

## Research Article

# Classification and Design of HIV-1 Integrase Inhibitors Based on Machine Learning

Junlin Zhou,<sup>1</sup> Juan Hao,<sup>1</sup> Lianxin Peng,<sup>1</sup> Huaichuan Duan,<sup>1</sup> Qing Luo,<sup>1</sup> Hailian Yan,<sup>1</sup> Hua Wan,<sup>2</sup> Yichen Hu,<sup>1</sup> Li Liang,<sup>1</sup> Zhenjian Xie,<sup>1</sup> Wei Liu ,<sup>1</sup> Gang Zhao ,<sup>1</sup> and Jianping Hu <sup>1</sup>

<sup>1</sup>Key Laboratory of Coarse Cereal Processing, Ministry of Agriculture and Rural Affairs, School of Pharmacy, Key Laboratory of Medicinal and Edible Plants Resources Development of Sichuan Education Department, Sichuan Industrial Institute of Antibiotics, Chengdu University, Chengdu 610106, China

<sup>2</sup>College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510106, China

Correspondence should be addressed to Wei Liu; [liuwei@cdu.edu.cn](mailto:liuwei@cdu.edu.cn), Gang Zhao; [2530813104@qq.com](mailto:2530813104@qq.com), and Jianping Hu; [hjpcdu@163.com](mailto:hjpcdu@163.com)

Received 14 January 2021; Revised 2 March 2021; Accepted 9 March 2021; Published 2 April 2021

Academic Editor: Lei Chen

Copyright © 2021 Junlin Zhou et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A key enzyme in human immunodeficiency virus type 1 (HIV-1) life cycle, integrase (IN) aids the integration of viral DNA into the host DNA, which has become an ideal target for the development of anti-HIV drugs. A total of 1785 potential HIV-1 IN inhibitors were collected from the databases of ChEMBL, Binding Database, DrugBank, and PubMed, as well as from 40 references. The database was divided into the training set and test set by random sampling. By exploring the correlation between molecular descriptors and inhibitory activity, it is found that the classification and specific activity data of inhibitors can be more accurately predicted by the combination of molecular descriptors and molecular fingerprints. The calculation of molecular fingerprint descriptor provides the additional substructure information to improve the prediction ability. Based on the training set, two machine learning methods, the recursive partition (RP) and naive Bayes (NB) models, were used to build the classifiers of HIV-1 IN inhibitors. Through the test set verification, the RP technique accurately predicted 82.5% inhibitors and 86.3% noninhibitors. The NB model predicted 88.3% inhibitors and 87.2% noninhibitors with correlation coefficient of 85.2%. The results show that the prediction performance of NB model is slightly better than that of RP, and the key molecular segments are also obtained. Additionally, CoMFA and CoMSIA models with good activity prediction ability both were constructed by exploring the structure-activity relationship, which is helpful for the design and optimization of HIV-1 IN inhibitors.

## 1. Introduction

Acquired immune deficiency syndrome (AIDS) is a systemic immune dysfunction syndrome caused by the infection of human immunodeficiency virus (HIV) infection, inducing the destruction of CD4<sup>+</sup> T lymphocytes [1–3]. HIV can be divided into two subtypes: HIV-1 (i.e., the main pathogen of AIDS) and HIV-2. HIV-1 is characterized by strong infection, rapid mutation, and high mortality and can be transmitted through blood, mother-infant, sexual intercourse, etc. [4–8]. Since the first case of HIV-1 infection in 1981, the number of AIDS patients has exploded worldwide [9]. According to World Health Organiza-

tion (WHO) data of 2019, more than 38 million people have been infected, and 7.1 million of them have died [10]. Highly active antiretroviral therapy (HAART) is the main strategy in the clinical treatment of AIDS—the combination of drugs inhibiting both reverse transcriptase (RT) and the protease (PR), which can reduce the damage of virus to immune system [11]. However, the high variability of HIV-1 results in poor efficacy of HAART treatment, leading to the emergence of drug-resistant virus strains. It is urgent to identify new targets and develop novel structural inhibitors [12–14].

As such an attractive and important target, HIV-1 integrase (IN) is an essential enzyme in the HIV-1 lifecycle responsible

for inserting the reverse-transcribed viral genome into the host DNA through 3' processing (3'-P) and strand transfer (ST) reaction [15, 16]. Unlike PR and RT, there is neither known functional analog of IN in human cells nor apparent cellular toxicity for IN inhibitors [17, 18]. Encoded by the pol gene, HIV-1 IN is composed of 288 residues with molecular weight of 32 kDa, which can be divided into three domains: N-terminal domain (NTD, residues 1-49), catalytic core domain (CCD, residues 50-212), and C-terminal domain (CTD, residues 213-288) [19]. The zinc finger in NTD is conducive to the stability of the whole IN enzyme; proper chelation of DDE motif (i.e., Asp64, Asp116, and Glu152) in CCD with two  $Mg^{2+}$  ions is essential to maintain high enzymatic activity; CTD serves as the nonspecific binding to viral DNA [20–26].

There are ten main types of HIV-1 IN inhibitors currently reported: diketoacids, diazonaphthalene derivatives, quinolone acids, pyrimidine ketone, sulfur nitrogen thiozapine, polyhydroxy arylcyclic compounds, disulfoxide compounds, benzene sulfonamides, coumarin derivatives, salicylhydrazide derivatives, etc. [27–33]. Diketoacids are the most fully studied and most promising inhibitors against HIV-1 IN, showing high efficiency, high selectivity, and low toxicity [34–36]. In terms of inhibitory mechanism of diketoacid compounds, the carbonyl and carboxyl groups both are, respectively, chelated with two different  $Mg^{2+}$  ions, which significantly weakens ST reaction by destroying metal-DDE recognition. Raltegravir (RLT) was the first approved IN inhibitor drug through FDA in 2007, followed by elvitegravir (EVG) and dolutegravir (DTG) for clinical use [37–39]. Here, all three belong to diketoacid compounds.

A lot of experimental and theoretical studies have involved IN-ligand recognition, inhibition mechanism, and molecular modification of diketoacid compounds. Two important scientific problems remain unclear: (1) for many IN inhibitors reported, is there a good classification method to determine their activity? (2) How to effectively modify the diketoacid inhibitor by obtaining the key groups that affect molecular activity and then combining 3-dimensional quantitative structure-activity relationship (3D-QSAR) results? In this work, HIV-1 IN inhibitors were first collected to establish a personalized database; the activity prediction model and the key groups affecting the activity both were obtained with recursive partition (RP) and naive Bayes (NB) model; finally, based on the structure and activity data of pyruvic acid inhibitors against HIV-1 IN, a 3D-QSAR model with good predictive ability was proposed [40–42]. In particular, the quantitative relationship between molecular structure (such as spatial conformation, electrostatic characteristics, hydrophobicity, and H-bond) and its inhibitory activity was explored, which will provide theoretical guidance for the design of effective anti-AIDS drugs.

## 2. Methods

**2.1. Preparation of HIV-1 IN Inhibitor Database.** The established HIV-1 IN inhibitor database contains 682 inhibitors and 1103 noninhibitors (1785 molecules in total). All data on the structure and activity of small molecules were obtained from ChEMBL, Binding Database, DrugBank, PubMed, and 40 recent references. The  $IC_{50}$  value of 4600  $\mu m$  was set as

the criterion for defining an inhibitor. In data processing, 1 and 0 were adopted to characterize inhibitors and noninhibitors, respectively. All the small molecules were generated using ChemOffice package with Gasteiger-Hückel charge attached and then optimized by the steepest descent (1000 steps) and the conjugate gradient (1000 steps) algorithms based on Tripos force field of SYBYL package [43]. The convergence criterion is less than 4.182  $kJ \cdot mol^{-1} \cdot nm^{-1}$  for energy gradient. The optimized structure is the basis of the subsequent molecular descriptor and molecular fingerprint calculations.

**2.2. Calculation of Molecular Descriptors and Molecular Fingerprints.** Molecular descriptors and molecular fingerprints of personalized database elements both were calculated with Discovery Studio 3.5 (DS 3.5) package. Here, a total of 13 molecular descriptors widely used in ADME prediction were adopted for calculation: apparent partition coefficient (logD), octanol-water partitioning coefficient (AlogP), the number of rotatable bonds ( $n_{rot}$ ), molecular weight (MW), the number of H-bond donors ( $n_{HBD}$ ), the number of H-bond acceptors ( $n_{HBA}$ ), the sum of oxygen and nitrogen atoms ( $n_{O+N}$ ), polar surface area (PSA), the number of aromatic rings ( $n_{AR}$ ), the number of rings ( $n_R$ ), molecular solubility (logS), molecular fraction polar surface area (MFPSPA), and molecular surface area (MSA) [44–46].

The SciTegic extended link fingerprints (i.e., FCFP, ECFP, and LCFP) and path-based ones (i.e., FFPF, EPFP, and LEFP) both were calculated using Morgan algorithm [47, 48]. The first letter of molecular fingerprint, F/E/L, respectively, represents atomic functional role code, the properties used in the Daylight atomic invariants rule, and atomic type code of AlogP. Atomic functional role code (i.e., letter F) mainly includes the combinations of H-bond acceptor, H-bond donor, positive ionization, negative ionization, and aromatic with halogen elements. Letter E consists of the sum of connection number among atoms, element types, atomic charge, atomic weight, etc. Letter L is used to characterize the 120 atomic types involved in AlogP calculation. The second letter, C/P, respectively, stands for extended-connection molecular fingerprint and path-based one. The third and fourth letters are derived from the initial capitalization of the "finger print" word. Four-letter molecular fingerprints are often followed by Arabic numerals 4 or 6, indicating the maximum distance between atoms. As an important complement to molecular descriptors for drug-like compounds, molecular fingerprint parameters have been widely used in the classification and prediction of inhibitors and noninhibitors.

**2.3. Recursive Partitioning Classifiers.** Recursive partitioning (RP) is a classification statistical method which can directly predict inhibitors and noninhibitors in the form of decision tree, on the basis of compound data processing and biological activity threshold criteria. Depth and node both are two important parameters of decision tree, respectively, corresponding to the complexity of the whole event and the judgment process [49–51]. For example, when a compound is at the logD node, its calculated data can be compared with the threshold value and partitioned and then reclassified and repartitioned at the next RP node, until the RP is

infeasible to continue at the bottom of tree. The criterion for stopping partitioning is that the classification effect cannot be improved or the remaining samples are too small. As far as the tree depth is concerned, the larger the value, the better the classification of the training data, despite the risk of overfitting; a smaller tree depth indicates that the accuracy of feature recognition in the training set is slightly lower, and the tree shows good adaptability to the new dataset; generally, the tree depth of 3 to 10 is a more appropriate. In accordance with the golden section ratio, the database was divided into the training set containing 1485 compounds and the test set containing 300 compounds. The decision tree was established based on the training set, and the accuracy of model prediction was evaluated from the test set data.

**2.4. Naive Bayesian Classifiers.** In addition to RP method, naive Bayesian (NB) model was also performed to develop classifiers to distinguish HIV-1 IN inhibitors from noninhibitors [52–55]. Firstly, the  $f$  vector ( $f = \langle f_1, f_2, \dots, f_n \rangle$ ) was set, and the component vectors ( $f_1, f_2, \dots$  and  $f_n$ ) were, respectively, calculated to obtain the eigenvectors ( $F_1, F_2, \dots$  and  $F_n$ ), which can be used to represent the corresponding molecular descriptors or molecular fingerprints. According to Bayes' theorem, the conditional probability and marginal probability of two events can be correlated, as shown in formula (1):

$$p(C | F_1, F_2, \dots, F_n) = \frac{p(C)p(F_1, \dots, F_n | C)}{p(F_1, \dots, F_n)}. \quad (1)$$

Here,  $C$  stands for the classification of compounds;  $p(C | F_1, F_2, \dots, F_n)$  is the posteriori probability after classification;  $p(C)$  is the prior probability obtained from the training set;  $p(F_1, \dots, F_n | C)$  is for the conditional probability of compounds having a specific molecular descriptor;  $p(F_1, \dots, F_n)$  represents the marginal probability (or total probability) for the occurrence of all particular molecular descriptors. In NB model, each molecular descriptor is independent of each other, from which formula (2) can be obtained:

$$p(F_1, \dots, F_n | C) = p(F_1 | C) \cdots p(F_n | C) = \prod_{i=1}^n p(F_i | C). \quad (2)$$

Based on the data of training set, all the coefficients required in formula (2) can be calculated by formulae (3) and (4):

$$p(F_i = f_i | +) = \frac{\text{count}(F_i = f_i \cap C = +)}{\text{count}(C = +)}, \quad (3)$$

$$p(F_i = f_i | -) = \frac{\text{count}(F_i = f_i \cap C = -)}{\text{count}(C = -)}. \quad (4)$$

In this work, all compounds are divided into either HIV-1 IN inhibitors or noninhibitors.  $p(+)$  and  $p(-)$ , respectively, represent the prior probability grouped into

inhibitors and noninhibitors. The posterior probability of the compound being an inhibitor ( $p$ ) or a noninhibitor ( $q$ ) is calculated as follows:

$$p = \frac{p(+)}{p(F_1 = f_1, \dots, F_n = f_n)} \prod_{i=1}^n p(F_i = f_i | +), \quad (5)$$

$$q = \frac{p(-)}{p(F_1 = f_1, \dots, F_n = f_n)} \prod_{i=1}^n p(F_i = f_i | -),$$

where the marginal probability  $p(F_1 = f_1, \dots, F_n = f_n)$  is a constant, and the sum of  $p$  and  $q$  is equal to 1.

**2.5. Evaluation of Classification Model Quality.** The prediction ability of NB and RP models can be evaluated with lots of parameters, including true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), sensitivity (SE), specificity (SP), prediction accuracy of TP (PRE1), prediction accuracy of TN (PRE2), and Matthews correlation coefficient (C). Their specific calculation formula is as follows:

$$SE = \frac{TP}{TP + FN}, \quad (6)$$

$$SP = \frac{TN}{TN + FP}, \quad (7)$$

$$PRE1 = \frac{TP}{TP + FP}, \quad (8)$$

$$PRE2 = \frac{TN}{TN + FN}, \quad (9)$$

$$C = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}. \quad (10)$$

**2.6. Three-Dimensional Quantitative Structure Activity Relationship.** The quinolinone acid compounds—a class of HIV-1 IN inhibitors—were randomly divided into a training set (18 molecules in total) and a test set (4 molecules in total), and their experimental  $IC_{50}$  values were transformed into negative logarithmic form (i.e.,  $pIC_{50}$ ). According to the three-dimensional quantitative structure activity relationship (3D-QSAR) theory, molecules with similar conformations tend to have similar biological activity. In this work, compound # 2, which has been resolved and has good activity, is selected as the template molecule [56]. Before constructing the 3D-QSAR models, all molecules are aligned according to the principle that common substructures overlap each other, which is done in SYBYL-X1.3. In addition, comparative molecular similarity index analysis (CoMSIA) and comparative molecular field analysis (CoMFA) both are currently the two most widely used 3D-QSAR methods. In order to establish CoMSIA model, the superimposed inhibitors were placed in the spatial grid, and a series of  $sp^3$ -hybridized probe particles (such as  $C^+$ ,  $CH_4$ ,  $H^+$ , and  $H_2O$ ) were rolled to calculate the interactions between probe and inhibitor. Based on different spatial coordinates of probes, all the field data of inhibitors were obtained, including steric field (S),

electrostatic field (E), hydrophobic field (HD), H-bond acceptor (A), and H-bond donor (D) [57–61]. Compared with CoMSIA model, CoMFA only provides information on S and E fields.

Partial least squares (PLS) method was used for regression analysis of the training set. Leave-one-out (LOO) method was adopted for cross validation to gain the optimal numbers of component (ONC) and determination coefficient  $q^2$ . Based on the ONC values, 3D-QSAR models were established by noncross validation, and a series of parameters including correlation coefficient  $r^2$ , estimated standard error Es, root mean square error (RMSE), and  $F$ -test values were obtained accordingly. These parameters can be used to evaluate the stability and predictive ability of the models and predict biological activity of the molecules in test set [62–64]. The prediction correlation coefficient for the test set is calculated by equation (11):

$$r_p^2 = \frac{SD - \text{PRESS}}{SD}. \quad (11)$$

In equation (11), SD represents the deviation-square sum between experimental biological activity data in the test set and the average biological activity in the training set, and PRESS indicates the error-square sum of the predictive biological activity in the test set with experimental biological activity.

### 3. Results and Discussion

**3.1. Classification Based on Molecular Descriptors.** Compounds with similar biological activity usually have some similar molecular descriptors, such as reasonable hydrophilicity and H-bond number and volume. In theory, molecular descriptors can be partially used to classify inhibitors and noninhibitors. Figure 1 shows the distributions of eight molecular descriptors (i.e., AlogP, logD, MW,  $n_{\text{HBA}}$ ,  $n_{\text{HBD}}$ , MSA,  $n_{\text{AR}}$ , and MFPSA). The distribution of AlogP ranged from -13.707 to 13.333, with an average of 2.523. Specifically, the average values for 682 inhibitors and 1103 noninhibitors were 2.608 and 2.470, respectively. Then,  $t$ -test was used to evaluate the significant difference of AlogP between inhibitors and noninhibitors. At 95% confidence level, the  $P$  value related to the difference between the two types of molecules was 0.274, indicating that there was no significant difference between the two distributions. Similarly, logD and  $n_{\text{AR}}$  both also have high  $P$  values of 0.236 and 0.332, respectively. Obviously, these three molecular descriptors with higher  $P$  values cannot be used to distinguish HIV-1 IN inhibitors from noninhibitors.

In addition, the  $P$  value of the other five molecular descriptors (i.e., MW,  $n_{\text{HBA}}$ ,  $n_{\text{HBD}}$ , MSA, and MFPSA) were relatively small, with a minimum of  $6.78e^{-14}$  and a maximum of  $4.9e^{-4}$ . Given the relatively scattered distribution and small overlap of these parameters, it is obvious that they cannot be used to accurately distinguish inhibitors from noninhibitors.

**3.2. Classification Based on Recursive Partition Model.** According to the above analysis, using a single molecular

descriptor cannot classify compounds well. In order to establish a more accurate and understandable classification model, IC<sub>50</sub> values were set as the classification basis and recursive partitioning (RP) model was adopted. In this model, molecules were divided into smaller and smaller subsets and finally presented in the form of decision tree. According to our experiments, the classification performance of the RP model containing 12 molecular descriptors and molecular fingerprints is better than that of the model only with molecular descriptors. Of all the molecular fingerprints, ECFP\_6 and FCFP\_6 both have better classification effect in training set and test set. The corresponding highest  $C$  value of Matthews correlation coefficient was 0.717 and 0.733, respectively (see figure S1). Based on molecular descriptors and ECFP\_6, the sensitivity and specificity of RP model were 0.848 and 0.852, respectively. The prediction accuracy of inhibitors and noninhibitors was 70.3% and 90.4%, respectively.

In RP model, depth is an important parameter determining the complexity of decision tree. Generally, the larger the tree depth, the more accurate the recognition of important features in the training set, but it also increases the risk of overfitting; the smaller the tree depth, the higher the tree applicability to datasets. Figure 2 shows the  $C$  value changes of the training and test sets along with tree depth, which are used to evaluate the response ability of the models. From the training set, the  $C$  value increases with the growing of tree depth. For the test set, the  $C$  value reaches a maximum of 0.748, when the tree depth is 9. In order to avoid overfitting phenomenon, setting the tree depth of RP model to 9 is the best choice.

Table S1 lists the decision tree reports with tree depth of 9. In the mixed matrix, experimental data and prediction results are filled in the vertical and horizontal columns, respectively, where 0 and 1 represent HIV-1 IN inhibitor and noninhibitor, respectively. Based on the above equations (8) and (9), the prediction accuracies of inhibitors and noninhibitors are 0.825 and 0.862, respectively, with Matthews correlation coefficient of 0.722.

Figure 3 shows all the details of a decision tree with a depth of 9. It can be seen that the decision tree has 25 internal nodes and 26 leaves. The discriminant descriptors consist of 7 molecular properties (ECFP\_6, logD, MSA, AlogP,  $n_{\text{HBD}}$ , MFPSA, and MW) and 16 structural fragments (F1 ... F16). These 16 molecular fingerprints are helpful to distinguish inhibitors from noninhibitors and have positive reference value to the following drug design (see figure S2).

**3.3. Classification Based on Naive Bayesian Model.** Although the decision tree obtained by RP model is concise and explicit, this method is highly sensitive to predetermined parameters and it will lead to false positives (FP) and false negatives (FN). To further improve model accuracy and make comparison, we used naive Bayesian (NB) method to establish another new classification model. The process of NB classification is to find features with separation ability in an unbiased way, which does not involve parameter fitting and adjustment as an unsupervised learning method.

Table S2 shows the influence of different parameter combinations in the test and training sets on NB classification.

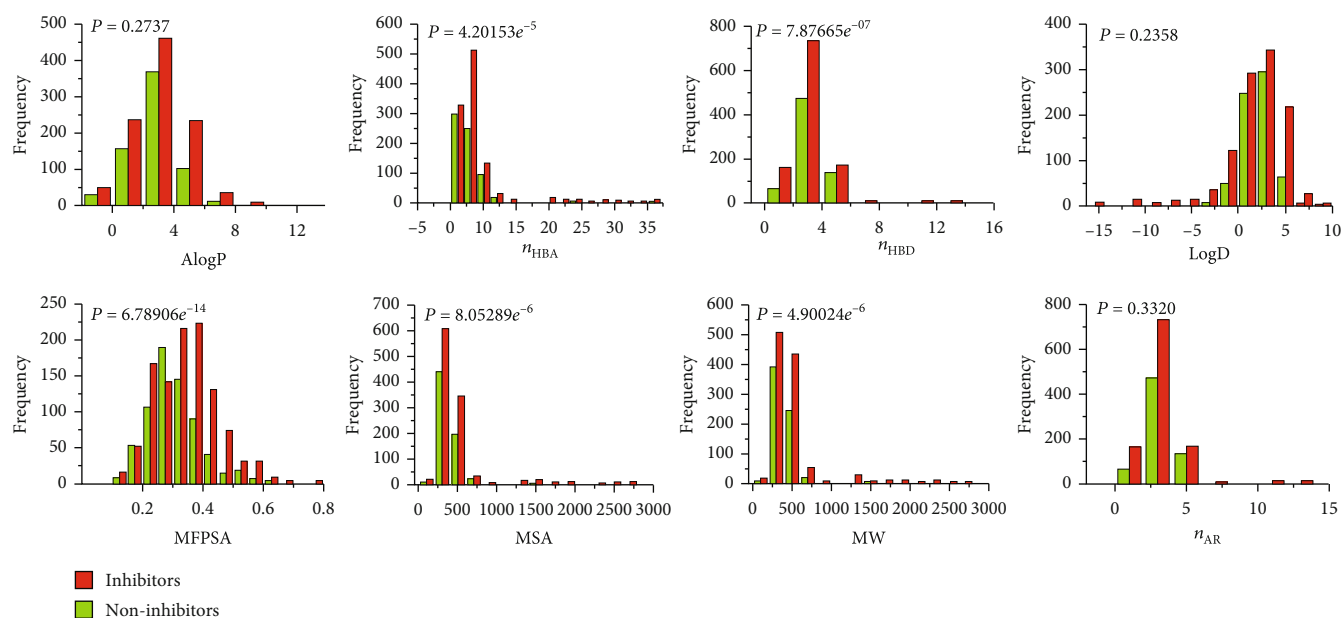


FIGURE 1: Distributions of eight molecular descriptors of both inhibitors and noninhibitors.

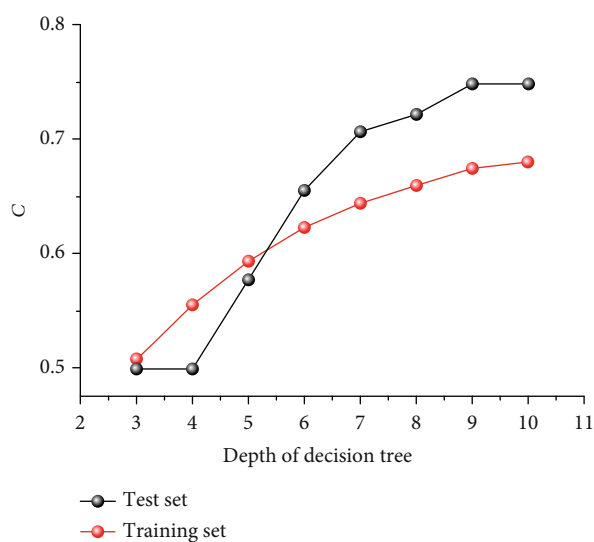


FIGURE 2: The  $C$  value changes of the training and test sets along with tree depth.

As for the model only based on molecular descriptors, the sensitivity, specificity, and  $C$  value of training set were 66.9%, 71.9% and 0.561; while those of test set were 60.5%, 78.4%, and 0.57, respectively. Considering that Matthews coefficient  $C$  is unsatisfactory, several other types of NB models have also been constructed by combining molecular fingerprint and molecular descriptor, and the classification performance is significantly improved.

Meanwhile, ECFP\_6 was selected to compare the prediction performances between RP and NB models. Table 1 shows the cross validation of NB classification, from which the TP, FN, FP, and TN values of NB model are 627, 55, 210, and 893, respectively. It turns out that the prediction accuracy of inhibitors is 0.883, and that of noninhibitors is

0.872; the correlation coefficient  $C$  value of NB model is 0.852, which is slightly higher than RP model (see Table S1), indicating that the prediction ability of NB model is better.

Like RP model, NB classification can also provide the unique key fragment structure (i.e., molecular fingerprints) of certain compounds. Molecular fingerprints can be transformed into two-dimensional fragments, which aids the design of HIV-1 IN inhibitors. Figure 4 shows the top 20 potential favorable molecular fingerprints of HIV-1 IN inhibitors obtained by NB classification. Most of the advantageous fragments contain nitrogen-oxygen heterocycles (such as oxazole rings and pyridine rings), providing implications for molecular design based on inhibition mechanism and ligand structure. In addition, oxygen-sulfur double bond, nitrogen-nitrogen double bond, and pyrimidine ring appear in the unfavorable fragments (see Figures S3), which should be filtered out to improve the screening efficiency of HIV-1 IN inhibitors.

#### 3.4. Molecular Design for Quinolinone Acid Inhibitors.

Figure 5 shows structures and  $\text{pIC}_{50}$  of 22 quinolinone acid inhibitors, where 18 inhibitors were randomly selected into the training set to establish three-dimensional quantitative structure activity relationship (3D-QSAR) model. In the preprocessing step of 3D-QSAR analysis, the conformations of quinolinone acid inhibitors overlap quite well (see figure S4), which lays the foundation for the subsequent establishment of a good model.

As for the CoMFA models (see Table S3), the cross-validated correlation coefficient ( $q^2$ ) and noncross-validated correlation coefficient ( $R^2$ ) were 0.864 and 0.969, respectively. The root mean square error (RMSE) was 0.020, and the combination with high predicted correlation coefficient ( $r_p^2 = 0.918$ ) confirms the reasonability and reliability of this model. According to the CoMFA model, the contribution

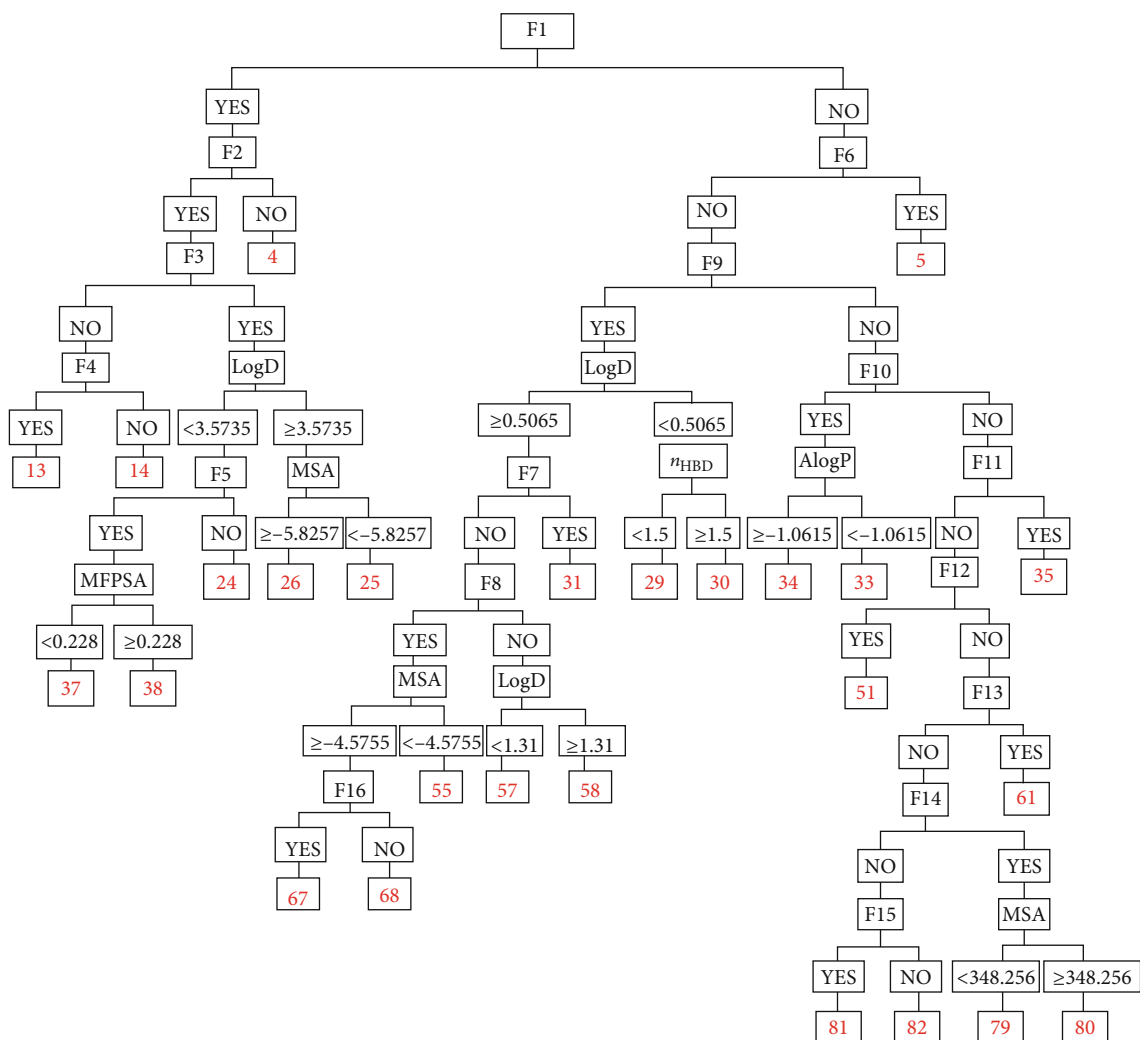


FIGURE 3: Decision tree with a depth of 9.

TABLE 1: Cross validation of naive Bayesian classification.

Model name	ROC score	ROC rating	TP	FN	FP	TN	SE	SP	C
Naive Bayesian model	0.897	Good	627	55	210	893	0.919	0.81	0.852

rates of steric field (S) and electrostatic file (E) are 68.6% and 31.4%, respectively. It indicates that S has an important influence on the inhibitory activity of quinolinone acid inhibitors. In CoMSIA model, the contribution rates of S, E, H, D, and A were 7.4%, 13.2%, 16.9%, 45.7%, and 16.8%, respectively, which shows that the H-bond donors of quinolinone acid inhibitors have great influence on their activity. Then, the trained CoMFA and CoMSIA models both are used to predict molecular activity in the test set. Figure 6 shows the correlation between experimental  $pIC_{50}$  and the predicted values by two models. It can be found that the correlation coefficient  $R^2$  was greater than 0.9, and the deviation between the predicted and experimental data was less than 1, which proves the reliability of the two models. In addition, some individual results are very consistent with

their experimental data, such as compounds 4, 7, 10, and 14 in CoMFA model as well as compounds 4 and 10 in CoMSIA model.

It has been mentioned above that the two models complement and verify each other, which provides an important idea for the design of HIV-1 IN inhibitors. Figure S5 shows the CoMFA and CoMSIA contour map of quinolinone acid inhibitors, where compound 2 is used as the template and the contour line truncation is 80%: 20%. In the S field, there are large yellow and green blocks around the inhibitor, indicating that the introduction of large volume groups in the corresponding region is not conducive to and conducive to the enhancement of inhibitory activity. The yellow blocks are widely distributed around the ketone, carboxyl, and chlorine atoms; the green blocks are clustered

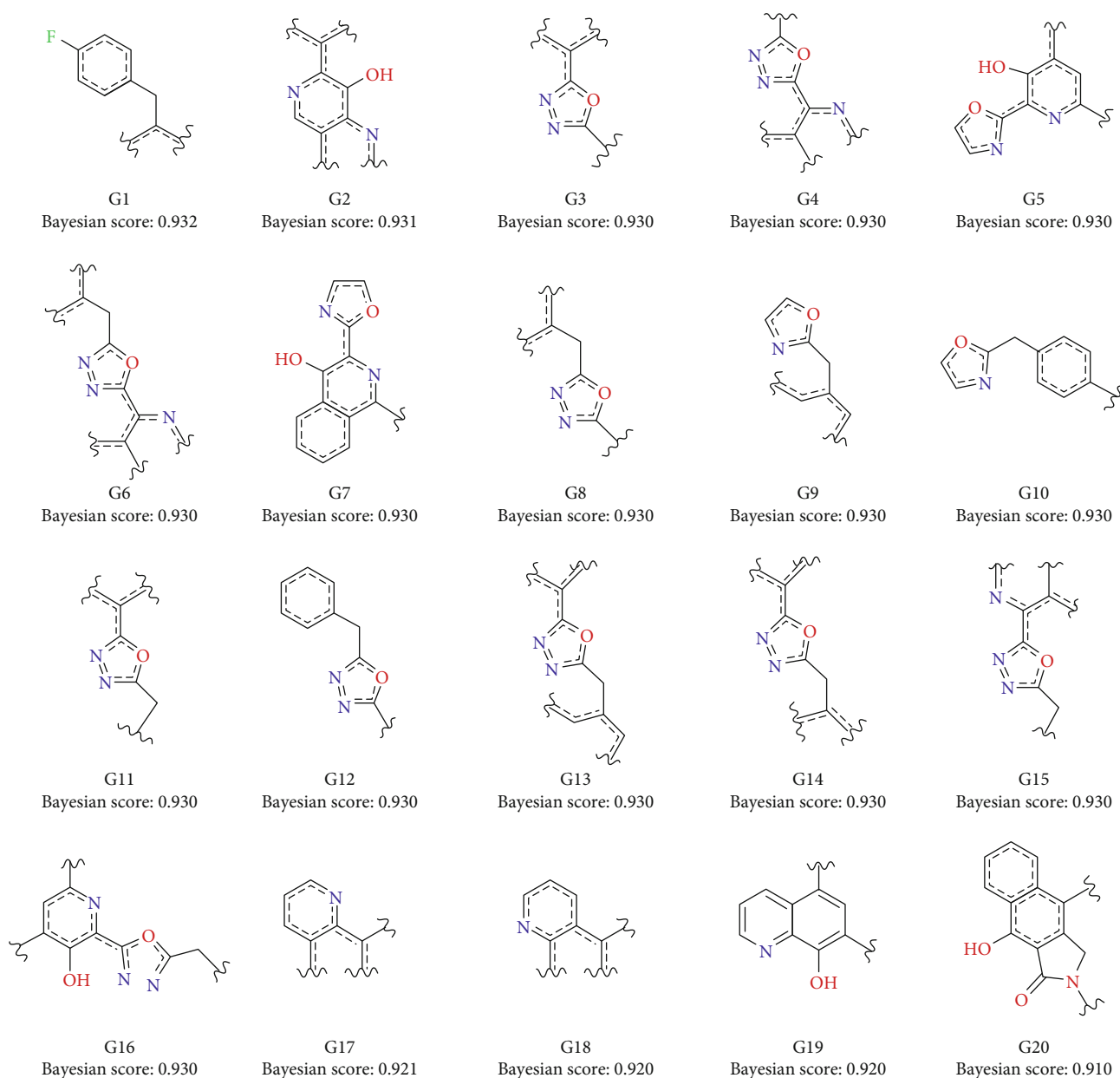


FIGURE 4: Potentially advantageous molecular fingerprint structures for HIV-1 IN inhibitors derived from naive Bayesian classification.

near the nitrogen atom; it partly confirms why compounds 10 and 11 have better inhibitory activities. In the E field, the blue area near the ketone group indicates the introduction of a positively charged groups is conducive to improving inhibitory activity; there is a red block near the amino group, so the introduction of negatively charged group should be fully considered in molecular design. H-bond is one of the most important nonbonding interactions between drug molecules and receptors, which is the key to the drug-target specific recognition and biological activity. In the H field, almost all the hydrophobic chains are surrounded by gray blocks, and the introduction of hydrophobic groups will reduce the inhibitory activity. As for the D field, cyan represents the donor region of H-bond, where introducing hydroxyl or carboxyl group helps to improve inhibitory activity. In the A field of CoMSIA

contour map, the red blocks near the carboxylic acid group and the contralateral nitrogen atom are mainly H-bond receptor rejection regions, and the introduction of H-bond receptors is strictly prohibited.

Based on the above analyses and previous studies, several design suggestions on improving the activity of quinolinone acid inhibitors are proposed: (1) the structure of beta-carbonyl carboxylic acid is strictly preserved, which is the key to maintain its activity; (2) at the N atom of amino group, long fatty chains (especially negatively charged one, such as carboxy group) are recommended; (3) H-bond donors (such as amino or hydroxyl group) may be considered for addition to the diphenylmethane side of quinolinone. To be objective, molecular dynamics (MD) simulation and biochemical enzyme experiments both are still needed to further verify the above molecular design ideas.

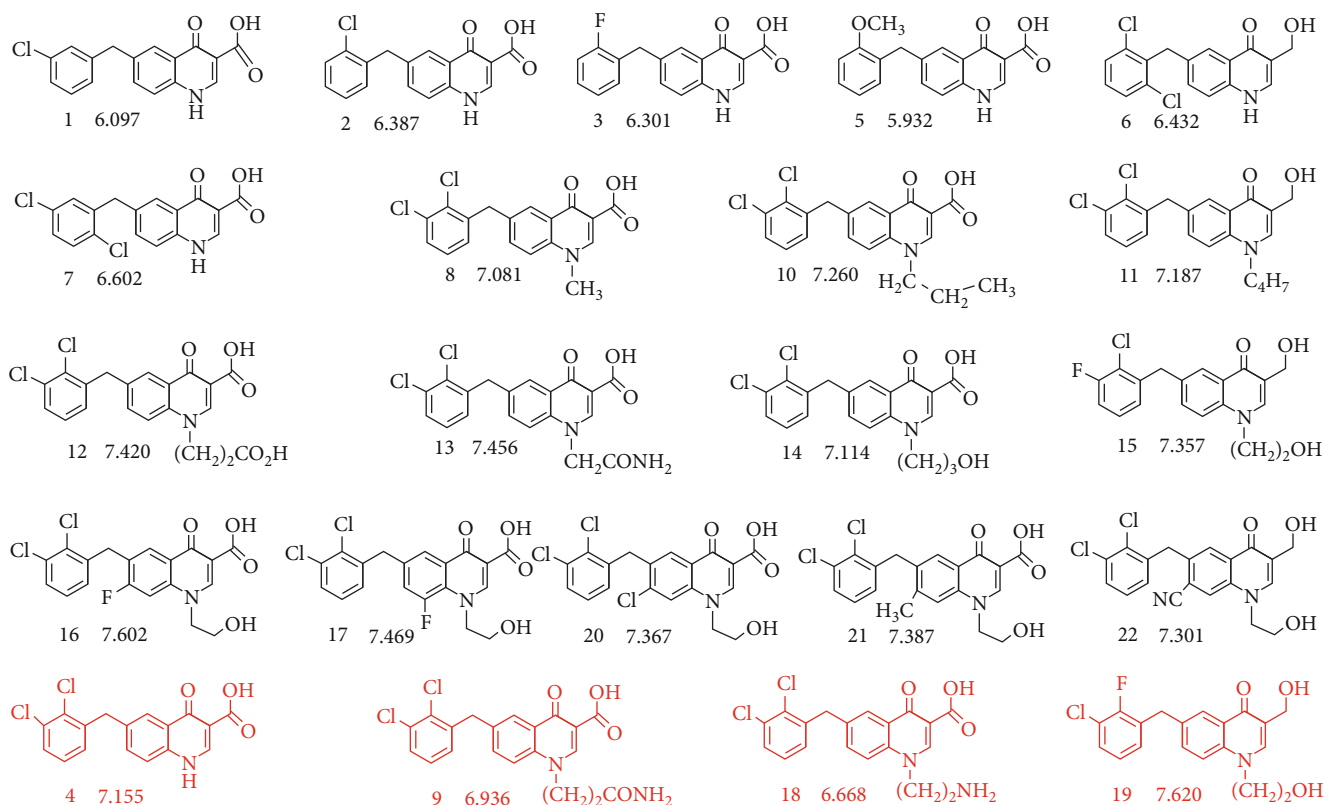


FIGURE 5: Structures and  $pIC_{50}$  values of quinolinone acid inhibitors against HIV-1 IN. The training set and test set are shown in black and red, respectively. The  $IC_{50}$  value is units of  $\mu M$ .

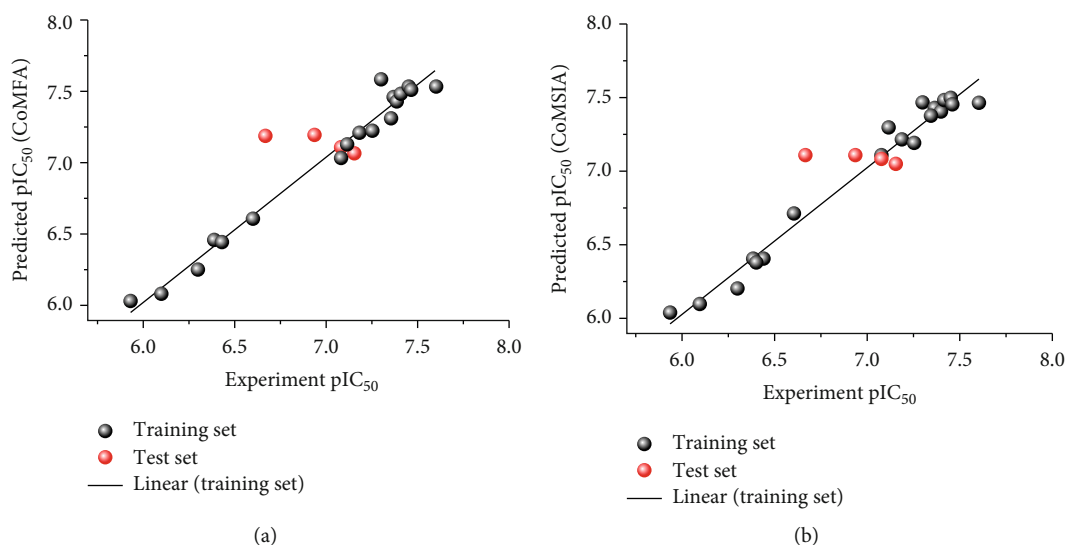


FIGURE 6: The correlation between experimental  $pIC_{50}$  and the predicted values by (a) CoMFA and (b) CoMSIA models.

#### 4. Conclusion

A database of HIV-1 IN inhibitors containing 1785 molecular structure and biological activity data was established first. The relationship between molecular descriptors and their inhibitory activities was systematically studied through the RP and NB methods. The prediction performance of the

two classification models based on the combination of molecular descriptor with molecular fingerprint than that based on the individual molecular descriptor. By analyzing the key fragments transformed from molecular fingerprints, the nitrogen-containing ring including oxazole and pyridine rings is suggested to be introduced into subsequent inhibitor modification.



Finally, CoMSIA and CoMFA models with good predictive ability ( $R^2 > 0.9$ ) both were established by selecting quinolinone acid inhibitors against HIV-1 IN. According to the contour maps and the favorable groups given by NB classification, several design suggestions on improving the activity of inhibitors are proposed. In particular, it is recommended to introduce long fatty chains (especially negatively charged one, such as carboxy group) into the N atom of amino group, as well as H-bond donors (such as amino group, hydroxyl group, and nitrogen-oxygen heterocycles) into the diphenylmethane side of quinolinone. This work provides some theoretical guidance for classification and molecular design of HIV-1 IN inhibitors.

### Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

### Ethical Approval

No animals/humans were used for studies that are base of this research.

### Conflicts of Interest

The authors declare no conflict of interest, financial, or otherwise.

### Authors' Contributions

Junlin Zhou, Juan Hao, and Lianxin Peng contributed equally to this work.

### Acknowledgments

This work was supported by the National Key Research and Development Program of China (2018YFC1602101), the funding from the National Natural Science Foundation of China (31600591 and 31870655), the project of Chongqing Key Laboratory of Environmental Materials and Restoration Technology (CEK1803), the Scientific Special Fund of Sichuan Traditional Chinese Medicine Administration (2018KF006), and the project of Chengdu Science and Technology Bureau (2016-XT00-00023-GX).

### Supplementary Materials

Figure S1: the  $C$  values of decision trees based on molecular descriptors and molecular fingerprints in the (a) training set and (b) test set. Figure S2: the 16 dominant fingerprints obtained from decision tree with depth of 9. Figure S3: potentially disadvantageous molecular fingerprint structures for HIV-1 IN inhibitors derived from naive Bayesian classification. Figure S4: structural superimposition of quinolinone acid inhibitors in the training set. Figure S5: CoMFA and CoMSIA contour maps of quinolinone acid inhibitors: (a) steric field, (b) electrostatic field, (c) hydrophobic field, (d) H-bond donor field, and (e) H-bond acceptor field. Table S1: decision tree report with tree depth of 9. Table S2: effects

of different parameter combinations in the test set and training sets on naive Bayesian classification. Table S3: statistical parameters of CoMFA and CoMSIA models. (*Supplementary Materials*)

### References

- [1] R. A. Frost, J. Fuhrer, R. Steigbigel, P. Mariuz, C. Lang, and M. Gelato, "Wasting in the acquired immune deficiency syndrome is associated with multiple defects in the serum insulin-like growth factor system," *Clinical Endocrinology*, vol. 44, no. 5, pp. 501–514, 1996.
- [2] A. K. Steele, E. J. Lee, J. A. Manuzak et al., "Microbial exposure alters HIV-1-induced mucosal CD4<sup>+</sup> T cell death pathways *Ex vivo*," *Retrovirology*, vol. 11, no. 1, 2014.
- [3] Y. Y. Qin, Y. H. Zhou, Y. Q. Lu et al., "Effectiveness of glucocorticoid therapy in patients with severe coronavirus disease 2019: protocol of a randomized controlled trial," *Chinese Medical Journal*, vol. 133, no. 9, pp. 1080–1086, 2020.
- [4] T. X. CHU and J. A. LEVY, "Injection drug use and HIV/AIDS transmission in China," *Cell Research*, vol. 15, no. 11–12, pp. 865–869, 2005.
- [5] M. Laplana, A. Caruz, J. A. Pineda, T. Puig, and J. Fibla, "Association of BST-2 gene variants with HIV disease progression underscores the role of BST-2 in HIV type 1 infection," *Journal of Infectious Diseases*, vol. 207, no. 3, pp. 411–419, 2013.
- [6] J. Mabuka, L. Goo, M. M. Omenda, R. Nduati, and J. Overbaugh, "HIV-1 maternal and infant variants show similar sensitivity to broadly neutralizing antibodies, but sensitivity varies by subtype," *AIDS*, vol. 27, no. 10, pp. 1535–1544, 2013.
- [7] T. N. do Prado, D. B. Brickley, N. K. Hills, E. Zandonade, S. F. Moreira-Silva, and A. E. Miranda, "Factors associated with maternal-child transmission of HIV-1 in southeastern Brazil: a retrospective study," *AIDS and Behavior*, vol. 22, Supplement 1, pp. 92–98, 2018.
- [8] A. Wagner, J. Slyker, A. Langat et al., "High mortality in HIV-infected children diagnosed in hospital underscores need for faster diagnostic turnaround time in prevention of mother-to-child transmission of HIV (PMTCT) programs," *BMC Pediatrics*, vol. 15, no. 1, 2015.
- [9] P. S. Sullivan, O. Hamouda, V. Delpuch et al., "Reemergence of the HIV epidemic among men who have sex with men in North America, Western Europe, and Australia, 1996–2005," *Annals of Epidemiology*, vol. 19, no. 6, pp. 423–431, 2009.
- [10] S. KWC and D. KO, "Decay of soluble CD30 and HIV-1 plasma viral load during early highly active antiretroviral therapy: a short-term longitudinal study," *Journal of Microbial & Biochemical Technology*, vol. 8, no. 2, pp. 90–96, 2016.
- [11] H. M. Sebitloane, J. Moodley, and B. Sartorius, "Associations between HIV, highly active anti-retroviral therapy, and hypertensive disorders of pregnancy among maternal deaths in South Africa 2011–2013," *International Journal of Gynecology & Obstetrics*, vol. 136, no. 2, pp. 195–199, 2017.
- [12] F. Raffi, V. Reliquet, D. Podzamczar, and R. Pollard, "Efficacy of nevirapine-based HAART in HIV-1-infected, treatment-naive persons with high and low baseline viral loads," *HIV Clinical Trials*, vol. 2, no. 4, pp. 317–322, 2001.
- [13] F. Abdoel Wahid, F. R. Sno, E. Darcissac, A. Lavergne, M. R. Adhin, and V. Lacoste, "HIV-1 genetic diversity and drug resistance mutations among treatment-naive adult patients in

- suriname,” *AIDS Research and Human Retroviruses*, vol. 32, no. 12, pp. 1223–1228, 2016.
- [14] M. Chen, Y. Ma, S. Duan et al., “Genetic diversity and drug resistance among newly diagnosed and antiretroviral treatment-naive HIV-infected individuals in western Yunnan: a hot area of viral recombination in China,” *BMC Infectious Diseases*, vol. 12, no. 1, pp. 75–83, 2012.
- [15] M. Putz, N. Dudaş, and A. Isvoran, “Double variational binding—(SMILES) conformational analysis by docking mechanisms for anti-HIV pyrimidine ligands,” *International Journal of Molecular Sciences*, vol. 16, no. 8, pp. 19553–19601, 2015.
- [16] M. Su, J. Tan, and C. Y. Lin, “Development of HIV-1 integrase inhibitors: recent molecular modeling perspectives,” *Drug Discovery Today*, vol. 20, no. 11, pp. 1337–1348, 2015.
- [17] S. Tekeste, T. Wilkinson, E. Weiner et al., “Interaction between reverse transcriptase and integrase is required for reverse transcription during HIV-1 replication,” *Journal of Virology*, vol. 89, no. 23, pp. 12058–12069, 2015.
- [18] N. Deng, W. Zheng, E. Gallicchio, and R. M. Levy, “Insights into the dynamics of HIV-1 protease: a kinetic network model constructed from atomistic simulations,” *Journal of the American Chemical Society*, vol. 133, no. 24, pp. 9387–9394, 2011.
- [19] J.-Y. Wang, H. Ling, W. Yang, and R. Craigie, “Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein,” *The EMBO Journal*, vol. 20, no. 24, pp. 7333–7343, 2001.
- [20] X. Ni, S. Abdel-Azeim, E. Laine et al., “In silico and in vitro comparison of HIV-1 subtypes B and CRF02\_AG integrases susceptibility to integrase strand transfer inhibitors,” *Advances in Virology*, vol. 2012, no. 1, Article ID 548657, p. 13, 2012.
- [21] D. Bonnard, E. le Rouzic, S. Eiler et al., “Structure-function analyses unravel distinct effects of allosteric inhibitors of HIV-1 integrase on viral maturation and integration,” *Journal of Biological Chemistry*, vol. 293, no. 16, pp. 6172–6186, 2018.
- [22] J. Park, J. Yun, Y. Shi et al., “Non-cryogenic structure and dynamics of HIV-1 integrase catalytic core domain by X-ray free-electron lasers,” *International Journal of Molecular Sciences*, vol. 20, no. 8, p. 1943, 2019.
- [23] Z. Hobaika, L. Zargarian, Y. Boulard, R. G. Maroun, O. Mauffret, and S. Fermandjian, “Specificity of LTR DNA recognition by a peptide mimicking the HIV-1 integrase  $\alpha$ 4 helix,” *Nucleic Acids Research*, vol. 37, no. 22, pp. 7691–7700, 2009.
- [24] G. Goodarzi, G. J. Im, K. Brackmann, and D. Grandgenett, “Concerted integration of retrovirus-like DNA by human immunodeficiency virus type 1 integrase,” *Journal of Virology*, vol. 69, no. 10, pp. 6090–6097, 1995.
- [25] E. le Rouzic, D. Bonnard, S. Chasset et al., “Dual inhibition of HIV-1 replication by integrase-LEDGF allosteric inhibitors is predominant at the post-integration stage,” *Retrovirology*, vol. 10, no. 1, pp. 144–189, 2013.
- [26] W. Xue, Y. Yang, X. Wang, H. Liu, and X. Yao, “Computational study on the inhibitor binding mode and allosteric regulation mechanism in hepatitis C virus NS3/4A protein,” *PloS One*, vol. 9, no. 2, article e87077, 2014.
- [27] G. Li, N. A. Meanwell, M. R. Krystal et al., “Discovery and optimization of novel pyrazolopyrimidines as potent and orally bioavailable allosteric HIV-1 integrase inhibitors,” *Journal of Medicinal Chemistry*, vol. 63, no. 5, pp. 2620–2637, 2020.
- [28] M. Patel, B. N. Naidu, I. Dicker et al., “Design, synthesis and SAR study of bridged tricyclic pyrimidinone carboxamides as HIV-1 integrase inhibitors,” *Bioorganic & Medicinal Chemistry*, vol. 28, no. 13, article 115541, 2020.
- [29] L. Q. al-Mawsawi, R. Dayam, L. Taheri, M. Witvrouw, Z. Debysier, and N. Neamati, “Discovery of novel non-cytotoxic salicylhydrazide containing HIV-1 integrase inhibitors,” *Bioorganic & Medicinal Chemistry Letters*, vol. 17, no. 23, pp. 6472–6475, 2007.
- [30] L. Pescatori, M. Métifiot, S. Chung et al., “N-Substituted quinolinonyl diketo acid derivatives as HIV integrase strand transfer inhibitors and their activity against RNase H function of reverse transcriptase,” *Journal of Medicinal Chemistry*, vol. 58, no. 11, pp. 4610–4623, 2015.
- [31] M. Sato, T. Motomura, H. Aramaki et al., “Novel HIV-1 integrase inhibitors derived from quinolone antibiotics,” *Journal of Medicinal Chemistry*, vol. 49, no. 5, pp. 1506–1508, 2006.
- [32] H. Sirous, A. Fassihi, S. Brogi et al., “Synthesis, molecular modelling and biological studies of 3-hydroxypyrrane-4-one and 3-hydroxy-pyridine-4-one derivatives as HIV-1 integrase inhibitors,” *Medicinal Chemistry*, vol. 15, no. 7, pp. 755–770, 2019.
- [33] R. B. Patil and S. D. Sawant, “4D-QSAR studies of coumarin derivatives as HIV-1 integrase 3'-processing inhibitors,” *Medicinal Chemistry Research*, vol. 24, no. 7, pp. 3062–3076, 2015.
- [34] M. Sechi, L. Sannia, F. Carta et al., “Design of novel bioisosteres of  $\beta$ -diketo acid inhibitors of HIV-1 integrase,” *Antiviral Chemistry & Chemotherapy*, vol. 16, no. 1, pp. 41–61, 2005.
- [35] H. Li, C. Wang, T. Sanchez et al., “Amide-containing diketoaids as HIV-1 integrase inhibitors: synthesis, structure-activity relationship analysis, and biological activity,” *Bioorganic & Medicinal Chemistry*, vol. 17, no. 7, pp. 2913–2919, 2009.
- [36] H. Sharma, T. W. Sanchez, N. Neamati et al., “Synthesis, docking, and biological studies of phenanthrene  $\beta$ -diketo acids as novel HIV-1 integrase inhibitors,” *Bioorganic & Medicinal Chemistry Letters*, vol. 23, no. 22, pp. 6146–6151, 2013.
- [37] Y. Li, S. Xuan, Y. Feng, and A. Yan, “Targeting HIV-1 integrase with strand transfer inhibitors,” *Drug Discovery Today*, vol. 20, no. 4, pp. 435–449, 2015.
- [38] L. J. Waters and T. J. Barber, “Dolutegravir for treatment of HIV: SPRING forwards?,” *The Lancet*, vol. 381, no. 9868, pp. 705–706, 2013.
- [39] J. D. Croxtall and L. J. Scott, “Raltegravir,” *Drugs*, vol. 70, no. 5, pp. 631–642, 2010.
- [40] K. Zuo, L. Liang, W. du et al., “3D-QSAR, molecular docking and molecular dynamics simulation of Pseudomonas aeruginosa LpxC inhibitors,” *International Journal of Molecular Sciences*, vol. 18, no. 5, p. 761, 2017.
- [41] G. B. Gonzales, J. van Camp, M. Zotti et al., “Two- and three-dimensional quantitative structure–permeability relationship of flavonoids in Caco-2 cells using stepwise multiple linear regression (SMLR), partial least squares regression (PLSR), and pharmacophore (GALAHAD)-based comparative molecular similarity index analysis (COMSIA),” *Medicinal Chemistry Research*, vol. 24, no. 4, pp. 1696–1706, 2015.
- [42] H. Duan, X. Liu, W. Zhuo et al., “3D-QSAR and molecular recognition of Klebsiella pneumoniae NDM-1 inhibitors,” *Molecular Simulation*, vol. 45, no. 9, pp. 694–705, 2019.
- [43] M. I. Marzouk, S. A. Shaker, A. Abdel Hafiz, and K. Z. el-Baghady, “Design and synthesis of new phthalazinone derivatives containing benzyl moiety with anticipated antitumor activity,”

- Biological & Pharmaceutical Bulletin*, vol. 39, no. 2, pp. 239–251, 2016.
- [44] K. O. Alfarouk, C. M. Stock, S. Taylor et al., “Resistance to cancer chemotherapy: failure in drug response from ADME to P-gp,” *Cancer Cell International*, vol. 15, no. 1, 2015.
- [45] N. N. Wang, J. Dong, Y. H. Deng et al., “ADME properties evaluation in drug discovery: prediction of Caco-2 cell permeability using a combination of NSGA-II and boosting,” *Journal of Chemical Information and Modeling*, vol. 56, no. 4, pp. 763–773, 2016.
- [46] Y. Fu, Y. N. Sun, K. H. Yi et al., “3D pharmacophore-based virtual screening and docking approaches toward the discovery of novel HPPD inhibitors,” *Molecules*, vol. 22, no. 6, p. 959, 2017.
- [47] D. Zhou, Y. Alelyunas, and R. Liu, “Scores of extended connectivity fingerprint as descriptors in QSPR study of melting point and aqueous solubility,” *Journal of Chemical Information and Modeling*, vol. 48, no. 5, pp. 981–987, 2008.
- [48] C. Lu, H. Ke, G. Zhang, H. Xu, and Y. Mei, “An improved weighted extreme learning machine for imbalanced data classification,” *Memetic Computing*, vol. 11, no. 1, pp. 27–34, 2019.
- [49] P. Miniyar, A. Mahajan, D. Anuse et al., “Recursive partitioning analysis and anti-tubercular screening of 3-aminopyrazine-2-carbohydrazide derivatives,” *Letters in Drug Design & Discovery*, vol. 16, no. 11, pp. 1264–1275, 2019.
- [50] H. Seibold, A. Zeileis, and T. Hothorn, “Model-based recursive partitioning for subgroup analyses,” *The International Journal of Biostatistics*, vol. 12, no. 1, pp. 45–63, 2016.
- [51] F. Wickelmaier and A. Zeileis, “Using recursive partitioning to account for parameter heterogeneity in multinomial processing tree models,” *Behavior Research Methods*, vol. 50, no. 3, pp. 1217–1233, 2018.
- [52] L. Chen, Y. Li, Q. Zhao, T. Hou, and H. Peng, “ADME evaluation in drug discovery. 10. Predictions of P-glycoprotein inhibitors using recursive partitioning and naive Bayesian classification techniques,” *Molecular Pharmaceutics*, vol. 8, no. 3, pp. 889–900, 2011.
- [53] J. Yu, P. Ping, L. Wang, L. Kuang, X. Li, and Z. Wu, “A novel probability model for LncRNA–disease association prediction based on the naïve Bayesian classifier,” *Genes*, vol. 9, no. 7, p. 345, 2018.
- [54] I. Rish, “An empirical study of the naive Bayes classifier,” *Journal of Universal Computer Science*, vol. 1, no. 2, p. 127, 2001.
- [55] D. Syafira, S. Suwilo, and P. Sihombing, “Analysis of attribute reduction effectiveness on the naive Bayes classifier method,” *Journal of Physics: Conference Series*, vol. 1566, article 012060, 2020.
- [56] Z. Cheng, Y. Zhang, and W. Fu, “QSAR study of carboxylic acid derivatives as HIV-1 integrase inhibitors,” *European Journal of Medicinal Chemistry*, vol. 45, no. 9, pp. 3970–3980, 2010.
- [57] D. Speller, G. Nusz, and H. D. Hallen, “Cell nucleus manipulation: hydrophobic probe and electric field driven motion,” *Bio-medical Physics & Engineering Express*, vol. 4, no. 4, article 045031, 2018.
- [58] S. J. Pike, J. J. Hutchinson, and C. A. Hunter, “H-Bond acceptor parameters for anions,” *Journal of the American Chemical Society*, vol. 139, no. 19, pp. 6700–6706, 2017.
- [59] F. Russman, S. Marini, E. Peter, G. I. de Oliveira, and F. B. Rizzato, “Self focusing in a spatially modulated electrostatic field particle accelerator,” *Physics of Plasmas*, vol. 25, no. 2, article 023110, 2018.
- [60] J. Chavda and H. Bhatt, “3D-QSAR (CoMFA, CoMSIA, HQSAR and topomer CoMFA), MD simulations and molecular docking studies on purinylpyridine derivatives as B-Raf inhibitors for the treatment of melanoma cancer,” *Structural Chemistry*, vol. 30, no. 6, pp. 2093–2107, 2019.
- [61] L. Farmakis, J. Sakellarakis, A. Koliadima, D. Gavril, and G. Karaiskakis, “Size analysis of barley starch granules by sedimentation/steric field flow fractionation,” *Starch/Staerke*, vol. 52, no. 8-9, pp. 275–282, 2000.
- [62] M. Hou, W. Hu, H. Qiao, W. Li, and X. Yan, “Application of partial least squares (PLS) regression method in attribution of vegetation change in eastern China,” *Journal of Natural Resources*, vol. 30, no. 3, pp. 409–422, 2015.
- [63] J. Hulland, “Use of partial least squares (PLS) in strategic management research: a review of four recent studies,” *Strategic Management Journal*, vol. 20, no. 2, pp. 195–204, 1999.
- [64] J. A. Gil and R. Romera, “On robust partial least squares (PLS) methods,” *Journal of Chemometrics*, vol. 12, no. 6, pp. 365–378, 1998.