Comparing Genetic Evolutionary Algorithms on Three Enzymes of HIV-1: Integrase, Protease, and Reverse Transcriptome

Nafiseh Masroor *masro001@cougars.csusm.edu*
Student / Computer Science and Information System
California State University San Marcos
San Marcos, 92096, USA

Jack Wang *wang071@cougars.csusm.edu*
"Student / Computer Science and Information System
California State University San Marcos
San Marcos, 92096, USA

Bita Pouyanfar *pouya001@cougars.csusm.edu*
"Student / Computer Science and Information System
California State University San Marcos
San Marcos, 92096, USA

Yanyan Li *yali@csusm.edu*
Faculty / Computer Science and Information System
California State University San Marcos
San Marcos, 92096, USA

Ahmad Hadaegh *hadaegh@csusm.edu*
Faculty / Computer Science and Information System
California State University San Marcos
San Marcos, 92096, USA

Abstract

In this work, we utilized Quantitative Structure-Activity Relationship (QSAR) techniques to develop predictive models for inhibitors of the HIV-1 enzymes Integrase, HIV-Protease, and Reverse Transcriptase. Each predictive model was composed of quantitative drug characteristics that were selected by genetic evolutionary algorithms, such as Genetic Algorithm (GE), Differential Evolutionary Algorithm (DE), Binary Particle Swarm Optimization (BPSO), and Differential Evolution with Binary Particle Swarm Optimization (DE-BPSO). After characteristic selection, each model was tested with machine-learning algorithms such as Multiple Linear Regression (MLR), Support Vector Machine (SVM), and Multi-Layer Perceptron neural networks (MLP/ANN). We found that a combination of DE-BPSO combined with Multi-Layer Perceptron produced the most accurate predictive models as measured by R^2, the statistical measure of proportion of variance in prediction values, and root-mean-square-error (RMSE) of prediction values compared to observed values. As for the models themselves: the best predictors for Integrase inhibitor included mass-weighted centred Broto-Moreau autocorrelation values, Moran autocorrelations, and eigenvalues of Burden matrices weighted by I-states; the best predictors for HIV-Protease inhibitors included the second Zagreb index value, the normalized spectral positive sum from Laplace matrix, and the connectivity-like index of order 0 from edge adjacency mat; and the best predictors for Reverse Transcriptase inhibitors included the number of hydrogen atoms, the molecular path count of order 7, the centred Broto-Moreau autocorrelation of lag 2 weighted by Sanderson electronegativity, the P_VSA-like on ionization potential, and the frequency of C – N bonds at topological distance 3.

Keywords: Genetic Evolutionary Algorithms, HIV, Data Predictive Data Mining.

1. INTRODUCTION

The human immunodeficiency virus type 1 (HIV-1) is a retrovirus that is the causative agent for the life-threatening acquired immunodeficiency syndrome (AIDS) [1]. HIV originated in non-human primates in Central and West Africa in 1920. It infects vital cells in the human immune system, such as Helper T cell [2], and destroys CD 4 cells [3]. These cells help the body to fight infections. In 2019, an estimated 38 million people were infected across the globe [4].

In this research, we studied 3 types of HIV-1 enzymes to identify the best novel drug candidates for this virus. To date, many anti-HIV-1 drugs have been approved by the U.S. Food and Drug Administration(FDA). However, mutations arising in the HIV-l genome that confer resistance to existing anti-HIV-1 inhibitors drive the need to develop new anti-HIV-l drugs with an acceptable mutation profile [5].

There are two types of HIV, HIV-1, and HIV-2. HIV-1 is more dangerous because it can be transmitted more easily, and it causes most HIV infections annually. Therefore, this thesis will study HIV-1. HIV-1 contains three important enzymes which are essential for virus replication such as Integrase (IN), Protease (PR), and Reverse Transcriptase (RT). After HIV has bound to the target cell, The HIV- RNA and various enzymes, including these three enzymes are injected into the cell.

Integrase is a 288-amino-acid, 32-kDa viral enzyme that mediates the linkage of double-stranded viral DNA into the host cell genome. After the translocation of a large complex derived from the viral core from the cytoplasm into the nucleus Integration occurs. Once integrated, the provirus can be considered for most purposes to be a stable genetic element remaining for the life of the cell and, through cellular replication, for the life of the individual [6].

HIV-1 Protease is a dimeric enzyme from the family of aspartic proteases with C2 symmetric in the free form, containing 99 amino acids in both of its chains. This enzyme is an essential enzyme of HIV replication and is a vital target for drug design strategies to fight AIDS. HIV-1 Protease converts the immature virions into a mature virion through the cleavage of precursor polypeptides [7].

Reverse Transcriptase is responsible for the conversion of the single-stranded genomic RNA into double-stranded DNA. Reverse Transcriptase is an essential step in retroviral replication. The sequences of the genomes show how Reverse Transcriptase is pervasive, not only do these genomes contain a large number of endogenous retroviruses, but also a variety of retropulsion and reverse-transcribed elements [8].

In this research, we have selected Quantitative Structure - Activity Relationship (QSAR) models because they are the best tools for regression or classification models used widely in biological engineering. Predictor variables X relate to the potency of the response variable Y in the regression models. In QSAR modeling the predictors consist of molecular descriptors. QSAR models usually apply to drug discovery and lead optimization. The accuracy of input data, the selection of appropriate descriptors, and statistical tools have a significant impact on the results of the QSAR models.

In this research, evolutionary algorithms such as the Genetic Algorithm(GA), the Differential evolution Algorithm (DE), the Binary Particle Swarm Optimization Algorithm (BPSO), and the Differential evolution- Binary Particle Swarm Optimization Algorithm (DE-BPSO) are used to predict the best inhibitors for three HIV-1 enzymes. These evolutionary algorithms analyze the data with the help of the Multiple Linear Regression, the Support Vector Machine (SVM), and the Artificial Neural Network (ANN) machine learning models.

After getting all the data and the cleaning process, we had accurate data to work on the research and our development. To converge the models to their best forms. We executed each model for 1000 iterations and compared the results in each HIV-1 enzyme to find the ones that fit the best.

In the second part of this thesis, we get the results for all three HIV-1 enzymes with the help of the Orange Data Mining software and compare the results with the previous results.

The rest of this paper is organized as follows. We explain the related work in section 2. The architecture of the optimization schemes will be illustrated in section 3. Next, we summarize the implementation of the programs using the machine learning tools. Finally, we present the results of the 36 combinations we described above and compare them with each other.

2. RELATED WORKS

The work of Gene M. Ko and others [10] concentrates on the HIV-1 Integrase Inhibitors. In their study, they developed a hybridized EA-based feature selection method for developing QSAR models using DE and BPSO. The DE-BPSO algorithm is used to develop MLR based QSAR models for the analysis of inhibition of HIV-1 integrase. There are some differences between their study with this work. First, they only worked on HIV-1 integrase. This work compares all three enzymes: HIV-1 Protease, HIV-Integrase, and HIV, and Reverse Transcriptase. The second difference is that they just used DE-BPSO for their analysis but in this study, we used all the Genetic Evolutionary Optimization schemes (GA, DE, BPSO, and DE-BPSO) to compare the results with each together. Further, they only used linear MLR in their analysis. This work used two linear models, MLR and SVM, and one non-linear model ANN. We looked at the results of 36 different combinations and each combination only required around 1000 iteration. Models were converging in less than 2 hours.

Another project by Jiali Tang [15] developed a repository database system of drugs, drug features, and drug targets where data can be mined and analyzed. In their study, they used the Genetic Evolution(GE) algorithm. They used Multiple Linear Regression(MLR), Partial Least Square Regression(PLSR), and Support Vector Machine(SVM) for their implementations. In this database service and web application, anyone even without computer science experience can utilize a data mining application for drug discovery. This includes upload the experimental input, run tests, and download test results.

The third work is the Galvan [16] study. In this study, he implemented a DE-BPSO feature selection algorithm and AdaBoost Random Forest Regression learner [17] to develop a non-linear QSAR model for the analysis of Aryl β-Diketo Acids targeting the inhibition of HIV-1 integrase protein enzyme by identifying the physiochemical molecular descriptors that exhibit the greatest influence on the crystallization of the integrase enzyme’s binding mechanism. An experiment was run on two sets of 37 and 91 dimeric Aryl β-Diketo Acids partitioned into training, validation, and test sets. Results found descriptors with the greatest inhibitory effect on the biological activity of β-diketo acids are those related to molecular volume, topology, and electrostatic effects, with large molecular volumes having the greatest impact.

3. METHODOLOGY

Using preexisting data on drug efficacy, we built a machine learning setup that inductively determines the best drug characteristics to compose predictive models that can predict the possible efficacy of as yet untested HIV drugs.

3.1. Data Preparation

Our data for for inhibitors of Integrase, HIV-1 Protease, and Reverse Transcriptase were obtained from the Binding Database [9]. The efficacy data was downloaded in TSV (Tab-Separated Value) format and the raw molecular structural data was downloaded in SDF (Spatial Data File) format. Drug efficacy is delineated by IC50, the quantitative measure that indicates how much of an inhibitory substance (e.g. drug) is needed to inhibit, in vitro, a given biological process or biological component by 50%.

To obtain more accurate results the descriptors were normalized and IC50 values were converted to PIC50.

$$PIC(50) = -log\big(IC(50)\big)$$

As for our descriptor values, we used the DRAGON software [14] to extract those values and remove overly linearly dependent descriptor types to reduce superfluous dimensionality. In addition to the preprocessing by DRAGON, our program eliminates all the data that are not accurate such as empty rows or columns, rows without PCI(50), columns with null or zero values, etc.

3.2. Machine Learning Algorithm

This research used four different optimization techniques: Genetic Algorithm(GA), Differential Evolution Algorithm (DE), the Binary Particle Swarm Optimization (BPSO), and Differential evolution - Binary Particle Swarm Optimization (DE-BPSO) to identify a model that can predict the best drug (inhibitors) candidates.

3.3. Genetic Algorithm

Introduced by John Holland, the genetic algorithm is a search-based optimization technique that finds the best solution by mimicking the process of natural selection, in which each slight variation that is useful is passed onto the next generation.

In the genetic algorithm, we first initialize a matrix of N by M matrix where N is the number of rows (models) and M is the number of descriptors. In this work, we used 50 as the number of rows for all methods. We noticed that with more than 50 rows, the program becomes complex and takes longer than usual. With less than 50 rows, we did not have enough models to compare with others. Each row is an indication of one model in which each time 1.5% of the descriptors are selected randomly for training. We did this execution with 1000 iterations and 50 rows in each iteration. The indicators to compare the models were based on five factors: R^2 of training, R^2 of testing, R^2 of validation, Root Mean Square Error (RMSE), and fitness of each model. The closer the value of R^2s to 1, the better the model. The closer the values of fitness and RMSE to 0, the better is the prediction. All values should be positive. Even though all factors should be included when we compare the results, fitness is the best element to move from one iteration to another. The fitness algorithm is:

$$f = \sqrt{\frac{(m_t - n - 1) \cdot RMSE_t{}^2 + m_v \cdot RMSE_v{}^2}{m_t - n - 1 + m_v}}$$

where n is the total number of data (compounds), m_t is the number of samples in the training set, m_v is the number of samples in the validation set and $RMSE_t$ and $RMSE_v$ are the RMSE of training and validation, respectively.

Based on the value of the fitness received from each row of the population, the two best rows are picked to move to the next round of calculation. The two best models are called “Parent 1” and “Parent 2” in the GA algorithm. Since Parent 1 and Parent 2 refer to the best models of the first iteration, we automatically move them to the second iteration. Next, based on the GA algorithm, if parent 1 and parent 2 are the best rows of the first iterations, it is assumed that any children created from these rows could be good rows as well. The GA algorithm created the children and goes from one iteration to another to find the best fitness.

3.4. Differential Evolution Algorithm

The differential evolution algorithm(DE) is an algorithm that uses evolution, where the fittest individual of a population are the ones that produce more offspring that inherit the good traits of the parents. This algorithm is a strong algorithm for finding the minimum of a function calls it black-box optimization. In DE the fitness function which must be minimized is $f(x): R^n \rightarrow R$. This function produces a fitness of the output. The goal is to find the global minimum. DE does not guarantee to find the global minimum. In a complex function, we need to have more iterations to find a good approximation.

3.5. Binary Particle Swarm Optimization Algorithm

The binary particle swarm optimization (BPSO) algorithm is the binary form of particle swarm optimization that is a meta-Heuristic algorithm. This algorithm is generated based on the social behavior of the school of fish or flock of birds. For example, in the flock of birds if one bird can find the location of food then the bird will give the information to other birds in the folk. In this algorithm, the birds are the particles, and the folks are the population of particles.

3.6. Differential Evolution - Binary Particle Swarm Optimization (DE-BPSO)

DE-BPSO is a feature selection algorithm that attempts to search a multi-dimensional feature space to find the optimal features of a dataset that can be used to build the best model. To do this, the algorithm attempts to mimic the social behavior of animal life. The algorithm iteratively searches the feature space, provided by the dataset, with everyone remembering its current and best local position, while the swarm maintains the global best position. A model is created using the features that an individual, which is based on its current position. After using the trained model to make predictions, the fitness function is used to determine the strength of the model and the individual. All the optimization algorithms are detailed in [10].

4. IMPLEMENTATION

To build our scripts for this research, we utilized the Python programming language. We decided on Python because of its concise, easily understandable syntax and its wide variety of machine learning libraries. The machine learning algorithms used in this research are the Multiple Linear Regression(MLR), the Support Vector Machine (SVM), and the Artificial Neural Network (ANN) models.

4.1. Multiple Linear Regression(MLR)

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression [19].

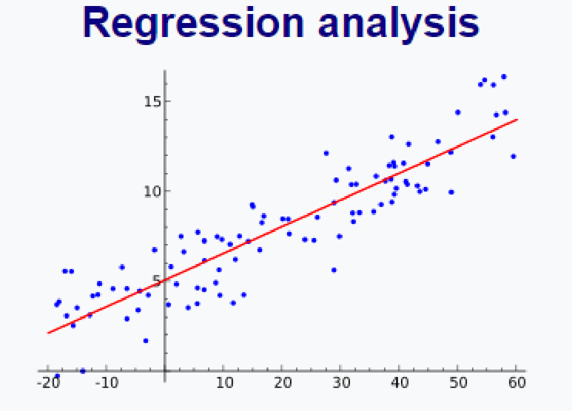


FIGURE 1: Multiple Linear Regression.

Given a data set

$$\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^{n} \qquad (1)$$

of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressors x is linear. This relationship is modeled through a disturbance term or error variable ε — an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and regressors. Thus, the model takes the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \cdots = X_i^T \beta + \cdots., for i = 1, \ldots n \tag{2}$$

where ^{T} denotes the transpose so that $x_i^T\beta$ is the inner product between vectors x_i and β. This is shown in Figure 1.

4.2. Support Vector Machine (SVM)

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. When data are unlabeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The support vector clustering algorithm applies the statistics of support vectors that are developed in the support vector machines algorithm, to categorize unlabeled data. The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

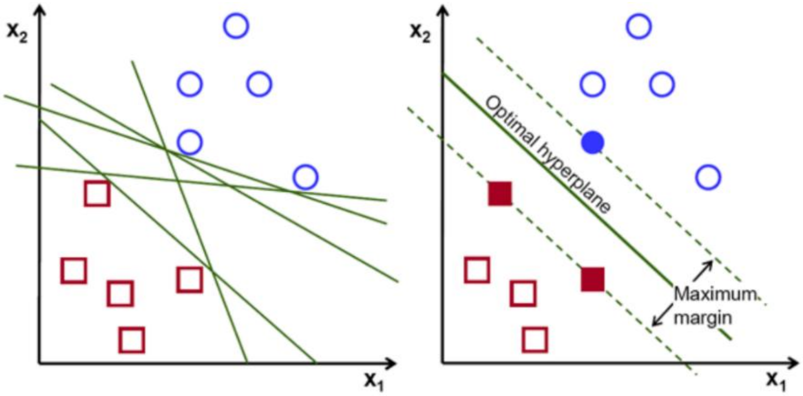


FIGURE 2: Support-Vector Clustering Algorithm.

To separate the two classes of data points, many possible hyperplanes could be chosen. Our objective is to find a plane that has the maximum margin, that is the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. This is shown in Figure 2.

4.3. Artificial Neural Network (ANN)

An artificial neural network (ANN) is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron receives a signal, processes it, and can signal neurons connected to it. The "signal" at a connection is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs. The connections are called edges. Neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Neurons may have a threshold such that a signal is sent only if the aggregate signal crosses that threshold.

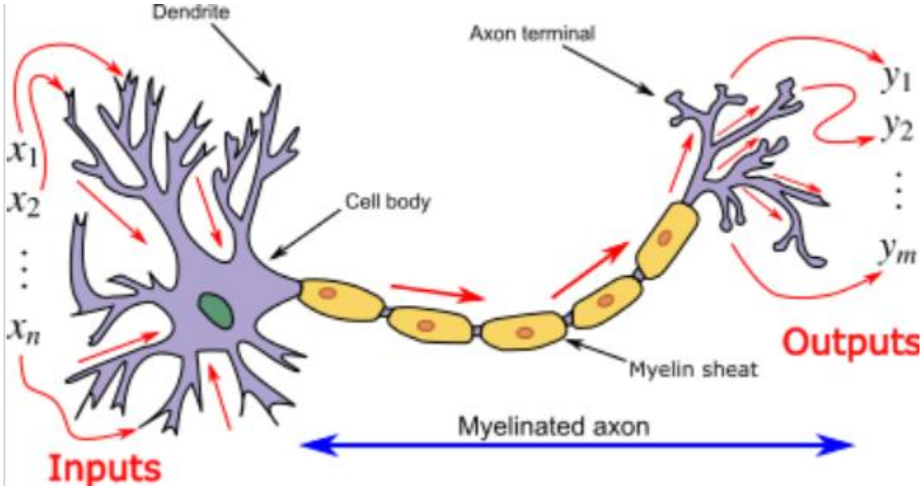


FIGURE 3: Artificial Neural Network.

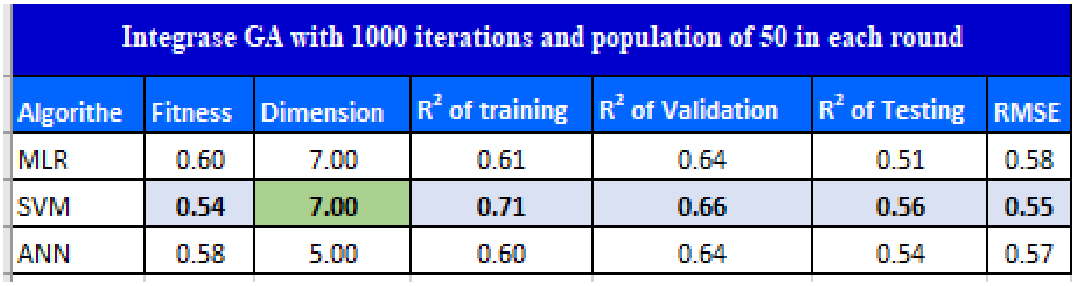
Supervised learning uses a set of paired inputs and desired outputs. The learning task is to produce the desired output for each input. In this case, the cost function is related to eliminate incorrect deductions (See Figure 3).

A commonly used cost is the mean-squared error, which tries to minimize the average squared error between the network's predicted and the desired output. Tasks suited for supervised learning are pattern recognition and regression. Supervised learning is also applicable to sequential data.

5. ANALYSIS OF THE RESULTS

5.1. Integrase

5.1.1. Genetic Algorithm

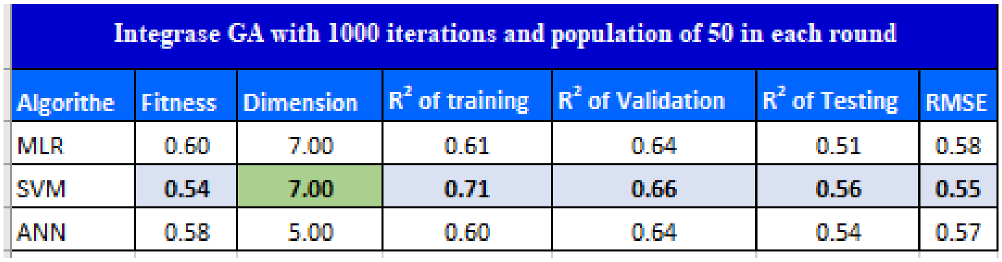


Integrase GA with 1000 iterations and population of 50 in each round

Algorithe	Fitness	Dimension	R^2 of training	R^2 of Validation	R^2 of Testing	RMSE
MLR	0.60	7.00	0.61	0.64	0.51	0.58
SVM	**0.54**	**7.00**	**0.71**	**0.66**	**0.56**	**0.55**
ANN	0.58	5.00	0.60	0.64	0.54	0.57

TABLE 1: Result of Integrase-GA.

Table 1 shows the result of Integrase-GA. The result for the SVM is relatively better than the MLR and ANN. As mentioned earlier, a better model will have lower values for the “Fitness”, “Dimensions”, and “RMSE” and have a “R^2” value closer to one. We also need to make sure that the values of R^2 for training, validation, and testing are similar to within < 0.2. Otherwise, it could be an indicator that the model is overfitted.

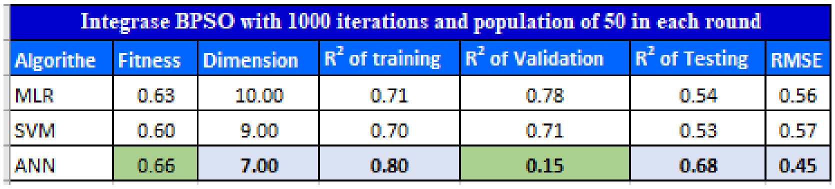


Integrase GA with 1000 iterations and population of 50 in each round						
Algorithe	Fitness	Dimension	R^2 of training	R^2 of Validation	R^2 of Testing	RMSE
MLR	0.60	7.00	0.61	0.64	0.51	0.58
SVM	**0.54**	**7.00**	**0.71**	**0.66**	**0.56**	**0.55**
ANN	0.58	5.00	0.60	0.64	0.54	0.57

TABLE 2: Result of Integrase DE.

5.1.2. Differential Evolution Algorithm

As it is shown in Table 2 in the Integrase-DE, ANN produces better results in 5 out of 6 areas. The only area where ANN does not produce the best result is the root-mean-square error. However, the result of RMSE is quite close to MLR and SVM.

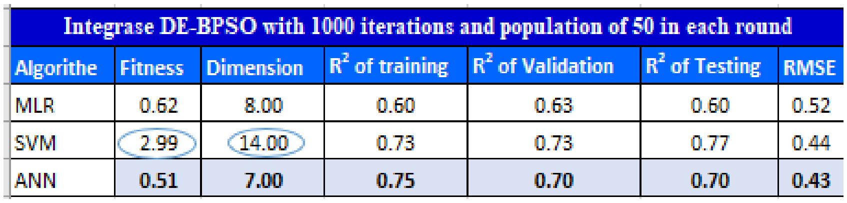


Integrase BPSO with 1000 iterations and population of 50 in each round						
Algorithe	Fitness	Dimension	R^2 of training	R^2 of Validation	R^2 of Testing	RMSE
MLR	0.63	10.00	0.71	0.78	0.54	0.56
SVM	0.60	9.00	0.70	0.71	0.53	0.57
ANN	0.66	**7.00**	**0.80**	**0.15**	**0.68**	**0.45**

TABLE 3: Result of Integrase BPSO.

5.1.3. Binary Particle Swarm Optimization Algorithm

Table 3 illustrates the result of Integrase-BPSO. Again, ANN performs better in four out of 6 areas compared to SVM and MLR.



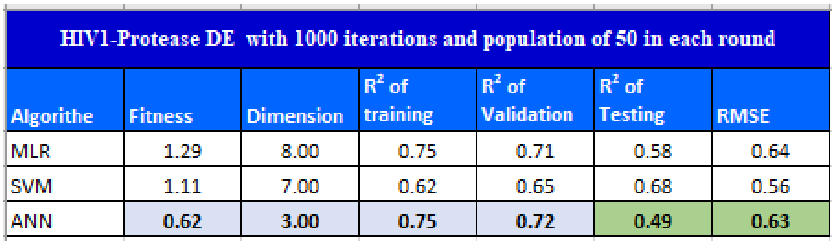
Integrase DE-BPSO with 1000 iterations and population of 50 in each round						
Algorithe	Fitness	Dimension	R^2 of training	R^2 of Validation	R^2 of Testing	RMSE
MLR	0.62	8.00	0.60	0.63	0.60	0.52
SVM	2.99	14.00	0.73	0.73	0.77	0.44
ANN	**0.51**	**7.00**	**0.75**	**0.70**	**0.70**	**0.43**

TABLE 4: Result of Integrase DE-BPSO.

5.1.4. Differential evolution- Binary Particle Swarm Optimization Algorithm

Table 4 displays the result of Integrase with DE-BPSO. As shown, we don’t need to compare SVM with the ANN and MLR because it has comparatively high fitness and very high dimensionality, which indicates that the SVM model is very complex and not robust. Therefore, comparing ANN with MLR, ANN does better in all areas than MLR

5.2. HIV-1 Protease

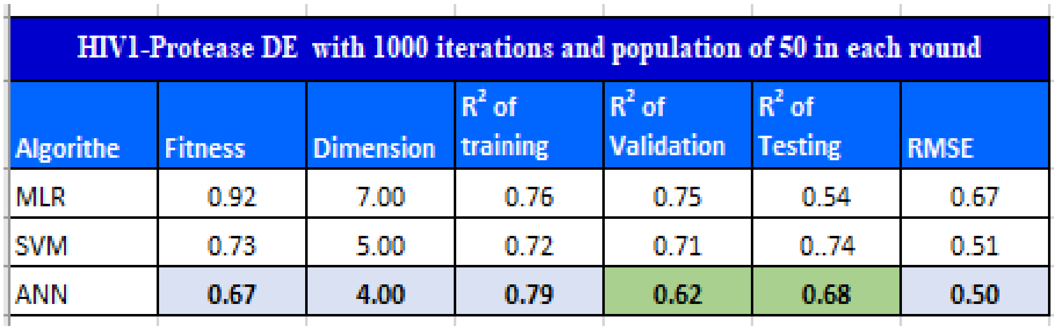


HIV1-Protease DE with 1000 iterations and population of 50 in each round						
Algorithe	Fitness	Dimension	R^2 of training	R^2 of Validation	R^2 of Testing	RMSE
MLR	1.29	8.00	0.75	0.71	0.58	0.64
SVM	1.11	7.00	0.62	0.65	0.68	0.56
ANN	**0.62**	**3.00**	**0.75**	**0.72**	**0.49**	**0.63**

TABLE 5: Result of HIV-1 Protease GA-BPSO.

5.2.1. Genetic Algorithm

Table 5 shows the result of HIV-1 Protease with GA. As shown, at least after 1000 iterations, ANN demonstrates better results in 4 out of 6 areas.

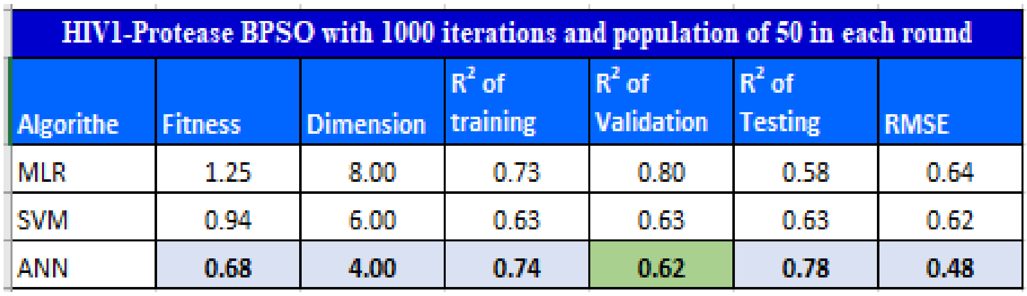


HIV1-Protease DE with 1000 iterations and population of 50 in each round						
Algorithe	Fitness	Dimension	R^2 of training	R^2 of Validation	R^2 of Testing	RMSE
MLR	0.92	7.00	0.76	0.75	0.54	0.67
SVM	0.73	5.00	0.72	0.71	0..74	0.51
ANN	**0.67**	**4.00**	**0.79**	**0.62**	**0.68**	**0.50**

TABLE 6: Result of HIV-1 Protease DE.

5.2.2. Differential Evolution Algorithm

Similarly, table 6, demonstrates that the result of HIV-1 Protease with DE produces better results in 4 out of 6 areas compare to MLR and SVM.

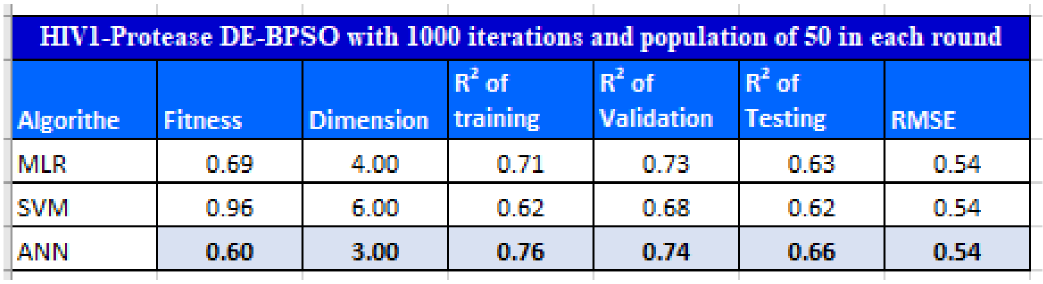


HIV1-Protease BPSO with 1000 iterations and population of 50 in each round						
Algorithe	Fitness	Dimension	R^2 of training	R^2 of Validation	R^2 of Testing	RMSE
MLR	1.25	8.00	0.73	0.80	0.58	0.64
SVM	0.94	6.00	0.63	0.63	0.63	0.62
ANN	**0.68**	**4.00**	**0.74**	**0.62**	**0.78**	**0.48**

TABLE 7: Result of HIV-1 Protease BPSO.

5.2.3. Binary Particle Swarm Optimization Algorithm

In the HIV-1 Protease BPSO (Table 7), again the result for the ANN is better in 5 out of 6 areas compare to SVM and MLR. shows the best results. Note that even though the result of ANN for R^2 of validation does not have the best result, it is still relatively close to SVM and MLR.

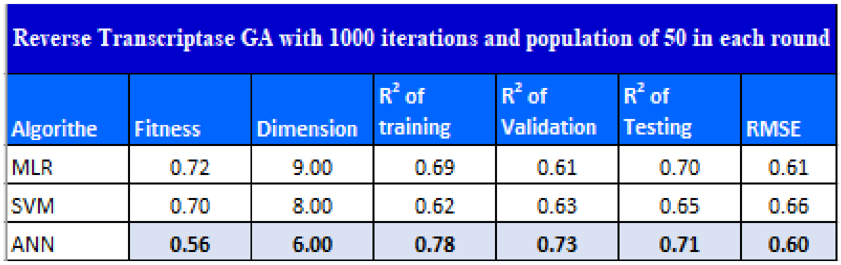


HIV1-Protease DE-BPSO with 1000 iterations and population of 50 in each round						
Algorithe	Fitness	Dimension	R^2 of training	R^2 of Validation	R^2 of Testing	RMSE
MLR	0.69	4.00	0.71	0.73	0.63	0.54
SVM	0.96	6.00	0.62	0.68	0.62	0.54
ANN	**0.60**	**3.00**	**0.76**	**0.74**	**0.66**	**0.54**

TABLE 8: Result of HIV-1 Protease DE-BPSO.

5.2.4. Differential Evolution- Binary Particle Swarm Optimization Algorithm

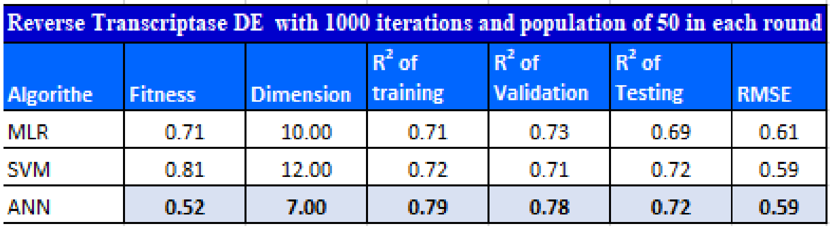
In the HIV-1 Protease DE-BPSO (Table 8), the ANN model demonstrates the best results in all areas. Especially notable is its dimensionality of 3, which means that the model is very simple and robust.



Reverse Transcriptase GA with 1000 iterations and population of 50 in each round

Algorithe	Fitness	Dimension	R^2 of training	R^2 of Validation	R^2 of Testing	RMSE
MLR	0.72	9.00	0.69	0.61	0.70	0.61
SVM	0.70	8.00	0.62	0.63	0.65	0.66
ANN	**0.56**	**6.00**	**0.78**	**0.73**	**0.71**	**0.60**

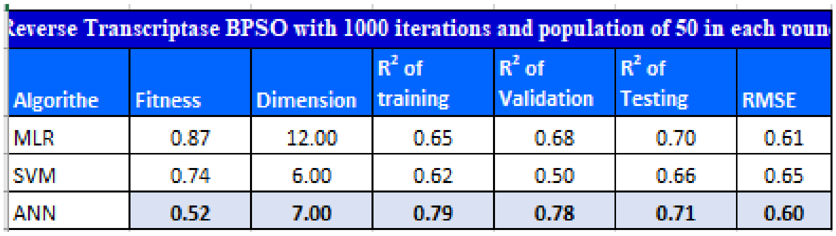
TABLE 9: Result of Reverse Transcriptase GA.



Reverse Transcriptase DE with 1000 iterations and population of 50 in each round

Algorithe	Fitness	Dimension	R^2 of training	R^2 of Validation	R^2 of Testing	RMSE
MLR	0.71	10.00	0.71	0.73	0.69	0.61
SVM	0.81	12.00	0.72	0.71	0.72	0.59
ANN	**0.52**	**7.00**	**0.79**	**0.78**	**0.72**	**0.59**

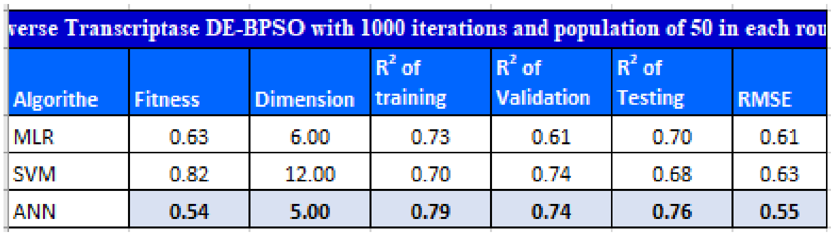
TABLE 10: Result of Reverse Transcriptase DE.



Reverse Transcriptase BPSO with 1000 iterations and population of 50 in each roun

Algorithe	Fitness	Dimension	R^2 of training	R^2 of Validation	R^2 of Testing	RMSE
MLR	0.87	12.00	0.65	0.68	0.70	0.61
SVM	0.74	6.00	0.62	0.50	0.66	0.65
ANN	**0.52**	**7.00**	**0.79**	**0.78**	**0.71**	**0.60**

TABLE 11: Result of Reverse Transcriptase BPSO.



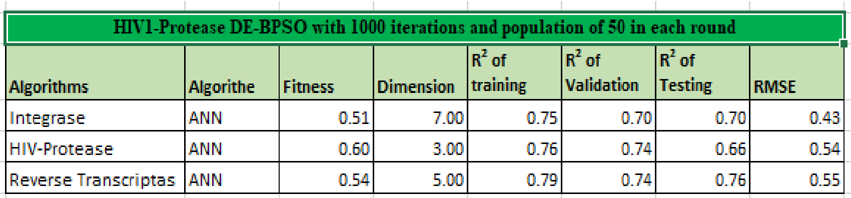
verse Transcriptase DE-BPSO with 1000 iterations and population of 50 in each rou

Algorithe	Fitness	Dimension	R^2 of training	R^2 of Validation	R^2 of Testing	RMSE
MLR	0.63	6.00	0.73	0.61	0.70	0.61
SVM	0.82	12.00	0.70	0.74	0.68	0.63
ANN	**0.54**	**5.00**	**0.79**	**0.74**	**0.76**	**0.55**

TABLE 12: Result of Reverse Transcriptase DE-BPSO.

5.3. Reverse Transcriptase

Tables 9-12 show that, in the Reverse Transcriptase, ANN consistently shows the best result in all 6 areas for all Genetic Evaluation Optimization algorithms.



HIV1-Protease DE-BPSO with 1000 iterations and population of 50 in each round

Algorithms	Algorithe	Fitness	Dimension	R^2 of training	R^2 of Validation	R^2 of Testing	RMSE
Integrase	ANN	0.51	7.00	0.75	0.70	0.70	0.43
HIV-Protease	ANN	0.60	3.00	0.76	0.74	0.66	0.54
Reverse Transcriptas	ANN	0.54	5.00	0.79	0.74	0.76	0.55

TABLE 13: Integrase all the results.

5.4. Summarizing the Results

Summarizing the above results, we realize that all three enzymes produce better results with DE-BPSO and ANN as is displayed in table 13.

Going through the complexity of each descriptor and their biological activities is beyond the scope of this paper; however, we list the meaning of the descriptors in case some researchers plan to take this research to the next steps.

The 7 descriptors for Integrase are:

[169 183 184 237 309 319 435]

- **169**: VE1_B(s): The coefficient sum of the last eigenvector (absolute values) from Burden matrix weighted by I-State.
- **183**: ATSC7m: Centred Broto-Moreau autocorrelation of lag 7 weighted by mass.
- **184**: ATSC8m: Centred Broto-Moreau autocorrelation of lag 8 weighted by mass.
- **237**: MATS5s: Moran autocorrelation of lag 5 weighted by I-state.
- **309**: SpMax5_Bh(v): Largest eigenvalue n. 5 of Burden matrix weighted by van der Waals volume.
- **319**:SpMax7_Bh(s): Largest eigenvalue n. 7 of Burden matrix weighted by I-state.
- **435**:SpMin4_Bh(s): Smallest eigenvalue n. 4 of Burden matrix weighted by I-state

The 3 descriptors of HIV-Protease are:

[46 107 407]

- **46: ZM2:** Second Zagreb index
- **107: SpPosA_L:** Normalized spectral positive sum from Laplace matrix
- **407: Chi0_EA:** connectivity-like index of order 0 from edge adjacency mat.

The 5 descriptors of reverse Transcriptome are:

[17 90 215 365 849]

- **17: NH**: number of Hydrogen atoms
- **90: MPC07**:molecular path count of order 7
- **215:ATSC2e:** Centred Broto-Moreau autocorrelation of lag 2 weighted by Sanderson electronegativity.
- **365:P_VSA_i_3**: P_VSA-like on ionization potential, bin 3
- **849:F03[C-N]:** Frequency of C – N at topological distance 3

6. CONCLUSION AND FUTURE WORK

In this study, we used QSAR techniques and machine learning to build predictive models for inhibitors' efficacy against three HIV-1 enzymes. Our setup combined evolutionary algorithms such as GA, DE, BPSO, and DEBPSO with machine learning algorithms such as MLR, SVM, and MLP (ANN) to produce and test predictive models in an iterative process.

The results show that in all three enzymes we had the best results were DE-BPSO with the ANN. Even though similar work on these three enzymes was done in the past decade, the amount of data has been relatively very low compare to what we used in this work and none of that research utilized as many evolutionary algorithms to optimize their model-building as we did. We have at least four times as many rows (compounds) and columns (descriptors) to provide more accurate results. Furthermore, the previous works were being executed against relatively fast computers for at least two million iterations to converge and obtain models with acceptable results, whereas our results were obtained within a few thousand iterations, requiring several hours rather than taking days to complete. We could use simple computers such as our laptops to run the machine-learning program instead of needing to utilize supercomputers with powerful CPUs.

This study's contribution is thus twofold: it offers three promising predictive models for HIV-1 drugs and a lightweight, efficient machine learning setup to create new predictive models.

Future work can be done in several directions. One is to get data for other diseases such as Parkinson's, breast cancer, thyroid, etc can be tested with our program. Another direction is to optimize the Genetic Evolutionary algorithms further and test the results with other diseases.

7. REFERENCES

[1] Youcef Mehellou, and Erik De Clercq, "Twenty-Six Years of Anti-HIV Drug Discovery: Where Do We Stand and Where Do We Go?", vol. 53(2), pp: 521–538, 2010.

[2] Jinfang Zhu. "T Helper Cell Differentiation Heterogeneity, and Plasticity", vol. 10(10), 2018

[3] E. Dukhanina, T. Lukyanova, A. Dukhanin, S. Georgieva, "The role of S100A4 protein in anticancer cytotoxicity: its presence is required on the surface of CD 4+ CD 25+ PGRPs + S100A4 + lymphocyte and undesirable on the surface of target cells" . vol. 17(4), pp: 1-20, 2017.

[4] Joint United Nations Program on HIV/AIDS(UNAIDS), UNAIDS Report on the Global AIDS Epidemic 2020, Geneva, 2020.

[5] Annemarie M Wensing, Vincent Calvez, Francesca Ceccherini-Silberstein, Charlotte Charpentier, Huldrych F Günthard, Roger Paredes, Robert W Shafer and Douglas D Richman, "2019 update of the drug resistance mutations in HIV-1", vol. 27(3) pp:111-121. 2019.

[6] M.J. Pucci, Christian Callebaut, Andrea Cathcart and Karen Bush, "5.17 - Recent Epidemiological Changes in Infectious Diseases", in book "Comprehensive Medicinal Chemistry III", pp: 511-552, 2017.

[7] Biswa Ranjan Meher, Megha Vaishnavi, Venkata Satish Kumar Mattaparthi and Seema Patel, "Mutation Effects on 3D-Structural Reorganization Using HIV-1 Protease as a Case Study", In book: Encyclopedia of Bioinformatics and Computational Biology, 2018.

[8] Wei-Shau Hu and Stephen H. Hughes, "HIV-1 reverse transcription", Cold Spring Harb Perspect Med, vol. 2(10), 2012.

[9] Michael K Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang and Jenny Chong, "BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology", Nucleic Acids Res, vol. 44(D1), pp: D1045-53, 2016.

[10] Gene M. Ko, Srinivas Reddy, Sunil Kumar, Rajni Garg, Barbara A. Bailey and Ahmad R. Hadaegh, "Differential evolution-binary particle swarm optimization algorithm for the analysis of aryl β-diketo acids for HIV-1 integrase inhibition", 2012 IEEE Congress on Evolutionary Computation, Brisbane, pp. 1-7, 2012.

[11] Falguni. Thakor, Ahmad. Hadaegh and Xiaoyu. Zhang. "Comparative study of Differential Evolutionary-Binary Particle Swarm Optimization (DE-BPSO) algorithm as a feature selection technique with different linear regression models for the analysis of HIV-1 Integrase Inhibition features of Aryl β-Diketo Acids". Proceedings of 9th International Conference on Bioinformatics and Computational Biology (BICOB 2017) Honolulu, Hawaii, USA, pp: 179-184, 2017.

[12] Ian Kane, and Ahmad. Hadaegh. “Non-linear Quantitative Structure-Activity Relationship (QSAR) Models for the Prediction of HIV Drug Performance”. 24th International Conference on Software Engineering and Data Engineering (SEDE-2015), vol 1, pp: 63-68, 2015.

[13] Matineh. Kashani, Richard. Galvan, and Ahmad. Hadaegh, “Improving the Feature Selection for the Development of Linear Model for Discovery of HIV-1 Integrase Inhibitors”. ABDA'15 International Conference on Advances in Big Data Analytics. In Proceeding of the 2015 International Conferences on Advances on Big Data Analyses, pp: 150-154. Las Vegas, Nevada, 2015.

[14] Andrea Mauri, Viviana Consonni, Manuela Pavan, Roberto Todeschini, "DRAGON software: An easy approach to molecular descriptor calculations", vol. 56(2), pp:237-248, 2006.

[15] Jiali. Tang, Jack. Wang, and Ahmad .R. Hadaegh, “A Web Repository System for Data Mining in Drug Discovery”, 2020 International Journal of Data Mining & Knowledge Management Process (IJDKP) vol. 10, No. 1, 2020.

[16] Richard Galvan, Matineh Kashani, and Ahmad Hadaegh. “Improving Pharmacological Research of HIV-1 Integrase Inhibition Using Differential Evolution-Binary Particle Swarm Optimization and Non-Linear Adaptive Boosting Random Forest Regression”, 2015 IEEE International Workshop on Data Integration and Mining San Francisco, pp: 485-490. San Francisco, CA. 2015.

[17] Julia. Mikulski, “The Ultimate Guide to AdaBoost, random forests and XGBoost. Towards Data Science”, 2020.

[18] Stanford Encyclopedia of Philosophy. Darwin: From Origin of Species to Descent of Man. First published Mon Jun 17, 2019.

[19] David Freedman, in book “Statistical Models: Theory and Practice”. Cambridge University Press. p. 26. Cambridge University Press, 2009.