Chemometrics and Intelligent Laboratory Systems. 2014; 138: 161-170

# Mapping networks of anti-HIV drug cocktails vs. AIDS epidemiology in the US counties

Diana María Herrera-Ibatá[a], Alejandro Pazos[a], Ricardo Alfredo Orbegozo-Medina[b], Humberto González-Díaz[c,d]

[a] *Department of Information and Communication Technologies, University of A Coruña UDC, 15071, A Coruña, Spain*
[b] *Department of Microbiology and Parasitology, University of Santiago de Compostela (USC), 15782, Santiago de Compostela, A Coruña, Spain*
[c] *Department of Organic Chemistry II, Faculty of Science and Technology, University of the Basque Country UPV/EHU, 48940, Leioa, Spain*
[d] *IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain*

**Abstract**
The implementation of the highly active antiretroviral therapy (HAART) and the combination of anti-HIV drugs have resulted in longer survival and a better quality of life for the people infected with the virus. In this work, a method is proposed to map complex networks of AIDS prevalence in the US counties, incorporating information about the chemical structure, molecular target, organism, and results in preclinical protocols of assay for all drugs in the cocktail. Different machine learning methods were trained and validated to select the best model. The Shannon information invariants of molecular graphs for drugs, and social networks of income inequality were used as input. The nodes in molecular graphs represent atoms weighed by Pauling electronegativity values, and the links correspond to the chemical bonds. On the other hand, the nodes in the social network represent the US counties and have Gini coefficients as weights. We obtained the data about anti-HIV drugs from the ChEMBL database and the data about AIDS prevalence and Gini coefficient from the AIDSVu database of Emory University. Box–Jenkins operators were used to measure the shift with respect to average behavior of drugs from reference compounds assayed with/in a given protocol, target, or organism. To train/validate the model and predict the complex network, we needed to analyze 152,628 data points including values of AIDS prevalence in 2310 counties in the US vs. ChEMBL results for 21,582 unique drugs, 9 viral or human protein targets, 4856 protocols, and 10 possible experimental measures. The best model found was a linear discriminant analysis (LDA) with accuracy, specificity, and sensitivity above 0.80 in training and external validation series.

**Abbreviations**
$a^{th}$, County; ANNs, Artificial neural networks; $BI_k$, Balaban's information indices; $b_j$, Experimental conditions; $CDR_{ac}(b_j)$, Cocktail–AIDS disease ratio; $c^{th}$, Cocktails; $D_a$, AIDS prevalence rate for the county; $d_{ij}$, Topological distance between atoms i and j; $d^{th}$, Drug; $G_a$, Gini measure in the county; $I^a_0$, Shannon information index of income inequality; $IC_k$, Neighborhood symmetry indices; $L_{ac}(b_j)_{obs}$, Observed variable of formation of links between counties and cocktails; MA, Moving average operators; $MI_k$, Molecular information indices; ML, Machine learning; nAT, Number of molecule atoms; $n_g$, Number of elements in the $g^{th}$ class; nSK, Number of non-H atoms; $v_d(b_j)$, Value of biological activity; W, Wiener index; $z_c$, Average of the z-scores of the biological activity of each drug; $z_d(b_j)$, z-score of the biological activity; $\eta_i$, Atom eccentricity; $\sigma_i$,, $i^{th}$ vertex distance degree

**Keywords**
ChEMBL; AIDSVu; anti-HIV drug cocktails; HAART therapy; Gini coefficient; Multiscale models; Box–Jenkins operators; Shannon entropy

## 1. Introduction

The rates of disease progression, opportunistic infections, and mortality have decreased with the implementation of the highly active antiretroviral therapy (HAART), and the combination of anti-HIV drugs has resulted in longer survival and a better quality of life for the people infected with the virus [1]. The infections with the HIV are commonly treated with drug combinations consisting of at least three different antiretroviral drugs. The most common drug treatment administered to patients consists of two nucleoside reverse transcriptase inhibitors combined with either a non-nucleoside reverse transcriptase inhibitor, or a "boosted" protease inhibitor or an integrase strand transfer inhibitors (INSTIs)-based regimen. These treatments have all resulted in decreased HIV RNA levels (< 50 copies/ml) at 48 weeks and increased CD4 cell counts in the majority of patients [2]. The targets of anti-HIV drugs are proteins present in the virus or in the host. The most important are: the reverse transcriptase enzyme (RT) that converts viral RNA genomes into DNA [3], the integrase enzyme (IN) that facilitates the incorporation of HIV-1 proviral DNA into the host cell genome, and HIV protease (PR), which is essential for viral maturation [4] and [5]. Other important viral proteins are envelope glycoprotein (Env), responsible for binding to specific target cell receptors and facilitating HIV entry [6]. On the other hand, chemokine co-receptors like CXCR4 and/or CCR5, necessary for HIV-1 entry [7], and C-C chemokine receptor types 3 and 2 (alternatives with CD4 for HIV-1 infection) [8] are important targets in the human host.

Subsequently, the antiretroviral therapy includes: the fusion and entry inhibitors, whose use is normally reserved for people who have taken a lot of anti-HIV drugs in the past. The enfuvirtide belongs to the fusion inhibitors; it inhibits the entry of HIV into the CD4 cell [9]. The CCR5 inhibitor, Maraviroc, is an entry inhibitor; it binds to the CCR5 receptor on the membrane of human cells such as CD4 cells. This binding prevents the interaction of HIV-1 gp120 and human CCR5, which is necessary for entry into the cell [10]. The nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs) are another type of anti-HIV drugs. When the HIV virus enters a healthy cell, it makes replicas of itself by using an enzyme called RT, which is responsible for transcribing viral RNA into double stranded DNA. The NRTIs work because they block that enzyme. Some examples of this class of drugs are zidovudine, didanosine, zalcitabine, stavudine, lamivudine, abacavir, tenofovir, and emcitrabine [11]. There are also non-nucleoside reverse transcriptase inhibitors (NNRTIs), whose interaction with RT induces conformational changes that inhibit the catalytic activities of the enzyme. They are characterized by their specificity for HIV-1, which makes them very selective inhibitors of the virus [12]. Five NNRTIs (nevirapine, delavirdine, efavirenz, etravirine, and rilpivirine) are currently approved by the FDA. Moreover, all of them except for delavirdine have been approved by the European Union [2]. The integrase inhibitors are another important class of anti-HIV drugs. The HIV-1 IN transfers the viral encoded DNA into the host chromosome, which is a necessary event in retrovirus replication [13]. The raltegravir and dolutegravir are examples of integrase inhibitors [14] and [15]. Lastly, the protease inhibitors are important compounds; they prevent maturation of the virus protein by competitively inhibiting HIV PR, because in HIV-1, as in all retroviruses, the production of infectious virus invariably requires an active viral protease [16]. Some examples of this kind of drugs are amprenavir, atazanavir, indinavir, nelfinavir, lopinavir, saquinavir, tipranavir, and ritonavir [17] and [18].

Some examples of combination of anti-HIV drugs approved by the FDA are Atripla®, which contains two NRTIs, Emtriva® (emtricitabine) and Viread® (tenofovir disoproxil fumarate) and an NNRTI, Sustiva® (efavirenz) [19]. Complera® contains a combination of two NRTIs (emtricitabine and tenofovir disoproxil fumarate) and an NNRTI (rilpivirine) [20]. Stribild® contains a combination of an INSTI (elvitegravir), a pharmacokinetic enhancer (cobicistat), an NRTI (emtricitabine), and a nucleotide reverse transcriptase inhibitor N(t)RTI (tenofovir disoproxil fumarate) [21]. Combivir® contains two NRTIs (zidovudine and lamivudine) [22]. Truvada® contains two NRTIs (emtricitabine/tenofovir) [23]. Kaletra® contains two protease inhibitors (lopinavir and ritonavir) [24]. Trizivir® contains a fixed-dose combination of three NRTIs (abacavir sulfate, lamivudine, and zidovudine) [25]. Epzicom® or Kivexa® in Europe contains two NRTIs (abacavir sulfate, lamivudine) [26].

In this context, the computational methods such as QSAR models are used to predict the property of a chemical compound, using information obtained from its structure [27]. To increase the accuracy, artificial intelligence techniques have been applied to a quantitative structure–activity relationships (QSAR)- or quantitative structure–property relationships (QSPR)-analysis since the late 1980s [28], [29] and [30]. Gupta et al. [31] studied the curcumine derivatives as HIV-1 integrase inhibitors, and they concluded that their model has a good predictive power for the screening of new molecules. Muthukumaran et al. [32] developed anti-HIV activity models, identifying compounds with favorable

interactions. Debnath [33] studied the applications of 3D-QSAR studies in anti-HIV-1 drug design and he stated that the structure-based drug design had been successful in identifying several drugs that were available at the time for the treatment of HIV-1, and other applications such as the design of effective analogs. Some authors [34], [35], [36] and [37] indicated that the results of their *in silico* studies provided a contribution to the design of novel active molecules for the inhibition of some target proteins involved in the HIV.

A useful model must be multi-level to account for molecular and population structure. Different types of input data are needed. At the beginning, we need the information about the chemical structure of the antiretroviral drugs and preclinical information, such as targets, organisms, assay protocols, etc. Afterwards, we need to incorporate the population structure descriptors that quantify the social and economic factors affecting the population selected for the study. Lastly, as populations in modern society are not close systems we should quantify also the effect of interaction of the population under study with other populations that may influence the pharmacoepidemiology study. We should focus on three characteristics of the problem resultant from the connection of chemical, pharmacological, and epidemiological information: (1) multi-targeting, (2) multi-objective, and/or (3) multi-scaling features. The interaction of the molecules with more than one target refers to the term multi-targeting [38], [39] and [40]. Multi-objective optimization problem (MOOP) [41], [42], [43], [44] and [45] refers to the necessity of prediction/optimization of results for different experimental measures obtained in different assays. Lastly, multi-scaling refers to the different structural levels of the organization of matter, the input variables. It means that we need to develop models able to link the changes in the AIDS prevalence in a given ($a^{th}$) population with the changes in the biological activity of the drug ($d^{th}$), due to variations in the chemical structure, detected in preclinical assays carried out under a set of $j^{th}$ boundary conditions of assay ($b_j$).

There are online resources containing epidemiological data of AIDS prevalence. One of these databases is AIDSVu (http://aidsvu.org), created by researchers at the Rollins School of Public Health at Emory University. They collected state and county-level information for AIDS prevalence in the United States. AIDSVu gathers the information from the US Centers for Disease Control and Prevention's (CDC) national surveillance database. On the other hand, there is ChEMBL (https://www.ebi.ac.uk/chembl/) [46], [47] and [48], which is one of the biggest bioactivity database with a large number of drug-like bioactive compounds. It includes data from life science experiments. In addition, there are now > 1.3 million distinct compound structures and 12 million bioactivity data points. The data are mapped to > 9000 targets, out of which 2827 are human protein targets [48].

In addition, Shannon's entropy measures are universal parameters used to codify biologically relevant information in many systems. The seminal paper "A Mathematical Theory of Communications," written by Claude Elwood Shannon [49], led to the creation of concept of information theory (IT). The IT established a connection with theoretical physics and chemistry through the concept of entropy, a link that today is firmly established. It has also been applied with some success to other disciplines [50]. Information theory in systems biology has been successfully applied to the identification of optimal pathway structures, mutual information and entropy as system response in sensitivity analysis, and quantification of input and output information [51].

## 2. Materials and methods

Quantitative descriptors of the molecular graph of the drug can be used. In particular, some of these parameters are useful to quantify information about the properties of biological, molecular, and/or social systems (information measures). We used the information indices implemented in the DRAGON software version 5.3 [52]. This software calculates different information indices, such as molecular information indices ($MI_k$) [52], Balaban's information indices ($BI_k$) [53] and [54], and neighborhood symmetry indices ($IC_k$) [52] and [55]. In this work, only the $MI_k$ information indices were used. The calculation of the $MI_k$ requires the use of different input parameters. Some of these parameters are the number of elements or nodes (atoms) of the molecular graph **G**, the number of different classes of equivalence G, and $n_g$ is the number of elements in the $g^{th}$ class, the logarithm is taken at base 2 for measuring the information content in bits, nAT is the number of molecule atoms (hydrogen included). Other parameters are ${}^{g}f_i$, which is the number of distances from the vertex $v_{i,}$ equal to g, $\eta_i$ is the atom eccentricity (i.e., the maximum topological distance from the vertex $v_i$). The parameter nSK is the number of non-H atoms. The symbol $\sigma_i$, which is the $i^{th}$ vertex distance degree (i.e., sum of topological distances from the considered atom to

any other atom), W is the Wiener index, $d_{ij}$ is the topological distance between atoms i and j. In addition, there are two basic criteria in several information indices. The first one is the equality criterion, which implies that elements are considered equivalent if their values are equal (according to this criterion $n_g$ is the number of equivalent elements, n is the total number of elements and the sum runs over all the equivalence classes). The second one is the magnitude criterion, where each element is considered as an equivalence class whose cardinality, i.e., number of elements, is equal to the magnitude of the element (according to this criterion, $n_g$ is the value of each element, n is the sum of the values of all the elements and the sum runs over all the elements). The names, symbols, and formula for the calculation of different $MI_k$ descriptors is summarized in Table 1, see details on the following references [52], [56], [57], [58], [59], [60] and [61].

**Table 1**. Names, symbols, and formula for the calculation of different $Ml_k$ descriptors.

| Symbol | D-symbol | Name | Formula | Ref. |
|---|---|---|---|---|
| $I_{tot}$ | I | Total information content | $I = n\log_2 n - \sum_{g=1}^{G} n_g \log_2 n_g$ | [56] |
| $I_{avg}$ | $\bar{I}$ | Mean information content | $\bar{I} = -\sum_{g=1}^{G} \frac{n_g}{n} \log_2 \frac{n_g}{n}$ | [56] |
| $I_{siz}$ | ISIZ | Information index on molecular size | $\mathrm{ISIZ} = \mathrm{nAT}.\log_2 n\,\mathrm{AT}$ | [57] |
| $I_{ac}$ | IAC | Total information index on atomic composition | $I = n\log_2 n - \sum_{g=1}^{G} n_g \log_2 n_g$ | [58] |
| $I_{aac}$ | AAC | Mean information index on atomic composition | $\bar{I} = -\sum_{g=1}^{G} \frac{n_g}{n} \log_2 \frac{n_g}{n}$ | [58] |
| $I_{det}$, $I_{de}$ | IDET, IDE | Total and mean information content on the distance equality | Equality of topological distances in an H-depleted molecular graph | [59] |
| $I_{dmt}$, $I_{dm}$ | IDMT, IDM | Total and mean information content on the distance magnitude | Distribution of topological distances according to their magnitude in an H-depleted molecular graph | |
| $I_{dde}$ | IDDE | Mean information content on the distance degree equality | Partition of vertex distance degrees according to their equality | |
| $I_{ddm}$ | IDDM | Mean information content on the distance degree magnitude | Partition of vertex distance degrees according to their magnitude | |
| $I_{vde}$ | IVDE | Mean information content on the vertex degree equality | Partition of vertices according to vertex degree equality | |
| $I_{vdm}$ | IVDM | Mean information content on the vertex degree magnitude | Partition of vertices according to the vertex degree magnitude | [60] |
| $I_{hvcpx}$ | HVcpx | Graph vertex complexity index | $\mathrm{HVcpx} = \frac{1}{\mathrm{nSK}} \cdot \sum_{i=1}^{\mathrm{nSK}} \left( -\sum_{g=0}^{\eta_l} {}^{g}\mathrm{f}\,\mathrm{nSK}.\log_2 {}^{g}\mathrm{f}\,\mathrm{nSK} \right)$ | [60] |
| $I_{hdcpx}$ | HDcpx | Graph distance complexity index | $\mathrm{HDcpx} = \sum_{i=1}^{\mathrm{nSK}} \frac{\sigma_i}{2W} \cdot \left( -\sum_{j=1}^{\mathrm{nSK}} \frac{d_{ij}}{\sigma_i} \cdot \log_2 \frac{d_{ij}}{\sigma_i} \right)$ | [60] and [61] |

### *2.1. ALMA models*

We have developed a similar approach called ALMA (Assessing Links with Moving Averages) using also Moving Average (MA) operators. We have data about a large number of experiments developed in very different assay conditions ($b_j$) (targets, organisms, protocols, experimental measures, etc.).The use of MA operators is a potential solution; these operators were used in a time-series analysis with a similar purpose [62] in the same line of thinking as the Autoregressive Integrated Moving-Average (ARIMA) conducted by Box and Jenkins [63].

We used as inputs of the model the $MI_k$ of a given drug ($d^{th}$) and the Shannon information indices ($I^a{}_0$) for the population, i.e., the US County ($a^{th}$). This model may predict the formation of links ($L_{ac} = 1$) or not ($L_{ac} = 0$) in a complex network of AIDS pharmacoepidemiology in the US. In the present context, we can use MA of networks (drugs, proteins, organisms, etc.) nodes properties to predict the observed variable $L_{ac}(b_j)_{obs}$ in a specific sub-set of boundary conditions of assay ($b_j$). This variable quantifies the formation of links between nodes. There are two different types of nodes making up this specific network. The first node represents the US counties ($a^{th}$) and the second type of node characterizes the drugs ($d^{th}$). The value is $L_{ac}(b_j)_{obs} = 1$ when the cocktail–disease ratio = $CDR_{ac}(b_j) >$ cutoff = 0.001 and $L_{ac}(b_j)_{obs} = 0$ otherwise. In our previous work [64], we have used a drug–disease ratio $DDR_{ac}(b_j)$ for a single drug to calculate $L_{ac}(b_j)$ values, as this parameter is not applicable to drug cocktails. In the present work we have defined $CDR_{ac}(b_j) = [z_c/D_a]$. The term $z_c = (z_1 + z_2 + z_3)/3$ is the average of the z-scores $z_1$, $z_2$, $z_3$ of the biological activity for each drug ($d^{th}$) present in the cocktail assayed in the sets of conditions ($b_j$).The term $D_a$ is the AIDS prevalence rate for the county ($a^{th}$). We calculated each zeta as: $z_d(b_j) = \delta_j \cdot z_d(b_j) = \delta_j \cdot [v_d(b_j) - AVG(v(b_j))]/SD(v(b_j))$. In this operator, $v_d(b_j)$ is the value of biological activity ($EC_{50}$, $IC_{50}$, $K_i$, etc.) reported in the ChEMBL database for the drug assayed in the set of conditions. The parameter $\delta_j$ is similar to a Kronecker delta function. The parameter $\delta_j = 1$ when the $v_d(b_j)$ is directly proportional to the biological effect (e.g., $K_i$ values, Activity (%) values, etc.). Conversely, $\delta_j = -1$ when $v_d(b_j)$ is in inverse proportion to the biological effect (e.g., $EC_{50}$ values, $IC_{50}$ values, etc.). The parameter $z_d(b_j)$ is the z-score of the biological activity that depends on the AVG and SD functions. These functions are the average and standard deviation of $v_d(b_j)$ for all drugs assayed under the same conditions. The general formula for a linear model developed using the average values of $MI_k$ of the compounds used in a given drug cocktail was as follows:

$$S_{ac} = \sum_{k=1}^{k=13} e_k \cdot \left(\frac{1}{3}\sum_{d=1}^{d=3} I^d{}_k\right) + \sum_{k=1}^{k=13}\sum_{j=1}^{j=4} e_{kj} \cdot \left[\frac{1}{3}\sum_{d=1}^{d=3}\left(\Delta I^d{}_{kj}\right)\right] + e_a \cdot I^a{}_0 + e_0 \quad (12)$$

$$= \sum_{k=1}^{k=4} e_k \cdot \left(\frac{1}{3}\sum_{d=1}^{d=3} I^d{}_k\right) + \sum_{j=1}^{j=4} e_{kj} \cdot \left[\frac{1}{3}\sum_{d=1}^{d=3}\left(I^d{}_k - <I^d{}_k>_j\right)\right] + e_a \cdot I^a{}_0 + e_0$$

$$= \sum_{k=d=1}^{k=13,c,d=3} {}'e_k \cdot I^d{}_k + \sum_{k=j=d=1}^{k=13,j=4,d=3} {}'e_{kj} \cdot \left(I^d{}_k - <I^d{}_k>_j\right) + e_a \cdot I^a{}_0 + e_0$$

The reader should note that the predicted output, or dependent variable $S_{acj}$ is not a discrete variable, but a real-valued numerical score. However, the variable is directly proportional to the observed variable ($L_{ac}$). In general, $b_1$, $b_2$, $b_3$, and $b_4$ refer to different sets of boundary conditions for the assay, targets, cellular lines, organisms, experimental measures, etc. Therefore, $b_1$ = represents the experimental measures of activity for the cocktail drugs. In analogy, $b_2$ refers to the protein targets. In addition, $b_3$ refers to the organisms that expressed the targets of these compounds. Lastly, $b_4$ represents different assay protocols used to test the activity of these compounds *per se*. The inputs used to perform the model were the $MI_k$ (13 information indices) of each anti-HIV drug making up the cocktail (152,628 anti-HIV cocktails), and with these data, we calculated the average of the three molecular information indices of the drug cocktail. In addition, we used as input the average of the MA operators of the drugs that make up the cocktail. Consequently, to calculate the MA, we needed the value and the average of the drug information indices under the same conditions. Fig. 1 shows a scheme with some examples that describe

the methodology used to calculate the inputs corresponding to the drugs. The $MI_k$ of the molecules, the average values of the different boundary conditions, and the information on the US counties are in Table SM1, Table SM2 and Table SM3 of the supplementary material, respectively.

$$\Delta I^{d}_{kj} = I^{d}_{k} - \langle I^{d}_{k} \rangle_{j} \tag{13}$$

$$\langle I^{d}_{k} \rangle_{j} = \frac{1}{n_j} \sum_{d=1}^{d=n_j} I^{d}_{k} . \tag{14}$$

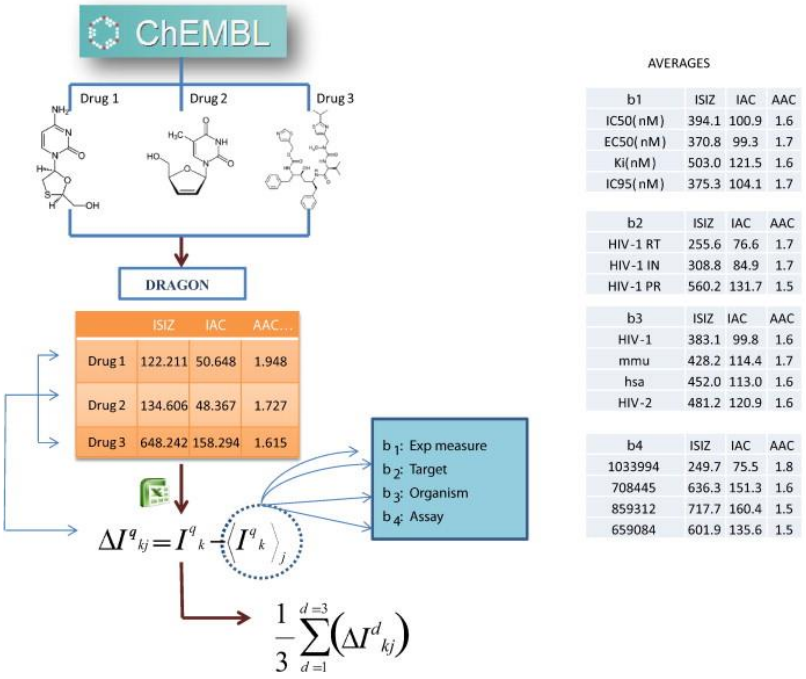


**Fig. 1**. Calculation details of the inputs of the anti-HIV drugs (left branch of Fig. 2).

### *2.2. Shannon information indices of income inequality*

We can calculate an information index to quantify the possibility of spreading/prevalence of AIDS in different US counties. Let be an initial situation in which each county has a value of AIDS prevalence rate $D_a$ at the initial time ($t_0$ = 2010). A simple information index ($I^{a}_{0}$) was used herein for income inequality in the different counties that year. This index depends on the probability $^{0}p_a$, with which the county presents certain income inequality. This probability $^{0}p_a = G_a$ was set herein. In this definition, $G_a$ is the Gini measure of income inequality in the county ($a^{th}$) of a given state in the US [65]. The class of information index selected was the Shannon entropy index [66].

$$I^{a}_{0} = - {}^{0}p_a \cdot log( {}^{0}p_a) \tag{7}$$

### *2.3. Machine learning models*

The dataset used to *train* the model includes N = 91,578 statistical cases. The dataset used to *validate* the model includes N = 30,525 statistical cases. The dataset used for *selection* consisted of 30,525 statistical cases. The cases used in the *validation set* (external validation set) were never used to train the model. Overall, training + validation + selection sets include N = 152,628 statistical cases. The amount of cases with $L_{ac}(b_j)_{obs} = 1$ was 17,381 and that with $L_{ac}(b_j)_{obs} = 0$ was 135,247. In order to seek the coefficients of the model, we can use linear or non-linear classification techniques. In this work, we used two different machine learning (ML) algorithms, a linear discriminant analysis (LDA) and artificial neural networks (ANNs). In some cases, the machine learning algorithms are carried out using as input the drug information indices and their Box–Jenkins MA operators. However, in other cases, a pre-processing of data with dimensionality reduction techniques was performed. The dimensionality reduction techniques used are of the type determined by the factor analysis. We carried out a factor analysis using two different methods to extract the principal components. The methods used were the principal components analysis (PCA) and minimum residual method (MINRES). The combination of these pre-processing algorithms with machine learning resulted in two different techniques PCA-LDA and MINRES-LDA. We never combined PCA and MINRES with ANNs. We also trained different topologies of ANNs including multilayer perceptrons (MLPs) and linear neural networks (LNNs). We also used the LDA as variable selection strategy to make a selection out of the 66 input variables, and afterwards we trained the MLP network. We summarized the previous steps of the algorithm in Fig. 2. The statistical parameters used to support the model were number of cases in training (N), and overall values of, specificity (Sp), sensitivity (Sn), and accuracy (Ac). All these methods are implemented in the STATISTICA 6.0 [67] and [68] software package.

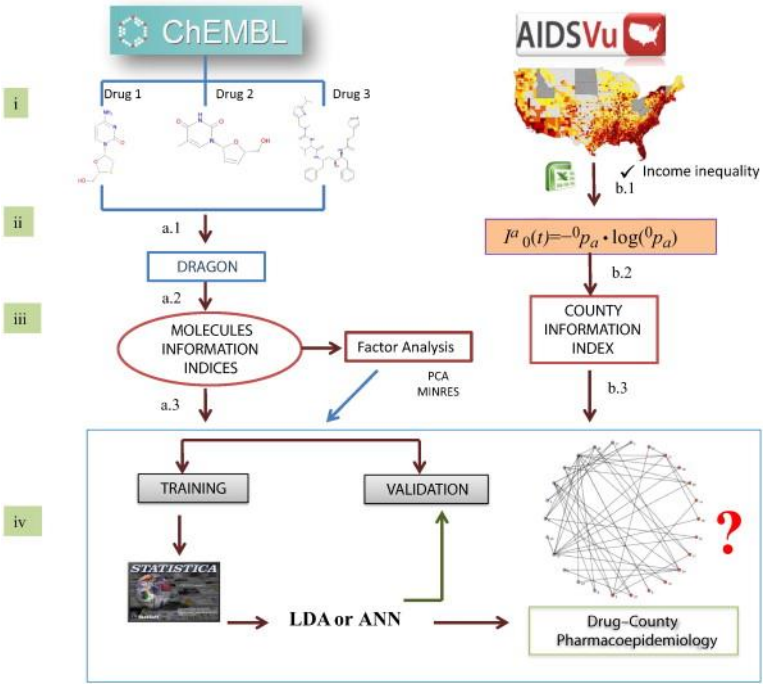


**Fig. 2**. Flowchart to construct the ML methods for the AIDS pharmacoepidemiology model in the US.

## 3. Results and discussion

### *3.1. Training and validation of the model*

In our previous work [64], we have developed a linear model using Balaban information indices for each anti-HIV drug from the ChEMBL database (unique drugs = 21,582, total data points = 43,249) and Shannon entropy based on income inequality of the US counties. The model has values of Ac, Sp, and Sn above 0.76 in training and external validation series. However, this previous model can predict outputs for only one drug each time. This previous model is unable to predict outputs for cocktails of two or three drugs. In this work, we obtained the first model useful to map the effect of cocktails of anti-HIV drugs vs. AIDS epidemiology using the present methodology based on ML-ALMA classifiers. We used 13 $MI_k$, 52 MA operators $\Delta I^d_{kj}$ for the different assay conditions for drugs and 1 $I^a_0$ operator for the US counties. First, we used LDA to seek linear models. The LDA was used as pattern classification technique, using a forward stepwise procedure as variable selection strategy. The LDA model has 23 variables, an accuracy rate of 80.39% in the training set, and an accuracy rate of 80.53% in the external validation set (see Table 2). In Table 3, we depict the description of the variables included in the LDA and the coefficients of these variables in the model.

**Table 2**. Machine learning classifiers based on $MI_k$ information indices.

| Models | Model | | Training | | Selection | | Validation | |
|---|---|---|---|---|---|---|---|---|
| | profile [a] | Observed | $L_{ac} = 0$ | $L_{ac} = 1$ | $L_{ac} = 0$ | $L_{ac} = 1$ | $L_{ac} = 0$ | $L_{ac} = 1$ |
| | | Parameter [a] | Sn | Sp | Sn | Sp | Sn | Sp |
| LDA | 66-23-1 | Predicted | 83.64 | 77.15 | – | – | 83.69 | 77.37 |
| | | $L_{ac} = 0$ | 67971 | 2356 | – | – | 45183 | 1599 |
| | | $L_{ac} = 1$ | 13292 | 7959 | – | – | 8801 | 5467 |
| | | Parameter[a] | Sn | Sp | Sn | Sp | Sn | Sp |
| MLP | 66-26-1 | Predicted | 61.31 | 60.97 | 61.47 | 62.13 | 60.77 | 59.36 |
| | | $L_{ac} = 0$ | 49830 | 4025 | 16618 | 1354 | 16381 | 1452 |
| | | $L_{ac} = 1$ | 31433 | 6290 | 10414 | 2139 | 10571 | 2121 |
| | | Parameter [a] | Sn | Sp | Sn | Sp | Sn | Sp |
| LDA-MLP | 19-10-1 | Predicted | 77.07 | 76.52 | 77.42 | 76.0 | 76.88 | 76.77 |
| | | $L_{ac} = 0$ | 62626 | 2422 | 20928 | 838 | 20722 | 830 |
| | | $L_{ac} = 1$ | 18637 | 7893 | 6104 | 2655 | 6230 | 2743 |
| | | Parameter [a] | Sn | Sp | Sn | Sp | Sn | Sp |
| LNN | 66-1 | Predicted | 82.27 | 81.31 | 82.57 | 81.93 | 82.11 | 81.52 |
| | | $L_{ac} = 0$ | 66856 | 1927 | 22322 | 631 | 22132 | 660 |
| | | $L_{ac} = 1$ | 14407 | 8388 | 4710 | 2862 | 4820 | 2913 |
| | | Parameter [a] | Sn | Sp | Sn | Sp | Sn | Sp |
| PCA-LDA | 8-7-1 | Predicted | 50.98 | 70.66 | – | – | 50.94 | 70.93 |
| | | $L_{ac} = 0$ | 41434 | 3026 | – | – | 27504 | 2054 |
| | | $L_{ac} = 1$ | 39829 | 7289 | – | – | 26480 | 5012 |
| | | Parameter [a] | Sn | Sp | Sn | Sp | Sn | Sp |
| MINRES-LDA | 8-5-1 | Predicted | 49.80 | 72.06 | – | – | 50.02 | 72.06 |
| | | $L_{ac} = 0$ | 40476 | 2882 | – | – | 27007 | 1974 |
| | | $L_{ac} = 1$ | 40787 | 7433 | – | – | 26977 | 5092 |

[a] Parameter: Sp = Specificity, Sn = Sensitivity. Columns: Observed classifications Rows: Predicted classifications

**Table 3**. Variables included in the LDA and coefficients of the model.

| Index | Function | Description |
|---|---|---|
| AAC | 74.35 | Mean information index on atomic composition |
| IDE | 1634.66 | Mean information content on the distance equality |
| IVDM | 432.67 | Mean information content on the vertex degree magnitude |
| HVcpx | − 1988.73 | Graph vertex complexity index |
| HDcpx | − 759.29 | Graph distance complexity index |
| $\Delta AAC(c_1)$ | − 97.34 | MA for AAC of drugs with the same experimental measure |
| $\Delta IDE(c_1)$ | − 1111.82 | MA for IDE of drugs with the same experimental measure |
| $\Delta IDM(c_1)$ | − 955.75 | MA for IDM of drugs with the same experimental measure |
| $\Delta IVDM(c_1)$ | 1472.82 | MA for IVDM of drugs with the same experimental measure |
| $\Delta HVcpx(c_1)$ | 1028.83 | MA for HVcpx of drugs with the same experimental measure |
| $\Delta HDcpx(c_1)$ | 3011.83 | MA for HDcpx of drugs with the same experimental measure |
| $\Delta HVcpx(c_2)$ | 0.45 | MA for HDcpx of drugs with the same protein |
| $\Delta AAC(c_3)$ | 23.96 | MA for AAC of drugs with the same organism |
| $\Delta IDE(c_3)$ | − 519.39 | MA for IDE of drugs with the same organism |
| $\Delta IDM(c_3)$ | 954.21 | MA for IDM of drugs with the same organism |
| $\Delta IVDM(c_3)$ | − 1901.33 | MA for IVDM of drugs with the same organism |
| $\Delta HDcpx(c_3)$ | 955.72 | MA for HDcpx of drugs with the same organism |
| $\Delta HDcpx(c_3)$ | − 2256.14 | MA for HDcpx of drugs with the same organism |
| $\Delta AAC(c_4)$ | − 1.46 | MA for AAC of drugs with the same assay protocol |
| $\Delta IDE(c_4)$ | − 8.43 | MA for IDE of drugs with the same assay protocol |
| $\Delta IVDE(c_4)$ | 1.28 | MA for IVDE of drugs with the same assay protocol |
| $\Delta HVcpx(c_4)$ | 9.22 | MA for HVcpx of drugs with the same assay protocol |
| $I^a_0$ | 89.14 | Information index based on the Gini coefficient |
| $e_0$ | − 15.07 | Independent term |

We also explored the possibility of training non-linear models. In so doing, we used two options implemented on the STATISTICA software: (1) neural networks, intelligent problem solver and (2) custom network designer, which are specialized tools to analyze the data and generate ANNs. These tools are available in the STATISTICA 6.0 [68] computer program. As it can be seen below in Table 4, we described the parameters of the generated neural networks. The results obtained show that the MLP trained with the 66 input variables fails to generate good predictions models, it presents an accuracy rate of 60% [67]. However, the LNN classifies correctly above 82% of the cases in the training, selection and external validation sets with 66 input variables (see Table 2). This LNN model presented values of Sn = 82.27 and Sp = 81.31 in training, and Sn = 82.11 and Sp = 81.52 in the external validation sets, but it uses 43 variables more than the LDA model. Additionally, we used the variables selected on the LDA analysis as input to train a non-linear MLP. This LDA-MLP [69] method presented values of Sp and Sn close to 77%. The LNN and the LDA-MLP networks show values of AUROC (Area Under Receiver Operating Characteristic) = 0.88 and 0.84 in training respectively, and 0.88 and 0.83 for the external validation set respectively (see Fig. 3).

**Table 4**. Parameters of neural networks.

| Details | MLP | LDA-MLP | LNN |
|---|---|---|---|
| ANN module[a] | IPS | Custom network designer | IPS |
| Training details | BP10b, iterative training | BP11741b | Pseudo-invert (PI) linear least squares optimization<br>Dot product training algorithms |
| Inputs | 66 | 19 | 66 |
| Hidden (1) | 26 | 10 | 0 |
| Hidden (2) | 0 | 0 | 0 |
| Activation function | Sigmoid | Sigmoid | Identity |
| Classification error function[b] | Entropy | Entropy | Entropy |
| Epochs | – | 10000 | – |
| Learning rate | – | 0.01 | – |
| Threshold[c] | 1.0 | 1.0 | 1.0 |
| Criteria to select retained networks | Best performance | Best performance | Balance performance against diversity |
| Stopping conditions | Target error | Target error | Target error |
| Training target error | 0.0 | 0.0 | 0.0 |
| Selection target error | 0.0 | 0.0 | 0.0 |

[a] Module for ANN analysis implemented on the STATISTICA software. IPS = intelligent problem solver. BP = back-propagation.
[b] Classification tasks ANN uses, the so called cross-entropy error, to train the neural networks, but the selection criteria for evaluating the best network is actually based on the classification rate, which can be easily interpreted as compared to the entropy-based error function.
[c] This is available only if the dependent variable is nominal with two values. A single threshold (accept = reject) is determined to minimize expected loss. A loss coefficient of 1.0 indicates that the two classes are equally important.

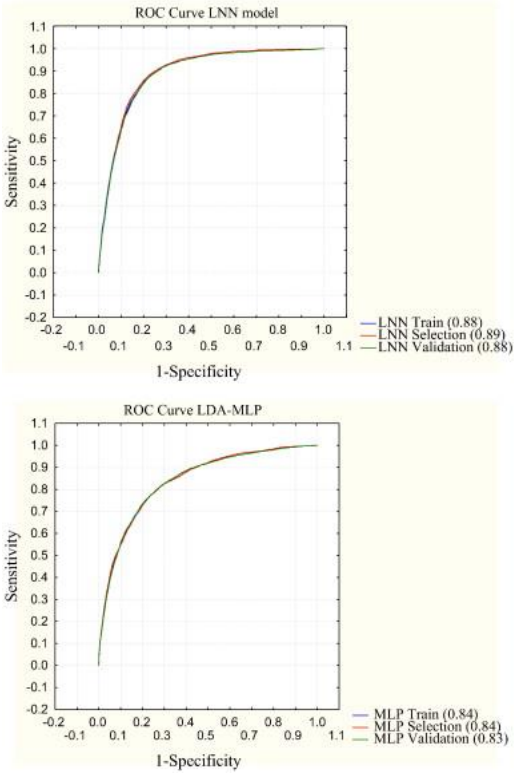


**Fig. 3.** AUROC curve values for the ANNs**.**

We also carried out a PCA and MINRES of data. The PCA and MINRES for the bio-molecular factors were conducted with 65 input variables. The analyses showed seven eigenvalues for the bio-molecular factors that account for the 90% with PCA and 80.55% with MINRES of the information. These analyses include mainly factors such as drug structure, experimental measure, organism, assay, and target (see Fig. 4). Table 5 depicts the eigenvalues obtained with these techniques. The eigenvalues generated give an indication of the amount of information carried by each component. Additional information about the extraction of the principal components with PCA and MINRES is in Tables SM4 and SM5 of the supplementary material. Next, with the extraction of the principal components (seven factors) and with the $I^a_0$, we carried out a PCA-LDA and a MINRES-LDA separately, but they failed to generate good prediction models, since they presented values of specificity and sensitivity close to 50% (values for a random classifier) (see Table 2).

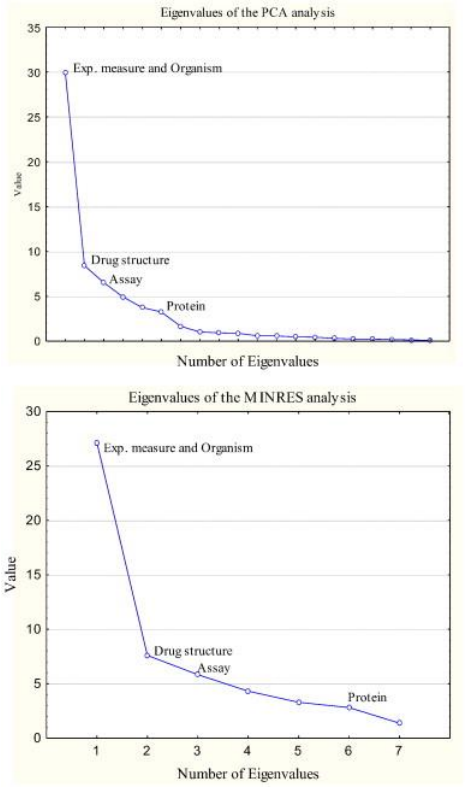


**Fig. 4.** Plot of bio-molecular eigenvalues for PCA and MINRES.

**Table 5**. Eigenvalues of the factor PCA analysis.

| Extraction method | Principal factors | Eigenvalue | % Total variance | Cumulative eigenvalue | Cumulative % |
|---|---|---|---|---|---|
| PCA | 1 | 29.97708 | 46.11858 | 29.97708 | 46.11858 |
| | 2 | 8.41775 | 12.95038 | 38.39482 | 59.06896 |
| | 3 | 6.53269 | 10.05028 | 44.92751 | 69.11924 |
| | 4 | 4.95234 | 7.61899 | 49.87985 | 76.73823 |
| | 5 | 3.75442 | 5.77603 | 53.63427 | 82.51426 |
| | 6 | 3.25460 | 5.00707 | 56.88887 | 87.52134 |
| | 7 | 1.66932 | 2.56819 | 58.55819 | 90.08952 |
| MINRES | 1 | 27.12600 | 41.73230 | 27.12600 | 41.73230 |
| | 2 | 7.60284 | 11.69667 | 34.72883 | 53.42897 |
| | 3 | 5.83629 | 8.97891 | 40.56512 | 62.40788 |
| | 4 | 4.31739 | 6.64214 | 44.88251 | 69.05002 |
| | 5 | 3.29302 | 5.06618 | 48.17553 | 74.11619 |
| | 6 | 2.81078 | 4.32427 | 50.98630 | 78.44047 |
| | 7 | 1.37669 | 2.11798 | 52.36299 | 80.55845 |

Consequently, the LDA model is better here with Ac, Sn, and Sp rates of 80%, similar to the LDA-MLP performance. Considering that both models LDA-MLP and LDA have similar performance and a similar number of inputs, we should consider the simpler LDA (23 variables and 0 hidden neurons) model as a good model. Because the LDA-MLP needs 10 hidden neurons to increase performance and even its performance is slightly lower compared to the LDA model. All in all, the LDA was the best model in terms of accuracy and simplicity.

### *3.2. Construction of complex networks*

In our previous work [64], we have also used a linear-ALMA model to create a complex network. The network had two classes of nodes (counties vs. drugs). The drug nodes contained information about the chemical structure, as well as, all the assay conditions (target protein, organism, assay protocol, experimental measure). On the other hand, the county nodes contained the information about the income inequality. However, because of the type of model used, these complex networks are unable to represent drug cocktails. In the present paper, we propose to use the predicted values ($L_{ac}(b_j)_{pred} = 1$) of the LDA-ALMA classifier to generate different sub-networks. These sub-networks are maps of the AIDS prevalence with respect to the preclinical activity of anti-HIV drug cocktails in each state of the US at county level. This type of sub-network may have different classes of nodes. There are three main classes: counties $a^{th}$ of the state, the $c^{th}$ drug cocktails, and the $d^{th}$ drugs (chemical compounds) making up the cocktail. We may also include other classes of nodes for the different boundary conditions of assay $b_j$. In doing so, we may include the following classes of nodes: experimental measures ($b_1$), protein targets ($b_2$), organisms of assay ($b_3$), or assay protocols ($b_4$). In these sub-networks we draw arcs connecting the nodes of the different classes when $L_{ac}(b_j)_{pred} = 1$ or do not draw these arcs when the model predict $L_{ac}(b_j)_{pred} = 0$. Fig. 5 shows the previous type of sub-network of AIDS prevalence vs. anti-HIV drug preclinical activity for the state of California. The sub-network has three types of nodes: anti-HIV drugs (blue), cocktails (red) and US counties. It is important to understand that here $L_{ac}(b_j)_{pred} = 1$ expresses the existence of a sub-graph that connects several nodes of all classes by means of various arcs and there is no single arc which connects two nodes. For instance, let us see a simple sub-network including only nodes for drugs, cocktails, and counties. In this case, when $L_{ac}(b_j)_{pred} = 1$ we connect each node of the compounds making up the cocktail with the node ($c^{th}$) that represents this cocktail. Consequently, $L_{ac}(b_j)_{pred} = 1$ expresses the existence of the sub-graph $(d^1 \rightarrow c_1)(d^2 \rightarrow c_1)d^3 \rightarrow c_1 \rightarrow a_1$ for all the drugs in the cocktail, see Fig. 6.

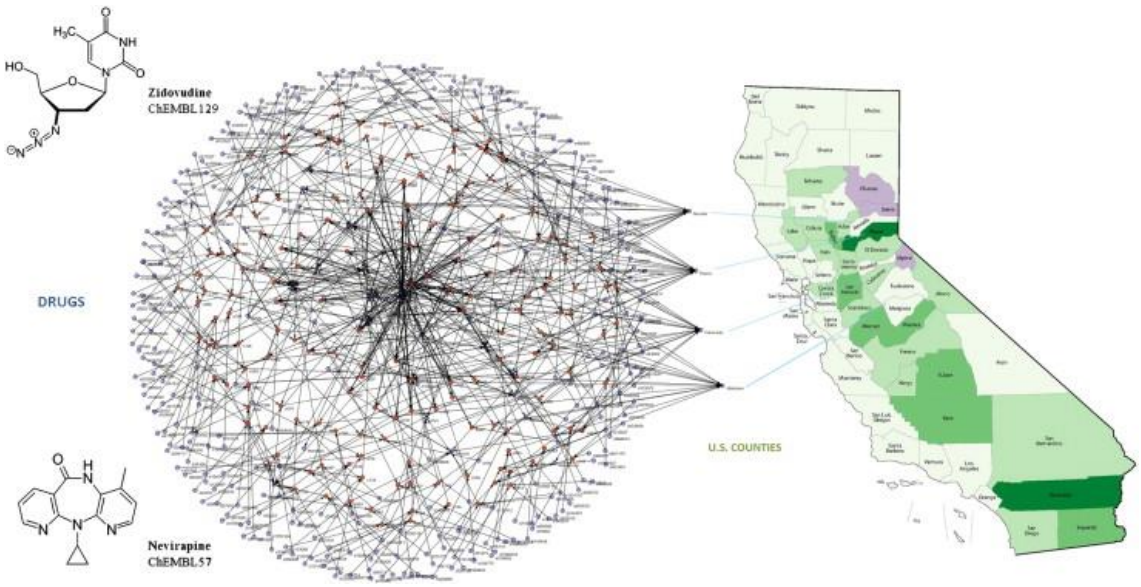


**Fig. 5**. Sub-network of anti-HIV drug cocktails vs. AIDS prevalence for the US state of California (CA).

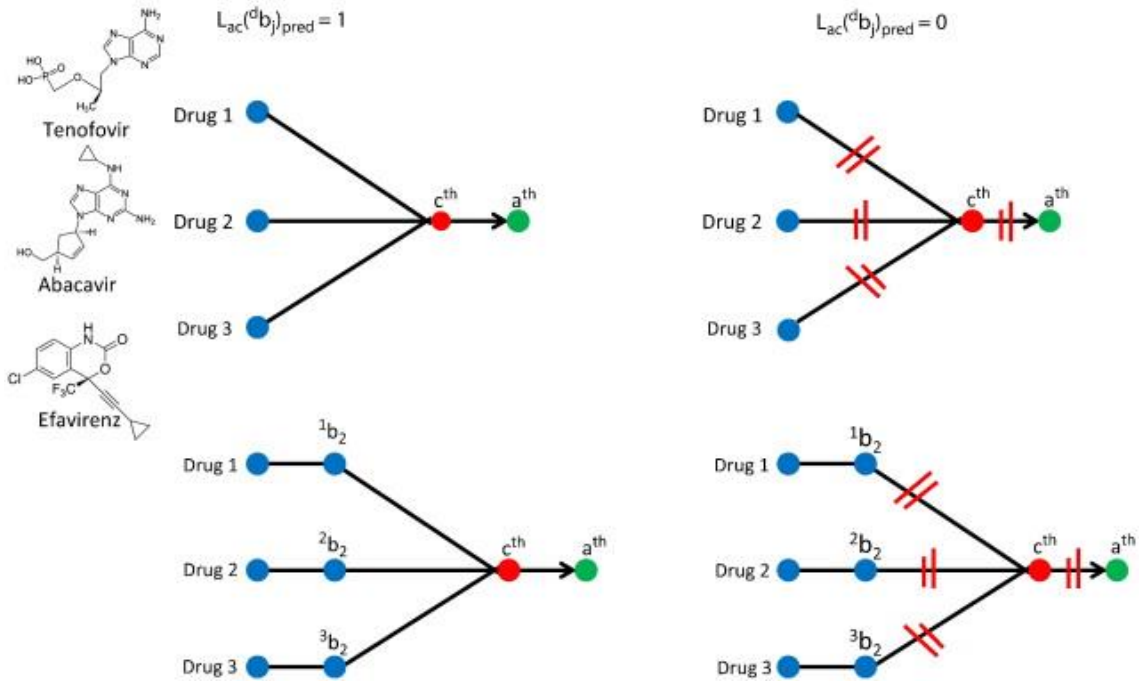


**Fig. 6**. Sub-network node connection $L_{ac}(^{d}b_j)_{pred} = 1$ and non-connection $L_{ac}(^{d}b_j)_{pred} = 0$. The $b_2$ represents the drug target, $c^{th}$ represents the drug cocktails, $a^{th}$ represents the US counties.

In a more complicated example including also the boundary condition of assay $b_2$ = target, for each drug, the situation is similar. $L_{ac}(^{d}b_j)_{pred} = 1$ expresses the existence of the sub-graph $(d^1 \rightarrow b_2)(d^2 \rightarrow b_2)d^3 \rightarrow b_2 \rightarrow c_1 \rightarrow a_1$ for all the drugs in the cocktail, see also Fig. 6. Additionally, Table 6 shows the LDA prediction for some cases of drug cocktails vs. US counties. We included some examples of antiretroviral cocktails with observed $L_{ac}(b_j)_{obs}$ and predicted $L_{ac}(b_j)_{pred}$ effects over AIDS prevalence in several counties of the same state in the US. Table SM6 of the supplementary material shows the results predicted with the LDA model for all the cases in the training and external validation sets.

**Table 6**. LDA model prediction of some cases of drug cocktails vs. different counties.

| $L_{ac}(b_j)_{Obs}$ | $L_{ac}(b_j)_{Pred}$ | c-Level | Drug name or ChEMBL ID | | | $ID_i$ | $ID_{ii}$ | $ID_{iii}$ | State, county |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0.916 | Zalcitabine | Nevirapine | Ritonavir | 38347 | 38201 | 32404 | KS, Montgomery |
| 0 | 0 | 0.795 | Nevirapine | Delavirdine | Indinavir | 38207 | 38322 | 32336 | PA, Westmoreland |
| 0 | 0 | 0.925 | Zidovudine | Nevirapine | Darunavir | 38307 | 38265 | 32427 | KY, Boyd |
| 0 | 0 | 0.89 | Nevirapine | Delavirdine | Amprenavir | 38280 | 38337 | 32362 | PA, Northampton |
| 0 | 0 | 0.913 | Delavirdine | Nevirapine | Ritonavir | 38316 | 38285 | 32392 | KS, Riley |
| 0 | 0 | 0.828 | Delavirdine | Nevirapine | Indinavir | 38326 | 38211 | 32341 | PA, Montgomery |
| 0 | 0 | 0.58 | Lamivudine | Stavudine | Ritonavir | 38310 | 38350 | 32386 | KS, Pottawatomie |
| 0 | 0 | 0.928 | 593 | 57 | 115 | 38325 | 38276 | 32339 | TX, Milam |
| 0 | 0 | 0.918 | 129 | 57 | 116 | 38305 | 38280 | 32375 | TX, Kaufman |
| 0 | 0 | 0.912 | 129 | 57 | 729 | 38308 | 38236 | 32275 | GA, Berrien |
| 0 | 0 | 0.833 | 160 | 593 | 115 | 38311 | 38334 | 32304 | GA, Chattooga |
| 0 | 0 | 0.872 | 57 | 991 | 114 | 38249 | 38348 | 32288 | GA, Columbia |
| 0 | 0 | 0.86 | 57 | 853 | 114 | 38220 | 38346 | 32257 | VA, Albemarle |
| 0 | 0 | 0.887 | 798 | 57 | 115 | 38343 | 38192 | 32302 | VA, Nelson |
| 0 | 0 | 0.885 | 57 | 129 | 114 | 38241 | 38288 | 32268 | TX, Lee |
| 1 | 1 | 0.911 | 129 | 593 | 114 | 38297 | 38332 | 32277 | WY, Uinta |
| 1 | 1 | 0.96 | 57 | 798 | 115 | 38260 | 38345 | 32301 | TX, Leon |
| 1 | 1 | 0.995 | 57 | 798 | 114 | 38204 | 38345 | 32250 | GA, Lamar |
| 1 | 1 | 0.956 | 57 | 593 | 116 | 38207 | 38339 | 32362 | GA, Cherokee |
| 1 | 1 | 0.939 | 593 | 129 | 1323 | 38340 | 38307 | 32427 | GA, Whitfield |
| 1 | 1 | 0.883 | 625 | 57 | 114 | 38342 | 38245 | 32276 | OR, Lincoln |
| 1 | 1 | 0.983 | 593 | 57 | 114 | 38341 | 38235 | 32270 | GA, Franklin |
| 1 | 1 | 0.779 | 991 | 593 | 115 | 38351 | 38320 | 32297 | AL, Randolph |
| 1 | 1 | 0.983 | 57 | 593 | 116 | 38218 | 38340 | 32356 | IN, Floyd |
| 1 | 1 | 0.998 | 593 | 57 | 163 | 38341 | 38256 | 32382 | AR, Franklin |

ChEMBL IDs are the identifiers of a drug in ChEMBL database. Some ChEMBL IDs used in this table are Nevirapine = 57, Delavirdine = 593, Atazanavir = 1163, AZT Triphosphate = 798, Amprenavir = 116, Zidovudine = 129, Indinavir = 115, Stavudine = 991, Saquinavir = 114, Ritonavir = 163. $ID_i$, $ID_{ii}$, and $ID_{iii}$, are the identifiers used in this work for the set of assay conditions for each drug of the cocktail according to supplementary material Table SM4 (these are not ChEMBL IDs).

## 4. Conclusions

This work presents the development of a model called LDA-ALMA to map networks of cocktails of anti-HIV drugs vs. AIDS epidemiology in the US counties. We used as inputs molecular information indices of drugs and Shannon entropy based on county-level income inequality. Machine learning techniques, such as LDA and ANNs, were used. The LDA classifier presented good values of sensitivity/specificity (80%) compared to the MLP, with values close to 60%. Therefore, this LDA-ALMA model may be useful to design effective antiretroviral cocktails to treat HIV in the US counties with a given AIDS prevalence rate.

## Acknowledgments

R.O.M. acknowledges financial support of FPI fellowship associated to research project (AGL2011-30563-C03-01) funded by MECD (Spanish Ministry of Education, Culture and Sport).

## References


[1]. G.L. Colombo, A. Castagna, S. Di Matteo, L. Galli, G. Bruno, A. Poli, S. Salpietro, A. Carbone, A. Lazzarin. Cost analysis of initial highly active antiretroviral therapy regimens for managing human immunodeficiency virus-infected patients according to clinical practice in a hospital setting. Ther. Clin. Risk Manag., 10 (2014), pp. 9–15.

[2]. I. Usach, V. Melis, J.E. Peris. Non-nucleoside reverse transcriptase inhibitors: a review on pharmacokinetics, pharmacodynamics, safety and tolerability. J. Int. AIDS Soc., 16 (2013), pp. 1–14.

[3]. W.S. Hu, S.H. Hughes. HIV-1 reverse transcription. Cold Spring Harb. Perspect. Med., 2 (2012) (pii: a006882).

[4]. X. Qiu, Z.P. Liu. Recent developments of peptidomimetic HIV-1 protease inhibitors. Curr. Med. Chem., 18 (2011), pp. 4513–4537.

[5]. H.C. Castro, P.A. Abreu, R.B. Geraldo, R.C. Martins, R. dos Santos, N.I. Loureiro, L.M. Cabral, C.R. Rodrigues. Looking at the proteases from a simple perspective. J. Mol. Recognit., 24 (2011), pp. 165–181.

[6]. G. Alkhatib. The biology of CCR5 and CXCR4. Curr. Opin. HIV AIDS, 4 (2009), pp. 96–103.

[7]. C. Blanpain, F. Libert, G. Vassart, M. Parmentier. CCR5 and HIV infection. Recept. Channels, 8 (2002), pp. 19–31.

[8]. J.H. Tan, J.P. Ludeman, J. Wedderburn, M. Canals, P. Hall, S.J. Butler, D. Taleski, A. Christopoulos, M.J. Hickey, R.J. Payne, M.J. Stone. Tyrosine sulfation of chemokine receptor CCR2 enhances interactions with both monomeric and dimeric forms of the chemokine monocyte chemoattractant protein-1 (MCP-1). J. Biol. Chem., 288 (2013), pp. 10024–10034.

[9]. K. Qian, S.L. Morris-Natschke, K.H. Lee. HIV entry inhibitors and their potential in HIV therapy. Med. Res. Rev., 29 (2009), pp. 369–393.

[10]. T.J. Wilkin, R.M. Gulick. CCR5 antagonism in HIV infection: current concepts and future opportunities. Annu. Rev. Med., 63 (2012), pp. 81–93.

[11]. C.F. Perno. The discovery and development of HIV therapy: the new challenges. Ann. Ist. Super. Sanita, 47 (2011), pp. 41–43.

[12]. M.P. de Bethune. Non-nucleoside reverse transcriptase inhibitors (NNRTIs), their discovery, development, and use in the treatment of HIV-1 infection: a review of the last 20 years (1989-2009). Antivir. Res., 85 (2010), pp. 75–90.

[13]. C. Hicks, R.M. Gulick. Raltegravir: the first HIV type 1 integrase inhibitor. Clin. Infect. Dis., 48 (2009), pp. 931–939.

[14]. W.G. Powderly. Integrase inhibitors in the treatment of HIV-1 infection. J. Antimicrob. Chemother., 65 (2010), pp. 2485–2488.

[15]. J.L. Adams, B.N. Greener, A.D. Kashuba. Pharmacology of HIV integrase inhibitors. Curr. Opin. HIV AIDS, 7 (2012), pp. 390–400.

[16]. J.J. Eron Jr.. HIV-1 protease inhibitors. Clin. Infect. Dis., 30 (Suppl. 2) (2000), pp. S160–S170.

[17]. E.J. Arts, D.J. Hazuda. HIV-1 antiretroviral drug therapy. Cold Spring Harb. Perspect. Med., 2 (2012), p. a007161.

[18]. I. Chougrani, D. Luton, S. Matheron, L. Mandelbrot, E. Azria. Safety of protease inhibitors in HIV-infected pregnant women. HIV AIDS (Auckl), 5 (2013), pp. 253–262.

[19]. J. King, M. McCall, A. Cannella, M.A. Markiewicz, A. James, C.B. Hood, E.P. Acosta. A randomized crossover study to determine relative bioequivalence of tenofovir, emtricitabine, and efavirenz (Atripla) fixed-dose combination tablet compared with a compounded oral liquid formulation derived from the tablet. J. Acquir. Immune Defic. Syndr., 56 (2011), pp. e130–e132.

[20]. R. O'Neal. Rilpivirine and complera: new first-line treatment options. BETA, 23 (2011), pp. 14–18.

[21]. C.M. Perry. Elvitegravir/cobicistat/emtricitabine/tenofovir disoproxil fumarate single-tablet regimen (Stribild((R))): A review of its use in the management of HIV-1 infection in adults. Drugs, 74 (2014), pp. 75–97.

[22]. S.D. Portsmouth, C.J. Scott. The renaissance of fixed dose combinations: Combivir. Ther. Clin. Risk Manag., 3 (2007), pp. 579–583.

[23]. B. Coutinho, R. Prasad. Emtricitabine/tenofovir (Truvada) for HIV prophylaxis. Am. Fam. Physician, 88 (2013), pp. 535–540.

[24]. E. Lopez Aspiroz, D. Santos Buelga, S. Cabrera Figueroa, R.M. Lopez Galera, E. Ribera Pascuet, A. Dominguez-Gil Hurle, M.J. Garcia Sanchez. Population pharmacokinetics of lopinavir/ritonavir (Kaletra) in HIV-infected patients. Ther. Drug Monit., 33 (2011), pp. 573–582.

[25]. M. Shey, E.J. Kongnyuy, J. Shang, C.S. Wiysonge. A combination drug of abacavir-lamivudine-zidovudine (Trizivir) for treating HIV infection and AIDS. Cochrane Database Syst. Rev. (2009), p. CD005481.

[26]. P.E. Sax, C. Tierney, A.C. Collier, M.A. Fischl, K. Mollan, L. Peeples, C. Godfrey, N.C. Jahed, L. Myers, D. Katzenstein, A. Farajallah, J.F. Rooney, B. Ha, W.C. Woodward, S.L. Koletar, V.A. Johnson, P.J. Geiseler, E.S. Daar. Abacavir-lamivudine versus tenofovir-emtricitabine for initial HIV-1 therapy. N. Engl. J. Med., 361 (2009), pp. 2230–2240.

[27]. R. Guha. On exploring structure-activity relationships. Methods Mol. Biol., 993 (2013), pp. 81–94.

[28]. R. Burbidge, M. Trotter, B. Buxton, S. Holden. Drug design by machine learning: support vector machines for pharmaceutical data analysis. Comput. Chem., 26 (2001), pp. 5–14.

[29]. J. Patel. Science of the science, drug discovery and artificial neural networks. Curr. Drug Discov. Technol., 10 (2013), pp. 2–7.

[30]. A. Speck-Planche, V.V. Kleandrova, F. Luan, M.N. Cordeiro. A ligand-based approach for the in silico discovery of multi-target inhibitors for proteins associated with HIV infection. Mol. Biosyst., 8 (2012), pp. 2188–2196.
[31]. P. Gupta, A. Sharma, P. Garg, N. Roy. QSAR study of curcumine derivatives as HIV-1 integrase inhibitors. Curr. Comput. Aided Drug Des., 9 (2013), pp. 141–150.
[32]. R. Muthukumaran, B. Sangeetha, R. Amutha, P.P. Mathur. Development of anti-HIV activity models of lysine sulfonamide analogs: a QSAR perspective. Curr. Comput. Aided Drug Des., 8 (2012), pp. 70–82.
[33]. A.K. Debnath. Application of 3D-QSAR techniques in anti-HIV-1 drug design–an overview. Curr. Pharm. Des., 11 (2005), pp. 3091–3110.
[34]. U. Debnath, S. Verma, S. Jain, S.B. Katti, Y.S. Prabhakar. Pyridones as NNRTIs against HIV-1 mutants: 3D-QSAR and protein informatics. J. Comput. Aided Mol. Des., 27 (2013), pp. 637–654.
[35]. X.H. Sun, J.Q. Guan, J.J. Tan, C. Liu, C.X. Wang. 3D-QSAR studies of quinoline ring derivatives as HIV-1 integrase inhibitors. SAR QSAR Environ. Res., 23 (2012), pp. 683–703.
[36]. K. Swiderek, S. Marti, V. Moliner. Theoretical studies of HIV-1 reverse transcriptase inhibition. Phys. Chem. Chem. Phys., 14 (2012), pp. 12614–12624.
[37]. Y. Marrero-Ponce. Linear indices of the “molecular pseudograph's atom adjacency matrix”: definition, significance-interpretation, and application to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors. J. Chem. Inf. Comput. Sci., 44 (2004), pp. 2010–2026.
[38]. Y. Hu, J. Bajorath. Molecular scaffolds with high propensity to form multi-target activity cliffs. J. Chem. Inf. Model., 50 (2010), pp. 500–510.
[39]. D. Erhan, P.J. L'Heureux, S.Y. Yue, Y. Bengio. Collaborative filtering on a family of biological targets. J. Chem. Inf. Model., 46 (2006), pp. 626–635.
[40]. V. Namasivayam, Y. Hu, J. Balfer, J. Bajorath. Classification of compounds with distinct or overlapping multi-target activities and diverse molecular mechanisms using emerging chemical patterns. J. Chem. Inf. Model., 53 (2013), pp. 1272–1281.
[41]. M. Cruz-Monteagudo, M.N. Cordeiro, E. Tejera, E.R. Dominguez, F. Borges. Desirability-based multi-objective QSAR in drug discovery. Mini-Rev. Med. Chem., 12 (2012), pp. 920–935.
[42]. A. Machado, E. Tejera, M. Cruz-Monteagudo, I. Rebelo. Application of desirability-based multi(bi)-objective optimization in the design of selective arylpiperazine derivates for the 5-HT1A serotonin receptor. Eur. J. Med. Chem., 44 (2009), pp. 5045–5054.
[43]. L. Saiz-Urra, A.J. Bustillo Perez, M. Cruz-Monteagudo, C. Pinedo-Rivilla, J. Aleu, R. Hernandez-Galan, I.G. Collado. Global antifungal profile optimization of chlorophenyl derivatives against Botrytis cinerea and Colletotrichum gloeosporioides. J. Agric. Food Chem., 57 (2009), pp. 4838–4843.
[44]. M. Cruz-Monteagudo, F. Borges, M.N. Cordeiro, J.L. Cagide Fajin, C. Morell, R.M. Ruiz, Y. Canizares-Carmenate, E.R. Dominguez. Desirability-based methods of multiobjective optimization and ranking for global QSAR studies. Filtering safe and potent drug candidates from combinatorial libraries. J. Comb. Chem., 10 (2008), pp. 897–913.
[45]. C.A. Nicolaou, N. Brown, C.S. Pattichis. Molecular optimization using computational multi-objective methods. Curr. Opin. Drug Discov. Devel., 10 (2007), pp. 316–324.
[46]. K. Heikamp, J. Bajorath. Large-scale similarity search profiling of ChEMBL compound data sets. J. Chem. Inf. Model., 51 (2011), pp. 1831–1839.
[47]. A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J.P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res., 40 (2012), pp. D1100–D1107.
[48]. A.P. Bento, A. Gaulton, A. Hersey, L.J. Bellis, J. Chambers, M. Davies, F.A. Kruger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, J.P. Overington. The ChEMBL bioactivity database: an update. Nucleic Acids Res., 42 (2013), pp. D1083–D1090.
[49]. C.E. Shannon. A mathematical theory of communication. Bell Syst. Tech. J., 27 (1948), pp. 379–423.
[50]. J.S. Gonzalez-Garcia, J. Diaz. Information theory and the ethylene genetic network. Plant Signal. Behav., 6 (2011), pp. 1483–1498.
[51]. C. Waltermann, E. Klipp. Information theory based approaches to cellular signaling. Biochim. Biophys. Acta, 1810 (2011), pp. 924–932.
[52]. R. Todeschini, V. Consonni. Handbook of Molecular Descriptors. Wiley-VCH Verlag GmbH, Weinheim, Germany (2000).
[53]. A.T. Balaban, T.S. Balaban. New vertex invariants and topological indices of chemical graphs based on information on distances. J. Math. Chem., 8 (1991), pp. 383–397.
[54]. O. Ivanciuc, T.S. Balaban, A.T. Balaban. Chemical graphs with degenerate topological indices based on information on distances. J. Math. Chem., 14 (1993), pp. 21–33.
[55]. V.R. Magnuson, D.K. Harriss, S.C. Basak. Studies in Physical and Theoretical Chemistry. R.B. King (Ed.)Elsevier, Amsterdam (The Netherlands) (1983), pp. 178–191.
[56]. C. Shannon, W. Weaver. The Mathematical Theory of Communication. University of Illinois Press, Urbana United States (1949).
[57]. S.H. Bertz. The first general index of molecular complexity. J. Am. Chem. Soc., 103 (1981), pp. 3599–3601.
[58]. S.M. Dancoff, H. Quastler. Essays on the Use of Information Theory in Biology. University of Illinois, Urbana (1953).
[59]. D. Bonchev, N. Trinajstic. On topological characterization of molecular branching. Int. J. Quantum Chem. Quantum Chem. Symp., 12 (1978), pp. 293–303.

[60]. C. Raychaudhury, S.K. Ray, J.J. Ghosh, A.B. Roy, S.C. Basak. Discrimination of isomeric structures using information theoretic topological indices. J. Comput. Chem., 5 (1984), pp. 581–588.
[61]. G. Klopman, C. Raychaudhury, R.V. Henderson. A new approach to structure-activity using distance information content of graph vertices: a study with phenylalkylamines. Math. Comput. Model., 11 (1988), pp. 635–640.
[62]. E. Tenorio-Borroto, X. Garcia-Mera, C.G. Penuelas-Rivas, J.C. Vasquez-Chagoyan, F.J. Prado-Prado, N. Castanedo, H. Gonzalez-Diaz. Entropy model for multiplex drug-target interaction endpoints of drug immunotoxicity. Curr. Top. Med. Chem., 13 (2013), pp. 1636–1649.
[63]. G.E.P. Box, G.M. Jenkins. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco, California (1970).
[64]. H. González-Díaz, D.M. Herrera-Ibatá, A. Duardo-Sanchez, C.R. Munteanu, R.A. Orbegozo-Medina, A. Pazos. Model of the multiscale complex network of AIDS prevalence in US at county level vs. preclinical activity of anti-HIV drugs based on information indices of molecular graphs and social networks. J. Chem. Inf. Model., 54 (2014), pp. 744–755.
[65]. R. Pabayo, I. Kawachi, S.E. Gilman. Income inequality among American states and the incidence of major depression. J. Epidemiol. Community Health, 68 (2014), pp. 110–115.
[66]. P. Riera-Fernandez, C.R. Munteanu, M. Escobar, F. Prado-Prado, R. Martin-Romalde, D. Pereira, K. Villalba, A. Duardo-Sanchez, H. Gonzalez-Diaz. New Markov-Shannon entropy models to assess connectivity quality in complex networks: from molecular to cellular pathway, parasite-host, neural, industry, and legal-social networks. J. Theor. Biol., 293 (2012), pp. 174–188.
[67]. T. Hill, P. Lewicki. STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining. StatSoft, Tulsa (2006).
[68]. STATISTICA. version 6.0. StatSoft Inc., Tulsa, Oklahoma (2001).
[69]. F. Rosenblatt. Principles of Neurodynamics; Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington (1962).