Prediction of HIV-1 protease resistance using genotypic, phenotypic, and molecular information with artificial neural networks

Huseyin Tunc [1] , **Berna Dogan** [2] , **Büşra Nur Darendeli Kiraz** [3, 4] , **Murat Sari** [5] , **Serdar Durdagi** [6, 7] , **Seyfullah Kotil** [Corresp. 3]

[1] Department of Biostatistics and Medical Informatics, School of Medicine, Bahcesehir University, Istanbul, Turkey

[2] Department of Medicinal Biochemistry, School of Medicine, Bahcesehir University, Istanbul, Turkey

[3] Department of Biophysics, School of Medicine, Bahcesehir University, Istanbul, Turkey

[4] Department of Bioengineering, Yildiz Technical University, Istanbul, Turkey

[5] Department of Mathematics Engineering, Faculty of Science and Letters, Istanbul Technical University, Istanbul, Turkey

[6] Computational Biology and Molecular Simulations Laboratory, Department of Biophysics, School of Medicine, Bahcesehir University, Istanbul, Turkey

[7] Department of Pharmaceutical Chemistry, School of Pharmacy, Bahcesehir University, Istanbul, Turkey

Corresponding Author: Seyfullah Kotil
Email address: enesseyfullah.kotil@med.bau.edu.tr

Drug resistance is a primary barrier to effective treatments of HIV/AIDS. Calculating quantitative relations between genotype and phenotype observations for each inhibitor with cell-based assays requires time and money-consuming experiments. Machine learning models are good options for tackling these problems by generalizing the available data with suitable linear or nonlinear mappings. The main aim of this paper is to construct drug isolate fold (DIF) change-based artificial neural network (ANN) models for estimating the resistance potential of molecules inhibiting the HIV-1 protease (PR) enzyme. Throughout the study, seven of eight protease inhibitors (PIs) have been included in the training set and the remaining ones in the test set. We have obtained 11803 genotype-phenotype data points for eight PIs from Stanford HIV drug resistance database. Using the leave-one-out (LVO) procedure, eight ANN models have been produced to measure the learning capacity of models from the descriptors of the inhibitors. Mean R^2 value of eight ANN models for unseen inhibitors is 0.732, and the 95% confidence interval (CI) is [0.613,0.850]. Predicting the fold change resistance for hundreds of isolates allowed a robust comparison of drug pairs. These eight models have predicted the drug resistance tendencies of each inhibitor pair with the mean 2D correlation coefficient of 0.933 and 95% CI [0.930,0.938]. A classification problem has been created to predict the ordered relationship of the PIs, and the mean accuracy, sensitivity, specificity, and Matthews correlation coefficient (MCC) values are calculated as 0.954, 0.791, 0.791, and 0.688, respectively. Furthermore, we have created an external test dataset consisting of 51 unique known HIV-1 PR inhibitors and 87 genotype-phenotype relations. Our developed ANN model has accuracy and area under the curve (AUC) values of 0.749 and 0.818 to predict the ordered relationships of molecules on the same strain for the external dataset. The currently derived ANN models can accurately predict the drug resistance tendencies of PI pairs. This observation could help test new inhibitors with various isolates.

Prediction of HIV-1 protease resistance using genotypic, phenotypic, and molecular information with artificial neural networks

Huseyin Tunc[1], Berna Dogan[2], Büşra Nur Darendeli Kiraz[3,4], Murat Sari[5] , Serdar Durdagi[6,7], Seyfullah Enes Kotil[3,*]

[1]Department of Biostatistics and Medical Informatics, School of Medicine, Bahcesehir University, Istanbul, Turkey

[2]Department of Medicinal Biochemistry, Bahcesehir University Medical School, Istanbul, Turkey

[3]Department of Biophysics, School of Medicine, Bahcesehir University, Istanbul, Turkey

[4]Department of Bioengineering, Yildiz Technical University, Istanbul, Turkey

[5]Department of Mathematics Engineering, Faculty of Science and Letters, Istanbul Technical University, Istanbul, Turkey

[6]Computational Biology and Molecular Simulations Laboratory, Department of Biophysics, School of Medicine, Bahcesehir University, Istanbul, Turkey

[7]Department of Pharmaceutical Chemistry, School of Pharmacy, Bahcesehir University, Istanbul, Turkey

*Corresponding author

enesseyfullah.kotil@med.bau.edu.tr

Abstract

Drug resistance is a primary barrier to effective treatments of HIV/AIDS. Calculating quantitative relations between genotype and phenotype observations for each inhibitor with cell-based assays requires time and money-consuming experiments. Machine learning models are good options for tackling these problems by generalizing the available data with suitable linear or nonlinear mappings. The main aim of this paper is to construct drug isolate fold (DIF) change-based artificial neural network (ANN) models for estimating the resistance potential of molecules inhibiting the HIV-1 protease (PR) enzyme. Throughout the study, seven of eight protease inhibitors (PIs) have been included in the training set and the remaining ones in the test set. We have obtained 11803 genotype-phenotype data points for eight PIs from Stanford HIV drug resistance database. Using the leave-one-out (LVO) procedure, eight ANN models have been produced to measure the learning capacity of models from the descriptors of the inhibitors. Mean R^2 value of eight ANN models for unseen inhibitors is 0.732, and the 95% confidence interval (CI) is [0.613,0.850]. Predicting the fold change resistance for hundreds of isolates allowed a robust comparison of drug pairs. These eight models have predicted the drug resistance tendencies of each inhibitor pair with the mean 2D correlation coefficient of 0.933 and 95% CI [0.930,0.938]. A classification problem has been created to predict the ordered relationship of the PIs, and the mean accuracy, sensitivity, specificity, and Matthews correlation coefficient (MCC) values are calculated as 0.954, 0.791, 0.791, and 0.688, respectively. Furthermore, we have created an external test dataset consisting of 51 unique known HIV-1 PR inhibitors and 87 genotype-phenotype relations. Our developed ANN model has accuracy and area under the curve (AUC) values of 0.749 and 0.818 to predict the ordered relationships of molecules on the same strain for the external dataset. The currently derived ANN models can accurately predict the drug resistance tendencies of PI pairs. This observation could help test new inhibitors with various isolates.

Keywords: Machine learning; Artificial neural networks; HIV/AIDS; Drug resistance; Protease inhibitors

Introduction

Acquired immunodeficiency syndrome (AIDS) disease caused by the human immunodeficiency viruses, HIV-1 and HIV-2, began to spread in the 1970s and came into focus in the early 1980s as one of the most severe public health threats in history [1]. Detection of reverse transcription activity in cultures of lymph node cells from AIDS patients in the early 1980s revealed that AIDS was caused by a retrovirus later called human immunodeficiency virus (HIV) [2]. Zidovudine (AZT), the first nucleotide reverse transcriptase inhibitor (NRTI) that inhibits the reverse transcription enzyme of HIV, was approved in 1987, and today there are nearly thirty approved drugs [3]. HIV-1 has affected approximately 38 million people today, and just about 26 million people are receiving "Highly Active Antiretroviral Treatment" (HAART) [4]. The HAART therapy proposed in the mid-1990s was defined as the procedure of using three or four different drugs that act on various targets in the virus's life cycle [5]. With HAART therapy, the death rate fell to 47% in 1997, just ten years after the first AIDS case was detected [6].

Drug resistance is the primary barrier to the effective treatment of HIV/AIDS [7,8]. Single-drug treatments for HIV yield rapid resistance due to the high genetic diversity and error-prone replication of the virus [8,9]. Hence, the use of drug combinations through the HAART protocols increases the efficacy of the treatment [10]. However, cross-resistant isolates for available drugs encourage researchers to find novel inhibitors [11-15]. To combat drug-resistant isolates, novel drug design methodologies have been adopted for HIV-1 protease enzyme such as phosphonate-mediated solvent anchoring [11], lysine sulfonamide-based molecular core [12], allophenylnorstatine containing inhibitors [13], nonpeptic inhibitor GRL-02031 [14], bis-tetrahydrofuranylurethane containing nonpeptidic inhibitor UIC-94017 [15]. Testing novel inhibitors with various drug-resistant isolates need experimental or computational mechanisms.

HIV protease enzyme plays a vital role in forming infectious viruses by regulating immature viruses' synthesized gag and gag-pol polyproteins [16]. Protease inhibitors are generally included in the scope of HAART therapy, and eight approved drug molecules are used effectively today [17]. Dose-response curves of protease inhibitors were shown that they have higher Hill coefficient values than the fusion (FI), integrase (II), nucleoside reverse transcriptase (NRTI), and non-nucleotide reverse transcriptase (NNRTI) inhibitors [18]. Even if a person is infected with the wild-type virion, resistant variants may emerge with dosing disruptions or the use of inappropriate combinations in the HAART therapy [19]. The success rate of HAART therapy can be increased by measuring the efficacy of existing and novel inhibitors over resistant genotypes [20-21]. Observing drug-efficacy relations with cell-based assays is expensive and time-consuming in the presence of genotype information. Mathematical models are essential to tackle this important problem [22-24].

Various mathematical models have been calibrated using genotype-phenotype change data proposed in the Stanford HIV database to predict mutational effects on viral dynamics in the literature [25-40]. The life span of patients can be considerably extended by constructing reliable mathematical models that accurately predict suitable drugs for existing isolates. Most existing prediction models are knowledge-based and require predetermined rules on mutations and drugs [25-28]. The most commonly used genotype interpretation algorithms have been observed to be Stanford HIVdb [25], HIV-grade [26], REGA [27], and ANRS [28]. In addition to these genotype interpretation algorithms, various machine learning models have recently been proposed to predict genotype-phenotype change relationships in the presence of a predetermined inhibitor [29-40]. Artificial neural network [29-34], random forest algorithm [35-41], support vector machine [37,41-42], decision trees [43], k-nearest neighbors (kNN) [36], restricted Boltzmann machine [44], support vector regression [40] and linear regression [45] are the methods used in the literature to model the efficacy of different drugs against HIV-1 variants. All the works mentioned above focus on predicting the fold change of mutant fitness under a single drug. Fold change values for each molecule are treated as disjoint and used to construct a drug-specific model. Those type of models does not need to take molecular descriptors of a drug as input, hence are indifferent to chemical structure. Such models cannot predict the effects of resistance mutations for a novel drug. Therefore, a model that predicts fold change of mutant fitness for multiple molecules is needed.

Here, a machine learning model was constructed that simultaneously takes molecular fingerprints and mutational information jointly as inputs to estimate the fold change values. For training and testing sets, we used data from eight approved protease inhibitors atazanavir (AZT), darunavir (DRV), fosamprenavir (FPV), indinavir (IDV), lopinavir (LPV), nelfinavir (NFV), saquinavir (SAV) and tipranavir (TPV) in the Stanford HIV drug resistance database. By imposing leave one out (LVO) test procedure, our drug-isolate-fold (DIF) change-based artificial neural network (ANN) models are seen to have the ability to learn both from inhibitor descriptors and mutational genotype information to predict fold-change values. The model can predict the fold change of hundreds of isolates. To that end, the learned hundreds of predictors (fold-change of isolates) can be successfully used to assess the resistance potential of inhibitors. We used pairs of drugs to predict the more resistance-prone molecule. We called these pairwise comparisons the resistance tendencies. Our DIF-based ANN models predicted each protease inhibitor (PI) pair's drug resistance tendencies accurately, and these quantitative results support our central arguments.

Methods and Material

Dataset Description

Filtered genotype-phenotype data on the Stanford HIV drug resistance database was retrieved for PIs [2]. We have organized this dataset with respect to isolates and inhibitors, and 498 protease mutations have been observed. For the HIV-1 PI: 1218 isolates for atazanavir (ATV), 678 isolates for darunavir (DRV), 1809 isolates for fosemprenavir (FPV), 1860 isolates for indinavir (IDV), 1562 isolates for lopinavir (LPV), 1907 isolates for nelfinavir (NFV), 1861 isolates for saquinavir (SQV) and 908 isolates for tipranavir (TPV) have been analyzed for PI susceptibility. In the dataset, 436, 336, 480, 483, 472, 486, 489 and 409 different mutations have been observed for ATV, DRV, FPV, IDV, LPV, NFV, SQV, and TPV, respectively.

Representation of Isolates

Four hundred ninety-eight unique mutations were observed in the eight protease inhibitors dataset. The binary barcoding technique was applied here to represent the isolates that occurred in the dataset, as also used in several studies of modeling genotype-phenotype data for various HIV-1 inhibitors [3]. Thus, a 498-dimensional vector of binary entries with 0s and 1s uniquely representing any existing isolates is considered. Assume that the 498 unique mutations produce the vector $X = [x_1, x_2, ..., x_{498}]$ where x_i is a mutation pattern that occurred in the dataset. For instance, x_1 denotes the occurrence of the mutation A22S or x_{478} denotes the occurrence of the mutation V82A. For example, the isolate $I_j = [\text{A22S}, \text{V82A}]$ can be barcoded as $X = [1,0,...,0,1,0,...,0]$ in which only the first and four hundred seventy-eighth position take value one, and the remaining entries have value zero. Any isolate can be obtained from any combination of these mutations, and the isolate j can be defined as $I_j = \{a_1, a_2, ..., a_n\}$ with

$$a_k = \begin{cases} 1, & if\ x_k \in I_j \\ 0, & otherwise. \end{cases}$$

In this way, each isolate can be transformed into a unique 498-dimensional input vector used for the machine learning. The binary barcoding approach has the advantage of representing two or more mutational changes in the amino acids since each mutation has a unique position in the 498-dimensional input vector. For example, assuming that x_{480} and x_{481} denotes mutations V82F and V82I, the isolate $I = [\text{V82F}, \text{V82I}]$ can be represented as $X = [0,0,...,0,1,1,0,...,0]$. It is important to note that, the genotype-fold change measurements are made on population of viruses, so that it is possible to find two separate mutations for a single residue.

Representation of Inhibitors

To construct a drug-isolate-fold change model for the HIV-1 protease inhibitors, the molecular representations of the inhibitors have been built with binary Morgan fingerprints. The Morgan fingerprints provide an effective way of the vector representations of molecules and are widely used in machine learning models [4]. The RDKit environment of the Python program has been used to convert the smile representations of ATV, DRV, FPV, IDV, LPV, NFV, SQV, and TPV inhibitors to a binary 512-bit vector representation. 234 out of 512 bits have been seen to provide unique characteristics for 8 PI. Thus, the molecular representation of each PI needs 234-dimensional vectors.

Artificial Neural Network (ANN) Model for Regression

An ANN model has been constructed with isolate-inhibitor inputs and fold change outputs with the Machine Learning and Deep Learning toolbox of the MATLAB program. Since isolates and inhibitors are uniquely represented by 498- and 234- dimensional vectors, the ANN model has 732-dimensional input. The ANN architecture includes 732-dimensional input, five hidden layer neurons, and one output neuron with hyperbolic tangent-sigmoid and linear activation functions. Logarithms of fold-change values in the dataset are taken as output variables of the neural network models. In the training process, the scaled conjugate gradient algorithm with MATLAB built-in function “trainscg” is utilized over GPU [5].

Ensemble Processing

Since we have only eight inhibitors, measuring the molecular learning capacity of our ANN model is crucial. In this way, an ensemble learning procedure is used to improve the molecular learning performance of the model. For each PI, the 100×50 model has been trained with the data of the remaining seven inhibitors. From every 50 models, a model is chosen that yields the minimum mean square error for the interior test set of the corresponding PI data. Thus, 100 optimal models are obtained, and the final model is calculated as the average of these models.

RESULTS

Regression performance of molecular learning models

Eight feed-forward neural network models have been constructed with drug-isolate-fold- change (DIF) data by excluding one of the drugs from training in each case. The ANN model was trained with the remaining seven DIF data predicted the excluded results. The sizes of the training and test sets were changed according to the excluded PIs (mean values are 10328 and 1475 for training and test sets, respectively). The regression performances of each model are illustrated in Figure 1 with corresponding R^2 values (square of the linear correlation coefficient). The best and worst results are obtained by predicting the outcomes of the drugs LPV and TPV with $R^2 = 0.837$ and $R^2 = 0.393$, respectively. Similarly, predicting the fold-change results of the inhibitor TPV was observed to be the worst in the literature [23]. The mean R^2 value of all predictions is 0.732 and the 95% confidence interval is $[0.613, 0.850]$. The DIF-based ANN model provides accurate estimations even if the test data consists of unseen drugs. This observation implies that our ANN models accurately learn molecular information from the Morgan fingerprints. The detailed performance results of our DIF-based ANN models are presented in Table 1.

An inevitable question is how molecular information changes the regression and classification performance of our ANN models. To clarify this, we trained isolate-fold-change (IF) based ANN models for each inhibitor and compared them with the current DIF-based models. In Figures S1-S2, the model performances have been compared by measuring R^2 and area under the curve (AUC) values for each PI. Table S1 also shows the accuracy, sensitivity, specificity, and Matthews correlation coefficient (MCC) scores of both DIF-based and IF-based models. These findings suggest that the DIF-based ANN model performs slightly better in terms of regression and classification for six of the eight PIs. The two models are compatible with the remaining two inhibitors (DRV and TPV). Therefore, the molecular information used by the DIF-based model provides a better predictive capability. It is very important to note that IF-based models can never predict fold-change values for novel molecules. The real distinction between DIF and IF-based models is that the DIF-based model can predict fold-change values for a novel drug. Hence, the DIF-based ANN model has both better learning capability from mutant information (by comparing to the IF-based model) and the ability to test novel molecules for a given isolate.

Prediction of drug resistance tendencies for each PI pair

The inhibition potential of each PI in the presence of various genotypes is known to be variable. Tendencies of the logarithmic fold change values for each PI pair provide valuable information about the resistance profiles of the inhibitors, as seen in Figure 2. Prediction of these tendencies by the DIF-based ANN models and the corresponding 2D correlation coefficients are presented in Figure 2 in a comparative way for each PI pair. For each PI, prediction has been made with the ANN model trained by the data of the remaining seven inhibitors with an ensemble learning approach. This procedure shows the molecular learning capacity of our ANN models from the Morgan fingerprints. The minimum and maximum 2D correlation coefficients are 0.892 and 0.954 for TPV-DRV and LPV-DRV couples (95% CI [0.930, 0.938]), respectively. Thus, the current DIF-based ANN models can distinguish the inhibitory potentials of each PI pair.

Classification of PIs with respect to possible common isolates

Our DIF-based ANN models can distinguish the fold change values of each PI in the presence of any isolate. In this way, a classification problem measuring the relationship $log(Fold\ Change\ [A,Isolate]) > log(Fold\ Change\ [B,Isolate])$ has been constructed, where A and B are possible protease inhibitors. These relations take values 0 and 1 depending on the inhibitors and isolates. Therefore, our ANN models have been trained with the data from seven inhibitors, with the exception of one particular inhibitor considered as test data. The corresponding receiver operating characteristic (ROC) curves are illustrated in Figure 3. Area under the ROC curve (AUC) values are included in the figure. The best and worst AUC values have been obtained for the IDV-LPV and DRV-LPV pairs with 0.992 and 0.818 (95% CI: [0.950, 0.978]), respectively. In this context, the current DIF-based ANN models have ability to capture the binary relations between any PI pair with high approximation performance.

Performance metrics of the current ANN models for capturing binary relations of PI pairs are presented in Table 2. As indicated in the table, the DIF-based ANN models have a high rate of true prediction for each PI pair. The mean accuracy, sensitivity, specificity and Matthews correlation coefficient (MCC) values have been computed as 0.954, 0.791, 0.791 and 0.688 (95% CI [0.932, 0.952], [0.719, 0.863], [0.719, 0.863] and [0.600, 0.776]), respectively. The most conspicuous result here is that the neural network models can classify the inhibitors for resistance profiles, even if that model did not see the corresponding inhibitors in the training process.

Testing the molecular learning model with the external data set

For the external dataset, we conducted a search for compounds that were comparable to the eight HIV PIs that were already available in the ChEMBL database [46]. An initial set of 1305 compounds and their biological activity values were extracted. First, compounds having 70% or less similarity to each drug were filtered out. Then, molecules with determined IC_{50} values in mutant viruses were collected. The compounds with determined IC_{50} values in mutant viruses were then gathered. Furthermore, the maximum Tanimoto similarities of these molecules to the current eight PIs were computed (using 512-bit Morgan fingerprints), and molecules with 40% or less structural similarity were filtered out. Finally, molecules with 1s in the discarded 278-dimensional fingerprint vectors, that is, molecules with information loss, were filtered out. A final dataset of 87 genotype-phenotype relationships involving 51 different molecules was obtained (see Supplemental Data S1). Only six of these molecules are in the existing PI set (DRV, IDV, NFV, ATV, LPV) and only 20 of 87 genotype-phenotype relations belong to these six PIs. In the presence of eight distinct strains, there is a fold change difference in IC_{50} values for these molecules. Fortunately, the data includes different molecules tested on the same protein, so we can evaluate the efficiency of our ANN model on ranking of these molecules.

We have used the eight existing PIs (see Figure 4 for chemical structures) to construct an ANN model as described in the *Methods and Material* section. The main difference is the consideration of all molecules rather than the LVO procedure. The Morgan fingerprints of the new 51 molecules were determined, and then the similar mapping was used to reduce the 512-dimensional vectors into 234-dimensional inputs. It should be noted that the molecules were chosen in such a way that there are no 1s in the discarded 278 bits. This condition was specifically designed into the external dataset. In this approach, we ensure that no meaningful information for novel molecules is lost during the machine learning process.

Two different classification performances of our ANN model were observed to test its molecular learning potential. The first is the classification of resistant and susceptible strains for a given PI, with a threshold fold chance value of 3 [36]. The goal of this classification task is to evaluate the resistance labeling performance of the current model for diverse molecules. As demonstrated in Table 3, the accuracy, sensitivity, specificity, MCC, and AUC metrics of our model for labeling the resistance in external data are 0.678, 0.536, 0.935, 0.468, and 0.843, respectively. Our ANN model performs satisfactorily statistically on the external dataset as a result.

Our major point is that the built ANN model is capable of ranking the efficiency of PIs for a given strain. The external dataset contains 853 pair of resistance scores for 51 different molecules. Our ANN model predicted these 853 pair of resistance scores as well, and we measured the ranking performance of the model. This testing approach is identical to the previously described tendency and ranking assessments of the eight existing PIs. For this classification task, the accuracy, sensitivity, specificity, MCC, and AUC scores have been 0.749, 0.767, 0.730, 0.497, and 0.819, respectively (see Table 3). As a result, our ANN model appropriately ranks the compounds based on their resistance profiles (see Figure 5 for a representative example). The ROCs for both resistance classification and ranking classification performances of the ANN model can be seen in Figure 6. The ANN model learned considerable information from the molecular structure of the eight PIs, according to the ROCs. In summary, if the external PIs have no information loss (no 1s in the discarded 278 bits within 512 bits) in comparison to the existing eight PIs, our ANN model has a high ability to rank these PIs.

One method for reducing and visualizing high-dimensional data is principal component analysis (PCA). It is widely used to describe the chemical space occupied by a set of molecules. When the descriptors of the compounds are used for analysis, it allows for the clustering of similar molecules as well as the distinguishing of diverse molecules. To verify the applicability of our model in terms of chemical similarity and diversity, we have performed the PCA on the external set molecules using 234-dimensional vectors employed for fold-change prediction. We only evaluated the unique molecules in the external set, and for molecules that were tested in several strains and had multiple fold-change prediction values, we have taken the values with the largest absolute prediction error (AE). As the first two components, PC1 and PC2, were able to represent more than half of the variance in the data and accurately depict the correlations between the similarities of the molecules, 234-dimensional vectors for each unique molecule were projected to 2D-space. Figure 7 displays the PCA plot for both training and external set compounds and colored according to the AE for fold-change prediction or training compound name. According to the figure, other training compounds other than the IDV were not involved in cluster formation with external set molecules. Though the same external set molecules were not forming clusters, still there were three clusters that contain three or more molecules. For each of these cluster, we have selected one representative structure and additionally we have found the maximum common substructure (MCS) using the FMCS algorithm implemented in RDKit. We discovered that for all compounds in the cluster with the representative structure A in Figure 7, the fold-changes were predicted with low errors (AE < 1.0). The compounds in this cluster have certain substructures with HIV inhibitors, such as a sulfonamide group, a bis-THF alcohol moiety, and a benzodioxole group [46]. The only structural difference in this cluster of compounds was observed on the side chain connecting to the phenoxy group. Our model, on the other hand, did not perform well for another cluster of compounds represented by structure B in Figure 7. Despite the fact that the MCS for three compounds in this cluster were similar to the preceding cluster, represented by structure A, there were subtle differences that caused compounds to be in distinct clusters. For these, the cyclohexyl hydroxy or cyclopentyl hydroxy group was attached to the N of the sulfonamide group. Furthermore, the varying side group of the phenoxy group in the first cluster was shorter in the second cluster since there was only phenyl group. Compounds having structure C are represented in another cluster, as seen in Figure 7. These compounds had a high resemblance to the IDV (see Figure 4 for a 2D representation of the IDV), and the fold-change was predicted with minimal errors (AE < 1.0). These compounds, like the IDV, have substructure groups such as piperazinecarboxamide, indenol derivative, and phenyl group. However, the compounds in this cluster have different attachment groups connected to piperazine ring instead of pyridine group in the IDV. We have also discovered that for the compound where the fold change estimation error above the threshold (AE ≥ 1.0), the exact value was 1.01, indicating that it is negligible. It should be noted, however, that this compound contains a Fluorine substituent, which is known to cause a significant increase in the inhibitory activity of compounds [47] and it may not be easily learnt using our model. This chemical similarity analysis revealed that our model can predict the fold change for similar compounds with low error, despite the fact that some of the external set molecules based on our descriptors were not similar (or in close proximity in the PCA plot in Figure 7) to internal set molecules.

Discussions

This paper presents a machine learning approach for predicting fold-change values using HIV-1 protease inhibitor and isolate characteristics. The filtered PhenoSense assay results made accessible in the Stanford HIV drug resistance database have been used training and testing machine learning models. Seven of the eight inhibitors have been used to train drug-isolate-fold change-based feed-forward artificial neural networks, with the remaining one serving as test data (LVO). In this context, the LVO procedure produces an objective testing approach for determining the learning capacity of models from the descriptors of the inhibitors. Both inhibitors and isolates have been encoded using binary mappings, which have been shown to be computationally effective. Because of their acknowledged advantages in molecular machine learning models, the Morgan fingerprints have been exploited as binary mappings of protease inhibitors [41-43]. An efficient ensemble process has been proposed and verified through various quantitative experiments to handle the overfitting trouble.

The most significant contribution of this research is the construction of drug-isolate-fold change (DIF)-based ANN models, as opposed to the widely studied isolate-fold change (IF)-based models [29-40]. Because the IF models do not take the molecular fingerprints as input, they are insensitive to molecular structure. This study shows the possibility of achieving such a generalized model by feeding models with adequate data from various PIs in the presence of isolates. With the use of the LVO procedure throughout our investigation, the current DIF-based models have been shown to be capable of predicting the drug resistance profiles of unknown inhibitors. Even though the Stanford HIV database only has eight available inhibitors, having many isolates for each inhibitor has contributed to the learning process, and reasonable predictions have been found in the regression performance of remaining inhibitors.

The prediction of drug resistance tendencies for each PI pair is an unavoidable expectation from our DIF-based ANN models. Our generalized models can predict resistance trends with high 2D correlation scores, as demonstrated here. The DIF-based models have provided satisfactory accuracy, sensitivity, specificity, and MCC values by creating classification problems from the tendency relations of each PI pair. The DIF-based ANN model utilizes valuable information from the Morgan fingerprints to predict the fold change values of hidden inhibitors, according to our all-quantitative observations.

We have shown that our ANN model can categorize resistant and susceptible strains as well as rank inhibitors based on resistance profiles for an external dataset. The external dataset is designed in such a way that the unique molecules are comparable enough to any of the primary eight PI and there is no bit loss in the reduction procedure of the Morgan fingerprints from 512 to 234 dimensions. Thus, whenever a molecule satisfies these conditions, our DIF-based model has ability to compare this molecule with existing PIs in terms of their resistance scores.

Instead of building independent separate models for each inhibitor, this study offers a fresh viewpoint on the field by incorporating inhibitor characteristics on the input side of machine learning models. The most obvious drawback of our model is the dearth of protease inhibitors with sufficient genotype-phenotype information. Nevertheless, our encouraging findings have demonstrated that including genotype-phenotype information of novel protease inhibitors will help build more generic drug-isolate-fold change-based machine learning models. Additionally, feeding the DIF-based models with data from many conventional and nonconventional inhibitors may result in a unified model for forecasting drug resistance tendencies for any PI pair in the presence of known genotypes. The drug development process for evolvable diseases, such as HIV, bacterial infections, and cancer should be fundamentally different from diseases such as blood-pressure regulators. A drug needs to be effective and stay effective through the test of evolution. Predicting resistance potentials for drugs is becoming a necessity.

Conclusions

This study has revealed the advantages of developing DIF-based models to predict drug resistance profiles. Instead of IF-based models, the current approach has allowed us to investigate a new model that can predict the drug resistance tendencies of PI pairs. Even with only eight PIs available, internal and external test results show that the DIF-based model takes significant information from inhibitor descriptors and leads to satisfactory regression performance. As a result, after finishing this study, it is highlighted on the research agenda to train ANN models with more inhibitors by expanding the existing dataset. In this context, it will be feasible to track the drug resistance profiles of any novel protease inhibitor, and it is strongly believed that these insightful forecasts will be a right direction for moving forward.

Author Contributions

Conceptualization: [Huseyin Tunc, Enes Seyfullah Kotil]; Methodology: [Huseyin Tunc, Enes Seyfullah Kotil, Berna Dogan, Busra Nur Darendeli Kiraz]; Formal analysis and investigation: [Huseyin Tunc]; Writing - original draft preparation: [Huseyin Tunc]; Writing - review and editing: [Huseyin Tunc, Murat Sari, Serdar Durdagi, Enes Seyfullah Kotil, Berna Dogan]; Funding acquisition: [Enes Seyfullah Kotil]; Resources: [Enes Seyfullah Kotil]; Supervision: [Murat Sari, Serdar Durdagi, Enes Seyfullah Kotil].

Data and Software Availability

All data and necessary codes are deposited to:
https://github.com/tnchsyn/hivdrugisolatefoldchange_model
https://doi.org/10.5281/zenodo.7527918

References

1. Sharp P, Hahn BH (2011) Origins of HIV and the AIDS Pandemic. Cold Spring Harb Perspect 1:006841. https://doi. org/10.1101/cshperspect.a006841
2. Das K, Arnold E HIV-1 (2013) Reverse transcriptase and antiviral drug resistance (Part 1 of 2). Curr Opin Virol 3(2):111–118. https://doi. org/10.1016/j.coviro.2013.03.014
3. Lu DY, Wu HY, Yarla, NS, Xu B, Ding J, Lu TR (2018) HAART in HIV/AIDS Treatments: Future Trends. Infect Disord Drug Targets 18(1):15-22. https://doi. org/10.2174/1871526517666170505122800
4. Jespersen NA, Axelsen F, Dollerup J, Nørgaard M, Larsen CS (2021) The burden of non-communicable diseases and mortality inpeople living with HIV (PLHIV) in the pre-, early-andlate-HAART era. HIV Medicine. https://doi. org/10.1111/hiv.13077
5. Palmisano L, Vella S (2011) A brief history of antiretroviral therapy of HIV infection: success and challenges. Ann Ist Super Sanità 47(1):44-48. https://doi.org/10.4415/ANN_11_01_10
6. World Heath Organization. Global HIV/AIDS response: epidemic update and health sector progress towards universal access: progress report. https://apps.who.int/iris/handle/10665/44787
7. Günthard HF, Calvez V, Paredes R, Pillay D, Shafer RW, Wensing AM, Jacobsen DM, Richma DD (2019) Human Immunodeficiency Virus Drug Resistance: 2018 Recommendations of the International Antiviral Society–USA Panel. Clin Infect Dis 68(2):177–187. https://doi.org/0.1093/cid/ciy463
8. Kuritzkes DR (2011) Drug resistance in HIV-1. Curr Opin Virol 1(6):582-589. https://doi.org/10.1016/j.coviro.2011.10.020
9. Oroz M, Begovac J, Planinic A, Rokic F, Lunar MM, Zorec TM, Beluzić R, Korać P, Vugrek O, Poljak M, Lepej SZ (2019) Analysis of HIV-1 diversity, primary drug resistance

and transmission networks in Croatia. Sci Rep 9:17307. https://doi.org/10.1038/s41598-019-53520-8

10. Lagnese M, Daar ES (2008) Antiretroviral regimens for treatment-experienced patients with HIV-1 infection. Expert Opin Pharmacother 9:5:687-700. https://doi.org/ 10.1517/14656566.9.5.687
11. Cihlar T, He GX, Liu X, Chen JM, Hatada M, Swaminathan S, McDermott MJ, Yang ZY, Mulato AS, Chen X, Leavitt SA, Stray KM, Lee WA (2006) Suppression of HIV-1 protease inhibitor resistance by phosphonate-mediated solvent anchoring. J Mol Biol 363:635–647. https://doi.org/ 10.1016/j.jmb.2006.07.073
12. Stranix BR, Sauve G, Bouzide A, Cote A, Sevigny G, Yelle J (2003) Lysine sulfonamides as novel HIV-protease inhibitors: Optimization of the Nepsilon-acyl-phenyl spacer. Bioorg Med Chem Lett 13:289–4292. https://doi.org/ 10.1016/j.bmcl.2003.09.058
13. Nakatani S, Hidaka K, Ami E, Nakahara K, Sato A, Nguyen JT, Hamada Y, Hori Y, Ohnishi N, Nagai A, Kimura T, Hayashi Y, Kiso Y (2008) Combination of non-natural D-amino acid derivatives and allophenylnorstatine-dimethylthioproline scaffold in HIV protease inhibitors have high efficacy in mutant HIV. J Med Chem 51:2992–3004. https://doi.org/10.1021/jm701555p
14. Koh Y, Das D, Leschenko S, Nakata H, Ogata-Aoki H, Amano M, Nakayama M, Ghosh AK, Mitsuya H (2009) GRL-02031, a novel nonpeptidic protease inhibitor (PI) containing a stereochemically defined fused cyclopentanyltetrahydrofuran potent against multi-PI-resistant human immunodeficiency virus type 1 in vitro. Antimicrob Agents Chemother 53(3):997-1006. https://doi.org/10.1128/AAC.00689-08
15. Koh Y, Nakata H, Maeda K, Ogata H, Bilcer G, Devasamudram T, Kincaid JF, Boross P, Wang YF, Tie Y, Volarath P, Gaddis L, Harrison RW, Weber IT, Ghosh AK, Mitsuya H (2003) Novel bis-tetrahydrofuranylurethane-containing nonpeptidic protease inhibitor (PI) UIC-94017 (TMC114) with potent activity against multi-PI-resistant human immunodeficiency virus in vitro. Antimicrob Agents Chemother 47(10):3123-9. https://doi.org/ 10.1128/AAC.47.10.3123-3129.2003
16. Zhang S, Kaplan AH, Tropsha A (2008) HIV-1 protease function and structure studies with the simplicial neighborhood analysis of protein packing method. Proteins 73(3): 742–753. https://doi.org/ 10.1002/prot.22094

17. World Heath Organization. Updated recommendations on first-line and second-line antiretroviral regimens and post-exposure prophylaxis and recommendations on early infant diagnosis of HIV, https://www.who.int/publications/i/item/WHO-CDS-HIV-18.51.

18. Rosenbloom DIS, Hill AL, Rabi SA (2012) Antiretroviral dynamics determines HIV evolution and predicts therapy outcome. Nat Med 18(9):1378-1386. https://doi.org/10.1038/nm.2892

19. Jilek BL, Zarr M, Sampah ME, Rabi SA, Bullen CK, Lai J, Shen L, Siliciano RF (2011) A quantitative basis for antiretroviral therapy for HIV-1 infection. Nat Med 18(3):446-452. https://doi.org/ 10.1038/nm.2649

20. Xing H, Ruan Y, Li J, Shang H, Zhong P, Wang X, Liao L, Li H, Zhang M, Xue Y, Wang Z, Su B, Liu W, Dong Y, Ma Y, Li H, Qin G, Chen L, Pan X, Chen X, Peng G, Fu J, Chen RY, Kang L, Shao Y (2013) HIV Drug Resistance and Its Impact on Antiretroviral Therapy in Chinese HIV-Infected Patients. PLoS ONE 8(2): e54917. https://doi.org/10.1371/journal.pone.0054917

21. Lima VD, Gill VS, Yip B, Hogg RS, Montaner JS, Harrigan PR (2008) Increased Resilience to the Development of Drug Resistance with Modern Boosted Protease Inhibitor-Based Highly Active Antiretroviral Therapy. J Infect Dis 198(1):51–58. https://doi.org/10.1086/588675

22. Wei Y, Li J, Chen Z, Wang F, Huang W, Hong Z, Lin J (2015) Multistage virtual screening and identification of novel HIV-1 protease inhibitors by integrating SVM, shape, pharmacophore and docking methods. Eur J Med Chem 101:409–418. https://doi.org/10.1016/j.ejmech.2015.06.05

23. Yu X, Weber IT, Harrison RW (2014) Prediction of HIV drug resistance from genotype with encoded three-dimensional protein structure. BMC Genom 15:1-13. https://doi.org/10.1186/1471-2164-15-S5-S1

24. Hosseini A, Alibés A, Noguera-Julian M (2016) Computational Prediction of HIV-1 Resistance to Protease Inhibitors. J Chem Inf Model 56(5): 915–923. https://doi.org/10.1021/acs.jcim.5b00667

25. Talbot A, Grant P, Taylor J, Baril JG, Liu TF, Charest H, Brenner B, Roger M, Shafer R, Cantin R, Zolopa A (2010) Predicting tipranavir and darunavir resistance using genotypic, phenotypic, and virtual phenotypic resistance patterns: an independent cohort analysis of

clinical isolates highly resistant to all other protease inhibitors. Antimicrob Agents Chemother 54:2473-2479. https://doi.org/10.1128/AAC.00096-10

26. Obermeier M, Pironti A, Berg T, Braun P, Daumer M, Eberle J, Ehret R, Kaiser R, Kleinkauf N, Korn K, Kücherer C, Müller H, Noah C, Stürmer M, Thielen A, Wolf E, Walter H (2012) HIVGRADE: a publicly available, rules-based drug resistance interpretation algorithm integrating bioinformatic knowledge. Intervirology 55:102-107. https://doi.org/10.1159/000331999
27. Van Laethem K, De Luca A, Antinori A, Cingolani A, Perno CF (2002) A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients. Antivir Ther 7:123–129. https://doi.org/10.1177/135965350200700206
28. Meynard JL, Vray M, Morand-Joubert L, Race E, Descamps D, Peytavin G, Matheron S, Lamotte C, Guiramand S, Costagliola D, Brun-Vezinet F, Clavel F, Girard PM (2002) Phenotypic or genotypic resistance testing for choosing antiretroviral therapy after treatment failure: a randomized trial. AIDS 16:727–736. https://doi.org/10.1097/00002030-200203290-00008
29. Amamuddy OS, Bishop NT, Bishop OT (2017) Improving fold resistance prediction of HIV-1 against protease and reverse transcriptase inhibitors using artificial neural networks. BMC Bioinform 18:369-376. https://doi.org/10.1186/s12859-017-1782-x
30. Amamuddy OS, Bishop NT, Bishop OT (2018) Characterizing early drug resistance-related events using geometric ensembles from HIV protease dynamics. Sci Rep 8:17938. https://doi.org/10.1038/s41598-018-36041-8
31. Wang D, Larder B (2003) Enhanced Prediction of Lopinavir Resistance from Genotype by Use of Artificial Neural Networks. J Infect Dis 88(5):653–660. https://doi.org/ 10.1086/377453
32. Drăghici S, Potter RR (2002) Predicting HIV drug resistance with neural networks. Bioinformatics 19(1):98-107. https://doi.org/ 10.1093/bioinformatics/19.1.98
33. Kjaer J, Høj L, Fox Z, Lundgren J (2008) Prediction of phenotypic susceptibility to antiretroviral drugs using physiochemical properties of the primary enzymatic structure combined with artificial neural networks. HIV Medicine 9:642-652. https://doi.org/ 10.1111/j.1468-1293.2008.00612.x

34. Steiner MC, Gibson KM, Crandall KA (2020) Drug Resistance Prediction Using Deep Learning Techniques on HIV-1 Sequence Data. Viruses 12:560. https://doi.org/10.3390/v12050560
35. Wang D, Larder B, Revell A, Montaner J, Harrigan R, De Wolf F, Lange J, Wegner S, Ruiz L, Pérez-Elías MJ, Emery S, Gatell J, Monforte AD, Torti C, Zazzi M, Lane C (2009) A Comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy. Artif Intell Med 47:63-74. https://doi.org/10.1016/j.artmed.2009.05.002
36. Shen CH, Yu X, Harrison RW, Weber IT (2016) Automated prediction of HIV drug resistance from genotype data. BMC Bioinform 17:278-285. https://doi.org/10.1186/s12859-016-1114-6
37. Shah D, Freas C, Weber IT, Harrison RW (2020) Evolution of drug resistance in HIV protease. BMC Bioinform 21:497-512. https://doi.org/10.1186/s12859-020-03825-7
38. Tarasova O, Biziukova N, Kireev D, Lagunin A, Ivanov S, Filimonov D, Poroikov V (2020) A Computational Approach for the Prediction of Treatment History and the Effectiveness or Failure of Antiretroviral Therapy. Int J Mol Sci 21(3):748. https://doi.org/10.3390/ijms21030748
39. Tarasova O, Biziukova N, Filimonov D, Poroikov V (2018) A Computational Approach for the Prediction of HIV Resistance Based on Amino Acid and Nucleotide Descriptors. Molecules 23(11):2751. https://doi.org/10.3390/molecules23112751
40. Ota R, So K, Tsuda M, Higuchi Y, Yamashita F (2021) Prediction of HIV drug resistance based on the 3D protein structure: Proposal of molecular field mapping. PLoS ONE 16(8):e0255693. https://doi.org/10.1371/journal.pone.0255693
41. Cai Q, Yuan R, He J, Menglong L, Yanzhi G (2021) Predicting HIV drug resistance using weighted machine learning method at target protein sequence-level. Mol Divers, 25:1541–1551. https://doi.org/10.1007/s11030-021-10262-y
42. Beerenwinkel N, Daumer M, Oette M, Korn K, Hoffmann D, Kaiser R, Lengauer T, Selbig J, Walter H (2003) Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. Nucleic Acids Res 200331:3850-3855. https://doi.org/10.1093/nar/gkg575
43. Beerenwinkel N, Schmidt B, Walter H, Kaiser R, Lengauer T, Hoffmann D, Korn K, Selbig J (2002) Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to

predicting phenotype from genotype. PNAS 99:8271-8276. https://doi.org/10.1073/pnas.112177799

44. Pawar SD, Freas C, Weber IT, Harrison RW (2018) Analysis of drug resistance in HIV protease. BMC Bioinform 19:362-368. https://doi.org/10.1186/s12859-018-2331-y
45. Rhee SY, Taylor J, Fessel WJ (2010) HIV-1 Protease Mutations and Protease Inhibitor Cross-Resistance. Antimicrob Agents Chemother 54(10):4253–4261. https://doi.org/10.1128/AAC.00574-10
46. Sevenich A, Liu GQ, Arduengo III AJ, Gupton BF, Opatz T. (2017) Asymmetric One-Pot Synthesis of (3 R, 3a S, 6a R)-Hexahydrofuro [2, 3-b] furan-3-ol: A Key Component of Current HIV Protease Inhibitors. J Org Chem Jan 82(2):1218-23. https://doi.org/10.1021/acs.joc.6b02588
47. Amano M, Yedidi RS, Salcedo-Gómez PM, Hayashi H, Hasegawa K, Martyr CD, Ghosh AK, Mitsuya H. (2022) Fluorine Modifications Contribute to Potent Antiviral Activity against Highly Drug-Resistant HIV-1 and Favorable Blood-Brain Barrier Penetration Property of Novel Central Nervous System-Targeting HIV-1 Protease Inhibitors In Vitro. Antimicrob Agents Chemother 66(2):e01715-21. https://doi.org/10.1128/aac.01715-21

Table 1(on next page)

Predictive performance of DIF-based ANN models.

Mean square error (MSE) and R^2 values of the DIF-based ANN models for predicting the logarithmic fold change values of eight PIs are presented. Drug-isolate-fold change models are constructed as a general neural network model taking drug fingerprints and mutation information as inputs. For each row, the corresponding drug has not been included in the training process. The test set performance of each model has been evaluated with respect to the excluded drugs. 100*50 simulations with random weights have been done, and 100 neural network models that yield minimum MSE for interior test set among 50 trials are obtained. The final neural network model is achieved by taking the mean of 100 models. Abbreviations: ATV, atazanavir; DRV, darunavir; FPV, fosamprenavir; IDV, indinavir; LPV, lopinavir; NFV, nelfinavir; SQV, saquinavir; TPV, tipranavir.

Table 1. Mean square error (MSE) and R^2 values of the DIF-based ANN models for predicting the logarithmic fold change values of 8 PIs are presented[b].

ARVs[a]	R^2		MSE	
	Whole dataset	Test set	Whole dataset	Test set
ATV	0.865	0.778	0.087	0.166
DRV	0.857	0.736	0.092	0.227
FPV	0.849	0.738	0.097	0.160
IDV	0.861	0.811	0.090	0.131
LPV	0.852	0.822	0.096	0.188
NFV	0.845	0.757	0.101	0.215
SQV	0.833	0.731	0.109	0.283
TPV	0.821	0.359	0.116	0.560

[a] Abbreviations: ATV, atazanavir; DRV, darunavir; FPV, fosamprenavir; IDV, indinavir; LPV, lopinavir; NFV, nelfinavir; SQV, saquinavir; TPV, tipranavir.

[b] Drug-isolate-fold change models are constructed as a general neural network model taking drug fingerprints and mutation information as inputs. For each row, the corresponding drug has not been included in the training process. The test set performance of each model has been evaluated with respect to the excluded drugs. 100×50 simulations with random weights have been done, and 100 neural network models that yield minimum MSE for interior test set among 50 trials are obtained. The final neural network model is achieved by taking the mean of 100 models.

Table 2(on next page)

Accuracy, sensitivity, specificity and MCC values of the DIF-based ANN models for predicting the drug resistance tendencies for each couple of ARVs.

Accuracy, sensitivity and specificity values represent the rate of true predictions, true positive rate and true negative rate, respectively. The common genotype data is used for each PI pair by eliminating the observations satisfying |log(A)-log(B)|≤log2 where A and B are the fold change values of drugs A and B for a specified genotype.

Table 2. Accuracy, sensitivity, specificity and MCC values of the DIF-based ANN models for predicting the drug resistance tendencies for each couple of ARVs[a].

ARVs		ATV	DRV	FPV	IDV	LPV	NFV	SQV	TPV
ATV	Accuracy	-	0.970 (224/231)	0.932 (438/470)	0.923 (264/286)	0.913 (303/332)	0.948 (361/381)	0.896 (301/336)	0.983 (404/411)
	Sensitivity	-	0.500 (7/14)	0.772 (78/101)	0.850 (85/100)	0.978 (178/182)	0.986 (291/295)	0.720 (90/125)	0.000 (0/6)
	Specificity	-	1.000 (217/217)	0.976 (360/369)	0.962 (179/186)	0.833 (125/150)	0.814 (70/86)	1.000 (211/211)	0.998 (404/405)
	MCC	-	0.696	0.791	0.829	0.828	0.856	0.786	0.000
DRV	Accuracy	0.970 (224/231)	-	0.968 (184/190)	0.917 (222/242)	0.960 (215/224)	0.976 (321/329)	0.936 (206/220)	0.897 (156/174)
	Sensitivity	1.000 (217/217)	-	0.972 (172/177)	0.966 (198/205)	0.982 (214/218)	1.000 (308/308)	0.972 (173/178)	0.714 (40/56)
	Specificity	0.500 (7/14)	-	0.923 (12/13)	0.649 (24/37)	0.167 (1/6)	0.619 (13/21)	0.786 (33/42)	0.983 (116/118)
	MCC	0.696	-	0.792	0.662	0.162	0.777	0.788	0.761
FPV	Accuracy	0.932 (438/470)	0.968 (184/190)	-	0.936 (677/723)	0.952 (511/537)	0.964 (878/911)	0.930 (705/758)	0.932 (369/396)
	Sensitivity	0.976 (360/369)	0.923 (12/13)	-	0.993 (552/556)	0.996 (465/467)	0.996 (817/820)	0.975 (502/515)	0.511 (24/47)
	Specificity	0.772 (78/101)	0.972 (172/177)	-	0.749 (125/167)	0.657 (46/70)	0.670 (61/91)	0.835 (203/243)	0.989 (345/349)
	MCC	0.791	0.792	-	0.816	0.771	0.782	0.838	0.630
IDV	Accuracy	0.923 (264/286)	0.917 (222/242)	0.936 (677/723)	-	0.952 (399/419)	0.952 (498/523)	0.929 (562/605)	0.957 (404/422)
	Sensitivity	0.962 (179/186)	0.649 (24/37)	0.749 (125/167)	-	0.989 (270/273)	0.994 (468/471)	0.874 (221/253)	0.280 (7/25)
	Specificity	0.850 (85/100)	0.966 (198/205)	0.993 (552/556)	-	0.884 (129/146)	0.577 (30/52)	0.969 (341/352)	1.000 (397/397)
	MCC	0.829	0.662	0.816	-	0.895	0.702	0.854	0.518
LPV	Accuracy	0.913 (303/332)	0.960 (215/224)	0.952 (511/537)	0.952 (399/419)	-	0.944 (526/557)	0.929 (509/548)	0.982 (429/437)
	Sensitivity	0.833 (125/150)	0.167 (1/6)	0.657 (46/70)	0.884 (129/146)	-	0.979 (375/383)	0.836 (173/207)	0.632 (12/19)
	Specificity	0.978 (178/182)	0.982 (214/218)	0.996 (465/467)	0.989 (270/273)	-	0.868 (151/174)	0.985 (336/341)	0.998 (417/418)
	MCC	0.828	0.162	0.770	0.895	-	0.869	0.850	0.755
NFV	Accuracy	0.948 (361/381)	0.976 (321/329)	0.964 (878/911)	0.952 (498/523)	0.944 (526/557)	-	0.935 (735/786)	0.966 (477/494)
	Sensitivity	0.814 (70/86)	0.619 (13/21)	0.670 (61/91)	0.577 (30/52)	0.868 (151/174)	-	0.451 (41/91)	0.188 (3/16)
	Specificity	0.986 (291/295)	1.000 (308/308)	0.996 (817/820)	0.994 (468/471)	0.979 (375/383)	-	0.999 (694/695)	0.992 (474/478)
	MCC	0.846	0.777	0.782	0.702	0.869	-	0.639	0.268
SQV	Accuracy	0.896 (301/336)	0.936 (206/220)	0.930 (705/758)	0.929 (562/605)	0.929 (509/548)	0.935 (735/786)	-	0.898 (359/400)
	Sensitivity	1.000 (211/211)	0.786 (33/42)	0.835 (203/243)	0.969 (341/352)	0.985 (336/341)	0.999 (694/695)	-	0.146 (7/48)
	Specificity	0.720	0.972	0.975	0.874	0.836	0.451	-	1.000

		(90/125)	(173/178)	(502/515)	(221/253)	(173/207)	(41/91)		(352/352)
	MCC	0.786	0.788	0.838	0.854	0.850	0.639	-	0.361
TPV	Accuracy	0.983 (404/411)	0.897 (156/174)	0.932 (369/396)	0.957 (404/422)	0.982 (429/437)	0.966 (477/494)	0.898 (359/400)	-
	Sensitivity	0.998 (404/405)	0.983 (116/118)	0.989 (345/349)	1.000 (397/397)	0.998 (417/418)	0.992 (474/478)	1.000 (352/352)	-
	Specificity	0.000 (0/6)	0.714 (40/56)	0.511 (24/47)	0.280 (7/25)	0.632 (12/19)	0.188 (3/16)	0.146 (7/48)	-
	MCC	0.000	0.761	0.630	0.518	0.755	0.268	0.361	-

[a] Accuracy, sensitivity and specificity values represent the rate of true predictions, true positive rate and true negative rate, respectively. The common genotype data is used for each PI pair by eliminating the observations satisfying $|\log(A) - \log(B)| \leq log2$ where A and B are the fold change values of drugs A and B for a specified genotype.

Table 3(on next page)

Classification performance of the ANN model on the external test data.

In resistance classification procedure, the ANN model is used to classify the genotypes as drug resistant (Fold Change >=3) and susceptible (Fold Change<3) for each external test data. On the other hand, for ranking classification procedure, the ANN ranks the drugs for a given genotype in terms of their logarithmic fold change values.

Table 3. Classification performance of the ANN model on the external test data. In resistance classification procedure, the ANN model is used to classify the genotypes as drug resistant ($Fold\ Change \geq 3$) and susceptible ($Fold\ Change < 3$) for each external test data. On the other hand, for ranking classification procedure, the ANN ranks the drugs for a given genotype in terms of their logarithmic fold change values.

Model / Metric	Accuracy	Sensitivity	Specificity	MCC	AUC
Resistance Classification	0.678	0.536	0.935	0.468	0.843
Ranking Classification	0.749	0.767	0.730	0.497	0.819

Figure 1

Data versus the predicted fold change values from DIF-based ANN models.

DIF-based ANN regression models are constructed with the LVO testing methodology. For each figure, the fold-change results are estimated by an ANN model, which is trained with the remaining data of the seven PIs. The R^2 values correspond to the square of the linear correlation coefficient of the data and prediction.

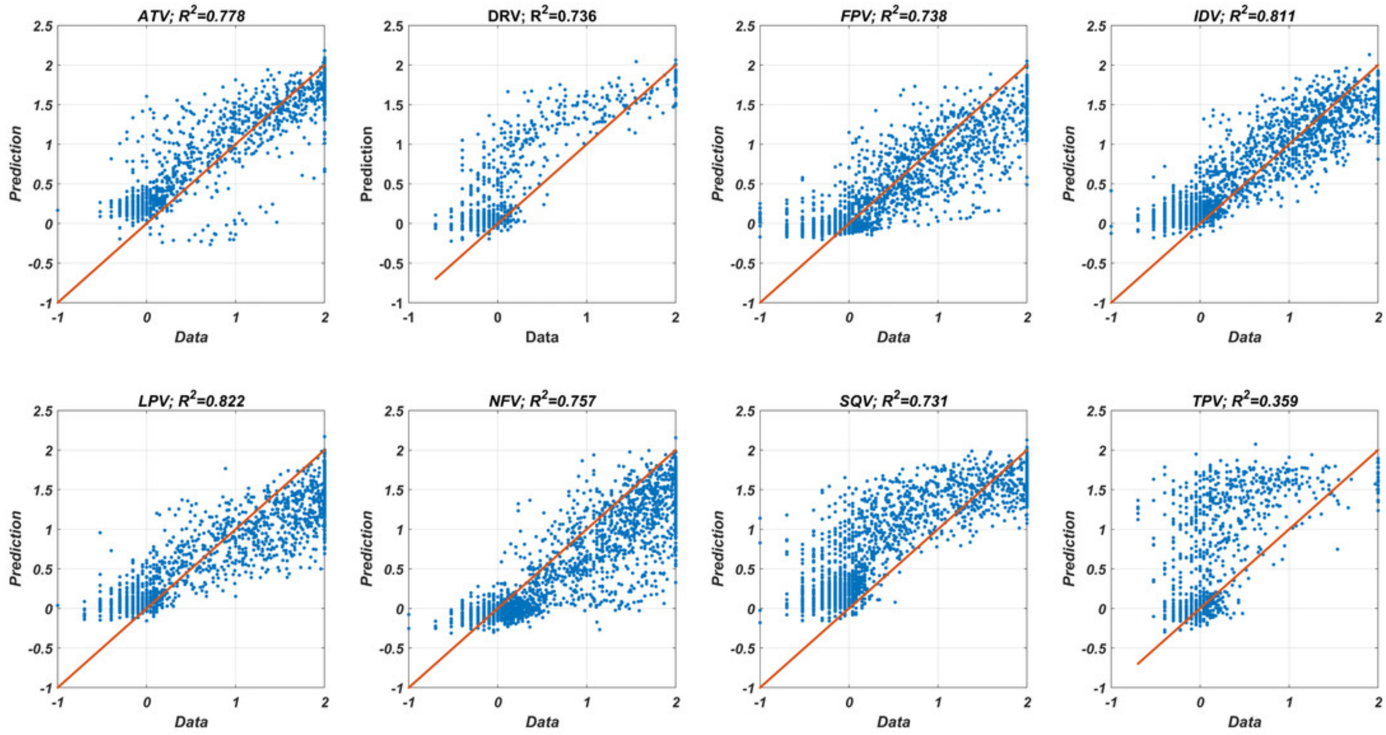


Figure 2

Prediction of the fold-change tendencies with the DIF-based ANN model for each PI pair.

The common isolate data of each PI pair and the corresponding DIF-based ANN model predictions are illustrated with 2D correlation coefficients. For each PI, the prediction is constructed using the DIF-based ANN model, which is trained with the remaining seven PI data. The illustrations show the tendencies of the drug resistances for each PI pair for common genotypes.

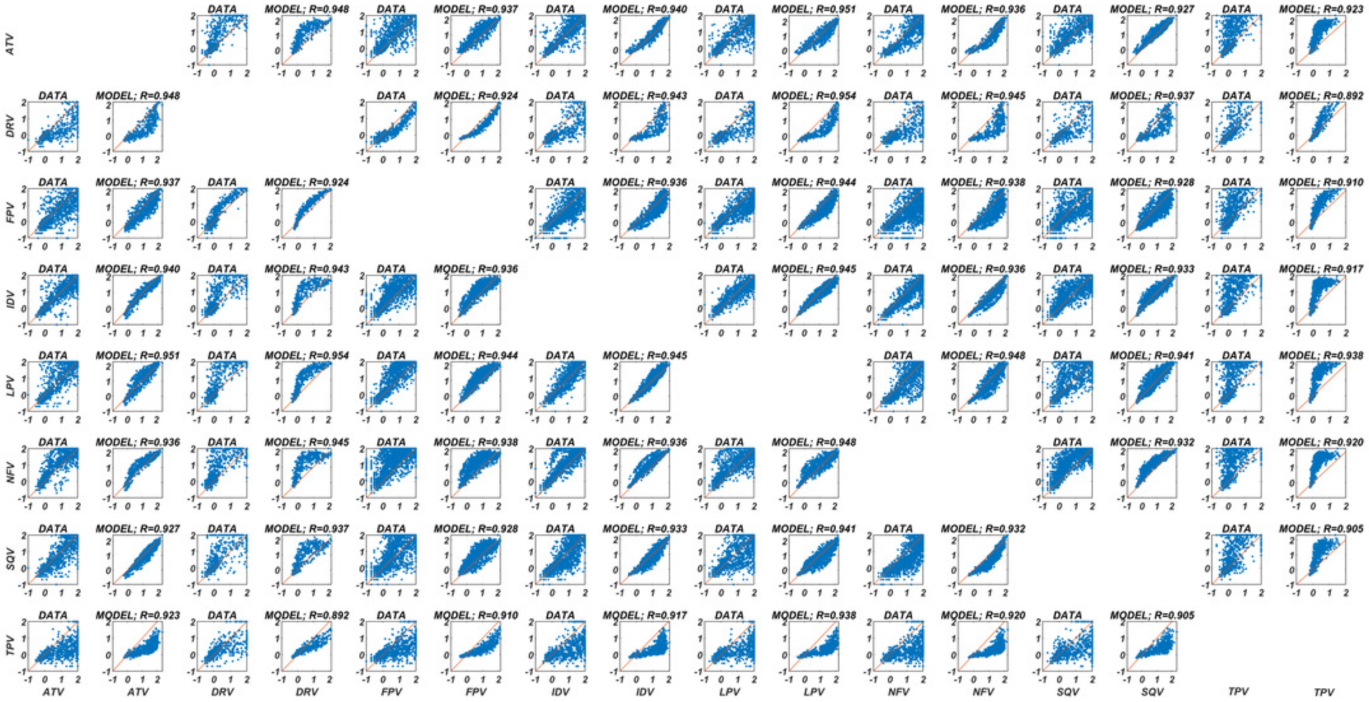


Figure 3

Classification performances of the DIF-based ANN models

The DIF-based ANN classification models are constructed with the LVO methodology. For each PI, the classification of resistant and non-resistant isolates is estimated by an ANN model trained with the remaining data of the seven PI. The AUC values correspond to the area under the ROC curves, and the accuracy is evaluated with the true estimation rate.

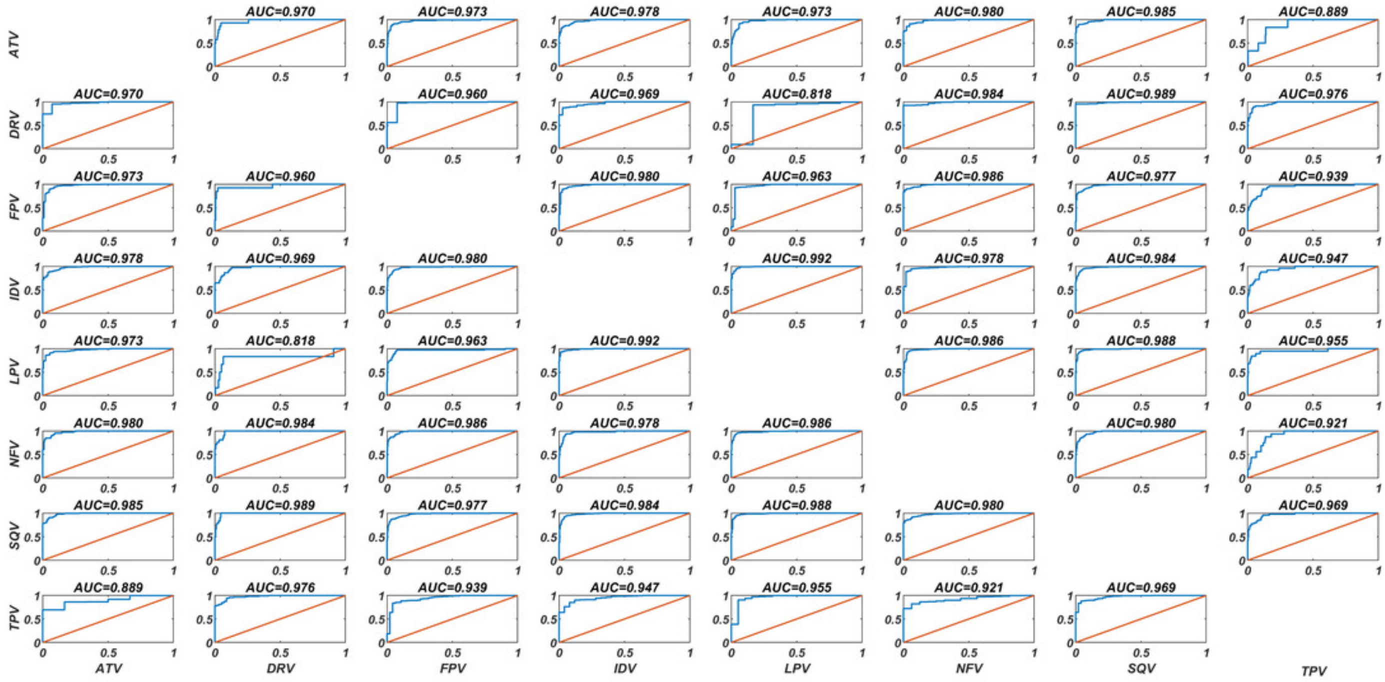


Figure 4

Chemical structures of eight PIs utilized in the DIF-based ANN model training.

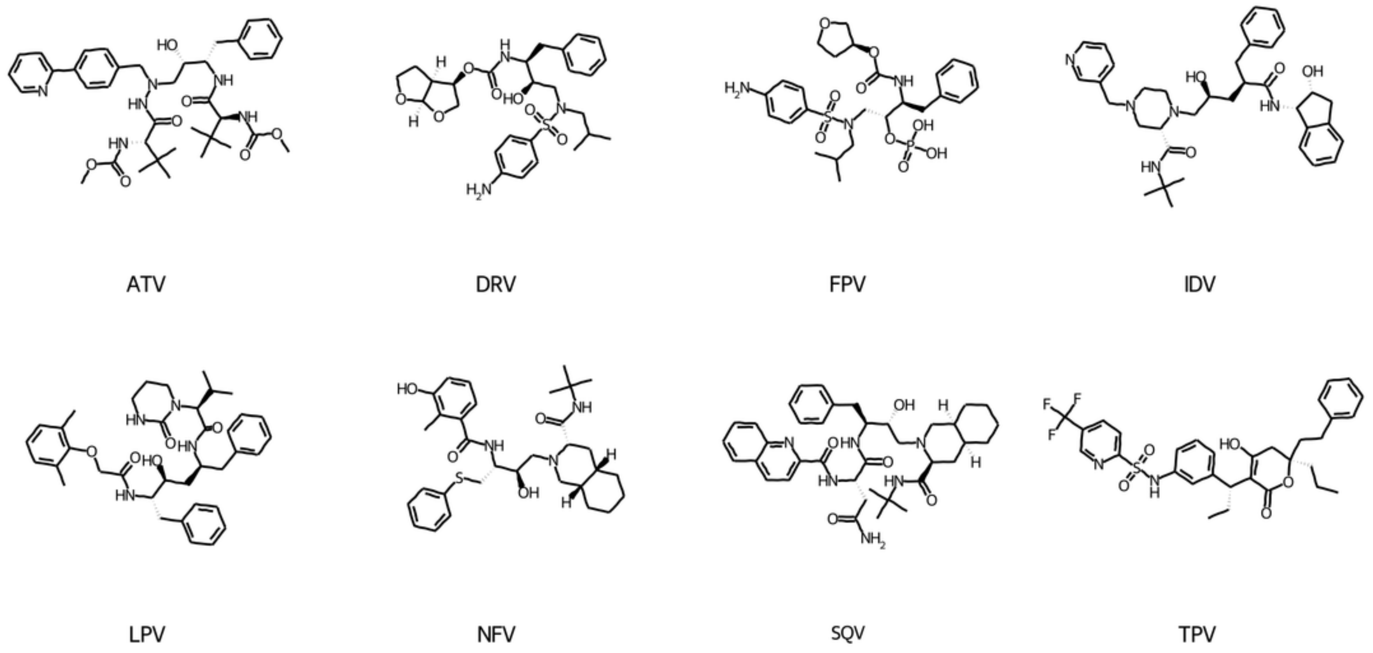
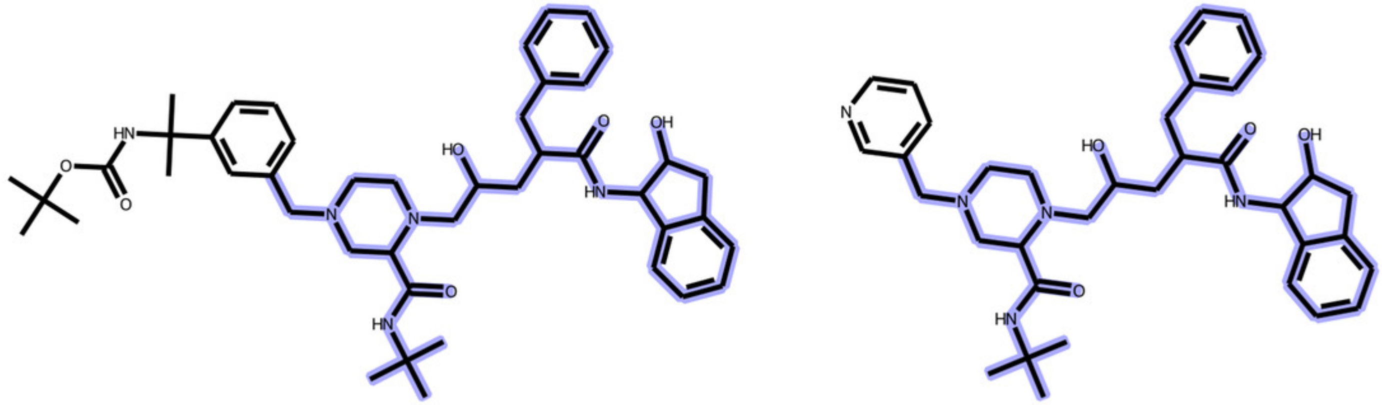


Figure 5

Resistance classification for two compounds having similar structures.

A ranking classification example for two compounds that have similar structures with different resistance profiles for the same strain (L10I, L19Q, K20R, E35D, M36I, S37N, M46I, I50V, I54V, I62V, L63P, A71V, V82A, L90M, E28K, K32E, V35I, T39T, E40DV, M41L, K43E, Y181Y).



ChEMBL ID = CHEMBL405134
Experimental Fold-change = 436.52
Predicted Fold-change = 2.82

ChEMBL ID = CHEMBL115
Experimental Fold-change = 1.15
Predicted Fold-change = 1.74

Figure 6

The ROCs corresponding to resistance and ranking classifications for the test data.

The DIF-based ANN model has been used to (A) classify the resistance and susceptible strains (B) classify the rankings of 853 pair of resistance scores, for various molecules existing our external data. The AUC ratings associated with the ROC curves measure how threshold probabilities affect FPR-TPR pairs in classification tasks.

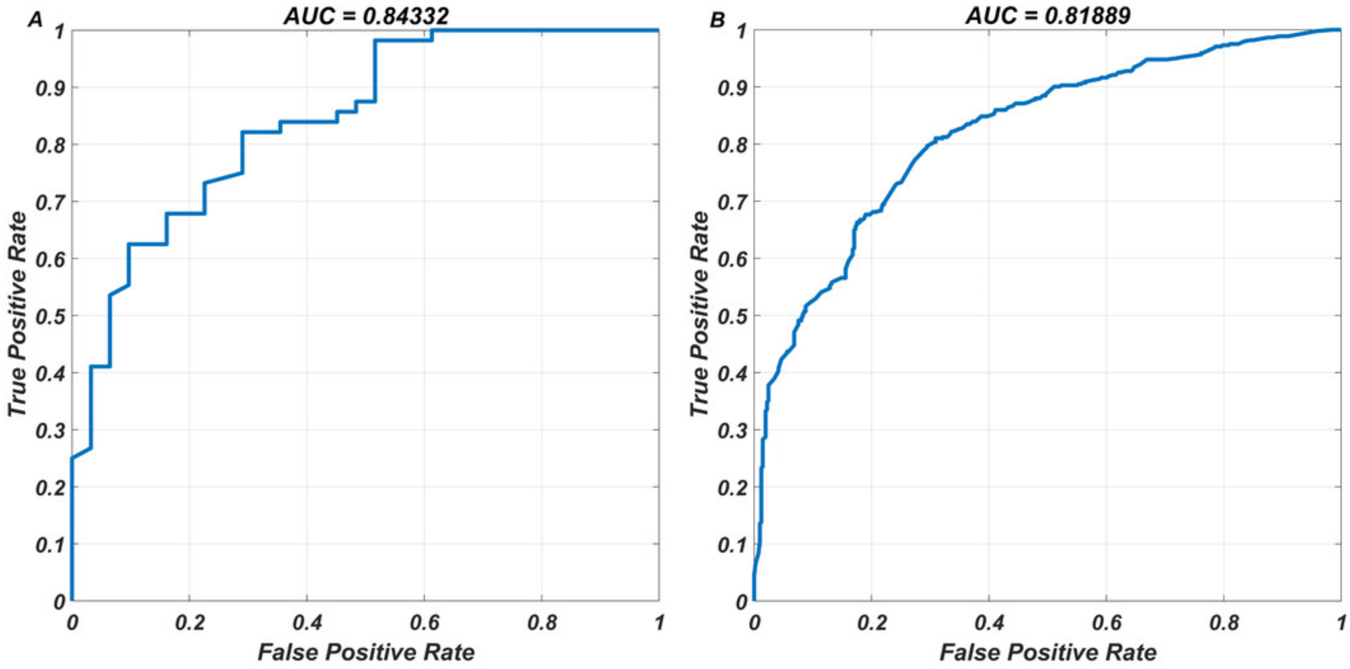
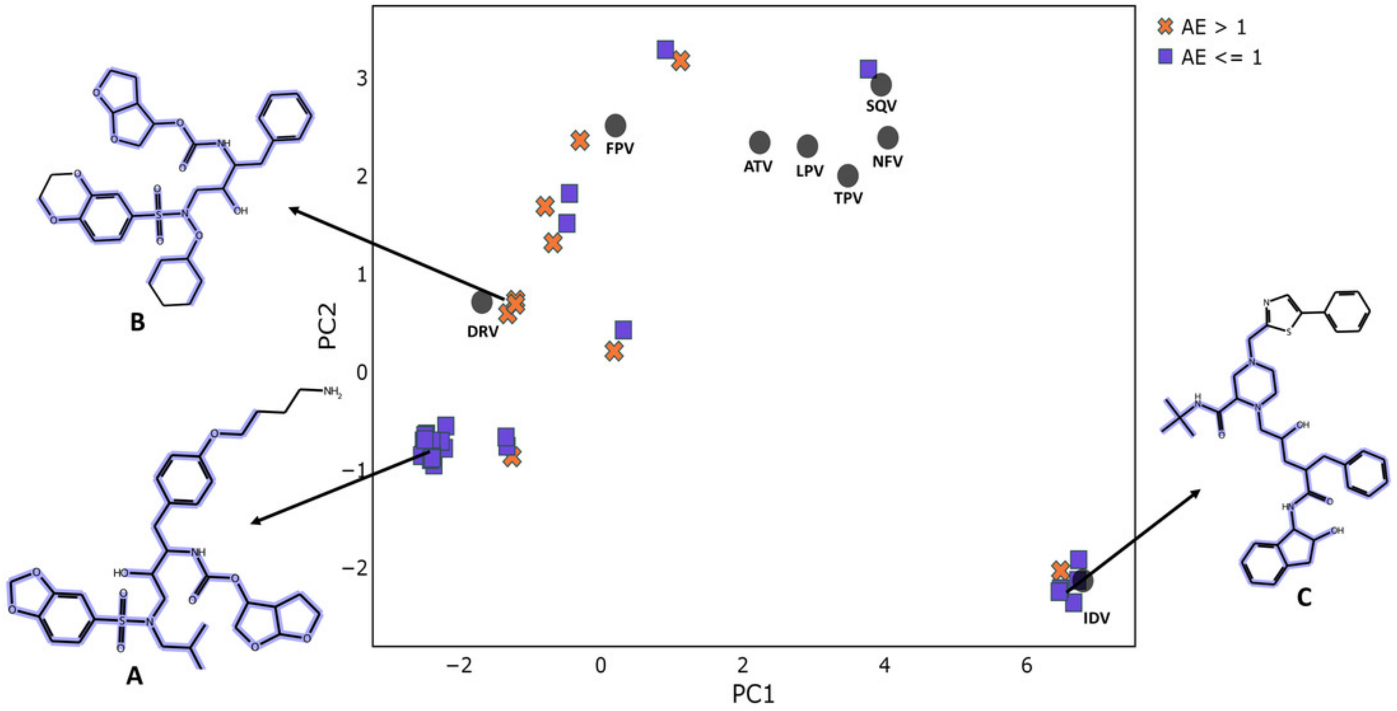


Figure 7

PCA plot for external set molecules obtained using the unique characteristic 234 bits fingerprint descriptors.

PCA plot for external set molecules produced with the unique 234-bit fingerprint descriptors. The absolute error (AE) for fold change prediction using the model is calculated for each molecule. Molecules are classified with respect to corresponding AE values (AE=1 is selected as threshold). The data points in black are for the existing eight PIs, which have their names indicated. As demonstrated, example structures from clusters are exhibited and designated A, B, and C. In the 2D depiction, the most prevalent substructures in the same clusters have been highlighted in dark blue.


AE > 1
AE <= 1
PC1
PC2
SQV
FPV
ATV
LPV
NFV
TPV
DRV
IDV
A
B
C
NH
OH
O
S
N
NH$_2$
HO
HN
N
S