of the data.

The reduced-dimensional data was performed for model selection to select an optimal algorithm for model development. Ten different algorithms were selected for this step, including logistic regression (Logic), k-nearest neighbor (KNN), support vector machine (SVM), random forest (RF), extra tree (ExT), Adaboost (ADA), Gradient Boosting (Grad), XGBoost (XGB), CatBoost (Catbst), and Multilayer Perceptron (MLP). 3x10 RepeatedStratifiedKFold cross-validation with Wilcoxon signed-rank test[24] were also performed for these two stages.

Moreover, the Tree-structured Parzen Estimator (TPE) algorithm from the Optuna library was implemented for Bayesian Optimization (BO) to optimize the hyperparameter. BO is a global optimization technique that builds a surrogate model of the objective function and uses an acquisition function to suggest the next sample point[25, 26].

Finally, the applicability domain (AD) of this model was developed based on the local density deviation of a given compound with respect to its neighbors, from which the local outlier factor (LOF) algorithm was implemented to identify "novelty" data or data out of the AD.

### 2.4. Model evaluation

The performance of a model can be evaluated by its learnability from data and generalizability on unseen datasets, performed through internal and external validation, respectively. Internal validation (IV) involves cross-validation techniques for training models and hyperparameter tuning. External validation (EV), on the other hand, utilizes a validation dataset from an independent source to assess the model's performance unbiasedly. As such, the results of EV provide crucial evidence for the generalizability of a QSAR model[27].

The models' performance in this study were evaluated using statistical parameters such as F1-score, average precision, precision, and recall. Precision is calculated as the ratio of true positive predictions to the sum of true positive and false positive predictions[28].

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}$$

Recall is a statistical measure that quantifies the proportion of true positive instances that are correctly identified by a predictive model[28].

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}$$

Average Precision (AP) is calculated as the weighted mean of precision at each threshold, the weight is the increase in recall from the prior threshold[29].

$$\text{AP} = \sum_{n} (\text{Recall}_n - \text{Recall}_{n-1}) \times \text{Precision}_n$$

The F1-score is calculated as the harmonic mean of precision and recall, providing a measure of the trade-off between them[28].

$$\text{F1} - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

### 2.5. Molecular docking

In this study, the structure of HIV Integrase was retrieved from the Protein Data Bank (PDB ID: 6PUW)[30], and underwent the standardization step, including missing residues restoration, local energy minimization. Then, MGLTools was utilized to add hydrogens atom to structure, including both polar hydrogens and non-polar hydrogens, as well as the Gasteiger charges. Finally, the gridbox was defined as a cube of 60 × 60 × 60 grid points, with coordinates x = 143.399 Å, y = 159.402 Å, and z = 177.382 Å.

The performance of the docking model was validated by redocking to validate docking power (RMSD ≤ 2 Å) and retrospective control (enrichment analysis) to validate screening power. DeepCoy was utilized to generate decoy (active: decoy = 1:50) for the latter step[31], and the performance was measured by the receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC). Additionally, the Geometric Mean (G-Mean) was employed to determine the optimal cut-off point for the ROC curve[32]. The G-Mean is a metric measuring the balance between classification performance for majority and minority classes.

## 3. Results

### 3.1. Molecular features set selection

The meta-analysis utilized the Wilcoxon signed-rank test to identify the optimal features set. The F1-score was the primary metric to compare the performance among 16 molecular feature sets. The evaluation results are detailed in S1 Table (Supporting information), while the summarized comparison is illustrated in Fig 2.

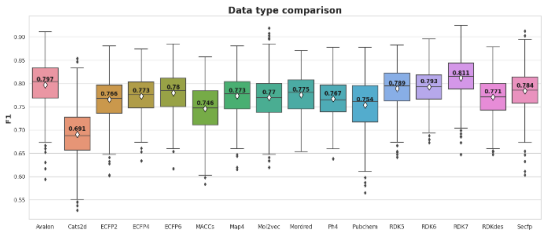


**Figure 2.** The meta-analysis of 16 types of fingerprints and descriptors utilizing F1-score metric

From Fig 2, the RDK7 fingerprint experienced the highest average F1-score in accordance with cross-validation (0.811), calculated from 10 models with 300 observations the whole. Meta-analysis was also conducted for pairwise comparison among fingerprints and descriptors using the Wilcoxon signed-rank test, illustrated in S2 Table (Supporting information). As shown in Fig 3, RDK7 showed a statistically higher significance of average F1-score compared to other datasets ($p < 0.05$).

Therefore, RDK7 was selected as the optimal molecular feature to develop a machine-learning model.

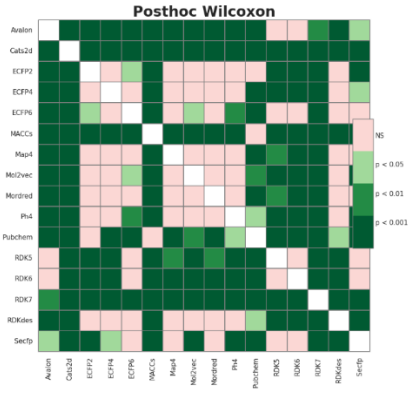


**Figure 3.** The heatmap illustrated the Wilcoxon signed-rank test 16 types of fingerprints and descriptors for meta-analysis

### 3.2. Model selection

#### Feature extraction

In this stage, our main objective was to select the optimal subset of RDK7 fingerprint, building the prediction model based on two criteria. Firstly, the model used for selection had to achieve the highest average F1-score in cross-validation with significant differences based on the Wilcoxon signed-rank test. Secondly, the model used for feature selection should yield the result with the minimum number of fingerprints to accelerate the optimization step. Fig 4 illustrates the results of internal cross-validation among the fingerprint selection methods.

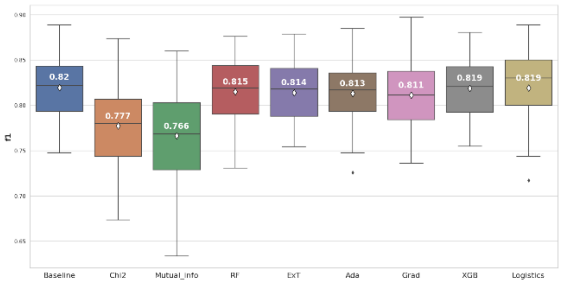


**Figure 4.** The feature extraction methods comparision for RDK7 dataset

According to the box and whisker plot in Fig 4, feature selection methods were stable except for mutual information, Logistics, and AdaBoost returning F1-score outliers after 30 times cross-validation. While using the Wilcoxon signed-rank test for F1-score comparison among 8 models, only the chi-squared and mutual information gave statistically significantly lower results than the baseline model ($p < 0.05$). Other models had no statistically significant difference compared to the baseline model, so the feature extraction methods did not meet the first criterion.

On the other hand, the XGBoost and the Logistic Regression achieved the highest average F1-score among all the models, except for the baseline model. However, according to pairwise assessment of these two models (S3 Figure Supporting information), there was no statistically significant difference ($p > 0.05$). The second criterion, aimed at reducing computational resources by minimizing the number of features, was taken into consideration. The XGBoost algorithm had 533 bits, a lower number of features than the Logistic Regression algorithm, which had 744 bits. Therefore, the XGBoost algorithm was selected to reduce the dimension of the RDK7 dataset. The results of the features selection comparison are illustrated in S4 Table (Supporting information).

#### Machine learning model selection

Ten different machine learning models were employed, and internal cross-validation along with the Wilcoxon signed-rank test was utilized to identify the most efficient machine learning model based on two criteria. The first criterion focused on selecting the model with an average F1-score derived from cross-validation that was significantly higher than the other models. As for the second criterion, the model with an average precision (AP) derived from cross-validation showed significantly higher performance compared to the other models and required a shorter training time.

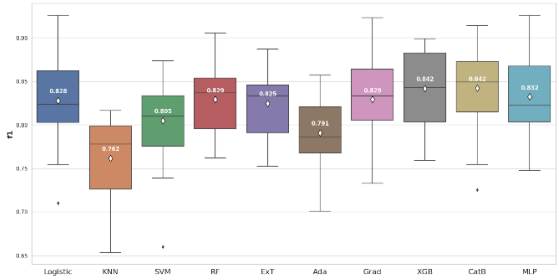


**Figure 5.** The machine learning algorithms comparision for RDK7 dataset utilizing F1-score

Based on the box and whisker plot in Fig 5, XGBoost ($0.842 \pm 0.039$), and CatBoost ($0.842 \pm 0.044$) achieved the highest average F1-score. However, when the Wilcoxon signed-rank test was applied, these differences were not statistically significant ($p > 0.05$) compared to Logistic Regression, Random Forest, Gradient Boosting and Multilayer Perceptron (Fig 6).

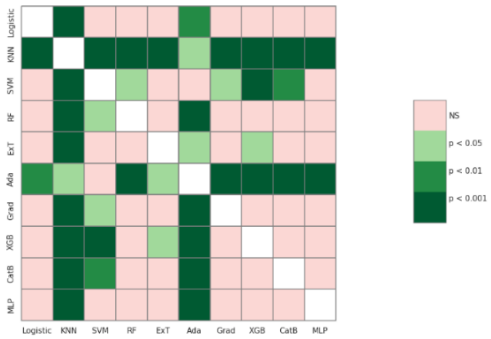


**Figure 6.** The Wilcoxon signed-rank test compared 10 machine learning models using F1-score

We continued to use average precision (AP) to evaluate model performance (Fig 7). Two models including XGBoost ($0.929 \pm 0.029$), and CatBoost ($0.927 \pm 0.030$) remained the best performers. In this case, XGBoost showed significant difference when being compared to most of the others ($p < 0.05$). In addition, XGBoost had a shorter training time than CatBoost. Thus, in terms of the second criterion for this step, XGBoost model was selected for the optimization.

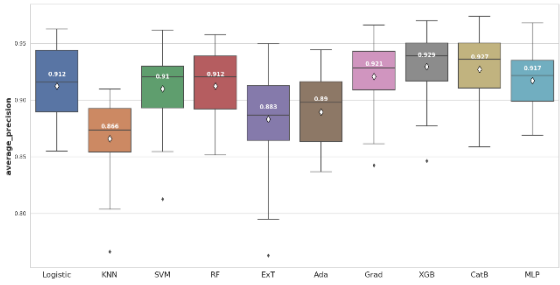


**Figure 7.** The machine learning algorithms comparision for RDK7 dataset utilizing average precision score

## Machine learning model optimization

In the study, hyperparameters were optimized using Bayesian Optimization through 1000 trials. After analyzing the results, it was found that the highest average F1-score across all trials in the cross-validation was 0.854. Therefore, the hyperparameters associated with this trial were selected to be used for the XGBoost model. The results of this step were shown in S5 Table (Supporting informaiton).

The results derived from the cross-validation and external validation were consistent. This external validation result was highly generalizable and can be applied in virtual screening.

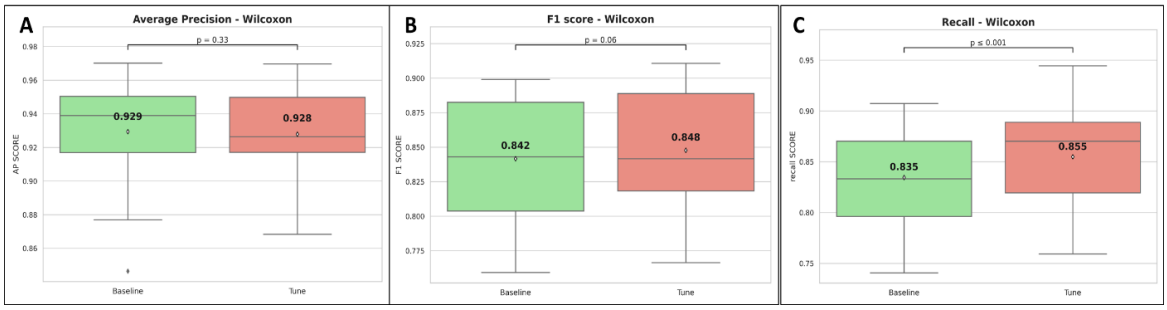


**Figure 8.** The internal cross-validation results of the model, both before and after hyperparameter optimization, (A) Average precision, (B) F1-score, (C) Recall.

## Evaluating the generalizability of the model

The external validation dataset (20%) divided from the beginning was used to evaluate the generalizability of the model. The results were illustrated in Table 1.

**Table 1.** Internal and external validation results for the Gradient Boosting model

| | Cross-validation | | | External validation | | |
|---|---|---|---|---|---|---|
| | AP | F1 | Recall | AP | F1 | Recall |
| **Baseline** | $0.929 \pm 0.030$ | $0.842 \pm 0.038$ | $0.835 \pm 0.045$ | 0.921 | 0.856 | 0.878 |
| **Optimize** | $0.928 \pm 0.027$ | $0.848 \pm 0.041$ | $0.855 \pm 0.052$ | 0.921 | 0.889 | 0.921 |

According to Fig 8, the CV-recall values are significantly higher after optimization compared to the default hyperparameter model, with a p-value ≤ 0.01. However, the average CV-AP and CV-F1 scores do not show a statistically significant improvement after optimization, with p-values of 0.33 and 0.06 respectively.

Compared with the study of Lucas A et al., a state-of-the-art machine learning model targeting HIV integrase utilizing Mordred descriptor, our model could not outperform in external validation, with F1-score being 0.89 lower than the 0.93 of their study. However, our model development procedure is more rigid, with several decision-making stages, supported by the Wilcoxon signed-rank test of cross-validation. Moreover, our study performed cross-validation in the development pipeline for selection and optimization. At the same time, external validation was conducted in the final stage to prove the generalization of the machine learning model. The applicability domain was also investigated to remove five substances in the external validation set to ensure the interpolation of our model (Fig 9). From Fig 9, the red point was detected as a novelty, or outside applicability domain, which was far from the training set (grey points). This could solve the problem of sparse space in the bounding box approach.

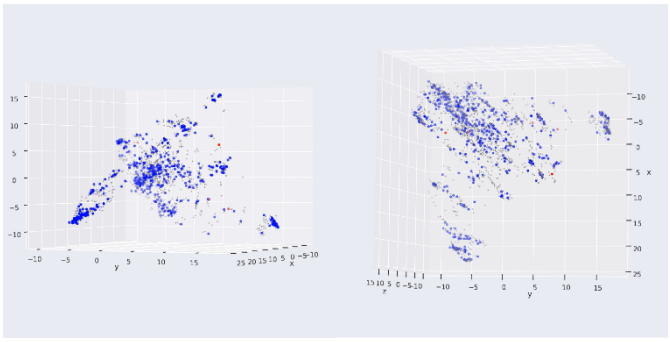

**Figure 9.** Application of LOF in applicability domain. The grey, blue, and red points describes the training set, external validation set in and out the applicability domain.

The model evaluation was conducted based on redocking data collected from Autodock-GPU [33]. Based on the largest cluster of the redocking procedure, which comprised approximately 20% of all generated conformations, the RMSD value of the best-docked conformation (the most negative) did not exceed 2 Å (0.63 Å), which is illustrated in Fig 10A.

### 3.3. Molecular doking

According to the AUC-ROC curve in Fig 10B, the AUC value was 0.876 for the most negative conformation (ad_gpu_min) with the G-mean value reaching 0.841. The docking threshold extrapolated from the G-mean was -9.71 kcal/mol.

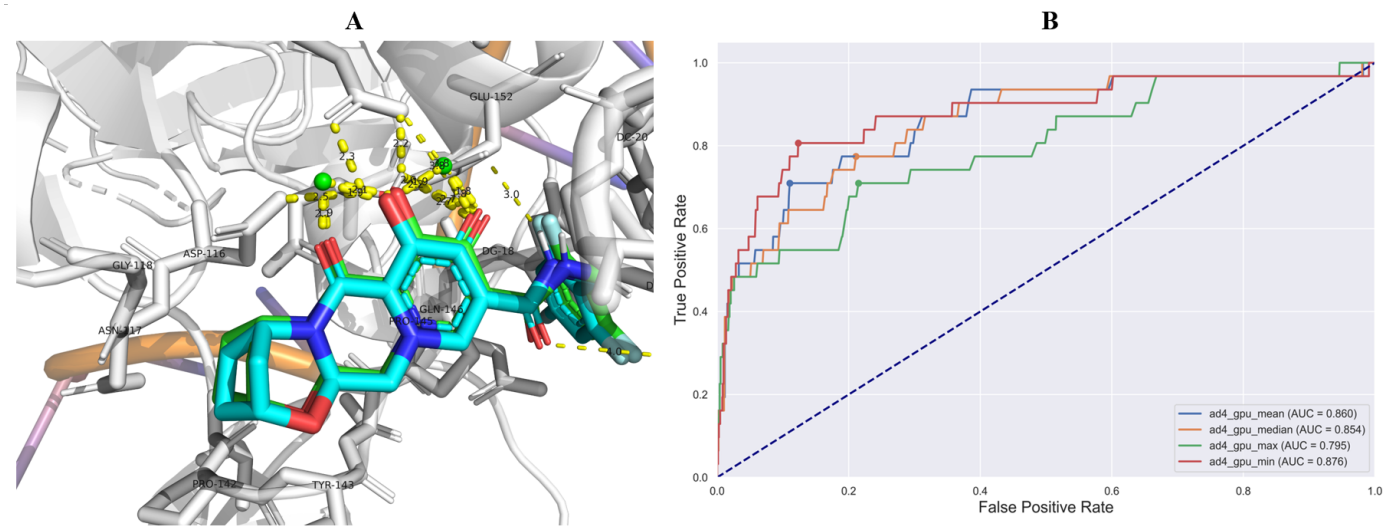


**Figure 10.** The results of redocking (A) and retrospective control (B) evaluation. The retrospective control was conducted utilizing four types of conformations, including the most negative (ad4_gpu_min), the most positive (ad4_gpu_max), the median (ad4_gpu_median), and mean (ad4_gpu_mean) of docked conformation distribution.

### 3.4. Virtual screening process

15235 substances from DrugBank were screened through medicinal chemistry filter, obtained 8333 structures. These structures then underwent further screening using a 2D-QSAR classification model, resulting in 44 compounds that were identified as active. Next, these 44 compounds underwent molecular docking, resulting in the discovery of 3 compounds, including two medicines and one hit (PSI-697). The results of the virtual screening process were shown in Fig 11.

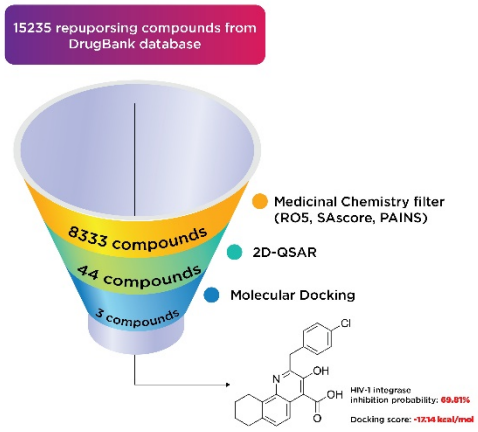


**Figure 11.** The results of virtual screening process

In general, the results obtained from the molecular docking process showed that all three compounds in Fig 12 formed electrostatic interactions with both Mg2+ ions. This is the first and highly important pharmacophore characteristic shared by all currently available INSTIs on the market. Additionally, all three selected conformations formed hydrogen bonds with the sidechain carbonyl group of Asp64, while fitting within the binding pocket between the two subunits of HIV integrase (matching the binding site of Bictegravir in the initial protein-ligand complex with PDB ID 6PUW).

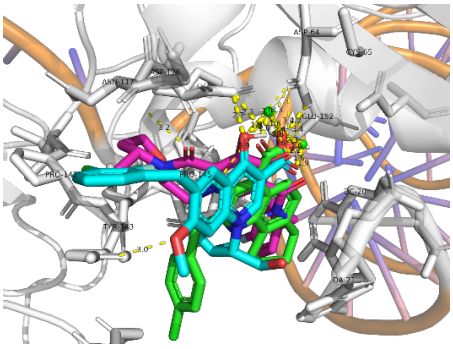

**Figure 12.** The binding modes of three potential candidates from the QSAR model. Blue: Elvitegravir (-17.32, 98,82%). Red: Dolutegravir (-11.42, 98,63%). Green: PSI-697 (-17.14, 69,81%).

Regarding PSI-697 (green) in Fig 13, the binding mode was similar to Bictegravir, but the docking score was more negative, with the figures being –17,14 kcal/mol and –11 kcal/mol, respectively. This could be explained by the hydrogen bonds with sidechain of Glu152, which was observed in complex of Bictegravir and protein. Moreover, the docking score of PSI-697 was also approximately equal to Elvitegravir (-17,32 kcal/mol). The HIV integrase inhibitory probability of PSI-697 from the QSAR model was also good, with the figure being around 69%. As a result, PSI-697 was the most promising candidate targeting HIV integrase for both inhibition and binding ability.

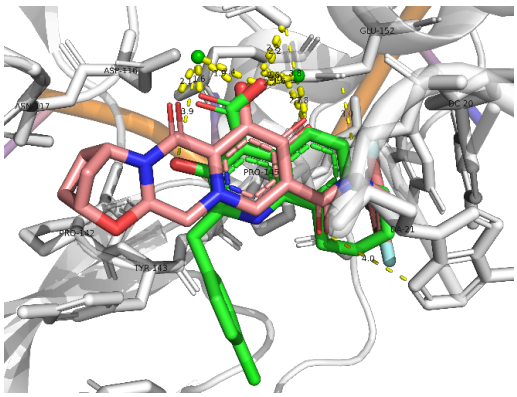

**Figure 13.** Bictegravir (red) and PSI-697 (green) in the active site.

## 4. Conclusion

Our study introduced a novel approach to machine learning, where decision-making stages were made based on statistical tests. We utilized 16 different molecular fingerprints and descriptors, and employed the Wilcoxon signed-rank test of cross-validation to determine the optimal one for feature and model selection. The LOF algorithm was implemented to establish the applicability domain, outperforming the bounding box technique in sparse areas.

The RDK7 fingerprint proved the most suitable, and XGBoost was the best model. External validation yielded impressive results with an F1-score of 0.889, average precision of 0.921, and recall of 0.921. These findings are highly generalizable and valuable for the virtual screening of potential HIV-1 integrase inhibitors.

The repurposing structure library was screened, resulting in the identification of one potential compound. We recommend the synthesis and biological activity testing of this potential compound.

## Data availability

The source code, notebooks, and all datasets are available at: https://github.com/Medicine-Artificial-Intelligence/HIV_IN_Classification_ML

## References

1. A. Van Heerden, H. Humphries, E. J. C. O. i. H. Geng and AIDS, *Ingenta connect*, 2023, **18**, 46-51.

2. P. A. Furman, J. A. Fyfe, M. H. St Clair, K. Weinhold, J. L. Rideout, G. A. Freeman, S. N. Lehrman, D. P. Bolognesi, S. Broder and H. J. P. o. t. N. A. o. S. Mitsuya, *PNAS*, 1986, **83**, 8333-8337.

3. E. J. A. r. De Clercq, *ScienceDirect*, 1998, **38**, 153-179.

4. M. Mahdi, J. A. Mótyán, Z. I. Szojka, M. Golda, M. Miczi and J. J. V. j. Tőzsér, *Virology Journal*, 2020, **17**, 1-8.

5. D. J. McColl and X. J. A. r. Chen, *ScienceDirect*, 2010, **85**, 101-118.

6. N. Ray, L. A. Blackburn and R. W. J. J. o. v. Doms, *Journal of Virology*, 2009, **83**, 2989-2995.

7. J. L. Blanco, G. Whitlock, A. Milinkovic and G. Moyle, *Expert Opinion on Pharmacotherapy*, 2015, **16**, 1313-1324.

8. T. A. Stern, G. L. Fricchione and J. F. Rosenbaum, *Massachusetts General Hospital Handbook of General Hospital Psychiatry-E-Book*, Elsevier Health Sciences, 2010.

9. A. K. Pau and J. M. J. I. D. C. George, *Infectious Disease Clinics*, 2014, **28**, 371-402.

10. D. Baptista, P. G. Ferreira and M. J. B. i. b. Rocha, *Briefings in Bioinformatics*, 2021, **22**, 360-379.

11. L. Zhang, H. Zhang, H. Ai, H. Hu, S. Li, J. Zhao and H. J. C. t. i. m. c. Liu, *Ingenta connect*, 2018, **18**, 987-997.

12. J. Kong, H. Lee, D. Kim, S. K. Han, D. Ha, K. Shin and S. J. N. c. Kim, *Nature Communications*, 2020, **11**, 1-13.

13. N. Ståhl, G. Falkman, A. Karlsson, G. Mathiason, J. J. J. o. c. i. Bostrom and modeling, *ACS Publications*, 2019, **59**, 3166-3176.

14. Y. Li, Y. Wu and A. J. M. I. Yan, 2017, **36**, 1600127.

15. A. Kurczyk, D. Warszycki, R. Musiol, R. Kafel, A. J. Bojarski, J. J. J. o. c. i. Polanski and modeling, 2015, **55**, 2168-2177.

16. L. A. Machado, E. Krempser and A. C. R. J. F. i. D. D. Guimarães, 2022, **2**, 954911.

17. S. C. Gad, in *Encyclopedia of Toxicology (Third Edition)*, ed. P. Wexler, Academic Press, Oxford, 2014, DOI: https://doi.org/10.1016/B978-0-12-386454-3.00971-4, pp. 1-9.

18. E. McCrum-Gardner, *British Journal of Oral and Maxillofacial Surgery*, 2008, **46**, 38-41.

19. G. J. R. Landrum, *ACS Publications*, 2013, **1**, 4.

20. H. Moriwaki, Y.-S. Tian, N. Kawashita and T. J. J. o. c. Takagi, *Journal of Cheminformatics*, 2018, **10**, 1-14.

21. T. Alice Capecchi, Richard Gowers, Map4, https://github.com/reymond-group/map4, (accessed Nov 13, 2022).

22. D. Probst and J.-L. J. J. o. c. Reymond, *Springer Link*, 2018, **10**, 1-12.

23. O. Kramer, in *Machine learning for evolution strategies*, Springer, 2016, pp. 45-53.

24. J. J. T. J. o. M. l. r. Demšar, *Journal of Machine Learning Research 7*, 2006, **7**, 1-30.

25. M. Kuss, C. E. Rasmussen and R. J. J. o. m. l. r. Herbrich, 2005, **6**.

26. V. Nguyen, 2019.

27. S. Y. Ho, K. Phua, L. Wong and W. W. B. J. P. Goh, *ScienceDirect*, 2020, **1**, 100129.

28. A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk and F. Herrera, *Learning from imbalanced data sets*, Springer, 2018.

29. J. Davis and M. Goadrich, 2006.

30. D. O. Passos, M. Li, I. K. Jóźwik, X. Z. Zhao, D. Santos-Martins, R. Yang, S. J. Smith, Y. Jeon, S. Forli and S. H. J. S. Hughes, 2020, **367**, 810-814.

31. F. Imrie, A. R. Bradley and C. M. Deane, *Bioinformatics*, 2021, **37**, 2134-2141.

32. J. Akosa, 2017.