WEP Warwick Evans Publishing

Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 1
FBB 2023

Machine Learning-Assisted Discovery of Novel Anti-HIV Drug Candidates: An Analysis Using Molecular Datasets

Zeyu Gou

Chengdu Golden Apple Jin Cheng Number One Secondary School, Chengdu, China

Abstract. Human Immunodeficiency Virus (HIV) serves as a crisis of global public health, necessitating new anti-HIV agents due to the virus's rapid mutation and subsequent drug resistance. While current Combination Antiretroviral Therapy (CART) has helped control the infection and mortality rates, traditional drug development approaches are costly and inefficient. This study aims to address this issue by applying machine learning algorithms for lead compound discovery using the extensive, quality-assured DTP Antiviral Screen Databases. Three molecular datasets—Extended-Connectivity Fingerprints (ECFP), Simplified Molecular-Input Line-Entry System (SMILES), and 2D molecular IMAGES—were processed using Principal Component Analysis (PCA), train-test splitting, and dataset balancing. Six machine learning algorithms were employed, including linear and nonlinear models, optimized through 5-fold cross-validation. The Area Under the Receiver Operating Characteristic (AUROC) curve was utilized to evaluate the models' performance, as well as the macro averaged precision, averaged recall, averaged F1 score, and balanced accuracy metrics. The ensemble models were constructed from the top-performing individual models. The best individual model, a SVM model trained on the ECFP dataset, achieved performance metrics of 0.78 on macro-averaged precision; 0.68 on macro-averaged recall; 0.72 on macro-averaged F1 score; 0.71 on balanced accuracy; and 0.75 on AUROC when evaluated on the testing data. The best ensemble model, fused with SVM, kNN, and logistic regression trained on the ECFP dataset, achieved performance metrics of 0.82 on macro-averaged precision; 0.67 on macro-averaged recall; 0.72 on macro-averaged F1 score; and 0.70 on balanced accuracy when evaluated on the testing data. The models were then applied to a Pubmed-extracted drug dataset, identifying several promising anti-HIV drug candidates, fulfilling the study's objective to improve the efficiency and success rate of new anti-HIV drug screening and discovery. In summary, this research demonstrates the transformative potential of machine learning in accelerating and optimizing the drug discovery process for HIV treatment.

Keywords: Machine Learning, Anti-HIV Drug Prediction, Extended-Connectivity Fingerprints (ECFP), Simplified Molecular-Input Line-Entry System (SMILES), DTP Antiviral Screen Databases, Principal Component Analysis, Support Vector Machine, Ensemble Models.

1. Introduction

1.1 HIV prevalence, ART, and drug resistance

Since the first publicly reported case of HIV infection in June 1981 in the United States [1], HIV has evolved into a persistent pandemic. According to the WHO, HIV has claimed over 40 million lives, with about 39 million people currently living with HIV globally. In 2022, about 1.3 million people became newly infected, while approximately 630,000 died [2]. Although HIV continues to significantly impact our world, substantial progress has been made. New HIV infections have been reduced by 59% since the peak in 1995, and AIDS-related deaths have been reduced by 69% since the peak in 2004 [3].

Antiretroviral therapy (ART) has been the primary contributor to this progress, proven to markedly reduce both the death rate and infection rate [4]. Currently, roughly 76% of all individuals living with HIV have access to ART [3]. Since 1995, the FDA has approved more than 30 HIV drugs that block various HIV replicative stages [5]. The major classes of HIV drugs include Nucleoside Reverse Transcriptase Inhibitors (NRTI) [6] and None Nucleoside Reverse Transcriptase Inhibitors (NNRTI) [7], both of which inhibit reverse transcription; Integrase Inhibitors [8], which prevent the integration

of viral DNA into the host genome; Protease Inhibitors [9], which inhibit viral protein maturation; and Fusion Inhibitors [10], which block the attachment of HIV to host cells. Combination antiretroviral therapy (CART) [11], consisting of at least two drugs from two different classes, has also been developed, significantly extending the life expectancy of HIV-infected patients.

Despite these available drugs, the urgent need for new medications persists. RNA viruses like HIV exhibit high mutation rates [12], lacking the proofreading process present in DNA-based organisms. Over time, drug-resistant mutants emerge under selective pressure from prolonged chemotherapy, diminishing the efficacy of existing drugs and underscoring the continual need for new drug development [13].

1.2 Machine learning and novel drug identification

The design and development of drugs is an intricate and laborious process encompassing stages such as target selection, validation, lead compound discovery, optimization, pre-clinical and clinical trials, and manufacturing [14]. Of all these steps, identifying novel drug candidates is particularly challenging and disheartening. Generally, pre-clinical trials require testing thousands of new compounds, with usually less than ten qualifying for clinical trials; of those, nine out of ten often fail to pass phase two regulatory approvals [15]. Conventionally, resource-intensive in vitro and in vivo experiments are conducted to identify suitable drug candidates, contributing to the high cost and low efficiency of drug development [16].

Computational techniques like virtual screening and molecular docking have been employed to mitigate these challenges [17]. These methods, known as traditional computational approaches, however, are often criticized for their inaccuracy and inefficiency. Machine learning, a subset of artificial intelligence, has emerged as a solution to these shortcomings, offering an unparalleled alternative [18]. It enables researchers to move away from expensive, time-consuming trial-and-error processes, enhances drug development precision, and prioritizes experimentation on the most promising drug candidates. For example, machine learning was used to identify a new drug named MDL-001, which was an effective candidate against the currently circulating virus SARS-COV-2.[19] Also, the Support Vector Machine (SVM) algorithm was used to screen and discover potential anti-depression drug candidates, by in silico screening of serotonin inhibitors from large compound libraries. [20]

However, the success of machine learning in drug discovery relies heavily on the availability of large, high-quality datasets. Public literature data often suffers from noise, heterogeneity, and rarely exceeds one thousand samples, limiting the predictive capability of machine learning models. Fortunately, the extensive, quality-assured DTP Antiviral Screen Databases [21], sponsored by the National Cancer Institute, significantly bolster the development of robust machine learning models for novel HIV drug candidate identification.

1.3 DTP Antiviral Screen Database

The Developmental Therapeutics Program (DTP) AIDS Antiviral Screen assessed 43,850 novel compounds for their anti-HIV capability using a soluble formazan assay [21]. Among them, Compounds were classified as either active, moderately active, or inactive based on their ability to protect human CEM cells from HIV-1 infection. Those providing 100% protection were confirmed active (CA), while those offering at least 50% protection upon retest were listed as moderately active (CM). Compounds neither active nor moderately active were classified as confirmed inactive (CI). Compounds both affirmed and moderately active were regarded as active and labeled as “1”, while compounds confirmed inactive were labeled as “0”, leading to a binary classification task. Information on some evaluated compounds was not uploaded to the datasets, so there is a total number of 40946 compounds available. This large-scale database contains anti-HIV efficacy data on more than 40,000 compounds, providing a solid foundation for developing machine-learning models.

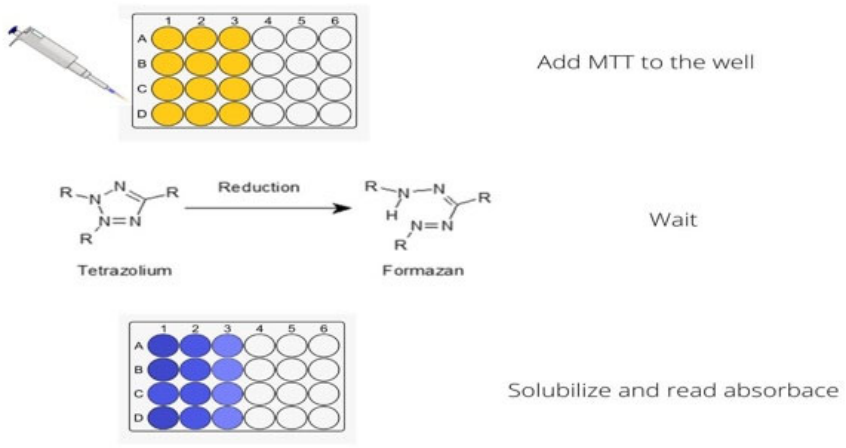

Figure 1. The mechanism of soluble formazan assay for anti-HIV capability validation. Tetrazolium is reduced by viable cells to form soluble, colored formazan product, which is measureable by colorimetric techniques. Therefore, the darker the absorbace, the higher the concentration of viable cells, and thus the more protective of the given compound. https://cellculture.altervista.org/cell-viability-assay-mtt/

1.4 Objectives

Utilizing machine learning algorithms including Linear Regression, Logistic Regression, k-Nearest-Neighbor (kNN), Random Forest, and Support Vector Machine (SVM), based on the large-scale and high-quality DTP Antiviral Screen Databases, this paper aims to develop an effective machine learning model to accurately predict novel drug candidates for anti-HIV treatment. The model performance will be evaluated by various metrics like balanced accuracy, and the best models based on global performance will be applied to a new dataset extracted from Pubmed to identify potential drug candidates. The success of this study could provide valuable tools for the in-silico discovery of anti-HIV drug candidates, fostering the efficiency and success of HIV drug discovery and development.

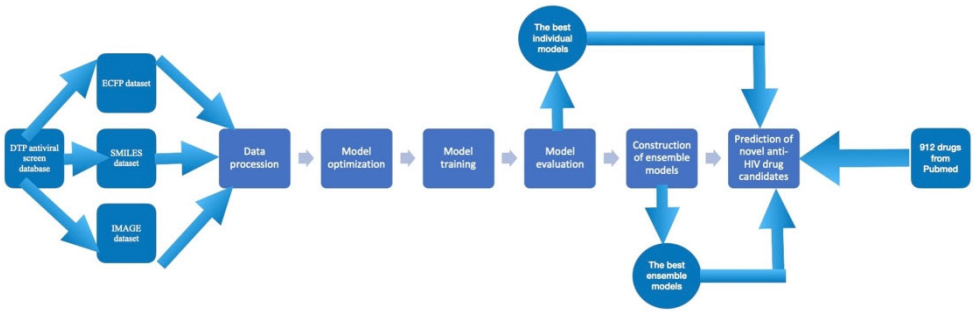

Figure 2. The overall workflow of this study.

2. Methods

2.1 Data Preparation and Preprocessing

2.1.1 Dataset Collection and Initial Processing

Data for this study were meticulously collected from the DTP Antiviral Screen Databases. Three distinct datasets, namely ECFP, SMILES, and IMAGE, were chosen to provide a comprehensive understanding of the molecular properties of 40,946 novel compounds. This multi-dataset approach facilitates a well-rounded analysis by highlighting different molecular attributes.

2.1.2 Comprehensive Feature Engineering and Description

This study employs three distinct but complementary datasets for feature engineering, crucial for the machine learning models. These datasets are Extended-Connectivity Fingerprints (ECFP), Simplified Molecular-Input Line-Entry System (SMILES), and the IMAGE dataset. Each offers a unique approach to capture molecular characteristics: ECFP focuses on generating multi-dimensional feature vectors that encapsulate various molecular attributes; SMILES provides a string-based representation focusing on the arrangement and connectivity of atoms; and the IMAGE dataset converts the molecular structures into grayscale images to capture spatial relationships. The combination of these datasets allows for a robust and comprehensive representation of molecular structures, thereby enhancing the predictive performance of the machine learning algorithms employed. The following paragraphs describe each method in detail.

ECFP is a state-of-the-art fingerprinting technique widely used in computational biology and drug discovery. In this project, ECFP was employed to generate a 2048-dimensional binary feature vector for each of the 40,946 novel compounds. These features encapsulate a wide array of molecular characteristics, such as types of atoms, bonding configurations, and unique ring structures, thereby providing a comprehensive molecular signature.

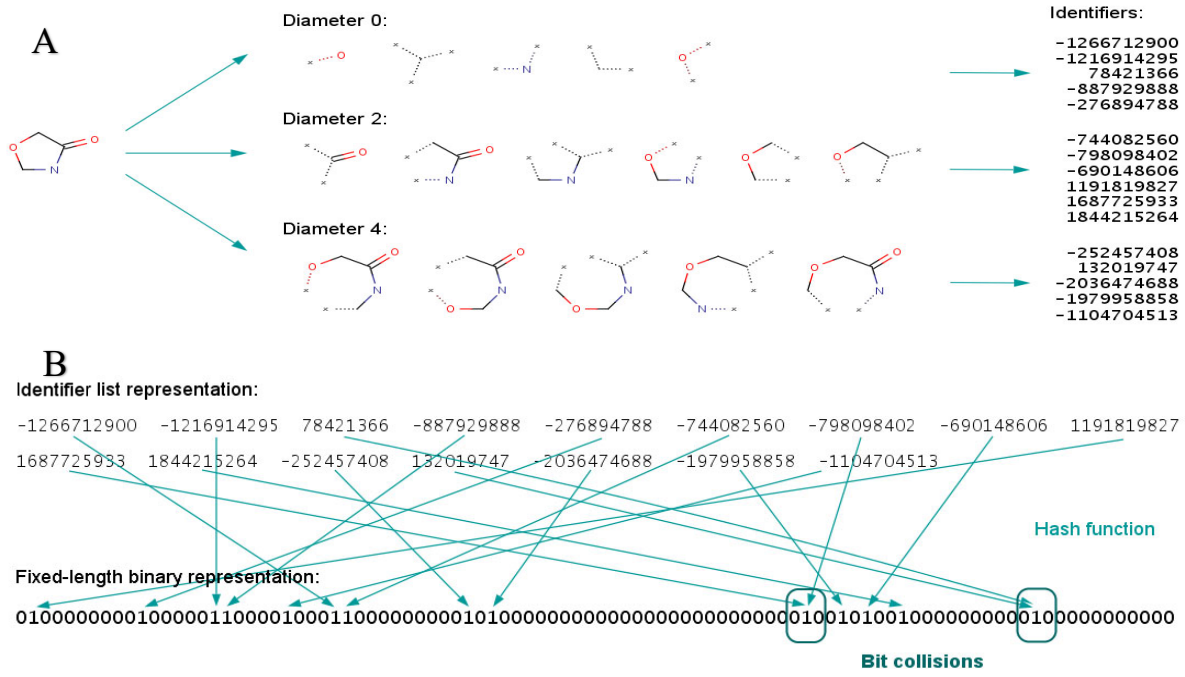

Figure 3. ECFP generation process. (A) Firstly, identifiers were assigned to structures with different diameter levels. (B) Identifiers were converted to the fixed-length binary presentation using a hash function. https://docs.chemaxon.com/display/docs/extended-connectivity-fingerprint-ecfp.md.

The SMILES dataset provides a compact, ASCII string representation of the molecular structure. (Tabel 1) The SMILES feature was embedded into a binary form using fingerprinting from rdkit. (Figure 1) Each string encodes crucial information about the arrangement and connectivity of atoms. This encoding method is highly efficient and captures the core structural elements of each molecule.

Table 1. Examples of the SMILES presentation of 3 compounds: Zidovudine, Zalcitabine, and Molnupiravir.

Compound name	SMILES structure
Zidovudine	CCC1=[O+][Cu-3]2([O+]=C(CC)C1)[O+]=C(CC)CC(CC)=[O+]2
Zalcitabine	CC(=O)N1c2ccccc2Sc2c1ccc1ccccc21
Molnupiravir	Nc1ccc(C=Cc2ccc(N)cc2S(=O)(=O)O)c(S(=O)(=O)O)c1

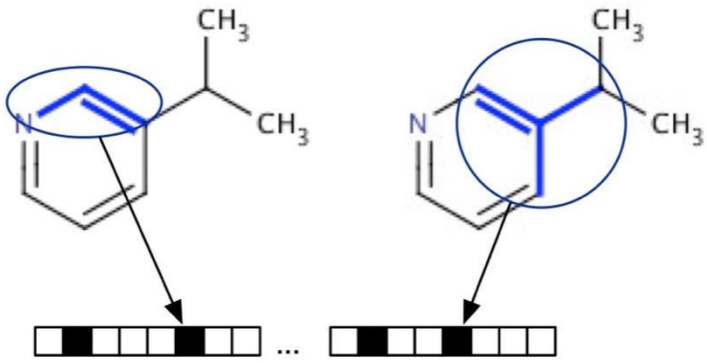

Figure 4. Fingerprinting of the SMILES dataset. A kernel was applied to a molecule to generate a bit vector. Each fingerprint bit corresponds to a fragment of the molecule.

The IMAGE dataset is that molecular structures were converted into 80x80 pixel grayscale images, resulting in 6400 binary-encoded features. This visual modality provides insights into spatial relationships between atoms and bonds, offering a unique perspective that traditional encoding methods might overlook.

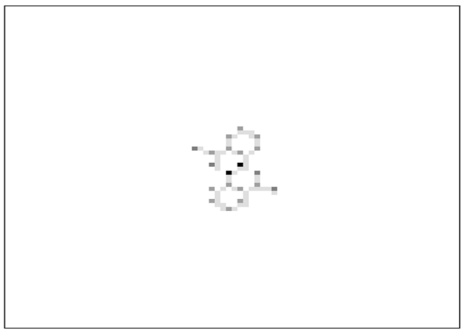

Figure 5. An example of the 80x80 pixel grayscale image of Abacavir succinate

2.1.3 Advanced Dimensionality Reduction Using PCA

To address the curse of dimensionality and associated computational challenges, Principal Component Analysis (PCA) was applied. This statistical technique transforms the original high-dimensional space into a new coordinate system. The primary aim was to reduce the feature set to a manageable size while retaining the maximum possible variance from the original data. This technique also serves to increase the difference between the number of features and the number of instances across the datasets, reducing the possibility of overfitting. The explained variance ratio was

carefully calculated and analyzed to ensure that the new reduced feature set was still representative of the original data.

2.1.4 Stratified Data Splitting and Balancing

The train_test_split method from scikit-learn was adapted to divide the datasets into training and testing sets at a 3:1 ratio. A special stratification protocol was followed to ensure that compounds with and without benzo rings were equally represented in both sets. The rdkit.chem library was employed to achieve this stratification, creating two subsets of compounds based on the presence of benzo rings. This stratification strategy is essential as it ensures that both the training and testing datasets maintain a consistent ratio of compounds with and without benzo rings, thereby minimizing biased learning and enhancing model generalizability. Because the training dataset is significantly unbalanced (drug labelled as inactive= 29675; drug labelled as active=1035), risking the classifications to biases towards the overrepresented group and thus will adversely affect the reliability of the models, drug balancing was conducted by increasing the number of active compounds ten-fold to 11385. The performance of prediction on both the original and the manually balanced datasets was evaluated.

2.2 Model Development and Optimization

2.2.1 Classification Models and Their Characteristics

In our quest for optimized drug discovery, a diversified approach to model selection was applied, consisting of linear, nonlinear, and hybrid models. Linear models such as Linear Regression and Logistic Regression were primarily utilized for their simplicity and interpretability, serving as benchmark models. On the other hand, Nonlinear models like K-Nearest-Neighbors (KNN) and Random Forest were chosen for their proficiency in capturing complex relationships within the data. Finally, the Support Vector Machine (SVM) offers a unique adaptability, capable of functioning as both a linear and nonlinear model depending on the chosen kernel. The ensemble of these models aims to capitalize on the strengths of each, offering a more robust and accurate predictive system for anti-HIV drug discovery. The following paragraphs describe each model in detail.

(1) Linear Models. Both Linear Regression and Logistic Regression were used, given their simplicity and ease of interpretation. While they are generally less flexible and suffer from low predictive accuracy, they serve as good baseline models against which more complex models can be compared.

(2) Nonlinear Models. K-Nearest-Neighbors (KNN) and Random Forest were chosen for their ability to capture complex relationships in the data. These models are generally more flexible and can achieve higher accuracy but at the cost of computational intensity and interpretability.

(3) Support Vector Machine (SVM) is unique in its flexibility; it can operate as both a linear and nonlinear model depending on the kernel used. This makes it a versatile choice for various kinds of data patterns.

2.2.2 Hyperparameter Optimization and Cross-Validation

An exhaustive grid search was performed over a defined hyperparameter space for each model type. Five-fold cross-validation was employed during this optimization phase. The performance was primarily judged based on accuracy and F1-score, and the models with the best performance were selected for model training and further evaluation.

2.3 Final Testing and Candidate Drug Prediction

2.3.1 Rigorous Evaluation of Testing Dataset

The well-trained and established models underwent thorough testing using a dataset that includes 10,236 chemical compounds. We tested our final models on a dataset of 10,236 compounds. We used different ways to measure how well our models can classify the compounds. One way is called AUROC, which shows how good our models are at finding true positives (TPs) and avoiding false

positives (FPs). Another way is called precision, which shows how often our models are correct when they say something is positive. A third way is called recall, which shows how often our models can find all the positive cases. A fourth way is called F1 score, which combines precision and recall into one number. We also used macro-averaging and balanced accuracy to make sure our models are fair and not biased by the unequal number of positive and negative cases in the dataset. By examining all these metrics, we can gauge the model's overall ability to correctly classify data.

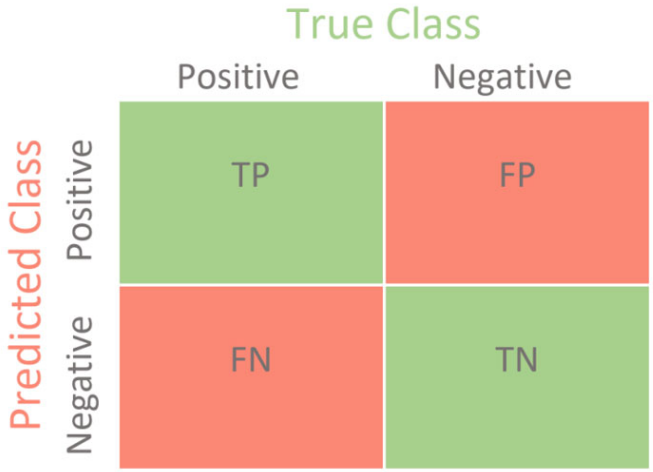

Figure 6. The confusion matrix. TP, FP, TN, and FN's definition is demonstrated.

https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

2.3.2 Ensemble Learning and Selection of Best Models

A vote ensemble approach was employed, combining the strengths of the 3 top-performing models. Global performance was considered to select the best ML models for constructing ensemble models. We gave each model a score based on how well it did for each way of measuring. The best model got 3 points, the second best got 2 points, and the third best got 1 point. We did not count the F1 score. We added up the scores for each model and ranked them from highest to lowest. This strategy aims to mitigate the weaknesses of individual models, thereby enhancing the overall predictive accuracy. The best-performing models among the individual and ensemble models were then used for potential drug prediction.

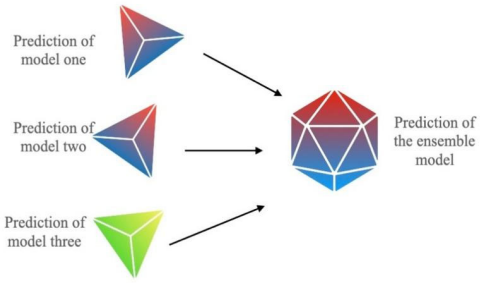

Figure 7. The mechanism of vote ensemble. A vote ensemble decides the prediction by the prediction of most of the models(in this case, two or three)

2.4 Identification of New Drug Candidates

An additional set of 912 antiviral compounds, previously not known for their efficacy against HIV, were extracted from PubChem. These compounds were evaluated using the best-performing models to flag potential new candidates for HIV treatment.

3. Results

3.1 Data Preparation and Feature Engineering

3.1.1 Dataset Stratification and Partitioning

The ECFP, SMILES, and IMAGE datasets underwent a systematic preprocessing phase. A stratified sampling technique was employed to partition the compounds into two distinct categories: those containing benzo rings (30,296) and those without (10,650).

3.1.2 Dimension Reduction with Principal Component Analysis

To combat the curse of dimensionality, PCA was applied as a dimensionality reduction technique. For the ECFP dataset, which originally contained 2048 features, the top 200 principal components were found to retain over 99% of the total variance and were adopted. This is an important finding, confirming that while the dimensionality was drastically reduced, the integrity of the dataset remained largely intact. For the SMILES and IMAGE datasets, with 167 and 6400 features respectively, a reduced feature space using the top 100 principal components was adopted.

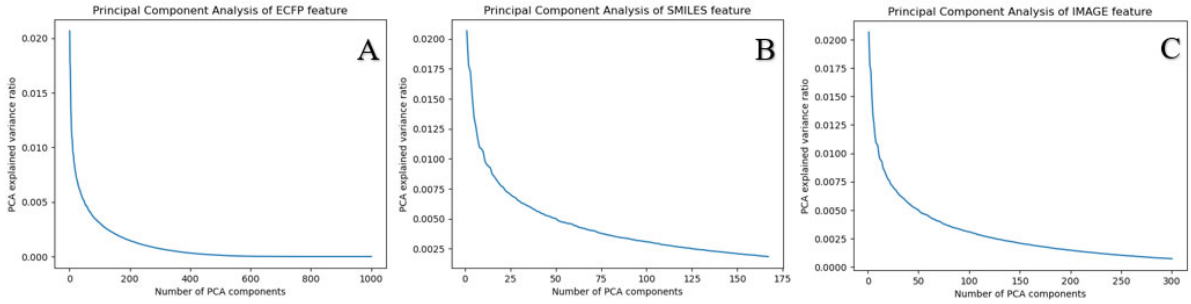

Figure 8. PCA explained variance ratio for each of the three datasets. For the ECFP dataset, explained variance ratio of the first 1000 components are shown (A); for the SMILES dataset, the first 175 components are shown (B); and for the IMAGE dataset, 300 components are shown (C).

3.2 Hyperparameter Tuning and Model Construction

A wide array of machine learning algorithms were evaluated, including Lasso and Ridge (linear regression models), Logistic Regression, KNN, Random Forest, and SVM (both linear and non-linear models). An exhaustive grid search methodology combined with 5-fold cross-validation was employed to optimize the hyperparameters for each model. For example, in the case of SVM on the ECFP dataset, multiple 'C' values and kernels were tested, ranging from 0.01 to 3 for "C", and linear, poly, rbf, and sigmoid for kernel. The most accurate SVM model for the ECFP dataset used a polynomial kernel and a 'C' value of 0.5 (Figure 9).

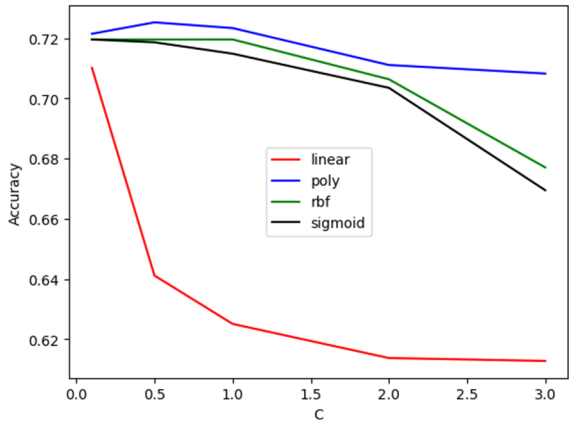

Figure 9. 5-fold cross validation for SVM model on the ECFP dataset. Different colored lines indicate different kernels. C is a hyperparameter used to control the margin of the hyperplane

3.3 Training and Performance Metrics

3.3.1 Training Set Accuracy

After hyperparameter tuning, the selected models were then trained using their respective training sets. The training accuracy was subsequently calculated, falling within a range of 0.72 to 0.76 across all datasets. (Figure 10) This tight range of training accuracy indicates a low likelihood of model overfitting. SVM emerged with the highest training accuracy consistently across all three datasets, which could imply its suitability for this specific type of classification task. However, the training accuracy alone can not sufficiently explain the ML model's authentic performance, which was further evaluated on the testing data.

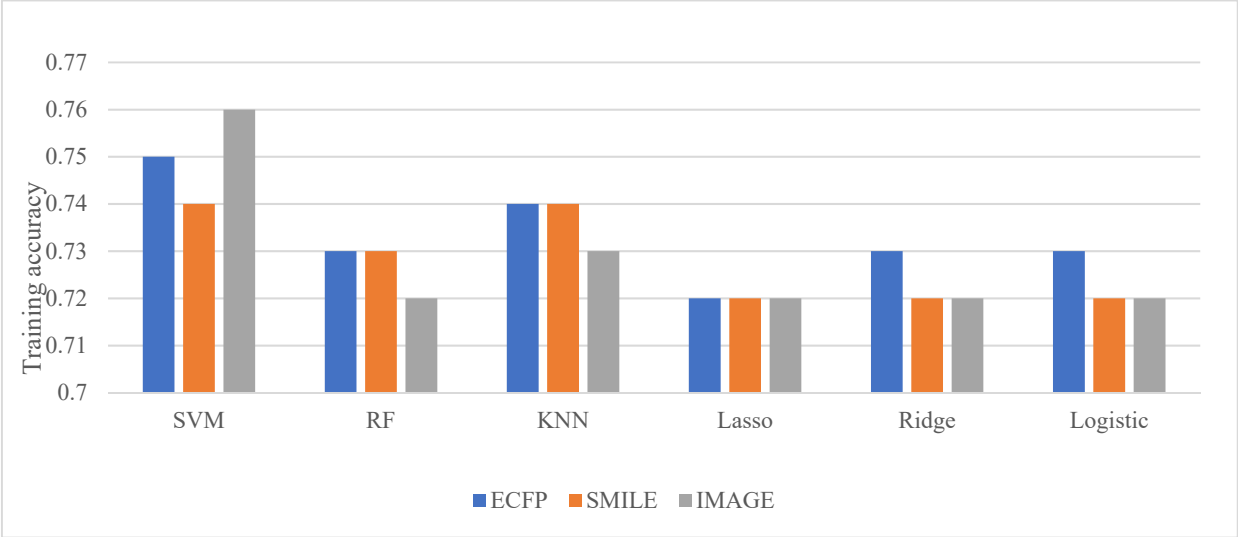

Figure 10. Training accuracy for ML models. Blue bars indicate training accuracy for models trained on the ECFP dataset. Orange bars indicate training accuracy for models trained on the SMILES dataset. Gray bars indicate training accuracy for models trained on the IMAGE dataset.

3.3.2 Model Evaluation Metrics

Models were rigorously evaluated using a plethora of metrics, AUROC (The ROC curves of SVM are depicted in Figure 11, serving as an example), precision, recall, F1 score, and balanced accuracy (Figure 12). For instance, the SVM model achieved a laudable F1 score of 0.71 on the ECFP dataset, which speaks volumes about its discriminative power between active and inactive compounds.

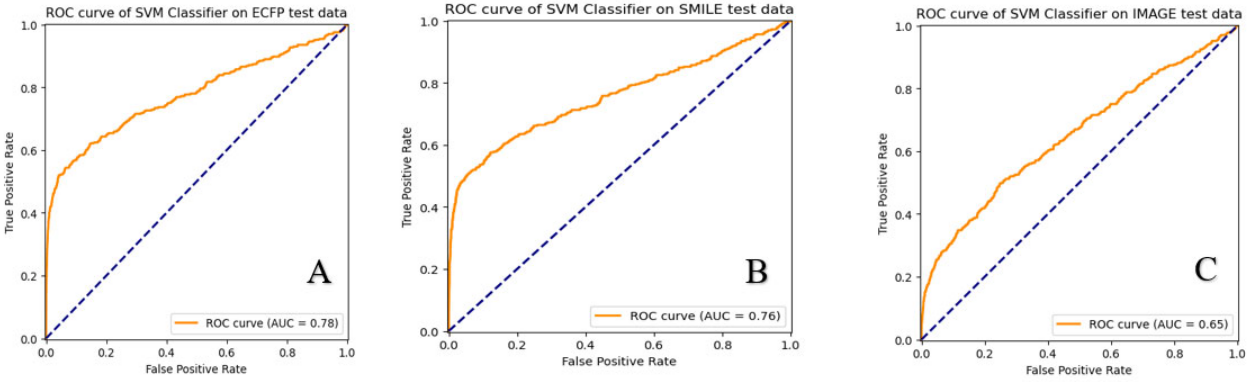

Figure 11. ROC curves of the SVM classifiers on the ECFP (A), SMILES (B), and IMAGE (C) datasets. The left image depicts the ROC curve of SVM on the ECFP dataset. The middle image depicts the ROC curve of SVM on the SMILES dataset. The right image depicts the ROC curve of SVM on the IMAGE dataset. The yellow curve is and the ROC curve, the AUROC scores are shown

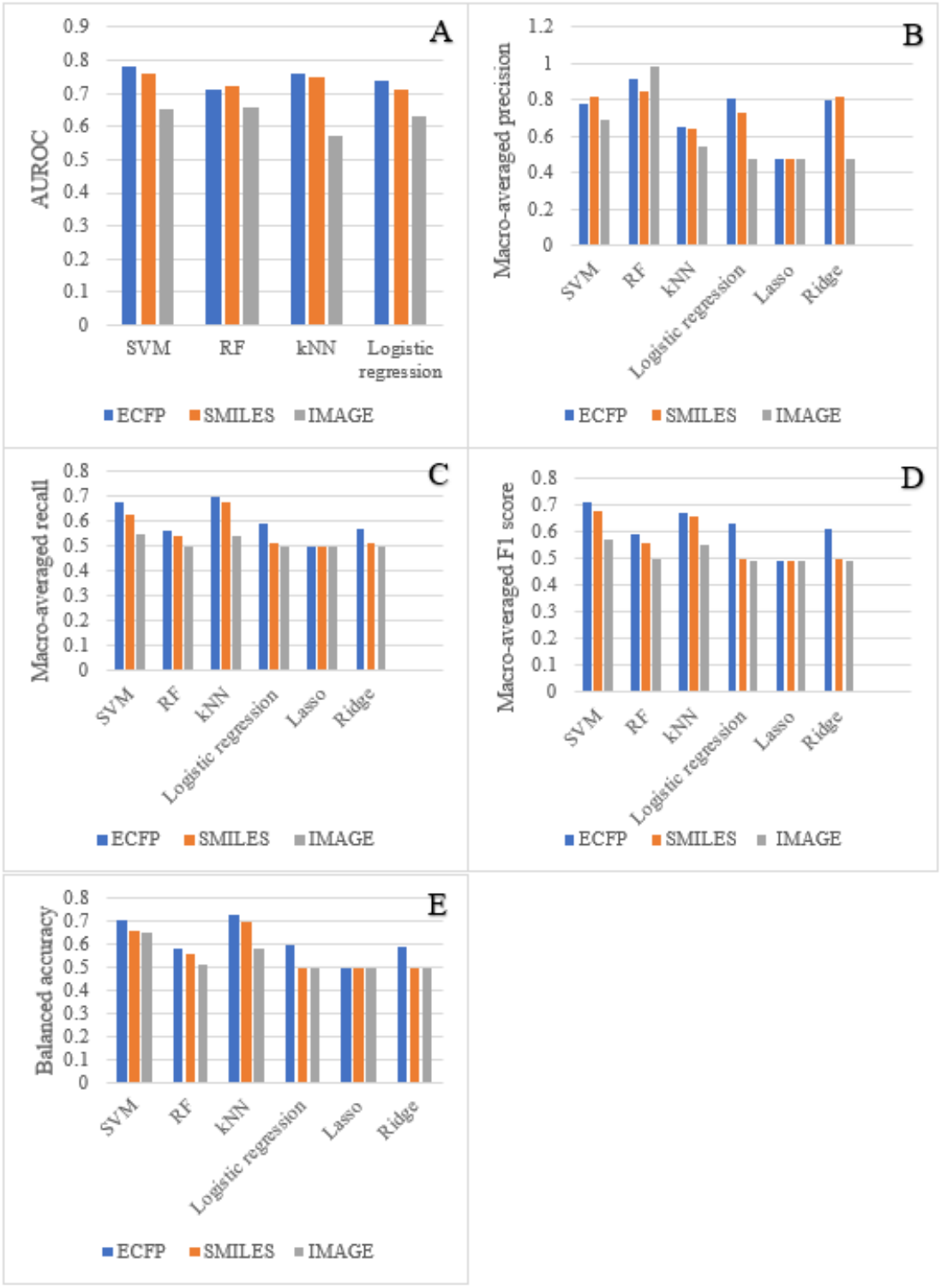

Figure 12. AUROC (A), macro-averaged precision (B), macro-averaged recall (C), macro-averaged F1 score (D) and balanced accuracy (E) for ML models. The 6 ML models were: support vector machine (SVM), random forest (RF), k-nearest neighbors (kNN), logistic regression, ridege regression and lasso regression.

3.4 Ensemble Approaches and Performance

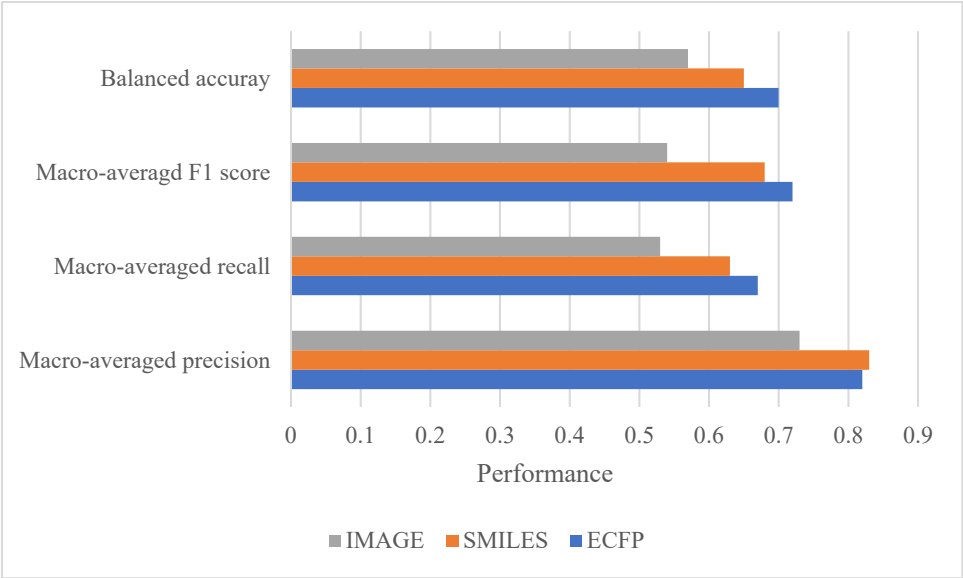

Figure 13. Performance metrics for the ensemble models. Four performance metrics were examined, and the bar colors indicate the datasets.

Ensemble models were formulated using the top three individual models based on their ranks in global performance. For example, an ensemble model for the ECFP dataset was crafted using SVM, kNN, and Logistic Regression. These ensemble models recorded F1 scores of 0.72, 0.68, and 0.54 for the ECFP, SMILES, and IMAGE datasets respectively. The complete sets of performance of ensemble models are demonstrated below. (Figure 13)

3.5 External Validation: Identifying New Drug Candidates

An independent dataset comprising 912 antiviral drugs, sourced from PubMed and not previously tested for HIV efficacy, was used for external validation. Due to the subpar performance of the IMAGE-based models (see section 4.5), they were excluded from this analysis. In total, four models were deployed for this crucial validation step, two each from the ECFP and SMILES datasets based on their global performance: the SVM models and ensemble models. When tested on the dataset, the top-performing individual model (using SVM on ECFP) yielded scores of 0.78, 0.68, 0.72, 0.71, and 0.75 for macro-averaged precision, recall, F1 score, balanced accuracy, and AUROC, respectively. Meanwhile, the highest-rated ensemble model (combining SVM, kNN, and logistic regression with ECFP) recorded scores of 0.82 for macro-averaged precision, 0.67 for recall, 0.72 for F1 score, and 0.70 for balanced accuracy.

The results were highly promising, with 21 drugs being flagged by multiple models as potential anti-HIV candidates. Among them, 3 are constantly predicted as active by all the validating models: Zidovudine, [[(2S,3R,5R)-3-azido-5-(5-methyl-2,4-dioxo-pyrimidin-1-yl)tetrahydrofuran-2-yl]methoxy-hydroxy-phosphoryl] phosphonic hydrogen phosphate, and Combivir. The RF models were then used to supplement this finding due to their extremely high precision (0.92 for ECFP, 0.85 for SMILES), which further confirmed the second and third compound as active. Fozivudine tidoxil was also confirmed by the RF models while only not predicted as active by the SVM model trained on the SMILES dataset.

Table 2. Prediction of the potential anti-HIV drug candidates. 6 models were used to perform this prediction task. However, since the RF model trained on SMILES does not predict any compound as "active", it is not shown on this table. The check mark indicates that the given compound is predicted as "active" by the given model. Only a fraction of some compounds' full name are shown. The full name of all the compounds in this table and their SMILES sturture are exhibited in the appendix. The most promising drug candidates were bolded

Compound n	SVM trained on SMILES	SVM trained on ECFP	Ensemble trained on SMILES	Ensemble trained on ECFP	RF trained on ECFP
Zidovudine	✓	✓	✓	✓	
3'-Azido-2',3'	✓	✓	✓		
Fosamprenavi	✓		✓		
Lamivudine ar	✓		✓		
Amprenavir	✓		✓		
Fosamprenavi	✓		✓		
Darunavir	✓		✓		
Darunavir etha	✓		✓		
Fosamprenavi	✓		✓		
4-Chloro-8-m	✓	✓	✓		
Elvitegravir	✓	✓	✓		
3'-Azido-2',3'	✓		✓		
5-[1-(3-carbo	✓	✓	✓		
[[(2S,3R,5R)-	✓	✓	✓	✓	✓
Cosalane	✓	✓	✓		
5-[1-(3-carbo	✓	✓	✓		
Navuridine	✓		✓		
Fozivudine ti	✓	✓	✓		✓
Combivir	✓	✓	✓	✓	✓
Tivirapine	✓		✓		
(2-decoxy-3-c	✓		✓		

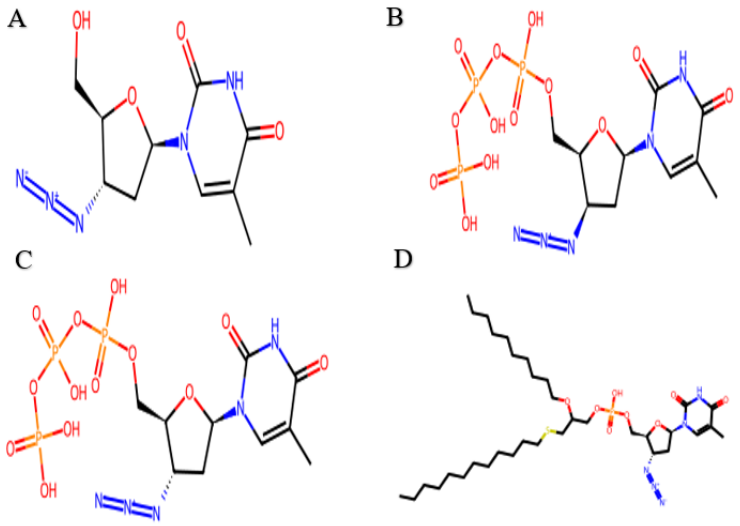

Figure 14. The 4 most promising drug candidates. (A) Zidovudine; (B) [[(2S,3R,5R)-3-azido-5-(5-methyl-2,4-dioxo-pyrimidin-1-yl) tetrahydrofuran-2-yl]methoxy-hydroxy-phosphoryl] phosphono hydrogen phosphate; (C) Combivir; (D) Fozivudine tidoxil.

4. Discussion

4.1 PCA visualization

Besides reducing the features' dimensionality, PCA is a tool to visualize how the compounds' ECFP, SMILES and IMAGE features relate to their status in anti-HIV activeness. The PCA revealed that active compounds have similar ECFP, SMILES, and IMAGE features to those that are inactive. Since distinct clusters of active/inactive drugs are not apparent in Figure 15, it is indispensable to investigate supervised machine learning algorithms for the prediction task.

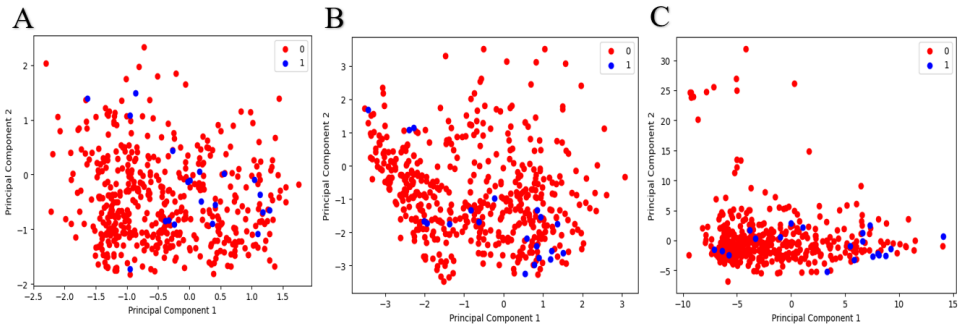

Figure 15. Visulization of the relationship between compounds. ECFP (A), SMILES (B), and IMAGE (C) features and whether they are active, using principal component analysis. PC1: principal component 1, PC2: principal component 2.

4.2 Comparative Analysis of Feature Representations

The study employed three diverse molecular encodings—ECFP, SMILES, and IMAGE. These datasets encoded complementary aspects of the molecular structures, each with its unique advantages and limitations. ECFP, with its 2D topological properties, was especially conducive for linear models like Logistic and Ridge Regression. These models assume a linear relationship between features, making them naturally compatible with ECFP. On the other hand, the simplified connectivity patterns in SMILES and the spatial visual features in IMAGE proved more challenging for linear models, which performed significantly worse on these datasets.

Specifically, Logistic Regression and Ridge Regression ranked 3rd and 4th in global performance when trained on ECFP but fell to the bottom ranks on SMILES and IMAGE datasets. The gap in performance of evaluating metrics on ECFP and SMILES datasets is also the highest for linear models. (Figure 12) This suggests that ECFP's feature set contains more linear relationships that are predictive of activity, corroborating the importance of feature selection based on the nature of the model.

4.3 Importance of Data Preprocessing

4.3.1 stratification

One crucial step was the preprocessing of data with stratification based on the presence or absence of benzo rings. For all performance metrics, models trained on possessed datasets performed better than those trained on datasets without stratification (Figure 16). This indicates that by maintaining identical ratios of this important substructure across training and testing sets, the models were better equipped to generalize to new, unseen data, thus fulfilling a key aim of the study.

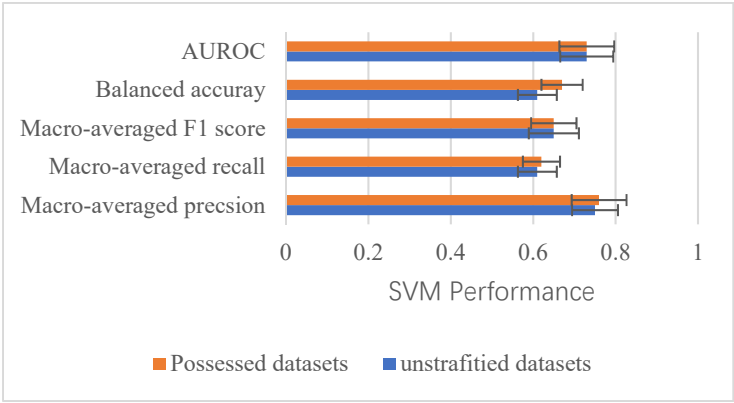

Figure 16. Average performance across all datasets for SVM on the possessed datasets and the unstratified datasets. Orange bars indicate average performance for the SVM model trained on the possessed datasets. Blue bars indicate average performance for the SVM model trained on the unstratified datasets. Relationship between performance trained on possessed and unstratified datasets for other models follow the same trend as the SVM model

4.3.2 Dataset balancing

There is a change in performance resulting from increasing the balance between the two classes within the training datasets (active vs. inactive). However, across the models, while dataset balancing increases some performance metrics, others decrease, which suggests that dataset balancing does not efficiently lead to better performance. SVM serves as an example. (Figure 17)

However, it is generally recognized that dataset balancing is required to produce reliable ML models to prevent biases toward the overrepresented group.

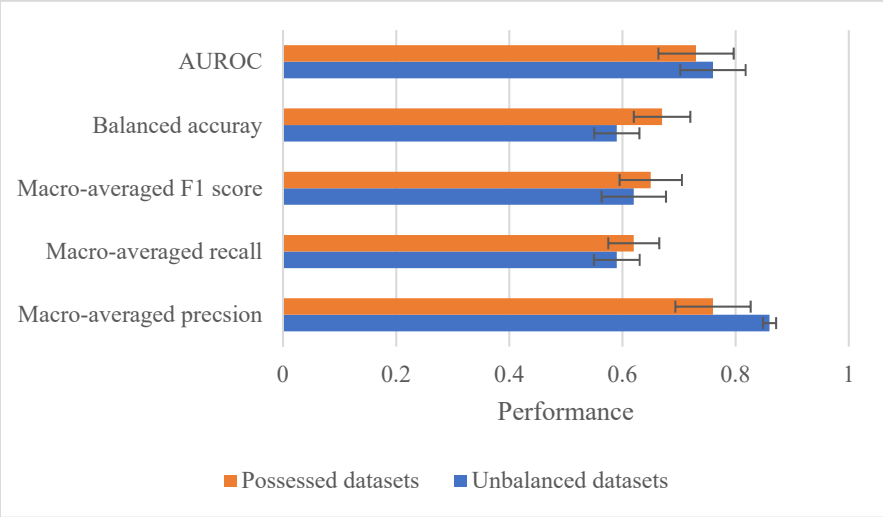

Figure 17. Average performance across all datasets for SVM on the possessed datasets and the unbalanced datasets. The orange bars are average performance for the SVM models being trained on the possessed datasets. Blue bars indicate average performance for the SVM models being trained on the datasets which are unbalanced. While balanced accuracy, F1 score, and recall increases, AUROC and precision decreases after dataset balancing

There are two methods commonly adopted for dataset balancing: depletion of instances in the overrepresented class or replication of instances in the underrepresented class. The second method was used in this study since it makes full use of all the available data.

4.4 Model Evaluation

4.4.1 Analysis of the Performance Metrics

Due to the imbalance of the instances of the two classes within the testing dataset (Drugs labelled as inactive = 9863; drugs labelled as active = 374), the testing accuracy rate significantly overestimates the models' performance (larger than 95 for all models), and cannot help distinguishing the ML models. Therefore, it is essential to use other evaluating metrics to testify the ML models' performance. In essence, the measures show that our top-performing models (refer to Section 3.5) are adept at forecasting potential anti-HIV drug candidates. Notably, the precision values, especially in the RF models, highlight a minimal likelihood of falsely identifying an inactive drug as active. This accuracy in avoiding type 2 errors is crucial in our research; it means researchers can rely on the model's drug activity predictions, minimizing the chances of investing in drugs that yield false positives. RF models' extremely high precision explains why they are used to supplement prediction.

4.4.2 Non-linear v.s. Linear Models

Given the study's focus on predictive power over interpretability, non-linear models like SVM and KNN naturally excelled. The high dimensionality and inherent complexity of the datasets made it unlikely for linear models to capture all nuances, which is also true for the ECFP dataset, although it is relatively more linear than the other datasets. This was evident from the consistently high performance of SVM (tuned with non-linear kernels) across all datasets. The SVM model tuned with

non-linear kernels performed far better than when tuned with the linear kernel on all three datasets, which is exemplified by the one trained on the ECFP dataset (Figure 9), highlighting the importance of non-linear effects in the models.

4.4.3 Ensemble Modeling: A Double-Edged Sword

Although ensembling was expected to enhance model performance, it provided only marginal improvements on the ECFP dataset and even led to worse results on the IMAGE dataset compared with the best individual models. This discrepancy suggests that the top models might contain redundant predictive information, offering minimal benefits when combined.

4.5 Limitations of the IMAGE Dataset

The models trained on the IMAGE dataset notably underperformed, with an average macro-averaged F1 score and average balanced accuracy below 0.6, and average AUROC below 0.7. Models trained on the IMAGE dataset also performed the worst among those trained on all datasets (Figure 18). This points to the dataset's poor generalization capabilities, likely due to the low-resolution 2D visual patterns being insufficient to capture key structural features related to the activity. Thus, models trained on the IMAGE dataset were excluded from the prediction task.

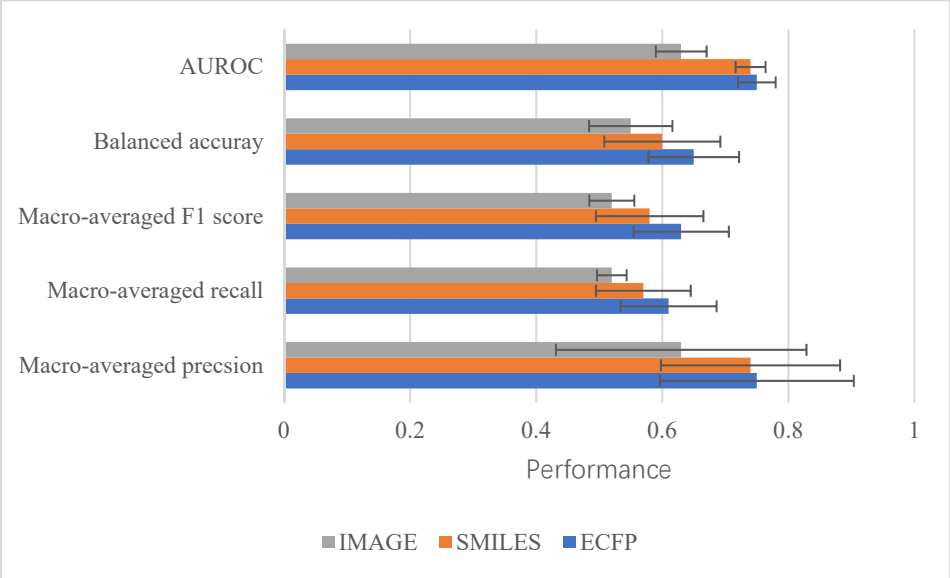

Figure 18. Average performance of machine learning models for the IMAGE, SMILES, and ECFP datasets. Models trained on the IMAGE dataset performed the poorest in all metrics.

4.6 Application to Real-world Data

Applying the top-performing models to an external drug dataset led to the identification of 21 promising anti-HIV candidates. Interestingly, there is a significant overlap in the predictions across the applied models. All of the predicted potential drug candidates were validated by at least two models, indicating a high probability of activateness of the drugs and thus ensuring the success of this study. Besides, the four most promising drugs (Figure 14), along with the other 7 drugs predicted as active, all contain a nucleoside-like structure, which shows that they might function as nucleoside reverse transcriptase inhibitors (NRTI).

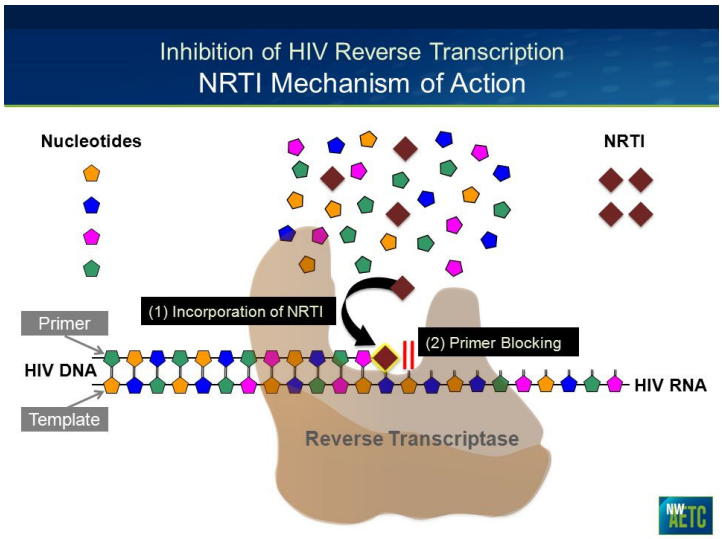

Figure 19. The mechanism of Nucleoside Reverse Transcriptase Inhibitor. NRTIs block HIV replication by inserting themselves into the new forming viral DNA strand, preventing the formation of the phosphodiester bond and thus stop reverse transcription.

https://twitter.com/ID_fellows/status/1296490837996601344

4.7 Future Directions

Future work could involve leveraging advanced techniques like Deep Graph Library and Graph Recurrent Neural Networks to expand datasets synthetically. Transfer learning from related targets could provide a boost in predictive power. The use of alternative molecular representations, such as graph-based features, could also be explored to enhance learning capabilities. One promising direction is the utilization of more advanced machine learning methodologies like Deep Graph Library (DGL) and Graph Recurrent Neural Networks (GRNN). These techniques could enable the synthetic expansion of molecular datasets, making the machine learning models more robust and accurate. DGL and GRNN offer the advantage of capturing complex topological features and dynamic changes in molecular structures, thereby potentially improving predictive performance. Furthermore, transfer learning from related targets or diseases could be a game-changer. Learning the underlying patterns in one context and transferring that knowledge to a similar but not identical scenario could significantly enhance the model's predictive accuracy. This is particularly beneficial when dealing with rare or less-studied diseases where data might be sparse. Expanding beyond the scope of ECFP, SMILES, and IMAGE datasets, the use of alternative molecular representations such as graph-based features could be explored. These graph-based approaches could offer a more nuanced understanding of molecular characteristics, potentially leading to more accurate predictions. Building upon the success of the ensemble models in this study, future work could also focus on constructing more sophisticated ensemble techniques that incorporate a blend of linear, nonlinear, and advanced models. This would aim to capitalize on the strengths of each type of model, thereby creating a more robust predictive system. Additionally, as more antiviral databases become publicly available, incorporating them into the existing pipeline would allow for a more diversified and comprehensive dataset. This would improve the generalizability of the model, which is critical for real-world applications. The future directions in this field should focus on the integration of advanced computational methods, diversification of feature representation, and refinement of ensemble strategies, all aiming to advance the efficiency and accuracy of machine learning-assisted drug discovery for HIV treatment.

4.8 Conclusion

In this study, 3 comprehensive molecular datasets containing 40946 compounds were extraced from Pubmed. Drugs were labeled as two classes (active/inactive). Six machine learning algorithms were

applied to those datasets after PCA, dataset stratification and balancing. 5-fold CV was used to optimize model performance, and the performance of ML models were analyzed by a variety of metrics. The best models according to global performance were used to build ensemble models, which minorly increased the models' performance. The best 2 models trained on the ECFP and SMILES datasets after ensembling were applied for potential anti-HIV drug prediction, supplemented by the RF models with particularly high precision. 21 drugs were predicted as potential anti-HIV drug candidates, and 4 among them were extremely promising. This study provides a tool for machine learning to accelerate drug discovery. It allows experimental efforts to focus on a more narrowed down, missing subset of compounds, thereby increasing the efficiency and success rate of the expensive and time-consuming drug discovery pipeline.

Acknowledgement

My first visit to a local hospital that is specialized for HIV-infected patients 8 years ago opened my curiosity for HIV. Since then, I have dived to explore it, both within and beyond the school curriculum. I have visited similar hospitals for many times, and I found that many patients suffer from drug resistance and high-drug price. The way to counteract both, as I think, is to increase the efficiency and decrease the costs of the laborious drug development process. And machine learning can significantly help. I learnt machine learning from JHU's biomedical engineering innovation summer school, and with the encouragement and help from my biology teacher, Miss Tong, I decided to apply machine learning for this study that aims to facilitate lead compound identification for HIV drug development. I strengthened my understanding of machine learning through the Edx course: introduction to data science with python, and the book An Introduction to Statistical Learning with Applications in Python. I would like to appreciate Miss Tong's patient and compassionate guide, which helps me to choose the topic, solve problems that I encountered, and write this scientific paper.

References

[1] Vergis, E. N., & Mellors, J. W. (2000). Natural history of HIV-1 infection. Infectious disease clinics of North America, 14(4), 809–vi. https://doi.org/10.1016/s0891-5520(05)70135-5 10. Menéndez-Arias, L., & Delgado, R. (2022). Update and latest advances in antiretroviral therapy. Trendsinpharmacologicalsciences, 43(1),16–29. https://doi.org/10.1016/j.tips.2021.10.004.

[2] World Health Organization. HIV and AIDS. https://www.who.int/data/gho/data/themes/hiv-aids. Published 13 July 2023.

[3] UNAIDS. The Path That Ends AIDES. https://unaids.org/en. Published 2023.

[4] Menéndez-Arias, L., & Delgado, R. (2022). Update and latest advances in antiretroviral therapy. Trendsinpharmacologicalsciences, 43(1),16–29. https://doi.org/10.1016/j.tips.2021.10.004

[5] Tompa, D. R., Immanuel, A., Srikanth, S., & Kadhirvel, S. (2021). Trends and strategies to combat viral infections: A review on FDA approved antiviral drugs. International journal of biological macromolecules, 172, 524–541. https://doi.org/10.1016/j.ijbiomac.2021.01.076

[6] Arribas J. R. (2004). The rise and fall of triple nucleoside reverse transcriptase inhibitor (NRTI) regimens. TheJournalofantimicrobialchemotherapy, 54(3),587–592. https://doi.org/10.1093/jac/dkh384

[7] Li, G., Wang, Y., & De Clercq, E. (2022). Approved HIV reverse transcriptase inhibitors in the pastdecade. ActapharmaceuticaSinica.B, 12(4),1567–1590. https://doi.org/10.1016/j.apsb.2021.11.009

[8] Scarsi, K. K., Havens, J. P., Podany, A. T., Avedissian, S. N., & Fletcher, C. V. (2020). HIV-1 Integrase Inhibitors: A Comparative Review of Efficacy and Safety. Drugs, 80(16), 1649–1676. https://doi.org/10.1007/s40265-020-01379-9

[9] Walmsley S. (2007). Protease inhibitor-based regimens for HIV therapy: safety and efficacy. Journal of acquired immune deficiency syndromes (1999), 45 Suppl 1, S5–S31. https://doi.org/10.1097/QAI.0b013e3180600709

[10] Xiao, T., Cai, Y., & Chen, B. (2021). HIV-1 Entry and Membrane Fusion Inhibitors. Viruses, 13(5), 735. https://doi.org/10.3390/v13050735

[11] Domingo, P., & Vidal, F. (2011). Combination antiretroviral therapy. Expert opinion on pharmacotherapy, 12(7), 995–998. https://doi.org/10.1517/14656566.2011.567001

[12] Nomaguchi, M., Doi, N., Koma, T., & Adachi, A. (2018). HIV-1 mutates to adapt in fluxing environments. Microbesandinfection, 20(9-10),610–614. https://doi.org/10.1016/j.micinf.2017.08.003.

[13] Bandera, A., Gori, A., Clerici, M., & Sironi, M. (2019). Phylogenies in ART: HIV reservoirs, HIV latency and drug resistance. Current opinion in pharmacology, 48, 24–32. https://doi.org/10.1016/j.coph.2019.03.003.

[14] Berdigaliyev, N., & Aljofan, M. (2020). An overview of drug discovery and development. Future medicinal chemistry, 12(10), 939–947. https://doi.org/10.4155/fmc-2019-0307.

[15] Umscheid, C. A., Margolis, D. J., & Grossman, C. E. (2011). Key concepts of clinical trials: a narrativereview. Postgraduatemedicine, 123(5),194–204. https://doi.org/10.3810/pgm.2011.09.2475

[16] . Saeidnia, S., Manayi, A., & Abdollahi, M. (2015). From in vitro Experiments to in vivo and Clinical Studies; Pros and Cons. Current drug discovery technologies, 12(4), 218–224. https://doi.org/10.2174/1570163813666160114093140

[17] Giordano, D., Biancaniello, C., Argenio, M. A., & Facchiano, A. (2022). Drug Design by Pharmacophore and Virtual Screening Approach. Pharmaceuticals (Basel, Switzerland), 15(5), 646. https://doi.org/10.3390/ph15050646

[18] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nature reviews. Drug discovery, 18(6), 463–477. https://doi.org/10.1038/s41573-019-0024-5.

[19] Model medicine. Model Medicines' Oral Anti-COVID-19 Drug Candidate MDL-001 Found to Significantly Reduce Viral Load in Lungs; Accepted into NIH's Antiviral Program for Pandemics (APP). https://www.prnewswire.com/news-releases/model-medicines-oral-anti-covid-19-drug-candidate-mdl-001-found-to-significantly-reduce-viral-load-in-lungs-accepted-into-nihs-antiviral-program-for-pandemics-app-301477223.html. Published 08 Feb, 2022.

[20] Shi, Z., Ma, X. H., Qin, C., Jia, J., Jiang, Y. Y., Tan, C. Y., & Chen, Y. Z. (2012). Combinatorial support vector machines approach for virtual screening of selective multi-target serotonin reuptake inhibitors from large compound libraries. Journal of molecular graphics & modelling, 32, 49–66. https://doi.org/10.1016/j.jmgm.2011.09.002.

[21] Weislow, O. S., Kiser, R., Fine, D. L., Bader, J., Shoemaker, R. H., & Boyd, M. R. (1989). New soluble-formazan assay for HIV-1 cytopathic effects: application to high-flux screening of synthetic and natural products for AIDS-antiviral activity. Journal of the National Cancer Institute, 81(8), 577–586. https://doi.org/10.1093/jnci/81.8.577