UNIVERSIDADE DA CORUÑA

Facultad de Informática

Departamento de Tecnologías de la Información
y las Comunicaciones

Modelos Multi-escala de Inteligencia Artificial para Diseño Quimio-Informático y Fármaco-Epidemiológico de Terapias anti-VIH en Condados de Estados Unidos

Tesis Doctoral

Doctoranda

DIANA MARÍA HERRERA IBATÁ

Directores
Prof. Dr. Alejandro Pazos Sierra
Prof. Dr. Humberto González Díaz
Prof. Dr. Cristian Robert Munteanu

A Coruña, 2015

Prof. Dr. D. Alejandro Pazos Sierra, Catedrático del Departamento de Tecnologías de la Información y las Comunicaciones, Facultad de Informática, Universidade da Coruña, UDC.

Prof. Dr. D. Humberto González Díaz, Prof. Investigador Ikerbasque del Departamento de Química Orgánica II, Facultad de Ciencia y Tecnología, Universidad del País Vasco, UPV/EHU.

Prof. Dr. D. Cristian Robert Munteanu, Profesor Contratado Doctor del Departamento de Tecnologías de la Información y las Comunicaciones, Facultad de Informática, Universidade da Coruña, UDC.

CERTIFICAN:

Que la memoria titulada "**Modelos Multi-escala de Inteligencia Artificial para Diseño Quimio-Informático y Fármaco-Epidemiológico de Terapias anti-VIH en Condados de Estados Unidos**" presentada por **D. DIANA MARÍA HERRERA IBATÁ**, ha sido realizada bajo nuestra dirección, en el Departamento de Tecnologías de la Información y las Comunicaciones, Facultad de Informática, Universidade da Coruña, UDC y en el Departamento de Química Orgánica II, Facultad de Ciencia y Tecnología, Universidad del País Vasco UPV/EHU.

Considerando que el trabajo constituye tema de Tesis Doctoral, se autoriza su presentación en la Universidade da Coruña.

Y para que conste, se expide el presente certificado en A Coruña, a Junio de dos mil quince.

Prof. Dr. Alejandro Pazos Sierra

Prof. Dr. Humberto González Díaz

Prof. Dr. Cristian Robert Munteanu

A mi familia

Agradecimientos

En primer lugar deseo expresar mi agradecimiento a los directores de esta tesis los Profesores Dr. Alejandro Pazos Sierra, Dr. Humberto González Díaz, y Dr. Cristian Robert Munteanu, por toda la ayuda y dedicación que han brindado a este trabajo. Asímismo, quiero agradecer a sus colaboradores de la Universidade da Coruña, especialmente al grupo de Redes de Neuronas Artificiales y Sistemas Adaptativos RNASA-IMEDIR.

Deseo expresar también todo mi agradecimiento a mis padres y a mi hermana por su apoyo incondicional a lo largo de este trayecto y por su confianza.

A mis amigos y a todas aquellas personas que han estado presentes y han dedicado parte de su tiempo al desarrollo de este trabajo, por su colaboración y paciencia.

¡Muchas gracias a todos!

Resumen

Los métodos que relacionan la estructura química con la actividad biológica se conocen como “relaciones cuantitativas estructura-actividad” (en adelante QSAR). Es fundamental entender y cuantificar la relación entre la estructura y la actividad biológica de los potenciales fármacos para realizar su estudio eficiente. Este tipo de estudio consiste en correlacionar, por medio de descriptores moleculares, distintas propiedades químicas o fisicoquímicas de las moléculas en cuestión con valores de actividad biológica. Actualmente, el desarrollo de medicamentos más seguros y efectivos en el tratamiento de enfermedades como el SIDA es un objetivo que requiere del esfuerzo de un elevado número de especialistas en diferentes campos de la Ciencia, y donde el azar ha tenido un gran protagonismo. Sin embargo, parece razonable pensar que nunca se obtendrán medicamentos eficaces y seguros con sólo acudir al azar.

Para ser más eficientes en el desarrollo de nuevos fármacos, la investigación en el tratamiento de las enfermedades requiere poseer mecanismos predictivos de algunas actividades. Los modelos basados en “redes de neuronas artificiales” (en adelante RRNNAA) son un ejemplo de modelos teóricos de predicción, ampliamente utilizados en muchas áreas de la Ciencia, como medicina, química, bioquímica…, así como también en el desarrollo de medicamentos. En esto último, son muy útiles para la predicción de propiedades de los potenciales fármacos. Las RRNNAA se aproximan a la forma de operar que usa el cerebro humano, con habilidad para abordar con éxito los datos, las informaciones y los conocimientos naturales, o del mundo real, que están afectados por lo que se conoce como la “maldición de la cuádruple I”, por ser datos: inciertos, inconsistentes, incompletos e imprecisos. Esta particularidad hace que sean difíciles de gestionar adecuadamente por las técnicas computacionales convencionales, haciendo precisa la utilización de técnicas de Inteligencia Artificial, como son las ya citadas RRNNAA. La mayor ventaja de estos modelos inteligentes de predicción es que permiten evitar costes innecesarios producidos por desarrollos de nuevos compuestos con potencialidad terapéutica que resultarán estériles.

Por lo tanto, el objetivo principal de la tesis aquí presentada es el desarrollo, con técnicas de inteligencia artificial, de una metodología “quimioinformática multi-escala” que permita relacionar cuantitativamente datos químicos y pre-clínicos con datos epidemiológicos, para llevar a cabo predicciones “fármaco-epidemiológicas”, teniendo en cuenta la imposibilidad práctica y legal de obtener datos experimentales, en la fase IV del proceso de desarrollo de nuevos compuestos.

Resumo

Os métodos que relacionan a estrutura química coa actividade biolóxica son chamados “relacións cuantitativas estrutura – actividade” (en adiante QSAR). É esencial para entender e cuantificar a relación entre a estrutura e a actividade biolóxica dos potenciais fármacos para realizar o seu estudio eficiente. Este tipo de estudo consiste en correlacionar, a través de descritores moleculares, distintas propiedades químicas ou fisicoquímicas de las moleculas en cuestión, con valores de actividade biolóxica. Actualmente, o desenvolvemento de medicamentos máis seguros e efectivos no tratamento de enfermidades como o SIDA é un obxectivo que require do esforzo de un gran número de especialistas en diferentes campos da ciencia, e onde o azar tivo un gran protagonismo. Nembergantes, parece razoable pensar que nunca se obterían medicamentos eficaces e seguros con só acudir ao azar.

Para ser máis eficaces no desenvolvemento de novos farmacos, a investigación para o tratamento de enfermidades require mecanismos preditivos de algunhas actividades. Os modelos baseados en redes neurais artificiais (en adiante RRNNAA) son un exemplo de modelos teóricos de predición amplamente utilizado en moitas áreas da ciencia, como medicina, química, bioquímica..., así como tamén no desenvolvemento de medicamentos. Nesto último, son moi útiles para a predición de propiedades dos potenciais medicamentos. As RRNNAA achegánse ao xeito de funcionar do cerebro humano, coa capacidade para abordar con éxito los datos, las informaciones y los conocimientos naturales, o del mundo real, que están afectados polo que se coñece como a “maldición da cuadrúple I”, por ser dados: incertos, inconsistentes, incompletos e imprecisos. Esta particularidade fai que sexan díficiles de xestionar axeitadamente coas técnicas computacionais convencionais, facendo preciso o uso de técnicas de Intelixencia Artificial, como son as xa citadas RRNNAA. A maior vantaxe destes modelos preditivos intelixentes é que permiten evitar custos innecesarios producidos polos desenvolvementos de novos compostos con potencial terapéutico que resultaran esteriles.

Polo tanto o obxectivo principal da tese aquí presentada é o desenvolvemento, con tecnicas de intelixencia artificial dunha metodoloxía “quimioinformática multi-escala” que permita relacionar cuantitativamente datos químicos e pre-clínicos con datos epidemiolóxicos, para levar a cabo predicións fármaco-epidemiolóxicas, tendo en conta a imposibilidade práctica e legal de obter datos experimentais na fase IV do proceso de desenvolvemento de novos compostos.

Abstract

The methods relating chemical structure to biological activity are called “Quantitative Structure Activity Relationships” (QSAR). It is essential to understand and quantify the relationships between the structure and biological activity of potential drugs to develop an efficient study on them. This kind of study consists of the correlation of the molecular descriptors based on several chemical or physicochemical properties with biological activity. Currently, the development of safer and more effective drugs in the treatment of diseases such as AIDS is a goal that requires a joint effort of a large number of specialists from different fields of science, and where chance also has a major role. However, it seems reasonable that no effective and safe drugs will be obtained based on chance only.

To be more efficient in developing new drugs, the research for the treatment of diseases requires predictive mechanisms of some biological activities. The models based on "Artificial Neural Networks" (ANNs) are an example of theoretical prediction models, widely used in many areas of science such as Medicine, Chemistry, Biochemistry, etc. as well as in Drug Development. In the latter, they are very useful for predicting properties of potential drugs. ANNs approach the modus operandi used by the human brain, being able to successfully manage data, information and natural knowledge, or from the real world, which are affected by the so-called "curse of the fourfold I", dealing with information which is uncertain, inconsistent, incomplete and inaccurate. This feature makes it difficult to properly manage by conventional computational techniques, making the use of Artificial Intelligence (AI) techniques necessary, such as the above-mentioned ANNs. The most important advantage of these intelligent prediction models is the fact that they avoid unnecessary production costs associated with the development of new compounds with therapeutic potential which proved to be inactive.

Therefore, the main objective of the thesis is the development of a chemoinformatics multi-scale methodology using artificial intelligence techniques to quantitatively relate chemical and pre-clinical data with epidemiological data, with the aim of performing "drug - epidemiological" predictions, taking into account the practical and legal impossibility of obtaining experimental data in Phase IV of the development process of new compounds.

ÍNDICE DE CONTENIDO

ABREVIATURAS

3D-QSAR	Relación estructura-actividad en tercera dimensión
a-ésimo	Numeración condados de Estados Unidos
CADD	Diseño de fármacos asistido por ordenador
CCR2	C-C quimiocina receptora tipo 2
CCR3	C-C quimiocina receptora tipo 3
CCR4	C-C quimiocina receptora tipo 4
CXCR4	C-X-C quimiocina receptora tipo 4
CCR5	C-C quimiocina receptora tipo 5
CDC	Center for Disease Control and Prevention
CoMFA	Comparative molecular field analysis
CoMSIA	Comparative molecular similarity indices analysis
G_a	Coeficiente de Gini de desigualdad de los ingresos
GP160	Glicoproteína de Envoltura
HAART	Highly active antiretroviral therapy
INNTR	Inhibidores no nucleósidos de la transcriptasa reversa
INTR	Inhibidores nucleósidos de la transcriptasa reversa
IN-VIH1	Integrasa VIH-1
I^q_k	Índices de Balaban
LDA	Análisis discriminante lineal
LNN	Red de neuronas artificiales lineal
Mm	especie de ratón Mus musculus
mt-QSAR	Modelos QSAR múltiple tarea
N(t)RTI	Inhibidor de la transcriptasa inversa análogos de nucleótidos
PLS	Mínimos cuadrados parciales
PR-VIH1	Proteasa VIH-1
QSAR	Relación cuantitativa estructura-actividad
RRNNAA	Redes de neuronas artificiales
RT-VIH1	Transcriptasa reversa VIH-1
SIDA	Síndrome de inmunodeficiencia humana
SVM	Maquinas de vectores de soporte
TARGA	Terapia antirretroviral de gran actividad

Tat VIH-1	Proteína Tat VIH-1
TIs	Índices topológicos
VIH	Virus de inmunodeficiencia humana

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

El virus de inmunodeficiencia humana (VIH) es un retrovirus que pertenece a la familia de los Lentivirus que causan el síndrome de inmunodeficiencia humana (SIDA) (Lindemann, Steffen, y Pohlmann, 2013). Los retrovirus pueden usar su ARN y el ADN del huésped para producir ADN viral, y son conocidos por sus largos periodos de incubación. Existen dos tipos de VIH: VIH-1 y VIH-2. El más predominante es el VIH-1, además el VIH-2 parece progresar más lentamente (Rinaldo, 2013). En el SIDA, el sistema inmune se ve severamente afectado y el organismo es susceptible de contraer a una gran variedad de infecciones por gérmenes como: bacterias, parásitos, virus, etc., que pueden resultar fatales en las personas afectadas (Noorizadeh, Sajjadifar, y Farmany, 2013).

El SIDA se considera una importante enfermedad que constituye una amenaza para la vida y la pandemia del VIH continúa propagándose. Desde el comienzo de la epidemia, más de 60 millones de personas se han infectado con el VIH y superan los 35 millones los fallecidos a causa de la enfermedad. Desde el primer caso de SIDA reportado por los Estados Unidos, en 1981, se han realizado enormes progresos en la prevención y tratamiento del VIH/SIDA, especialmente en el desarrollo de una terapia antirretroviral que ha demostrado salvar la vida a millones de personas ("HIV surveillance--United States, 1981-2008," 2011; Moss, 2013; Piot y Quinn, 2013). Los antirretrovirales pueden brindar una mejor calidad de vida a un portador del VIH y aumentan sus posibilidades de supervivencia. Por lo tanto, el descubrimiento y desarrollo de fármacos innovadores y altamente potentes contra la enfermedad siguen siendo imprescindibles, aunque la erradicación continua mostrándose como una meta difícil de lograr debido a un alto nivel de persistencia viral en los sujetos tratados.

Asimismo, el VIH ha llevado a muchos investigadores en todo el mundo a descubrir nuevos compuestos y/o dianas moleculares o celulares útiles contra la enfermedad. Actualmente, existen diferentes bases de datos con resultados experimentales de las interacciones de los compuestos antirretrovirales con sus respectivas dianas. ChEMBL (https://www.ebi.ac.uk/chembl/) (Bento et al., 2013a; Gaulton et al., 2012; Heikamp y Bajorath, 2011) es una de las más grandes bases de datos abierta de bioactividad a gran escala, extraída en gran parte de forma manual desde la literatura química y médica. La información relativa a los compuestos ensayados (incluyendo sus estructuras), los ensayos biológicos o físico-químicos realizados sobre éstos y las dianas de estos ensayos se registran en una forma estructurada. Existen, en la actualidad, más de 1,3 millones de estructuras de compuestos distintos y 12 millones de puntos de datos de bioactividad. Los datos se asignan a más de 9000 dianas, de las cuales 2827 son dianas de proteínas humanas (Bento et al., 2013b). Específicamente, ChEMBL contiene más de 43.000 resultados para los ensayos de compuestos anti-VIH. Igualmente, la base de datos contiene un conjunto de datos generales de compuestos potenciales contra el VIH de más de 8.000 puntos finales de ensayo de multiplexación (resultados de múltiples ensayos).

En este contexto, las diferentes técnicas de diseño de fármacos asistido por ordenador (CADD), útiles para predecir el comportamiento de los compuestos contra el VIH, pueden desempeñar un papel importante en la reducción del número de estudios preclínicos y clínicos. Por ejemplo, se podrían usar modelos quimio-informáticos que enlazan la estructura química de los fármacos con su actividad biológica. De hecho, hay muchos informes de modelos quimio-informáticos, útiles para predecir la actividad de fármacos anti-VIH en ensayos preclínicos. En principio, se podrían actualizar estos modelos para predecir la actividad anti-VIH de los fármacos no sólo en la detección preclínica, sino también en los estudios clínicos y en el ámbito de la farmacoepidemiología. Tal modelo puede llegar a ser una herramienta muy útil para la industria farmacéutica con el objetivo de reducir el tiempo y recursos empleados en los ensayos clínicos. Igualmente, los modelos deben ser idealmente útiles para las entidades públicas encargadas de la aplicación de las políticas de salud en la fase IV de desarrollo de fármacos. Sin embargo, no hay informes de modelos útiles para predecir el rendimiento de los medicamentos contra el VIH, tanto en estudios preclínicos como de farmacoepidemiología en grandes poblaciones, sin necesidad de llevar a cabo estudios clínicos. Esto significa que, con el fin de desarrollar este tipo de modelos computacionales, se necesitan procesar distintos tipos de datos de entrada procedentes de muchos niveles de organización de la materia, por lo que se trata de datos provenientes de fuentes heterogéneas que serán difícilmente abordables con técnicas de computación convencional para su gestión y análisis. Por ello, se recurren a técnicas y procedimientos de Inteligencia Artificial, como son las RRNNAA, que permitirán gestionar y analizar datos del mundo real y procesar conocimiento natural, el cual está afectado por lo que se conoce como la maldición de la “cuádruple I”, puesto que son: inciertos, imprecisos, inconsistentes e incompletos.

Por un lado, se tiene que introducir la información sobre los compuestos anti-VIH, que incluye al menos la estructura química del compuesto (i) e información del ensayo preclínico, tales como dianas biológicas (nivel ii), organismos (nivel iii), o protocolos de ensayo (iv nivel). Además, se necesita incorporar descriptores de la estructura de la población (nivel v) que cuantifican los factores epidemiológicos, sociales y económicos que afectan a la población seleccionada para el estudio. Por consiguiente, como las poblaciones en la sociedad actual no son sistemas cerrados, también se debería cuantificar el efecto de la interacción de la población en análisis con otras poblaciones que pueden influir en el estudio farmacoepidemiológico (vi nivel). Por último, un modelo quimioinformático-farmacoepidemiológico útil debe ser por definición multi-nivel, ya que se espera unir tanto la estructura molecular como la estructura de la población.

1.1 OBJETIVOS

- Desarrollar con técnicas de inteligencia artificial, una metodología “quimioinformática multi-escala” que permita relacionar cuantitativamente datos químicos y pre-clínicos con datos epidemiológicos, para llevar a cabo predicciones “fármaco-epidemiológicas”, teniendo en cuenta la imposibilidad práctica y legal de obtener datos experimentales, en la fase IV del proceso de desarrollo de nuevos compuestos.

- Describir el flujo de trabajo para una metodología general que combine métodos quimioinformáticos y farmacoepidemiológicos usando las bases de datos ChEMBL del Instituto Europeo de Biología Molecular y AIDSVu de la Universidad de Emory de Estados Unidos.

- Aplicar esta metodología al desarrollo de modelos teóricos de la efectividad de compuestos contra el VIH en distintos condados de Estados Unidos con una prevalencia de SIDA dada. Es decir, modelos capaces de vincular los cambios en la prevalencia del SIDA en una población con los cambios en la actividad biológica de los compuestos activos, detectados en ensayos preclínicos.

- Extender esta metodología al diseño computacional de terapias farmacológicas basadas en combinaciones de fármacos “cócteles de fármacos”.

- Generación de modelos predictivos basados en RRNNAA utilizando como entrada los índices de información de redes sociales y grafos moleculares.

- Predecir redes complejas de prevalencia del SIDA en los condados de Estados Unidos, teniendo en cuenta los determinantes sociales y la relación estructura-actividad de los compuestos anti-VIH en ensayos preclínicos.

1.2 HIPÓTESIS DE TRABAJO

Los índices de información moleculares y sociales son variables de entrada útiles en modelos de RRNNAA, capaces de vincular los cambios en la prevalencia del SIDA en una población dada (con valores determinados de parámetros socio-económicos) con los cambios en la relación estructura-actividad biológica de los compuestos anti-VIH, detectados en ensayos preclínicos realizados en el marco de un conjunto de condiciones de ensayo.

2. FUNDAMENTOS TEÓRICOS

2.1 Problemática del conocimiento en el mundo real

La finalidad de los sistemas inteligentes es desarrollar su tarea no sólo en problemas de laboratorio, sino con casos que utilicen conocimiento natural, que se deriva y utiliza entornos del mundo real. Por ello, en muchos casos este conocimiento padece lo que se viene a denominar la “maldición de la cuádruple I”, por ser datos: inciertos, incompletos, imprecisos e inconsistentes.

La incertidumbre se produce porque los datos en los que se basan los razonamientos para tomar las decisiones son inciertos; es decir, son verdaderos o falsos, pero no hay forma de saberlo. También se puede producir incertidumbre porque las reglas de inferencia se obtienen directamente de la experiencia o son heurísticas y, por lo tanto, no son completamente fiables. Asimismo, gran parte del razonamiento del mundo real se realiza bajo condiciones de incertidumbre, como en diagnosis los falsos (+) o (-) de las pruebas, síntomas y signos.

La incompletitud se produce debido a que no todo lo que pertenece al caso en cuestión puede observarse o recopilarse, o porque no existen suficientes recursos para efectuar todas las deducciones potencialmente interesantes, o aún, porque las teorías del mundo sólo son aproximaciones.

La inconsistencia surge en parte por el manejo de la incompletitud. Por ejemplo, al pasar por alto una omisión se puede provocar una contradicción. Además, la información puede proceder de fuentes contradictorias, resultando en un conocimiento contradictorio. También puede producir inconsistencia el razonamiento no monótono o no monotónico. O sea, cuando el entorno que influye sobre los conocimientos no es constante en el tiempo y cambian algunas de las variables a tener en cuenta para encontrar la solución más adecuada.

Por último, la imprecisión se produce porque las expresiones son borrosas, vagas o, aún, difusas si en ellas no se define exactamente al menos alguna de las variables que la engloban. Un ejemplo es cuando se recurre a la utilización de expresiones con cuantificadores para expresar cantidades tales como: algún, mucho, la mayoría, etc.

Una de las formas más eficientes de tratar los problemas asociados al conocimiento natural es por medio de las RRNNAA, en las cuales no existe el problema de la incertidumbre, ya que la representación del conocimiento no la hace el experto, sino que es el propio sistema quién la hace a partir de los ejemplos que se le proporcionen. Igualmente, se han mostrado útiles para trabajar con información incompleta e inconsistente gracias a su sistema de representación del conocimiento distribuido y a trabajar adecuadamente incluso en ambientes con "ruido" y, o, con datos incompletos. Además, se han demostrado hábiles para soslayar el problema de la imprecisión siempre que se disponga de un adecuado y completo conjunto de entrenamiento. También existen paradigmas de RRNNAA que son adecuados para funcionar con valores difusos, utilizando funciones difusas en lugar de datos concretos, tanto en las funciones que manejan los valores transmitidos en y entre sus elementos, como en los pesos de las conexiones que unen a sus elementos de procesamiento.

2.2 Sistemas adaptativos inteligentes

Una de las principales características de los sistemas adaptativos inteligentes es que tienen capacidad de aprendizaje automático (en adelante, AA). El campo del AA se aplica en construir programas computacionales que son capaces de mejorar con las experiencias pasadas o con ejemplos, y de una forma automática, para resolver un problema dado. El AA es una rama de la Inteligencia Artificial (IA) que tiene por objetivo desarrollar técnicas mediante las cuales los ordenadores puedan aprender y ayudar en la solución de problemas complejos. Los algoritmos de aprendizaje más utilizados pueden clasificarse, de acuerdo a la cantidad de ejemplos disponibles para su entrenamiento, en dos categorías dependiendo de si existe una intervención externa en el proceso de entrenamiento de las RRNNAA: algoritmos supervisados y no supervisados. La aplicación de uno u otro algoritmo depende de las características del asunto a resolver, pues cada uno de ellos puede ser útil en determinadas circunstancias. Las principales ventajas de las técnicas de AA son:

I. Algunas tareas no pueden ser bien definidas excepto utilizando ejemplos; es decir, se podría ser capaz de especificar pares de entradas/salidas, pero no una relación concisa entre las entradas y las salidas deseadas. Por lo tanto, se buscan sistemas capaces de ajustar su estructura interna para producir salidas correctas para un gran número de muestras de entrada y así restringir adecuadamente su función salida/entrada para aproximar la relación entre ellas, la cual que se encuentra representada implícitamente en los ejemplos.

II. Es posible que, en grandes cantidades de datos, estén ocultas importantes relaciones y correlaciones que no se sea capaz de adivinar. Los métodos de aprendizaje automático son muy útiles para extraer esas relaciones.

III. Estas técnicas de AA se han mostrado muy eficientes en problemas de: clasificación, agrupamiento, regresión, diagnóstico y predicción.

2.3 Introducción a los modelos RRNNAA

Las RRNNAA están compuestas por un elevado número de elementos de procesamiento o *neuronas artificiales* organizadas en capas y muy interconectadas entre sí que trabajan al unísono para resolver un problema específico. En las RRNNAA cada neurona recibe un conjunto de entradas (x_1, x_2,…x_D), procesa la información y devuelve una única salida. Las conexiones entre los elementos de procesamiento simulan las conexiones interneuronales del cerebro y, al igual que estas, pueden establecerse con mayor o menor intensidad (Palma-Méndez y Marín-Morales, 2008). En el caso de las RRNNAA esta intensidad la determinan los pesos sinápticos. De este modo, cada entrada x_i de una neurona se encuentra afectada por un peso w_{ij} que identifica la unión entre la neurona receptora x_i y la neurona emisora x_j.

La primera neurona formal fue concebida por McCulloch y Pitts de la Universidad de Chicago en 1943, los autores propusieron uno de los primeros modelos matemáticos de una neurona, del que se basan las redes neuronales actuales (Figura 1) (McCulloch y Pitts, 1943). Se trata de un modelo binario cuyo estado es 1 (activo) o 0 (inactivo).

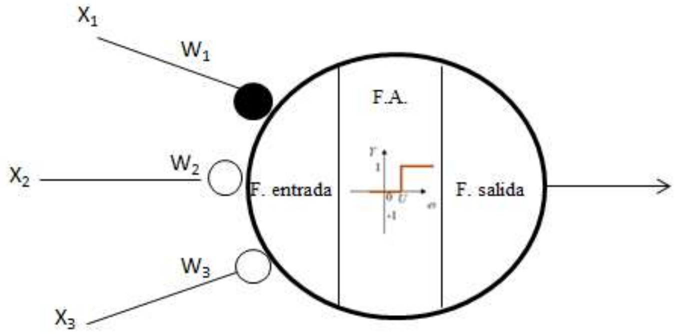


Figura 1. Neurona artificial de McCulloch y Pitts

Algunas de las ventajas de estas redes es que son capaces de aprender de la experiencia, generalizar a partir de ejemplos previos y abstraer las características principales de una serie de ejemplos. Asimismo, las RRNNAA pueden aprender; es decir, adquirir el conocimiento de una cosa por medio de un conjunto completo de ejemplos que hacen que puedan cambiar su comportamiento en función del entorno. Se les muestra un conjunto de entradas y ellas mismas se ajustan para producir una salida consistente y adaptada a la entrada recibida.

La distribución de neuronas dentro de la red se realiza formando niveles o capas, con un número determinado de dichas neuronas en cada una de ellas. A partir de su situación dentro de la red, se pueden distinguir tres tipos de capas:

I. *De entrada*: es la capa que recibe directamente la información proveniente del entorno y la emite a otro elemento del sistema.

II. *Ocultas*: son internas a la red y no tienen contacto directo con el entorno exterior. Solo reciben información de otros elementos del sistema y solo emiten información a otros elementos del sistema y nunca reciben o emiten información al entorno. El número de niveles ocultos puede estar entre cero y el número que se desee. Las neuronas de las capas ocultas pueden estar interconectadas de distintas maneras, lo que determina, junto con su número, las distintas topologías de RRNNAA.

III. *De salida*: reciben información de otros elementos del sistema y emiten información de la red hacia el entorno.

La neurona artificial fue diseñada para "emular" las características del funcionamiento básico de la neurona biológica, esto es, reproduce su función y su estructura. En esencia, se aplica un conjunto de entradas a una neurona, cada una de las cuales representa una salida de otra neurona o un estímulo del entorno. Cada entrada se multiplica por su "peso" o ponderación correspondiente análoga al grado de intensidad de conexión de la sinapsis. Todas las entradas ponderadas se suman y se determina el nivel de excitación o activación de la neurona. Una representación vectorial del funcionamiento básico de una neurona artificial se indica según la siguiente expresión de la ecuación

$$\mathbf{NET} = X * Wij$$

Siendo NET obtenida a partir de la ponderación del sumatorio de los valores del vector de entrada X ponderado por los valores del vector de pesos de las conexiones W_{ij}. Posteriormente la señal NET es procesada por una función de activación *f* para producir la señal de salida de la neurona OUT. Una neurona biológica puede estar activa o inactiva; es decir, que tiene un "*estado de activación*". Las neuronas artificiales también tienen diferentes estados de activación; algunas de ellas solamente dos, pero otras pueden tomar cualquier valor dentro de un conjunto determinado. La *función de activación* calcula el estado de actividad de una neurona; transformando en su caso la entrada global en un valor o estado de activación, cuyo rango normalmente va de [0 a 1] o de [–1 a 1]. La función *f* puede ser una función umbral como la propuesta por McCulloch y Pitts, una función lineal o una función no lineal que simula con mayor exactitud las características de transferencia no lineales de las neuronas biológicas.

La función de salida determina qué valor se transfiere a las neuronas que reciben la señal por ella emitida. Si la función de activación está por debajo de un valor determinado, la neurona no se activa y ninguna salida se pasa a la neurona subsiguiente. Normalmente, no se permite cualquier valor como una entrada para las neuronas, por lo tanto, los valores de salida usualmente están comprendidos en el rango [0, 1] o en el [-1, 1], pero también pueden ser binarios.

2.3.1 Elaboración de la información

El proceso de elaboración de la información recibida depende de las distintas características, tanto estructurales como funcionales, de la red. Existen modelos muy diversos de RRNNAA en los cuales se siguen filosofías de diseño, reglas de aprendizaje y funciones de construcción de las respuestas muy distintas. Una primera clasificación se hace en función del recorrido que sigue la información dentro de la red entre sus elementos y, así, se distinguen: redes alimentadas hacia adelante y redes con retroalimentación.

2.3.1.1 Redes alimentadas hacia adelante (Feedforward)

Las redes alimentadas hacia adelante son aquellas en las que, como su nombre indica, la información se mueve en un único sentido, desde la capa de entrada hacia la capa de salida de la RNA. Estas redes están clásicamente organizadas en capas. Cada capa agrupa a un conjunto de neuronas que reciben sinapsis de las neuronas del entorno o de la capa anterior y emiten salidas hacia neuronas de la capa siguiente o hacia el entorno. Entre las neuronas de una misma capa no hay sinapsis. En este tipo de redes existe al menos una capa de entrada, formada por las neuronas que reciben las señales de entrada a la red y una capa de salida, formada por una o mas neuronas que emiten la respuesta de la red al exterior. Entre la capa de entrada y la de salida existen una o más capas intermedias. En redes así construidas es evidente que la información sólo puede moverse en un sentido: desde la capa de entrada hasta la capa de salida, atravesando todas y cada una de las capas intermedias una sóla vez. El hecho de que no haya conexión entre las neuronas de una misma capa hace que no haya tiempos de espera en los que las neuronas estén interactuando unas sobre otras hasta que toda la capa adquiera un estado estable. Se trata, por tanto, de redes rápidas en sus cálculos.

En este tipo de red no existen interconexiones entre capas más allá de la conexión directa hacia adelante para propagar la información. No hay rutas de retroalimentación para desempeñar la función de “memoria” de la red, o de condicionamiento del funcionamiento previo.

2.3.1.2 Redes con retroalimentacion total o parcial (Feedback)

En este tipo de redes los elementos pueden enviar estímulos a neuronas de capas anteriores, de su propia capa o a ellos mismos, por lo que desaparece el concepto de agrupamiento de las neuronas en capas. Cada neurona puede estar conectada a todas las demás; de este modo, cuando se recibe información de entrada a la red, cada neurona tendrá que calcular y recalcular su estado varias veces, hasta que todas las neuronas de la red alcancen un estado estable. Un estado estable es aquel en el que no ocurren cambios en la salida de ninguna neurona. No habiendo cambios en las salidas, las entradas de todas las neuronas serán también constantes, por lo que no tendrán que modificar su estado de activación ni su respuesta, manteniéndose así un estado global estable.

2.3.2 Tipos de neuronas

A pesar de la gran variedad de funciones de transferencia y de activación que se pueden implementar, hay un criterio de clasificación de todas ellas que las divide en dos grandes grupos: funciones lineales y no lineales, siendo esta la característica más definitoria del comportamiento de una neurona.

2.3.2.1 Neuronas lineales

Una neurona es lineal cuando su salida es linealmente dependiente de sus entradas; es decir, proporcional a las funciones de transferencia y de activación. Estas neuronas tienen funciones de activación y de transferencia lineales por lo que la composición de ambas funciones da lugar a otra función lineal que regirá la elaboración de las respuestas en función de las entradas. Esto conlleva ciertos problemas como la falta de persistencia en las respuestas, de modo que cambios muy pequeños en las entradas puede producir fluctuaciones bastante grandes en las respuestas, o la falta de adecuación simultánea.

2.3.2.2 Neuronas no lineales

En estas neuronas, o bien la función de activación, o bien la función de transferencia (o ambas) son funciones no lineales, dando lugar a que la respuesta de la neurona no sea lineal respecto de sus entradas. Este tipo de neuronas va a producir respuestas acotadas, atenuando los problemas de fluctuación y la falta de adecuación a señales pequeñas y grandes. Como ejemplo de funciones no lineales se pueden destacar como más habituales en los modelos clásicos la función umbral, la función sigmoide y la función hiperbólica tangente. Si se trata de una neurona tipo Umbral (función de activación o transferencia tipo umbral), su salida es discreta, despareciendo así los problemas de fluctuación de la respuesta y de no adecuación a señales grandes o pequeñas.

Las neuronas con función sigmoide o hiperbólica tangente o, en general, con función de transferencia con límites de saturación superior e inferior, permiten que las respuestas solo varíen con elevada sensibilidad ante los cambios de las entradas cuando la neurona está a medio camino entre los dos niveles de saturación. Pero cuando la neurona alcanza uno de los niveles de saturación, pequeñas fluctuaciones en las señales de entrada, no alterarán o alterarán mínimamente la respuesta, que permanecerá estable en dicho nivel de saturación. Este efecto es muy deseable, pues garantiza una cierta persistencia de la respuesta de las neuronas. Durante la emisión de la respuesta, la neurona amortigua, hasta cierto punto, los efectos de nuevas señales de entrada, que no fueran muy diferentes de las que estuviera tratando.

2.3.3 Mecanismos de aprendizaje más utilizados

2.3.3.1 Aprendizaje supervisado

El aprendizaje supervisado se caracteriza porque el proceso de aprendizaje se realiza mediante un entrenamiento controlado por un agente externo (supervisor o maestro) que supervisa la respuesta que genera la red a partir de una entrada determinada. El supervisor controla la salida de la red y, en caso de que ésta no se acerque lo suficiente a la deseada, se procederá a modificar los pesos de las conexiones, con el fin de conseguir que la salida obtenida se aproxime a la salida deseada, en mayor medida que la anterior.

2.3.3.2 Aprendizaje no supervisado.

Las redes con aprendizaje no supervisado (también conocido como autosupervisado) no requieren influencia externa para ajustar los pesos de las conexiones entre sus neuronas. La red no recibe ninguna información por parte del entorno que le indique si la salida generada en respuesta a una determinada entrada es o no correcta. Estas redes deben encontrar las características, regularidades, correlaciones o categorías que se puedan establecer entre los datos que se presenten en su entrada.

2.3.4 Modelos RRNNAA-QSAR

Métodos computacionales como el AA se están aplicando cada vez más en el descubrimiento y evaluación de fármacos para la predicción de compuestos con actividad farmacológica, farmacodinamia específica y propiedades ADMET (absorción, distribución, metabolismo, excreción y toxicidad). Recientemente, las técnicas de aprendizaje automático, tales como las RRNNAA, máquinas de soporte vectorial (Support Vector Machines) y programación genética se han explorado para la predicción de inhibidores, antagonistas, agonistas, bloqueantes, activadores y sustratos de proteínas relacionadas con objetivos terapéuticos específicos. Estos métodos son particularmente útiles para el cribado de bibliotecas de compuestos de diversas estructuras químicas, datos "ruidosos" y para complementar los métodos QSAR.

Las RRNNAA han sido usadas para un amplio rango de aplicaciones quimioinformáticas. Entre los estudios buscando predecir bioactividad se encuentran los trabajos sobre agonistas estrogénicos (Li et al., 2006), y el estudio de esteroides (So y Karplus, 1997). También, las RRNNAA estaban entre un número de métodos identificando posibles inhibidores kinasa (Briem y Gunther, 2005). Las RRNNAA son, a menudo, usadas para predecir propiedades toxicológicas, farmacológicas y fisicoquímicas tales como toxicidad acuática, aclaramiento, pKa, solubilidad, etc.

Un modelo teórico fue desarrollado para discriminar entre compuestos activos e inactivos contra el VIH con cuatro diferentes mecanismos de acción para las drogas activas. El modelo fue construido utilizando una RNA probabilística y una base de datos de 2720 compuestos. El modelo no solo clasificó correctamente una serie de compuestos orgánicos muy heterogéneos, sino que también discriminó entre químicos activos/no activos muy similares que pertenecen a la misma familia de compuestos. Más específicamente, el modelo reconoció el 96.02% de compuestos no activos, el 94.24% de compuestos activos que inhiben la transcriptasa inversa, el 97.24% de inhibidores de proteasa, y el 90.32% de inhibidores de la integrasa. Los resultados indican que esta aproximación puede representar una herramienta útil para modelar grandes bases de datos en QSAR con aplicaciones en química médica (Vilar, Santana, y Uriarte, 2006).

Asimismo, se han realizado métodos no lineales para construir una relación cuantitativa entre la actividad biológica anti-HIV de ligandos HEPT y drogas retrovirales y sus descriptores calculados (Noorizadeh et al., 2013). Los resultados obtenidos por la RNA Levenberg–Marquardt (L-M ANN) fueron comparados con los resultados obtenidos por el algoritmo genético-Kernel cuadrados mínimos parciales (GA-KPLS). Los resultados demostraron que el primer modelo (R^2=calibración, predicción, y test fueron 0.916, 0.894, 0.868) es más poderoso que el segundo (R^2 entrenamiento y test 0.861, 0.748). Este modelo podía predecir exactamente la actividad biológica anti-VIH de compuestos que no existían en el procedimiento de modelado.

Un estudio QSAR anti-VIH no lineal se ha investigado en una serie de derivados HEPT, en total 79, que actúan como inhibidores no nucleósidos de la transcriptasa reversa (INNTR). Se usaron como datos iniciales 20 descriptores moleculares para generar un nuevo modelo QSAR. Este estudio QSAR ha sido llevado a cabo por una RNA de tres capas utilizando descriptores moleculares conocidos, por ser responsables de la actividad anti-VIH-1. La utilidad del modelo y la no linealidad de la relación entre los descriptores moleculares y la actividad anti-VIH-1 han sido claramente demostradas. El análisis de RRNNAA produce actividades predictivas en excelente acuerdo con los valores obtenidos experimentalmente (R^2 = 0.977, predictivo r^2 = 0.862). El modelo tuvo excelentes resultados prediciendo la actividad anti-VIH-1 de los derivados HEPT que no fueron incluidos en el conjunto de desarrollo del modelo (Douali, Villemin, y Cherqaoui, 2003).

2.4 Otros algoritmos de aprendizaje automático

De entre otros muchos algoritmos de AA, por su importancia y aplicación en este campo, se van a referenciar los dos siguientes: bosques aleatorios y máquinas de soporte vectorial.

2.4.1 Bosques aleatorios

El bosque aleatorio (en inglés, Random Forest) es un clasificador conjunto usando muchos modelos de árbol de decisión que puede ser usado para clasificación y regresión. El árbol de decisión es una representación gráfica de un procedimiento para clasificar o evaluar un concepto. Intuitivamente, un árbol de decisión es una colección de condiciones organizadas jerárquicamente. Formalmente, es un árbol donde cada nodo representa una condición o test sobre algún atributo y cada rama que parte de ese nodo corresponde a un posible valor para ese atributo, las hojas son las clases. Para clasificar una instancia se comienza en el nodo raíz, se aplica el test al atributo especificado por este nodo y se sigue la rama que corresponde al valor que dicho atributo tiene en el ejemplo. Este proceso se repite hasta alcanzar una de las hojas la cual indica la clase de la instancia. Si el atributo clase es discreto, el árbol recibe el nombre de clasificación, mientras que si la clase toma valores en un rango continuo es un árbol de regresión (Palma-Méndez y Marín-Morales, 2008).

Un árbol de decisión se puede utilizar como un modelo para problemas secuenciales de decisión bajo incertidumbre. Un árbol de decisión describe gráficamente las decisiones a tomar, los eventos que pueden ocurrir y los resultados asociados con la combinación de las decisiones y eventos. Las probabilidades se asignan a los hechos, y los valores se determinan para cada resultado. Un objetivo importante del análisis es determinar las mejores decisiones.

2.4.2 Máquinas de soporte vectorial

La teoría de las máquinas de soporte vectorial (cuyo acrónimo es SVM por su nombre en inglés Support Vector Machines) es una nueva técnica de clasificación y ha recabado mucha atención en los años recientes. La teoría de la SVM está basada en la idea de minimización de riesgo estructural (SRM) (Betancourt, 2005). En muchas aplicaciones, las SVM han mostrado tener gran eficiencia y han sido introducidas como herramientas poderosas para resolver problemas de clasificación. Una SVM primero “mapea” los puntos de entrada a un espacio de características de una dimensión mayor (i.e., si los puntos de entrada están en $\Re^2$ entonces son “mapeados” por la SVM a $\Re^3$) y encuentra un hiperplano que los separe y maximice el margen “m” entre las clases en este espacio.

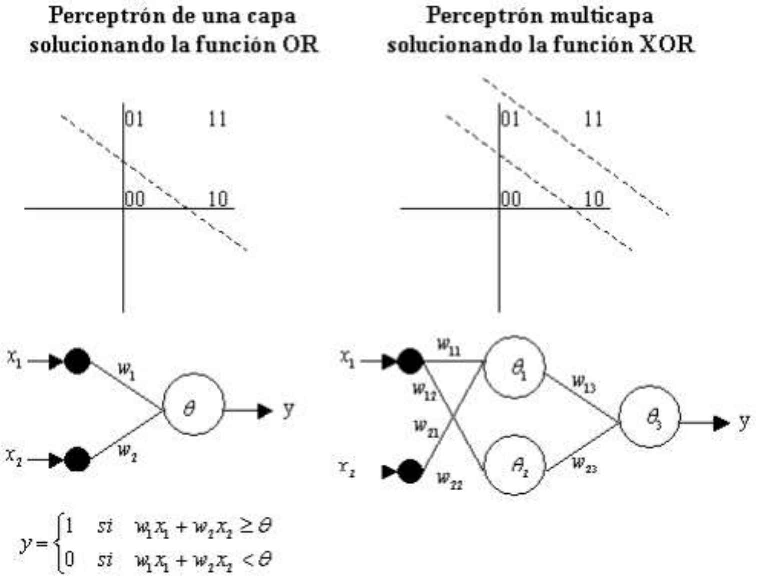
Los fundamentos de las SVM fueron desarrollados por Vapnik. Las SVM están fundamentadas en sólidos principios teóricos y proporcionan un buen rendimiento en gran variedad de aplicaciones prácticas. Las SVM no se centran en construir sistemas que comentan pocos errores, sino que intentan construir modelos fiables en los cuales se pueda obtener una gran confianza, aunque se comentan más errores en el entrenamiento del sistema. Por lo tanto, las SVM tratan de obtener modelos que estructuralmente tengan poco riesgo de cometer errores ante datos futuros (Palma-Méndez y Marín-Morales, 2008), a costa de un entrenamiento más costoso.

2.5 Representación del conocimiento en sistemas adaptativos inteligentes

Se parte de que hay dos hipótesis básicas de representación del conocimiento: La representación puntual, donde cada neurona es, o significa, un concepto determinado (Sin embargo, tiene el problema de la necesidad de recurrir a la redundancia para solucionar la tolerancia a fallos) y la representación holográfica, donde no importa la actividad de un determinado elemento o de una determinada conexión, sino que lo importante es el patrón de actividad de un gran grupo de elementos o de conexiones, o del estado global del sistema en el límite (tiene el inconveniente de que se necesitan muchos elementos y estructuras libres para poder organizar el conocimiento).

Así, lo más natural parece una posición intermedia entre las dos anteriores, consistente en una representación distribuida pero limitada a grupos más o menos grandes de elementos, donde cada grupo representa un concepto conjunto de conceptos relacionados. Esto, además, se correlaciona bastante bien con lo que ocurre en el cerebro donde, ante cualquier tipo de estímulo todos sus elementos reaccionan en mayor o menor medida, pero existiendo unas zonas de mayor respuesta ante un determinado tipo de estímulo.

Un ejemplo básico de esta representación mixta del conocimiento puede apreciarse en la función XOR ("o exclusiva"), la cual puede tomarse como un caso particular del problema de la clasificación. Los primeros sistemas conexionistas tenían importantes limitaciones técnicas. Una de las más importantes es que una RNA tipo Perceptrón con solamente una capa permite discriminar entre dos clases linealmente separables; es decir, cuyas regiones de decisión pueden ser separadas mediante una única recta o hiperplano (dependiendo del número de entradas). Otra importante limitación era la carencia de técnicas para la modificación de conexiones en sistemas de múltiples estratos. Este problema se puede ilustrar con las conocidas funciones OR y OR-Exclusiva (XOR). En el caso de la función OR, un Perceptrón de una sola capa de conexiones modificables permite solucionar esta función debido a que el problema es linealmente separable (ver figura 2 izquierda). En cambio, en el caso de la función OR-Exclusiva, un Perceptrón de este tipo no permite solucionar esta función debido a que no existe ninguna recta que separe los patrones de una clase de los de la otra. Para ello, es necesario que se introduzca una capa intermedia que determine dos rectas en el plano (ver figura 2 derecha) (Montaño Moreno, 2002). La puerta lógica XOR realiza una comparación de las entradas siendo el resultado 0 si las entradas son iguales o 1 cuando son diferentes.



Perceptrón de una capa solucionando la función OR
Perceptrón multicapa solucionando la función XOR
01
11
00
10
01
11
00
10
x_1
x_2
w_1
w_2
θ
y
x_1
x_2
w_{11}
w_{12}
w_{21}
w_{22}
θ_1
θ_2
w_{13}
w_{23}
θ_3
y
$y=\begin{cases}1 & si \quad w_1x_1+w_2x_2 \geq \theta \\ 0 & si \quad w_1x_1+w_2x_2 < \theta\end{cases}$

Figura 2. Perceptrones resolviendo las funciones OR y XOR.

2.6 Quimioterapia y epidemiología del SIDA

A pesar de los avances científicos y de los buenos resultados obtenidos con las distintas medidas terapéuticas implementadas, la infección por VIH-SIDA continúa siendo un grave problema de salud a nivel mundial y es considerado como un tema prioritario dentro de los programas de Salud Pública. El Center for Disease Control and Prevention, www.cdc.gov en Estados Unidos (en adelante, CDC) estima que 1.144.500 personas mayores de 13 años, viven con la infección por el VIH, incluidos 180.900 (15,8%) que no son conscientes de su infección. Durante la última década, el número de personas que viven con el VIH ha aumentado, mientras que el número anual de nuevas infecciones por el VIH se ha mantenido relativamente estable. Sin embargo, el ritmo de nuevas infecciones continúa a un nivel demasiado alto. La incidencia estimada del VIH se ha mantenido estable en general en los últimos años, con cerca de 50.000 nuevas infecciones de VIH por año. Dentro de las estimaciones globales, sin embargo, algunos grupos se ven más afectados que otros. Hombres que tienen sexo con hombres siguen soportando la mayor carga de la infección por el VIH, y entre las razas/etnias, los afroamericanos siguen estando afectados de manera desproporcionada.

Por otro lado, existen diversos recursos con datos epidemiológicos sobre la prevalencia del VIH/SIDA. Por ejemplo, investigadores de la Escuela Rollins de Salud Pública de la Universidad de Emory han compilado en la base de datos AIDSVu (http://aidsvu.org/about-aidsvu/), la información de la prevalencia del SIDA a nivel de condado y estado para EE.UU. La información de AIDSVu proviene de la base de datos nacional de vigilancia, CDC. AIDSVu es un mapa interactivo que está disponible en línea y muestra la prevalencia de VIH en los Estados Unidos (Figura 3).

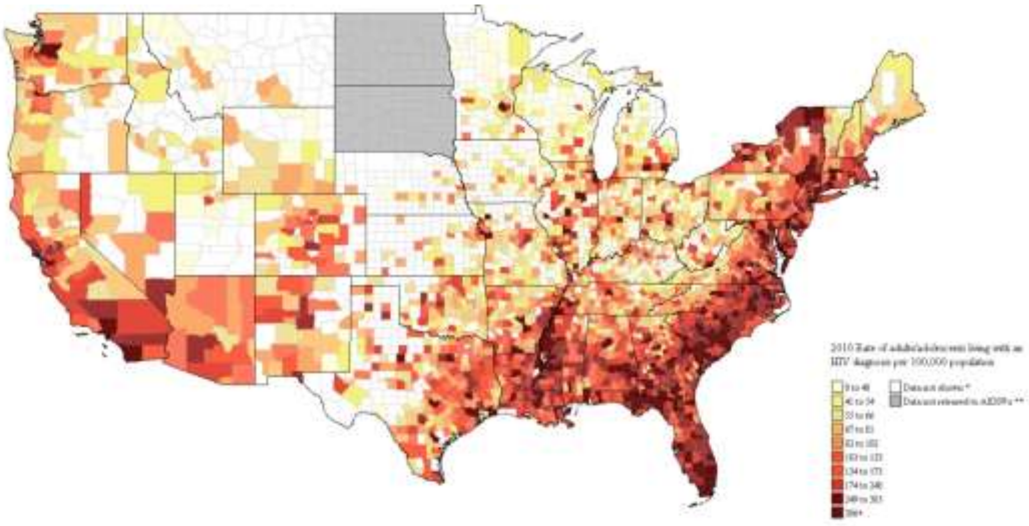

2010 Rate of adults/adolescents living with an HIV diagnosis per 100,000 population
0 to 40
40 to 54
55 to 66
67 to 83
83 to 102
103 to 133
134 to 173
174 to 248
249 to 383
384+
Data not shown *
Data not released to AIDSVu **

Figura 3. Mapa AIDSVu de la frecuencia de VIH en EE.UU. a nivel de condado en el año 2010.

Fuente: AIDSVu (www.aidsvu.org) Emory University, Rollins School of Public Health.

2.6.1 Dianas moleculares de fármacos anti-VIH

Algunas de las dianas más importantes son proteínas presentes en el virus o en el huésped. En la superficie del virus se encuentra la "glicoproteína de envoltura". Esta glicoproteína facilita la entrada del VIH por un proceso de fusión directa entre el virión y la célula diana (Alkhatib, 2009). La fusión de la glicoproteína de envoltura y las membranas celulares diana es iniciada por la unión de la subunidad viral gp120 al receptor CD4 en la superficie de las células T CD4+ (Alkhatib, 2009). Esta interacción crea un sitio de unión de alta afinidad para un correceptor de quimioquinas, como el CXCR4 y/o el CCR5, necesario para la entrada del VIH-1 en la célula diana y la subsiguiente infección (Blanpain, Libert, Vassart, y Parmentier, 2002). Esta fusión introduce el contenido del virión en el citoplasma de la célula, preparando el escenario para la transcripción inversa y por lo tanto para convertir el genoma viral ARN en ADN. La transcriptasa reversa es un paso esencial en la replicación retroviral (Hu y Hughes, 2012). Una vez ocurrido este paso, la enzima integrasa facilita la incorporación del ADN proviral VIH-1 en el genoma de la célula huésped y cataliza una función vital para la replicación viral (Karmon y Markowitz, 2013). Después, la proteasa juega un rol crucial en el ciclo de vida viral por procesamiento de poliproteínas en proteínas estructurales y funcionales esenciales para la maduración viral (Castro et al., 2011; Qiu y Liu, 2011). En la Figura 4 se pueden observar algunas dianas virales y humanas implicadas en la terapia antirretroviral. Los modelos estructurales 3D de las dianas para los medicamentos anti-VIH ilustrados en esta figura fueron descargados de PDBe (http://www.ebi.ac.uk/pdbe/). Las proteínas ilustradas en la Figura 4 son: CXCR4 humano (file: 3oe8) (Wu et al., 2010a), CCR5 humano (file: 4mbs) (Tan et al., 2013), Transcriptasa reversa VIH-1 (file:4mfb) (Lee et al., 2013), integrasa VIH-1 (file:3vqe) (Wielens et al., 2013), proteasa VIH-1 (file:4he9) (Zhang et al., 2013). En la Tabla 1, se pueden observar los distintos compuestos ensayados contra diferentes dianas del VIH.

Tabla 1. Revisión de los resultados del ChEMBL para ensayos de fármacos anti-VIH

ID ChEMBL	*diana*	*acceso*	*tipo diana*	*organismo*	*compuestos*	*actividades*
CHEMBL274	CCR5	P51681	proteína	*H. sapiens*	2922	4740
CHEMBL4015	CCR2	P41597	proteína	*H. sapiens*	2567	4674
CHEMBL2414	CCR4	P51679	proteína	*H. sapiens*	1335	2489
CHEMBL2107	CXCR4	P61073	proteína	*H. sapiens*	550	967

CHEMBL3473	CCR3	P51677	proteína	*H. sapiens*	1276	1550
CHEMBL5412	CCR2	P51683	proteína	*Mm*	74	90
CHEMBL3676	CCR5	P51682	proteína	*Mm*	63	65
CHEMBL378	VIH-1		organismo	*H. sapiens*	16547	41411
CHEMBL243	PR-VIH1	Q72874	proteína	*H. sapiens*	5503	8268
CHEMBL247	RT- VIH1	Q72547	proteína	*H. sapiens*	3292	7187
CHEMBL380	VIH-2		organismo	*H. sapiens*	1703	2834
CHEMBL613758	VIH		organismo	*H. sapiens*	1421	2718
CHEMBL3471	IN-VIH1	Q7ZJM1	proteína	*H. sapiens*	1269	2910
CHEMBL612359	VIH-3		organismo	*H. sapiens*	104	136
CHEMBL613498	VIH-1		organismo	*H. sapiens*	76	76
CHEMBL4609	Tat VIH-1	P04326	proteína	*H. sapiens*	71	75
CHEMBL3520	GP160	P04578	proteína	HIV-1	129	330

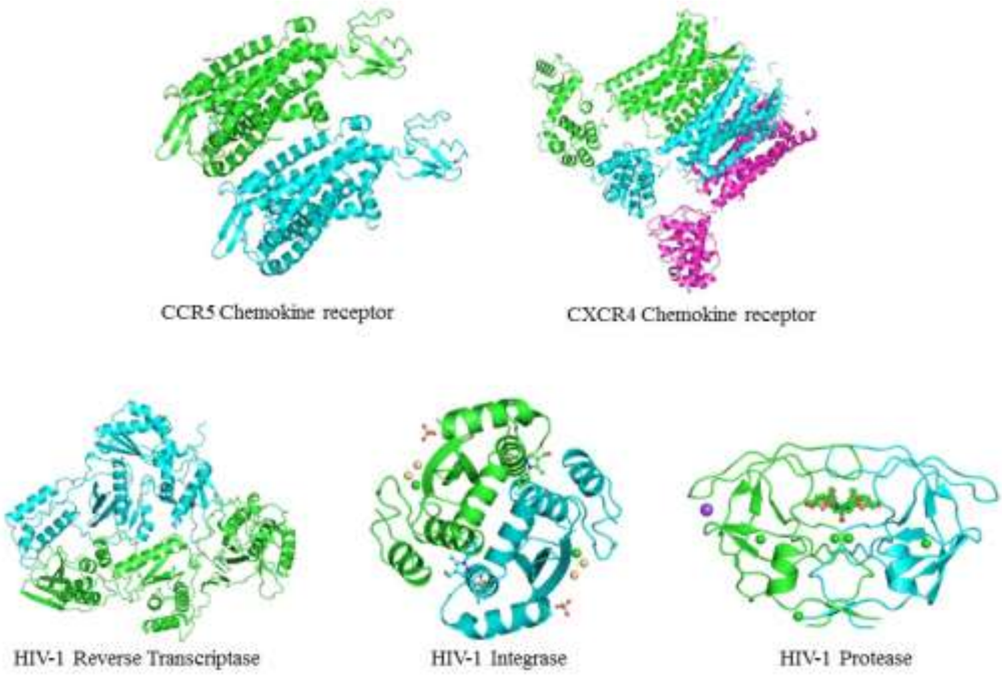


Figura 4. Modelos estructurales 3D de dianas para medicamentos anti-VIH descargados desde el PDBe.

(http://www.ebi.ac.uk/pdbe/): CXCR4 humano (file: 3oe8) (Wu et al., 2010b) , CCR5 humano (file: 4mbs)(Tan et al., 2013), Transcriptasa reversa VIH-1 (file:4mfb) (Lee et al., 2013), integrasa VIH-1 (file:3vqe) (Wielens et al., 2013), proteasa VIH-1 (file:4he9) (Zhang et al., 2013).

2.6.2 Fármacos anti-VIH y terapia TARGA

Las tasas de progresión de la enfermedad, las infecciones oportunistas y la mortalidad han disminuido con la aplicación de la terapia antirretroviral de gran actividad (en adelante, TARGA), reconocida en la literatura en inglés como HAART. Asimismo, la combinación de medicamentos contra el VIH se ha traducido en una mayor supervivencia y una mejor calidad de vida de las personas infectadas con el virus (Colombo et al., 2014).

Las infecciones debidas al VIH se tratan comúnmente con combinaciones de fármacos que consisten en al menos tres fármacos antirretrovirales diferentes. El tratamiento administrado más común a los pacientes consta de dos inhibidores nucleósidos de la transcriptasa reversa en combinación con un inhibidor no nucleósido de la transcriptasa inversa, un inhibidor de proteasa "potenciado" o un inhibidor de la transferencia de cadenas de la integrasa (en adelante, INSTIs). Todos estos tratamientos se han traducido en niveles reducidos de ARN viral (<50 copias / mL) a las 48 semanas y aumento de los recuentos de células CD4 en la mayoría de los pacientes (Usach, Melis, y Peris, 2013).

La terapia antirretroviral incluye: los inhibidores de fusión y de entrada, cuyo uso está normalmente reservado para las personas que han tomado una gran cantidad de medicamentos antirretrovirales en el pasado. La Enfuvirtida pertenece a los inhibidores de fusión, impidiendo la entrada del VIH en la célula CD4 (Qian, Morris-Natschke, y Lee, 2009). El inhibidor del receptor CCR5, Maraviroc, es un inhibidor de entrada el cual se une al receptor CCR5 en la membrana de las células humanas tales como células CD4. Esta unión impide la interacción del VIH-1 gp120 y CCR5 humano, que es necesaria para la entrada en la célula (Wilkin y Gulick, 2012). Los inhibidores nucleósidos/nucleótidos de la transcriptasa reversa (en adelante, INTR) son otra clase de medicamentos contra el VIH. Cuando el virus entra en una célula sana, hace réplicas de sí mismo mediante el uso de la enzima transcriptasa reversa, que es responsable de la transcripción de ARN viral en ADN bicatenario. Los INTR funcionan porque bloquean esa enzima. Algunos ejemplos de este tipo de medicamentos son la Zidovudina, Didanosina, Zalcitabina, Estavudina, Lamivudina, Abacavir, Tenofovir, Emcitrabine (Perno, 2011).

También existen los inhibidores no nucleósidos de la transcriptasa reversa (en adelante, INNTR) cuya interacción con la transcriptasa reversa induce cambios conformacionales que inhiben las actividades catalíticas de la enzima. Ellos se caracterizan por su especificidad para el VIH-1, lo que les hace inhibidores muy selectivos del virus (de Bethune, 2010). Cinco INNTR (Nevirapina, Delavirdina, Efavirenz, Etravirina y Rilpivirina) están actualmente aprobados por la FDA. Además, todos ellos a excepción de la Delavirdina, han sido aprobados por la Unión Europea (Usach et al., 2013). Los inhibidores de la integrasa son otra clase importante de fármacos anti-VIH. La integrasa transfiere el ADN viral codificado en el cromosoma del huésped, este es un evento necesario en la replicación del retrovirus (Hicks y Gulick, 2009). El Raltegravir y Dolutegravir son ejemplos de inhibidores de la integrasa (Adams, Greener, y Kashuba, 2012; Powderly, 2010). Por último, los inhibidores de la proteasa son compuestos que impiden la maduración de la proteína viral inhibiendo competitivamente la proteasa VIH-1, porque en el VIH-1, como en todos los retrovirus, la producción de virus infeccioso requiere invariablemente una proteasa viral activa (Eron, 2000). Algunos ejemplos de este tipo de fármacos son Amprenavir, Atazanavir, Indinavir, Nelfinavir, Lopinavir, Saquinavir, Tipranavir, y Ritonavir (Arts y Hazuda, 2012; Chougrani, Luton, Matheron, Mandelbrot, y Azria, 2013). En la Figura 5, se ilustran algunas estructuras químicas de los distintos fármacos implicados en la terapia antirretroviral.

Algunos ejemplos de combinación de medicamentos contra el VIH aprobados por la FDA son: Atripla®, que contiene dos INTR, Emtriva® (Emtricitabina) y Viread® (Tenofovir disoproxil fumarato) y un INNTR, Sustiva® (Efavirenz) (King et al., 2011); Complera®, una combinación de 2 INTR (Emtricitabina y Tenofovir disoproxil fumarato) y un INNTR (Rilpivirina) (O'Neal, 2011); Stribild®, una combinación de un INSTI (Elvitegravir), un potenciador farmacocinético (cobicistat), un INTR (Emtricitabina), y un inhibidor de la transcriptasa inversa análogos de nucleótidos N(t)RTI (Tenofovir disoproxil fumarato) (Perry, 2014). Combivir® contiene 2 INTR (Zidovudina y Lamivudina) (Portsmouth y Scott, 2007); Truvada® contiene 2 INTR (Emtricitabina/Tenofovir) (Coutinho y Prasad, 2013). Kaletra® contiene dos inhibidores de la proteasa (Lopinavir y Ritonavir) (Lopez Aspiroz et al., 2011); Trizivir® contiene una combinación de dosis fija de tres INTR (sulfato de Abacavir, Lamivudina y Zidovudina) (Shey, Kongnyuy, Shang, y Wiysonge, 2009); Epzicom® o Kivexa® que en Europa contiene dos INTR (sulfato de Abacavir y Lamivudina) (Sax et al., 2009).

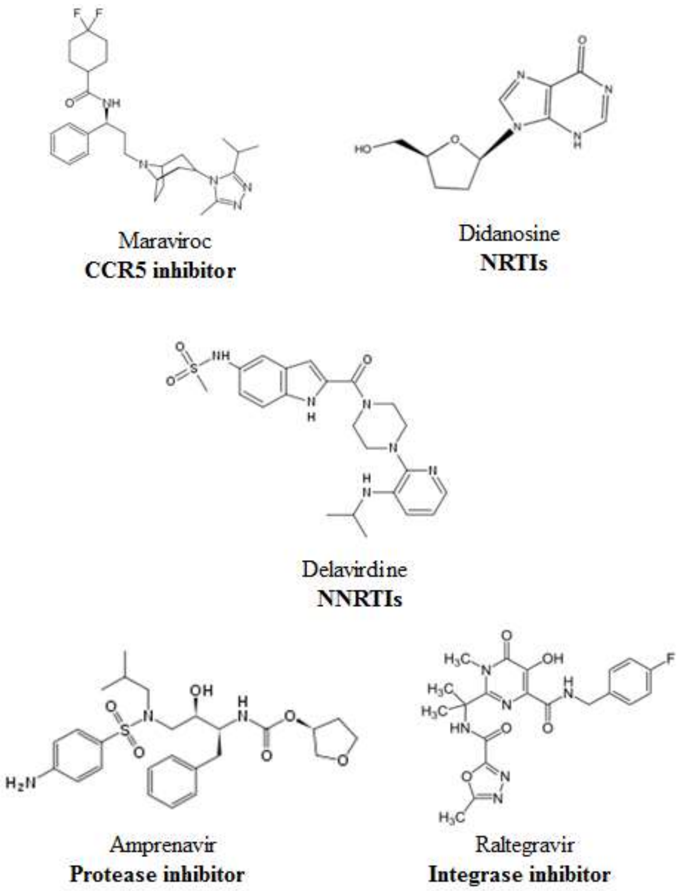


Figura 5. Estructuras químicas de fármacos contra el VIH

2.6.3 Mutaciones asociadas a resistencia en la terapia antirretroviral

El desarrollo de la terapia antirretroviral de gran actividad (TARGA) ha sido de gran importancia en el tratamiento contra el VIH. La terapia actual ha reducido en forma significativa la morbi-mortalidad asociada a la infección por VIH, llevando la carga viral a niveles indetectables y restaurando secundariamente el sistema inmune, evidenciado por el aumento de linfocitos T CD4 (+). Asimismo, como efecto indirecto, la TARGA puede disminuir la transmisión del VIH. Sin embargo, la aparición de resistencias es una de las principales causas del fracaso en el tratamiento antirretroviral. La disminución de la sensibilidad del virus a los fármacos antirretrovirales se debe a mutaciones o cambios en el genoma viral. Dichas mutaciones se producen como consecuencia de una replicación viral persistente en presencia de concentraciones subóptimas de los fármacos antirretrovirales.

Las mutaciones se denominan según el aminoácido que sea reemplazado en un codón determinado; así por ejemplo, en la mutación M184V el aminoácido original (cepa salvaje) es metionina (M), el número representa la posición del codón en el genoma (184) y, a continuación, sigue el aminoácido reemplazante (mutante), en este caso valina (V). La cantidad de mutaciones necesarias para que aparezca resistencia puede variar entre una (como en el caso de resistencia a la Lamivudina con la mutación M184V) o varias (como en el caso de los inhibidores de la proteasa). De este modo, los antirretrovirales pueden clasificarse en fármacos de barrera genética baja si una sola mutación es suficiente para inducir resistencia, o de barrera genética alta si por el contrario se precisa la acumulación de varias mutaciones para reducir la sensibilidad al fármaco (Carmona, 2014).

En diversos estudios poblacionales, entre 30 y 63% de los pacientes fallan al año de haber iniciado tratamiento. La resistencia no es sólo un problema en pacientes con fracaso virológico, sino también en pacientes vírgenes en la toma de tratamiento, debido a transmisión de virus resistentes. Comprender la resistencia es la clave para enfocar una TARGA óptima. A continuación, en las Figuras 6 a 9 se observan las mutaciones presentes en las principales dianas de los fármacos antirretrovirales (Wensing et al., 2014):

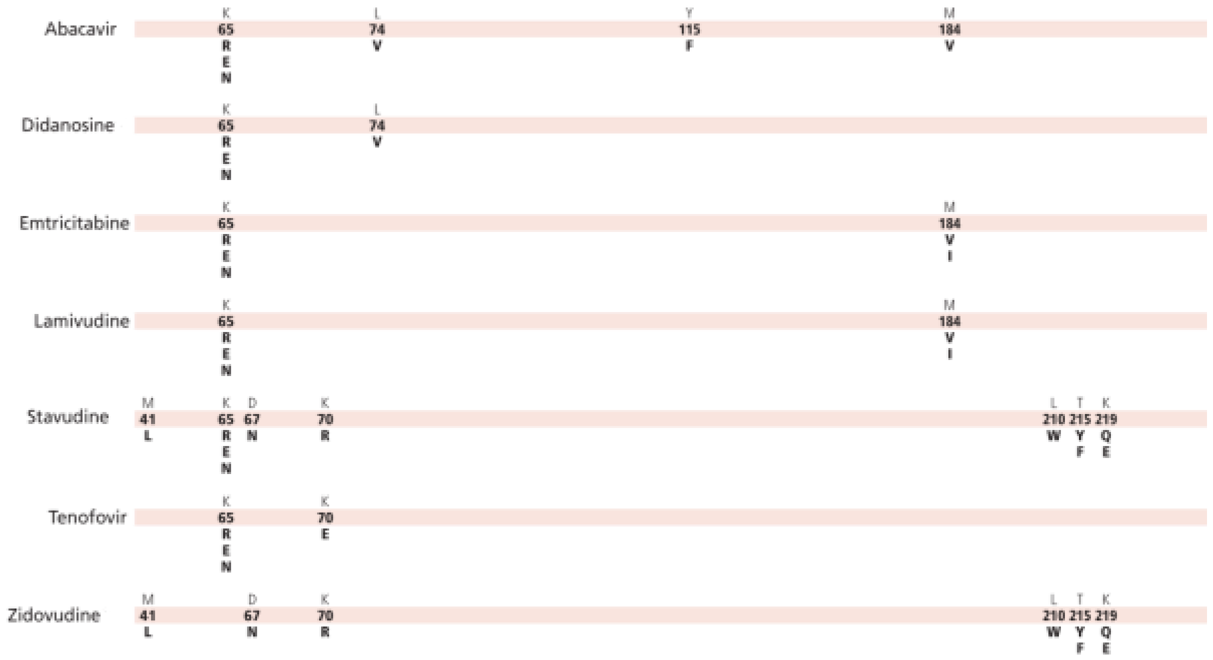


Figura 6. Mutaciones asociadas a los INTR

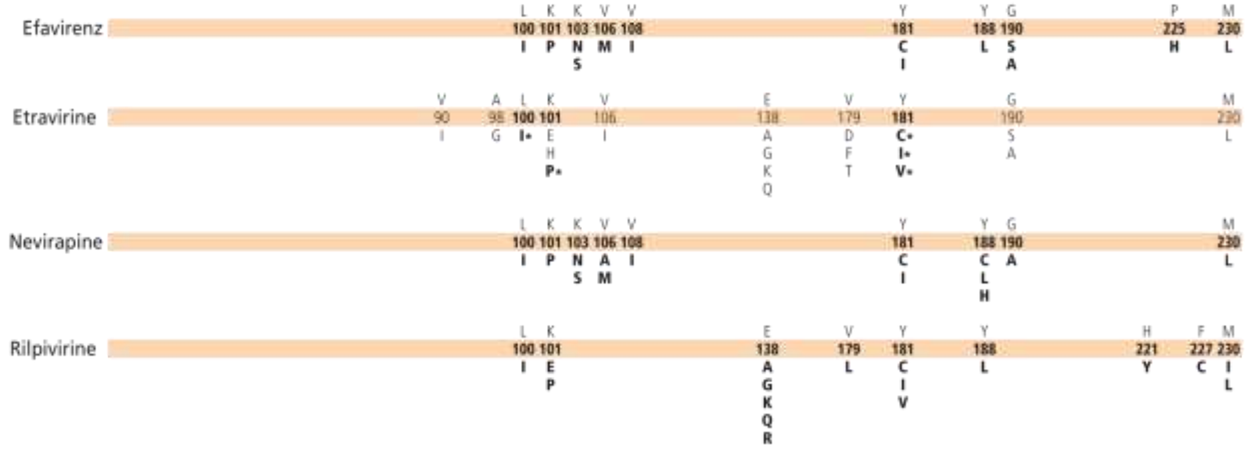


Figura 7. Mutaciones asociadas a los INNTR

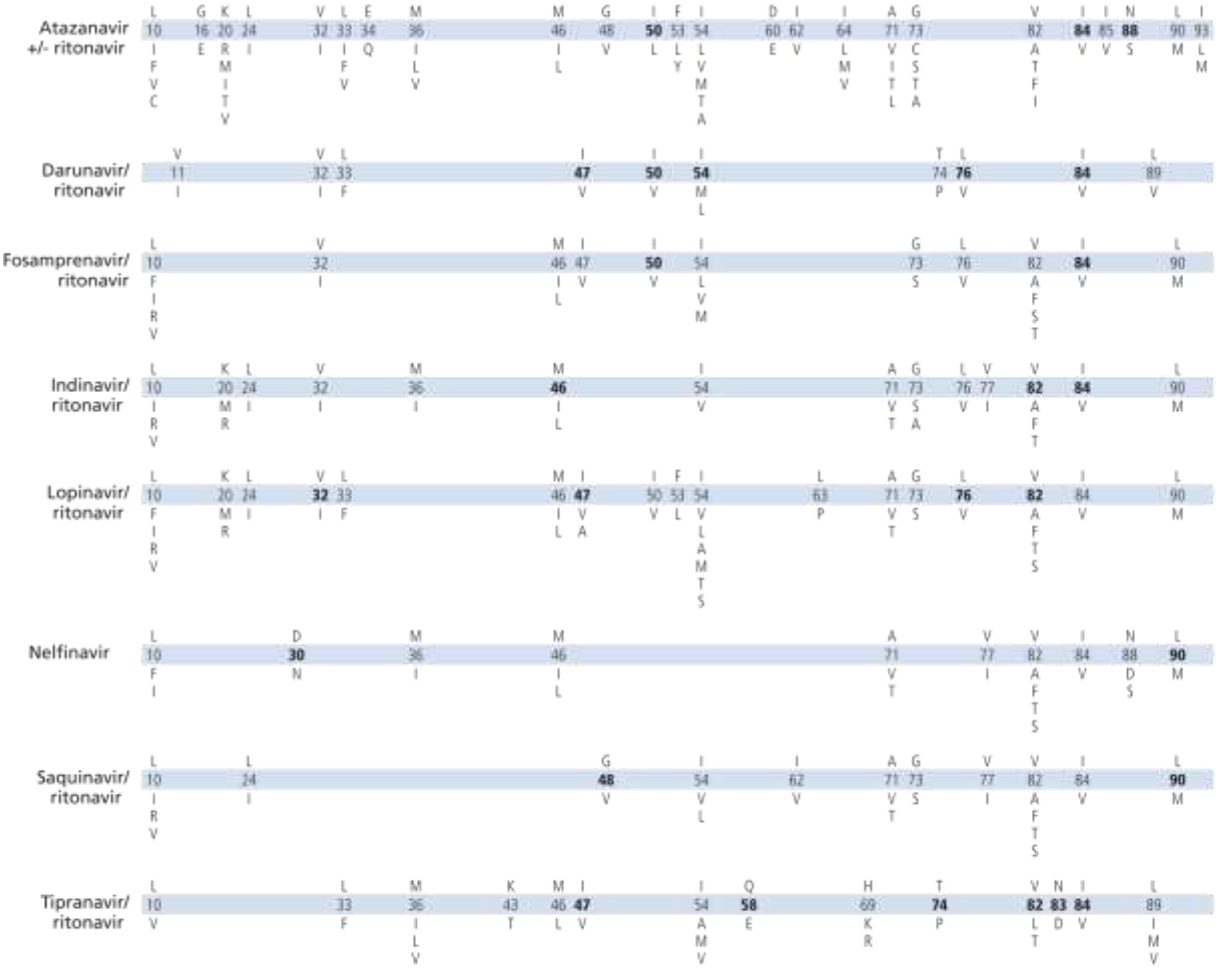


Figura 8. Mutaciones asociadas a los inhibidores de proteasa

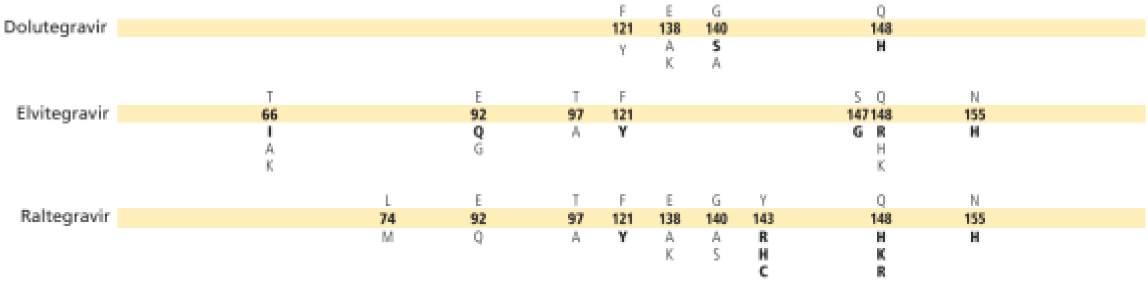


Figura 9. Mutaciones asociadas a los inhibidores de la integrasa

2.6.4 Factores socio-económicos en la epidemiología del SIDA

La relación entre los determinantes sociales y la prevalencia del SIDA desempeña un papel importante en el estudio de la eficacia de la terapia TARGA. Sin embargo, la evidencia de asociación entre la transmisión del VIH y el nivel socioeconómico es aún rudimentaria y diversa. Algunos estudios reportan un menor nivel socioeconómico, que se asocia con mayor mortalidad por SIDA (McFarland, Chen, Hsu, Schwarcz, y Katz, 2003). La evidencia reciente indica que el SIDA es una enfermedad de desigualdad, a menudo asociada con la transición económica, en vez de una enfermedad de pobreza en sí misma (Piot, Greener, y Russell, 2007). Actualmente, muchos investigadores no señalan a la pobreza en sí, sino a las desigualdades económicas y de género, y una cohesión social debilitada (Barnett y Whiteside, 2006) incluyendo los factores que influyen en el comportamiento sexual y, por lo tanto, en la posibilidad de transmisión del VIH. Sin lugar a dudas, más personas viven con el VIH en los países pobres que en los ricos. Más del 60% de las personas que viven con el VIH habitan en la región más pobre del mundo: África subsahariana. Sin embargo, los estudios durante la fase inicial de la epidemia sugieren que la incidencia del VIH se produjo inicialmente entre los miembros más acomodados de la sociedad en esta región, y no entre los más pobres. Una década más tarde, la infección se presenta más concentrada en los entornos urbanos y entre las personas con mayor capacidad de movilidad de la sociedad, y en consecuencia en los grupos con mayor riqueza (Piot et al., 2007).

Los factores sociales pueden afectar las posibilidades de una persona de contraer una enfermedad crónica o infecciosa, como el VIH, a través de las influencias sobre el comportamiento, el acceso limitado a las medidas de prevención y el acceso limitado a los proveedores de cuidado de la salud o sitios de pruebas para el VIH. La investigación muestra que el entorno social en el que vive un individuo tiene un gran impacto sobre la infección por el VIH (Gant et al., 2012). Hay una serie de factores sociales y estructurales que aumentan el riesgo de infectarse con el virus, incluyendo el contexto cultural, las redes sociales, los efectos de vecindad, la violencia estructural, la discriminación y el cambio demográfico de una persona. Además, el nivel socioeconómico y el estado de pobreza afectan en gran medida las tasas de infecciones por transmisión sexual, incluyendo la infección por el VIH.

Resumiendo, la revisión sistemática de la evidencia disponible no proporciona un soporte concluyente de la existencia de una clara asociación entre el estatus socioeconómico y la adherencia al tratamiento antirretroviral en pacientes adultos infectados por el VIH/SIDA en los países de ingresos bajos y medianos. Sin embargo, parece que hay una tendencia positiva entre los componentes del nivel socioeconómico (ingresos, educación, situación laboral) y la adherencia a la terapia antirretroviral (Gillespie, Kadiyala, y Greener, 2007; Hajizadeh, Sia, Heymann, y Nandi, 2014; Lim et al., 2013; McDavid Harrison, Ling, Song, y Hall, 2008; McFarland et al., 2003; McMahon, Wanke, Terrin, Skinner, y Knox, 2011; Peltzer y Pengpid, 2013).

2.6.4.1 Índices de información de Shannon

Se puede calcular un índice de información para cuantificar los riesgos de propagación/prevalencia de SIDA en diferentes condados de Estados Unidos, en una situación inicial en la que cada condado tiene un valor de la tasa de prevalencia del SIDA (D_a) en el momento inicial (t_0 = 2010). Un índice de información sencillo ($I^a{}_0$) se usa para la desigualdad de ingresos en los diferentes condados de ese año. Este índice depende de la probabilidad ^0p_a, con la cual el condado presenta cierta desigualdad de ingresos. Esta probabilidad $^0p_a = G_a$ se estableció en el presente trabajo. En esta definición, G_a es la medida de Gini de desigualdad de los ingresos en el condado de un estado dado en los Estados Unidos (Pabayo, Kawachi, y Gilman, 2014). La clase de índice de información seleccionada fue el índice de entropía de Shannon (Riera-Fernandez et al., 2012).

$$I^a{}_0 = -{}^0p_a \cdot \log \left({}^0p_a\right) \tag{1}$$

2.6.4.2 Índices de información Markov-Shannon

Las cadenas de Markov han sido usadas para calcular los índices de información de Shannon de diferentes sistemas, incluyendo simulaciones relevantes en Epidemiología de transmisión de la enfermedad (Riera-Fernandez et al., 2012). Se puede definir el vector de probabilidades absolutas iniciales $^{0}p \equiv [^{0}p_{1}, {}^{0}p_{2}, {}^{0}p_{3}..., {}^{0}p_{a}..., {}^{0}p_{nt}]$ para los n_t condados en el mismo estado. Se puede calcular la probabilidad absoluta de ocurrencia de la enfermedad en un condado dado ($^{0}p_{a}$) en un tiempo inicial (t_0) como en otros modelos de cadenas de Markov (Gonzalez-Diaz, Aguero, et al., 2005; Gonzalez-Diaz, Cruz-Monteagudo, Molina, Tenorio, y Uriarte, 2005; Van Waterbeemd, 1995):

$$^{0}p_{a} = \frac{G_{a}}{\sum_{a=1}^{nt} G_{c}} \quad \textbf{(2)}$$

Aquí, G_a es el coeficiente de Gini de desigualdad de los ingresos (Pabayo, Kawachi, y Gilman, 2013) en el condado a-ésimo de un estado dado en los EE.UU. Se debe considerar que el único factor epidemiológico utilizado como entrada para calcular los índices de Shannon del condado era el G_a. Utilizando la ecuación de Chapman-Kolmogorov, se puede calcular el vector $^{k}p^{t} \equiv [^{k}p_{1}, {}^{k}p_{2}, {}^{k}p_{3}..., {}^{k}p_{a}..., {}^{k}p_{na}]$ para las probabilidades absolutas $^{k}p_{a}$ a lo largo de tiempo t_k:

$$^{k}\mathbf{p} = \left(^{1}\mathbf{\Pi}\right)^{k} \cdot {}^{0}\,\mathbf{p} = {}^{k}\mathbf{\Pi} \cdot {}^{0}\,\mathbf{p} \quad \textbf{(3)}$$

En consecuencia, los elementos de la matriz estocástica $^{k}\mathbf{\Pi} = (^{1}\mathbf{\Pi})^{k}$ son las probabilidades $^{k}p_{ab}$ de transmisión del SIDA de un condado a otro en t_k = k años (pasos). Se calcularon los elementos de la matriz estocástica $^{1}\mathbf{\Pi}$ como sigue (Gonzalez-Diaz, Aguero, et al., 2005; Gonzalez-Diaz, Cruz-Monteagudo, et al., 2005; Gonzalez-Diaz, Prado-Prado, Santana, y Uriarte, 2006; Van Waterbeemd, 1995).

$$^{1}p_{ab} = \frac{\left(G_{a} + G_{b}\right) \cdot \exp\left(G_{b}\right)}{\sum_{c=1}^{c=n} \left(G_{a} + G_{c}\right) \cdot \exp\left(G_{c}\right)} \quad \textbf{(4)}$$

Posteriormente, se calcularon los índices de información $I^{a}_{k}(s)$ para cuantificar la desigualdad en los ingresos/epidemiología esperada en los condados y sus vecinos mediante la fórmula de la entropía de Shannon (Riera-Fernandez et al., 2012).

2.7 Quimio-informática y diseño de terapias anti-VIH

2.7.1 Modelos QSAR para fármacos anti-virales

El uso de métodos computacionales en el diseño de nuevos ligandos para una determinada diana terapéutica se denomina comúnmente CADD. La genómica, proteómica y bioinformática están descubriendo nuevas dianas terapéuticas y contribuyendo cada vez más al descubrimiento de fármacos. Asimismo, el desarrollo y lanzamiento de un nuevo fármaco al mercado requiere a la industria farmacéutica una media de 12 a 20 años y unos costes de aproximadamente 850 millones de euros (Beas Zárate, Ureña Guerrero, Rivera Cervantes, Pallàs i Llibería, y Camins Espuny, 2010). Las técnicas de cribado virtual ahorran en la compra de reactivos, el tiempo, la robotización y la disminución de experimentación en animales. Asimismo son rápidas, y permiten considerar un gran número de compuestos *in silico*.

Los modelos QSAR se utilizan como modalidad inicial de "screening" en el descubrimiento y desarrollo de fármacos, estableciendo una relación cuantitativa entre la estructura química de los fármacos y las posibles dianas moleculares, además de la actividad biológica y/o capacidad específica de interaccionar. Una limitación de casi todos los modelos QSAR/QSPR es que predicen la actividad biológica de los medicamentos sólo para un sistema biológico (organismo, diana, etc.). La solución viene con el desarrollo de modelos de múltiple tarea QSAR/QSPR (mt-QSAR/mt-QSPR) para predecir la actividad/propiedades de los fármacos ante diferentes sistemas biológicos. Estos mt-QSAR/mt-QSPRs ofrecen también una buena oportunidad para la construcción de redes complejas que se pueden utilizar para explorar grandes y heterogéneas bases de datos de medicamentos/sistemas biológicos. En esta sección, se van a revisar algunos de los modelos mt-QSAR/QSPR propuestos en la literatura.

Las técnicas QSAR se utilizan de forma rutinaria por especialistas en química computacional en el descubrimiento y desarrollo de fármacos, para analizar conjuntos de datos de compuestos. Métodos numéricos cuantitativos como las RRNNAA se han utilizado en QSAR para establecer correlaciones entre las propiedades moleculares y la bioactividad. En la literatura se reportan diversos modelos para el descubrimiento de fármacos antivirales contra determinadas enfermedades. Por ejemplo, el estudio de Sirois y colaboradores, se centra en los inhibidores de proteasa del VIH-1 pertenecientes a la clase de compuestos peptidomiméticos (Sirois et al., 2005). El principal objetivo fue seleccionar los descriptores moleculares con el mejor valor predictivo para la potencia antiviral (Ki). Se utilizaron técnicas como: Mínimos cuadrados parciales (PLS) y RRNNAA para predecir la actividad Ki de los inhibidores de proteasa contra VIH-1 y se compararon los resultados. Para abordar la cuestión de la reducción de dimensionalidad, se utilizaron algoritmos genéticos (GA) para la selección de variables y su rendimiento se comparó con las RRNNAA. La formación y la validación de los modelos se realizaron en 15 conjuntos divididos al azar de la base de datos maestra conformada por 231 compuestos. Para cada compuesto, se consideraron 192 descriptores moleculares. Los hallazgos encontrados en el estudio sugieren que las interacciones no covalentes tales como la hidrofobicidad, la forma y el enlace de hidrógeno describen bien la actividad antiviral de los compuestos de la proteasa del VIH-1.

El resveratrol es un polifenol natural encontrado en el vino tinto que puede inhibir la replicación *in vivo* e *in vitro* del virus de la influenza A, sin producir ninguna toxicidad importante. Li y colaboradores estudiaron la relación de 35 derivados del resveratrol y su actividad inhibitoria, mediante el uso de dos métodos 3D-QSAR (CoMFA Y CoMSIA) (Li et al., 2014). Los resultados indicaron que los derivados del resveratrol son compuestos antivirales potencialmente útiles para el diseño y desarrollo de nuevos fármacos para el tratamiento de la influenza. Además, otros autores han estudiado compuestos naturales (bioflavonoides) contra el virus de la hepatitis C (Mathew et al., 2014), creando el método VEGA-QSAR, el cual fue aplicado para predecir las diversas propiedades bioquímicas de moléculas ligando. Los resultados obtenidos por medio de los modelos computarizados QSAR son potencialmente efectivos en evaluar las propiedades químicas de los compuestos seleccionados, y de esta forma reduciendo la experimentación en animales. Los flavonoides seleccionados mostraron una predicción positiva siendo no mutagénicos y no carcinogénicos.

Se ha puesto en práctica la teoría de la cadena de Markov para calcular nuevos momentos espectrales con el fin de ajustar un modelo mt-QSAR para medicamentos activos contra 40 especies virales (Prado-Prado, Borges, Uriarte, Perez-Montoto, y Gonzalez-Diaz, 2009). El modelo está basado en 500 fármacos (compuestos activos e inactivos) analizados como agentes antivirales en la literatura reciente; aunque no todos los fármacos fueron evaluados contra todos los virus, sólo aquellos con valores experimentales. Se emplearon técnicas como el Análisis Discriminante Lineal (LDA) para clasificar todos estos medicamentos en dos clases: compuestos activos o inactivos contra las diferentes especies virales analizadas. El modelo clasificó correctamente 5129 de los 5594 compuestos inactivos (sensibilidad = 91,69%) y 412 de los 422 compuestos activos (especificidad = 97,63%). La ecuación del modelo es la siguiente:

$$\begin{aligned} Actv = {} & -0.95 \cdot {}^{0}\mu_s(\mathrm{H\text{-}Het}) + 1.50 \cdot {}^{2}\mu_s(\mathrm{H\text{-}Het}) - 3.23 \cdot {}^{0}\mu_s(\mathrm{C_{uns}}) \\ & - 4.02 \cdot {}^{0}\mu_s(\mathrm{C_{sat}}) - 0.47 \cdot {}^{1}\mu_s(\mathrm{T}) + 10.34 \cdot {}^{0}\mu_s(\mathrm{T}) + 0.74 \cdot {}^{5}\mu_s(\mathrm{X}) \\ & - 8.88 \\ & \qquad \lambda = 0{,}51; \quad \chi^2 = 4024.83; \quad \mathrm{p} < 0.001 \end{aligned} \qquad \textbf{(5)}$$

donde λ es la estadística de Wilk; χ2 chi cuadrado y *p* el nivel de error. En la ecuación, ${}^{k}\mu_{s}$ es el momento espectral de una cierta especie después de *k* etapas. Se ha calculado para el total (T) de los átomos en la molécula o para asociaciones específicas de átomos. Estas asociaciones son átomos con una característica común: H-Het = hidrógeno unido a heteroátomos, C_{uns} = átomos de carbono insaturados, C_{sat} = átomos de carbono saturados, X = átomos de halógeno.

2.7.2 Modelos QSAR para fármacos anti-VIH

Algunos autores indican que los resultados de sus estudios *in silico* proveen una contribución importante para el diseño de nuevas moléculas activas para la inhibición de algunas dianas implicadas en el VIH (Debnath, Verma, Jain, Katti, y Prabhakar, 2013; Swiderek, Marti, y Moliner, 2012). Existen cuantiosos estudios aplicando técnicas QSAR a la epidemia del VIH/SIDA, predominando los estudios QSAR en tercera dimensión: CoMFA, CoMSIA (Debnath et al., 2013; Gadhe, Kothandan, y Cho, 2013; Ma, Ye, Ji, y Chen, 2013; Moonsamy, Dash, y Soliman, 2014; Sun, Guan, Tan, Liu, y Wang, 2012; Wu, Yao, He, y Chen, 2014).

Asimismo, se han realizado estudios QSAR en derivados de curcumina como potenciales inhibidores de la enzima integrasa del VIH-1, usando regresión lineal múltiple. El modelo estadísticamente significativo fue desarrollado con coeficientes de correlación al cuadrado r^2 = 0,891 validación cruzada r^2= 0.825. El modelo desarrollado reveló que la forma, el tamaño, la geometría, la información de la sustitución y la hidrofilicidad eran importantes propiedades atómicas para la determinación de la actividad inhibidora de estas moléculas. El modelo también fue probado con éxito para la validación externa (r^2 pred = 0,849), así como la prueba de Tropsha para la predictibilidad del modelo. El modelo fue estadísticamente robusto y tenía un buen poder predictivo que puede ser utilizado con éxito para la detección de nuevas moléculas (Gupta, Sharma, Garg, y Roy, 2013).

En el presente trabajo, se llevaron a cabo estudios de QSAR 3D y de “docking” en una serie de derivados del Danuravir, el inhibidor de proteasa del VIH más potente conocido hasta ahora (Ul-Haq, Usmani, Shamshad, Mahmood, y Halim, 2013). Estudios combinados de 3D-QSAR se aplicaron para los derivados del Danuravir utilizando protocolos basados en ligando y basados en receptor (ligand-based and receptor-based), y comparando los modelos generados. En los resultados, se observó una relación positiva con los resultados experimentales. Este conjunto de información podría ser utilizada para diseñar fármacos candidatos altamente potentes tanto para la forma salvaje como para la mutada del virus.

2.8 Cálculo de índices de información

Los índices de información calculados por el software DRAGON son descriptores moleculares definidos como contenido total y contenido de información de las moléculas. Este tipo de parámetro de información es muy importante porque puede aplicarse también al estudio de sistemas no moleculares. Por lo tanto, en el presente trabajo se usan únicamente los índices de información calculados por el DRAGON y no se usan otras familias de descriptores. En estos índices de información calculados para moléculas se utilizan diferentes criterios para definir las clases de equivalencia; es decir, la equivalencia de átomos en una molécula tal como la identidad química, formas de enlace a través del espacio, la topología molecular y la simetría.

2.8.1 Software DRAGON

DRAGON versión 5.3 (Figura 10) es una aplicación para el cálculo de descriptores moleculares originalmente desarrollado por el Milano Chemometrics and QSAR Research Group (http://michem.disat.unimib.it/chm/). Estos descriptores se pueden utilizar para evaluar las relaciones moleculares estructura-actividad o estructura-propiedad, así como para el análisis de similitud y cribado de alto rendimiento de bases de datos de moléculas. El primer lanzamiento de DRAGON se remonta a 1997, las actualizaciones y las inclusiones de nuevos descriptores moleculares se hacen regularmente con el fin de avanzar en la investigación en QSAR (Todeschini, Consonni, Mauri, y Pavan, 2005). El software DRAGON ofrece más de 1.600 descriptores moleculares que se dividen en 20 bloques lógicos. El usuario puede calcular no sólo el tipo de átomo, grupo funcional y fragmentos de recuentos simples, sino también varios descriptores topológicos y geométricos. Algunas propiedades moleculares tales como log P, refractividad molar, el número de enlaces rotables, H-donantes, H-aceptantes, y el área de superficie topológica (TPSA) también se calculan mediante el uso de algunos modelos comunes tomados de la literatura.

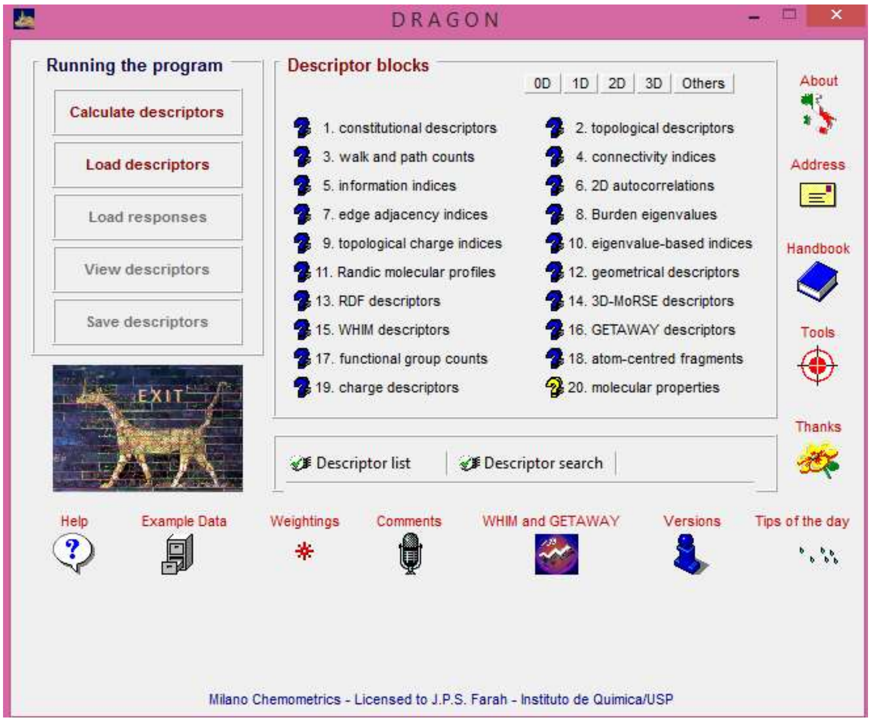


Figura 10. Interfaz software DRAGON versión 5.3

A continuación, se detallan los índices de información usados en esta tesis:

2.8.1.1 Índices de información de Balaban

Los índices de información de Balaban U, V, X e Y (Balaban y Balaban, 1991) son muy útiles para cuantificar información sobre la estructura química de los fármacos (Ivanciuc, Balaban, y Balaban, 1993). Estos índices incorporan en sus ecuaciones los siguientes parámetros: σ_x=grado de distancia del vértice del átomo x-ésimo (suma de las distancias topológicas desde el átomo considerado a cualquier otro átomo), d_{xy} es la distancia topológica entre los átomos x-ésimo e y-ésimo; nSK es el número de átomos diferentes al Hidrógeno. Otros parámetros usados son ${}^{g}f_{x}$ = el número de distancias desde el vértice x-ésimo igual a *g* y η_x = la excentricidad del átomo x-ésimo (la distancia máxima topológica del átomo considerado). Se indican estos índices en el presente trabajo como I^q_k. En esta notación, la letra I representa el índice de información, q indica el compuesto y k el tipo de índice. Las fórmulas matemáticas para el cálculo de estos índices son:

$$I^{q}{}_{1} = U_q = -\sum_{j=1}^{nSK} \frac{d_{xy}}{\sigma_x} \cdot \log_2 \frac{d_{xy}}{\sigma_x} \qquad \textbf{(6)}$$

$$I^{q}{}_{2} = V_q = \sigma_x \cdot \log_2(\sigma_x) - u_q = \sigma_x \cdot \log_2(\sigma_x) + \sum_{j=1}^{nSK} \frac{d_{xy}}{\sigma_x} \cdot \log_2 \frac{d_{xy}}{\sigma_x} \qquad \textbf{(7)}$$

$$I^{q}{}_{3} = Y_q = \sum_{g=1}^{\eta_i} {}^{g} f_q \cdot g \cdot \log_2(g) \qquad \textbf{(8)}$$

$$I^{q}{}_{4} = X_q = \sigma_q \cdot \log_2(\sigma_q) - \sum_{g=1}^{\eta_i} {}^{g} f_q \cdot g \cdot \log_2(g) \qquad \textbf{(9)}$$

2.8.1.2 Índices de información de simetría

Son índices de información topológicos calculados para un grafo molecular incluyendo Hidrógeno y basados en los grados de vecindad y multiplicidad de enlace químico (aristas del grafo molecular) (Magnuson, Harriss, y Basak, 1983; Todeschini y Consonni, 2000). Estos índices son calculados haciendo una partición del grafo molecular en clases de equivalencias; dos átomos (nodos) se consideraron como pertenecientes a la misma clase de equivalencia topológica cuando sus vecinos correspondientes de orden k-ésimo son los mismos. El software DRAGON versión 5.3, calcula los índices de simetría de vecindad desde el orden cero hasta cinco (Todeschini et al., 2005). Los nombres, símbolos y fórmula para calcular los índices se explica a continuación:

Hay 5 tipos de índices de simetría, el primer índice es el contenido de información de vecindad (IC*k*) (en inglés *neighborhood information content*). El IC*k* usa los siguientes parámetros: A_g es la cardinalidad de la equivalencia de clase g-*ésimo* y *n*AT es el número total de átomos. Este índice representa una medida de complejidad estructural por vértice. La fórmula de IC*k* es la siguiente:

$$IC_k = -\sum_{g=1}^{G} \frac{A_g}{nAT} \cdot \log_2 \frac{A_g}{nAT} \quad \textbf{(10)}$$

El siguiente índice es el contenido de información total de vecindad (TIC*k*) (en inglés *neighborhood total information content).* Este índice representa una medida de complejidad del grafo.

$$TIC_k = nAT \cdot IC_k \quad \textbf{(11)}$$

El tercer índice es el contenido de información estructural (*SICk*) *(*en inglés *structural information content).* Se calcula en una forma normalizada del índice IC*k* para eliminar la influencia del tamaño de grafo

$$SIC_k = \frac{IC_k}{\log_2 nAT} \quad \textbf{(12)}$$

El cuarto índice es el contenido de información de enlace (BIC*k*) *(*en inglés *bonding information content).* Este índice se calcula en una forma normalizada del IC*k* teniendo en cuenta el número de enlaces y su multiplicidad. Utiliza el parámetro *n*BT, que es el número de enlaces, y π* es el orden de enlace convencional (1 individual, 2 doble, 3 triples y de 1,5 para los enlaces aromáticos).

$$BIC_k = \frac{IC_k}{\log_2\left(\sum_{b=1}^{nBT} \pi_b^*\right)} \quad \textbf{(13)}$$

Por último, el quinto índice es el contenido de información complementaria (CIC*k*) *(*en inglés *complementary information content).* Mide la desviación del IC*k* de su valor máximo, que corresponde a la partición de vértice en clases de equivalencia que contienen un elemento cada uno.

$$CIC_k = \log_2 nAT - IC_k \qquad \textbf{(14)}$$

2.8.1.3 Otros índices de información

El cálculo de los índices de información requiere el uso de diferentes parámetros de entrada. Algunos de estos parámetros son el número de elementos o nodos (átomos) del grafo molecular **G**, el número de diferentes clases de equivalencia G, y n_g es el número de elementos en la clase g-ésima, el logaritmo es en base 2 para la medición del contenido de información en bits, nAT es el número de átomos (incluyendo hidrógeno). Otros parámetros son ${}^{g}f_{i,}$ el cual es el número de distancias desde el vértice v_i, igual a g, η_i es la excentricidad del átomo (la distancia topológica máxima desde el vértice v_i). El parámetro nSK es el número de átomos distinto al hidrógeno. El símbolo σ_i es el grado de distancia del vértice i-ésimo (suma de las distancias topológicas desde el átomo considerado a cualquier otro átomo), W es el índice de Wiener, d_{ij} es la distancia topológica entre los átomos *i* y *j*.

Además, hay dos criterios básicos en varios índices de información. El primero es el criterio de la igualdad, lo que implica que los elementos se consideran equivalentes si sus valores son iguales (de acuerdo con este criterio n_g es el número de elementos equivalentes, *n* el número total de elementos y la suma se ejecuta sobre todas las clases de equivalencia). El segundo criterio es de magnitud, donde cada elemento se considera como una clase de equivalencia cuya cardinalidad; es decir, el número de elementos es igual a la magnitud del elemento. Los índices de información calculados por el software DRAGON son los siguientes:

- Contenido total de información. El contenido total de información representa la información residual contenida en el sistema después de las relaciones G que se definen entre los *n* elementos (Shannon y Weaver, 1949).

$$I = n\log_2 n - \sum_{g=1}^{G} n_g \log_2 n_g \qquad \textbf{(15)}$$

- El índice descrito como la media del contenido de información también llamado entropía de Shannon (Shannon y Weaver, 1949)

$$\bar{I} = -\sum_{g=1}^{G} \frac{n_g}{n} \log_2 \frac{n_g}{n} \quad \textbf{(16)}$$

- El índice de información sobre el tamaño molecular (ISIZ) se calcula como el contenido total de información sobre el número de átomos (Bertz, 1981)

$$ISIZ = nAT.\, \log_2 nAT \quad \textbf{(17)}$$

- El índice de información total en la composición atómica (IAC) se calcula como el contenido total de información. Las relaciones de equivalencia se basan en tipos de átomos químicos (Dancoff y Quastler, 1953).

$$I = n \log_2 n - \sum_{g=1}^{G} n_g \log_2 n_g \quad \textbf{(18)}$$

- El índice de información medio en la composición atómica (AAC) se calcula como la media del contenido de información.

$$\bar{I} = -\sum_{g=1}^{G} \frac{n_g}{n} \log_2 \frac{n_g}{n} \quad \textbf{(19)}$$

- El contenido total de información en la igualdad de distancia (IDET) y la media de contenido de información en la igualdad de distancia (IDE) se basan en la igualdad de distancias topológicas en un grafo molecular agotado en Hidrógeno (Bonchev y Trinajstic, 1978).

- El contenido total de información sobre la magnitud de distancia (IDMT) y la media de contenido de información sobre la magnitud de distancia (IDM) se basan en la distribución de las distancias topológicas de acuerdo con su magnitud en un grafo molecular agotado en Hidrógenos, siendo la distancia topológica entre dos átomos el camino más corto que conecta los dos átomos.

- El contenido de información medio sobre el grado de igualdad de distancia (IDDE) se basa en la partición de los grados de distancia de los vértices de acuerdo con su igualdad, siendo el grado de distancia del vértice de un átomo, la suma de las distancias topológicas desde el átomo considerado a cualquier otro átomo en el grafo molecular agotado en Hidrógeno. El contenido de información medio en la magnitud del grado de distancia (IDDM) se basa en la partición de los grados de distancia de los vértices de acuerdo con su magnitud, siendo el grado de distancia del vértice de un átomo, la suma de las distancias topológicas desde el átomo considerado a cualquier otro átomo en el grafo molecular agotado en Hidrógeno.

- La media del contenido de información sobre el grado de igualdad del vértice (IVDE) se basa en la partición de vértices de acuerdo con el grado de igualdad del vértice, siendo el grado del vértice de un átomo, el número de átomos de Hidrógeno no conectados. El contenido medio de información sobre la magnitud de grado del vértice (IVDM) se basa en la partición de vértices de acuerdo con la magnitud de grado del vértice. Este índice fue propuesto como una medida de la complejidad molecular (Raychaudhury, Ray, Ghosh, Roy, y Basak, 1984).

- El índice de complejidad grafo-vértice (HVcpx) (Raychaudhury et al., 1984), se calcula por medio de la siguiente fórmula:

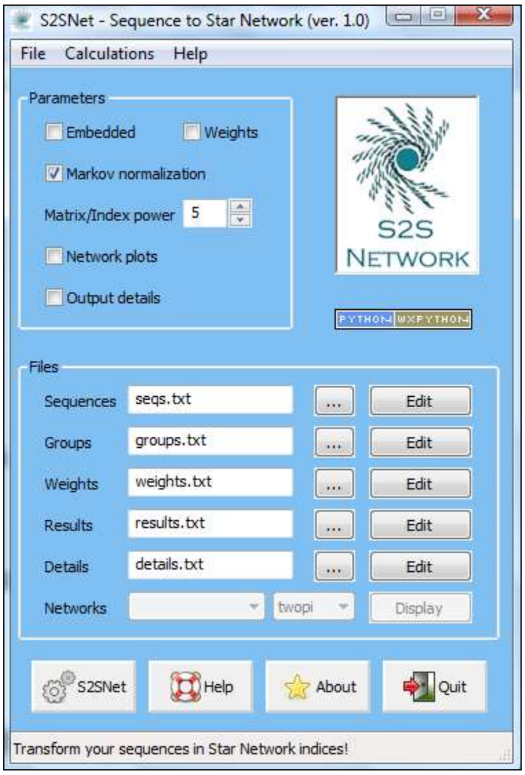
$$HVcpx = \frac{1}{nSK}.\sum_{i=1}^{nSK}\left(-\sum_{g=0}^{\eta_i}\frac{{}^{g}f_j}{nSK}.\log_2\frac{{}^{g}f_j}{nSK}\right) \qquad \textbf{(20)}$$

- El índice de complejidad en la distancia del grafo (Klopman, Raychaudhury, y Henderson, 1988; Raychaudhury et al., 1984), se calcula de la siguiente forma:

$$HDcpx = \sum_{i=1}^{nSK}\frac{\sigma_i}{2W}.\left(-\sum_{j=1}^{nSK}\frac{d_{ij}}{\sigma_i}.\log_2\frac{d_{ij}}{\sigma_i}\right) \qquad \textbf{(21)}$$

2.8.2 Software S2SNet

S2SNet transforma secuencias de caracteres en índices topológicos (TIs) de redes complejas de tipo estrella (*Star Network*, SN). Con estos índices se pueden realizar diversos análisis estadísticos o crear modelos QSAR. Las cadenas de aminoácidos de las proteínas y los ácidos nucleicos son algunos ejemplos de secuencias. El software S2SNet (Figura 11), se puede utilizar para estudiar distintos sistemas, desde sistemas de átomos simples en pequeñas moléculas anti-cancerígenas, hasta sistemas complejos de redes metabólicas, sociales, computacionales o sistemas biológicos.



Parameters

□ Embedded □ Weights

☑ Markov normalization

Matrix/Index power 5

□ Network plots

□ Output details

Files

Sequences seqs.txt

Groups groups.txt

Weights weights.txt

Results results.txt

Details details.txt

Networks twopi

Transform your sequences in Star Network indices!

Figura 11. Interfaz del programa S2SNet

Un ejemplo de cálculo es utilizar una secuencia proteica de las dianas del VIH de la base de datos, Protein Data Bank (http://www.rcsb.org/). La S2SNet transforma la secuencia en una lista de índices topológicos específicos para el grafo de tipo estrella y también puede generar las imágenes de los grafos con la ayuda del Graphviz. En la Figura 12, se presentan los resultados para los cálculos de grafos “non-embedded” (grafo situado a la izquierda con *neato*, grafo situado a la derecha con *twopi*). En la Figura 13, se presenta el caso de los grafos “embedded” (grafo situado a la izquierda con *fdp*, grafo situado a la derecha con *twopi*).

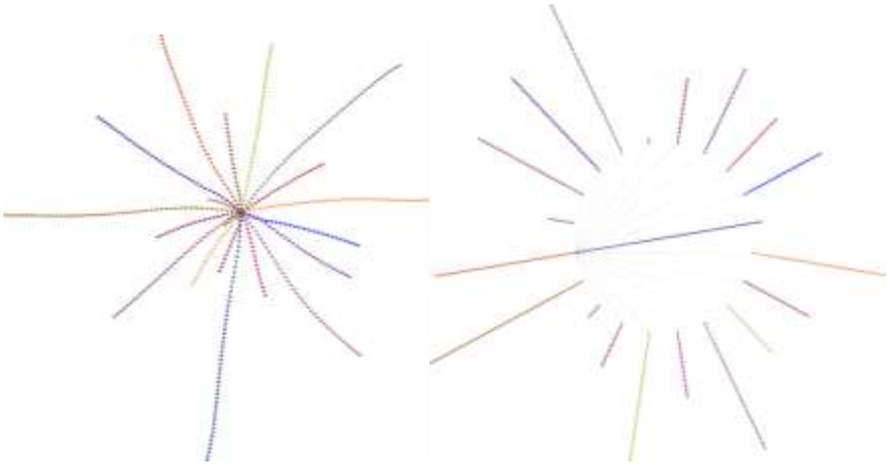


Figura 12. Ejemplo de resultados “non-embedded” con la S2SNet: dibujos de los grafos de tipo estrella

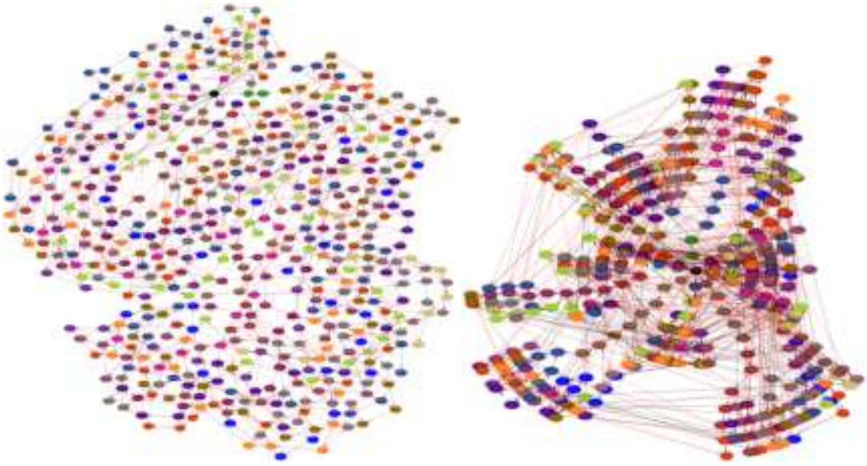


Figura 13. Ejemplo de resultados “embedded” con la S2SNet: dibujos de los grafos de tipo estrella

2.8.2.1 Índices de las redes de tipo estrella

Los datos se utilizarán para calcular los siguientes índices: **Entropía de Shannon** de la n Matrices de Markov (Sh)

$$Sh_n = -\sum_i p_i * log(p_i) \qquad \textbf{(22)}$$

Los p_i son los elementos n_i del vector p resultado desde la multiplicación vectorial entre la matriz Markov normalizada (n_i x n_i) de orden n y el vector (n_i x 1) con cada elemento igual a $1/n_i$.

2.9 Modelos con múltiples entradas, dianas y escalas

Un modelo quimioinformático - farmacoepidemiológico útil debe ser multi-nivel por definición, ya que se espera unir tanto la estructura molecular como la estructura de la población. Esto significa que, con el fin de desarrollar este tipo de modelos computacionales, se necesita procesar diferentes tipos de datos de entrada, procedentes de muchos diferentes niveles de organización de la materia. Por un lado, se tiene que introducir la información sobre los compuestos anti-VIH, que incluye al menos la estructura química del compuesto (I) e información del ensayo preclínico, tales como dianas biológicas (nivel II), organismos (nivel III), o protocolos de ensayo (nivel IV). Por otro lado, se necesitan incorporar descriptores de la estructura de la población (nivel V) que cuantifican los factores epidemiológicos, sociales y económicos que afectan a la población seleccionada para el estudio. Además, como las poblaciones en la sociedad actual no son sistemas cerrados también se debería cuantificar el efecto de la interacción de la población en estudio con otras poblaciones que pueden influir en el estudio farmacoepidemiológico (nivel VI).

Los datos para los niveles de I-IV se obtuvieron de bases de datos públicas de actividad biológica de los compuestos orgánicos. Estas bases de datos acumulan inmensos conjuntos de datos de resultados experimentales de ensayos farmacológicos para muchos compuestos. Por ejemplo, ChEMBL (https://www.ebi.ac.uk/chembl/) es uno de los más grandes con más de 11.420.000 datos de actividad para más de 1.295.500 compuestos y 9844 dianas. Asimismo, se obtuvieron los datos para los niveles V y VI de bases de datos epidemiológicas de acceso público. Por ejemplo, AIDSVu (http://aidsvu.org/about-aidsvu/) es una base de datos detallada de la prevalencia del VIH en los EE.UU a disposición del público.

Se puede hablar de tres características del problema resultante de la unión de información química, farmacológica y epidemiológica: (1) multi-diana, (2) multi-objetivo y/o multi-escala. La naturaleza multi-diana del problema se refiere a la existencia de compuestos que pueden interactuar con más de una diana molecular o celular. El problema de optimización multi-objetivo (MOOP) se refiere a la necesidad de predicción/optimización de resultados para diferentes medidas experimentales obtenidas en diferentes ensayos farmacológicos. Por último, multi-escala se refiere a los diferentes niveles estructurales de organización de la materia (variables de entrada). Esto quiere decir que se necesitan desarrollar modelos capaces de vincular los cambios en la prevalencia del SIDA en una determinada población, con cambios en la actividad biológica del fármaco, debido a variaciones en la estructura química, detectadas en ensayos preclínicos llevados a cabo bajo un conjunto de condiciones experimentales (medida experimental, dianas, organismo, etc.).

2.9.1 Operadores de medias móviles de Box y Jenkins

La codificación de la estructura química de los compuestos es el primer paso. Los datos se basan en un gran número de ensayos desarrollados en condiciones (c_j) muy diferentes para dianas iguales o diferentes (moleculares o no). La información de carácter no estructural se refiere a las diferentes condiciones de ensayo (c_j), tales como concentraciones, temperatura, dianas, organismos, etc. Una solución puede estar en el uso de operadores de medias móviles (MA) (Tenorio-Borroto et al., 2013). Muchos autores han desarrollado MA autorregresivas integradas (ARIMA) y otros modelos MA basados en el trabajo inicial de Box y Jenkins (Box y Jenkins, 1970). Algunos trabajos donde se reporta el uso de MA son los siguientes: Modelos ARIMA para determinar los efectos de la intervención hipnótica en su estudio sobre el control del dolor relacionado con el VIH/SIDA (Langenfeld, Cipani, y Borckardt, 2002). Así como, el análisis de MA para la predicción de la actividad anti-VIH utilizando un nuevo descriptor topológico: el índice de adyacencia excéntrico, concluyendo que el índice propuesto ofrece un gran potencial para los estudios de estructura-actividad/propiedad (Gupta, Singh, y Madan, 2001).

El estudio del momento y la magnitud de la tendencia de los casos de tuberculosis en los Estados Unidos (Chen, Shang, Winston, y Becerra, 2012), utilizando una combinación de ARIMA y métodos bayesianos, resume las ventajas de estos métodos para la estimación e interpretación de los datos operacionales en materia de Salud Pública u otras áreas. Gupta y Madan, estudiaron el desarrollo de modelos para predecir la actividad inhibidora de la integrasa del VIH mediante análisis MA (Gupta y Madan, 2012).

Dos tipos de parámetros D^q_k y $<D^q_k>_j$ son necesarios para calcular una MA. La variable D^q_k es uno de los parámetros de entrada de tipo (k), con valores promedio $<D^q_k>_j$ para todos los casos (q-ésimo) medidos en un conjunto de condiciones experimentales (c_j) (no necesariamente un descriptor molecular). La fórmula general de un operador de MA Box-Jenkins es:

$$\Delta D^{q}{}_{kj} = D^{q}{}_{k} - \left\langle D^{q}{}_{k} \right\rangle_{j} \quad \textbf{(23)}$$

2.9.2 Modelos ALMA

El desarrollo de un enfoque similar llamado "Evaluación de los vínculos con las medias móviles" (en adelante, ALMA), utilizando también los operadores MA. Son adaptables a todos los descriptores moleculares y/o grafos invariantes o descriptores para redes complejas. En conformidad con el apartado anterior, se utiliza una terminología similar. Las entradas de un modelo ALMA son los descriptores $D^q{}_k$ de tipo k-ésimo del sistema q-ésimo (compuesto o fármaco d^q en este caso) representado por una matriz **M**. Además, las salidas de un modelo ALMA son los enlaces (L_{aq} = 1 ó L_{aq} = 0) de una red compleja con la matriz Booleana L y formada por diferentes pares de sistemas de entrada. En consecuencia, la ecuación lineal general del modelo utilizando un descriptor genérico o grafo teórico invariante $D^q{}_k$ tiene la siguiente forma general:

$$\begin{aligned} S_{aqj} &= \sum_{k=1}^{k=k\max} e_k \cdot D^{q}{}_{k} + \sum_{k=1}^{k=k\max} \sum_{j=1}^{j=j\max} e_{kj} \cdot \Delta D^{q}{}_{kj} + e_0 \\ &= \sum_{k=1}^{k=k\max} e_k \cdot D^{q}{}_{k} + \sum_{k=1}^{k=k\max} \sum_{j=1}^{j=j\max} e_{kj} \cdot \left(D^{q}{}_{k} - \left\langle D^{q}{}_{k} \right\rangle_{j} \right) + e_0 \end{aligned} \quad \textbf{(24)}$$

La variable dependiente de salida es $S_{aqj} = S_{aq}\,(c_j) = S_{aqj}\,(c_l, c_2, c_3, ...\ c_{max})$. Esta variable es una puntuación (score) numérica de la formación de enlaces (L_{aq} = 1 ó L_{aq} = 0) en la red compleja que se predijo. En el caso particular de las redes de fármaco-diana es la puntuación de la actividad biológica de la droga q-ésima *vs.* la diana a-ésima medida en un ensayo llevado a cabo en el marco del conjunto de condiciones. Se han publicado diferentes documentos con esta metodología estudiando las relaciones cuantitativas estructura-toxicidad (QSTR) de fármacos con modelos de entropía para puntos finales de interacción fármaco-diana (Tenorio-Borroto et al., 2013). Igualmente, se han introducido nuevos modelos QSAR para evaluar la neurotoxicidad/neuroprotección de fármacos para el tratamiento de enfermedades neurodegenerativas (Alonso et al., 2013; Luan et al., 2013) . Otros autores han utilizado la misma metodología ALMA, en modelos multi-objetivo con descriptores moleculares de diferentes tipos (Speck-Planche, Kleandrova, y Cordeiro, 2013; Speck-Planche, Kleandrova, Luan, y Cordeiro, 2012b).

Se han desarrollado modelos QSAR multitarea para la predicción simultánea de actividad anti-estreptococo y toxicidad de los compuestos (Speck-Planche et al., 2013), acercamientos multi-diana para la predicción de fármacos contra diferentes clases de cáncer (Speck-Planche, Kleandrova, Luan, y Cordeiro, 2011, 2012a). También propusieron modelos ALMA de inhibidores multi-diana contra diferentes proteínas asociadas con la enfermedad de Alzheimer y la Tuberculosis (Speck-Planche, Kleandrova, Luan, y Cordeiro, 2012c, 2012d).

En el presente contexto se pueden usar MA de propiedades de nodos de redes complejas (fármacos, proteínas, organismos, condados, etc.) para predecir la variable $L_{aq}(c_j)_{obs}$ en un conjunto específico de condiciones de ensayo (c_j). Así, c_1 = es la medida experimental, c_2 = es la proteína diana, c_3 = es el organismo que expresan la diana, y c_4 = es el protocolo de ensayo per se. La variable $L_{aq}(c_j)_{obs}$ cuantifica la formación de enlaces entre los nodos. Existen dos tipos de nodos formando la red específica. El primer nodo representa los condados de EE.UU. y el segundo tipo de nodo representa los compuestos. El valor es $L_{aq}(c_j) = 1$ cuando la proporción fármaco-enfermedad (Drug-Disease Ratio) = $DDR_{aq}(c_j) >$ cutoff y $L_{aq}(c_j)_{obs} = 0$ de la otra forma. La fórmula general para un modelo linear usando índices de información de Balaban es la siguiente: (González-Díaz et al., 2014):

$$\begin{aligned} S_{aqj} = & \sum_{k=1}^{k=4} e_k \cdot I^q{}_k + \sum_{k=1}^{k=4}\sum_{j=1}^{j=4} e_{kj} \cdot \Delta I^q{}_{kj} \\ & + e_{ak} \cdot I^a{}_k(t) + e_0 \\ & = \sum_{k=1}^{k=4} e_k \cdot I^q{}_k + \sum_{k=1}^{k=4}\sum_{j=1}^{j=4} e_{kj} \cdot \left(I^q{}_k - \left\langle I^q{}_k \right\rangle_j \right) \\ & + e_{ak} \cdot I^a{}_k(t) + e_0 \end{aligned} \tag{25}$$

El lector debe notar que la variable dependiente, output o predicha S_{aqj} no es una variable discreta, pero si una puntuación numérica de valor real. Sin embargo, la variable es directamente proporcional a la variable observada (L_{aq}). Con el fin de buscar los coeficientes del modelo se puede utilizar una técnica de clasificación lineal como son las RRNNAA disponibles en el paquete de software STATISTICA 6.0 (StatSoft.Inc., 2002). Los parámetros estadísticos utilizados para apoyar el modelo fueron: Número de casos en entrenamiento (N), y los valores globales de sensibilidad, especificidad y precisión (Hill y Lewicki, 2006). En cualquier caso, para calcular la MA se tiene que sumar los valores de I^q_k o todos los compuestos n_j con condiciones de ensayo c_j. A continuación, se divide esta suma por el número de compuestos n_j con esta condición.

$$\Delta I^q{}_{kj} = I^q{}_k - \left\langle I^q{}_k \right\rangle_j \qquad \textbf{(26)}$$

$$\left\langle I^q{}_k \right\rangle_j = \frac{1}{n_j} \sum_{q=1}^{q=n_j} I^q{}_k \qquad \textbf{(27)}$$

2.10 Representaciones generales de redes complejas

Se denomina grafo G al par de conjuntos V(G) y E(G), donde V(G) es un conjunto de vértices o nodos unidos por enlaces llamados aristas E(G). Se dice que un grafo es bipartito si V está compuesto por dos subconjuntos V_A y V_B tal que cada arista conecta un nodo de V_A con un nodo de V_B. En algunos casos, es necesario asignar un sentido a las aristas; por ejemplo, si se quiere representar la red de las calles de una ciudad con sus direcciones únicas (Diestel, 2000). En este caso se habla de un grafo dirigido y se usa el término arco para referirse a las aristas. En un grafo dirigido $a_{ij} \neq a_{ji}$ dos conceptos a tener en cuenta son los de: adyacencia e incidencia.

La matriz de adyacencia se trata de una matriz cuadrada de *n* filas por *n* columnas (siendo *n* el número de vértices del grafo). Para construir la matriz de adyacencia, cada elemento a_{ij} vale 1 cuando haya una arista que una los vértices i y j. En caso contrario, el elemento a_{ij} vale 0. La matriz de adyacencia, por tanto, estará formada por ceros y unos. Adyacencia es la relación entre dos vértices. En el caso de un grafo no dirigido, si dos de ellos están conectados se dice que son adyacentes (la relación de adyacencia es simétrica), mientras que en el caso de un grafo dirigido, si v_1 se conecta con v_2, se dice que v_2 es adyacente a v_1. Así, si en un grafo todos los vértices son adyacentes entre si se dice que es completo.

Incidencia se refiere a la relación que se establece entre un vértice y una arista. En un grafo dirigido de v_1 a v_2 se dice que el arco es incidente positivo con respecto a v_1 (sale de v_1)

y que es incidente negativo respecto a v_2 (llega a v_2). Al número de aristas incidentes sobre un vértice se le conoce como grado y, en el caso de los grafos dirigidos, se divide en dos tipos: el grado positivo, que es el número de arcos que parten del vértice, y el grado negativo, que es el número de arcos que llegan al vértice. La distancia entre dos nodos se define como el número mínimo de aristas que los separa y la excentricidad como la distancia máxima que se puede recorrer en *G* partiendo de un vértice determinado. El diámetro de *G* se corresponde con la máxima excentricidad y el radio con la mínima. La densidad se define como el número de aristas del grafo dividido entre el número total de aristas posibles.

Generalmente, existen dos formas de representar la información de los grafos (ver Figura 14). La primera de ellas es el uso de matrices, que pueden ser:

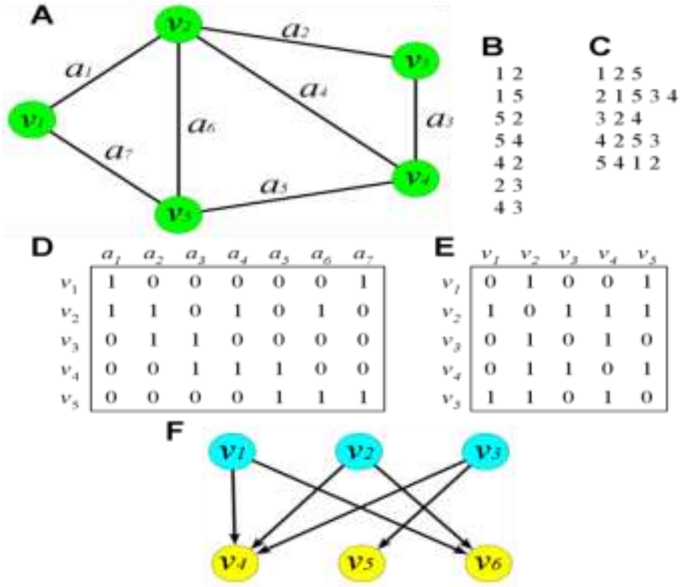
a) Matrices de incidencia: El grafo (*G*) está representado por una matriz de *N* vértices por *K* aristas, donde el par i-j (arista-vértice) contiene la información de la arista (1 - conectado, 0 - no conectado).
b) Matrices de adyacencia: *G* está representado por una matriz cuadrada de tamaño N^2, donde *N* es el número de vértices. Si hay una arista entre un vértice i y un vértice j, entonces el elemento a_{ij} es 1, de lo contrario, es 0.

La segunda es el uso de listas, que pueden ser:

a) Listas de incidencia: Se utiliza una lista de pares de vértices (ordenados si el grafo es dirigido), donde cada par representa una de las aristas.
b) Listas de adyacencia: Cada vértice tiene una lista de vértices que son adyacentes a él. Esto causa redundancia en un grafo no dirigido (ya que si v_1 y v_2 están conectados v_1 existe en la lista de adyacencia de v_2 y viceversa).

Desde el punto de vista del manejo de los datos, las listas son preferidas en grafos dispersos porque tienen un uso de la memoria más eficiente. Las matrices permiten un acceso rápido a los datos, pero pueden consumir grandes cantidades de memoria. En muchos casos, los términos grafo y red se usan indistintamente, aunque no significan exactamente lo mismo. Cuando se habla de grafo se refiere a un objeto matemático, mientras que cuando se usa el término red se refiere a un sistema real (como, por ejemplo, una red de ordenadores o una red social) en el que los vértices se corresponden con las entidades reales que se pretenden representar (ordenadores o personas) y las aristas con las relaciones de distinta naturaleza que se establecen entre ellos (conexiones por cable, relaciones de amistad, etc.).

Durante las dos últimas décadas, el avance en el conocimiento de distintos sistemas complejos, la creación de bases de datos disponibles para la comunidad científica y el desarrollo de herramientas informáticas capaces de manejar de forma eficiente grandes cantidades de datos ha permitido caracterizar de forma eficiente la estructura de redes complejas de sistemas muy diversos. Así, se ha observado que las características de muchas redes reales son diferentes de las que presentan las redes totalmente regulares (en las que todos sus vértices presentan el mismo grado) o totalmente aleatorias usadas normalmente como modelos en el ámbito de la teoría de grafos.



B

1 2
1 5
5 2
5 4
4 2
2 3
4 3

C

1 2 5
2 1 5 3 4
3 2 4
4 2 5 3
5 4 1 2

D

	a_1	a_2	a_3	a_4	a_5	a_6	a_7
v_1	1	0	0	0	0	0	1
v_2	1	1	0	1	0	1	0
v_3	0	1	1	0	0	0	0
v_4	0	0	1	1	1	0	0
v_5	0	0	0	0	1	1	1

E

	v_1	v_2	v_3	v_4	v_5
v_1	0	1	0	0	1
v_2	1	0	1	1	1
v_3	0	1	0	1	0
v_4	0	1	1	0	1
v_5	1	1	0	1	0

Figura 14. A: Grafo no dirigido. *v:* vértices, a: aristas. B: Lista de incidencia. C: Lista de adyacencia. D: Matriz de incidencia. E: Matriz de adyacencia. F: Grafo bipartito y dirigido.

En concreto, el estudio de varias redes reales de naturaleza muy distinta y de sus propiedades ha llevado a proponer dos tipos de redes:

a) Redes de mundo pequeño (*Small world networks*):

Se trata de redes que presentan un coeficiente de agrupamiento elevado (este coeficiente cuantifica la interconexión o agrupamiento de un nodo con sus vecinos) y una distancia media entre nodos relativamente pequeña. Las primeras redes de mundo pequeño fueron empleadas en experimentos de sociología (Amaral, Scala, Barthelemy, y Stanley, 2000). De forma especial, se pueden destacar los experimentos llevados a cabo por Milgram en la década de los 60, en los que se intentaba averiguar el número de conocidos que separan a dos personas determinadas (Milgram, 1967). Estos experimentos dieron lugar a los conceptos de “mundo pequeño” y de los “seis grados de separación” (dos personas cualesquiera están separadas por una cadena de conocidos de distancia media 6). Posteriormente Watts y Strogatz, estudiaron la red de neuronas del nemátodo *Caenorhabditis elegans,* la red de colaboraciones de actores en películas y la red eléctrica del oeste de EE.UU., demostrando que estas redes presentan una estructura de mundo pequeño (es en este estudio donde se acuña el término “red de mundo pequeño”) y que dicha estructura se encuentra entre los extremos de lo totalmente regular y lo totalmente aleatorio (Watts y Strogatz, 1998). A partir de este estudio, se llevaron a cabo una gran cantidad de trabajos encaminados a comprender este tipo de redes y sus propiedades.

b) Redes libres de escala (*Scale free networks*):

Las redes libres de escala se caracterizan porque en ellas algunos vértices están altamente conectados (a estos nodos con grado alto se les denomina *hubs*), aunque en general el grado de conexión de casi todos los vértices de la red es bastante bajo (Barabasi y Bonabeau, 2003) (Figura 15). En general, se puede decir que la probabilidad de que un vértice de la red esté conectado con *k* vértices *P(k)* es proporcional a $k^{-\gamma}$, es decir sigue una ley de potencias. El exponente γ es específico de la red estudiada y, generalmente, su valor se encuentra en el rango $2 < \gamma \leq 3$. El interés por el estudio de redes con estas características comenzó con los trabajos de Barabasi y colaboradores (Barabasi y Albert, 1999), acerca de la topología de internet y ha continuado con el estudio de otras redes complejas pertenecientes a distintos campos, como por ejemplo la red de colaboraciones de actores, las redes de aerolíneas de estados unidos, las redes metabólicas, las redes de relaciones sexuales, etc. Es importante aclarar que una red puede ser de mundo pequeño y libre de escala a la vez.

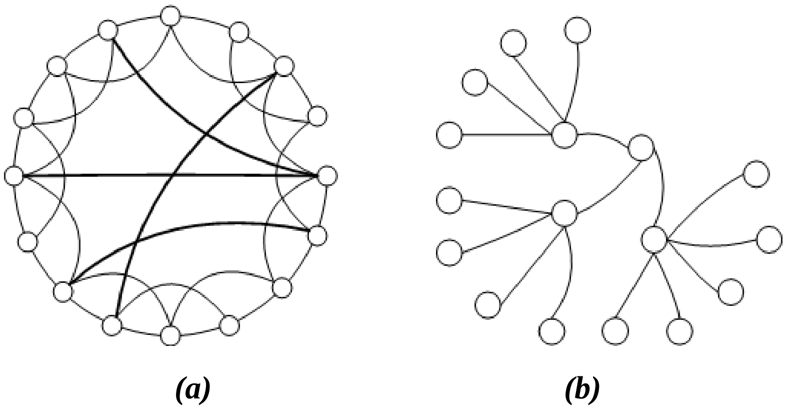


Figura 15. Ejemplos de redes complejas. (a): Red de mundo pequeño. (b): Red libre de escala.

2.10.1 Representación de redes complejas en Epidemiología

Las redes complejas describen un amplio rango de sistemas en la naturaleza y en la sociedad. La ciencia de las redes ha revolucionado la investigación sobre la dinámica de los elementos que interactúan. Se podría argumentar que la epidemiología, en particular, ha abrazado el potencial de la teoría de redes más que cualquier otra disciplina. Existe una relación muy estrecha entre la epidemiología y la teoría de las redes que se remonta a mediados de la década de 1980. Esto se debe a las conexiones de una red de individuos (o grupos de individuos) que permiten que una enfermedad infecciosa se propague naturalmente, mientras que la red que se genera proporciona información detallada sobre la dinámica epidemiológica. En particular, la comprensión de la estructura de la red de transmisión nos permite mejorar las predicciones de la probable distribución de la infección y el crecimiento temprano de la infección. Sin embargo, la interacción entre las redes y la epidemiología va más allá; porque la red define las posibles vías de transmisión, así como el conocimiento de su estructura se puede utilizar como parte de control de la enfermedad.

Los modelos de redes proveen una vía natural para la descripción de las poblaciones y sus interacciones. Los nodos del grafo representa a los individuos y los links describen interacciones entre individuos que podrían potencialmente conducir a la transmisión de la infección. Representaciones similares de redes en un numero de contextos, tales como redes de transporte, redes de comunicación incluyendo el internet, y redes sociales pueden ser observadas. El marco epidemiológico donde las redes tienen la historia más larga de uso implican las infecciones de transmisión sexual (ITS), como la gonorrea o el virus de la inmunodeficiencia humana (VIH) (Bauch y Rand, 2000; Ghani y Garnett, 2000). Aquí hay estructuras de redes naturales, bien definidas (redes de compañeros sexuales), que durante mucho tiempo han sido explotados por los organismos de Salud Pública en sus intentos de rastrear y controlar los brotes de enfermedades de transmisión sexual. Más recientemente, se han empleado modelos de redes para describir la propagación de una gama más amplia de infecciones como el sarampión, el SARS o la Fiebre Aftosa (FA) (Ferguson, Donnelly, y Anderson, 2001; Meyers, Pourbohloul, Newman, Skowronski, y Brunham, 2005)

Un estudio acerca de la reconstrucción de redes de transmisión del VIH-1 analizan las redes inferidas (Zarrabi et al., 2012). La Figura 16 muestra la red para toda la población que se compone de tres sub-redes correspondientes a los grupos de mayor riesgo de transmisión del VIH-1 (hombres que tienen sexo con hombres MSM, heterosexual, usuarios de drogas inyectables UDI). Había unos pocos pacientes con modo de infección por "productos de la sangre" que fueron aislados de otros grupos de riesgo. Se analizó el grado de distribución de la red como un todo (es decir, para todos los grupos de riesgo) y el grado de distribución de cada sub-red por separado. In-degree es el número de aristas entrantes a un nodo y out-degree es el número de bordes salientes de un nodo. El grado total es la suma de las entradas y salidas de grados. Sin embargo, desde las distribuciones se puede ver que el grado de pacientes altamente conectados en IDU es significativamente más alto que los de MSM y los pacientes adquiriendo la infección por contacto heterosexual.

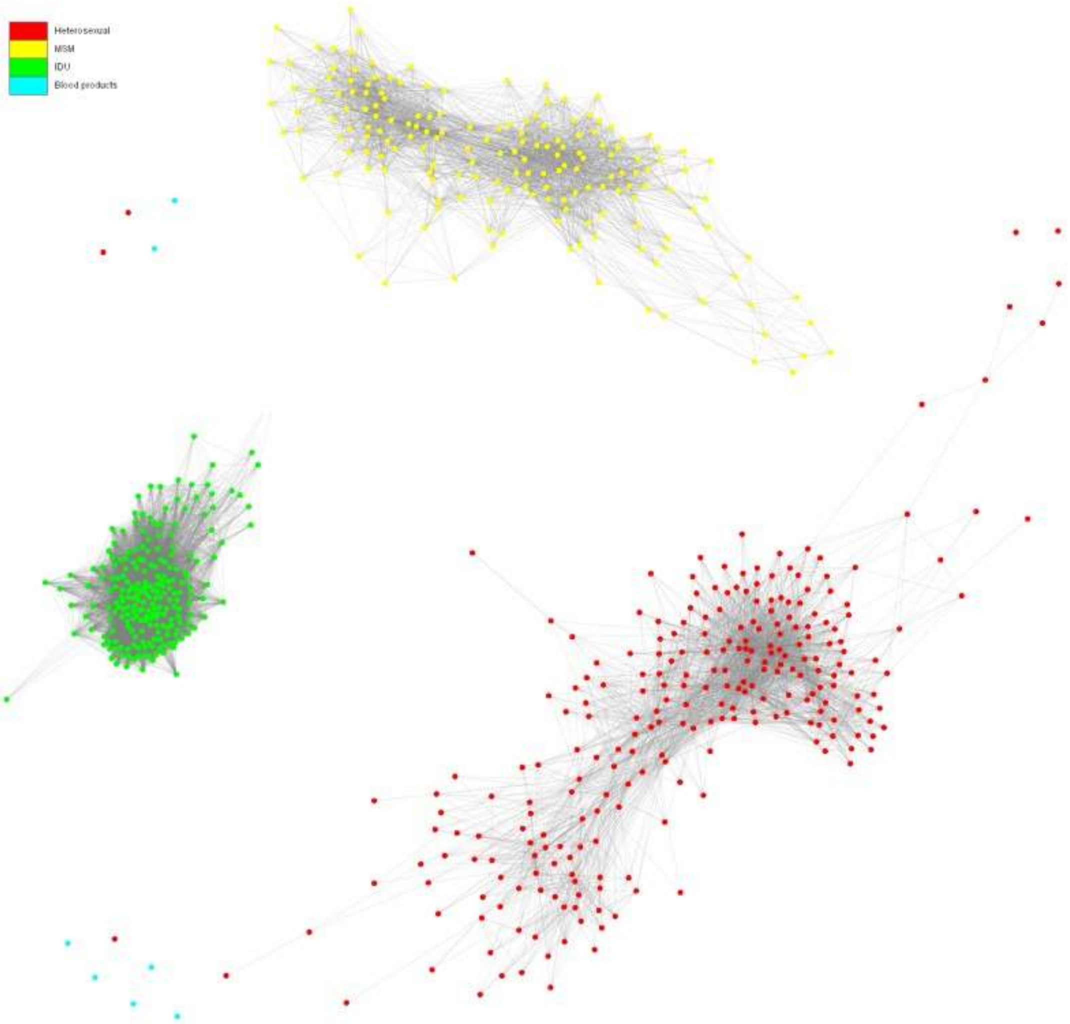


Figura 16. Red de contacto de los mayores grupos de riesgo de transmisión de VIH. MSM (amarillo), Heterosexual (rojo), y UDI (verde), productos de la sangre (azul).

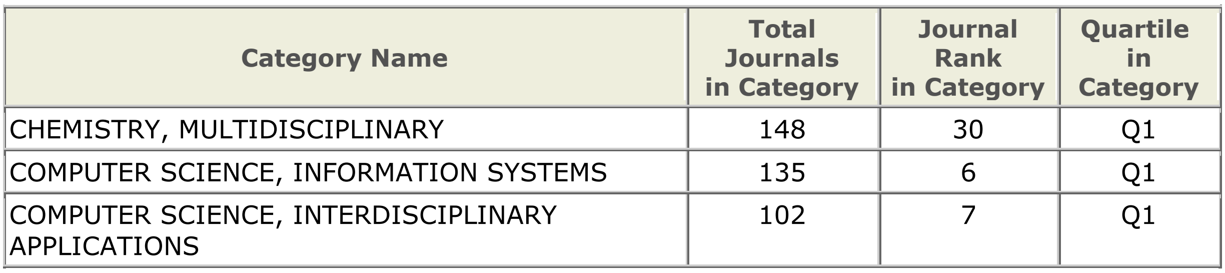
3. TRABAJO EXPERIMENTAL

3.1 Resultados y discusión artículo 1

González-Díaz H, **Herrera-Ibatá DM**, Duardo-Sánchez A, Munteanu CR, Orbegozo-Medina RA, Pazos A. ANN Multiscale Model of Anti-HIV Drugs Activity vs AIDS Prevalence in the US at County Level Based on Information Indices of Molecular Graphs and Social Networks. Journal of Chemical Information and Modeling 54 (3) 2014, 744-755.

Journal of Chemical Information and Modeling: factor de impacto 4.068

Tabla 2. Cuartiles Journal of Chemical Information and Modeling



Category Name	Total Journals in Category	Journal Rank in Category	Quartile in Category
CHEMISTRY, MULTIDISCIPLINARY	148	30	Q1
COMPUTER SCIENCE, INFORMATION SYSTEMS	135	6	Q1
COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS	102	7	Q1

Este trabajo (González-Díaz et al., 2014), tiene como objetivo describir el flujo de trabajo para una metodología que combina quimioinformática y farmacoepidemiología, y reportando el primer modelo predictivo basado en técnicas de inteligencia artificial como son las RRNNAA. El nuevo modelo es capaz de predecir redes complejas de la prevalencia del SIDA en los condados de Estados Unidos, teniendo en consideración las determinantes sociales y la estructura/actividad de compuestos contra el VIH en ensayos preclínicos.

Se construyó el primer modelo ALMA para estudios multiescala del VIH/SIDA en EE.UU. a nivel de condado. Las entradas del modelo fueron los índices de información de Balaban (I^{q}_{k}) de un compuesto dado para cuantificar los cambios en la estructura química de los mismos, y los índices de información de Shannon (Godden, Stahura, y Bajorath, 2000) para la población (a-ésima condado) explicados en el apartado de fundamentos teóricos. Se utilizaron los índices de información de Shannon para describir la información de la red social (la desigualdad de ingresos caracterizada por el coeficiente de Gini) (Haidich y

Ioannidis, 2004). Los datos de la prevalencia del SIDA y el coeficiente de Gini a nivel de condado se obtuvieron de la base de datos AIDSVu. El conjunto de datos utilizados para entrenar y validar el modelo incluye N= 43.249 casos estadísticos. El conjunto de datos incluye los valores de prevalencia de SIDA en 2310 condados de Estados Unidos vs. resultados del ChEMBL para 21.582 compuestos únicos, 9 dianas proteicas virales o humanas, 4856 protocolos, y 10 posibles medidas experimentales. Se entrenaron diferentes topologías de RRNNAA incluyendo perceptrones multicapa (MLP) y redes de neuronas artificiales lineales (LNN). El mejor modelo fue la LNN con precisión (Ac), especificidad (Sp), y sensibilidad (Sn) por encima del 75%. En la Figura 17, se muestra el flujo de trabajo utilizado en el presente trabajo para construir los modelos ALMA para este problema. La participación del doctorando en este estudio, y los pasos más importantes para las ramas A y B son los siguientes:

a.1 Recopilación de la estructura química e información de la actividad biológica de los compuestos con potencial actividad antirretroviral de la base de datos pública ChEMBL, para la creación de la base de datos objeto de estudio en esta tesis doctoral.

a.2 Procesar la información de las moléculas por medio de los códigos SMILES (Simplified Molecular Input Line Entry Specification o especificación de introducción lineal molecular simplificado) ya que las representaciones moleculares pueden traducirse a este lenguaje. Su objetivo es obtener un código de información muy detallada respecto a composición y estructura molecular que se adapte a diferentes motores de búsqueda, programas y lenguajes informáticos. Estos códigos son una especificación para describir sin ambigüedades la estructura de una molécula usando cadenas ASCII cortas (código de caracteres basado en el alfabeto latino). Los códigos SMILES se introdujeron en el software DRAGON (Todeschini et al., 2005), para calcular los diferentes índices de información de las moléculas. El software DRAGON es una aplicación para el cálculo de descriptores moleculares que se pueden utilizar para evaluar relaciones moleculares estructura-actividad o estructura-propiedad, así como para el análisis de similitud y la selección de alto rendimiento de las bases de datos de moléculas. En realidad DRAGON es ampliamente utilizado en estudios científicos, así como parte de varios proyectos QSAR http://www.talete.mi.it/products/dragon_projects.htm.

a.3. Posteriormente, a través de Excel, se realizó el cálculo de los operadores MA para las moléculas teniendo en cuenta las condiciones de ensayo (medida experimental, proteína, organismo, ensayo).

b.1. Se descargaron los datos de prevalencia del SIDA y el índice de Gini (desigualdad de los ingresos) de EE.UU. de la base de datos AIDSVu (Emory University).

b.2. Se realizó el cálculo del índice de información I^a_0 para la desigualdad de los ingresos en diferentes condados de EE.UU basado en la entropía de Shannon (Ver Fundamentos teóricos).

c. Para concluir, se llevó a cabo la búsqueda de modelos predictivos usando técnicas de inteligencia artificial, como las RRNNAA, por medio del software STATISTICA 6.0 ("STATISTICA," 2001). El programa STATISTICA posee una herramienta llamada “Intelligent Problem Solver” (en adelante, IPS), útil en el entrenamiento y validación de los modelos de predicción. Permite la clasificación automática al indicarle las variables input y output, los casos de distribución asignados, aceptando-rechazando los límites de confianza, las opciones de complejidad de la red, estableciendo los límites de duración del proceso de diseño, y los resultados que serán mostrados.

Con esta herramienta se pueden entrenar simultáneamente varias arquitecturas de redes y calcular las predicciones de cada una. Este método puede requerir tiempos de computación significativos, pero entrena automáticamente y prueba un gran número de diferentes tipos de redes. La herramienta permite el uso de algoritmos de búsqueda para determinar la selección de entradas, el número de capas ocultas, y otros importantes factores en el diseño de la red. Estos algoritmos de búsqueda se intercalan, así que la IPS busca redes óptimas de diferentes tipos simultáneamente (por ejemplo, perceptrones multicapa y funciones de base radial, etc.). El programa genera un nivel de resultados detallado de gran ayuda para escoger la red con mejor rendimiento (gráficas de la arquitectura de la red, resumen estadístico para la red entrenada, análisis de sensibilidad, curva ROC, estadística descriptiva, etc.).

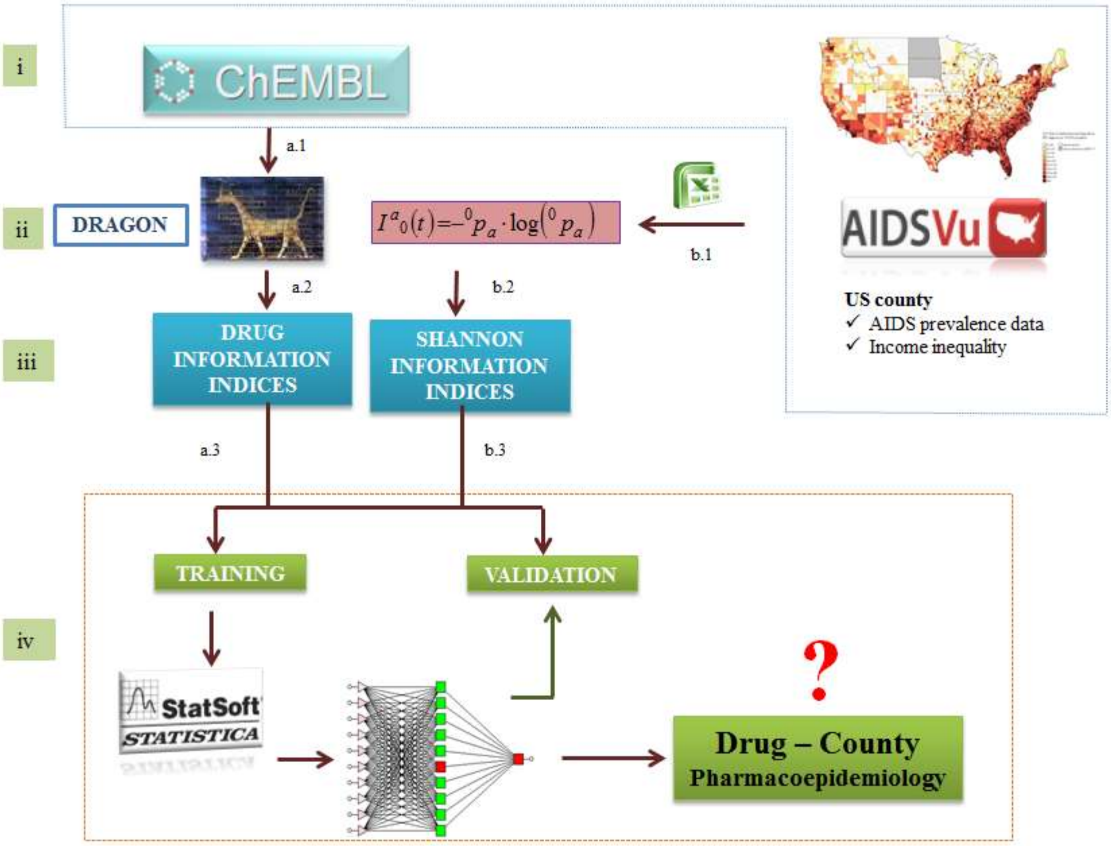


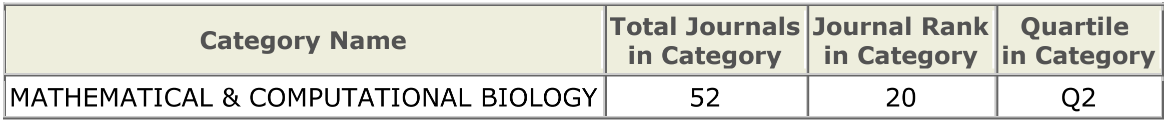
Figura 17. Diagrama de flujo para la construcción de redes de neuronas artificiales para el modelo Farmacoepidemiológico del SIDA en EE.UU.

3.2 Resultados y discusión artículo 2

Herrera-Ibatá DM, Orbegozo-Medina RA, and González-Díaz H. Multiscale Mapping of AIDS in U.S. Counties vs. Anti-HIV Drugs Activity with Complex Networks and Information Indices. Journal Current Bioinformatics. (2015), *en imprenta*.

Journal Current Bioinformatics: factor de impacto 1.726

Tabla 3. Cuartiles Journal Current Bioinformatics



Category Name	Total Journals in Category	Journal Rank in Category	Quartile in Category
MATHEMATICAL & COMPUTATIONAL BIOLOGY	52	20	Q2

En el presente trabajo, se utilizaron los índices de información de simetría ($^{q}IC_{kf}$) calculados con el software DRAGON y los índices Markov-Shannon $I^{a}_{k}(s)$ para caracterizar las diferentes poblaciones. El nuevo modelo ALMA se expresa en la siguiente ecuación:

$$\begin{aligned} S_{aqj} &= \sum_{k=0}^{k=5}\sum_{f=1}^{f=5} e_{kf}\cdot{}^{q}IC_{kf} + \sum_{k=0}^{k=5}\sum_{f=1}^{f=5}\sum_{j=1}^{j=4} e_{kfj}\cdot\Delta{}^{q}IC_{kfj} + \sum_{k=1}^{k=5} e_{ak}\cdot\Delta I^{a}{}_{ks} + e_{0} \\ &= \sum_{k=0}^{k=5}\sum_{f=1}^{f=5} e_{kf}\cdot{}^{q}IC_{kf} + \sum_{k=0}^{k=5}\sum_{f=1}^{f=5}\sum_{j=1}^{j=4} e_{kfj}\cdot\left({}^{q}IC_{kf} - \left\langle {}^{q}IC_{kf}\right\rangle_{j}\right) + \sum_{k=1}^{k=5} e_{k}\cdot\left(I^{a}{}_{k} - \left\langle I^{a}{}_{k}\right\rangle_{s}\right) + e_{0} \end{aligned} \qquad \textbf{(29)}$$

En este trabajo, se obtuvieron los modelos de RRNNAA a partir de índices de información de simetría y cadenas de Markov de la siguiente forma: se tenía la base de datos con las 43.249 moléculas, a cada una de ellas se le calculó los índices de simetría. Estos índices se caracterizan por ser cinco tipos o familias (f): IC_k, TIC_k, SIC_k, BIC_k y $CIC_{k,}$ y, a su vez, cada tipo o familia va de orden (*k*) cero a cinco. Por lo tanto, el número de entradas del modelo consistía en: $N_{entradas} = N_f{\cdot}N_k + N_f{\cdot}N_k\ {\cdot}N_j + {}^{a}N_k = 5{\cdot}6 + 5{\cdot}6{\cdot}4 + 5 = 30 + 120 + 5 = 155$; es decir, 30 índices de simetría ($^{q}IC_{kf}$), 120 MA de acuerdo a las diferentes condiciones de ensayo (c_1,c_2,c_3,c_4), y 5 MA ΔI^{a}_{ks} para los condados de EE.UU. Estos datos se generaron para cada molécula de la base de datos. Los resultados de este primer modelo arrojaron una RNA lineal (LNN) con 155 entradas, la cual clasificó correctamente el 76% de los casos en entrenamiento y validación. La red LNN mostró valores de AUROC = 0,82 en entrenamiento y 0,82 para el conjunto de validación externa. Se puede pensar que el modelo lineal parece mejor que el modelo no-lineal para ajustarse al presente conjunto de datos. Sin embargo, el número de entradas es muy alto para ser considerado un modelo simple. Teniendo en cuenta que el modelo anterior presenta un número de variables muy elevado con respecto al rendimiento obtenido, se decidió entrenar modelos LNN con cada familia de índices por separado; es decir: $N_{entradas} = N_f \cdot N_k + N_f \cdot N_k \cdot N_j + {}^aN_k = 1\cdot 6 + 1\cdot 6\cdot 4 + 5 = 6 + 24 + 5 = 35$.

Todos los modelos LNN para cada familia de índices clasificaron correctamente el 75% de los casos en entrenamiento y validación. Tras el análisis de los resultados basados principalmente en el análisis de sensibilidad generado por el módulo IPS del software STATISTICA 6.0, se decidió poner a prueba la capacidad de predicción de estos índices en un modelo más sencillo. Al hacerlo, se han entrenado los modelos de RNA lineales (LNN) utilizando el orden 5 de cada familia de índices de información (${}^qIC_{5f}$), sus operadores MA y el quinto operador MA de los condados de Estados Unidos. El modelo basado en el tipo de índice IC_k de orden 5, LNN ${}^qIC_{51}$ (LNN-IC51) presentó los valores más altos de Sn = 72.04 / 72.81 y Sp = 72.38 / 72.50 en el entrenamiento y conjuntos de validación externa. La ecuación para este modelo, se muestra a continuación:

$$
\begin{aligned}
S_{aq}(c_j) = \; & -25.48 \cdot {}^qIC_{51} + 1081.64 \cdot \Delta^q IC_{51}(c_1) + 29.36 \cdot \Delta^q IC_{51}(c_2) \\
& -1084.52 \cdot \Delta^q IC_{51}(c_3) - 0.7727 \cdot \Delta^q IC_{51}(c_4) \\
& -0.0792 \cdot \Delta I^a{}_5(s) - 0.5025
\end{aligned}
\qquad \textbf{(30)}
$$

Asimismo, el modelo basado en LNN ${}^qIC_{51,}$ se usó para generar/predecir una red compleja de la prevalencia del SIDA en EE.UU. a nivel de condado, con respecto a la actividad preclínica de compuestos anti-VIH. La red bipartita tiene dos tipos de nodos (condados vs. fármacos). De esta forma, esta es una red multiescala similar a las redes bipartitas de otros grupos (Araujo, Liotta, y Petricoin, 2007; Hecker et al., 2012; Prado-Prado et al., 2011; Prado-Prado et al., 2012; Vina, Uriarte, Orallo, y Gonzalez-Diaz, 2009). Sin embargo, los nodos en la presente red contienen información sobre las moléculas; es decir, la estructura química, así como las condiciones de ensayo (diana, organismo, medida experimental, etc.). Asimismo, el otro conjunto de nodos contiene información sobre factores socioeconómicos, tales como la desigualdad de los ingresos en el condado. Redes multiescala de este tipo han sido discutidas por Barabasi et al., (Barabasi, Gulbahce, y Loscalzo, 2011) como una herramienta importante para realizar investigaciones transdisciplinarias. Los enlaces de esta red compleja son las salidas $L_{aq}(c_j)_{pred} = 1$ de nuestro modelo. En la Figura 18, se ilustra la red de prevalencia del SIDA vs. actividad preclínica de compuestos anti-VIH para el estado de Florida. Por ejemplo, el modelo predice una alta efectividad para el compuesto Zidovudina (Shey, Kongnyuy, Alobwede, y Wiysonge, 2013) en el tratamiento del SIDA en el condado de Nassau.

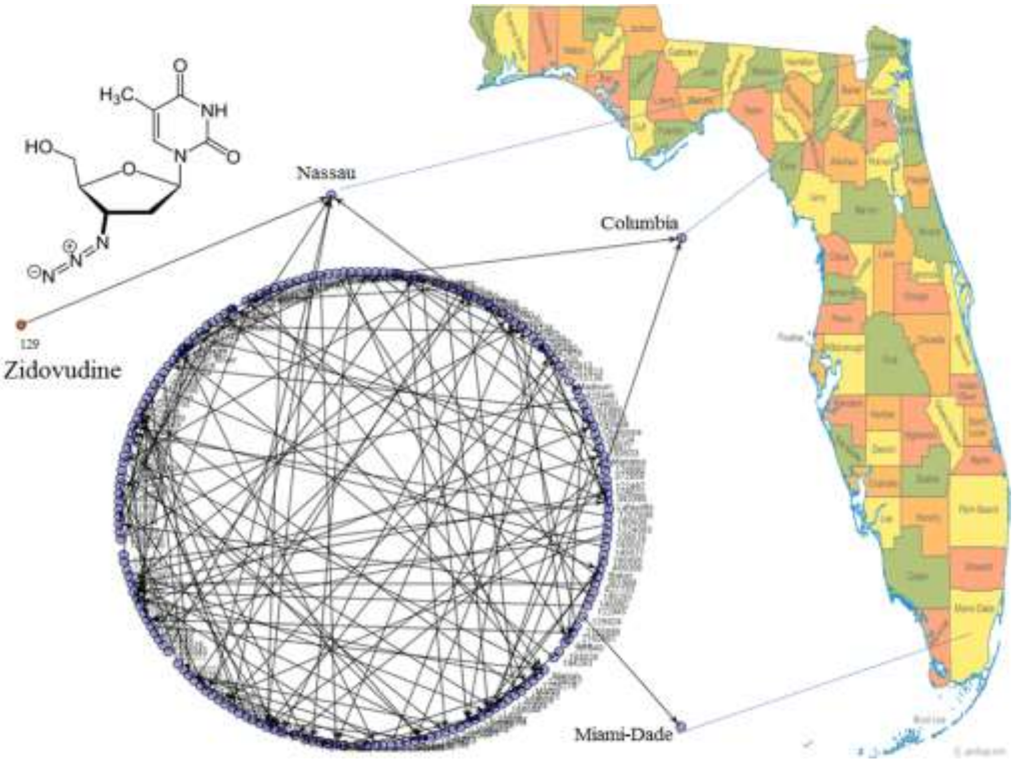


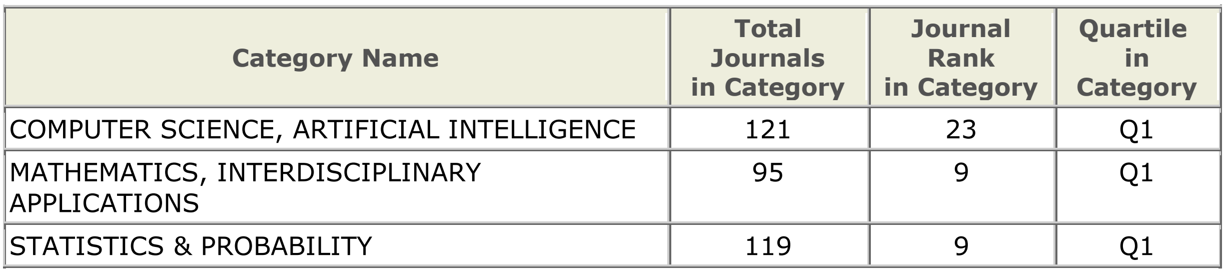
Figura 18. Red multiescala para el estado de Florida, EE.UU.

3.3 Resultados y discusión artículo 3

Herrera-Ibatá DM, Pazos A, Orbegozo-Medina RA, González-Díaz H. Mapping networks of anti-HIV drug cocktails vs. AIDS epidemiology in US Counties. Chemometrics and Intelligent Laboratory Systems. 138 (2014) 161-170.

Chemometrics and Intelligent Laboratory Systems: factor de impacto 2.381

Tabla 4. Cuartiles Chemometrics and Intelligent Laboratory Systems



Category Name	Total Journals in Category	Journal Rank in Category	Quartile in Category
COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE	121	23	Q1
MATHEMATICS, INTERDISCIPLINARY APPLICATIONS	95	9	Q1
STATISTICS & PROBABILITY	119	9	Q1

En este trabajo, se obtuvo el primer modelo útil para mapear el efecto de un cóctel de fármacos antirretroviral frente al SIDA utilizando clasificadores AA-ALMA. Se utilizaron dos algoritmos de aprendizaje automático diferentes: el análisis discriminante lineal (LDA) que trata de encontrar relaciones lineales entre las variables continuas que mejor discriminen en los grupos dados a los objetos, y las redes de neuronas artificiales (RRNNAA). También, se realizó un análisis factorial utilizando dos métodos diferentes para extraer los componentes principales con técnicas de reducción de dimensionalidad. Los métodos utilizados fueron el análisis de componentes principales (PCA) y el método residual mínimo (MINRES). La combinación de estos algoritmos de pre-procesamiento junto con AA resultó en dos técnicas diferentes PCA-LDA y MINRES-LDA, pero con resultados de clasificación alrededor del 50%.

Asimismo, se entrenaron diferentes topologías de RRNNAA incluyendo perceptrones multicapa (MLP) y redes de neuronas lineales (LNN) a través del módulo IPS del software STATISTICA 6.0 utilizando como variables de entrada los índices de información de los compuestos y sus operadores MA, con resultados de clasificación para la red MLP alrededor del 60% y la LNN con 80% pero usando una cantidad elevada de variables de entrada. Posteriormente, se realizó un LDA como estrategia de selección de variables para realizar una selección de las 66 variables de entrada. El modelo LDA presentó 23 variables, una tasa de precisión de 80,39% en el conjunto de entrenamiento y una tasa de precisión de 80,53% en el conjunto de validación externa. Usando las variables seleccionadas por el LDA, se exploró la posibilidad de generar modelos no lineales. Al hacerlo, se ha utilizado la opción implementada en el software STATISTICA: "*custom network designer*" para generar la red MLP, este módulo se utiliza para especificar el tipo y crear RRNNAA individuales. Los detalles de los parámetros para generar los clasificadores se pueden encontrar en la Tabla 4 de la siguiente referencia (Herrera-Ibatá, Pazos, Orbegozo-Medina, y Gonzalez-Diaz, 2014).

A continuación, en la Tabla 5 se describen los resultados de los modelos generados.

Tabla 5. Modelos predictivos generados

Modelos	Perfil modelo		Entrenamiento		Selección		Validación	
		Observado	$L_{ac} = 0$	$L_{ac} = 1$	$L_{ac} = 0$	$L_{ac} = 1$	$L_{ac} = 0$	$L_{ac} = 1$
LDA	66-23-1	Parámetro	Sn	Sp	Sn	Sp	Sn	Sp
		Predicho	83.64	77.15	-	-	83.69	77.37
		$L_{ac}= 0$	67971	2356	-	-	45183	1599
		$L_{ac} = 1$	13292	7959	-	-	8801	5467
MLP	66-26-1	Parámetro	Sn	Sp	Sn	Sp	Sn	Sp
		Predicho	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
LDA-MLP	19-10-1	Parámetro	Sn	Sp	Sn	Sp	Sn	Sp
		Predicho	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
LNN	66-1	Parámetro	Sn	Sp	Sn	Sp	Sn	Sp
		Predicho	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

Adicionalmente, se generó una sub-red compleja del estado de California con los resultados obtenidos para el clasificador LDA-ALMA. La sub-red tiene tres tipos de nodos: fármacos anti-VIH (azul), cócteles (rojo) y condados de EE.UU. Es importante entender que $L_{ac}(b_j)_{pred}$ = 1 expresa la existencia de un subgrafo que conecta varios nodos de todas las clases por medio de varios arcos y no un sólo arco que conecta dos nodos. Por ejemplo, una sub-red sencilla incluyendo solo nodos de fármacos, cócteles y condados. En este caso cuando $L_{ac}(b_j)_{pred}$ = 1 se conecta cada nodo de los compuestos y forman el cóctel y luego con el nodo que representa este cóctel. Por lo tanto, $L_{ac}(b_j)_{pred}= 1$ expresa la existencia del subgrafo $(d^1 \rightarrow c_1)(d^2 \rightarrow c_1)d^3 \rightarrow c_1 \rightarrow a_1$ para todas los fármacos en el cóctel (Figura 19).

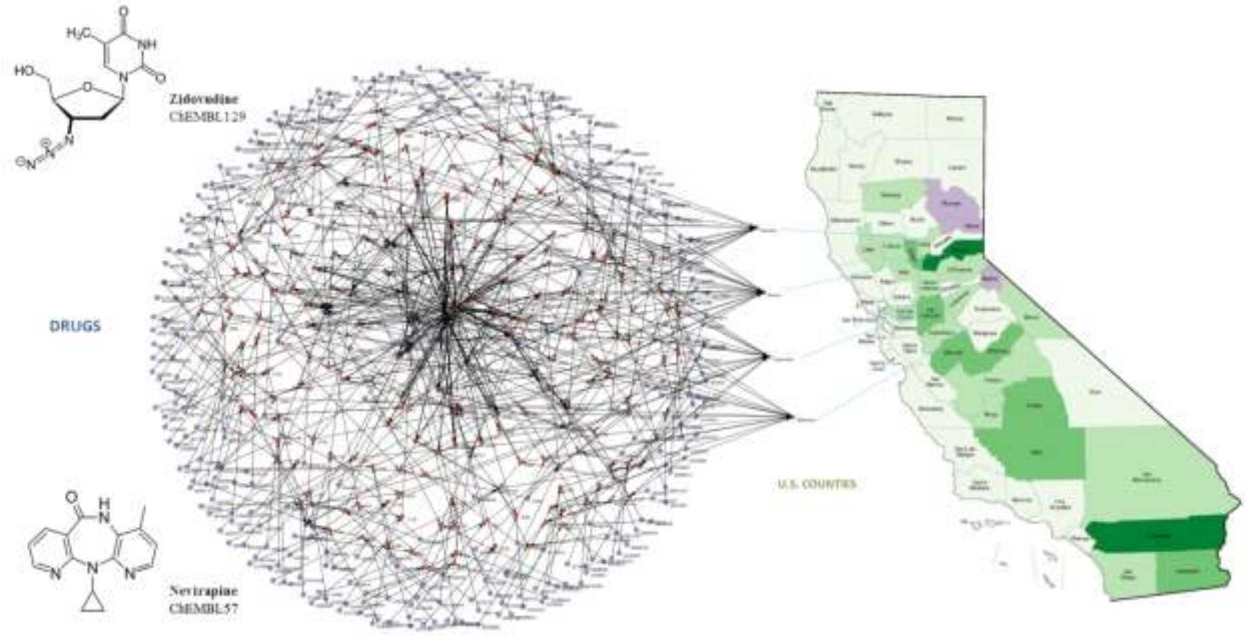


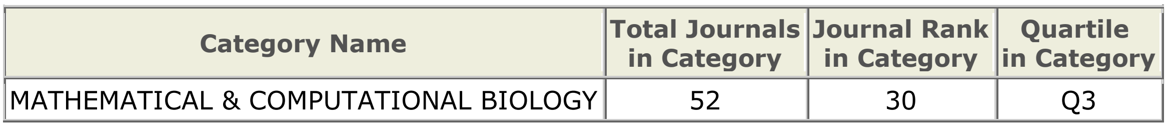
Figura 19. Sub-red de los cocteles anti-VIH para estado de California, EE.UU.

3.4 Resultados y discusión artículo 4

Herrera-Ibatá DM, Pazos A, Orbegozo-Medina RA, Romero-Durán FJ, González-Díaz H. Computational Algorithms for Network Epidemiology: Mapping Structure-Activity Relationships of HAART-Cocktail Drugs over AIDS and Socioeconomic Data of U.S. Counties. Biosystems. *Bajo revisión 2015*. (Ver sección Futuros desarrollos).

Biosystems: factor de impacto 1.472

Tabla 6. Cuartiles Biosystems



Category Name	Total Journals in Category	Journal Rank in Category	Quartile in Category
MATHEMATICAL & COMPUTATIONAL BIOLOGY	52	30	Q3

Se propuso el primer modelo RNA para la predicción de cócteles antirretrovirales para detener el SIDA en redes epidémicas de condados de EE.UU. usando índices de información que codifican varios factores socioeconómicos (educación, ingresos, empleo, migración, pobreza etc.) y factores biomoleculares. Se usaron tres escalas o niveles para agrupar los condados de acuerdo a la ubicación o códigos de la estructura de la población (Estado, RUCC y UIC).

Se analizaron más de 130.000 pares (enlaces de red) que corresponden a la prevalencia del SIDA en 2.310 condados en Estados Unidos vs. cócteles de medicamentos formados por combinaciones de compuestos reportados en el ChEMBL (21.582 medicamentos únicos, 9 dianas de proteínas virales o humanos, 4856 protocolos y 10 posibles medidas experimentales). El modelo generado con los datos originales era una red neuronal lineal (LNN) con AUROC > 0,80 y precisión, especificidad y sensibilidad ≈ 77% en entrenamiento y serie de validación externa. El cambio de la escala espacial y la estructura de la población (Estado, RUCC, UIC) no afectan a la calidad del modelo. Sin embargo, se detectó desequilibrio en todos los modelos comparando casos positivos/negativos y la precisión de los modelo lineales/no lineales. Por lo tanto, se usaron métodos de pre-procesamiento de datos como el algoritmo de sobremuestreo "SMOTE", en el que la clase minoritaria es sobremuestreada creando nuevos casos sintéticos. Lo que lo diferencia de otros algoritmos de sobremuestreo en los que los nuevos casos se generan replicando casos ya existentes. La idea básica es la de generar nuevos casos situados en algún punto de la línea que une cada caso de la clase minoritaria con alguno (o todos) de sus k vecinos más cercanos. Después de aplicar la técnica de sobremuestreo, se usaron algoritmos de aprendizaje automático implementados en el software WEKA (colección de algoritmos de aprendizaje automatico) (Hall et al., 2009), se encontraron modelos más balanceados y precisos. En particular, un MLP con AUROC = 97,4% y precisión > 90%. También se calcularon las probabilidades con las que el SIDA puede ser tratado con varios cócteles de fármacos en condados determinados. Como se puede ver en la fórmula, el condado “in-degree” es el número de vínculos positivos entre los diferentes cócteles y el condado. La frecuencia del condado se refiere al número total de veces que el condado está en la base de datos original, incluyendo los cócteles predichos como negativos. Es importante destacar que no todos los condados fueron modelados contra cada cóctel de drogas.

$$p(halt)=\frac{(county\ in-\deg ree)}{(county\ frequency)}=\left(\frac{county\quad in-\deg ree}{[(county\quad in-\deg ree)+(negativecocktails)]}\right) \qquad \textbf{(31)}$$

Además, en la Tabla 7, se representan algunos ejemplos de la compleja sub-red del estado de Nueva York, con datos de los condados en grado (*in-degree)* con varios cócteles TARGA. Por ejemplo, el condado de Bronx muestra una buena conexión en la compleja sub-red; por ejemplo, la probabilidad de que varios cócteles TARGA trabajen en este condado es más alta que el condado de Chemung, que presenta una menor probabilidad. Por lo tanto, este tipo de modelo podría ser útil para los procedimientos de vigilancia epidemiológica para entender la vulnerabilidad de las poblaciones en relación con la epidemia del SIDA.

Tabla 7. Probabilidades predichas con las cuales el SIDA puede ser detenido en un condado con un cóctel aleatorio

Condado NY	Frecuencia condado	Condado in-degree	p(halt)
Bronx	57	48	0.84
Queens	57	43	0.75
New York	56	41	0.73
Kings	56	39	0.70
Westchester	57	30	0.53
Jefferson	56	18	0.32
Orange	56	17	0.29
Rockland	56	16	0.29
Dutchess	57	16	0.28
Chemung	57	14	0.25

4. CONCLUSIONES

- Las variables de entrada basadas en índices de información molecular implementados en el software DRAGON y la entropía de Shannon basada en factores socioeconómicos a nivel de condado son herramientas útiles en la predicción de modelos RNA fármaco-epidemiológicos en la pandemia del SIDA.

- El cálculo de índices topológicos implementados en la herramienta informática DRAGON son de utilidad en el desarrollo de modelos AA-QSAR.

- Los modelos generados basados en algoritmos de aprendizaje automático, principalmente de RRNNAA, se han mostrado útiles para predecir los fármacos y cócteles de fármacos eficaces para el tratamiento del VIH en diferentes poblaciones de los condados de Estados Unidos con una prevalencia epidemiológica dada.

- Las herramientas informáticas basadas en técnicas y procedimientos de inteligencia artificial, han demostrado ser de gran utilidad como modalidad inicial de “screening” en el complejo proceso de desarrollo y descubrimiento de fármacos antirretrovirales.

- Gracias a la utilización de técnicas de IA se han podido predecir redes complejas de la prevalencia del SIDA en los EE.UU. a nivel de condado con respecto a la actividad de fármacos anti-VIH en ensayos preclínicos.

- La escala de la estructura de población (Estado, UIC, RUCC) no afecta a la calidad del modelo de aprendizaje automático. Esto puede indicar que la eficiencia de un cóctel desde el punto de vista epidemiológico no depende de la estructura demográfica de la población.

- Las técnicas y procedimientos de AA e IA utilizados en la presente tesis, no permiten establecer una relación definitiva entre la adherencia al tratamiento del VIH y el nivel socioeconómico, ya que esta es aún rudimentaria y no hay un soporte contundente de la existencia de una clara asociación.

5. FUTUROS DESARROLLOS

Desarrollo de modelos RNA-QSAR aumentando el número de variables de entrada respecto a factores socioeconómicos de poblaciones específicas (ingresos, nivel de empleo, presión fiscal, educación, migración, estructura de la población) para estudiar la asociación entre el estatus socioeconómico y el VIH/SIDA.

Generación de "scripts" propios en R y Phyton en las bases de datos para la inclusión de índices topológicos de redes de tipo estrella a través del software S2SNet, para las secuencias de proteínas de dianas involucradas en la terapia antirretroviral con mutaciones asociadas a resistencia viral, descargadas de la base de datos Protein Data Bank (www.rcsb.org/pdb/home/home.do) y de estudios reportados en la literatura (Wensing et al., 2014). Inclusión de estos índices en forma de Medias Móviles (MA) respecto a la proteína y el organismo, en las variables de entrada de nuevos modelos de predicción basados en RRNNAA y redes artificiales neurogliales (RRNNGG).

Definir otros descriptores moleculares y explorar su utilidad en estudios RNA-QSAR de redes complejas biomoleculares.

Crear "data warehouse" a partir de las bases de datos públicas disponibles, sobre los que aplicar técnicas de "Data Mining" y ontologías para hacer más eficiente la gestión del conocimiento y conseguir mejores niveles de convergencia en los resultados.

Creación de servidores "Web", para la predicción "online" de la actividad biológica de compuestos y sus correspondientes dianas moleculares. Esto tiene un gran interés, en el proceso de descubrimiento y desarrollo de fármacos.

Mapping Chemical Structure-Activity Information of HAART-Cocktail Drugs over Complex Networks of AIDS Epidemiology and Socioeconomic Data of U.S. Counties

Diana María Herrera-Ibatá[1,*], Alejandro Pazos[1], Ricardo Alfredo Orbegozo-Medina[2],
Francisco Javier Romero-Durán[3], and Humberto González-Díaz[4,5,*]

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

Abstract. Using computational algorithms to design tailored drug cocktails for highly active antiretroviral therapy (HAART) on specific populations is a goal of major importance for both pharmaceutical industry and public health policy institutions. In so doing, we need to predict new combinations of compounds to design HAART cocktails. We have the bio-molecular factors related to the drugs in the cocktail (experimental measure, chemical structure, drug target, organisms of assay, *etc*.). On the other side, we have the socio-economic factors of the specific population (income inequalities, employment levels, fiscal pressure, education, migration, population structure, *etc*.) to study the relation between the socioeconomic status and the disease. In this context, we need to use Machine Learning algorithms able to seek models for problems with multi-source data. In this work, we proposed the first ANN model for prediction of HAART cocktails to halt AIDS on epidemic networks of US Counties using information indices that codify both bio-molecular and several socioeconomic factors. We obtained the data from at least three major sources. The first dataset included assays of anti-HIV chemical compounds released to ChEMBL. The second data set is the AIDSVu database of Emory University. AIDSVu compiled AIDS prevalence for >2,300 U.S. counties. The third data set included socio-economic data from US Census Bureau. We used three scales or levels to group the counties according the location or population structure codes (state, RUCC AND UIC). We analyzed >130,000 pairs (network links) corresponding to AIDS prevalence in 2,310 counties in US vs. drug cocktails formed by combinations of ChEMBL results for 21,582 unique drugs, 9 viral or human protein targets, 4856 protocols, and 10 possible experimental measures. The best model found with the original data was a linear neural network (LNN) with AUROC > 0.80 and accuracy, specificity, and sensitivity ≈ 77% in training and external validation series. The change of the spatial and population structure scale (State, UIC, or RUCC codes) does not affect the quality of the model. We detected unbalance on all the models found comparing positive/negative cases and linear/non-linear model accuracy ratios. Using SMOTE data pre-processing and machine learning algorithms implemented in the software WEKA, we found models that are more balanced. In particular, we found one MLP with AUROC = 97.4% and Precision, Recall, and F-Measure >90%.

Keywords: Urban Influence Code; AIDS Epidemiology; Box-Jenkins operators; Shannon Entropy; Information Theory.

*** Corresponding authors:** DMHI (diana.herrera@udc.es) and HGD (humberto.gonzalezdiaz@ehu.es)

1. INTRODUCTION

Computational algorithms may play an important role in the process of elucidation of structure-activity relationships for many molecular systems and biological problems (Aguilera and Rodriguez-Gonzalez, 2014; Barresi et al., 2013; Gonzalez-Diaz et al., 2011; Munteanu et al., 2009). In particular, the theoretical biology has been very useful in the study of anti-HIV drugs and/or their molecular targets (Jain Pancholi et al., 2014; Ogul, 2009; Speck-Planche et al., 2012; Weekes and Fogel, 2003; Xu et al., 2013). However, classic algorithms useful to connect the structure of a single molecule with its biological properties are unable to study the effect of combinations (cocktails) of drugs over epidemiological outbreaks in large populations with different socio-economic factors. For instance, infections with HIV are commonly treated with antiretroviral drug combinations. These treatments could diminish the risk of HIV transmission (Castilla et al., 2005; Ping et al., 2013). In addition, the rates of disease progression, opportunistic infections, and mortality have decreased with the implementation of HAART, and the combination of anti-HIV drugs has result in longer survival and a better quality of life for the people infected with the virus (Colombo et al., 2014). 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), which have all resulted in decreased HIV RNA levels (<50 copies/mL) at 48 weeks and CD4 cell increases in the majority of patients (Usach et al., 2013). Research indicates (McMahon et al., 2011) that despite HAART therapy, HIV infected individuals who are poor, homeless, hungry, or have less education, continue to have a higher risk of death. Additionally, researchers (McMahon et al., 2011) suggest that HIV infected individuals with attributes of low socioeconomic status (SES) are more likely to have increased mortality than those who are not living under these adverse conditions. Therefore, resources for HIV testing, care, and proven economic interventions should be directed to areas of economically disadvantaged people (McDavid Harrison et al., 2008).

The case of the United States (US) is interesting for theoretical studies due to the abundance of epidemiological information. Holtgrave and Crosby (Holtgrave and Crosby, 2003) found strong correlation (r = 0.469, p <0.01) between the income inequality and the AIDS cases rates at state level in the United States. In addition, in 2010 the National HIV Behavioral Surveillance System of the US developed a study about HIV infection among heterosexuals at increased risk with a total of 12,478 persons. Within the 8,473 participants, 197 (2.3%) participants were positive for HIV infection, and prevalence was similar for men (2.2%) and women (2.5%). The research shows a higher prevalence for persons who reported less than a high school education (3.1%), compared with those with a high school education (1.8%). Income inequality, employment, and other social variables seem also to be relevant on AIDS epidemiology. Prevalence also was higher for those with an annual household income less than $10,000 (2.8%), compared with those with an income of $20,000 or more (1.2%).(CDC, 2013) Moreover, the percentage of HIV infected was higher for participants who reported being unemployed (1.1%) or disabled (and unemployed) (2.7%), compared with employed (0.4%). Furthermore, some authors as Mondal and Shitan (Mondal and Shitan, 2013) commented in their study connections among life expectancy, income, educational attainment, fertility, health facilities, and HIV prevalence.

Recently, large amounts of data have been accumulated in public databases about the scope of molecular biology. For instance, the ChEMBL database (https://www.ebi.ac.uk/chembl/) (Bento et al., 2013; Gaulton et al., 2012; Heikamp and Bajorath, 2011) provides data from life sciences experiments (Bento et al., 2013). In the same way, there are online resources containing epidemiological data of AIDS prevalence and data about socioeconomic factors at county level. These databases are AIDSVu (http://aidsvu.org), created by researchers at the Rollins School of Public Health at Emory University; and the US Center for Disease Control and Prevention (CDC). In this context, the search of computational chemistry algorithms useful to carry out a mapping of Structure-Activity data of HAART-cocktails drugs over AIDS Epidemiology networks and socioeconomic data is of the major importance. In a recent paper (González-Díaz et al., 2014), we used artificial neural networks (ANNs) to link data of AIDS in US counties with ChEMBL data about chemical structure and preclinical activity of anti-HIV compounds. ANNs are prediction models, widely used in many areas of science, such as medicine, chemistry, biochemistry, as well as drug development. In drug development are very useful for the prediction of properties of potential drugs. The ANNs approximate the operation of the human brain with the ability to get results from complicated or imprecise data, which are very difficult to appreciate by humans or other computer techniques (Burbidge et al., 2001; Guha, 2013; Patel, 2013; Speck-Planche et al., 2012). We used as input information indices of social networks and molecular graphs. We used a Shannon information index based on the Gini coefficient to quantify the effect of income inequality in the social network. We also used the Balaban information indices to quantify changes in the chemical structure of single anti-HIV drugs. Last, we used Box-Jenkins moving average operators (MA) to quantify information about the deviations of drugs with respect to data subsets of reference (targets, organisms, experimental parameters, protocols). In our previous paper,(González-Díaz et al., 2014) the model found was able to link the deviations in the AIDS prevalence in the ath county with the changes in the biological activity of the qth drug (d_q).

However, the previous computational chemistry algorithm fails on accounting for drug cocktails and many socio-economic factors. In this work, we developed by the first time a computational algorithm for Network Epidemiology able to map structure-activity data of HAART-cocktail drugs over complex networks of AIDS epidemiology and socio-economic for >2000 US counties.

2. MATERIALS AND METHODS

2.1. Socio-Economic Factors

2.1.1. Socio-economic Variables and Shannon-Entropy transformation into information indices

In total, 17 variables were withdrawn from AIDSVu, US Census Bureau databases (http://www.census.gov/) and Tax Foundation (http://taxfoundation.org/). See the symbols and details of these variables in **Table 1**. All the 17 socio-economic variables (v_a) discussed before come from very different original sources, describe different phenomena, and then use different scales.

Please, Table 1 comes about here

In order to perform an uniform and scale un-biased representation of information, we transformed all these variables into Shannon entropy information indices $I_a(v)$. These information indices depend on the values of variables re-scaled into probabilities as follow.

$$I_a(v) = -p_a(v) \cdot \log p_a(v) \qquad (1)$$

$$p_a(v) = \frac{(v - v_{\min} + \varepsilon)}{(v_{\max} - v_{\min} + \varepsilon)} \qquad (2)$$

This transformation guarantees that the new probability values become 1 for the maximal value v_{max} and approach to 0 for values near to minimal value v_{min}. The scaling parameter ε = 0.0001 was used to avoid values of $p_a(v) = 0$ with the subsequent undefined results of the entropy function for logarithm log(0). In the **Table 2,** we show a short example of the results of the consecutive probability and Shannon entropy scaling procedures for some variables.

Please, Table 2 comes about here

2.1.2. Demographic scale levels of socio-economic information indices

We studied the variability of these 17 socio-economic variables at two different demographic scales. One of these is geo-political level and the other level a local population structure level. The first level was identified as the grouping of counties into 51 different states (s). Actually, only 47 states are in our dataset. The second level was measured with two alternative codes: Rural-Urban Continuum Codes (RUCC): distinguishes metropolitan counties by the population size of their metro area, and nonmetropolitan counties by degree of urbanization and adjacency to a metro area. However, in this work we used the RUCC classification of 2003 to preserve causality relationships. RUCC 2003 is the classification reported before to the AIDSVu epidemic data, which is for year 2010, and RUCC 2013 is posterior and could cause-effect relationships. The standard Office of Management and Budget (OMB) metro and non-metro categories have been subdivided into three metro and six non-metro categories. Each county in the U.S. is assigned with one of the nine codes. This scheme allows researchers to break county data into finer residential groups, beyond metro and non-metro, particularly for the analysis of trends in non-metro areas that are related to population density and metro influence. In addition, the Urban Influence Codes (UIC) distinguishes metropolitan counties by population size of their metro area, and nonmetropolitan counties by size of the largest city or town and proximity to metro and micropolitan areas. The OMB metro and non-metro categories have been subdivided into two metro and 10 non-metro categories, resulting in a 12-part county classification. **Table 3** shows the RUCC and UIC classification codes.

Table 3 comes about here

2.1.4. Box-Jenkins MA operators of socio-economic variables at different levels

We used the Moving Average operators of Box-Jenkins (MA) in order to measure the variability of the $I_a(v)$ at two different demographic scales (state and population). In so doing, we calculated the average parameters $<I_a(v)_L>$ for different levels of population organization L = u, r, s. Consequently we obtained the average values $<I_a(v)_s>$, $<I_a(v)_r>$, $<I_a(v)_u>$ for all the $I_a(v)$ values. The parameters $<I_a(v)_s>$ are the averages of $I_a(v)$ for all the counties in the same state (L = s). The parameters $<I_a(v)_r>$ are the averages of $I_a(v)$ for all the counties with the same population structure according to RUCC code (L = r). The parameters $<I_a(v)_u>$ are the averages of $I_a(v)$ for all the counties with the same population structure according to UIC code (L = u)(Brown et al., 1976; Ghelfi and Parker, 1997).

After calculating the averages $<I_a(v)_L>$,we were able to determine the values of the MA operators for each county. We tabulated the values of $<I_a(v)_s>$ for 47 states. We also calculated the values of$<I_a(v)_r>$ and $<I_a(v)_u>$ for 9 and 12 different types of population structures according to RUCC and UIC codes, respectively, see **Table 1SM** of supplementary material. Some examples of MA operators for selected counties at State, RUCC, and UIC levels are in **Table 4**, see also other examples at state level in last columns of **Table 2**.The formulae of these MA operators are:

$$\Delta I_a(v)_L = I_a(v) - \langle I_a(v)_L \rangle \qquad (3)$$

$$\Delta I_a(v)_s = I_a(v) - \langle I_a(v)_s \rangle \qquad (4)$$

$$\Delta I_a(v)_r = I_a(v) - \langle I_a(v)_r \rangle \qquad (5)$$

$$\Delta I_a(v)_u = I_a(v) - \langle I_a(v)_u \rangle \qquad (6)$$

Table 4 comes about here

2.2. Bio-Molecular Factors

2.2.1. Shannon-Entropy transformation of chemical structure into information indices

We can use quantitative descriptors of the molecular graph of the drug to quantify the chemical structure of anti-HIV compounds. We used the molecular information indices $I_d(k)$ implemented in the software DRAGON version 5.3.(Todeschini and Consonni, 2000) In this work, only the $I_d(k)$information indices were used. The mathematical background of these descriptors has been explained in a previous work.(Herrera-Ibatá et al., 2014) The names, symbols, and formula for the calculation of different $I_d(k)$ descriptors is summarized in **Table 5,** see details on the following references.(Bertz, 1981; Bonchev and Trinajstic, 1978; Dancoff and Quastler, 1953; Klopman et al., 1988; Raychaudhury et al., 1984; Shannon and Weaver, 1949; Todeschini and Consonni, 2000)

Table 5 comes here

2.2.2. Box-Jenkins MA operators of molecular information indices for a single molecule

The molecular descriptors used were the $I_d(k)$ (13 information indices) of each anti-HIV drug forming the cocktail (131,252 anti-HIV cocktails). We used the $I_d(k)$ descriptors as input to calculate MA operators of the bio-molecular factors for the drugs. Consequently, to calculate the MA bio-molecular operators we needed the value of the drugs information indices $I_d(k)$ and the average of these indices for all drugs assayed with the same boundary conditions (c_j) of a given bio-molecular factor. In general, c_1, c_2, and c_3 refer to different sets of these boundary conditions for the same bio-molecular factor (type of assay, molecular targets, cellular lines, organisms, experimental measures, *etc.*) for a single molecule. A scheme with some examples that describes the methodology used to calculate the inputs corresponding to the drugs is in **Figure 1**.

$$\Delta I_d\left(k, {}^d c_j\right) = I_d(k) - \langle I_d(k) \rangle_{cj} \qquad (6)$$

$$\langle I_d(k) \rangle_{cj} = \frac{1}{n_j} \sum_{d=1}^{d=n_j} I_d(k) \qquad (7)$$

Figure 1 comes here

2.2.3. Box-Jenkins MA operators of molecular information indices for HAART drug cocktails

In the case of a MA operator for cocktail drugs (up to three molecules in HAART cocktails studied here), we used as input the MA operators of single drugs. These MA operators for cocktails take into consideration the sets of conditions ${}^dc_j$= [${}^dc_1$, ${}^dc_2$, ${}^dc_3$, ${}^dc_4$] for each drug. In general, ${}^1c$, ${}^2c$, and ${}^3c$ refer to different sets of these boundary conditions for the same bio-molecular factor (type of assay, molecular targets, cellular lines, organisms, experimental measures, *etc.*). Therefore, ${}^1c_1$, ${}^2c_1$, ${}^3c_1$ = are the experimental measures of activity for the first, second, and third drugs of the cocktail, respectively. In analogy, ${}^1c_2$, ${}^2c_2$, and${}^3c_2$ are the protein targets for the same drugs. In addition, ${}^1c_3$, ${}^2c_{3,}$, and ${}^3c_3$ are the organisms that express the targets of these compounds. Last, ${}^1c_4$, ${}^2c_{4,}$ and ${}^3c_4$ are the different assay protocols used to test the activity of these compounds *per se*. The MA operator for a drug cocktail was calculated as the arithmetic mean of the corresponding MA operator for each drug in the cocktail.

$$\Delta I_c\left(k,{}^d c_j\right) = \frac{1}{3}\sum_{d=1}^{d=3}\Delta I_d\left(k,c_j\right) = \frac{1}{3}\sum_{d=1}^{d=3}\left[I_d(k) - \left\langle I_d(k)\right\rangle_{c_j}\right] \qquad (8)$$

The information indices $I_d(k)$ of the molecules, the average values $<I_d(k)>_{cj}$ of these indices for different boundary conditions (c_j), and relevant information for bio-molecular factors of all drugs appear in **Table SM2** and **Table SM3** of the supplementary material, respectively. A scheme summarizing the above steps is depicted in **Figure 2.**

Figure 2 comes near about here

2.3. ALMA models of complex networks

2.3.1. Linear ALMA models

ALMA (Assessing Links with Moving Averages) is a technique developed by our group that has been used to construct complex multi-scale networks of AIDS and anti-HIV drugs before (González-Díaz et al., 2014; Herrera-Ibatá et al., 2014). In the previous study, we used MA operators of bio-molecular and socio-economic factors. In this work, we used the ALMA technique to fit a new class of dual models combining chemoinformatics and epidemiological data for HAART cocktails formed by combinations of 1-3 anti-HIV drugs. These new models are able to link AIDS epidemiology data with socio-economic and population structure data of US counties and preclinical structure-activity information of all compounds combined in each HAART cocktail. We used the MA of operators of nodes of networks (drugs, proteins, organisms, populations, *etc.*) to predict the variable $L_{ac}({}^d c_j)_{obs}$. The value is $L_{ac}(c_j) = 1$ when the Cocktail-Disease Ratio = $CDR_{ac}({}^d c_j)$ >cutoff and $L_{aq}({}^d c_j)_{obs} = 0$ otherwise. The term $CDR_{ac}(c_j) = [z_c/D_a]$; being $z_c = (z_1 + z_2 + z_3)/3$ = the average of the z-scores z_1, z_2, z_3 of the biological activity for each drug (dth) present in the cocktail. The term D_a is the AIDS prevalence rate for the ath county. We calculated each zeta as $z_d(c_j) = \delta_j \cdot z_d(c_j) = \delta_j \cdot [v_d(c_j) - AVG(v(c_j))]/SD(v(c_j))$. In this operator, $v_d(c_j)$ is the value of biological activity (EC_{50}, IC_{50}, K_i, … *etc.*) reported in the ChEMBL database for the dth drug assayed in the set of conditions. The parameter δ_j is similar to a Kronecker delta function. The parameter $\delta_j = 1$ when the biological activity parameter $v_d(c_j)$ is directly proportional to the biological effect (e.g., K_i values, Activity (%) values, etc.). Conversely, $\delta_j = -1$ when the biological activity parameter $v_d(c_j)$ is in inverse proportion to the biological effect (*e.g.*, EC_{50} values, IC_{50} values, *etc.*). The parameter $z_d(c_j)$ is the z-score of the biological activity that depends on the functions AVG and SD. These functions are the average and standard deviation of $v_d(c_j)$ for all drugs assayed in the same conditions. The reader should note that the predicted, output, or dependent variable of one ALMA model is not a discrete variable but a real-valued numerical score (S_{ac}). However, the variable is directly proportional to the observed variable. The general formula for a linear ALMA model developed using average values of $\Delta I_a(L, v)$ and ΔI_c indices of the counties and compounds used in a given drug cocktail was:

$$S_{ac} = \sum_{k=1}^{k=13}\sum_{j=1}^{j=4} e_{kj}\cdot\Delta I_c\left(k,{}^d c_j\right) + \sum_{L=1}^{L=3}\sum_{v=1}^{v=17} e_{Lv}\cdot\Delta I_a(L,v) + e_0 \qquad (9)$$

3. RESULTS AND DISCUSSION

3.1. Two-Way Joining Cluster Analysis and Principal components Analysis

Two-way joining cluster analysis (TWJCA) and principal components analysis (PCA) are useful to reduce the magnitude of data sets with many input variables. Two-way joining is useful in circumstances that expect that both cases and variables will simultaneously contribute to find meaningful patterns of clusters (Hill and Lewicki, 2006). We used here a dichotomist approach for both TWJCA and PCA. It means that we carried out the TWJCA and PCA of socio-economic and bio-molecular factors separately. These techniques were used in order to perform a preliminary exploratory study of the data and to determine the variability of them. In addition, we studied the discriminatory effect of the information indices on the different conditions of assay. First, we used TWJCA to analyze the bio-molecular data. The TWJCA algorithm reorganized the average values of the information indices with respect to those compounds with the same experimental measure, drug targets, and organism of assay into a total number of blocks (see **Table 6** and **Figure 3**). For example, the experimental measure present an initial input of blocks, 130, resulting in 49 output blocks after performing the TWJCA. As it can be seen, in the Hot Maps (HM) depicted in **Figure 3**, the experimental measure and bio-molecular targets show that there is not an information index that distinguishes well each condition for the experimental measure and targets, however some indices (IDET, IDMT, and ISIZ) represent clearly the CXCR-4 receptor. Moreover, in figure corresponding to the organism, the ISIZ index discriminates well each organism of assay (HIV-1, HIV-2, hsa, etc.). Next, we carried out the TWJCA of socio-economic data. In the specific case of TWJCA for socio-economic data, we also carried out separately the analysis of different levels of population distribution (UIC, RUCC, and State). The HM obtained by cluster analysis does not show significant differences between the metro and non-metro codes and the population structures (see **Figure 4**).

Table 6 and Figure 3 and Figure 4 comes near here

After that, we carried out a PCA of data. We used PCA in this work with two different aims. The first aim was to represent the complex data of anti-HIV drug cocktails *vs*. US counties in a compact form and analyze the results. The PCA for the socioeconomic factors was performed with 68 input variables, resulting in four factors that represent the 72% of the information (see **Table 6)**. The plot of socioeconomic eigenvalues is shown in **Figure 5**. The first factor represents the population and employment, the second factor shows the information about education and poverty, the third one is the domestic and net migration and the fourth factor refers to education level. On the other hand, the PCA for the bio-molecular factors was made with 65 input variables. In this case, the analysis showed four eigenvalues for the bio-molecular factors that account the 74% of the information, being the first factor the experimental measure and organism, the second factor the drug structure, the third factor the assay and the fourth factor the target **(Figure 5**). In **Table 6,** we depicted the eigenvalues obtained for the different principal components. The eigenvalues generated during PCA give an indication of the amount of information carried by each component.

Figure 5 comes about here

3.2. ANN calculation of parameters linear and non-linear ALMA model

In our previous work (González-Díaz et al., 2014), we developed a LNN model with Balaban information indices for anti-HIV compounds present in the ChEMBL database (unique drugs = 21,582 and total data points = 43,249). The model included also the Shannon entropy information indices based on values of Gini income inequality measure of the US counties (Pabayo et al., 2014). The model presented values of accuracy (Ac), specificity (Sp), and sensitivity (Sn) ≈ 0.75 in training and external validation series. In this work, we also trained different ANNs using the values of MA operators for the information indices of several socioeconomic and bio-molecular inputs. In total, we used 40 MA operators for the different bio-molecular conditions of drugs cocktails (experimental measures, targets and organism), and 50 MA operators of the socioeconomic factors in US counties. The MA of socioeconomic factors for each county was calculated in the form of deviations from all counties with the same populations, with the same structure (*i.e.,* RUCC or UIC code) or with the same geographic location (same State). Finally, we obtained different prediction models. The dataset used to perform the model includes N = 131,252 statistical cases. The data used to *train* the model include N = 78,752 statistical cases, *selection* were 26,250 statistical cases and *validation* were 26,250 statistical cases. Cases with $L_{ac}(c_j)_{obs}$=1 were 22,100 and cases with $L_{ac}(c_j)_{obs}$= 0 were 109,152. We employed the ANNs implemented in the software package STATISTICA 6.0 (Hill and Lewicki, 2006). The statistical parameters used to support the model were: Number of cases in training (N), and overall values of specificity (Sp), sensitivity (Sn), accuracy (Ac), and AUROC (area under receiver operating curve).We trained different topologies of ANNs including multilayer perceptrons (MLPs) and linear neural networks (LNN). Last, we also trained ALMA models using a PCA-ANN approach. In fact, the output of the PCA can be copied to the data set, and used to train the ANN with a notably lower number of input variables. We analyzed >130,000 pairs (network links) corresponding to AIDS prevalence in 2,310 counties in US vs. drug cocktails formed by combinations of ChEMBL results for 21,582 unique drugs, 9 viral or human protein targets, 4856 protocols, and 10 possible experimental measures. The parameters of the ANNs generated are depicted in **Table 7**.The best model found with original data was a linear neural network (LNN) with AUROC > 0.80 and Ac, Sp, and Sn ≈ 79% in training and external validation series. However, we chose the UIC-LNN model because its performance is Ac, Sp, and Sn ≈ 77% in training and external validation series using fewer variables. In addition, the Urban influence codes present a more specific classification scheme for structure population. Additionally, the population structure scale (State, UIC, or RUCC codes) does not affect the quality of the model (see **Table 8**). This may indicate that the efficiency of a cocktail from epidemiological point of view does not depend on the demographic structure of the population. However, the inclusion of different socio-economic factors seems to affect the accuracy of the model. The SES depends on a combination of variables including occupation, education, income, and place of residence, therefore the relationship between the social determinants and AIDS have a role to play in the adherence of HAART therapy (Falagas et al., 2008). Nevertheless, evidence of the association between adherence to HIV therapy and socioeconomic status is still rudimentary, varied and there is no a conclusive support for the existence of a clear association. Some studies (McFarland et al., 2003) found lower socioeconomic status (SES) to be associated with higher mortality from AIDS. Recent evidence indicates that AIDS is a disease of inequality, often associated with economic transition, rather than a disease of poverty in itself.(Piot et al., 2007) Additionally, many researchers now point not to poverty itself but to economic and gender inequalities and weakened "social cohesion" (Barnett and Whiteside, 2006) as factors influencing sexual behavior and hence the potential for HIV transmission. Undeniably, more people live with HIV in poor countries than in rich ones. More than 60% of people living with HIV inhabit the world's poorest region: sub-Saharan Africa. However, studies during the early stage of the epidemic suggested that HIV incidence initially occurred not amongst the poorest, but among better-off members of society in this region. A decade later, infections still appear more concentrated among the urban employed and more mobile members of society, and consequently the more wealthy groups (Piot et al., 2007).

We detected unbalance on all the models found comparing positive/negative cases and linear/non-linear model accuracy ratios. Using SMOTE data pre-processing and machine learning algorithms implemented in the software WEKA (Hall et al., 2009), we found models that are more balanced. In particular, as it can be seen in **Table 9**, we found one MLP with AUROC = 97.4% and Precision, Recall, and F-Measure >90%.

Table 7 and Table 8 comes near here

On the other hand, training ALMA models using PCA-ANN fails to generate good predictions classifiers, with Sp and Sn results close to 50% in MLP and LNN networks (see **Table 8).** In this work, we chose the UIC-LNN model, because is a more specific classification scheme of the population structure than the other ones and the LNN is the simplest type of classification model. The UIC LNN model shows values of AUROC = 0.85 in training and AUROC = 0.83 for external validation set (see **Figure 6**).

Figure 6 comes about here

Anyhow, we noted certain unbalance on the classification of positive / negatives cases as well as on the predictive power of linear vs. non-linear classifiers. In consequence, we transformed our dataset with data pre-processing algorithms and three different machine learning algorithms implemented in the software WEKA (Hall et al., 2009), a freeware package for data mining. In so doing, we used firstly a hybrid preprocessing approach called SMOTE (Chawla et al., 2002), based on oversampling and under-sampling our high imbalanced dataset in order to equilibrate the two output classes. This generates a substantial improvement of results on test set implemented on the non-linear models, being especially interesting the rise in precision or positive predictive value given, that it is the main goal of our research. We apply the MLP and Random Forest methods, a more computationally demanding schemes able to uncover underlying complex and non-linear functions between the variables. In conclusion, our data seem to be better modeled through a combination of previous preprocessing and the application of non-linear machine learning algorithms as reflected in **Table 9**.

Table 9 comes about here

3.3. Back-projection of the Computational Chemistry model over U.S. county sub-networks

We used the output values ($L_{ac}(^{d}c_{j})_{obs} = 1$) or not ($L_{ac}(^{d}c_{j})_{obs} = 0$) of the ALMA classifier to generate different sub-networks. This variable quantifies the formation of links between nodes in the core complex network.

This network maps the AIDS prevalence with respect to the preclinical activity of anti-HIV drug cocktails in each state of the US at county level. This network has two parts, the core, and the periphery. There are two different types of nodes forming the core of this specific network. The first node represents the US counties (a^{th}) and the second type of node represents HAART cocktails (c^{th}). In addition, each node cocktail has 2-3 nodes attached to it, which represents the drugs present in the cocktail (network periphery). The **Figure 7** shows a sub-network (of the previous type of network) for AIDS prevalence in the state of New York (NY) *vs.* Anti-HIV drug preclinical activity for all drugs combined in HAART cocktails designed from compounds reported in ChEMBL. The sub-network has the three types of nodes, the nodes of the core of the network are the US counties (red) and HAART cocktails (blue). The nodes of the periphery of the network are anti-HIV compounds combined to making up different cocktails (nodes hidden in the picture). It is important to understand that here $L_{ac}(c_j)_{pred}$= 1 expresses the existence of a sub-graph that connect several nodes of all classes by means of various arcs and no single arc that connect two nodes. It is possible to create a similar type of sub-network with a model reported in a previous work (Herrera-Ibatá et al., 2014). In that paper, the type of sub-network may have different classes of nodes. There are three main classes: counties of the state, the drug cocktails, and the chemical compounds making up the cocktail. However, this sub-network includes only one socioeconomic variable: Gini coefficient. Furthermore, we developed other type of sub-network in a preceding work (González-Díaz et al., 2014), It has two classes of nodes (counties vs. drugs). The drug nodes contained information about the chemical structure, as well as, all the assay conditions (target, organism, assay protocol, experimental measure). Additionally, the county nodes contained the information about the income inequality. However, because of the type of model used, those complex networks are unable to represent drug cocktails.

Please, Figure 7 near here

3.4. Computational chemistry modeling of AIDS Epidemiology in U.S. Counties Network

We calculated the probabilities with which AIDS could be halt with several drug cocktails in determined counties. As it can be seen in the formula, the county in-degree is the number of positive links between the different cocktails and the county. The county frequency refers to the total number of times that the county is in the original database, including the cocktails predicted as negative. It is important to explain, not all counties were modeled against every drug cocktails.

$$p(halt) = \frac{(county\ in - \deg ree)}{(county\ frequency)} = \left(\frac{county\, in - \deg ree}{[(county\, in - \deg ree) + (negative cocktails)]} \right) \qquad (10)$$

Moreover, in **Table 10** are depicted some examples from the complex sub-network of the state of New York, with data of counties in-degree with several HAART cocktails. For example, Bronx County shows a good in-degree in the complex subnetwork, e.g., the probability that several HAART cocktails work at this county is higher than Chemung county, that presents a lower probability. Thus, this type of model could be useful for epidemiological surveillance procedures to understand the vulnerability of the populations regarding AIDS epidemic.

Please, insert Table 10 near here

4. CONCLUSIONS

We used ALMA models to carry out a back-projection of the preclinical activity of drugs combined in one HAART cocktail over a complex network of AIDS in the US counties. In this work, we chose the UIC-LNN model, because is a more specific classification scheme of the population structure than the other ones and the LNN is the simplest type of classification model. However, we noted unbalance on the classification of positive /negatives cases as well as on the predictive power of linear vs. non-linear classifiers. In consequence, we transformed our dataset with data pre-processing algorithms and three different machine learning algorithms implemented in the software WEKA.(Hall et al., 2009) Firstly, we used a hybrid preprocessing approach called SMOTE (Chawla et al., 2002). This generates a substantial improvement of results on test set implemented on the non-linear models. We found models that are more balanced, such as, one MLP with AUROC = 97.4% and Precision, Recall, and F-Measure >90%.

Acknowledgements

R.O.M acknowledges financial support of FPI fellowship funded by MECD (Spanish Ministry of Education, Culture, and Sport).

References

Internal Revenue Service, Tax Foundation., http://taxfoundation.org/resources (February, 2014)

Aguilera, L. U., Rodriguez-Gonzalez, J., 2014. Studying HIV latency by modeling the interaction between HIV proteins and the innate immune response. J. Theor. Biol. 360, 67-77, doi:10.1016/j.jtbi.2014.06.025.

Barnett, T., Whiteside, A., 2006. AIDS in the twenty-first century: Disease and globalization. Palgrave Macmillan, New York.

Barresi, V., Bonaccorso, C., Consiglio, G., Goracci, L., Musso, N., Musumarra, G., Satriano, C., Fortuna, C. G., 2013. Modeling, design and synthesis of new heteroaryl ethylenes active against the MCF-7 breast cancer cell-line. Mol. Biosyst. 9, 2426-9, doi:10.1039/c3mb70151d.

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Kruger, F. A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., Overington, J. P., 2013. The ChEMBL bioactivity database: an update. Nucleic Acids Res., doi:10.1093/nar/gkt1031.

Bertz, S. H., 1981. The first general index of molecular complexity. . J. Am. Chem. Soc. 103, 3599-3601.

Bonchev, D., Trinajstic, N., 1978. On topological characterization of molecular branching. Int. J. Quantum Chem. Quant. Chem. Symp 12, 293-303.

Brown, D. L., Hines, F. K., Zimmer, J. M., 1976. Social and Economic Characteristics of the Population in Metro and Nonmetro Counties: 1970. U.S. Dept. Agr., Econ. Res. Serv. AER-272.

Burbidge, R., Trotter, M., Buxton, B., Holden, S., 2001. Drug design by machine learning: support vector machines for pharmaceutical data analysis. Comput. Chem. 26, 5-14.

Castilla, J., Del Romero, J., Hernando, V., Marincovich, B., Garcia, S., Rodriguez, C., 2005. Effectiveness of highly active antiretroviral therapy in reducing heterosexual transmission of HIV. J. Acquir. Immune Defic. Syndr. 40, 96-101.

CDC, 2013. HIV infection among heterosexuals at increased risk--United States, 2010. MMWR Morb. Mortal Wkly. Rep. 62, 183-8.

Colombo, G. L., Castagna, A., Di Matteo, S., Galli, L., Bruno, G., Poli, A., Salpietro, S., Carbone, A., Lazzarin, A., 2014. 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, 9-15, doi:10.2147/tcrm.s49428.

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Int. Res. 16, 321-357.

Dancoff, S. M., Quastler, H., 1953. Essays on the Use of Information Theory in Biology. University of Illinois, Urbana

Falagas, M. E., Zarkadoulia, E. A., Pliatsika, P. A., Panos, G., 2008. Socioeconomic status (SES) as a determinant of adherence to treatment in HIV infected patients: a systematic review of the literature. Retrovirology 5.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J. P., 2012. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 40, D1100-7, doi:10.1093/nar/gkr777.

Ghelfi, L. M., Parker, T. S., 1997. A County-Level Measure of Urban Influence. Rural Development Perspective 12.

Gonzalez-Diaz, H., Prado-Prado, F., Sobarzo-Sanchez, E., Haddad, M., Maurel Chevalley, S., Valentin, A., Quetin-Leclercq, J., Dea-Ayuela, M. A., Teresa Gomez-Munos, M., Munteanu, C. R., Jose Torres-Labandeira, J., Garcia-Mera, X., Tapia, R. A., Ubeira, F. M., 2011. NL MIND-BEST: a web server for ligands and proteins discovery--theoretic-experimental study of proteins of Giardia lamblia and new compounds active against Plasmodium falciparum. J. Theor. Biol. 276, 229-49, doi:10.1016/j.jtbi.2011.01.010.

González-Díaz, H., Herrera-Ibatá, D. M., Duardo-Sanchez, A., Munteanu, C. R., Orbegozo-Medina, R. A., Pazos, A., 2014. 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, 744-755.

Guha, R., 2013. On exploring structure-activity relationships. Methods Mol. Biol. 993, 81-94, doi:10.1007/978-1-62703-342-8_6

10.1007/978-1-62703-342-8_6.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H., 2009. The WEKA data mining software: an update. SIGKDD Explor. Newsl. 11, 10-18, doi:10.1145/1656274.1656278.

Heikamp, K., Bajorath, J., 2011. Large-scale similarity search profiling of ChEMBL compound data sets. J. Chem. Inf. Model. 51, 1831-9, doi:10.1021/ci200199u.

Herrera-Ibatá, D. M., Pazos, A., Orbegozo-Medina, R. A., Gonzalez-Diaz, H., 2014. Mapping networks of anti-HIV drug cocktails vs. AIDS epidemiology in the US counties. Chemometr. Intell. Lab. 138, 161-170.

Hill, T., Lewicki, P., 2006. STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining. StatSoft, Tulsa, Oklahoma.

Holtgrave, D. R., Crosby, R. A., 2003. Social capital, poverty, and income inequality as predictors of gonorrhoea, syphilis, chlamydia and AIDS case rates in the United States. Sex. Transm. Infect. 79, 62-4.

Jain Pancholi, N., Gupta, S., Sapre, N., Sapre, N. S., 2014. Design of novel leads: ligand based computational modeling studies on non-nucleoside reverse transcriptase inhibitors (NNRTIs) of HIV-1. Mol. Biosyst. 10, 313-25, doi:10.1039/c3mb70218a.

Klopman, G., Raychaudhury, C., Henderson, R. V., 1988. A new approach to structure-activity using distance information content of graph vertices: a study with phenylalkylamines. Math. Comput. Model. 11, 635-640.

McDavid Harrison, K., Ling, Q., Song, R., Hall, H. I., 2008. County-level socioeconomic status and survival after HIV diagnosis, United States. Ann. Epidemiol. 18, 919-27, doi:10.1016/j.annepidem.2008.09.003.

McFarland, W., Chen, S., Hsu, L., Schwarcz, S., Katz, M., 2003. Low socioeconomic status is associated with a higher rate of death in the era of highly active antiretroviral therapy, San Francisco. J. Acquir. Immune Defic. Syndr. 33, 96-103.

McMahon, J., Wanke, C., Terrin, N., Skinner, S., Knox, T., 2011. Poverty, hunger, education, and residential status impact survival in HIV. AIDS Behav. 15, 1503-11, doi:10.1007/s10461-010-9759-z.

Mondal, M. N., Shitan, M., 2013. Relative Importance of Demographic, Socioeconomic and Health Factors on Life Expectancy in Low- and Lower-Middle-Income Countries. J. Epidemiol.

Munteanu, C. R., Magalhaes, A. L., Uriarte, E., Gonzalez-Diaz, H., 2009. Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. J. Theor. Biol. 257, 303-11, doi:10.1016/j.jtbi.2008.11.017.

Ogul, H., 2009. Variable context Markov chains for HIV protease cleavage site prediction. Biosystems 96, 246-50, doi:10.1016/j.biosystems.2009.03.001.

Pabayo, R., Kawachi, I., Gilman, S. E., 2014. Income inequality among American states and the incidence of major depression. J. Epidemiol. Community Health 68, 110-5, doi:10.1136/jech-2013-203093.

Patel, J., 2013. Science of the science, drug discovery and artificial neural networks. Curr. Drug Discov. Technol. 10, 2-7.

Ping, L. H., Jabara, C. B., Rodrigo, A. G., Hudelson, S. E., Piwowar-Manning, E., Wang, L., Eshleman, S. H., Cohen, M. S., Swanstrom, R., 2013. HIV-1 transmission during early antiretroviral therapy: evaluation of two HIV-1 transmission events in the HPTN 052 prevention study. PLoS One 8, e71557, doi:10.1371/journal.pone.0071557.

Piot, P., Greener, R., Russell, S., 2007. Squaring the circle: AIDS, poverty, and human development. PLoS Med. 4, 1571-5, doi:10.1371/journal.pmed.0040314.

Raychaudhury, C., Ray, S. K., Ghosh, J. J., Roy, A. B., Basak, S. C., 1984. Discrimination of Isomeric Structures Using Information Theoretic Topological Indices. J. Comput. Chem. 5, 581-588.

Shannon, C., Weaver, W., 1949. The Mathematical Theory of Communication. University of Illinois Press, Urbana United States.

Speck-Planche, A., Kleandrova, V. V., Luan, F., Cordeiro, M. N., 2012. A ligand-based approach for the in silico discovery of multi-target inhibitors for proteins associated with HIV infection. Mol. Biosyst. 8, 2188-96, doi:10.1039/c2mb25093d.

Todeschini, R., Consonni, V., 2000. Handbook of Molecular Descriptors. Wiley-VCH Verlag GmbH, Weinheim, Germany.

Usach, I., Melis, V., Peris, J. E., 2013. Non-nucleoside reverse transcriptase inhibitors: a review on pharmacokinetics, pharmacodynamics, safety and tolerability. J. Int. AIDS Soc. 16, 1-14, doi:10.7448/ias.16.1.18567.

Weekes, D., Fogel, G. B., 2003. Evolutionary optimization, backpropagation, and data preparation issues in QSAR modeling of HIV inhibition by HEPT derivatives. Biosystems 72, 149-58.

Xu, L., Li, Y., Sun, H., Li, D., Hou, T., 2013. Structural basis of the interactions between CXCR4 and CXCL12/SDF-1 revealed by theoretical approaches. Mol. Biosyst. 9, 2107-17, doi:10.1039/c3mb70120d.

TABLES TO BE INSERTED IN THE TEXT

Table 1.US socio-economic variables

County Variables (v)	Description
G	Gini measure of income inequality in 2010
LIP	Percentage living in poverty in 2010
FIT	Federal Income Tax Burden as a Percentage of Adjusted Gross Income in 2004
LHS	Percent of persons with less than high school 2006-2010
OHS	Percent of persons with only a high school degree 2006-2010
SC	Percent of persons completing some college, 2006-2010
CD	Percent of persons with a college degree (at least a 4 year degree), 2006-2010
CPOP	4/1/2010 resident Census 2010 population
ChR	Numeric Change in resident total population 4/1/2010 to 7/1/2010
B	Births 2010
Nat	Natural increase in period 4/1/2010 to 6/30/2010
IntM	Net international migration in period 4/1/2010 to 6/30/2010
DMIG	Net domestic migration in period 4/1/2010 to 6/30/2010
NMIG	Net migration in period 4/1/2010 to 6/30/2010
CLF	Civilian labor force 2010
EMP	Employed 2010
UEMP	Unemployed 2010
RUC	2003 Rural Urban Continuum Code
UIC	2003 Urban Influence Code

Less than High School (LHS): In 1990, 2000, 2006-2010 the share includes those who did not receive a high school diploma or its equivalent (such as a GED), but did not report college experience. Only High School degree (OHS): In 1990, 2000, and 2006-2010 the share includes those who completed 12th grade and received a high school diploma or its equivalent (such as a GED), but did not report college experience. Some college (SC): In 1990, 2000, and 2006-2010 the share includes those who reported completing at least one year of college but did not receive a bachelor's degree. College graduate (CD): In 1990, 2000, and 2006-2010 the share includes those who received a bachelor's or higher degree.

Table 2. Process of transformation of original socio-economic variables into MA operators

U.S. County Name	U.S. State	Variable (v_a) (step 1)				Probability $p_a(v)$ (step 2)			
		G	LIP	FIT	LHS	$p_a(G)$	$p_a(LIP)$	$p_a(FIT)$	$p_a(LHS)$
Colfax County	NE	0.412	10.9	0.0604	26.7949	0.2558	0.1995	0.4224	0.5078
Dawson County	NE	0.403	11.7	0.0554	24.4983	0.2271	0.2200	0.4017	0.4631
Anoka County	MN	0.371	7.4	0.0968	7.2105	0.1249	0.1100	0.5732	0.1268
Beltrami County	MN	0.429	20.8	0.0767	10.9661	0.3101	0.4527	0.4898	0.1998
Baldwin County	GA	0.457	27	0.0634	21.9876	0.3996	0.6113	0.4346	0.4143
Fulton County	GA	0.529	17.7	0.1508	10.3541	0.6295	0.3734	0.7974	0.1879
Livingston County	IL	0.41	11.5	0.0836	15.9262	0.2494	0.2148	0.5187	0.2963
Greenbrier County	WV	0.45	20.8	0.0771	22.2466	0.3772	0.4527	0.4914	0.4193
Knox County	KY	0.507	33.9	0.0440	37.2290	0.5592	0.7877	0.3541	0.7108
Wexford County	MI	0.424	17	0.0739	12.5700	0.2942	0.3555	0.4785	0.2310
Becker County	MN	0.43	14.6	0.0787	10.0182	0.3133	0.2941	0.4981	0.1814
Benton County	MN	0.39	10.5	0.0895	9.7819	0.1856	0.1893	0.5429	0.1768
Choctaw County	MS	0.434	23.4	0.0360	19.2685	0.3261	0.5192	0.3210	0.3614
Lafayette County	MO	0.389	11.9	0.0707	15.0247	0.1824	0.2251	0.4653	0.2788
U.S. County Name	U.S. State	Shannon entropy $I_a(v)$ (step 3)				Box-Jenkins MA operator ΔI_a (L,v) (step 4)			
		$p_a(G)$	$p_a(LIP)$	$p_a(FIT)$	$p_a(LHS)$	$\Delta I_a(S,G)$	$\Delta I_a(S,LIP)$	$\Delta I_a(S,FIT)$	$\Delta I_a(S,LHS)$
Colfax County	NE	0.1515	0.1397	0.1581	0.1495	0.0060	-0.0006	0.0127	0.0158
Dawson County	NE	0.1462	0.1447	0.1591	0.1548	0.0007	0.0044	0.0138	0.0212
Anoka County	MN	0.1128	0.1054	0.1385	0.1137	-0.0286	-0.0274	-0.0036	-0.0181
Beltrami County	MN	0.1577	0.1558	0.1518	0.1398	0.0163	0.0230	0.0097	0.0079
Baldwin County	GA	0.1592	0.1307	0.1573	0.1585	0.0088	-0.0084	0.0045	0.0100
Fulton County	GA	0.1265	0.1598	0.0784	0.1364	-0.0239	0.0207	-0.0744	-0.0121
Livingston County	IL	0.1504	0.1435	0.1479	0.1565	0.0025	-0.0002	0.0045	0.0108
Greenbrier County	WV	0.1597	0.1558	0.1516	0.1583	0.0042	0.0040	-0.0011	0.0071
Knox County	KY	0.1412	0.0816	0.1597	0.1054	-0.0115	-0.0606	0.0077	-0.0411
Wexford County	MI	0.1563	0.1597	0.1532	0.1470	0.0041	0.0083	0.0090	0.0061
Becker County	MN	0.1579	0.1563	0.1508	0.1345	0.0165	0.0235	0.0087	0.0027
Benton County	MN	0.1357	0.1368	0.1440	0.1330	-0.0056	0.0040	0.0019	0.0012
Choctaw County	MS	0.1587	0.1478	0.1584	0.1597	0.0076	0.0199	0.0042	0.0099
Lafayette County	MO	0.1348	0.1458	0.1546	0.1547	-0.0150	-0.0042	0.0028	0.0044

Table 3. Values of RUCC and UIC codes in US in 2003

RUCC		**UIC**	
Code	Description	Code	Description
Metro counties:			
1	Counties in metro areas of 1 million population or more	1	In large metro area of 1+ million residents
2	Counties in metro areas of 250,000 to 1 million population	2	In small metro area of less than 1 million residents
3	Counties in metro areas of fewer than 250,000 population	-	---
Non-metro counties:			
4	Urban population of 20,000 or more, adjacent to a metro area	3	Micropolitan area adjacent to large metro area
5	Urban population of 20,000 or more, not adjacent to a metro area	4	Noncore adjacent to large metro area
6	Urban population of 2,500 to 19,999, adjacent to a metro area	5	Micropolitan area adjacent to small metro area
7	Urban population of 2,500 to 19,999, not adjacent to a metro area	6	Noncore adjacent to small metro area and contains a town of at least 2,500 residents
8	Completely rural or less than 2,500 urban population, adjacent to a metro area	7	Noncore adjacent to small metro area and does not contain a town of at least 2,500 residents
9	Completely rural or less than 2,500 urban population, not adjacent to a metro area	8	Micropolitan area not adjacent to a metro area
-	---	9	Noncore adjacent to micro area and contains a town of at least 2,500 residents
-	---	10	Noncore adjacent to micro area and does not contain a town of at least 2,500 residents
-	---	11	Noncore not adjacent to metro or micro area and contains a town of at least 2,500 residents
-	---	12	Noncore not adjacent to metro or micro area and does not contain a town of at least 2,500 residents

Table 4. Examples of MA operators for different scales and population structures for selected counties

U.S. County	U.S.		$\Delta I_a(L,v)$			
Name	State	AIDSCR	$\Delta I_a(S,G)_s$	$\Delta I_a(S,LIP)_s$	$\Delta I_a(S,FIT)_s$	$\Delta I_a(S,LHS)_s$
Perry County	PA	71	-0.066	-0.012	0.009	0.007
Sedgwick County	KS	177	0.011	0.011	-0.016	0.012
Mercer County	PA	66	0.004	0.014	0.004	0.002
Montgomery County	KS	50	0.006	0.012	0.006	0.017
Westmoreland County	PA	47	0.007	-0.010	-0.006	-0.021
Boyd County	KY	112	0.004	0.017	-0.015	0.010
Northampton County	PA	153	0.003	-0.008	-0.009	0.002
Riley County	KS	59	0.007	0.008	-0.009	-0.034
Montgomery County	PA	140	0.008	-0.066	-0.056	-0.031
Pottawatomie County	KS	48	0.006	-0.023	0.002	-0.019
Lebanon County	PA	101	-0.006	-0.006	0.004	0.009
Monroe County	PA	173	-0.007	0.005	0.004	-0.005
Wyoming County	PA	45	-0.013	0.004	0.003	-0.007
Boyle County	KY	89	0.001	0.017	-0.006	0.012
County Name	State	AIDSCR	$\Delta I_a(R,G)$	$\Delta I_a(R,LIP)$	$\Delta I_a(R,FIT)$	$\Delta I_a(R,LHS)$
Perry County	PA	71	-0.066	-0.014	0.010	0.009
Sedgwick County	KS	177	0.008	0.010	-0.008	0.002
Mercer County	PA	66	0.004	0.012	0.005	0.004
Montgomery County	KS	50	0.004	0.015	0.009	0.009
Westmoreland County	PA	47	0.013	0.002	0.005	-0.015
Boyd County	KY	112	0.004	0.011	-0.005	0.013
Northampton County	PA	153	0.003	-0.010	-0.009	0.005
Riley County	KS	59	0.004	0.011	-0.006	-0.042
Montgomery County	PA	140	0.014	-0.054	-0.046	-0.025
Pottawatomie County	KS	48	0.004	-0.023	-0.002	-0.034
Lebanon County	PA	101	-0.007	-0.010	0.000	0.010
Monroe County	PA	173	-0.008	0.002	-0.002	-0.007
Wyoming County	PA	45	-0.013	0.002	0.003	-0.005
Boyle County	KY	89	0.002	0.017	-0.005	0.014
County Name	State	AIDSCR	$\Delta I_a(U,G)$	$\Delta I_a(U,LIP)$	$\Delta I_a(U,FIT)$	$\Delta I_a(U,LHS)$
Perry County	PA	71	-0.066	-0.015	0.007	0.009
Sedgwick County	KS	177	0.008	0.009	-0.010	0.001
Mercer County	PA	66	0.003	0.011	0.002	0.003
Montgomery County	KS	50	0.003	0.015	0.006	0.007
Westmoreland County	PA	47	0.013	0.002	0.005	-0.015
Boyd County	KY	112	0.004	0.010	-0.007	0.012
Northampton County	PA	153	0.002	-0.011	-0.011	0.004
Riley County	KS	59	0.003	0.011	-0.008	-0.044
Montgomery County	PA	140	0.014	-0.054	-0.046	-0.025
Pottawatomie County	KS	48	0.004	-0.022	0.001	-0.033
Lebanon County	PA	101	-0.007	-0.009	0.002	0.011
Monroe County	PA	173	-0.008	0.000	-0.003	-0.009
Wyoming County	PA	45	-0.014	0.001	0.001	-0.005

Table 5. Names, symbols, and formula for the calculation of different I_d(k) descriptors

Symbol	D-symbol	Name	Formula	Ref.
I_d(tot)	I	Total information content	$\mathrm{I} = \mathrm{n}\log_2 \mathrm{n} - \sum_{\mathrm{g}=1}^{\mathrm{G}} \mathrm{n}_\mathrm{g} \log_2 \mathrm{n}_\mathrm{g}$	(Shannon and Weaver, 1949)
I_d(avg)	$\bar{I}$	Mean information content	$\bar{I} = -\sum_{g=1}^{G} \frac{n_g}{n} \log_2 \frac{n_g}{n}$	(Shannon and Weaver, 1949)
I_d(siz)	ISIZ	Information index on molecular size	$ISIZ = nAT.\log_2 nAT$	(Bertz, 1981)
I_d(ac)	IAC	Total information index on atomic composition	$\mathrm{I} = \mathrm{n}\log_2 \mathrm{n} - \sum_{\mathrm{g}=1}^{\mathrm{G}} \mathrm{n}_\mathrm{g} \log_2 \mathrm{n}_\mathrm{g}$	(Dancoff and Quastler, 1953)
I_d(aac)	AAC	Mean information index on atomic composition	$\bar{I} = -\sum_{g=1}^{G} \frac{n_g}{n} \log_2 \frac{n_g}{n}$	(Dancoff and Quastler, 1953)
I_d(det) I_d(de)	IDET, IDE	Total and mean information content on the distance equality, respectively	Equality of topological distances in a H-depleted molecular graph.	(Bonchev and Trinajstic, 1978)
I_d(dmt), I_d(dm)	IDMT, IDM	Total and mean information content on the distance magnitude, respectively	Distribution of topological distances according to their magnitude in a H-depleted molecular graph	
I_d(dde)	IDDE	Mean information content on the distance degree equality	Partition of vertex distance degrees according to their equality	
I_d(ddm)	IDDM	Mean information content on the distance degree magnitude	Partition of vertex distance degrees according to their magnitude	
I_d(vde)	IVDE	Mean information content on the vertex degree equality	Partition of vertices according to vertex degree equality	
I_d(vdm)	IVDM	Mean information content on the vertex degree magnitude	Partition of vertices according to the vertex degree magnitude	(Raychaudhury et al., 1984)
I_d(hvcpx)	HVcpx	Graph vertex complexity index	$HVcpx = \frac{1}{nSK}.\sum_{i=1}^{nSK} \left(-\sum_{g=0}^{\eta_i} \frac{{}^g f_j}{nSK}.\log_2 \frac{{}^g f_j}{nSK} \right)$	(Raychaudhury et al., 1984)
I_d(hdcpx)	HDcpx	Graph distance complexity index	$HDcpx = \sum_{i=1}^{nSK} \frac{\sigma_i}{2W}.\left(-\sum_{j=1}^{nSK} \frac{d_{ij}}{\sigma_i}.\log_2 \frac{d_{ij}}{\sigma_i} \right)$	(Klopman et al., 1988; Raychaudhury et al., 1984)

Table 6.TWJCA and PCA of average values of information indices for drugs $<I_d(k)>_{cj}$ and counties $<I_a(v)>_L$

TWJCA [a]	Inputs	Factor Name	IDB	ODB	Mean	SD
Bio-molecular factors						
HM1	$<I_d(k)>_{exp}$	Experimental Measure	130	49	$-0.1\cdot10^{-7}$	0.95
HM2	$<I_d(k)>_{target}$	Drug Targets	130	32	$0.1\cdot10^{-9}$	0.94
HM3	$<I_d(k)>_{org}$	Organism of Assay	65	23	$-0.1\cdot10^{-7}$	0.89
Socio-economic factors						
HM1	$<I_a(v)_r>$	RUCC	153	48	$0.1\cdot10^{-9}$	0.94
HM2	$<I_a(v)_u>$	UIC	204	58	$0.1\cdot10^{-7}$	0.96
HM3	$<I_a(v)_s>$	STATES	799	307	$-0.1\cdot10^{-7}$	0.99
PCA [b]	Inputs	Factor Name	EV(%)	VAR(%)	CEV(%)	CVAR(%)
Bio-molecular factors						
PC1	$<I_d(k)_{cj}>$	Exp. Measure *vs*. Organism	28.4	43.6	28.4	43.6
PC2		Drug Structure	8.3	12.8	36.7	56.4
PC3		Pharmacological Assay	6.8	10.4	43.4	66.8
PC4		Drug Target	5.1	7.8	48.5	74.6
Socio-economic factors						
PC1	$<I_a(v)_L>$	Population *vs*. Employment	24.9	36.6	24.9	36.6
PC2		Education *vs*. Poverty	7.8	11.5	32.7	48.1
PC3		Domestic *vs*. Net Migration	7.4	10.9	40.1	59.0
PC4		Education Level	5.3	7.8	45.4	66.8
PC5		Other Factors	3.9	5.8	49.3	72.5

[a] TWJCA = Two-Way Joining Cluster Analysis, HM = Hot Maps (**Figure 4** and **Figure 5**), IDB = Input Data Blocks, ODB = Output Data Blocks, SD = Standard Deviation (threshold value = SD/2).

[b] PCA = Principal Component Analysis (**Figure 6**), EV = Eigenvalue, CEV = Cumulative Eigenvalues, VAR = Variance, CVAR = Cumulative Variance

Table 7. Parameters of generated ANNs

Population	**Net. name**	**Training algorithm**	**Error function**	**Hidden activation**	**Output activation**
All	MLP 90-14-2	BFGS 193	Entropy	Logistic	Softmax
State	MLP 56-23-2	BFGS 68	SOS	Logistic	Logistic
RUCC	MLP 56-15-2	BFGS163	Entropy	Identity	Softmax
UIC	MLP 56-15-2	BFGS140	SOS	Tanh	Identity
Population	**Net. name**	**Training algorithm**	**Error function**	**Activation**	**Hidden layers**
All	LNN 87:87-1:1	Pseudoinverse	Entropy	Identity	0
State	LNN 52:52-1:1				
RUCC	LNN 53:53-1:1				
UIC	LNN 54:54-1:1				

BFGS= Broyden-Fletcher-Goldfarb-Shanno, or Quasi-Newton. SOS= sum of squares.

Table 8.ALMA models based on ANN classifiers found with STATISTICA using original data

Level	ANN models		Training		Selection		Validation	
State		Observed	L_{ac} = 1	L_{ac} = 0	L_{ac} = 1	L_{ac} = 0	L_{ac} = 1	L_{ac} = 0
	MLP 56-23-2	Parameter [a]	Sn	Sp	Sn	Sp	Sn	Sp
		Predicted	16.58	98.20	10.98	98.42	12.84	97.48
		L_{ac}= 1	2189	1177	503	342	556	551
		L_{ac} = 0	11006	64380	4074	21331	3772	21371
	LNN 52:52-1:1	Parameter [a]	Sn	Sp	Sn	Sp	Sn	Sp
		Predicted	77.16	75.32	73.89	75.77	73.23	76.32
		L_{ac} = 1	10182	16174	3198	5310	3352	5131
		L_{ac} = 0	3013	49383	1130	16612	1225	16542
RUCC		Observed	L_{ac} = 1	L_{ac} = 0	L_{ac} = 1	L_{ac} = 0	L_{ac} = 1	L_{ac} = 0
	MLP 56-15-2	Parameter [a]	Sn	Sp	Sn	Sp	Sn	Sp
		Predicted	31	96	28	96	30	96
		L_{ac}= 1	4213	2281	1320	681	1329	770
		L_{ac} = 0	8982	63276	3257	20992	2999	21152
	LNN 53:53-1:1	Parameter [a]	Sn	Sp	Sn	Sp	Sn	Sp
		Predicted	79.37	77.07	76.38	77.38	73.30	79.22
		L_{ac} = 1	10473	15031	3306	4957	3355	4502
		L_{ac} = 0	2722	50526	1022	16965	1222	17171
UIC		Observed	L_{ac} = 1	L_{ac} = 0	L_{ac} = 1	L_{ac} = 0	L_{ac} = 1	L_{ac} = 0
	MLP 56-15-2	Parameter [a]	Sn	Sp	Sn	Sp	Sn	Sp
		Predicted	34.99	97.15	26.69	96.89	28.14	97.13
		L_{ac}= 1	4618	1865	1222	674	1218	627
		L_{ac} = 0	8577	63692	3355	20999	3110	21295
	LNN 54:54-1:1	Parameter [a]	Sn	Sp	Sn	Sp	Sn	Sp
		Predicted	79.67	77.07	76.73	77.52	72.97	79.03
		L_{ac} = 1	10513	15027	3321	4926	3340	4544
		L_{ac} = 0	2682	50530	1007	16996	1237	17129
ALL		Observed	L_{ac} = 1	L_{ac} = 0	L_{ac} = 1	L_{ac} = 0	L_{ac} = 1	L_{ac} = 0
	MLP 90-14-2	Parameter [a]	Sn	Sp	Sn	Sp	Sn	Sp
		Predicted	58.55	96.18	49.09	94.80	48.93	94.32
		L_{ac}= 1	7726	2498	2247	1125	2118	1243
		L_{ac} = 0	5469	63059	2330	20548	2210	20679
	LNN 87:87-1:1	Parameter[a]	Sn	Sp	Sn	Sp	Sn	Sp
		Predicted	80.93	79.83	80.26	80.31	77.14	80.99
		L_{ac}= 1	10680	13221	3474	4316	3531	4118
		L_{ac} = 0	2515	52336	854	17606	1046	17555
PCA		Observed	L_{ac} = 1	L_{ac} = 0	L_{ac} = 1	L_{ac} = 0	L_{ac} = 1	L_{ac} = 0
	MLP 6:6-8-1:1	Parameter [a]	Sn	Sp	Sn	Sp	Sn	Sp
		Predicted	61.42	57.07	52.86	60.55	53.68	57.95
		L_{ac}= 1	8104	28146	2288	8649	2457	9113
		L_{ac} = 0	5091	37411	2040	13273	2120	12560
	LNN 7:7-1:1	Parameter[a]	Sn	Sp	Sn	Sp	Sn	Sp

		Predicted	58.54	56.84	51.15	58.46	55.25	52.34
		L_{ac}= 1	7725	28295	2214	9107	2529	10330
		L_{ac} = 0	5470	37262	2114	12815	2048	11343

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

Table 9. Results for models obtained with WEKA before and after obtaining the SMOTED data

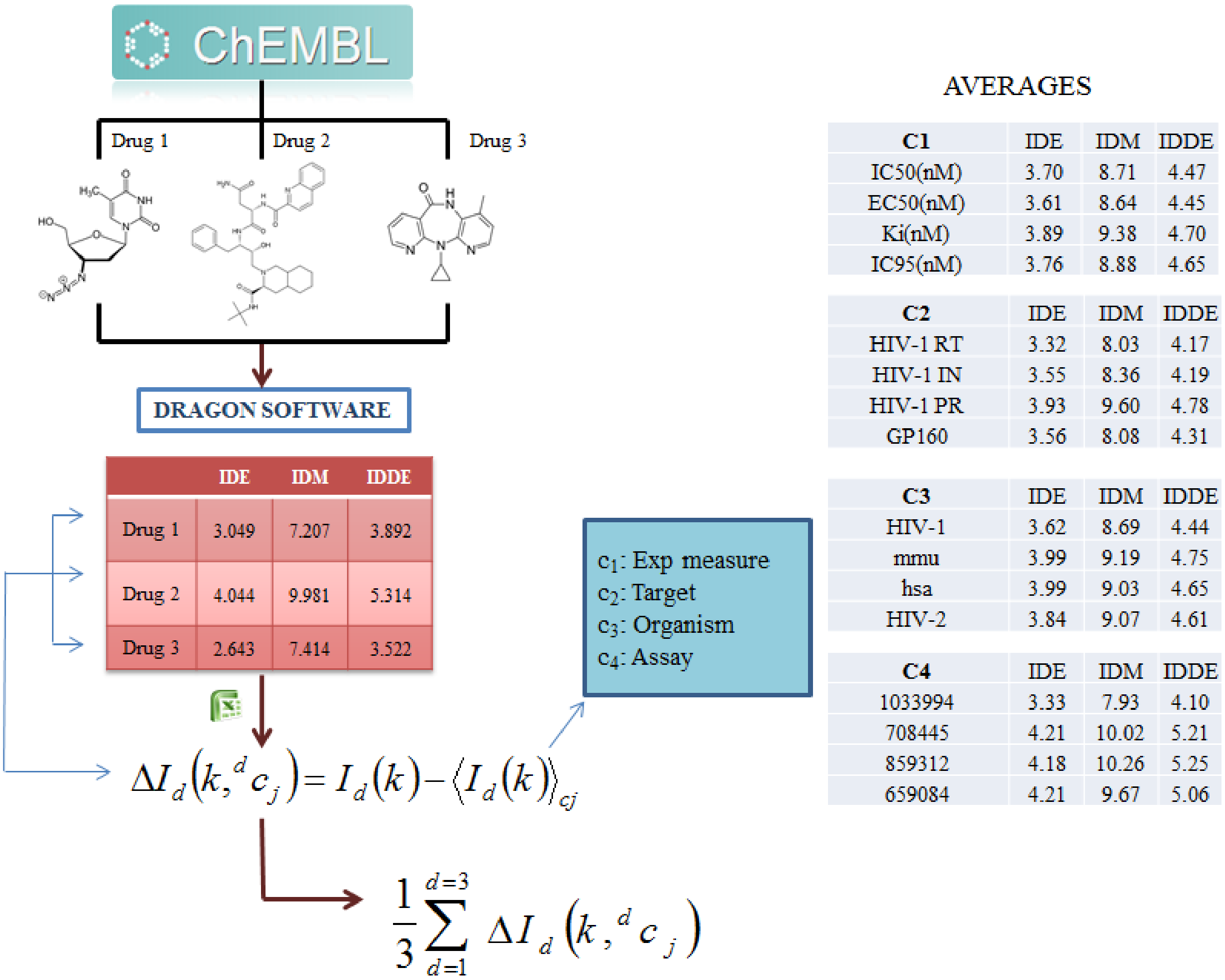
WEKA models [a]	Parameters			
Original data [b]	Precision	Recall	F-Measure	AUROC
VP	19.5	33.6	24.6	52.9
MLP	59.3	57.3	58.3	86.2
RNDF	60.3	43.3	50.4	82.1
SMOTE data filter [c]	Precision	Recall	F-Measure	AUROC
VP	50.1	74.9	60.0	49.9
MLP	94.2	90.1	92.1	97.4
RNDF	91.7	89.4	90.5	95.9

[a]VP = Voted Perceptron, MLP = Multi-Layer Perceptron, RNDF = Random Forest.
[b]Models obtained with WEKA before pre-processing of data.
[c] Models obtained with WEKA using SMOTED data.

Table 10.Predicted probabilities, p(halt), with which AIDS could be halt in a county with a HAART cocktail.

NY County	County frequency	County in-degree	p(halt)
Bronx	57	48	0.84
Queens	57	43	0.75
New York	56	41	0.73
Kings	56	39	0.70
Westchester	57	30	0.53
Jefferson	56	18	0.32
Orange	56	17	0.29
Rockland	56	16	0.29
Dutchess	57	16	0.28
Chemung	57	14	0.25

FIGURES TO BE INSERTED IN THE TEXT



	IDE	IDM	IDDE
Drug 1	3.049	7.207	3.892
Drug 2	4.044	9.981	5.314
Drug 3	2.643	7.414	3.522

$$\Delta I_d\left(k, {}^d c_j\right) = I_d(k) - \left\langle I_d(k)\right\rangle_{cj}$$

$$\frac{1}{3}\sum_{d=1}^{d=3} \Delta I_d\left(k, {}^d c_j\right)$$

AVERAGES

C1	IDE	IDM	IDDE
IC50(nM)	3.70	8.71	4.47
EC50(nM)	3.61	8.64	4.45
Ki(nM)	3.89	9.38	4.70
IC95(nM)	3.76	8.88	4.65

C2	IDE	IDM	IDDE
HIV-1 RT	3.32	8.03	4.17
HIV-1 IN	3.55	8.36	4.19
HIV-1 PR	3.93	9.60	4.78
GP160	3.56	8.08	4.31

C3	IDE	IDM	IDDE
HIV-1	3.62	8.69	4.44
mmu	3.99	9.19	4.75
hsa	3.99	9.03	4.65
HIV-2	3.84	9.07	4.61

C4	IDE	IDM	IDDE
1033994	3.33	7.93	4.10
708445	4.21	10.02	5.21
859312	4.18	10.26	5.25
659084	4.21	9.67	5.06

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

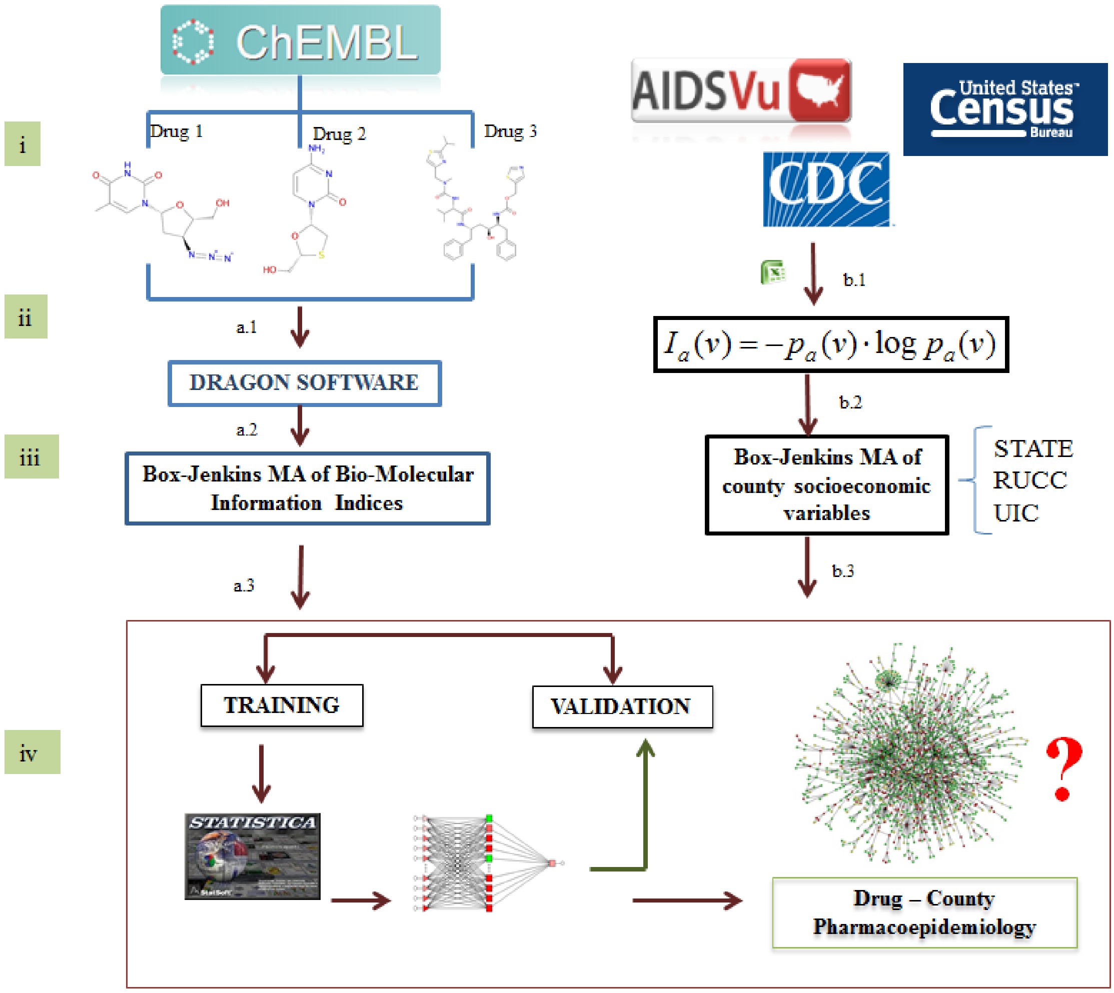

ChEMBL
i
Drug 1
Drug 2
Drug 3
AIDSVu
United States Census Bureau
CDC
ii
a.1
b.1
DRAGON SOFTWARE
$I_a(v) = -p_a(v) \cdot \log p_a(v)$
iii
a.2
b.2
Box-Jenkins MA of Bio-Molecular Information Indices
Box-Jenkins MA of county socioeconomic variables
STATE
RUCC
UIC
a.3
b.3
iv
TRAINING
VALIDATION
STATISTICA
?
Drug – County Pharmacoepidemiology

Figure 2. Flowchart used to construct the ANNs for the AIDS Pharmacoepidemiology model in U.S

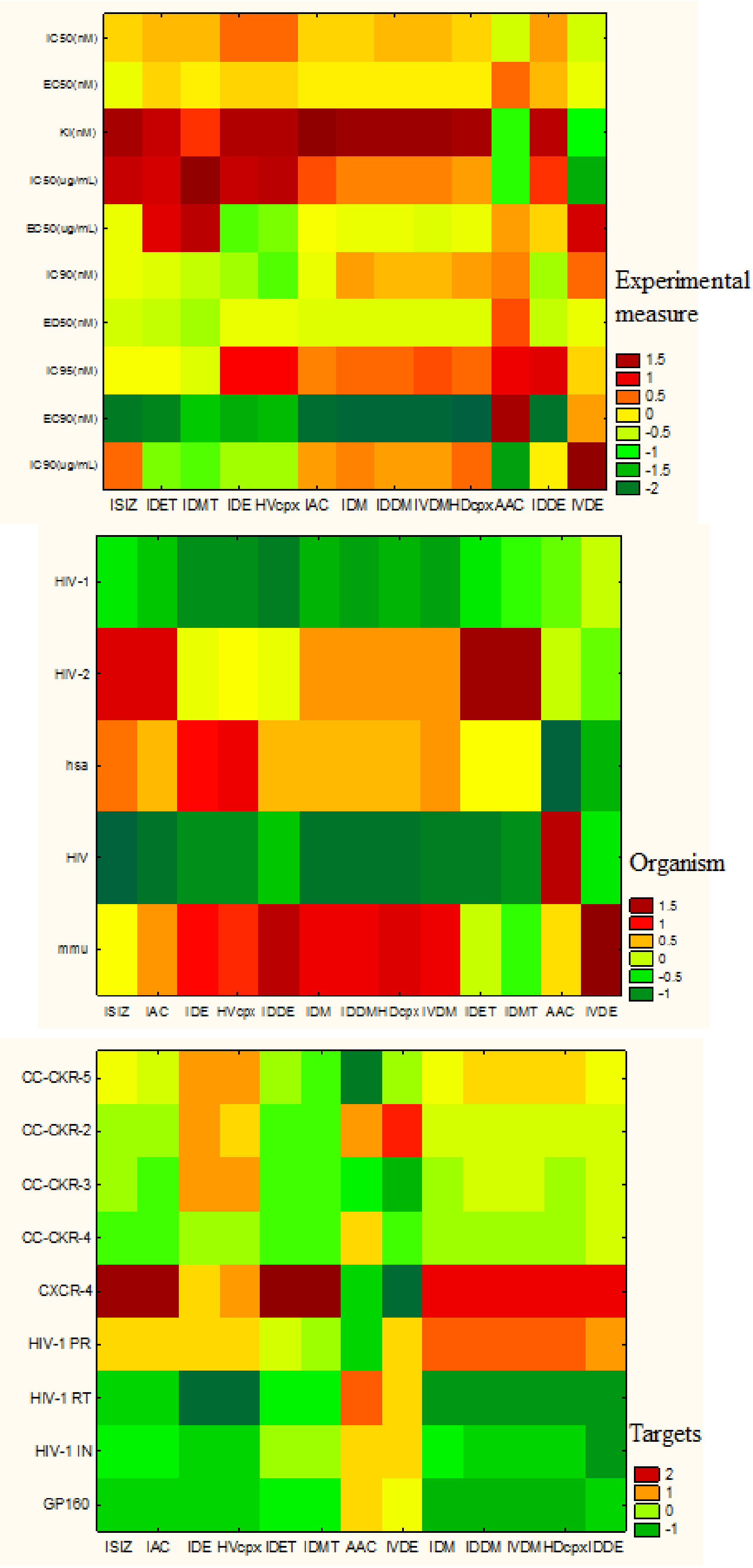


Figure 3.Hot Maps (HM) picture of TWJCA results withaverage values $<I_d(k)>_{cj}$ of the information indices $I_d(k)$ for different Bio-molecular factors (c_j)

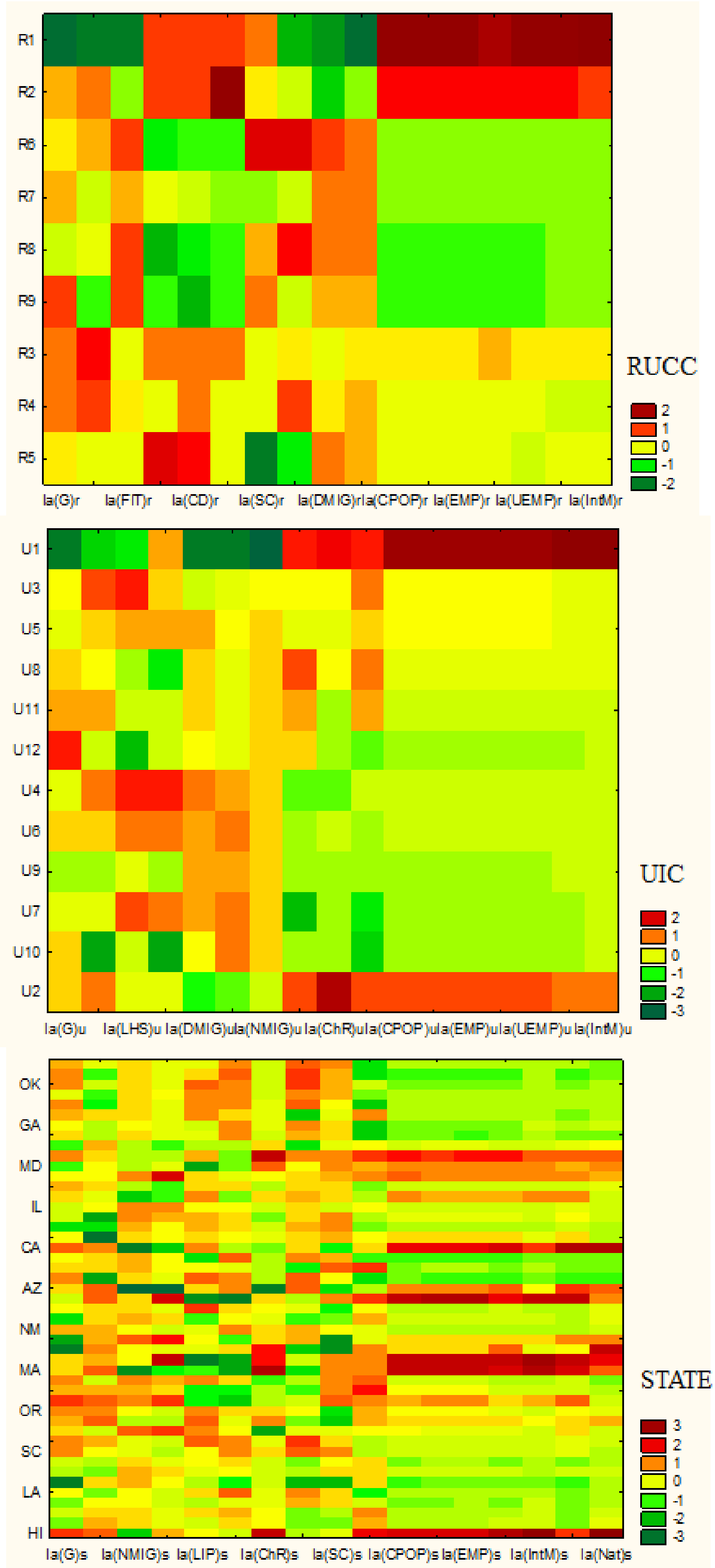
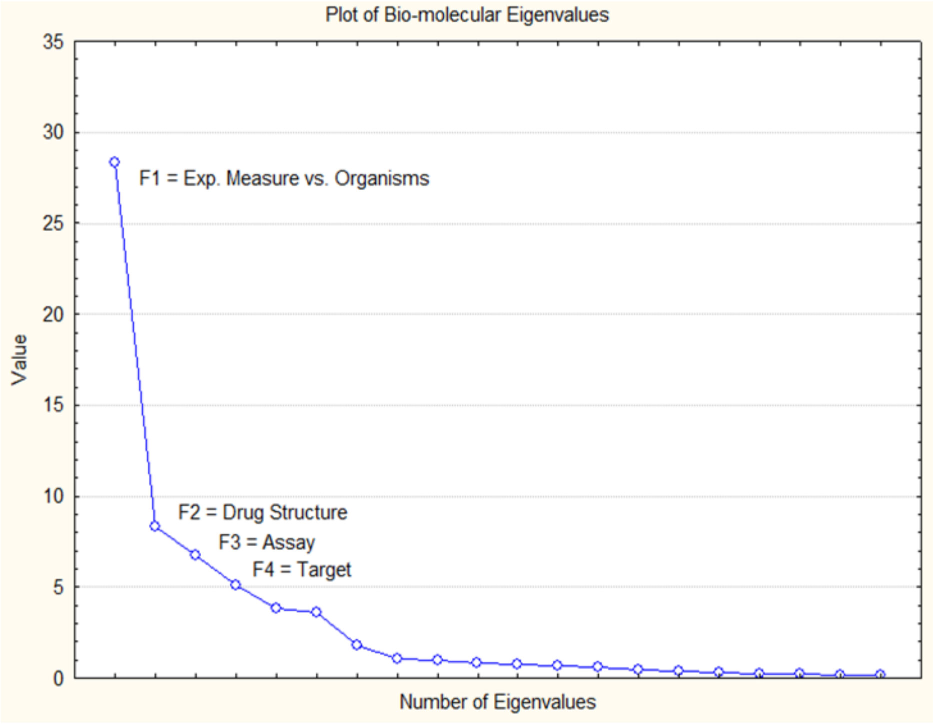


Figure4. Hot Maps (HM) of TWJCA results with average values $<I_a(k)>_{cj}$ of the information indices $I_a(k)$ for different Socio-Economic factors structure population(c_j = R = RUCC level, U = UIC level, S = State level)



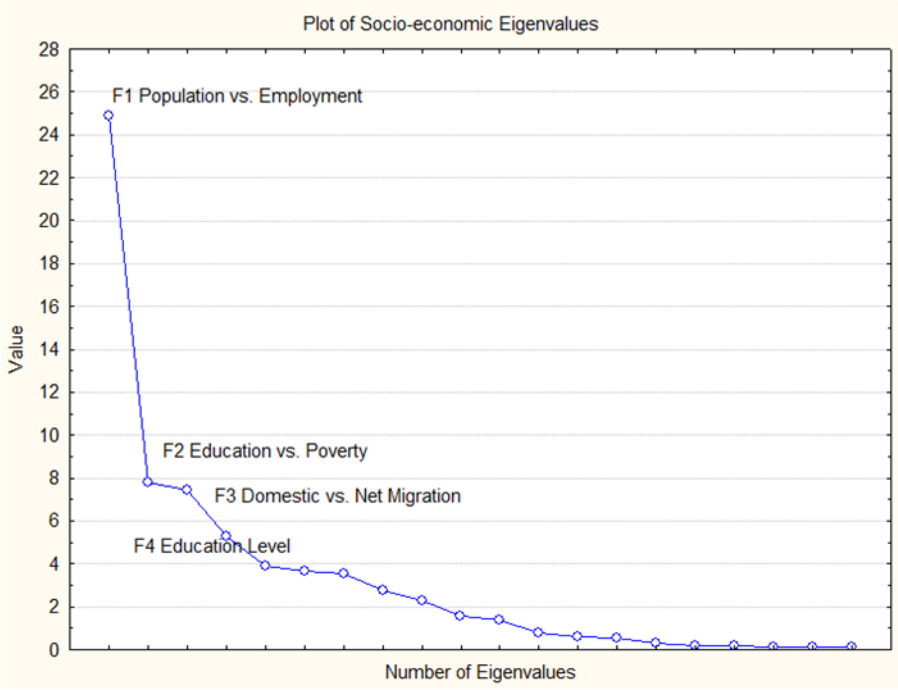


Figure 5. Plot of bio-molecular and socioeconomic eigenvalues for PCA of average values

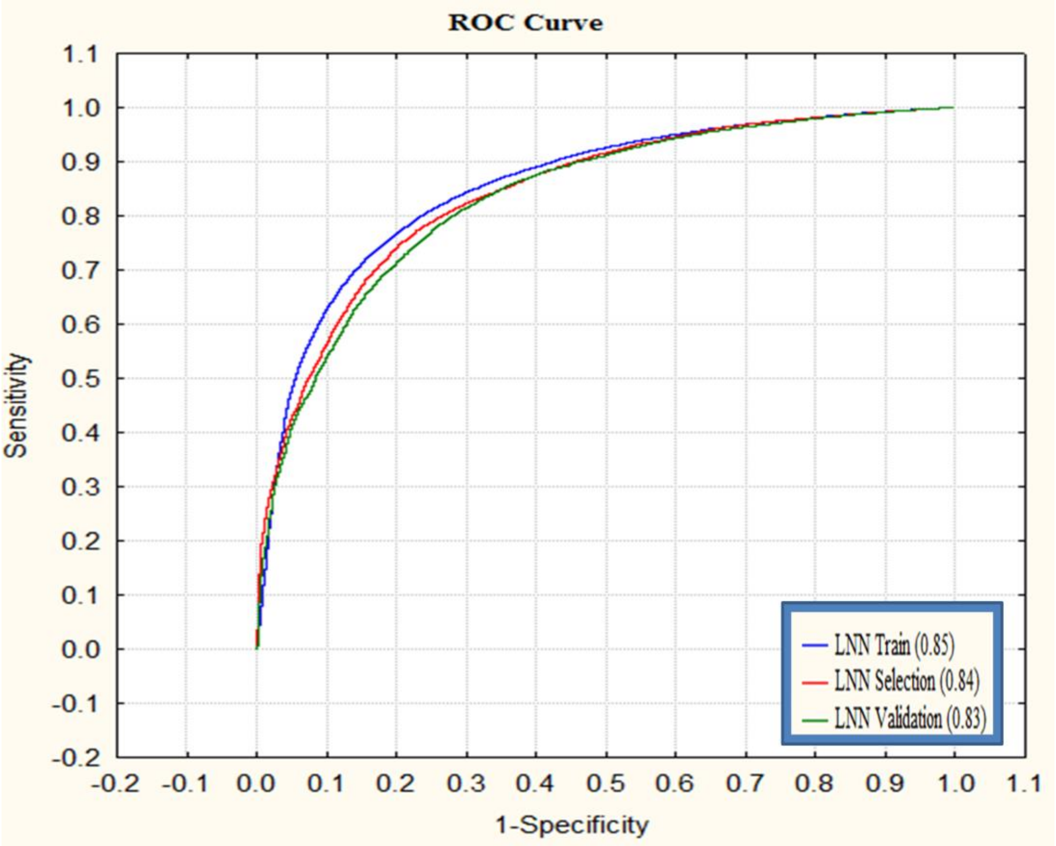


Figure 6.ROC for ALMA-LNN model with MA of Socio-Economic factors relative to UIC codes

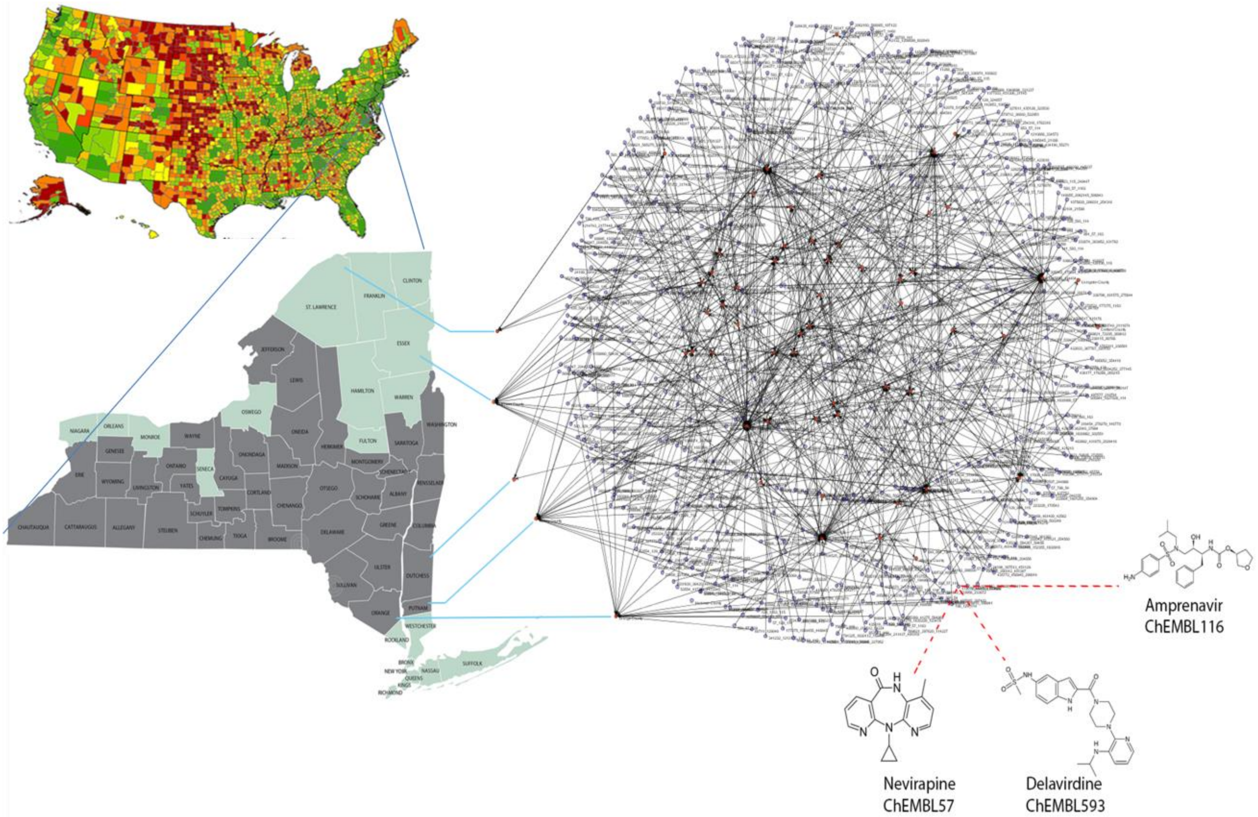


Figure 7. Predicted sub-network of HAART cocktails *vs.* AIDS prevalence for New York (NY) state

6. BIBLIOGRAFÍA

Adams, J. L., Greener, B. N., & Kashuba, A. D. (2012). Pharmacology of HIV integrase inhibitors. *Current opinion in HIV and AIDS, 7*(5), 390-400. doi: 10.1097/COH.0b013e328356e91c

Alkhatib, G. (2009). The biology of CCR5 and CXCR4. *Current opinion in HIV and AIDS, 4*(2), 96-103. doi: 10.1097/COH.0b013e328324bbec

Alonso, N., Caamano, O., Romero-Duran, F. J., Luan, F., MN, D. S. C., Yanez, M., . . . Garcia-Mera, X. (2013). Model for high-throughput screening of multitarget drugs in chemical neurosciences: synthesis, assay, and theoretic study of rasagiline carbamates. *ACS Chemical Neuroscience, 4*(10), 1393-1403. doi: 10.1021/cn400111n

Amaral, L. A., Scala, A., Barthelemy, M., & Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences of the United States of America, 97*(21), 11149-11152. doi: 10.1073/pnas.200327197

Araujo, R. P., Liotta, L. A., & Petricoin, E. F. (2007). Proteins, drug targets and the mechanisms they control: the simple truth about complex networks. *Nature reviews. Drug discovery, 6*(11), 871-880. doi: 10.1038/nrd2381

Arts, E. J., & Hazuda, D. J. (2012). HIV-1 antiretroviral drug therapy. *Cold Spring Harbor perspectives in medicine, 2*(4), a007161. doi: 10.1101/cshperspect.a007161

Balaban, A. T., & Balaban, T. S. (1991). New Vertex Invariants and Topological Indices of Chemical Graphs Based on Information on Distances. *Journal of Mathematical Chemistry, 8*, 383-397.

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science (New York, N.Y.), 286*(5439), 509-512.

Barabasi, A. L., & Bonabeau, E. (2003). Scale-free networks. *Scientific American, 288*(5), 60-69.

Barabasi, A. L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature reviews. Genetics, 12*(1), 56-68. doi: 10.1038/nrg2918

Barnett, T., & Whiteside, A. (2006). *AIDS in the twenty-first century: Disease and globalization.* (2nd ed.). New York: Palgrave Macmillan.

Bauch, C., & Rand, D. A. (2000). A moment closure model for sexually transmitted disease transmission through a concurrent partnership network. *Proceedings. Biological sciences / The Royal Society, 267*(1456), 2019-2027. doi: 10.1098/rspb.2000.1244

Beas Zárate, C., Ureña Guerrero, M., Rivera Cervantes, M., Pallàs i Llibería, M., & Camins Espuny, A. (2010). *Tópicos de Actualización en Neurobiologia: Excitotoxicidad y Cognición en Enfermedades Neurogenerativas: Aspectos Básicos, Clínicos y Sociales, Memorias Electrónicas* (pp. 493).

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., . . . Overington, J. P. (2013a). The ChEMBL bioactivity database: an update. *Nucleic Acids Research*. doi: 10.1093/nar/gkt1031

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., . . . Overington, J. P. (2013b). The ChEMBL bioactivity database: an update. *Nucleic Acids Research, 42*, D1083-1090. doi: 10.1093/nar/gkt1031

Bertz, S. H. (1981). The first general index of molecular complexity. . *Journal of the American Chemical Society, 103*, 3599-3601.

Betancourt, G. (2005). Las maquinas de soporte vectorial. *Scientia et Technica Año XI, 27*.

Blanpain, C., Libert, F., Vassart, G., & Parmentier, M. (2002). CCR5 and HIV infection. *Receptors & channels, 8*(1), 19-31.

Bonchev, D., & Trinajstic, N. (1978). On topological characterization of molecular branching. *International Journal of Quantum Chemistry, 12*, 293-303.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. San Francisco, California: Holden-Day.

Briem, H., & Gunther, J. (2005). Classifying "kinase inhibitor-likeness" by using machine-learning methods. *Chembiochem : a European journal of chemical biology, 6*(3), 558-566. doi: 10.1002/cbic.200400109

Carmona, A. (2014). Resistencia a los farmacos antirretrovirales. Retrieved 20/11/2014, from http://www.sefh.es/bibliotecavirtual/2_AF_VIH_2002/7_resistencias.pdf

Castro, H. C., Abreu, P. A., Geraldo, R. B., Martins, R. C., dos Santos, R., Loureiro, N. I., . . . Rodrigues, C. R. (2011). Looking at the proteases from a simple perspective. *Journal of molecular recognition : JMR, 24*(2), 165-181. doi: 10.1002/jmr.1091

Colombo, G. L., Castagna, A., Di Matteo, S., Galli, L., Bruno, G., Poli, A., . . . Lazzarin, A. (2014). Cost analysis of initial highly active antiretroviral therapy regimens for managing human immunodeficiency virus-infected patients according to clinical practice in a hospital setting. *Therapeutics and clinical risk management, 10*, 9-15. doi: 10.2147/tcrm.s49428

Coutinho, B., & Prasad, R. (2013). Emtricitabine/Tenofovir (Truvada) for HIV Prophylaxis. *American family physician, 88*(8), 535-540.

Chen, M. P., Shang, N., Winston, C. A., & Becerra, J. E. (2012). A Bayesian analysis of the 2009 decline in tuberculosis morbidity in the United States. *Statistics in medicine, 31*, 3278-3284. doi: 10.1002/sim.5343

10.1002/sim.5343. Epub 2012 Mar 13.

Chougrani, I., Luton, D., Matheron, S., Mandelbrot, L., & Azria, E. (2013). Safety of protease inhibitors in HIV-infected pregnant women. *HIV/AIDS (Auckland, N.Z.), 5*, 253-262. doi: 10.2147/hiv.s33058

Dancoff, S. M., & Quastler, H. (1953). *Essays on the Use of Information Theory in Biology*: University of Illinois, Urbana

de Bethune, M. P. (2010). 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). *Antiviral research, 85*(1), 75-90. doi: 10.1016/j.antiviral.2009.09.008

Debnath, U., Verma, S., Jain, S., Katti, S. B., & Prabhakar, Y. S. (2013). Pyridones as NNRTIs against HIV-1 mutants: 3D-QSAR and protein informatics. *Journal of computer-aided molecular design, 27*(7), 637-654. doi: 10.1007/s10822-013-9667-1

Diestel, R. (2000). *Graph Theory* (2nd ed.). New York: Springer-Verlag.

Douali, L., Villemin, D., & Cherqaoui, D. (2003). Neural networks: Accurate nonlinear QSAR model for HEPT derivatives. *Journal of chemical information and computer sciences, 43*(4), 1200-1207. doi: 10.1021/ci034047q

Emory University, R. S. o. P. H. AIDSVu. Retrieved accessed September 21, 2013, from http://aidsvu.org/

Eron, J. J., Jr. (2000). HIV-1 protease inhibitors. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 30 Suppl 2*, S160-170. doi: 10.1086/313853

Ferguson, N. M., Donnelly, C. A., & Anderson, R. M. (2001). Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature, 413*(6855), 542-548. doi: 10.1038/35097116

Gadhe, C. G., Kothandan, G., & Cho, S. J. (2013). Binding site exploration of CCR5 using in silico methodologies: a 3D-QSAR approach. *Archives of pharmacal research, 36*(1), 6-31. doi: 10.1007/s12272-013-0001-1

Gant, Z., Lomotey, M., Hall, H. I., Hu, X., Guo, X., & Song, R. (2012). A County-Level Examination of the Relationship Between HIV and Social Determinants of Health: 40 States, 2006-2008. *The open AIDS journal, 6,* 1-7. doi: 10.2174/1874613601206010001

10.2174/1874613601206010001. Epub 2012 Feb 21.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., . . . Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research, 40*(Database issue), D1100-1107. doi: 10.1093/nar/gkr777

Ghani, A. C., & Garnett, G. P. (2000). Risks of acquiring and transmitting sexually transmitted diseases in sexual partner networks. *Sexually transmitted diseases, 27*(10), 579-587.

Gillespie, S., Kadiyala, S., & Greener, R. (2007). Is poverty or wealth driving HIV transmission? *AIDS (London, England), 21 Suppl 7,* S5-S16. doi: 10.1097/01.aids.0000300531.74730.72

Godden, J. W., Stahura, F. L., & Bajorath, J. (2000). Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. *Journal of chemical information and computer sciences, 40*, 796-800.

Gonzalez-Diaz, H., Aguero, G., Cabrera, M. A., Molina, R., Santana, L., Uriarte, E., . . . Castanedo, N. (2005). Unified Markov thermodynamics based on stochastic forms to classify drugs considering molecular structure, partition system, and biological species: distribution of the antimicrobial G1 on rat tissues. *Bioorganic & medicinal chemistry letters, 15,* 551-557. doi: 10.1016/j.bmcl.2004.11.059.

Gonzalez-Diaz, H., Cruz-Monteagudo, M., Molina, R., Tenorio, E., & Uriarte, E. (2005). Predicting multiple drugs side effects with a general drug-target interaction thermodynamic Markov model. *Bioorganic and Medicinal Chemistry, 13*, 1119-1129. doi: 10.1016/j.bmc.2004.11.030

González-Díaz, H., Herrera-Ibatá, D. M., Duardo-Sanchez, A., Munteanu, C. R., Orbegozo-Medina, R. A., & Pazos, A. (2014). 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. *Journal of chemical information and modeling, 54*(3), 744-755.

Gonzalez-Diaz, H., Prado-Prado, F. J., Santana, L., & Uriarte, E. (2006). Unify QSAR approach to antimicrobials. Part 1: predicting antifungal activity against different species *Bioorganic and Medicinal Chemistry, 14*, 5973-5980. doi: 10.1016/j.bmc.2006.05.018.

Gupta, M., & Madan, A. K. (2012). Diverse models for the prediction of HIV integrase inhibitory activity of substituted quinolone carboxylic acids. *Archiv der Pharmazie, 345,* 989-1000. doi: 10.1002/ardp.201100316

10.1002/ardp.201100316. Epub 2012 Sep 4.

Gupta, P., Sharma, A., Garg, P., & Roy, N. (2013). QSAR study of curcumine derivatives as HIV-1 integrase inhibitors. *Current computer-aided drug design, 9*(1), 141-150.

Gupta, S., Singh, M., & Madan, A. K. (2001). Predicting anti-HIV activity: computational approach using a novel topological descriptor. *Journal of computer-aided molecular design, 15,* 671-678.

Haidich, A. B., & Ioannidis, J. P. (2004). The Gini coefficient as a measure for understanding accrual inequalities in multicenter clinical studies. *Journal of clinical epidemiology, 57*(4), 341-348. doi: 10.1016/j.jclinepi.2003.09.011

Hajizadeh, M., Sia, D., Heymann, S. J., & Nandi, A. (2014). Socioeconomic inequalities in HIV/AIDS prevalence in sub-Saharan African countries: evidence from the Demographic Health Surveys. *International journal for equity in health, 13*(1), 18. doi: 10.1186/1475-9276-13-18

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & I.H., W. (2009). The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10-18.

Hecker, N., Ahmed, J., von Eichborn, J., Dunkel, M., Macha, K., Eckert, A., . . . Preissner, R. (2012). SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Research, 40*(Database issue), D1113-1117. doi: 10.1093/nar/gkr912

Heikamp, K., & Bajorath, J. (2011). Large-scale similarity search profiling of ChEMBL compound data sets. *Journal of chemical information and modeling, 51*(8), 1831-1839. doi: 10.1021/ci200199u

Herrera-Ibatá, D. M., Pazos, A., Orbegozo-Medina, R. A., & Gonzalez-Diaz, H. (2014). Mapping networks of anti-HIV drug cocktails vs. AIDS epidemiology in the US counties. *Chemometr Intell Lab, 138,* 161-170.

Hicks, C., & Gulick, R. M. (2009). Raltegravir: the first HIV type 1 integrase inhibitor. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 48*(7), 931-939. doi: 10.1086/597290

Hill, T., & Lewicki, P. (2006). *STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining* (Vol. 1). Tulsa: StatSoft.

HIV surveillance--United States, 1981-2008. (2011). *MMWR. Morbidity and mortality weekly report, 60*, 689-693.

Hu, W. S., & Hughes, S. H. (2012). HIV-1 reverse transcription. *Cold Spring Harbor perspectives in medicine, 2*, pii: a006882. doi: 10.1101/cshperspect.a006882

Ivanciuc, O., Balaban, T. S., & Balaban, A. T. (1993). Chemical Graphs with Degenerate Topological Indices Based on Information on Distances. *Journal of Mathematical Chemistry, 14*, 21-33.

Karmon, S. L., & Markowitz, M. (2013). Next-generation integrase inhibitors : where to after raltegravir? *Drugs, 73*, 213-228. doi: 10.1007/s40265-013-0015-5

King, J., McCall, M., Cannella, A., Markiewicz, M. A., James, A., Hood, C. B., & Acosta, E. P. (2011). 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. *Journal of acquired immune deficiency syndromes (1999), 56*(5), e130-132.

Klopman, G., Raychaudhury, C., & Henderson, R. V. (1988). A new approach to structure-activity using distance information content of graph vertices: a study with phenylalkylamines. *Mathematical and Computer Modelling, 11*, 635-640.

Langenfeld, M. C., Cipani, E., & Borckardt, J. J. (2002). Hypnosis for the control of HIV/AIDS-related pain. *The International journal of clinical and experimental hypnosis, 50*, 170-188. doi: 10.1080/00207140208410097

Lee, W. G., Gallardo-Macias, R., Frey, K. M., Spasov, K. A., Bollini, M., Anderson, K. S., & Jorgensen, W. L. (2013). Picomolar Inhibitors of HIV Reverse Transcriptase Featuring Bicyclic Replacement of a Cyanovinylphenyl Group. *Journal of the American Chemical Society, 135*, 16705-16713. doi: 10.1021/ja408917n

Li, C., Fang, J. S., Lian, W. W., Pang, X. C., Liu, A. L., & Du, G. H. (2014). In vitro Antiviral Effects and 3D QSAR Study of Resveratrol Derivatives as Potent Inhibitors of Influenza H1N1 Neuraminidase. *Chemical Biology & Drug Design*. doi: 10.1111/cbdd.12425

Li, H., Ung, C. Y., Yap, C. W., Xue, Y., Li, Z. R., & Chen, Y. Z. (2006). Prediction of estrogen receptor agonists and characterization of associated molecular descriptors by statistical learning methods. *Journal of molecular graphics & modelling, 25*(3), 313-323. doi: 10.1016/j.jmgm.2006.01.007

Lim, T., Zelaya, C., Latkin, C., Quan, V. M., Frangakis, C., Ha, T. V., . . . Go, V. (2013). Individual-level socioeconomic status and community-level inequality as determinants of stigma towards persons living with HIV who inject drugs in Thai Nguyen, Vietnam. *Journal of the International AIDS Society, 16*(3 Suppl 2), 18637. doi: 10.7448/ias.16.3.18637

Lindemann, D., Steffen, I., & Pohlmann, S. (2013). Cellular entry of retroviruses. *Advances in experimental medicine and biology, 790*, 128-149. doi: 10.1007/978-1-4614-7651-1_7

Lopez Aspiroz, E., Santos Buelga, D., Cabrera Figueroa, S., Lopez Galera, R. M., Ribera Pascuet, E., Dominguez-Gil Hurle, A., & Garcia Sanchez, M. J. (2011). Population pharmacokinetics of lopinavir/ritonavir (Kaletra) in HIV-infected patients. *Therapeutic drug monitoring, 33*(5), 573-582. doi: 10.1097/FTD.0b013e31822d578b

Luan, F., Cordeiro, M. N., Alonso, N., Garcia-Mera, X., Caamano, O., Romero-Duran, F. J., . . . Gonzalez-Diaz, H. (2013). TOPS-MODE model of multiplexing neuroprotective effects of drugs and experimental-theoretic study of new 1,3-rasagiline derivatives potentially useful in neurodegenerative diseases. *Bioorganic and Medicinal Chemistry, 21*(7), 1870-1879. doi: 10.1016/j.bmc.2013.01.035

Ma, S., Ye, W., Ji, D., & Chen, H. F. (2013). Insight into the binding mode between HIV-1 integrase and pyrimidone analogue inhibitors with MD simulation and 3D-QSAR. *Medicinal chemistry (Shariqah (United Arab Emirates)), 9*(3), 420-433.

Magnuson, V. R., Harriss, D. K., & Basak, S. C. (1983) *Studies in Physical and Theoretical Chemistry; King, R.B.* (pp. 178-191). Amsterdam (The Netherlands): Elsevier.

Mathew, S., Faheem, M., Archunan, G., Ilyas, M., Begum, N., Jahangir, S., . . . Mathew, S. (2014). In silico studies of medicinal compounds against hepatitis C capsid protein from north India. *Bioinformatics and biology insights, 8*, 159-168. doi: 10.4137/bbi.s15211

McCulloch, W., & Pitts, W. (1943). A Logical Calculus of Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics, 5*(4), 115-133.

McDavid Harrison, K., Ling, Q., Song, R., & Hall, H. I. (2008). County-level socioeconomic status and survival after HIV diagnosis, United States. *Annals of epidemiology, 18*(12), 919-927. doi: 10.1016/j.annepidem.2008.09.003

McFarland, W., Chen, S., Hsu, L., Schwarcz, S., & Katz, M. (2003). Low socioeconomic status is associated with a higher rate of death in the era of highly active antiretroviral therapy, San Francisco. *Journal of acquired immune deficiency syndromes (1999), 33*(1), 96-103.

McMahon, J., Wanke, C., Terrin, N., Skinner, S., & Knox, T. (2011). Poverty, hunger, education, and residential status impact survival in HIV. *AIDS and behavior, 15*(7), 1503-1511. doi: 10.1007/s10461-010-9759-z

Meyers, L. A., Pourbohloul, B., Newman, M. E., Skowronski, D. M., & Brunham, R. C. (2005). Network theory and SARS: predicting outbreak diversity. *Journal of theoretical biology, 232*(1), 71-81. doi: 10.1016/j.jtbi.2004.07.026

Milgram, S. (1967). The small world problem. *Psychology today, 1*(1), 61-67.

Montaño Moreno, J. J. (2002). *Redes Neuronales Artificiales aplicadas al Análisis de Datos.* Universitat De Les Illes Balears, Palma de Mallorca.

Moonsamy, S., Dash, R. C., & Soliman, M. E. (2014). Integrated computational tools for identification of CCR5 antagonists as potential HIV-1 entry inhibitors: homology modeling, virtual screening, molecular dynamics simulations and 3D QSAR analysis. *Molecules (Basel, Switzerland), 19*(4), 5243-5265. doi: 10.3390/molecules19045243

Moss, J. A. (2013). HIV/AIDS Review. *Radiologic technology, 84*, 247-267; quiz p.268-270.

Noorizadeh, H., Sajjadifar, S., & Farmany, A. (2013). A quantitative structure-activity relationship study of anti-HIV activity of substituted HEPT using nonlinear models. *Medicinal chemistry research : an international journal for rapid communications on design and mechanisms of action of biologically active agents, 22*, 5442-5452. doi: 10.1007/s00044-013-0525-4

O'Neal, R. (2011). Rilpivirine and complera: new first-line treatment options. *BETA bulletin of experimental treatments for AIDS : a publication of the San Francisco AIDS foundation, 23*(4), 14-18.

Pabayo, R., Kawachi, I., & Gilman, S. E. (2013). Income inequality among American states and the incidence of major depression. *J Epidemiol Community Health*. doi: 10.1136/jech-2013-203093

Pabayo, R., Kawachi, I., & Gilman, S. E. (2014). Income inequality among American states and the incidence of major depression. *Journal of Epidemiology & Community Health, 68*(2), 110-115. doi: 10.1136/jech-2013-203093

Palma-Méndez, J. T., & Marín-Morales, R. (2008). *Inteligencia artificial: Técnicas, métodos y aplicaciones*. India: Mc Graw Hill.

Peltzer, K., & Pengpid, S. (2013). Socioeconomic factors in adherence to HIV therapy in low- and middle-income countries. *Journal of health, population, and nutrition, 31*(2), 150-170.

Perno, C. F. (2011). The discovery and development of HIV therapy: the new challenges. *Annali dell'Istituto superiore di sanita, 47*(1), 41-43. doi: 10.4415/ann_11_01_09

Perry, C. M. (2014). 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*(1), 75-97. doi: 10.1007/s40265-013-0158-4

Piot, P., Greener, R., & Russell, S. (2007). Squaring the circle: AIDS, poverty, and human development. *PLoS medicine, 4*(10), 1571-1575. doi: 10.1371/journal.pmed.0040314

Piot, P., & Quinn, T. C. (2013). Response to the AIDS pandemic--a global health model. *The New England journal of medicine, 368*, 2210-2218. doi: 10.1056/NEJMra1201533

Portsmouth, S. D., & Scott, C. J. (2007). The renaissance of fixed dose combinations: Combivir. *Therapeutics and clinical risk management, 3*(4), 579-583.

Powderly, W. G. (2010). Integrase inhibitors in the treatment of HIV-1 infection. *The Journal of antimicrobial chemotherapy, 65*(12), 2485-2488. doi: 10.1093/jac/dkq350

Prado-Prado, F., Garcia-Mera, X., Abeijon, P., Alonso, N., Caamano, O., Yanez, M., . . . Gonzalez-Diaz, H. (2011). Using entropy of drug and protein graphs to predict FDA drug-target network: theoretic-experimental study of MAO inhibitors and hemoglobin peptides from Fasciola hepatica. *European journal of medicinal chemistry, 46*(4), 1074-1094. doi: 10.1016/j.ejmech.2011.01.023

Prado-Prado, F., Garcia-Mera, X., Escobar, M., Alonso, N., Caamano, O., Yanez, M., & Gonzalez-Diaz, H. (2012). 3D MI-DRAGON: new model for the reconstruction of US FDA drug- target network and theoretical-experimental studies of inhibitors of rasagiline derivatives for AChE. *Current topics in medicinal chemistry, 12*(16), 1843-1865.

Prado-Prado, F. J., Borges, F., Uriarte, E., Perez-Montoto, L. G., & Gonzalez-Diaz, H. (2009). Multi-target spectral moment: QSAR for antiviral drugs vs. different viral species. *Analytica Chimica Acta, 651*(2), 159-164. doi: 10.1016/j.aca.2009.08.022

Qian, K., Morris-Natschke, S. L., & Lee, K. H. (2009). HIV entry inhibitors and their potential in HIV therapy. *Medicinal research reviews, 29*(2), 369-393. doi: 10.1002/med.20138

Qiu, X., & Liu, Z. P. (2011). Recent developments of peptidomimetic HIV-1 protease inhibitors. *Current medicinal chemistry, 18*(29), 4513-4537.

Raychaudhury, C., Ray, S. K., Ghosh, J. J., Roy, A. B., & Basak, S. C. (1984). Discrimination of Isomeric Structures Using Information Theoretic Topological Indices. *Journal of Computational Chemistry, 5*, 581-588.

Riera-Fernandez, P., Munteanu, C. R., Escobar, M., Prado-Prado, F., Martin-Romalde, R., Pereira, D., . . . Gonzalez-Diaz, H. (2012). 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. *Journal of theoretical biology, 293*, 174-188. doi: 10.1016/j.jtbi.2011.10.016

Rinaldo, C. R. (2013). HIV-1 Infection of CD4 T Cells by Professional Antigen Presenting Cells. *Scientifica, 2013*, 164203. doi: 10.1155/2013/164203

Sax, P. E., Tierney, C., Collier, A. C., Fischl, M. A., Mollan, K., Peeples, L., . . . Daar, E. S. (2009). Abacavir-lamivudine versus tenofovir-emtricitabine for initial HIV-1 therapy. *The New England journal of medicine, 361*(23), 2230-2240. doi: 10.1056/NEJMoa0906768

Shannon, C., & Weaver, W. (1949). *The Mathematical Theory of Communication*. United States: University of Illinois Press, Urbana

Shey, M., Kongnyuy, E. J., Shang, J., & Wiysonge, C. S. (2009). A combination drug of abacavir-lamivudine-zidovudine (Trizivir) for treating HIV infection and AIDS. *The Cochrane database of systematic reviews*(3), CD005481. doi: 10.1002/14651858.CD005481.pub2

Shey, M. S., Kongnyuy, E. J., Alobwede, S. M., & Wiysonge, C. S. (2013). Co-formulated abacavir-lamivudine-zidovudine for initial treatment of HIV infection and AIDS. *The Cochrane database of systematic reviews, 3*, CD005481. doi: 10.1002/14651858.CD005481.pub3

Sirois, S., Tsoukas, C. M., Chou, K. C., Wei, D., Boucher, C., & Hatzakis, G. E. (2005). Selection of molecular descriptors with artificial intelligence for the understanding of HIV-1 protease peptidomimetic inhibitors-activity. *Medicinal chemistry (Shariqah (United Arab Emirates)), 1*(2), 173-184.

So, S. S., & Karplus, M. (1997). Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 2. Applications. *Journal of medicinal chemistry, 40*(26), 4360-4371. doi: 10.1021/jm970488n

Speck-Planche, A., Kleandrova, V. V., & Cordeiro, M. N. (2013). Chemoinformatics for rational discovery of safe antibacterial drugs: simultaneous predictions of biological activity against streptococci and toxicological profiles in laboratory animals. *Bioorganic and Medicinal Chemistry, 21*(10), 2727-2732. doi: 10.1016/j.bmc.2013.03.015

Speck-Planche, A., Kleandrova, V. V., Luan, F., & Cordeiro, M. N. (2011). Multi-target drug discovery in anti-cancer therapy: fragment-based approach toward the design of potent and versatile anti-prostate cancer agents. *Bioorganic and Medicinal Chemistry, 19*, 6239-6244. doi: 10.1016/j.bmc.2011.09.015

Speck-Planche, A., Kleandrova, V. V., Luan, F., & Cordeiro, M. N. (2012a). Chemoinformatics in anti-cancer chemotherapy: multi-target QSAR model for the in silico discovery of anti-breast cancer agents. *European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences, 47*(1), 273-279. doi: 10.1016/j.ejps.2012.04.012

Speck-Planche, A., Kleandrova, V. V., Luan, F., & Cordeiro, M. N. (2012b). Chemoinformatics in multi-target drug discovery for anti-cancer therapy: in silico design of potent and versatile anti-brain tumor agents. *Anti-cancer agents in medicinal chemistry, 12*(6), 678-685.

Speck-Planche, A., Kleandrova, V. V., Luan, F., & Cordeiro, M. N. (2012c). In silico discovery and virtual screening of multi-target inhibitors for proteins in Mycobacterium tuberculosis. *Combinatorial chemistry & high throughput screening, 15*, 666-673.

Speck-Planche, A., Kleandrova, V. V., Luan, F., & Cordeiro, M. N. (2012d). Multi-target inhibitors for proteins associated with Alzheimer: in silico discovery using fragment-based descriptors. *Current Alzheimer research, 10*, 117-124.

. STATISTICA (Version version 6.0). (2001). Tulsa, Oklahoma: StatSoft Inc.

StatSoft.Inc. (2002). STATISTICA (data analysis software system), version 6.0, www.statsoft.com.Statsoft, Inc. (Version 6.0).

Sun, X. H., Guan, J. Q., Tan, J. J., Liu, C., & Wang, C. X. (2012). 3D-QSAR studies of quinoline ring derivatives as HIV-1 integrase inhibitors. *SAR and QSAR in Environmental Research, 23*(7-8), 683-703. doi: 10.1080/1062936x.2012.717541

Swiderek, K., Marti, S., & Moliner, V. (2012). Theoretical studies of HIV-1 reverse transcriptase inhibition. *Physical chemistry chemical physics : PCCP, 14*(36), 12614-12624. doi: 10.1039/c2cp40953d

Tan, Q., Zhu, Y., Li, J., Chen, Z., Han, G. W., Kufareva, I., . . . Wu, B. (2013). Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex. *Science, 341*(6152), 1387-1390. doi: 10.1126/science.1241475

Tenorio-Borroto, E., Garcia-Mera, X., Penuelas-Rivas, C. G., Vasquez-Chagoyan, J. C., Prado-Prado, F. J., Castanedo, N., & Gonzalez-Diaz, H. (2013). Entropy model for multiplex drug-target interaction endpoints of drug immunotoxicity. *Current topics in medicinal chemistry, 13*, 1636-1649.

Todeschini, R., & Consonni, V. (2000). *Handbook of Molecular Descriptors*. Weinheim, Germany: Wiley-VCH Verlag GmbH.

Todeschini, R., Consonni, V., Mauri, A., & Pavan, M. (2005). DRAGON [computer program]. Milano, Italy: Talete srl.

Ul-Haq, Z., Usmani, S., Shamshad, H., Mahmood, U., & Halim, S. A. (2013). A combined 3D-QSAR and docking studies for the In-silico prediction of HIV-protease inhibitors. *Chemistry Central journal, 7*(1), 88. doi: 10.1186/1752-153x-7-88

Usach, I., Melis, V., & Peris, J. E. (2013). Non-nucleoside reverse transcriptase inhibitors: a review on pharmacokinetics, pharmacodynamics, safety and tolerability. *Journal of the International AIDS Society, 16*(1), 1-14. doi: 10.7448/ias.16.1.18567

Van Waterbeemd, H. (1995). Discriminant Analysis for Activity Prediction. In H. Van Waterbeemd (Ed.), *Chemometric methods in molecular design* (Vol. 2, pp. 265-282). New York, NY: VCH.

Vilar, S., Santana, L., & Uriarte, E. (2006). Probabilistic neural network model for the in silico evaluation of anti-HIV activity and mechanism of action. *Journal of medicinal chemistry, 49*(3), 1118-1124. doi: 10.1021/jm050932j

Vina, D., Uriarte, E., Orallo, F., & Gonzalez-Diaz, H. (2009). Alignment-free prediction of a drug-target complex network based on parameters of drug connectivity and protein sequence of receptors. *Molecular pharmaceutics, 6*(3), 825-835. doi: 10.1021/mp800102c

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440-442.

Wensing, A. M., Calvez, V., Gunthard, H. F., Johnson, V. A., Paredes, R., Pillay, D., . . . Richman, D. D. (2014). 2014 Update of the drug resistance mutations in HIV-1. *Topics in antiviral medicine, 22*(3), 642-650.

Wielens, J., Headey, S. J., Rhodes, D. I., Mulder, R. J., Dolezal, O., Deadman, J. J., . . . Scanlon, M. J. (2013). Parallel screening of low molecular weight fragment libraries: do differences in methodology affect hit identification? *Journal of biomolecular screening, 18*(2), 147-159. doi: 10.1177/1087057112465979

Wilkin, T. J., & Gulick, R. M. (2012). CCR5 antagonism in HIV infection: current concepts and future opportunities. *Annual review of medicine, 63*, 81-93. doi: 10.1146/annurev-med-052010-145454

Wu, B., Chien, E. Y., Mol, C. D., Fenalti, G., Liu, W., Katritch, V., . . . Stevens, R. C. (2010a). Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science (New York, N.Y.), 330,* 1066-1071. doi: 10.1126/science.1194396

Wu, B., Chien, E. Y., Mol, C. D., Fenalti, G., Liu, W., Katritch, V., . . . Stevens, R. C. (2010b). Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science, 330*(6007), 1066-1071. doi: 10.1126/science.1194396

Wu, H. Q., Yao, J., He, Q. Q., & Chen, F. E. (2014). Docking-based CoMFA and CoMSIA studies on naphthyl-substituted diarylpyrimidines as NNRTIs. *SAR and QSAR in Environmental Research, 25*(10), 761-775. doi: 10.1080/1062936x.2014.955054

Zarrabi, N., Prosperi, M., Belleman, R. G., Colafigli, M., De Luca, A., & Sloot, P. M. (2012). Combining epidemiological and genetic networks signifies the importance of early treatment in HIV-1 transmission. *PloS one, 7*(9), e46156. doi: 10.1371/journal.pone.0046156

Zhang, H., Wang, Y. F., Shen, C. H., Agniswamy, J., Rao, K. V., Xu, C. X., . . . Weber, I. T. (2013). Novel P2 tris-tetrahydrofuran group in antiviral compound 1 (GRL-0519) fills the S2 binding pocket of selected mutants of HIV-1 protease. *Journal of medicinal chemistry, 56*, 1074-1083. doi: 10.1021/jm301519z

7. PUBLICACIONES

7.1 Publicación 1

González-Díaz H, **Herrera-Ibatá DM**, Duardo-Sánchez A, Munteanu CR, Orbegozo-Medina RA, Pazos A. ANN Multiscale Model of Anti-HIV Drugs Activity vs AIDS Prevalence in the US at County Level Based on Information Indices of Molecular Graphs and Social Networks. *Journal of Chemical Information and Modeling* 54 (3) 2014, 744-755.

Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

ANN Multiscale Model of Anti-HIV Drugs Activity vs AIDS Prevalence in the US at County Level Based on Information Indices of Molecular Graphs and Social Networks

Humberto González-Díaz,*[†,‡] Diana María Herrera-Ibatá,[§] Aliuska Duardo-Sánchez,[§] Cristian R. Munteanu,[§] Ricardo Alfredo Orbegozo-Medina,[‖] and Alejandro Pazos[§]

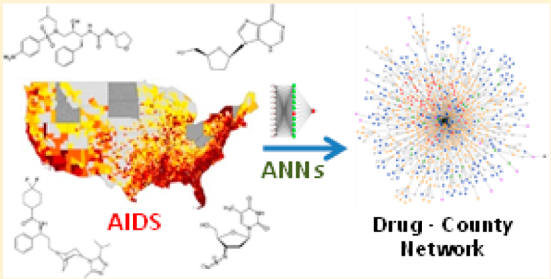
[†]Department of Organic Chemistry II, Faculty of Science and Technology, University of the Basque Country UPV/EHU, 48940, Leioa, Vizcaya, Spain

[‡]IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Vizcaya, Spain

[§]Department of Information and Communication Technologies, University of A Coruña UDC, 15071, A Coruña, A Coruña, Spain

[‖]Department of Microbiology and Parasitology, University of Santiago de Compostela (USC), 15782, Santiago de Compostela, A Coruña, Spain

S *Supporting Information*



ABSTRACT: This work is aimed at describing the workflow for a methodology that combines chemoinformatics and pharmacoepidemiology methods and at reporting the first predictive model developed with this methodology. The new model is able to predict complex networks of AIDS prevalence in the US counties, taking into consideration the social determinants and activity/structure of anti-HIV drugs in preclinical assays. We trained different Artificial Neural Networks (ANNs) using as input information indices of social networks and molecular graphs. We used a Shannon information index based on the Gini coefficient to quantify the effect of income inequality in the social network. We obtained the data on AIDS prevalence and the Gini coefficient from the AIDSVu database of Emory University. We also used the Balaban information indices to quantify changes in the chemical structure of anti-HIV drugs. We obtained the data on anti-HIV drug activity and structure (SMILE codes) from the ChEMBL database. Last, we used Box-Jenkins moving average operators to quantify information about the deviations of drugs with respect to data subsets of reference (targets, organisms, experimental parameters, protocols). The best model found was a Linear Neural Network (LNN) with values of Accuracy, Specificity, and Sensitivity above 0.76 and AUROC > 0.80 in training and external validation series. This model generates a complex network of AIDS prevalence in the US at county level with respect to the preclinical activity of anti-HIV drugs in preclinical assays. To train/validate the model and predict the complex network we needed to analyze 43,249 data points including values of AIDS prevalence in 2,310 counties in the US vs ChEMBL results for 21,582 unique drugs, 9 viral or human protein targets, 4,856 protocols, and 10 possible experimental measures.

1. INTRODUCTION

The acquired immunodeficiency syndrome (AIDS)[1] caused by the human immunodeficiency virus (HIV) is still considered as one of the most life-threatening diseases, and the HIV[2,3] pandemic continues to spread. Since the beginning of the epidemic, more than 60 million people have been infected with HIV, and over 25 million have died from the disease. Since the first case of AIDS was reported by the US in 1981, tremendous progress has been made in the prevention and treatment of HIV/AIDS,[4] especially in the development of antiretroviral therapy[5] that has proven to be life-saving to millions of people. Therefore, the discovery and development of novel, highly potent anti-HIV drugs remain imperative, although the eradication is still a difficult goal to achieve due to a low level of viral persistence in treated subjects.[6]

In this context, different Computer-Aided Drug Design (CADD) techniques, useful to predict the behavior of anti-HIV drugs, may play an important role in reducing the number of preclinical and clinical studies. For instance, we could use chemoinformatics models that link the chemical structure of drugs with their biological activity. In fact, there are many reports of chemoinformatics models, useful to predict anti-HIV activity in preclinical assays.[7] In principle, we could upgrade these models to predict the anti-HIV activity of drugs not only in preclinical screening but also in clinical and pharmacoepidemiology studies. Such a model may become a very useful tool not only for the Pharmaceutical Industry in order to reduce clinical assays. They should ideally be useful also for Public entities responsible for implementation of Health policies in the phase IV of drug development. However, there are no reports of models

Received: December 4, 2013
Published: February 12, 2014

useful to predict the performance of anti-HIV drugs in both preclinical and pharmacoepidemiology studies on large populations without carrying out clinical studies. We neither had at our disposal models able to extrapolate, at least, the performance of anti-HIV drugs from preclinical studies to epidemiology studies on large populations without carrying out clinical studies.

A useful chemoinformatics-pharmacoepidemiology model should be multilevel by definition as it is expected to account for both molecular and population structure. It means that, in order to develop such computational models, we need to process different types of input data coming from many different levels of organization of matter. On the one hand, we need to introduce information about the anti-HIV drugs including at least the chemical structure of the drug (level i) and the preclinical assay information, such as biological targets (level ii), organisms (level iii), or assay protocols (level iv). On the other hand, we need to incorporate population structure descriptors (level v) that quantify the epidemiological and social and economic factors affecting the population selected for the study. Last, as populations in modern society are not close systems we should also quantify the effect of interaction of the population under study with other populations that may influence the pharmacoepidemiology study (level vi). The data for levels i--iv were obtained from public databases of biological activity of organic compounds. These databases accumulated immense data sets of experimental results of pharmacological trials for many compounds. For instance, ChEMBL (https://www.ebi.ac.uk/chembl/)[8,9] is one of the biggest with more than 11,420,000 activity data for >1,295,500 compounds and 9,844 targets. Specifically, ChEMBL contains >43,000 outcomes for assays of anti-HIV compounds.

In addition, we obtained the data for levels v and vi from public epidemiological databases. For instance, AIDSVu[10] (http://aidsvu.org/about-aidsvu/) is the most detailed publicly available view of HIV prevalence in the US. AIDSVu is a compilation of interactive online maps that displays the HIV prevalence data at the national, state, and local levels and by different demographics, including age, race, and sex. Researchers at the Rollins School of Public Health at Emory University compiled the county-level data displayed on AIDSVu from the CDC (U.S. Centers for Disease Control and Prevention). State, county, and city health departments, depending on the entity responsible for HIV surveillance provided data on the HIV prevalence at the ZIP code and census tract. An Advisory Committee and a Technical Advisory Group guide the project with representatives from federal agencies, state health departments, and nongovernmental organizations working in HIV prevention, care, and research.

The formulation of mathematical models of this large data set from ChEMBL is very complex per se[9,11] but becomes an even more complicated problem when AIDSVu data are added. This is not only a problem of analysis of a huge number of data points (Big Data),[12−17] it is also a problem of dealing with the mathematical representation/codification of such diverse information from many different levels of organization of matter and areas of scientific knowledge. We can talk about three features of the problem resulting from the combination of chemical, pharmacological, and epidemiological information: (1) multitargeting, (2) multiobjective, and/or (3) multiscaling features. The multitargeting nature of the problem[18−20] refers to the existence of multitarget compounds that can interact with more than one molecular or cellular target. The multiobjective optimization problem (MOOP)[21−25] refers to the necessity of prediction/optimization of results for different experimental measures obtained in different pharmacological assays. Last, multiscaling refers to the different structural levels of the organization (i−vi) of matter that 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 q^{th} drug (d_q), due to variations in the chemical structure, detected in preclinical assays carried out under a set of j^{th} conditions (c_j).

We can use numerical descriptors of the molecular graph of the drug. In particular, some of these parameters are useful to quantify information about the properties of molecular, biological, and/or social systems (information measures). For instance, Shannon's entropy measures are universal parameters used to codify biologically relevant information in many systems. In the 1970s Bonchev and Trinajstic et al. published some works about the use of Shannon's entropy to calculate a structural information parameter.[26−29] Kier published other seminar works on the use of Shannon's entropy to encode molecular structure in chemoinformatics studies in 1980.[29] In this context, a drug molecule is considered an information source. Many other authors used Shannon's entropy parameters to encode small molecule structure.[30−35] Graham et al.[36−40] used entropy measures to study in depth the information properties of organic molecules. These concepts were extended to describe protein,[41,42] DNA sequences,[43] or protein−protein interaction networks.[44] Mikoláš et al.[45] reviewed several studies about the use of entropy measures in functional magnetic resonance. In a recent work we have used Shannon entropy measures and the idea of Moving Average (MA) operators in a time series analysis with a similar purpose.[46] Additionally, information indices are graph-theoretical invariants that view the molecular graph as a source of different probability distributions to which information theory definitions can be applied. They can be considered a quantitative measure of the lack of structural homogeneity or the diversity of a graph, in this way being related to the symmetry associated with structure.[47−49] Ivanciuc and Balaban[50] defined the indices for simple and weighted molecular graphs and tested the information theory-indices for modeling alkane densities. Moreover, Ivanciuc et al.[51] also found that the information indices were extended for any symmetric molecular matrix derived from vertex-and edge-weighted molecular graphs. Dehmer et al.[52−55] mentioned the Balaban information indices[56] in their work about novel topological descriptors for biological networks.

However, the codification of the molecular structure of the drug is only the first step here. We have information about a high number of assays carried out in very different conditions (c_j) for the same or different targets, which may be molecular or not. The nonstructural information herein refers to different assay conditions (c_j) like time, concentrations, temperature, cellular targets, tissues, organisms, etc. A possible solution may rely upon the use of the idea of MA operators used in a time series analysis with a similar purpose.[46] MA models became popular after the initial works conducted by Box and Jenkins.[57] In a time series analysis, MA models may combine other operators I = Integrated, AR = Autoregressive, N = Nonlinear operators, or X = Exogenous effects. In this sense, we can develop models like ARMA, ARIMA, VARIMA, ARIMAX, NARMA, etc., combining different operators. The MA operators used in time series are the average value of a characteristic of the system for different intervals of time or seasons. In multiobjective modeling, we calculate the MA operators as the average of the property of the system (molecular descriptors or others) for all drugs or targets with a specific response in an assay carried out at under a subset of conditions (c_j). Consequently, our MA operator does not act over a time domain but over a subset of conditions of the pharmacological assays. The idea of application of

MA operators to other domains different from time is gaining adepts due to its advantages. For instance, Botella-Rocamora et al.[58] developed a model map of diseases called SMARS: Spatial Moving Average Risk Smoothing. They applied the MA of time series theory to the spatial domain, making use of a spatial MA to define dependence on the risk of a disease occurring.

Certainly, we can see this entire problem as the prediction of a complex network represented by the Boolean matrix **L** with elements L_{aq}. That is, we have to seek a model able to assess the formation ($L_{aq} = 1$) or not ($L_{aq} = 0$) of links between nodes in a complex network of AIDS pharmacoepidemiology in the US. Two different classes of nodes make up this network, the first representing the US counties (a) and the other class of nodes representing drugs (d_q). In the present context, we can use MA of properties of network nodes (drugs, proteins, organisms, counties, etc.) that form links (L_{aq}) in a specific subset of conditions (c_j). For this reason, we decided to call this strategy ALMA (Assessing of Links with Moving Averages) models. Speck-Planche and Cordeiro[59−61] have reported different multitarget models using the same type of ALMA approach.

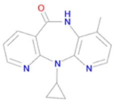
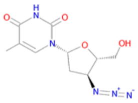
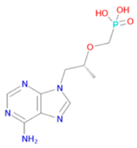
Last, we can use these information descriptors and MA operators as inputs for a Machine Learning (ML) algorithm. This ML has to seek the coefficients of the ALMA model able predict the correct links in **L**. The neural network approximates the operation of the human brain,[62,63] and this initially ″trained″ or fed large amounts of data and rules about data relationships. ANNs are in general nonlinear algorithms with a high number of processors (called neurons) which, in a classic picture, are distributed in layers and act in parallel (neurons in the same layer) or in series (pairs of neurons connected in different layers). In recent years, ANNs[64,65] have turned out to be a powerful method for various practical applications in a great variety of disciplines, and they can be used to find complex relationships between inputs and outputs or to find models in data. Another aspect of ANNs[66,67] is that there are different architectures, which require different types of algorithms for training; the trained ANN do not need to be reprogrammed.

2. MATERIALS AND METHODS

Linear and Nonlinear ALMA Models. The ALMA models are useful to assess the formation of links in different complex networks that are representations of complex systems. They are adaptable to all types of molecular descriptors and/or graphs invariants or descriptors for complex networks. In this work, we tried to seek a classification model. The overall output of this model is $L_{aq}(c_j)_{pred}$. This variable is a prediction of the observed variable $L_{aq}(c_j)_{obs}$. Both the observed and predicted variables are discrete Boolean variables (1, 0). The observed variable takes the value $L_{aq}(c_j)_{obs} = 1$ if the observed score $S_{aq}(c_j)_{obs} >$ input cutoff or $L_{aq}(c_j)_{obs} = 0$ otherwise. In analogy, the predicted variable $L_{aq}(c_j)_{pred} = 1$ if the predicted score $S_{aq}(c_j)_{pred} >$ output cutoff or $L_{aq}(c_j)_{pred} = 0$ otherwise.

More specifically, we can say that the value is $L_{aq}(c_j)_{obs} = 1$ when the $S_{aq}(c_j)_{obs}$ = Drug-Disease Ratio = $DDR_{aq}(c_j) >$ cutoff and $L_{aq}(c_j)_{obs} = 0$ otherwise. We defined the ratio as follows: $S_{aq}(c_j)_{obs} = DDR_{aq}(c_j) = [D_q(c_j)/D_a]$. We calculated the numerator term as $D_q(c_j) = \delta_j \cdot z_q(c_j) = \delta_j \cdot [v_q(c_j) - AVG(v(c_j))]/SD(v(c_j))$. In this operator, $v_q(c_j)$ is the value of biological activity (EC_{50}, IC_{50}, K_i, ..., etc.) reported in the ChEMBL database for the q^{th} drug assayed under the set of conditions $c_j = (c_1, c_2, c_3, c_4)$. The parameter δ_j is similar to a Kronecker delta function. The parameter $\delta_j = 1$ when the biological activity parameter $v_q(c_j)$ is directly proportional to the biological effect (e.g., K_i values, Activity (%) values, etc.). Conversely, $\delta_j = -1$ when the biological activity parameter $v_q(c_j)$ is in inverse proportion to the biological effect (e.g., EC_{50} values, IC_{50} values, etc.). The parameter $z_q(c_j)$ is the z-score of the biological activity that depends on the functions AVG and SD. These functions are the average and standard deviation of $v_q(c_j)$

Table 1. Examples about How To Calculate the Moving Average Operators for Some Compounds

Cmpd ID	Name	SMILE	Target	I^q_1	$<I^q_1>$	$\Delta I^q_1(c_2)$
57	Nevirapine 	Cc1ccnc2N(C3CC3)c4ncccc4C(=O)Nc12	HIV-1	25.731	31.007	-5.2763
129	Zidovudine 	CC1=CN([C@H]2C[C@H](N=[N+]=[N-])[C@@H](CO)O2)C(=O)NC1=O	HIV-1	33.488		2.48066
483	Tenofovir 	C[C@H](Cn1cnc2c(N)ncnc12)OCP(=O)(O)O	HIV-1	33.803		2.79566
964	Disulfiram	CCN(CC)C(=S)SSC(=S)N(CC)CC	CC-CKR-5	60.801	44.863	15.9376
39879	Dipyridyl	c1ccc(nc1)c2ccccn2	CC-CKR-5	17.119		-27.7443
1201187	Maraviroc	CC(C)c1nnc(C)n1[C@@H]2C[C@H]3CC[C@@H](C2)N3CC[C@H](NC(=O)C4CCC(F)(F)CC4)c5ccccc5	CC-CKR-5	56.67		11.8066
31574	Fisetin	OC1=C(Oc2cc(O)ccc2C1=O)c3ccc(O)c(O)c3	HIV-1 IN	32.01	33.507	-1.4976
28626	Morin	OC1=C(Oc2cc(O)cc(O)c2C1=O)c3ccc(O)cc3O	HIV-1 IN	34.236		0.72833
50	Quercetin	OC1=C(Oc2cc(O)cc(O)c2C1=O)c3ccc(O)c(O)c3	HIV-1 IN	34.277		0.76933

for all drugs assayed under the same conditions. In this sense, $c_1 =$ is the experimental measure of activity, $c_2 =$ is the protein target, $c_3 =$ is the organism that expresses the target, and $c_4 =$ is the assay protocol per se. In the denominator, we used the term D_a that is the AIDS prevalence rate for the a^{th} county. We can conclude that $L_{\text{aq}}(c_j)_{\text{obs}}$ and consequently $L_{\text{aq}}(c_j)_{\text{pred}}$ depend on both the prevalence of the disease and the effectiveness of the drug due to the definition of $\text{DDR}_{\text{aq}}(c_j)$. In Table 1, a simple example of calculation of MA operators is shown. In this example, we only use the condition (c_2), i.e., Balaban information index $\mathbf{U}= I^q{}_1$ and the target of the drug, to illustrate the method. First, we have the SMILE codes of the compounds obtained from ChEMBL. Next, using the DRAGON Software[68] we calculated the Balaban Information Indices (in this case only $\mathbf{U} = I^q{}_1$). Afterward, we calculated $\langle I^q{}_1\rangle$ the average of the information index $I^q{}_1$ for the compounds with the same targets. Last, we calculated the MA operators with the formula $\Delta I^q{}_1(c_2) = (I^q{}_1 - \langle I^q{}_1\rangle_{c2}$. In our work, this method was applied to the 43,249 molecules characterized by different Balaban Information indices $(\mathbf{U} = I^q{}_1, \mathbf{V} = I^q{}_2, \mathbf{X} = I^q{}_3, \mathbf{Y} = I^q{}_4)$ and assay conditions $c_j = (c_1, c_2, c_3, c_4)$. In addition, $\langle D_q(c_j)\rangle$ is the average value of the biological activity for all the drugs assayed under the same conditions. Consequently, $\Delta D_q(c_j)$ is an MA operator that accounts for the deviation of the biological activity of the drug $D_q(c_j)$ in a preclinical assay with respect to the average value $\langle D_q(c_j)\rangle$ of this activity for all drugs assayed under the same conditions c_j.

In order to seek a model able to predict $L_{\text{aq}}(c_j)_{\text{pred}}$, we used as input different information descriptors for drugs and populations. In general, we refer to an information index $I^q{}_k$ of type k^{th} for the system (drug or county in this case) represented by a matrix **L**. The aim of this model is to predict scores $S_{\text{aq}}(c_j)$ of the formation of links L_{aq} using as input the structural information quantified by the indices $I^a{}_0(t)$ for the population (county) and $I^q{}_k$ of a given compound d_q. The simplest model may be based on the additive hypothesis H_0. The hypothesis H_0 states that $S_{\text{aq}}(c_j) = {}^{q}S_k + {}^{qj}S_k + {}^{as}S_k + e_0$. It means that it can be calculated as a summation of different scores or measures of factors plus a model error e_0. We have three types of scores or factors divided into two subtypes. The first subtype includes the scores for drugs and the second subtype the scores for counties. The first scores ${}^{q}S_k \approx e_k\cdot p(c_1)\cdot I^q{}_k$ account for information on both the contributions of the k^{th} molecular descriptor and for the quality of raw data $p(c_1)$ to the final activity score $S_{\text{aq}}(c_j)$. In fact, we used the probability $p(c_1) = 1.0$; 0.75; or 0.5 for data curated in CHEMBL database at expert, intermediate, or autocuration levels, respectively. The second scores ${}^{qj}S_k \approx e_{kj}\cdot\Delta I^q{}_k(c_j)$ account for the contributions of deviations $\Delta I^q{}_k(c_j) = (I^q{}_k - \langle I^q{}_k\rangle_j)$ in the structure of the drug from the average of all those molecules assayed under the conditions c_j. In order to test this hypothesis we used the information indices and their MA operators $\Delta I^q{}_k(c_j) = I^q{}_k - \langle I^q{}_k(c_j)\rangle$ to express the different assay conditions for the drugs. We also used a simple information index $I^a{}_0(t)$ for income inequality in the different counties. The linear model ALMA has the following general form:

$$\begin{aligned} S_{\text{aq}}(c_j) &= {}^{q}S_k + {}^{qj}S_k + {}^{as}S_k + e_0 \\ &= \sum_{k=1}^{k=4} e_k\cdot I^q{}_k + \sum_{k=1}^{k=4}\sum_{j=1}^{j=4} e_{kj}\cdot\Delta I^q{}_k(c_j) + e_{ak}\cdot I^a{}_k(t) + e_0 \\ &= \sum_{k=1}^{k=4} e_k\cdot I^q{}_k + \sum_{k=1}^{k=4}\sum_{j=1}^{j=4} e_{kj}\cdot(I^q{}_k - \langle I^q{}_k\rangle_j) + e_{ak}\cdot I^{\text{a}}{}_k(t) + e_0 \end{aligned} \tag{1}$$

The reader should note that the predicted, output, or dependent variable $S_{\text{aq}}(c_j)$ is not a discrete variable but a real-valued numerical score. However, the variable $S_{\text{aq}}(c_j)$ is directly proportional to the observed variable (L_{aq}). Please, note that all the parameters $S_{\text{aq}}(c_j)$ => $L_{\text{aq}}(c_j)$ => $\text{DDR}_{\text{aq}}(c_j)$ => $D_q(c_j)$ form a series that in the last instance depends on (=>) the conditions of the initial preclinical assay used to measure the activity of the drug $c_j = (c_1, c_2, c_3, c_4)$. In general, c_j refers to different boundary conditions for the assay, e.g., targets, assays, cellular lines, organisms, organs, etc. In this sense, $c_1 =$ is the experimental measure of activity, $c_2 =$ is the protein target, $c_3 =$ is the organism that expresses the target, and $c_4 =$ is the assay protocol per se. Some inputs of the models depend on parameters of the type of deviations $\Delta I^q{}_k(c_j)$, which are similar to the MA operators used in the time series analysis for ARIMA models and others.[57] This means that, first, we add up for instance the values of $I^q{}_k$ for all the n_j drugs under the assay conditions c_j. Next, we divide this sum by the number of compounds n_j under this condition.

$$\langle I^q{}_k\rangle_j = \frac{1}{n_j}\sum_{i=1}^{i=n_j} I^q{}_k(c_j) \tag{2}$$

In order to seek the coefficients of the model, we can use a linear classification technique like ANN implemented in the STASTITICA 6.0 software package.[69] The statistical parameters used to corroborate the model were as follows: Number of cases in training (N), and overall values of Specificity (Sp), Sensitivity (Sn), and Accuracy (Ac).[70]

CHEMBL Data Set of Drugs. We downloaded from the public database CHEMBL a general data set composed of >8,000 multiplexing assay end points (results of multiple assays).[8,9] The data set used to perform the model included N = 43,249 statistical cases made up of N_{d} = 21,582 unique drugs. These drugs have been assayed one by one in at least one out of 10 possible standard type measures determined in at least one out of 4,856 different assays (experimental protocols reported as different in ChEMBL). Each assay involved, in turn, at least one out of 9 nonmolecular or protein targets expressed in tissues, cells, or viral particles of at least one out of 5 different organisms (including human cells lines).

Balaban Information Indices of Molecular Graphs of Drugs. The Balaban information indices[56] U, V, X, and Y are very useful to quantify information about the chemical structure of drugs.[71] These indices use some the following parameters: σ_x = vertex distance degree of x^{th} atom (i.e., sum of topological distances from the considered atom to any other atom), d_{xy} is the topological distance between atoms x^{th} and y^{th} atoms; n is the number of non-H atoms. Other parameters used are ${}^{g}f_x$ = the number of distances from the x^{th} vertex equal to g and η_x = the eccentricity of the x^{th} atom (i.e., the maximum topological distance from the considered atom). We denoted these indices in the present work as $I^q{}_k$. In this notation, the letter I stands for the information index, q indicates the number of order (label) of the drug in the data set, and k indicates the type of index. The mathematical formulas for calculation of these indices are

$$I^q{}_1 = U_q = -\sum_{j=1}^{n} \frac{d_{xy}}{\sigma_x}\cdot\log_2\frac{d_{xy}}{\sigma_x} \tag{3}$$

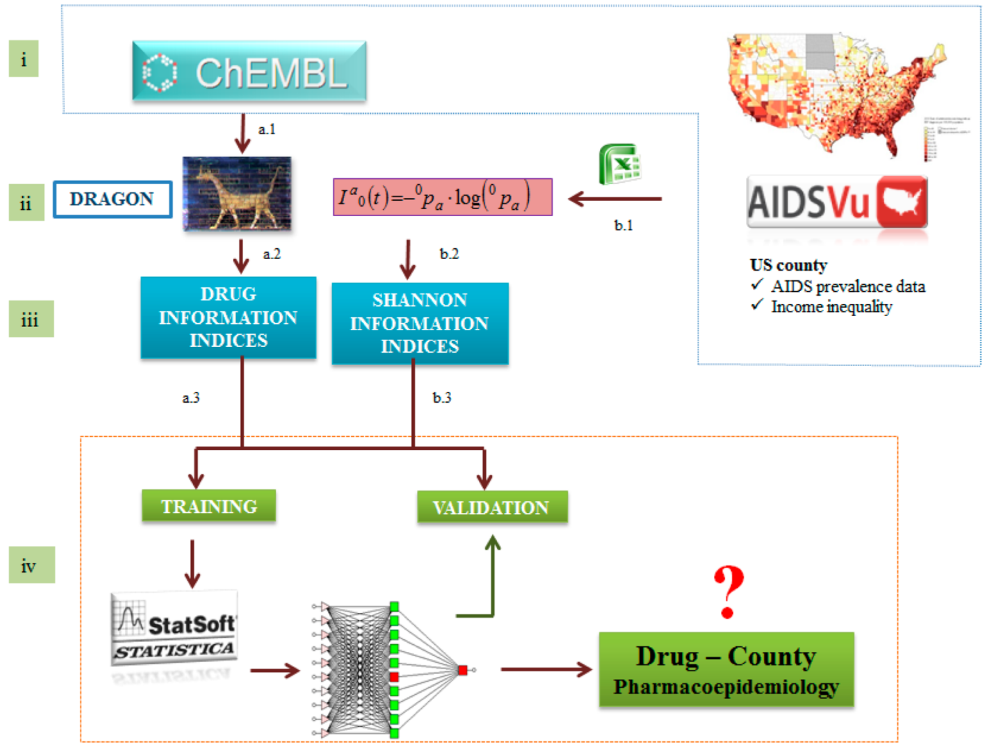


Figure 1. Flowchart of all steps given to construct the ANNs for the Drug-County Pharmacoepidemiology model in the United States.

$$I^q{}_2 = V_q = \sigma_x\cdot\log_2(\sigma_x) - u_q = \sigma_x\cdot\log_2(\sigma_x) + \sum_{j=1}^{n}\frac{d_{xy}}{\sigma_x}\cdot\log_2\frac{d_{xy}}{\sigma_x} \qquad (4)$$

$$I^q{}_3 = Y_q = \sum_{g=1}^{\eta_i} {}^g f_q\cdot g\cdot\log_2(g) \qquad (5)$$

$$I^q{}_4 = X_q = \sigma_q\cdot\log_2(\sigma_q) - \sum_{g=1}^{\eta_i} {}^g f_q\cdot g\cdot\log_2(g) \qquad (6)$$

AIDSvu Data Set of AIDS Prevalence in the US at County Level. Data were drawn from the AIDSVu database of the Rollins School of Public Health at Emory University (www.aidsvu.org). We downloaded the values of epidemiological variables for AIDS in the US at county level from the public database. The values used in this study included the percentage of adults/adolescents living with an HIV diagnosis in 2010 per 100,000 populations. The county-level HIV surveillance data displayed on AIDSVu are estimated data for persons aged 13 and older living with an HIV infection diagnosis. All race groups are non-Hispanic, and the Hispanic/Latino ethnicity is inclusive of all races. Sex is defined as "sex at birth". Data are not displayed at the county level for Asians, Native Hawaiians/Other Pacific Islanders, and American Indians/Alaska Natives because these data do not meet CDC's criteria for statistical reliability, data quality, or confidentiality due to small population denominators and HIV case counts. The total number of counties is $n_a = 2{,}310$.

Shannon Information Indices of Income Inequality. We can calculate an information index to quantify the possibility of AIDS spreading/prevalence in different counties (a) of the US. Let be an initial situation in which each county has a value of AIDS prevalence rate D_a at the initial time ($t_0 = 2010$). We used here a simple information index $I^a{}_0(t)$ for income inequality in the different counties that year. This index depends on the probability ${}^0p_a$ with which the county presents certain income inequality. We set here this probability ${}^0p_a = G_a$. In this definition, G_a is the Gini measure of income inequality in the a^{th} county of (a) given state(s) in the US.[72] The class of information indices selected by us was the Shannon entropy indices.[73]

$$I^a{}_0(t) = -{}^0p_a\cdot\log({}^0p_a) \qquad (7)$$

3. RESULTS AND DISCUSSION

Definition of the Algorithm. In this work, we report for the first time a model based on information indices of chemical structure, biological assay, and county level income inequality. The model is able to link the deviations in the AIDS prevalence in the a^{th} county with the changes in the biological activity of the q^{th} drug (d_q). In so doing, the model considers the biological activity of anti-HIV compounds detected in preclinical assays carried out under a set of j^{th} conditions (c_j). Using this type of model, we can predict the pharmacoepidemiology complex network for AIDS in the United States at county level.

First, we propose a new algorithm to construct this type of models. The algorithm/model used as input both drug structures and preclinical information as well as county income inequality data. We understand here as algorithm the series of all steps given in different stages in order to seek and use the model. We illustrate the different steps of this algorithm in Figure 1. The stages of the algorithm proposed are the following: (i) data compilation, (ii) data preprocessing, (iii) calculation of inputs, (iv) development, and use of the model. These stages are similar and divided into two parallel branches (A and B). Both branches have different steps, one for the chemical and biological information of drugs and the other for the information about county pharmacoepidemiology. Next, after the preprocessing stage (ii), the two branches are joined into a single branch (C) that enters a cycle of training vs validation of the different ANN models and ends with the selection and use of the best model found. In this context, we understand as model the ANN trained and validated in the final step of the algorithm. The most important steps for the branches A and B are the following (the software/databased used are between round brackets):

a.1. Gathering of the chemical structure and biological activity information from public sources (ChEMBL).

a.2. Processing of the information about molecular structure (SMILE codes) and biological activity (EXCEL).

a.3. Calculation of I^q_k values and MA operators for the molecules (DRAGON,[68] EXCEL).

b.1. Downloading the US AIDS prevalence and income inequality data (AIDSVu).

b.2. Calculation of the simple information index $I^a_0(t)$ for income inequality in the different counties.

c.5. Training and validation of ANN predictive models (STATISTICA).[69]

Model Training and Validation. In the first step, we calculated the values drug-disease ratio $\mathrm{DDR}_{\mathrm{aq}}(c_j)_{\mathrm{obs}}$ for the 43,249 drug-county pairs. After that, we carried out a cutoff scanning and found that we can split the data set into 11,089 cases with $L_{\mathrm{aq}}(c_j)_{\mathrm{obs}} = 1$ and 32,160 cases with $L_{\mathrm{aq}}(c_j)_{\mathrm{obs}} = 0$ using a cutoff = 500. This is 25.6% of the positive cases that ensure a ratio of above 1/4 of positive vs control cases. The data set used to train the model includes N = 32,437 statistical cases. The data set used to validate the model includes N = 10,812 statistical cases. The cases used in the validation set (external validation set) were never used to train the model. Overall, training + validation sets include N = 43,249 statistical cases. Next, we calculated the values of the Balaban information indices I^q_k for all the drugs/organic compounds present in our ChEMBL subset (step a.2). Table 2 shows some examples of these values I^q_k for known drugs. In addition, in Table SM1 of the Supporting Information content we list the values of I^q_k for all the drugs studied. We can see that the information indices I^q_k have different numerical values for different molecular structures of drugs. After that, we calculated the average values of these indices $\langle I^q_k\rangle$ for the different boundary conditions (c_j). In Table 3, you can see some examples of these average values for different boundary conditions like targets, organisms, etc. After a visual inspection, one can note that the $\langle I^q_1\rangle$ values seem to distinguish more clearly between the different boundary conditions. For instance, they have differences in the range of 10−100 units for 9 different protein targets (4 HIV vs five human proteins) present in the data set. However, the other averages $<I^q_k>$ with $k > 1$ seem to be worse at differentiating the proteins. In Table SM2 of the Supporting Information, we list the values of $<I^q_k>$ for all the organisms, assay protocols, protein targets, and experimental measures studied.

Next, we calculated the values of the information indices $I^a_0(t)$ for different US counties. Consequently, we used only the $I^a_0(t)$ as inputs for the model. After that, we obtained the ANN models using as input 19 descriptors: 4 Balaban information indices of the molecules (I^q_k), 14 MA operators ($\Delta I^q_k(c_j)$) for the different assay conditions for drugs (c_1,c_2,c_3,c_4), and the $I^a_0(t)$ of the US counties. In Table 4, we illustrate the values of $I^a_0(t)$ for some counties of different states. In Table SM3 of the Supporting Information, we list the values of $I^a_0(t)$ for the 2,310 US counties studied here.

The results obtained using the STATISTICA software show that the Multilayer Perceptron (MLP)[74] method fails to generate good prediction models, since it presents values of Specificity and Sensitivity close to 50% (values for a random classifier). On the other hand, the Linear Neural Network (LNN) predictor based on 19 descriptors (LNN 19:19-1:1) is able to classify correctly above 76% of the cases in training and validation (see Table 5). This model presented high values of Sensitivity = Sn = 76.46 and Specificity = Sp = 77.13 in training and Sn = 77.30 and Sp = 75.67 in the external validation sets.

Figure 2 shows the AUROC values for the different ANN models. The LNN network shows values of AUROC = 0.82 in the training and external validation set. These values are typical of a classifier with a classification behavior different from a random classifier (AUROC = 0.5).[70] The sensitivity analysis allowed us to quantify (rank) and order (ratio) into a sequence the importance

Table 2. Values of Balaban Information Indices for Some Anti-HIV Compounds

CMPD_ID	name	I^q_1	I^q_2	I^q_3	I^q_4
8	Ciprofloxacin	33.562	0.236	0.352	0.705
28	Apigenin	29.885	0.27	0.407	0.789
50	Quercetin	34.277	0.276	0.413	0.819
54	Haloperidol	44.833	0.185	0.303	0.477
57	Nevirapine	25.731	0.284	0.406	0.916
169	Ursolic Acid	47.503	0.217	0.323	0.663
58	Mitoxantrone	60.8	0.236	0.363	0.671
61	Podofilox	41.071	0.211	0.314	0.644
66	(+)-Taxifolin	34.277	0.276	0.413	0.819
76	Chloroquine	42.46	0.26	0.414	0.696
107	Colchicine	51.672	0.287	0.423	0.882
114	Saquinavir	86.9	0.151	0.237	0.415
115	Indinavir	76.144	0.147	0.233	0.403
116	Amprenavir	69.801	0.221	0.342	0.621
117	Chrysin	27.677	0.282	0.42	0.837
129	Zidovudine	33.488	0.331	0.497	0.973
141	Lamivudine	23.582	0.349	0.519	1.03
150	Kaempferol	32.004	0.278	0.416	0.828
151	Luteolin	32.088	0.267	0.404	0.78
160	Cyclosporine	582.739	0.44	0.689	1.214
163	Ritonavir	103.789	0.161	0.256	0.435
164	Myricetin	36.608	0.275	0.412	0.817
168	Oleanolic Acid	47.52	0.215	0.32	0.653
193	Nifedipine	50.628	0.377	0.547	1.204
413	Sirolimus	159.248	0.178	0.28	0.488
483	Tenofovir	33.803	0.293	0.455	0.81
484	Adefovir	31.274	0.282	0.443	0.764
593	Delavirdine	52.159	0.166	0.267	0.439
625	Thiabendazole	18.185	0.311	0.458	0.935
713	Entecavir	29.767	0.29	0.43	0.875
729	Lopinavir	90.532	0.173	0.271	0.477
853	Zalcitabine	23.582	0.349	0.519	1.03
885	Emtricitabine	25.889	0.35	0.52	1.041
964	Disulfiram	60.801	0.685	1.063	1.887
991	Stavudine	25.889	0.35	0.52	1.041
7187	Costatolide	39.569	0.262	0.381	0.826
1460	Didanosine	23.687	0.29	0.432	0.863
6246	Ellagic Acid	29.481	0.287	0.412	0.927
7187	Costatolide	39.569	0.262	0.381	0.826
8260	Baicalein	29.825	0.28	0.418	0.83
9352	Naringenin	29.885	0.27	0.407	0.789
12014	Harman	18.019	0.355	0.502	1.146
13134	Palinavir	95.629	0.142	0.226	0.386
16901	Honokiol	36.232	0.315	0.478	0.909

Table 3. Average Values of the Information Descriptors of Molecular Structure under Different Boundary Conditions

c_1	experimental measure	$N(c_j)$	$\langle I^q{}_1\rangle$	$\langle I^q{}_2\rangle$	$\langle I^q{}_3\rangle$	$\langle I^q{}_4\rangle$
IC_{50} (nM)	inhibitory concentration 50%	20332	64.303	0.209	0.324	0.587
EC_{50} (nM)	effective concentration 50%	14981	60.888	0.219	0.337	0.625
K_i (nM)	inhibitory constant	3736	78.878	0.180	0.282	0.501
IC_{95} (nM)	inhibitory concentration 95%	1290	59.295	0.189	0.296	0.521
IC_{90} (nM)	inhibitory concentration 90%	1118	54.730	0.226	0.338	0.682
ED_{50} (nM)	effective dose 50%	860	63.303	0.238	0.367	0.677
EC_{50} (μg·mL^{-1})	effective concentration	526	62.576	0.233	0.352	0.685
IC_{50} (μg·mL^{-1})	inhibitory concentration	335	147.952	0.254	0.406	0.687
EC_{90} (nM)	effective concentration	67	41.936	0.308	0.468	0.884
IC_{90} (μg·mL^{-1})	inhibitory concentration 90%	4	62.001	0.238	0.360	0.699
c_2	target protein	$N(c_j)$	$\langle I^q{}_1\rangle$	$\langle I^q{}_2\rangle$	$\langle I^q{}_3\rangle$	$\langle I^q{}_4\rangle$
CC-CKR-5	C−C chemokine receptor type 5	2304	62.466	0.152	0.243	0.405
CC-CKR-2	C−C chemokine receptor type 2	2009	64.050	0.170	0.273	0.448
CC-CKR-3	C−C chemokine receptor type 3	1206	56.723	0.156	0.253	0.410
CC-CKR-4	C−C chemokine receptor type 4	345	53.788	0.184	0.289	0.505
CXCR-4	C-X-C chemokine receptor type 4	332	147.452	0.178	0.278	0.497
HIV-1 RT	HIV-1 reverse transcriptase	4029	47.002	0.253	0.384	0.738
HIV-1 IN	HIV-1 integrase	1702	62.249	0.241	0.371	0.674
HIV-1 PR	HIV-1 protease	5946	89.711	0.184	0.288	0.513
GP160	envelope polyprotein GP160	34	45.879	0.224	0.353	0.611
c_3	organism	Ni(c_2)	$\langle I^q{}_1\rangle$	$\langle I^q{}_2\rangle$	$\langle I^q{}_3\rangle$	$\langle I^q{}_4\rangle$
HIV-1	HIV-1	34544	64.299	0.221	0.340	0.630
mmu	*Mus musculus*	68	64.004	0.157	0.251	0.423
hsa	*Homo sapiens*	6128	65.954	0.162	0.259	0.430
HIV-2	HIV-2	1030	81.747	0.198	0.311	0.547
HIV	HIV	1479	52.782	0.203	0.314	0.578
c_4	assay	$N(c_j)$	$\langle I^q{}_1\rangle$	$\langle I^q{}_2\rangle$	$\langle I^q{}_3\rangle$	$\langle I^q{}_4\rangle$
1033994	antiviral activity against HIV1	282	44.250	0.261	0.398	0.752
708445	effective concentration required for the inhibition of HIV-1 IIIB in MT-4 cells	176	102.090	0.158	0.251	0.424
859312	inhibitory activity was determined against HIV type 1 protease	175	112.916	0.164	0.258	0.450
659084	inhibitory conc for displacement of [125I]-MIP-1 alpha from human CCR5 in CHO cell	141	73.162	0.131	0.210	0.345
763303	inhibition of HIV-1 protease	118	72.588	0.177	0.269	0.515
974332	displacement of [125I]MIP1alpha from human CCR5 expressed in CHO cells	109	57.925	0.137	0.219	0.367
660813	inhibitory activity against recombinant human Chemokine receptor type 3 (CCR3) expressed in Chinese hamster ovary cells	108	57.228	0.154	0.248	0.406
833931	inhibitory activity against wild type HIV-1 LAI cell line	106	46.897	0.306	0.459	0.906

of the different chemoinformatics vs pharmacoepidemiology inputs. This kind of model may be useful to predict different situations of interest in pharmacoepidemiology. For instance, the model is able to identify when the same drugs present a strong effect on population epidemiology for different counties ($L_{aq}(c_j)_{pred} = 1$). Table 6 shows the predictions for some cases with the LNN model. In the table we can see that the model predicts $L_{aq}(c_j)_{pred}$ =1 for Nevirapine[75] in different counties, which is a drug $L_{aq}(c_j)_{obs} = 1$ for these counties. In Table SM4 of the Supporting Information, we provide the results predicted with the LNN model for all the cases in training and external validation series.

Last, we used this LNN-ALMA model to generate/predict a complex network of the AIDS prevalence in the US at county level with respect to the preclinical activity of anti-HIV drugs in preclinical assays. The network is bipartite with two classes of nodes (counties vs drugs). In this sense, it is a multiscale network similar to the bipartite networks of drugs vs target proteins reported by other groups.[76−80] However, the drug nodes of the present network contain information about the drug structure as well as all the assay conditions (target protein, organism, assay protocol, experimental measure). In addition, the other set of nodes is typical of a social network because they contain information about the income inequality in the county. Therefore, this complex network is multiscale, linking information about drugs, targets, assays, and society in the same line of thinking expressed by Barabasi et al.[81] The links of this complex network are the outputs $L_{aq}(c_j)_{pred} = 1$ of our model. That is why we analyzed 43,249 data points to fit the model and predict the complex network at the same time. Consequently, we have to include values of AIDS prevalence in 2,310 US counties vs ChEMBL results for 21,582 unique drugs, 9 viral or human protein targets, 4,856 protocols, and 10 possible experimental measures. In Figure 3, we illustrate the subnetwork of AIDS prevalence vs anti-HIV drug preclinical activity for the state of Texas. We include some examples of drugs like Efavirenz (ChEMBL223228) and Saquinavir (ChEMBL114) with observed and predicted $L_{aq}(c_j)_{obs} = L_{aq}(c_j)_{pred} = 1$ effects on AIDS prevalence in the counties of Kendall, Jasper, and Victoria, respectively.

We used the sensitivity analysis of the ANN module implemented in STATISTICA to detect the parameters with the higher contribution to the model. We can conduct the sensitivity analysis on the inputs to one ANN by a STATISTICA Neural Networks algorithm. The sensitivity analysis ranks the order of input importance by treating each input variable in turn as if it were "unavailable".[82] It is defined a missing value substitution procedure, which allows predictions to be made in

Table 4. Values of $I^a{}_0(t)$ for Some Counties of Different States

state(s)	county name	D_a[a]	G_a[b]	$I^a{}_0(t)$
AL	Autauga County	181	0.405	0.15898072
AL	Baldwin County	188	0.439	0.15695808
AR	Arkansas County	165	0.467	0.15442902
AR	Ashley County	97	0.447	0.15631254
AZ	Apache County	124	0.488	0.15205113
AZ	Cochise County	134	0.435	0.15725717
CA	Alameda County	396	0.456	0.15551203
CA	Amador County	114	0.399	0.15921181
CO	Adams County	179	0.403	0.15906207
CO	Alamosa County	78	0.474	0.15368107
CT	Fairfield County	375	0.537	0.14500381
CT	Hartford County	434	0.458	0.15532361
FL	Alachua County	383	0.516	0.14827275
FL	Baker County	380	0.429	0.15767582
GA	Appling County	105	0.422	0.15811815
GA	Atkinson County	256	0.447	0.15631254
HI	Hawaii County	199	0.458	0.15532361
HI	Honolulu County	201	0.422	0.15811815
IA	Boone County	58	0.407	0.15889508
Ia	Ada County	101	0.435	0.15725717
ID	Bannock County	100	0.429	0.15767582
IL	Adams County	65	0.453	0.15578751
IN	Adams County	21	0.380	0.15968223
IN	Allen County	136	0.428	0.15774207
KS	Allen County	44	0.394	0.15937449
KS	Atchison County	57	0.434	0.15732946
KY	Allen County	71	0.42	0.1582353
KY	Anderson County	76	0.376	0.15972937
KY	Barren County	56	0.455	0.15560481
LA	Acadia Parish	174	0.452	0.15587743
LA	Allen Parish	550	0.434	0.15732946
LA	Ascension Parish	178	0.409	0.15880517
MA	Berkshire County	102	0.462	0.15493541
MD	Allegany County	180	0.446	0.15639665
MD	Calvert County	124	0.369	0.15976727
ME	Hancock County	73	0.437	0.15710961
MI	Allegan County	74	0.402	0.15910113
MI	Barry County	44	0.392	0.15943186

[a] D_a is the AIDS prevalence rate in the county a^{th} in 2010. [b] G_a is the Gini income-inequality measure of the US county in 2010.

the absence of values for one or more inputs. To define the sensitivity of a particular variable X, the first run uses the network on a set of test cases and accumulates the network error. In the second step, the network is employed again using the same cases but replacing the observed values of X with the value estimated by the missing value procedure, and again it calculated the accumulated network error. By removing the variable X, it is expected for some deterioration in error to occur. Therefore, the measure of sensitivity is the ratio of the error with missing value substitution to the original error. The more sensitive the network is to a particular input, the greater the deterioration is expected, and therefore the greater the ratio. The elimination of a variable with ratio ≤1 improves or has no effect on the performance of the ANN. After the sensitivities are calculated, they are ranked in order. In Table 7 we can see that the model shows a higher relevance to the information about the molecular structure, parameters of type $I^q{}_k$. Second, the model ranks the information about the organism used to measure the biological activity, parameters of type $\Delta I^q{}_k(c_3)$. The third type of relevant input is the experimental measure used to quantify the activity of the drug, parameters of type $\Delta I^q{}_k(c_1)$. The fourth ranked inputs in order of importance are parameters of type $\Delta I^q{}_k(c_2)$, which quantify the target protein. The fifth type of input quantifies information about the assay protocol used to test the drug. The last effect introduced in the model was the information about income inequality in the county $I^a{}_0(t)$. Thus, the sensitivity analysis shows that the model is ranked according to the importance of factors in the following order (AIDS epidemiology/anti-HIV drug) ≈ structure of drug > organism in preclinical assay > experimental measure of activity > drug target > pharmacological assay > county income inequality. Table 7 depicts the parameters in decreasing order of their contribution to the model (higher contribution => higher ratio => lower rank). The five parameters with higher contribution are the following: $I^q{}_2$, $I^q{}_4$, $\Delta I^q{}_2(c_3)$, $\Delta I^q{}_2(c_1)$, $I^q{}_3$. The parameters of higher contribution for each type of information are the following: $I^q{}_2$ with rank = 1, $\Delta I^q{}_2(c_3)$ rank = 3, $\Delta I^q{}_2(c_1)$ with rank = 4, $\Delta I^q{}_2(c_2)$ with rank = 13, $\Delta I^q{}_2(c_4)$ with rank = 15, and $I^a{}_0(t)$ with rank = 17 (shown in boldface in Table 7).

We retrained the model using only these parameters, but the new ANN fails to generate good predictive models with Sp and Sn < 50%. It means that the model provides a greater importance to the chemical structure and pharmacological information (branch A), with respect to county information (branch B), but it needs all the parameters. This could be explained taking into consideration that branch A includes the higher number of input factors (information considered), whereas branch B includes only one input factor, the income-inequality in the county with respect to the state. We should also note that the only epidemiological feature used as input to calculate the Shannon information indices of the county was the G_a measure of income inequality. The G_a measure of income-inequality is widely used as a descriptor to approach the study of the epidemiology of different diseases.[83,84] The values of $G_a \approx 0$ are characteristic of societies with near-to-ideal equalitarian distribution of income, whereas values of $G_a \approx 1$ are typical of inequality in income distribution.[85] Gant et al.[86] found a positive value of the Pearson

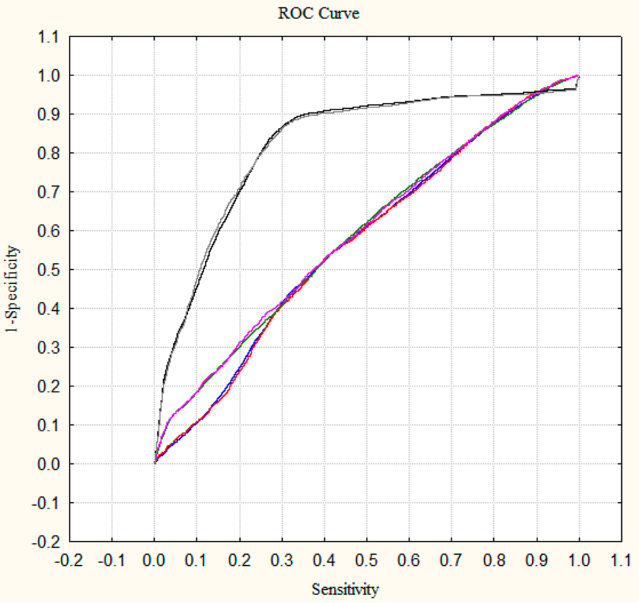
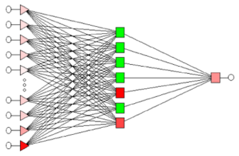
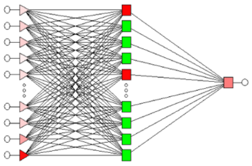
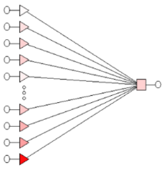


Figure 2. ROC curve analysis of LNN (gray color) vs MLPs (other colors) classifiers.

Table 5. Linear vs Nonlinear ANN Models

ANN models	Data sets	Training set		Validation set	
	Observed	$L_{aq} = 0$	$L_{aq} = 1$	$L_{aq} = 0$	$L_{aq} = 1$
MLP 13:13-7-1:1	Parameter [a]	Sn	Sp	Sn	Sp
	Predicted	55.79	56.17	56.52	55.67
	$L_{aq} = 0$	13469	3637	4534	1237
	$L_{aq} = 1$	10670	4661	3487	1554
	AUROC	0.57		0.57	
MLP 16:16-11-1:1	Parameter [a]	Sn	Sp	Sn	Sp
	Predicted	56.13	56.20	56.63	55.67
	$L_{aq} = 0$	13550	3634	4543	1237
	$L_{aq} = 1$	10589	4664	3478	1554
	AUROC	0.59		0.59	
LNN 19:19-1:1	Parameter [a]	Sn	Sp	Sn	Sp
	Predicted	76.46	77.13	77.30	75.67
	$L_{aq} = 0$	18458	1897	6201	679
	$L_{aq} = 1$	5681	6401	1820	2112
	AUROC	0.82		0.82	

[a] Parameter, Sp = Specificity, Sn = Sensitivity. Columns: observed classifications; Rows: predicted classifications.

Table 6. Predictions of Some Cases with the LNN Network

CMPD_ID	$L_{aq}(c_j)_{obs}$	$L_{aq}(c_j)_{pred}$	c-level	measure	name	target	organism	assay_ID	state_county
1172035	1	1	0.37192	IC$_{50}$ (nM)	Nifeviroc	CC-CKR-5	hsa	1174016	MO_Laclede
1172035	1	1	0.41989	IC$_{50}$ (nM)	Nifeviroc	CC-CKR-5	hsa	1174015	MO_Macon
1201187	1	1	0.43568	IC$_{50}$ (nM)	Maraviroc	CC-CKR-5	hsa	1034062	MI_Tuscola
1201187	1	1	0.44035	IC$_{50}$ (nM)	Maraviroc	CC-CKR-5	hsa	1019461	MN_Nicollet
129	1	1	0.36238	IC$_{50}$ (nM)	Zidovudine	HIV	HIV	640394	IN_Hancock
175691	1	1	0.55885	IC$_{95}$ (nM)	Rilpivirine	HIV	HIV	1930128	WA_Mason
175691	1	1	0.52729	IC$_{95}$ (nM)	Rilpivirine	HIV	HIV	1930283	WA_Pacific
308954	1	1	0.32563	IC$_{50}$ (nM)	Etravirine	HIV	HIV	1006144	GA_Gordon
308954	1	1	0.31279	IC50 (nM)	Etravirine	HIV	HIV	1006139	GA_Lumpkin
57	1	1	0.40037	ED$_{50}$ (nM)	Nevirapine	HIV-1	HIV-1	709947	TX_Dawson
57	1	1	0.45551	ED$_{50}$ (nM)	Nevirapine	HIV-1	HIV-1	709946	TX_Denton
114	1	1	0.34536	IC$_{50}$ (nM)	Saquinavir	HIV-1	HIV-1	755976	IL_Whiteside
114	1	1	0.34894	IC$_{50}$ (nM)	Saquinavir	HIV-1	HIV-1	868005	CA_Mono
114	1	1	0.32824	IC$_{50}$ (nM)	Saquinavir	HIV-1	HIV-1	866135	CA_Placer
129	0	0	0.17658	EC$_{50}$ (nM)	Zidovudine	HIV-1	HIV-1	884233	NC_Durham
129	0	0	0.18634	EC$_{50}$ (nM)	Zidovudine	HIV-1	HIV-1	688523	NC_Edgecombe
141	0	0	0.16086	EC$_{50}$ (nM)	Lamivudine	HIV-1	HIV-1	1263166	GA_Crisp
141	0	0	0.15857	EC$_{50}$ (nM)	Lamivudine	HIV-1	HIV-1	1263167	GA_DeKalb
141	0	0	0.13955	EC$_{50}$ (nM)	Lamivudine	HIV-1	HIV-1	1263157	GA_Dooly
484	0	0	0.26125	EC$_{50}$ (nM)	Adefovir	HIV-1	HIV-1	1831866	OH_Hocking
484	0	0	0.17849	EC$_{50}$ (nM)	Adefovir	HIV-1	HIV-1	1831858	OH_Jackson
1163	0	0	0.14076	EC$_{50}$ (nM)	Atazanavir	HIV2	HIV-2	991367	MO_Polk
1163	0	0	0.09766	EC$_{50}$ (nM)	Atazanavir	HIV2	HIV-2	991368	MO_Taney
1163	0	0	0.15377	IC$_{50}$ (nM)	Atazanavir	HIV2	HIV-2	1262836	TN_Putnam
222559	0	0	0.23747	IC$_{50}$ (nM)	Tipranavir	HIV2	HIV-2	1264851	TX_Camp
222559	0	0	0.20195	IC$_{50}$ (nM)	Tipranavir	HIV2	HIV-2	1262828	TX_Cass
625	0	0	0.26799	EC$_{50}$ (nM)	Thiabendazole	HIV-1	HIV-1	689145	WI_Jefferson

correlation coefficient $\rho = 0.40$ between AIDS diagnosis rates and G_a for 1,560 US counties between 2006 and 2008. However, they also found a positive correlation ($\rho = 0.52$) with proportion unmarried – ages >15 years. The AIDSVu data presented an average value of $G_a = 0.435$ and a standard deviation of only 0.03. The AIDSVu data set analyzed in this work presents an even weaker correlation ($\rho = 0.31$) between AIDS diagnosis rates in 2010 and G_a for the 2,310 US counties studied in this work. It

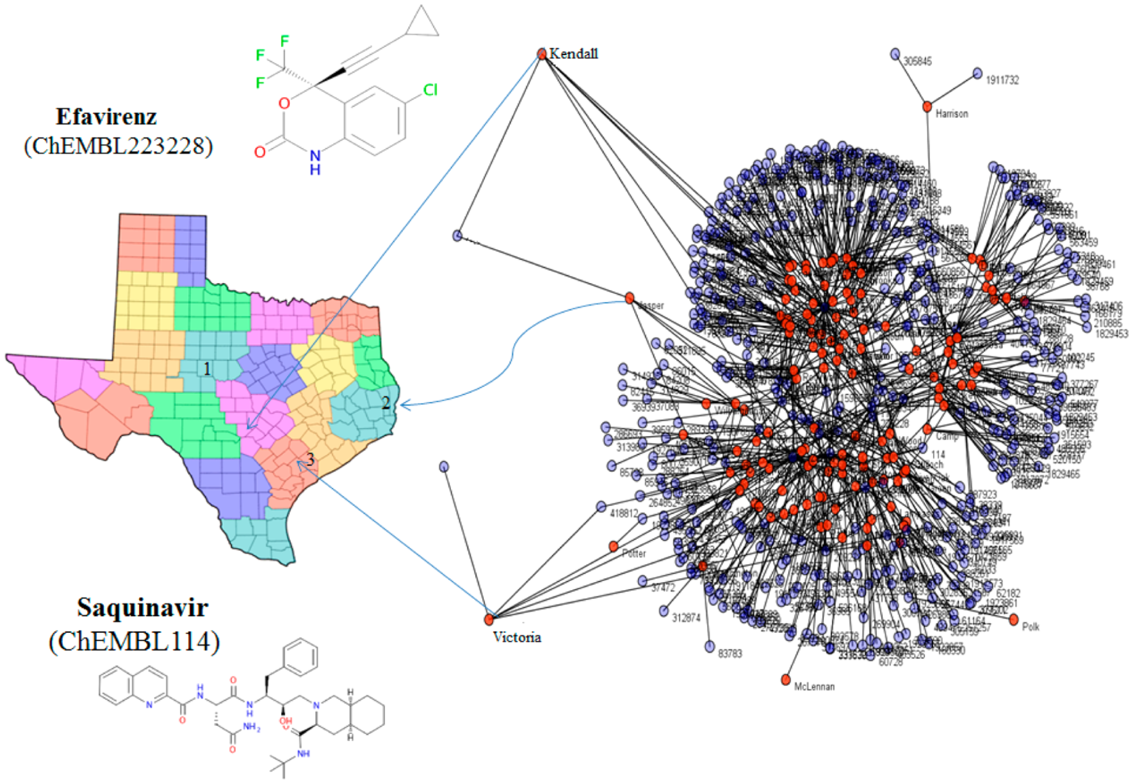


Figure 3. Predicted subnetwork of AIDS prevalence vs anti-HIV drug preclinical activity for Texas.

Table 7. Sensitivity Analysis for the LNN Network

index	name of information indices and their MA operators[a]	ratio	rank
$I^q{}_2$	Balaban V-index for drugs	532.48	**1**
$I^q{}_4$	Balaban X-index for drugs	336.91	2
$\Delta I^q{}_2(c_3)$	MA of V-index of drugs assayed in the same organism	263.34	**3**
$\Delta I^q{}_2(c_1)$	MA for V-index of drugs with the same experimental measure	254.03	**4**
$I^q{}_3$	Balaban Y-index for drugs	194.36	5
$\Delta I^q{}_4(c_3)$	MA of X-index of drugs assayed in the same organism	169.38	6
$\Delta I^q{}_4(c_1)$	MA for X-index of drugs with the same experimental measure	158.25	7
$\Delta I^q{}_3(c_3)$	MA for Y-index of drugs with the same organism	94.37	8
$\Delta I^q{}_3(c_1)$	MA for Y-index of drugs with the same experimental measure	94.09	9
$I^q{}_1$	Balaban U-index for drugs	10.56	10
$\Delta I^q{}_1(c_1)$	MA for U-index of drugs with the same experimental measure	5.55	11
$\Delta I^q{}_1(c_3)$	MA for U-index of drugs with the same organism	5.08	12
$\Delta I^q{}_2(c_2)$	MA for V-index of drugs with the same protein target	1.09	**13**
$\Delta I^q{}_4(c_2)$	MA for X-index of drugs with the same protein target	1.02	14
$\Delta I^q{}_2(c_4)$	MA for V-index of drugs tested in the same assay	1.01	**15**
$\Delta I^q{}_3(c_2)$	MA for Y-index of drugs tested with the same protein target	1.01	16
$I^a{}_0(t)$	Shannon information index based on the Gini coefficient	1.01	**17**
$\Delta I^q{}_4(c_4)$	MA for X-index of drugs tested in the same assay	1.01	18
$\Delta I^q{}_3(c_4)$	MA for Y-index of drugs tested in the same assay	1.0	19

[a] MA = Moving Average operator of Box-Jenkins.

may indicate that possibly we should include other factors in branch B in order to collect additional epidemiological information relevant to the present problem. In upcoming papers we will continue working on the strategy described here, including other information indices of the molecules, other epidemiological factors, different disease transmission matrices, and using different types of machine learning algorithms.

4. CONCLUSIONS

We developed a model called LNN-ALMA to generate complex networks of the AIDS prevalence in the US counties with respect to the preclinical activity of anti-HIV drugs. The best classifier found was the LNN; the inputs of this classifier are based on Balaban information indices. Consequently, this model may be useful to predict the most effective drugs to treat HIV in different populations (from the US counties) with a given epidemiological prevalence. In future work, we will continue to improve the models, and we will include other information indices, social and economic factors, machine-learning techniques, etc.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The additional tables include the information indices for all the molecules, averages of information indices of the molecules, information indices for the all the US counties, and the results of the LNN model. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: humberto.gonzalezdiaz@ehu.es.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors sincerely thank the kind attention of the JCIM Editor Anton J. Hopfinger, University of New Mexico, College of Pharmacy as well as excellent recommendations made by unknown reviewers of this manuscript. R.O.M acknowledges financial support of the FPI fellowship associated with the research project (AGL2011-30563-C03-01) funded by MECD (Spanish Ministry of Education, Culture and Sport).

■ REFERENCES

(1) Yu, F.; Lu, L.; Du, L.; Zhu, X.; Debnath, A. K.; Jiang, S. Approaches for identification of HIV-1 entry inhibitors targeting gp41 pocket. *Viruses* **2013**, *5*, 127−149.

(2) Gengiah, T. N.; Baxter, C.; Mansoor, L. E.; Kharsany, A. B.; Abdool Karim, S. S. A drug evaluation of 1% tenofovir gel and tenofovir disoproxil fumarate tablets for the prevention of HIV infection. *Expert Opin. Invest. Drugs* **2012**, *21*, 695−715.

(3) Cohen, M. S.; Hellmann, N.; Levy, J. A.; DeCock, K.; Lange, J. The spread, treatment, and prevention of HIV-1: evolution of a global pandemic. *J. Clin. Invest.* **2008**, *118*, 1244−1254.

(4) Zuo, T.; Liu, D.; Lv, W.; Wang, X.; Wang, J.; Lv, M.; Huang, W.; Wu, J.; Zhang, H.; Jin, H.; Zhang, L.; Kong, W.; Yu, X. Small-molecule inhibition of human immunodeficiency virus type 1 replication by targeting the interaction between Vif and ElonginC. *J. Virol.* **2012**, *86*, 5497−5507.

(5) Sun, L. Q.; Zhu, L.; Qian, K.; Qin, B.; Huang, L.; Chen, C. H.; Lee, K. H.; Xie, L. Design, synthesis, and preclinical evaluations of novel 4-substituted 1,5-diarylanilines as potent HIV-1 non-nucleoside reverse transcriptase inhibitor (NNRTI) drug candidates. *J. Med. Chem.* **2012**, *55*, 7219−7229.

(6) Deng, K.; Zink, M. C.; Clements, J. E.; Siliciano, R. F. A quantitative measurement of antiviral activity of anti-human immunodeficiency virus type 1 drugs against simian immunodeficiency virus infection: dose-response curve slope strongly influences class-specific inhibitory potential. *J. Virol.* **2012**, *86*, 11368−11372.

(7) Liao, C.; Nicklaus, M. C. Computer tools in the discovery of HIV-1 integrase inhibitors. *Future Med. Chem.* **2010**, *2*, 1123−1140.

(8) Heikamp, K.; Bajorath, J. Large-scale similarity search profiling of ChEMBL compound data sets. *J. Chem. Inf. Model.* **2011**, *51*, 1831−1839.

(9) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−1107.

(10) AIDSVu. http://aidsvu.org/ (accessed September 21, 2013).

(11) Mok, N. Y.; Brenk, R. Mining the ChEMBL database: an efficient chemoinformatics workflow for assembling an ion channel-focused screening library. *J. Chem. Inf. Model.* **2011**, *51*, 2449−2454.

(12) Chiolero, A. Big data in epidemiology: too big to fail? *Epidemiology* **2013**, *24*, 938−939.

(13) Hamilton, B. Impacts of big data. Potential is huge, so are challenges. *Health Manage. Technol.* **2013**, *34*, 12−13.

(14) Mallon, W. J. Big data. *J. Shoulder Elbow Surg.* **2013**, *22*, 1153.

(15) Moore, K. D.; Eyestone, K.; Coddington, D. C. The big deal about big data. *Healthc. Financ. Manage.* **2013**, *67* (60−66), 68.

(16) Toh, S.; Platt, R. Big data in epidemiology: too big to fail? *Epidemiology* **2013**, *24*, 939.

(17) Gijzen, H. Development: Big data for a sustainable future. *Nature* **2013**, *502*, 38.

(18) Hu, Y.; Bajorath, J. Molecular scaffolds with high propensity to form multi-target activity cliffs. *J. Chem. Inf. Model.* **2010**, *50*, 500−510.

(19) Erhan, D.; L'Heureux P, J.; Yue, S. Y.; Bengio, Y. Collaborative filtering on a family of biological targets. *J. Chem. Inf. Model.* **2006**, *46*, 626−635.

(20) Namasivayam, V.; Hu, Y.; Balfer, J.; Bajorath, J. Classification of compounds with distinct or overlapping multi-target activities and diverse molecular mechanisms using emerging chemical patterns. *J. Chem. Inf. Model.* **2013**, *53*, 1272−1281.

(21) Cruz-Monteagudo, M.; Cordeiro, M. N.; Tejera, E.; Dominguez, E. R.; Borges, F. Desirability-based multi-objective QSAR in drug discovery. *Mini-Rev. Med. Chem.* **2012**, *12*, 920−935.

(22) Machado, A.; Tejera, E.; Cruz-Monteagudo, M.; Rebelo, I. 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.* **2009**, *44*, 5045−5054.

(23) Saiz-Urra, L.; Bustillo Perez, A. J.; Cruz-Monteagudo, M.; Pinedo-Rivilla, C.; Aleu, J.; Hernandez-Galan, R.; Collado, I. G. Global antifungal profile optimization of chlorophenyl derivatives against Botrytis cinerea and Colletotrichum gloeosporioides. *J. Agric. Food Chem.* **2009**, *57*, 4838−4843.

(24) Cruz-Monteagudo, M.; Borges, F.; Cordeiro, M. N.; Cagide Fajin, J. L.; Morell, C.; Ruiz, R. M.; Canizares-Carmenate, Y.; Dominguez, E. R. Desirability-based methods of multiobjective optimization and ranking for global QSAR studies. Filtering safe and potent drug candidates from combinatorial libraries. *J. Comb. Chem.* **2008**, *10*, 897−913.

(25) Nicolaou, C. A.; Brown, N.; Pattichis, C. S. Molecular optimization using computational multi-objective methods. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 316−324.

(26) Mekenyan, O.; Bonchev, D.; Trinajstic, N. Chemical graph theory modeling the thermodynamic properties of molecules. *Int. J. Quantum Chem., Symp.* **1980**, *18*, 369−380.

(27) Bonchev, D.; Trinajstic, N. Information theory, distance matrix, and molecular branching. *J. Chem. Phys.* **1977**, *67*, 4517−4533.

(28) Bonchev, D.; Kamenski, D.; Kamenska, V. Symmetry and information content of chemical structures. *Bull. Math. Biol.* **1976**, *38*, 119−133.

(29) Kier, L. B. Use of molecular negentropy to encode structure governing biological activity. *J. Pharm. Sci.* **1980**, *69*, 807−810.

(30) Stahura, F. L.; Godden, J. W.; Bajorath, J. Differential Shannon entropy analysis identifies molecular property descriptors that predict aqueous solubility of synthetic compounds with high accuracy in binary QSAR calculations. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 550−558.

(31) Stahura, F. L.; Godden, J. W.; Xue, L.; Bajorath, J. Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245−1252.

(32) Agrawal, V. K.; Khadikar, P. V. Modelling of carbonic anhydrase inhibitory activity of sulfonamides using molecular negentropy. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 447−453.

(33) Katritzky, A. R.; Lomaka, A.; Petrukhin, R.; Jain, R.; Karelson, M.; Visser, A. E.; Rogers, R. D. QSPR correlation of the melting point for pyridinium bromides, potential ionic liquids. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 71−74.

(34) Katritzky, A. R.; Perumal, S.; Petrukhin, R.; Kleinpeter, E. Codessa-based theoretical QSPR model for hydantoin HPLC-RT lipophilicities. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 569−574.

(35) Roy, K.; Saha, A. Comparative QSPR studies with molecular connectivity, molecular negentropy and TAU indices. Part I: molecular thermochemical properties of diverse functional acyclic compounds. *J. Mol. Model.* **2003**, *9*, 259−270.

(36) Graham, D. J.; Schacht, D. V. Base information content in organic formulas. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 942−946.

(37) Graham, D. J. Information content in organic molecules: structure considerations based on integer statistics. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 215.

(38) Graham, D. J.; Schulmerich, M. V. Information content in organic molecules: reaction pathway analysis via Brownian processing. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1612−1622.

(39) Graham, D. J. Information content in organic molecules: Brownian processing at low levels. *J. Chem. Inf. Model.* **2007**, *47*, 376−389.

(40) Graham, D. J. Information content in organic molecules: aggregation states and solvent effects. *J. Chem. Inf. Model.* **2005**, *45*, 1223−1236.

(41) Strait, B. J.; Dewey, T. G. The Shannon information entropy of protein sequences. *Biophys. J.* **1996**, *71*, 148−155.

(42) Dima, R. I.; Thirumalai, D. Proteins associated with diseases show enhanced sequence correlation between charged residues. *Bioinformatics* **2004**, *20*, 2345−2354.

(43) Loewenstern, D.; Yianilos, P. N. Significantly lower entropy estimates for natural DNA sequences. *J. Comput. Biol.* **1999**, *6*, 125−142.

(44) Manke, T.; Demetrius, L.; Vingron, M. Lethality and entropy of protein interaction networks. *Genome Inform.* **2005**, *16*, 159−163.

(45) Mikolas, P.; Vyhnanek, J.; Skoch, A.; Horacek, J. Analysis of fMRI time-series by entropy measures. *Neuroendocrinol. Lett.* **2012**, *33*, 471−476.

(46) Tenorio-Borroto, E.; Garcia-Mera, X.; Penuelas-Rivas, C. G.; Vasquez-Chagoyan, J. C.; Prado-Prado, F. J.; Castanedo, N.; Gonzalez-Diaz, H. Entropy model for multiplex drug-target interaction endpoints of drug immunotoxicity. *Curr. Top. Med. Chem.* **2013**, *13*, 1636−1649.

(47) Bonchev, D.; Mekenyan, O.; Trinajstic, N. Isomer discrimation by topological information approach. *J. Comput. Chem.* **1981**, *2*, 127−148.

(48) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; Research Studies Press: Chichester, U.K., 1983.

(49) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2000.

(50) Ivanciuc, O.; Balaban, A. T. Design of topological indices. Part 20. Molecular structure descriptors computed with information on distances operators. *Rev. Roum. Chim.* **1999**, *44*, 479−489.

(51) Ivanciuc, O.; Ivanciuc, T.; Klein, D. J. Quantitative structure-property relationships generated with optimizable even/odd Wiener polynomial descriptors. *SAR QSAR Environ. Res.* **2001**, *12*, 1−16.

(52) Dehmer, M. M.; Barbarini, N. N.; Varmuza, K. K.; Graber, A. A. Novel topological descriptors for analyzing biological networks. *BMC Struct. Biol.* **2010**, *10*, 18.

(53) Dehmer, M.; Grabner, M.; Varmuza, K. Information indices with high discriminative power for graphs. *PLoS One* **2012**, *7*, e31214.

(54) Dehmer, M.; Mowshowitz, A. A history of graph entropy measures. *Inf. Sci. (N.Y.)* **2011**, *181*, 57−58.

(55) Emmert-Streib, F.; Dehmer, M. Information theoretic measures of UHG graphs with low computational complexity. *Appl. Math. Comp.* **2007**, *190*, 1783−1794.

(56) Balaban, A. T.; Balaban, T. S. New vertex invariants and topological indices of chemical graphs based on information on distances. *J. Math. Chem.* **1991**, *8*, 383−397.

(57) Box, G. E. P.; Jenkins, G. M. *Time series analysis: Forecasting and control*; Holden-Day: San Francisco, CA, 1970.

(58) Botella-Rocamora, P.; Lopez-Quilez, A.; Martinez-Beneito, M. A. Spatial moving average risk smoothing. *Stat. Med.* **2013**, *32*, 2595−2612.

(59) Speck-Planche, A.; Kleandrova, V. V.; Cordeiro, M. N. Chemoinformatics for rational discovery of safe antibacterial drugs: simultaneous predictions of biological activity against streptococci and toxicological profiles in laboratory animals. *Bioorg. Med. Chem.* **2013**, *21*, 2727−2732.

(60) Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. Chemoinformatics in multi-target drug discovery for anti-cancer therapy: in silico design of potent and versatile anti-brain tumor agents. *Anti-Cancer Agents Med. Chem.* **2012**, *12*, 678−685.

(61) Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. Chemoinformatics in anti-cancer chemotherapy: multi-target QSAR model for the in silico discovery of anti-breast cancer agents. *Eur. J. Pharm. Sci.* **2012**, *47*, 273−279.

(62) Goles, E.; Palacios, A. G. Dynamical complexity in cognitive neural networks. *Biol. Res.* **2007**, *40*, 479−485.

(63) Ramesh, A. N.; Kambhampati, C.; Monson, J. R.; Drew, P. J. Artificial intelligence in medicine. *Ann. R. Coll. Surg. Engl.* **2004**, *86*, 334−338.

(64) Wesolowski, M.; Suchacz, B. Artificial neural networks: theoretical background and pharmaceutical applications: a review. *J. AOAC Int.* **2012**, *95*, 652−668.

(65) Baykal, H.; Yildirim, H. K. Application of artificial neural networks (ANNs) in wine technology. *Crit. Rev. Food Sci. Nutr.* **2013**, *53*, 415−421.

(66) Ponulak, F.; Kasinski, A. Introduction to spiking neural networks: Information processing, learning and applications. *Acta Neurobiol. Exp.* **2011**, *71*, 409−433.

(67) Ghosh-Dastidar, S.; Adeli, H. A new supervised learning algorithm for multiple spiking neural networks with application in epilepsy and seizure detection. *Neural Netw.* **2009**, *22*, 1419−1431.

(68) *DRAGON*, version 5.3; Talete srl: Milano, Italy, 2005.

(69) *STATISTICA*, version 6.0; StatSoft Inc.: Tulsa, OK, 2001.

(70) Hill, T.; Lewicki, P. *STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining*; StatSoft: Tulsa, OK, 2006.

(71) Ivanciuc, O.; Balaban, T. S.; Balaban, A. T. Chemical graphs with degenerate topological indices based on information on distances. *J. Math. Chem.* **1993**, *14*, 21−33.

(72) Pabayo, R.; Kawachi, I.; Gilman, S. E. Income inequality among American states and the incidence of major depression. *J. Epidemiol. Community Health* **2014**, *68*, 110−115.

(73) Riera-Fernandez, P.; Munteanu, C. R.; Escobar, M.; Prado-Prado, F.; Martin-Romalde, R.; Pereira, D.; Villalba, K.; Duardo-Sanchez, A.; Gonzalez-Diaz, H. 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.* **2012**, *293*, 174−188.

(74) Rosenblatt, F. *Principles of neurodynamics; perceptrons and the theory of brain mechanisms*; Spartan Books: Washington, DC, 1962.

(75) Shubber, Z.; Calmy, A.; Andrieux-Meyer, I.; Vitoria, M.; Renaud-Thery, F.; Shaffer, N.; Hargreaves, S.; Mills, E. J.; Ford, N. Adverse events associated with nevirapine and efavirenz-based first-line antiretroviral therapy: a systematic review and meta-analysis. *AIDS* **2013**, *27*, 1403−1412.

(76) Hecker, N.; Ahmed, J.; von Eichborn, J.; Dunkel, M.; Macha, K.; Eckert, A.; Gilson, M. K.; Bourne, P. E.; Preissner, R. SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Res.* **2012**, *40*, D1113−1117.

(77) Prado-Prado, F.; Garcia-Mera, X.; Escobar, M.; Alonso, N.; Caamano, O.; Yanez, M.; Gonzalez-Diaz, H. 3D MI-DRAGON: new model for the reconstruction of US FDA drug- target network and theoretical-experimental studies of inhibitors of rasagiline derivatives for AChE. *Curr. Top. Med. Chem.* **2012**, *12*, 1843−1865.

(78) Prado-Prado, F.; Garcia-Mera, X.; Abeijon, P.; Alonso, N.; Caamano, O.; Yanez, M.; Garate, T.; Mezo, M.; Gonzalez-Warleta, M.; Muino, L.; Ubeira, F. M.; Gonzalez-Diaz, H. Using entropy of drug and protein graphs to predict FDA drug-target network: theoretic-experimental study of MAO inhibitors and hemoglobin peptides from Fasciola hepatica. *Eur. J. Med. Chem.* **2011**, *46*, 1074−1094.

(79) Araujo, R. P.; Liotta, L. A.; Petricoin, E. F. Proteins, drug targets and the mechanisms they control: the simple truth about complex networks. *Nat. Rev. Drug Discovery* **2007**, *6*, 871−880.

(80) Vina, D.; Uriarte, E.; Orallo, F.; Gonzalez-Diaz, H. Alignment-free prediction of a drug-target complex network based on parameters of drug connectivity and protein sequence of receptors. *Mol. Pharmaceutics* **2009**, *6*, 825−835.

(81) Barabasi, A. L.; Gulbahce, N.; Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **2011**, *12*, 56−68.

(82) Hunter, A.; Kennedy, L.; Henry, J.; Ferguson, I. Application of neural networks and sensitivity analysis to improved prediction of trauma survival. *Comput. Methods Programs Biomed.* **2000**, *62*, 11−19.

(83) Burns, J. K.; Tomita, A.; Kapadia, A. S. Income inequality and schizophrenia: Increased schizophrenia incidence in countries with high levels of income inequality. *Int. J. Soc. Psychiatry* **2013**, in press.

(84) Green, C.; Yu, B. N.; Marrie, R. A. Exploring the implications of small-area variation in the incidence of multiple sclerosis. *Am. J. Epidemiol.* **2013**, *178*, 1059−1066.

(85) Feigl, A. B.; Ding, E. L. Evidenced formal coverage index and universal healthcare enactment: A prospective longitudinal study of economic, social, and political predictors of 194 countries. *Health Policy* **2013**, *113*, 50−60.

(86) Gant, Z.; Lomotey, M.; Hall, H. I.; Hu, X.; Guo, X.; Song, R. A county-level examination of the relationship between HIV and social determinants of health: 40 states, 2006−2008. *Open AIDS J.* **2012**, *6*, 1−7.

7.2 Publicación 2

Herrera-Ibatá DM, Orbegozo-Medina RA, and González-Díaz H. Multiscale Mapping of AIDS in U.S. counties vs. Anti-HIV Drugs Activity with Complex Networks and Information Indices. *Journal Current Bioinformatics*, (2015), *en imprenta*.

Current Bioinformatics, **2015, x, xxx - xxx**

Multiscale Mapping of AIDS in U.S. Counties *vs.* Anti-HIV Drugs Activity with Complex Networks and Information Indices

Diana María Herrera-Ibatá[1*], Ricardo Alfredo Orbegozo-Medina[2], and Humberto González-Díaz[3,4]

[1] *Department of Information and Communication Technologies, University of A Coruña UDC, 15071, A Coruña, Spain*.

[2] *Department of Microbiology and Parasitology, University of Santiago de Compostela (USC), 15782, Santiago de Compostela, A Coruña, Spain.*

[3] *Department of Organic Chemistry II, Faculty of Science and Technology, University of the Basque Country UPV/EHU, 48940, Leioa, Spain.*

[4] *IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain.*

Abstract In this work, we reviewed different aspects about the epidemiology, drugs, targets, chem-bioinformatics, and systems biology methods, related to AIDS/HIV. Next, we developed a new model to predict complex networks of the prevalence of AIDS in U.S. counties taking into consideration the values of Gini coefficients of social income inequality. We also used activity/structure data of anti-HIV drugs in preclinical assays. First, we trained different Artificial Neural Networks (ANNs) using as input Markov and Symmetry information indices of social networks and of molecular graphs. We obtained the data about AIDS prevalence and Gini coefficient from the AIDSVu database of Emory University and the data about anti-HIV drugs from ChEMBL database. We used Box-Jenkins operators to measure the shift with respect to average behavior of counties from states and drugs from reference compounds assayed in a given protocol, target, or organism. To train/validate the model and predict the complex network we needed to analyzes 43,249 data points including values of AIDS prevalence in 2310 counties in U.S. *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 Neural Network (LNN) with Accuracy, Specificity, Sensitivity, and AUROC above 0.72-0.73 in training and external validation series. The new linear equation was shown to be useful to generate complex network maps of drug activity *vs.* AIDS/HIV epidemiology in U.S. at county level.

Keywords: Anti-HIV drugs; AIDS in U.S. at county level; Gini coefficient; Multiscale models; Box-Jenkins moving average operators; Shannon Entropy; indices of neighborhood symmetry.

INTRODUCTION

AIDS/HIV EPIDEMIOLOGY AND CHEMOTHERAPY

Human immunodeficiency virus (HIV) is a retrovirus belonging to the family of lentiviruses that causes AIDS. Retroviruses[1] can use their RNA and host DNA to make viral DNA, and are known for their long incubation periods. There are two types of HIV: HIV type 1 and HIV type 2. Worldwide, the predominant virus is HIV-1, the HIV-2 is less common and appears to progress more slowly [2]. In AIDS, the immune system is severely affected, and the body is susceptible to a variety of infections, such as bacteria, parasites, viruses, etc., which can cause fatal diseases in people with the syndrome [3]. Since the virus was discovered in the 1980s, many progresses have been made in the management of HIV/AIDS [4]. Antiretroviral treatment which can stop the replication of HIV, has been one of the most important discoveries [5].

Despite progresses, HIV [5] remains a public health challenge.After thirty years in the AIDS epidemic, there are over 34 million people living with HIV [6], and still 2.5 million new infections and 1.7 million deaths each year.

AIDS PREVALENCE IN U.S. AT COUNTY LEVEL

There are several databases with epidemiological data about AIDS prevalence. For instance, researchers at the Rollins School of Public Health at Emory University compiled state and county level information for AIDS prevalence in the U.S. in the database AIDSVu (http://aidsvu.org/about-aidsvu/). This information on AIDSVu comes from the U.S. Centers for Disease Control and Prevention's (CDC) national surveillance database. AIDSVu is an interactive map that is available online and it shows the prevalence of HIV in the United States. The values used in this study were rate of adults/adolescents living with an HIV diagnosis in 2010 *per* 100,000 populations (see **Figure 1**). The county-level HIV data is estimated for persons aged 13 and older living with an HIV infection diagnosis. All race groups are non-Hispanic, and the Hispanic/Latino ethnicity is inclusive of all races. The total number of counties is $n_a = 2310$.

***Corresponding author:** Herrera-Ibatá DM, Email: diana.herrera@udc.es, *Department of Information and Communication Technologies, University of A Coruña UDC, 15071, A Coruña, Spain.*

1574-8936/14 $58.00+.00

The Human immunodeficiency virus has prompted many researchers worldwide to discover new compounds and/or molecular or cellular targets to become useful against the disease. There are different dataset with experimental outcomes of the interactions of anti-HIV compounds with their respective targets. ChEMBL (https://www.ebi.ac.uk/chembl/) [7-9] is one of the biggest; ChEMBL is an open large-scale bioactivity database containing information largely manually extracted from the medicinal chemistry literature. Information regarding the compounds tested (including their structures), the biological or physicochemical assays performed on these and the targets of these assays are recorded in a structured form, there are now >1.3 million distinct compound structures and 12 million bioactivity data points. The data are mapped to >9000 targets, of which 2827 are human protein targets [9]. Specifically, ChEMBL contains > 43,000 outcomes for assays of anti-HIV compounds. CHEMBL contains a general data set of potential anti-HIV compounds composed of >8,000 multiplexing assay endpoints (results of multiple assays) [7, 8]. The dataset used to train and validate our model includes N = 43,249 statistical cases formed by N_d = 21,582 unique drugs; which have been assayed each one in at least one out of 10 possible standard type measures determined in at least one out of 4856 different assays (experimental protocols reported as different in ChEMBL). Each assay involves, in turn, at least one out of 9 non-molecular or protein targets expressed in the tissues, cells, or viral particles of at least one out of 5 different organisms (including human cells lines). In this work, we reviewed and tabulated a number of compounds, targets, biological activities present in ChEMBL for the HIV (see **Table 1**).

HIV TARGETS

Some of the more important targets are proteins present in the virus or in the host. The spikes projecting from the surface of HIV-1 are composed of the Envelope glycoprotein (Env), Env facilitates HIV entry by a process of direct fusion between the virion membrane and the target cell [10]. Envelope glycoprotein consists of two noncovalently associated subunits derived by proteolytic cleavage of the gp160 biosynthetic precursor: the external subunit gp120, which is responsible for binding to specific target cell receptors [10]. The fusion of the HIV-1 Envelope glycoprotein and target cell membranes is initiated by binding of the viral Envelope surface subunit gp120 to the CD4 receptor on the surface of the $CD4^+$ T cells [11]. This interaction creates a high affinity binding site for a chemokine coreceptor like CXCR4 and/or CCR5, necessary for HIV-1 entry into the target cell and subsequent infection [12]. The C-C chemokine receptor type 3 and C-C chemokine receptor type 2 are alternatives coreceptors with CD4 for HIV-1 infection [13]. This fusion introduces the contents of the virion into the cytoplasm of the cell, setting the stage for reverse transcription and thus to convert their RNA genomes into DNA.

The reverse transcription is an essential step in retroviral replication [14]. Once the reverse transcription has occurred, the integrase enzyme facilitates the incorporation of HIV-1 proviral DNA into the host cell genome and catalyses a function vital to viral replication [15]. After, the HIV protease plays a crucial role in the viral life cycle by processing polyproteins into structural and functional proteins essential for viral maturation [16, 17]. In **Figure 2,** you can see some viral and human targets involved in the antiretroviral therapy. The 3D structural models of targets for anti-HIV inhibitors illustrated in this figure were downloaded from PDBe (http://www.ebi.ac.uk/pdbe/). The proteins illustrated in Figure 2 were Human CXCR4 (file: 3oe8) [18] , Human CCR5 (file: 4mbs)[19], HIV-1 Reverse Transcriptase (file:4mfb) [20], HIV integrase (file:3vqe) [21], HIV protease (file:4he9) [22].

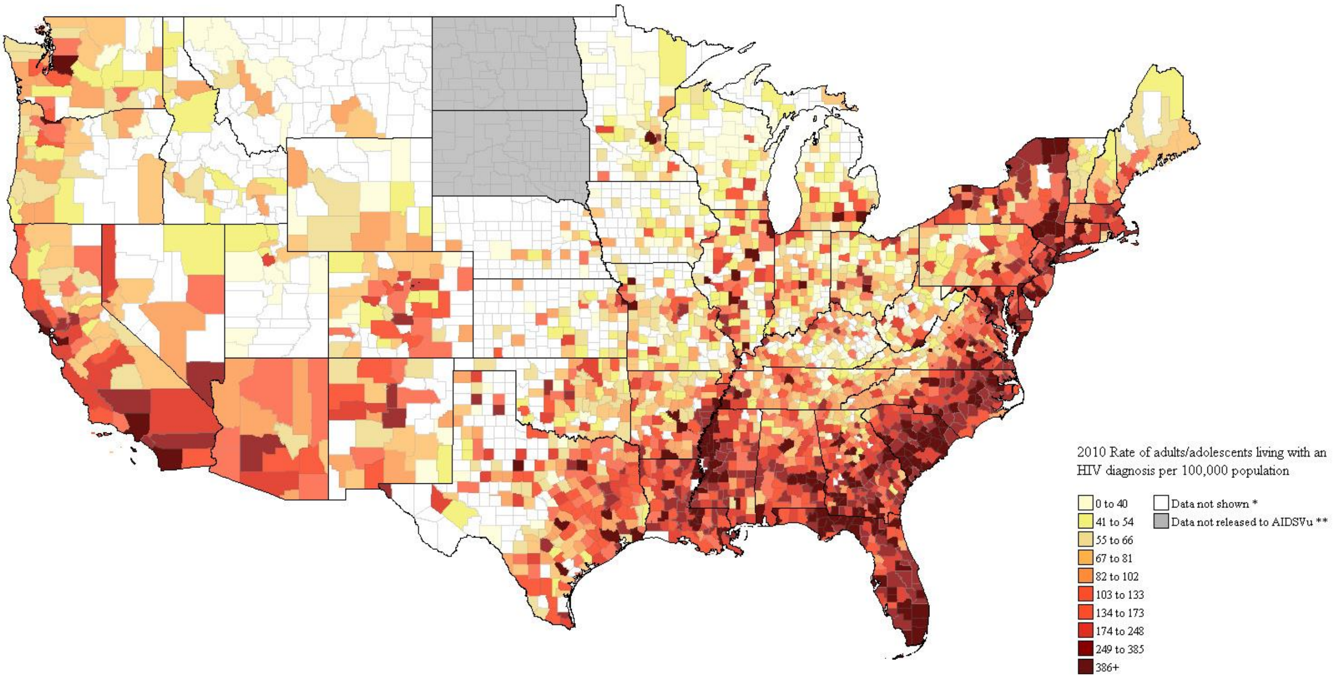

2010 Rate of adults/adolescents living with an HIV diagnosis per 100,000 population
0 to 40
41 to 54
55 to 66
67 to 81
82 to 102
103 to 133
134 to 173
174 to 248
249 to 385
386+
Data not shown *
Data not released to AIDSVu **

Fig. (1). AIDSVu map of HIV rate in U.S. at county level 2010

Table 1. Review of ChEMBL outcomes for assays of anti-HIV drugs

CHEMBLID	Target	Accession	Target type	Organism	Compounds	Activities
CHEMBL274	C-C chemokine receptor type 5	P51681	protein	*H. sapiens*	2922	4740
CHEMBL4015	C-C chemokine receptor type 2	P41597	protein	*H. sapiens*	2567	4674
CHEMBL2414	C-C chemokine receptor type 4	P51679	protein	*H. sapiens*	1335	2489
CHEMBL2107	C-X-C chemokine receptor type 4	P61073	protein	*H. sapiens*	550	967
CHEMBL3473	C-C chemokine receptor type 3	P51677	protein	*H. sapiens*	1276	1550
CHEMBL5412	C-C chemokine receptor type 2	P51683	protein	*M. musculus*	74	90
CHEMBL3676	C-C chemokine receptor type 5	P51682	protein	*M. musculus*	63	65
CHEMBL378	HIV type 1		organism	*H. sapiens*	16547	41411
CHEMBL243	HIV type 1 protease	Q72874	protein	*H. sapiens*	5503	8268
CHEMBL247	HIV type 1 reverse transcriptase	Q72547	protein	*H. sapiens*	3292	7187
CHEMBL380	HIV type 2		organism	*H. sapiens*	1703	2834
CHEMBL613758	HIV		organism	*H. sapiens*	1421	2718
CHEMBL3471	HIV type 1 integrase	Q7ZJM1	protein	*H. sapiens*	1269	2910
CHEMBL612359	HIV3		organism	*H. sapiens*	104	136
CHEMBL613498	HIV type 1		organism	*H. sapiens*	76	76
CHEMBL4609	HIV type 1 Tat protein	P04326	protein	*H. sapiens*	71	75
CHEMBL3520	Envelope polyprotein GP160	P04578	protein	HIV type 1	129	330

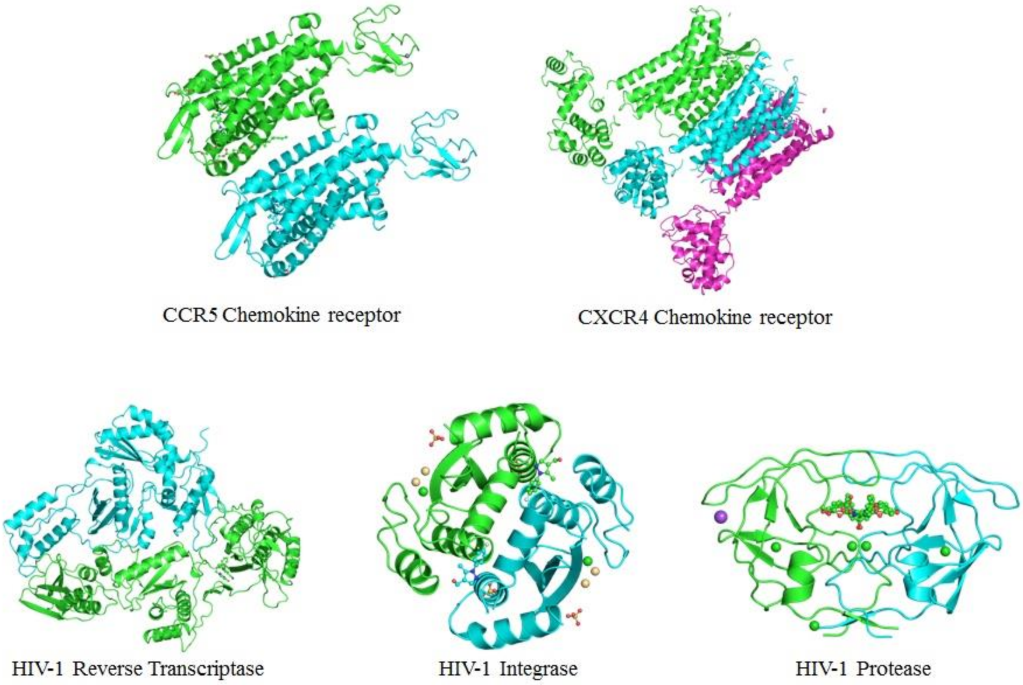
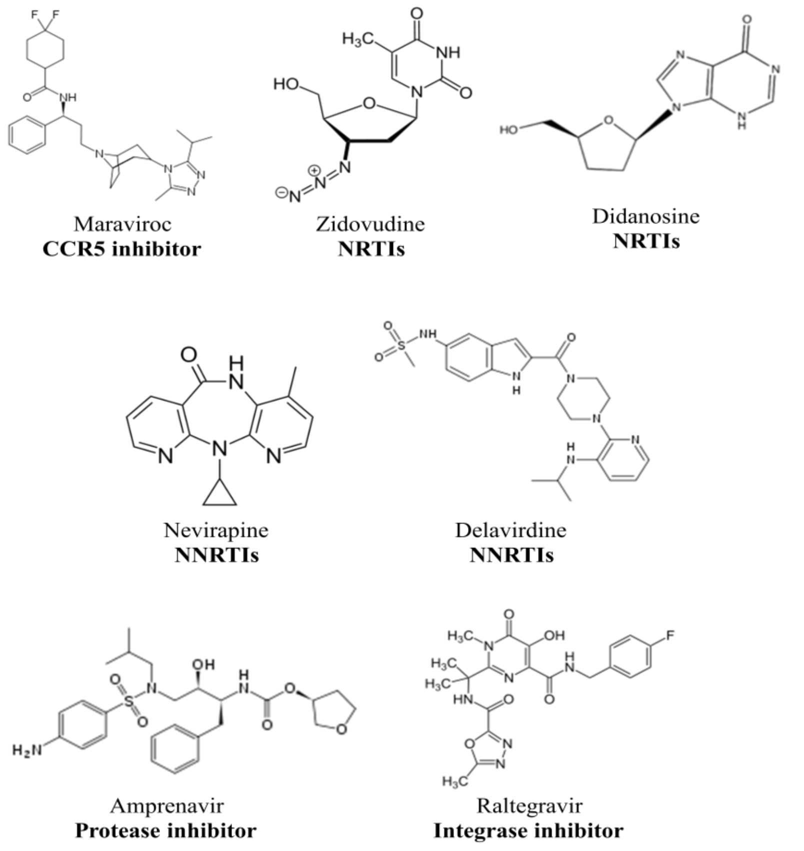


Fig. (2). 3D structural models of targets for anti-HIV inhibitors downloaded from PDBe (http://www.ebi.ac.uk/pdbe/): Human CXCR4 (file: 3oe8) [18] , Human CCR5 (file: 4mbs)[19], HIV-1 Reverse Transcriptase (file:4mfb) [20], HIV integrase (file:3vqe) [21], HIV protease (file:4he9) [22].

In **Figure 3**, we illustrate the structure of some compounds in the fight against HIV. Additionally, the antiretroviral therapy includes the entry inhibitors; they can be subdivided into distinct sub-classes. The first one is the fusion inhibitors, like Enfuvirtide, that inhibits entry of HIV into the CD4 cell because they bind to glycoprotein gp41 (a protein on the viral membrane). This sub-class prevents fusion of the virus and the CD4 cell membrane [23]. The second one is the CCR5 inhibitors, like Maraviroc, 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 [11]. Another type of anti-HIV drugs are the Nucleoside/Nucleotide Reverse Transcriptase Inhibitors (NRTI's). They inhibit the viral reverse transcriptase enzyme, which is responsible for transcribing viral RNA into double stranded DNA. Some examples of this class of drugs are: Zidovudine Didanosine, Zalcitabine, Stavudine, Lamivudine, Abacavir, Tenofovir, Emcitrabine [24]. There are also Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs) that inhibit the viral reverse transcriptase enzyme. NNRTIs bind directly to the reverse transcriptase enzyme [25]. Currently there are four: Nevirapine, Delavirdine, Efavirenz, Etravirine [26]. Among other important anti-HIV compounds are the integrase inhibitors. The integration is an essential element of the HIV type 1 (HIV-1) replication cycle, allowing the transfer of virally encoded DNA into the host chromosome before replication, the Raltegravir, Elvitegravir and Dolutegravir are integrase inhibitors [27, 28].

Last, the protease inhibitors are a different type of anti-HIV compounds, the protease is necessary to form a fully mature, functional virus that is capable to replicate and produce more virus. Some examples of this kind of drugs are Amprenavir, Atazanavir, Indinavir, Nelfinavir, Lopinavir, Saquinavir, Tipranavir, Ritonavir [26, 29].



Maraviroc
CCR5 inhibitor

Zidovudine
NRTIs

Didanosine
NRTIs

Nevirapine
NNRTIs

Delavirdine
NNRTIs

Amprenavir
Protease inhibitor

Raltegravir
Integrase inhibitor

Fig. (3). Chemical structure of some anti-HIV drugs

CHEMOINFORMATICS METHODS FOR ANTI-HIV DRUGS DISCOVERY

In this situation, *in silico* methods allow an efficient characterization of structure-activity relationships (SARs) and ease building of diverse models to capture and codify one or several SARs, which can be used to predict activities for new molecules [30]. To increase the accuracy, artificial intelligence techniques have been applied to Quantitative structure-activity relationship (QSAR) or quantitative structure–property relationships (QSPR) analysis since the late 1980s [31-33]. Magalhães *et al.* [34] studied two- and three-dimensional QSAR (2D/3D-QSAR) on a series of HIV-1 integrase inhibitors (HIV-1 IN) contributing to the design of new and more potent derivatives. Saranya and Selvaraj [35] developed QSAR studies on HIV-1 protease inhibitors and this models have potential application in the prediction of binding affinity for the newly synthesized compounds. Debnath [36] reviewed the applications of 3D-QSAR studies in anti-HIV-1 drug design during the last decade, and highlighted that the effort of the structure-based drug design was really successful in identifying several drugs that are currently available for the treatment of HIV-1, and other applications such as understanding the drug-receptor interactions and help in the design of effective analogs. Some authors [37-39] indicate that the results of their *in silico* studies provide a useful contribution to the design of novel active molecules for the inhibition of some target protein involved in the HIV.

MULTISCALE MODELS OF ANTI-HIV DRUG PRECLINICAL ACTIVITY *VS.* AIDS EPIDEMIOLOGY

A useful chemoinformatics-pharmacoepidemiology model must be multi-level to account molecular and population structure. We need to process diverse types of input data. Initially, we need the information about the anti-HIV drugs, such as chemical structure of the drug (level i) and preclinical information, like biological targets (level ii), organisms (level iii), or assay protocols (level iv). Afterwards, we need to incorporate population structure descriptors (level v) that quantify the epidemiological and socioeconomic factors affecting the population selected for the study. Last, 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 (level vi). Data for levels i-iv can be obtained from public databases of biological activity of organic compounds. In addition, we can obtain data of levels v and vi from public epidemiological databases like AIDSVu (mentioned above). We can talk about three characteristic 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. Multi-targeting [40-42], refers to the existence of compounds that can interact with more than one biological target. Multi-objective optimization problem (MOOP) [43-47] refers to the necessity of prediction/optimization of results for different experimental measures obtained in different assays. Last, multi-scaling refers to the different structural levels of the organization (i-vi) of matter the input variables. It means that we need to develop models able to link the changes in the prevalence of AIDS in a given ath population with the changes in the biological activity of the qth drug (d$_q$), due to variations in chemical structure, detected in preclinical assays carry out under a set of jth conditions (c$_j$).

REVIEW OF INFORMATION INDICES POTENTIALLY USEFUL FOR AIDS/HIV MULTISCALE STUDIES

Shannon entropy measures are universal information indices helpful in many disciplines. Shannon developed the original form of these parameters in a paper about information theory [48]. Godden et al. [49-51] have used Shannon entropy parameters as molecular descriptors to seek QSPR models with different applications. Other authors such as, Acharya *et al.* [52] used non-linear methods including Shannon's entropy in their study about Coronary artery disease (CAD) to evaluate the heart rate of normal and CAD-affected heart rate signals. Many other authors used Shannon's entropy parameters to encode small molecule structure [53-56]. Graham *et al.* [57-62] used entropy measures to study the information properties of organic molecules.

Symmetry information indices of molecular structure

We can use quantitative descriptors of the molecular graph of the drug. In particular, some of these parameters are useful to quantify information about the properties of biological, molecular, and/or social systems (information measures). We are going to use the information indices of neighborhood symmetry. In this work, we also call them symmetry information indices. Several authors in their publications report the use of some of this indices [63], such as Sharma et al. [64] used the bond and structural information content of five-order neighborhood symmetry, among others descriptors to optimize the caspase-3 inhibitory activity of isoquinoline-1,3,4-trione derivatives. Singh *et al.* [65] developed a statistical significant model considering the topological descriptors, including the indices of neighborhood symmetry in their study about the malonyl-CoA decarboxylase (MCD) inhibition activity. Prabhakar *et al.* [66] modeling the HIV-1 RT inhibitory activity of 2-(2,6-dihalophenyl)-3-(substituted pyridin-2-yl)-thiazolidin-4-ones with different topological descriptors obtained from Dragon software. Singh and Shekhawat [67] used the index of 5-order neighborhood symmetry, and highlighted the role of this descriptor in their paper of antimalarial activity of natural and synthetic prodiginines. These indices are calculated for H-included molecular graph and based on neighbor degrees and edge multiplicity [63, 68]. The symmetry information indices are calculated by partitioning graph vertices into equivalence classes; the topological equivalence of two vertices is that the corresponding neighborhoods of the kth order are the same. DRAGON version 5.3 Software [69], calculates information indices of neighborhood symmetry from order 0 up to 5 (see **Table 2**). The names, symbols (IC$_k$, TIC$_k$, SIC$_k$, BIC$_k$, CIC$_k$), and formula for the calculation of symmetry information indices of molecules appear at follow:

There are five types of neighborhood symmetry indices: the first one is the neighborhood information content (IC$_k$). The IC$_k$ uses the following parameters: A$_g$ is the cardinality of the gth equivalence class and *n*AT is the total number of atoms. This index represents a measure of structural complexity per vertex. The formula of IC$_k$ is:

$$IC_k = -\sum_{g=1}^{G} \frac{A_g}{nAT} \cdot \log_2 \frac{A_g}{nAT} \quad (1)$$

The next index is the neighborhood total information content (TIC$_k$), it represents a measure of the graph complexity. The formula is as follows:

$$TIC_k = nAT \cdot IC_k \quad (2)$$

The third index is the structural information content (SIC$_k$), it is calculated as you can see in the formula, in a normalized form of IC$_k$ to delete the influence of graph size:

$$SIC_k = \frac{IC_k}{\log_2 nAT} \quad (3)$$

The fourth index is the bonding information content (BIC$_k$), this index is calculated in a normalized form of the IC$_k$ taking into account the number of bonds and their multiplicity. It uses the *n*BT parameter, which is the number of bonds and π* is the conventional bond order (1 for single, 2 for double, 3 for triple and 1.5 for aromatic bonds). The formula is:

$$BIC_k = \frac{IC_k}{\log_2\left(\sum_{b=1}^{nBT} \pi_b^*\right)} \qquad (4)$$

Moreover, the last index is the complementary information content (CIC$_k$). It measures the deviation IC$_k$ from its maximum value that corresponds to the vertex partition into equivalence classes containing one element each. The formula CIC$_k$ is:

$$CIC_k = \log_2 nAT - IC_k \qquad (5)$$

Table 2. Neighborhood symmetry indices calculated by Dragon v5.3 software

k=order	f=family	Index $^q IC_{kf}$	Dragon symbol	Family [a]
0	1	$^q IC_{01}$	IC0	IC index
1		$^q IC_{11}$	IC1	IC index
2		$^q IC_{21}$	IC2	IC index
3		$^q IC_{31}$	IC3	IC index
4		$^q IC_{41}$	IC4	IC index
5		$^q IC_{51}$	IC5	IC index
0	2	$^q IC_{02}$	BIC0	Bond IC
1		$^q IC_{12}$	BIC1	Bond IC
2		$^q IC_{22}$	BIC2	Bond IC
3		$^q IC_{32}$	BIC3	Bond IC
4		$^q IC_{42}$	BIC4	Bond IC
5		$^q IC_{52}$	BIC5	Bond IC
0	3	$^q IC_{03}$	CIC0	Complementary IC
1		$^q IC_{13}$	CIC1	Complementary IC
2		$^q IC_{23}$	CIC2	Complementary IC
3		$^q IC_{33}$	CIC3	Complementary IC
4		$^q IC_{43}$	CIC4	Complementary IC
5		$^q IC_{53}$	CIC5	Complementary IC
0	4	$^q IC_{04}$	SIC0	Structural IC
1		$^q IC_{14}$	SIC1	Structural IC
2		$^q IC_{24}$	SIC2	Structural IC
3		$^q IC_{34}$	SIC3	Structural IC
4		$^q IC_{44}$	SIC4	Structural IC
5		$^q IC_{54}$	SIC5	Structural IC
0	5	$^q IC_{05}$	TIC0	Total IC index
1		$^q IC_{15}$	TIC1	Total IC index
2		$^q IC_{25}$	TIC2	Total IC index
3		$^q IC_{35}$	TIC3	Total IC index
4		$^q IC_{45}$	TIC4	Total IC index
5		$^q IC_{55}$	TIC5	Total IC index

[a] Neighborhood Symmetry Information Content (IC)

Markov-Shannon information indices of income-inequality complex networks

We have used Markov chains to calculate Shannon information indices of different systems including simulations of disease spreading relevant to epidemiology [70]. We can define the vector of initial absolute probabilities $^0\mathbf{p} \equiv [^0p_1, {}^0p_2, {}^0p_3..., {}^0p_a..., {}^0p_{nt}]$ for the n$_t$ counties in the same state. We calculate the absolute probability of occurrence of the disease in a given county ^0p_a at the initial time t$_0$ like in other Markov chain models [71-73]:

$$^0p_a = \frac{G_a}{\sum_{a=1}^{nt} G_c} \qquad (6)$$

Here, G$_a$ is the Gini coefficient of income inequality [74] in the ath county of a given state (s) of the U.S. We should to considerate that the only epidemiological factor used as input to calculate the Shannon information indices of the county was the G_a measure of income inequality. G_a measure of income-inequality is widely used as descriptor to approach the study of the epidemiology of different diseases [75, 76]. Haidicha and Ioannidis [77] conclude in their study about Gini coefficient in multicenter clinical studies, that this measure may be routinely incorporated in the description of the characteristics of a clinical study. Using the Chapman-Kolmogorov equation, we can calculate the vector $^k\mathbf{p}^t \equiv [^kp_1, {}^kp_2, {}^kp_3..., {}^kp_a..., {}^kp_{na}]$ for the absolute probabilities ^kp_a along time t_k:

$$^k\mathbf{p} = \left(^1\mathbf{\Pi}\right)^k \cdot {}^0\,\mathbf{p} = {}^k\mathbf{\Pi} \cdot {}^0\,\mathbf{p} \qquad (7)$$

Consequently, elements of the stochastic matrix $^k\mathbf{\Pi}$ = $(^1\mathbf{\Pi})^k$ are the probabilities $^kp_{ab}$ of transmission of AIDS from one county to other in t_k = k years (steps). We calculated the elements of the stochastic matrix $^1\mathbf{\Pi}$ as follow [71-73, 78]:

$$^1p_{ab} = \frac{(G_a + G_b)\cdot \exp(G_b)}{\sum_{c=1}^{c=n}(G_a + G_c)\cdot \exp(G_c)} \qquad (8)$$

Subsequently, we calculated the information indices $I^a{}_k(s)$ to quantify the expected income inequality/epidemiology in the counties and their neighbors using Shannon formula of entropy [70].

$$I^a{}_k(s) = -\left(^k p_a\right)\cdot \log\left(^k p_a\right) \qquad (9)$$

Moving Average (MA) operators

The codification of the chemical structure of the compounds is the first step here. We have data about a large number of assays developed in very different conditions (c_j) for equal or different targets (molecular or not). The non-structural information here refers to different assay conditions (c_j) like concentrations, temperature, targets, organisms, *etc.* A solution may rely upon the use of the idea of Moving Average (MA) operators used in time series analysis with a similar purpose [79]. Many authors have developed Autoregressive Integrated Moving-Average (ARIMA) and other MA models based on the initial work of Box and Jenkins [80]. Langenfeld *et al.* [81] use the ARIMA models to determine the effects of the hypnotic intervention in their study about control of HIV/AIDS-related pain. Gupta *et al.* [82] use moving average analysis for predicting anti-HIV activity using a novel topological descriptor: the eccentric adjacency index, and conclude that the proposed index offers a great potential for structure-activity/property studies. Chen *et al.* [83] studied the timing and magnitude in trend for tuberculosis cases in the United States, using a combination of ARIMA and Bayesian methods and summarize the advantages of this methods for the estimation and interpretation of operational data in public health or other areas. Gupta and Madan [84] studied the development of models for the prediction of HIV integrase inhibitory activity using MA analysis. For instance, Botella-Rocamora *et al.* [85] developed SMARS: Spatial Moving Average Risk Smoothing; a model to map diseases. Two type of parameters $D^q{}_{kj}$ and $<D^q{}_k>_j$ are necessary to calculate a MA. The variable $D^q{}_k$ is one input parameters of type (k) with average values $<D^q{}_k>_j$ for all q^{th} cases measured in a set of experimental conditions (c_j) (not necessarily a molecular descriptor). The general formula of a Box-Jenkins MA operator is:

$$\Delta D^q{}_{kj} = D^q{}_k - \left\langle D^q{}_k \right\rangle_j \qquad (10)$$

ALMA Models

We have developed a similar approach called ALMA (Assessing Links with Moving Averages) using also MA operators (see next section). ALMA models remember those used in ARIMA models of time series analysis . They are adaptable to all molecular descriptors and/or graphs invariants or descriptors for complex networks. In consonance with the previous section, we use a similar terminology. The inputs of one ALMA model are the descriptors $D^q{}_k$ of type k^{th} of the q^{th} system (compound or drug d_q in this case) represented by a matrix **M**. On the other hand, the outputs of one ALMA model are the links (L_{aq} = 1 or L_{aq} = 0) of a complex network with Boolean matrix **L** and formed by different pairs of input systems. Consequently, the general linear equation of the model using a generic descriptor or graph theoretical invariant $D^q{}_k$ has the following general form:

$$\begin{aligned} S_{aqj} &= \sum_{k=1}^{k=k\max} e_k \cdot D^q{}_k + \sum_{k=1}^{k=k\max}\sum_{j=1}^{j=j\max} e_{kj} \cdot \Delta D^q{}_{kj} + e_0 \qquad (11) \\ &= \sum_{k=1}^{k=k\max} e_k \cdot D^q{}_k + \sum_{k=1}^{k=k\max}\sum_{j=1}^{j=j\max} e_{kj} \cdot \left(D^q{}_k - \left\langle D^q{}_k \right\rangle_j \right) + e_0 \end{aligned}$$

The output dependent variable is $S_{aqj} = S_{aq}(c_j) = S_{aqj}(c_l, c_2, c_3, \ldots c_{max})$. This variable is a numerical score of of the formation of links (L_{aq} = 1 or L_{aq} = 0) in the complex network to be predicted. In the particular case of drug-target networks is the score for the biological activity of the q^{th} drug (d_q) vs. the a^{th} target measured in one assay carried out under the set of conditions c_j. We have published different papers with this methodology. Tenorio-Borroto *et al.* [79] studied the quantitative structure-toxicity relationships (QSTR) of drugs with entropy models for multiplex drug-target interaction endpoints. The authors Alonso *et al.* and Luan *et al.* [86, 87] introduced new QSAR models to evaluate the neurotoxicity/neuroprotective properties of drugs for the treatment of neurodegenerative diseases. Others authors have used the same methodology. For instance, Speck-Planche and Cordeiro [88-90] have reported different multi-target models using the same type of ALMA approach with molecular descriptors $D^q{}_k$ of different types. They developed a multitasking QSAR model for the simultaneous prediction of anti-streptococci activity and toxic effects of drugs [88]. They developed multi-target approaches for prediction of drugs in different classes of cancer [89-92]. They also proposed ALMA models of multitarget inhibitors against different proteins associated with Alzheimer [93] and Tuberculosis [94]. Last, these authors introduced models for inhibitors for C-C chemokine receptors using sub-structural descriptors [95]

ALMA models for AIDS/HIV multiscale studies

In a recent work [96] , we constructed the first ALMA model for AIDS/HIV multiscale studies of U.S. at county level. We used as inputs of the model the Balaban information indices ($I^q{}_k$) of a given compound d_q and the Shannon information indices for the population (ath county). We obtained the data about anti-HIV drugs activity from the database ChEMBL and we used the molecular smiles codes and the Balaban information indices to quantify information. We used Shannon information indices [49] to describe the information of the social network (income inequality characterized by Gini coefficient) [77]. The data about the AIDS prevalence and Gini coefficient at county level were obtained from the AIDSVu database. The ALMA parameters as inputs of ANNs trained/validated with 43,249 data points. The dataset included values of AIDS prevalence in 2310 U.S. counties *vs.* ChEMBL results for 21,582 unique drugs, 9 viral or human protein targets, 4856 protocols, and 10 possible experimental measures. We trained different topologies of ANNs including Multilayer Perceptrons (MLPs) and Linear Neural Networks (LNNs). The LNN with Accuracy (Ac), Specificity (Sp), and Sensitivity (Sn) above 75% was the best model found. In **Figure 4**, we show the workflow used in the present work to construct ALMA models for this problem.

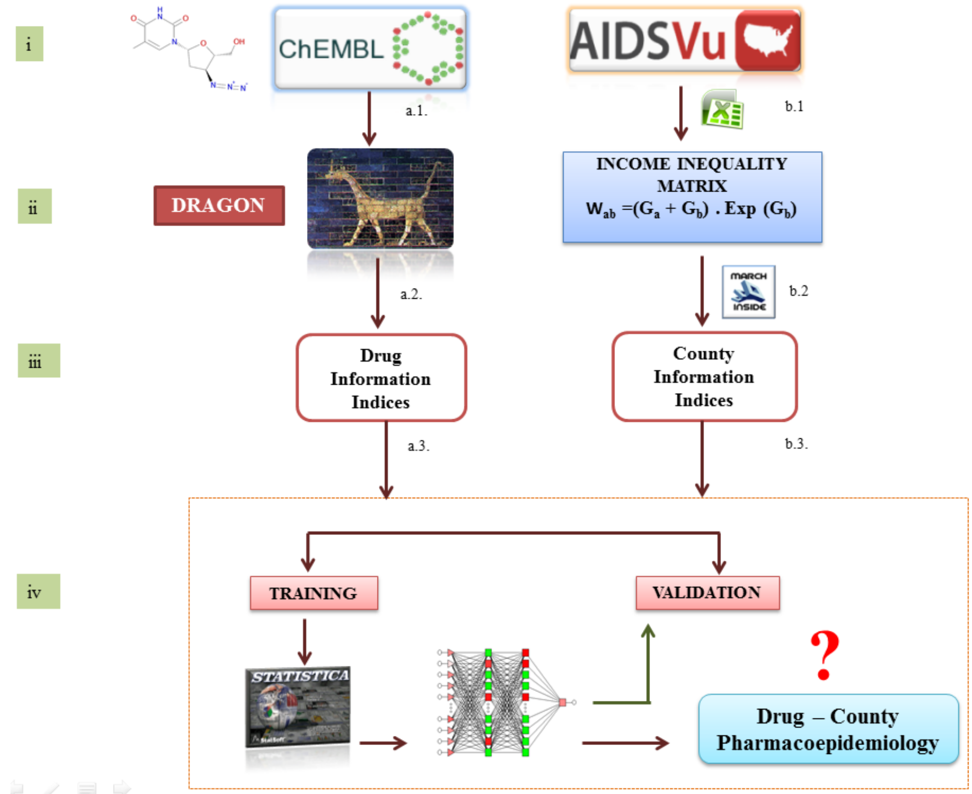


Fig. (4). Flowchart to construct the ANNs for the AIDS Pharmacoepidemiology model in U.S.

This type of models predict the formation of links (L_{aq}= 1) or not (L_{aq}= 0) in a complex network of AIDS pharmacoepidemiology in U.S. In the present context, we can use MA of properties of nodes of networks (drugs, proteins, organisms, counties, *etc.*) to predict the variable $L_{aq}(c_j)_{obs}$ in specific sub-set of conditions (c_j). This variable quantifies the formation of links between nodes. There are two different types of nodes forming this specific network. The first one represents the U.S. counties (a) and the second type of node characterizes the drugs (d_q). The value is $L_{aq}(c_j)$ = 1 when the Drug-Disease Ratio = $DDR_{aq}(c_j)$ > cutoff and $L_{aq}(c_j)_{obs}$ = 0 otherwise. In our previous work [96], we defined the ratio as follow $DDR_{aq}(c_j)$ = [$D_q(c_j/D_a)$]. The term D_a is the AIDS prevalence rate for the ath county and $D_q(c_j)$ is the biological activity of the qth drug assayed in the conditions c_j. The general formula for a linear model developed using information indices was [96]:

$$
\begin{aligned}
S_{aqj} = & \sum_{k=1}^{k=4} e_k \cdot I^q{}_k + \sum_{k=1}^{k=4}\sum_{j=1}^{j=4} e_{kj} \cdot \Delta I^q{}_{kj} \qquad (12) \\
& + e_{ak} \cdot I^a{}_k(t) + e_0 \\
& = \sum_{k=1}^{k=4} e_k \cdot I^q{}_k + \sum_{k=1}^{k=4}\sum_{j=1}^{j=4} e_{kj} \cdot \left(I^q{}_k - \left\langle I^q{}_k \right\rangle_j \right) \\
& + e_{ak} \cdot I^a{}_k(t) + e_0
\end{aligned}
$$

The reader should note that the predicted, output, or dependent variable S_{aqj} is not a discrete variable but a real-valued numerical score. However, the variable is directly proportional to the observed variable (L_{aq}). In general, c_j refers to different boundary conditions for the assay, *e.g.*, targets, cellular lines, organisms, experimental measures, *etc.* Therefore, c_1 = is the experimental measure of activity, c_2 = is the protein target, c_3 = is the organism that express the target, and c_4 = is the assay protocol *per se*. In order to seek the coefficients of the model we can use linear classification technique like ANNs implemented in the software package STASTITICA 6.0 [97]. The statistical parameters used to support the model were: Number of cases in training (N), and overall values of Sp, Sn, and Ac [98]. In the works of this series, we used specifically the type of descriptors called information indices (D = I). In any case, to calculate the MA we have to sum the values of $I^q{}_k$ or all the n_j drugs with assay conditions c_j. Next, we divide this sum by the number of compounds n_j with this condition.

$$\Delta I^q{}_{kj} = I^q{}_k - \left\langle I^q{}_k \right\rangle_j \qquad (13)$$

$$\left\langle I^q{}_k \right\rangle_j = \frac{1}{n_j} \sum_{q=1}^{q=n_j} I^q{}_k \qquad (14)$$

NEW ALMA MODEL FOR AIDS/HIV MULTISCALE STUDIES

In the present paper, we changed the Balaban information indices ($I^q{}_k$) by Symmetry information content indices (${}^qIC_{kf}$). However, we used the $I^a{}_k(s)$ indices to characterize the different populations. The new ALMA model developed using these other set of indices has the following general form:

$$\begin{aligned} S_{aqj} = & \sum_{k=0}^{k=5}\sum_{f=1}^{f=5} e_{kf} \cdot {}^q IC_{kf} \qquad (15) \\ & + \sum_{k=0}^{k=5}\sum_{f=1}^{f=5}\sum_{j=1}^{j=4} e_{kfj} \cdot \Delta^q IC_{kfj} \\ & + \sum_{k=1}^{k=5} e_{ak} \cdot \Delta I^a{}_{ks} + e_0 \\ = & \sum_{k=0}^{k=5}\sum_{f=1}^{f=5} e_{kf} \cdot {}^q IC_{kf} \\ & + \sum_{k=0}^{k=5}\sum_{f=1}^{f=5}\sum_{j=1}^{j=4} e_{kfj} \cdot \left({}^q IC_{kf} - \left\langle {}^q IC_{kf} \right\rangle_j \right) \\ & + \sum_{k=1}^{k=5} e_k \cdot \left(I^a{}_k - \left\langle I^a{}_k \right\rangle_s \right) + e_0 \end{aligned}$$

We used the software DRAGON [63] to calculate the ${}^qIC_{kf}$ indices for the molecules of the same dataset of anti-HIV drugs obtained from ChEMBL in the previous work [96]. In this case we calculated a total of $N_{indices} = N_k \cdot N_f = 6 \cdot 5 = 30$ values of ${}^qIC_{kf}$ indices with $N_k = 6$ different orders (k) that belong to $N_f = 5$ different families of descriptors (f). The families studied are the same reported in previous sections (see **Table 2**). We developed different ANN models using all the set of parameters as well as simple models using different sub-sets of descriptors. At follow we discuss some of the more relevant results found.

ANN models with all descriptors.

We obtained the ANN models using as input all descriptors. We used in total 30 ${}^qIC_{kf}$ indices of the molecules, 120 MA operators Δ^qIC_{kfj} for the different assay conditions for drugs (c_1,c_2,c_3,c_4), and 5 MA operators $\Delta I^a{}_{ks}$ for U.S. counties. It makes a total of $N_{inputs} = N_k \cdot N_f + N_k \cdot N_f \cdot N_j + {}^aN_k = 6 \cdot 5 + 6 \cdot 5 \cdot 4 + 5 = 30 + 120 + 5 = 155$ input values. In this counting formula N_k, N_f, N_j, and ${}^aN_k$ are the number of orders of molecular descriptors, families of descriptors, boundary conditions, and orders of county indices. The results obtained using the software STATISTICA show that the MLP [99] method fails to generate good prediction models. It presents values of Sp, Sn, and AUROC close to 50%. This values are typical of a random classifier and not the expected performance for a significant model [98]. Conversely, the LNN predictor based on the 155 descriptors classifies correctly above 76% of the cases in training and external validation sets (see **Table 3**).

This model presented values of Sn = 76.13 and Sp = 76.51 in training and Sn = 77.04 and Sp = 76.48 in external validation sets. The LNN network shows values of AUROC = 0.82 in training and 0.82 for external validation set. We can conclude that the linear models seem to be better than non-linear to fit the present dataset. However, the number of inputs is very high to be considered a simple model.

LNN models for each family of information indices

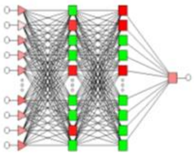
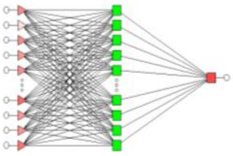
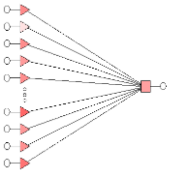
We decided to train LNN classifiers with each family of indices taking into consideration that the previous LNN model presented a very large number of inputs (155 in total) with respect to the performance obtained. In this sense, we trained ANN predictors for each family of neighborhood symmetry indices separately. It makes a total of $N_{inputs} = N_k \cdot N_f + N_k \cdot N_f \cdot N_j + {}^aN_k = 6 \cdot 1 + 6 \cdot 1 \cdot 4 + 5 = 6 + 24 + 5 = 35$ input values at least in each model for one specific family. All LNN models obtained (five in total) classified correctly above 75% of the cases and AUROC above 0.80 in train and validation series regardless of the family of indices used (see **Table 4**). This represents a spectacular 5-fold reduction of the number of parameters from 155 to above 30 parameters in all models. Interestingly, all models presented a very similar performance. It could be due to the co-linearity between the different information indices of drugs. A visual inspection of the formulae used to calculate the indices shows that all use as input the IC_k index. In fact, this index is the more relevant in because it is included in the formulas for calculate the remaining indices. We studied the correlation matrix of 435 pairs of information indices for the Anti-HIV drugs studied in this work. We found that 43% of these indices, *i.e.*, 187 pairs have a significant correlation $p < 0.05$. In addition, about 15% of the pairs of indices have a strong correlation $\geq \pm 0.7$.

LNN model with information index of five orders

After analysis of the previous results, we decided to test the predictive power of these indices in a simpler model. In so doing, we trained the LNN predictors using only each family of information indices of drugs (${}^qIC_{5f}$) of 5- order, their MA operators (Δ^qIC_{5fj})) and the fifth MA operator of the U.S. counties ($\Delta I^a{}_{5s}$). The LNN model based on ${}^qIC_{51}$ (LNN-IC_{51}) presented the higher values of Sn = 72.04/72.81 and Sp = 72.38/72.50 in training/ and external validation sets (see **Table 5**). LNN-IC_{51} presented also the higher values for the AUROC in train and validation series (0.73 and 0.74 respectively). Analyzing all the previous results for this dataset, we found that the IC$_k$ index appears to be the most important to predict the drug structure-activity relationships. We can conclude it by comparison to the other indices, which have lower values of classification. The equation of LNN-IC$_{51}$ this model is the following:

$$\begin{aligned}S_{aq}(c_j) = & -25.48 \cdot {}^{q}IC_{51} + 1081.64 \cdot \Delta^{q}IC_{51}(c_1) + 29.36 \cdot \Delta^{q}IC_{51}(c_2) \\ & -1084.52 \cdot \Delta^{q}IC_{51}(c_3) - 0.7727 \cdot \Delta^{q}IC_{51}(c_4) \\ & -0.0792 \cdot \Delta I^{a}{}_{5}(s) - 0.5025\end{aligned} \qquad (16)$$

Table 3. ANN classifiers based on neighbourhood symmetry information indices

ANN models	Observed	$L_{pq} = 1$	$L_{pq} = 0$	$L_{pq} = 1$	$L_{pq} = 0$	AUROC
		Training		**Validation**		
MLP2 155:155-12-29-1:1 	Parameter [a]	Sn	Sp	Sn	Sp	(T / V)
	Predicted	52.44	52.68	51.11	51.95	0.53 / 0.52
	$L_{pq} = 1$	6009	9845	1948	3334	
	$L_{pq} = 0$	5449	10961	1863	3605	
MLP1 155:155-18-1:1 	Parameter [a]	Sn	Sp	Sn	Sp	(T / V)
	Predicted	52.70	52.74	51.27	51.92	0.53 / 0.52
	$L_{pq} = 1$	6039	9831	1954	3336	
	$L_{pq} = 0$	5419	10975	1857	3603	
LNN 155:155-1:1 	Parameter [a]	Sn	Sp	Sn	Sp	(T / V)
	Predicted	76.13	76.51	77.04	76.48	0.82 / 0.82
	$L_{pq} = 1$	8723	4887	2936	1632	
	$L_{pq} = 0$	2735	15919	875	5307	

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

Table 4. LNN classifiers for the families of symmetry information indices from 0 to 5 order

Type of Index	LNN Profile	Observed	$L_{pq} = 1$	$L_{pq} = 0$	$L_{pq} = 1$	$L_{pq} = 0$	AUROC
			Training		**Validation**		
$^{q}IC_{k1}$	31:31-1:1	Parameter [a]	Sn	Sp	Sn	Sp	(T / V)
		Predicted	75.88	76.17	76.59	76.27	0.81 / 0.81
		$L_{pq} = 1$	8695	4958	2919	1646	
		$L_{pq} = 0$	2763	15848	892	5293	
$^{q}IC_{k2}$	28:28-1:1	Parameter [a]	Sn	Sp	Sn	Sp	(T / V)
		Predicted	75.85	76.19	76.75	76.33	0.81 / 0.81
		$L_{pq} = 1$	8691	4952	2925	1642	
		$L_{pq} = 0$	2767	15854	886	5297	
$^{q}IC_{k3}$	29:29-1:1	Parameter [a]	Sn	Sp	Sn	Sp	(T / V)
		Predicted	75.80	76.16	76.62	76.35	0.81 / 0.81
		$L_{pq} = 1$	8686	4959	2920	1641	
		$L_{pq} = 0$	2772	15847	891	5298	
$^{q}IC_{k4}$	30:30-1:1	Parameter [a]	Sn	Sp	Sn	Sp	(T / V)
		Predicted	75.87	76.18	76.69	76.27	0.80/ 0.81
		$L_{pq} = 1$	8694	4954	2923	1646	
		$L_{pq} = 0$	2764	15852	888	5293	
$^{q}IC_{k5}$	32:32-1:1	Parameter [a]	Sn	Sp	Sn	Sp	(T / V)
		Predicted	74.84	75.31	75.67	75.25	0.78 / 0.79
		$L_{pq} = 1$	8576	5135	2884	1717	
		$L_{pq} = 0$	2882	15671	927	5222	

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

The Sensitivity analysis of this model, allowed us to quantify (rank) and order (ratio) into a sequence, the different chemoinformatics *vs.* pharmacoepidemiology inputs, as you can see in the **Figure 5** and **Table 6**. In the **Figure 5,** we can see that the models give more relevance to the information about the molecular structure, parameters of type $^{q}IC_{kf}$. In second place the models rank the experimental measure used to measure the effectiveness of the drug $\Delta^{q}IC_{kf}(c_1)$. The third type of input is the organism $\Delta^{q}IC_{kf}(c_3)$, the fourth type of input in importance is the protein $\Delta^{q}IC_{kf}(c_2)$. The lasts ones are the information about the assay $\Delta^{q}IC_{kf}(c_4)$ and income inequality in the counties ΔI^{a}_{ks}. Whereas the LNN model with the $^{q}IC_{51}$ index gives a higher relevance to the information about the experimental measure $\Delta^{q}IC_{kf}(c_1)$, and in second place the model ranks information about the organism $\Delta^{q}IC_{kf}(c_3)$ used to measure the biological activity. Thus, the sensitivity analysis shows that the chosen model ranks the importance of factors in the following order (AIDS epidemiology / anti-HIV drug) ≈ organism in preclinical assay > experimental measure of activity > drug target > chemical structure of the drug > pharmacological assay> county income inequality.

Table 5. LNN classifiers for each family of symmetry information indices of 5-order

Type of Index	Observed	L_{pq} = 1	L_{pq} = 0	L_{pq} = 1	L_{pq} = 0	AUROC
		Training		**Validation**		
$^{q}IC_{51}$	Parameter [a]	Sn	Sp	Sn	Sp	(T / V)
	Predicted	72.04	72.38	72.81	72.50	0.73 / 0.73
	L_{pq} = 1	8255	5746	2775	1908	
	L_{pq} = 0	3203	15060	1036	5031	
$^{q}IC_{52}$	Parameter [a]	Sn	Sp	Sn	Sp	(T / V)
	Predicted	69.36	69.81	70.21	69.95	0.73 / 0.74
	L_{pq} = 1	7948	6281	2676	2085	
	L_{pq} = 0	3510	14525	1135	4854	
$^{q}IC_{53}$	Parameter [a]	Sn	Sp	Sn	Sp	(T / V)
	Predicted	66.51	66.94	67.04	66.73	0.70/ 0.70
	L_{pq} = 1	7621	6878	2555	2308	
	L_{pq} = 0	3837	13928	1256	4631	
$^{q}IC_{54}$	Parameter [a]	Sn	Sp	Sn	Sp	(T / V)
	Predicted	67.69	68.03	68.56	67.99	0.73 / 0.73
	L_{pq} = 1	7757	6651	2613	2221	
	L_{pq} = 0	3701	14155	1198	4718	
$^{q}IC_{55}$	Parameter [a]	Sn	Sp	Sn	Sp	(T / V)
	Predicted	54.90	54.80	55.81	55.16	0.52 / 0.51
	L_{pq} = 1	5167	11403	1684	3828	
	L_{pq} = 0	6291	9403	2127	3111	

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

Table 6. Parameters used for preprocessing data, carry out model predictions, and sensitivity analysis

Data Operation	[a] Parameter / Effect	Drug	Measure	Protein	Organism	Assay	County
		$^{q}IC_{51}$	$\Delta\,^{q}IC_{51}(c_1)$	$\Delta\,^{q}IC_{51}(c_2)$	$\Delta\,^{q}IC_{51}(c_3)$	$\Delta\,^{q}IC_{51}(c_4)$	$\Delta I^{a}_{5}(s)$
Preprocessing	Missing data value	0.9976678	0.9975952	0.9965166	0.9976395	0.9512190	0.5069762
	Linear shift parameter	0.0009935	0.0009935	0.0009937	0.0009935	0.0009945	0.0060280
	Scale coefficient	0.9925405	0.9975564	0.9964782	0.9976007	0.9512009	0.5070032
Model	LNN coefficient	-25.48	1081.64	29.36	-1084.52	-0.7727	0.0792
Sensitivity	Train ratio	1.047460	13.26668	1.062159	13.30156	1.000043	1.000075
	Train rank	4	2	3	1	6	5
	Validation ratio	1.139811	22.88760	1.172580	22.96505	1.000254	0.999974
	Validation rank	4	2	3	1	5	6

[a] Parameters used to transform data; LNN coefficients are the coefficients of the variables in the model, the independent term of the model is e_0 = -0.5026 (see equation in the text)

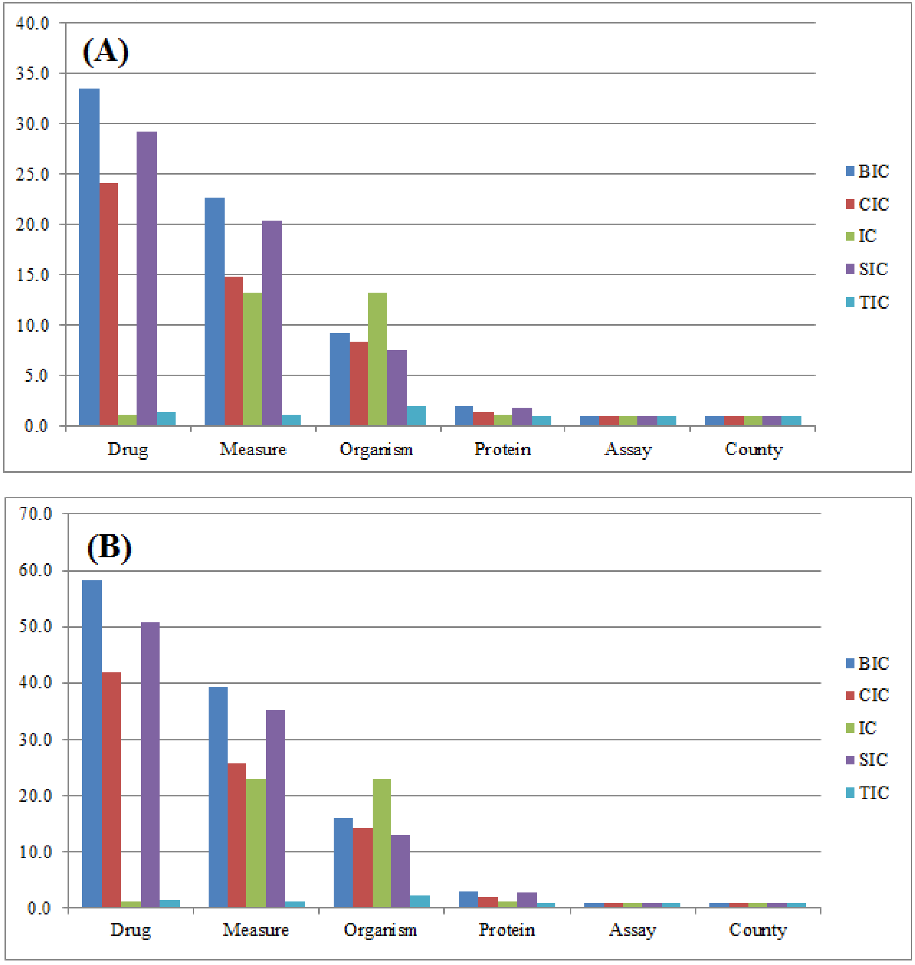


Fig. (5). Sensitivity analysis of LNN 6:6-1:1 classifiers. (A) training and (B) validation.

In order to use this model in the future, the $^{q}IC_{51}$ information indices of the molecules, the average values of $^{q}IC_{51j}$ with the different boundary conditions, and the information of the counties in the U.S. (ΔI^{a}_{5s}) are in the **Table SM1**, **Table SM2,** and **Table SM3** respectively. In **Table 7**, we illustrate some examples of values of $^{q}IC_{51}$ for drugs and ΔI^{a}_{5s} for counties of different states. In **Table 8,** we show some examples of average values of $^{q}IC_{51}$ information descriptors of molecular structure for different boundary conditions.

Use of the LNN-IC$_5$ model to construct the AIDS complex network

Last, we used this LNN-ALMA model to generate/predict a complex network of the prevalence of AIDS in the United States at county level with respect to the preclinical activity of anti-HIV drugs. The bipartite network has two types of nodes (counties *vs.* drug). Thus, this is a multiscale networksimilar to bipartite networks of drugs *vs.* target proteins reported by other groups [100-104]. However, the nodes in the present network contain information about the molecules, i.e., chemical structure as well as assay conditions (target protein, organism, experimental measure, etc.). Additionally, the other set of nodes contain information about socioeconomic factors, such as the income inequality in the county.

Multiscale networks of this type have been discussed by Barabasi *et al.* [105] as one of the more important tools to perform trans-disciplinary research. The links of this complex network are the outputs $L_{aq}(c_j)_{pred} = 1$ of our model. We studied 43,249 data points to fit the model and predict the complex network. Consequently, we have to add the values of AIDS prevalence in 2310 counties in U.S. *vs.* ChEMBL results for 21,582 unique drugs, 9 viral or human protein targets, 4856 protocols, and 10 possible experimental measures. In **Figure 6**, we illustrate the sub-network of AIDS prevalence *vs.* Anti-HIV drug preclinical activity for the state of Florida. For instance, the model predicts a high effectively for the drug Zidovudine [106] to treat AIDS in Nassau County. In **Table 9** we include some examples of antiretroviral drugs with observed $L_{aq}(c_j)_{obs}$ and predicted $L_{aq}(c_j)_{pred}$ effects over AIDS prevalence in several counties of the same state in U.S.

Table 7. Some values of IC^q_5 for drugs and $\Delta I^a_5(s)$ for counties of different states.

Name	ID	Target	Organism	${}^{q}IC_{51}$
Delavirdine	593	HIV-1	HIV-1	5.169
Zidovudine	129	HIV-1	HIV-1	4.726
Entecavir	713	HIV	HIV	4.901
Lopinavir	729	HIV-1	HIV-1	5.629
Lamivudine	1230	HIV-1	HIV-1	5.114
Ciprofloxacin	8	HIV-1	HIV-1	4.678
Apigenin	28	HIV-1	HIV-1	4.64
Stavudine	991	HIV-1 RT	HIV-1	4.566
Hept	31871	HIV-1 RT	HIV-1	4.703
Quercetin	50	HIV-1 IN	HIV-1	5
Loviride	37624	HIV-1 PR	HIV-1	4.785
Vicriviroc	82301	CC-CKR-5	hsa	5.241
Dipyridyl	39879	CC-CKR-5	hsa	3.322
Plerixafor	18442	CXCR-4	hsa	4.647
Disulfiram	964	CC-CKR-2	hsa	2.642
County name	State (s)	D_a [i]	G_a [ii]	$\Delta I^a_5(s)$
Autauga County	AL	181	0.405	9.1376
Arkansas County	AR	165	0.467	-1.4848
Apache County	AZ	124	0.488	-5.8417
Alameda County	CA	396	0.456	13.1466
Adams County	CO	179	0.403	2.1596
Fairfield County	CT	375	0.537	0.0
Kent County	DE	240	0.406	0.0
Baker County	FL	380	0.429	0.1173
Atkinson County	GA	256	0.447	-4.0347
Honolulu County	HI	201	0.422	0.0
Boone County	IA	58	0.407	-16.9929
Bannock County	ID	100	0.429	5.6071
Adams County	IL	65	0.453	-2.1575
Allen County	IN	136	0.428	2.738
Allen County	KS	44	0.394	0
Allen County	KY	71	0.42	-3.4678
Anderson County	KY	76	0.376	14.6872
Acadia Parish	LA	174	0.452	-4.7955
Ascension Parish	LA	178	0.409	9.3625
Berkshire County	MA	102	0.462	18.2985
Allegany County	MD	180	0.446	-38.5628
Calvert County	MD	124	0.369	12.8543
Hancock County	ME	73	0.437	0.2091

[i] D_a is the AIDS prevalence rate in the county p in 2010.

[ii] G_a is the Gini income-inequality measure of US county in 2010

Table 8. Average values of information descriptors of molecular structure for different boundary conditions

c_1	Experimental measure	N(c_1)	<^{q}IC$_{51}$>
IC$_{50}$(nM)	Inhibitory concentration 50%	20332	5.049
EC$_{50}$(nM)	Effective concentration 50%	14981	5.139
K$_i$(nM)	Inhibitory constant	3736	5.422
IC$_{95}$(nM)	Inhibitory concentration 95%	1290	5.368
IC$_{90}$(nM)	Inhibitory concentration 90%	1118	5.022
ED$_{50}$(nM)	Effective dose 50%	860	5.036
EC$_{50}$(ug.mL^{-1})	Effective concentration	526	5.109
IC$_{50}$(ug.mL^{-1})	Inhibitory concentration	335	5.111
EC$_{90}$(nM)	Effective concentration	67	4.763
c_2	Target protein	N(c_2)	
CC-CKR-5	C-C chemokine receptor type 5	2304	5.383
CC-CKR-2	C-C chemokine receptor type 2	2009	4.867
CC-CKR-3	C-C chemokine receptor type 3	1206	5.397
CC-CKR-4	C-C chemokine receptor type 4	345	5.256
CXCR-4	C-X-C chemokine receptor type 4	332	5.789
HIV-1 RT	HIV-1 reverse transcriptase	4029	4.922
HIV-1 IN	HIV-1 integrase	1702	3.769
HIV-1 PR	HIV-1 protease	5946	5.495
GP160	Envelope polyprotein GP160	34	4.952
c_3	Organism	N(c_3)	
HIV-1	HIV-1	34544	5.093
mmu	Mus musculus	68	5.536
hsa	Homo sapiens	6128	5.230
HIV-2	HIV-2	1030	5.330
HIV	HIV	1479	5.169
c_4	Assay	N(c_4)	
1033994	Antiviral activity against HIV1	282	4.874
708445	Effective conc. required for the inhibition of HIV-1 IIIB in MT-4 cells	176	5.835
859312	Inhibitory activity was determined against HIV type 1 protease	175	5.942
659084	Inhibitory conc. for displacement of [125I]-MIP-1 alpha from recombinant human CCR5 expressed in CHO cell	141	5.789
974332	Displacement of [125I]MIP1alpha from human CCR5 expressed in CHO cells	109	5.090

Table 9. Examples of antiretroviral drugs with observed $L_{aq}(c_j)_{obs}$ and predicted $L_{aq}(c_j)_{pred}$ effects.

Compound ID	$L_{aq}(c_j)_{obs}$	$L_{aq}(c_j)_{pred}$	c-level	Compound Name	Target	Organism	Assay ID	Measure	State, County
57	1	1	0.60	Nevirapine	HIV-1 RT	HIV-1	804261	EC_{50}(nM)	VA, Covington city
57	1	1	0.60	Nevirapine	HIV-1 RT	HIV-1	804184	EC_{50}(nM)	VA, Culpeper
114	1	1	0.57	Saquinavir	HIV-1	HIV-1	833275	EC_{50}(nM)	KY, Jackson
114	1	1	0.57	Saquinavir	HIV-1	HIV-1	833117	EC_{50}(nM)	KY, Montgomery
115	1	1	0.57	Indinavir	HIV-1	HIV-1	935845	EC_{50}(nM)	KY, Pulaski
115	1	1	0.58	Indinavir	HIV-1	HIV-1	952983	EC_{50}(nM)	KY, Rowan
116	1	1	0.58	Amprenavir	HIV-1	HIV-1	935841	EC_{50}(nM)	MA, Berkshire
116	1	1	0.58	Amprenavir	HIV-1	HIV-1	935840	EC_{50}(nM)	MA, Hampshire
116	1	1	0.28	Amprenavir	HIV-1 PR	HIV-1	828911	K_i(nM)	CA, Lake
116	1	1	0.28	Amprenavir	HIV-1 PR	HIV-1	830292	K_i(nM)	GA, Union
116	1	1	0.29	Amprenavir	HIV-1 PR	HIV-1	909837	K_i(nM)	IN, Clinton
116	1	1	0.28	Amprenavir	HIV-1 PR	HIV-1	909838	K_i(nM)	IN, Hancock
129	1	1	0.61	Zidovudine	HIV-1	HIV-1	1640111	EC_{50}(nM)	TN, Sevier
129	1	1	0.60	Zidovudine	HIV-1	HIV-1	1640126	EC_{50}(nM)	TN, Smith
57	1	0	0.70	Nevirapine	HIV-1	HIV-1	695511	IC_{50}(nM)	MS, Walthall
57	1	0	0.70	Nevirapine	HIV-1	HIV-1	695519	IC_{50}(nM)	MS, Wayne
38380	0	0	0.85	Fasudil	CC-CKR-2	hsa	915971	IC_{50}(nM)	GA, Baldwin
1185005	0	0	0.85	Cenicriviroc	CC-CKR-2	hsa	2184889	IC_{50}(nM)	MO, Laclede
39879	0	0	0.84	Dipyridyl	CC-CKR-5	hsa	2215079	IC_{50}(nM)	MS, Alcorn
82301	0	0	0.84	Vicriviroc	CC-CKR-5	hsa	1697613	IC_{50}(nM)	NJ, Essex
82301	0	0	0.84	Vicriviroc	CC-CKR-5	hsa	1697612	IC_{50}(nM)	NJ, Middlesex
1172035	0	0	0.84	Nifeviroc	CC-CKR-5	hsa	1174016	IC_{50}(nM)	MO, Laclede
1172035	0	0	0.84	Nifeviroc	CC-CKR-5	hsa	1174015	IC_{50}(nM)	MO, Macon

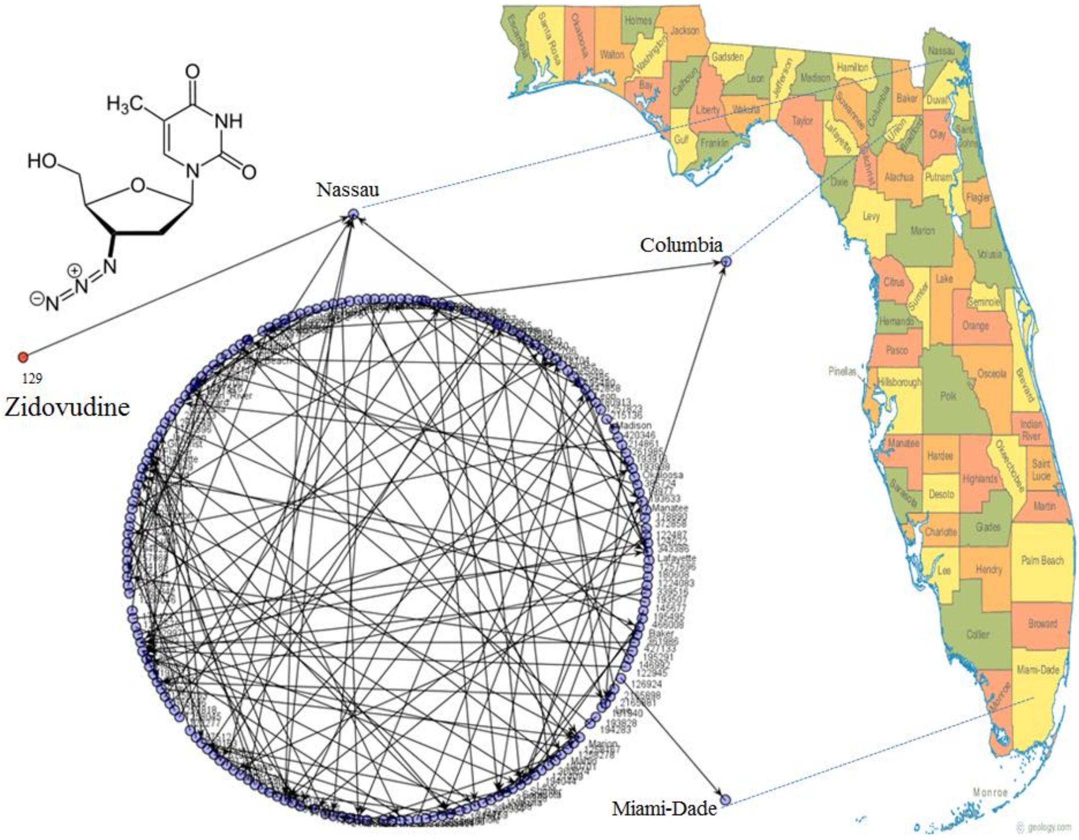


Figure 6. Sub-network of AIDS prevalence *vs.* Anti-HIV drug activity for U.S. state of Florida (FL)

CONCLUSIONS

This work presents a review of several aspects of the disease, including the epidemiology, pathophysiology, treatments, etc. We also developed a model called LNN-ALMA to generate complex networks of the prevalence of AIDS in the counties of the U.S. with respect to the preclinical activity of anti-HIV drugs. The best classifier found was the LNN-$IC_{51;}$ this classifier has only six inputs based on neighborhood information content indices, compared to the other models, the IC_k index seems to be the most important to predict the drug structure-activity relationships. The new model has similar performance but is notably simpler than a previous model based on Balaban's information indices with >20 inputs. In future work, we will continue to improve the models and we will include other information indices, socioeconomic factors, machine-learning techniques, etc.

ACKNOWLEDGEMENTS

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] Lindemann D, Steffen I, Pohlmann S. Cellular entry of retroviruses. Adv Exp Med Biol 2013; 790: 128-49.

[2] Rinaldo CR. HIV-1 Infection of CD4 T Cells by Professional Antigen Presenting Cells. Scientifica (Cairo) 2013; 2013: 164203.

[3] Noorizadeh H, Sajjadifar S, Farmany A. A quantitative structure-activity relationship study of anti-HIV activity of substituted HEPT using nonlinear models. Med Chem Res 2013; 22: 5442-52.

[4] HIV surveillance--United States, 1981-2008. MMWR Morb Mortal Wkly Rep 2011; 60: 689-93.

[5] Moss JA. HIV/AIDS Review. Radiol Technol 2013; 84: 247-67; quiz p.68-70.

[6] Piot P, Quinn TC. Response to the AIDS pandemic--a global health model. N Engl J Med 2013; 368: 2210-8.

[7] Heikamp K, Bajorath J. Large-scale similarity search profiling of ChEMBL compound data sets. J Chem Inf Model 2011; 51: 1831-9.

[8] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012; 40: D1100-7.

[9] Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, et al. The ChEMBL bioactivity database: an update. Nucleic Acids Res 2013.

[10] Alkhatib G. The biology of CCR5 and CXCR4. Curr Opin HIV AIDS 2009; 4: 96-103.

[11] Wilkin TJ, Gulick RM. CCR5 antagonism in HIV infection: current concepts and future opportunities. Annu Rev Med 2012; 63: 81-93.

[12] Blanpain C, Libert F, Vassart G, Parmentier M. CCR5 and HIV infection. Receptors Channels 2002; 8: 19-31.

[13] Tan JH, Ludeman JP, Wedderburn J, Canals M, Hall P, Butler SJ, et al. 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 2013; 288: 10024-34.

[14] Hu WS, Hughes SH. HIV-1 reverse transcription. Cold Spring Harb Perspect Med 2012; 2: pii: a006882.

[15] Karmon SL, Markowitz M. Next-generation integrase inhibitors : where to after raltegravir? Drugs 2013; 73: 213-28.

[16] Qiu X, Liu ZP. Recent developments of peptidomimetic HIV-1 protease inhibitors. Curr Med Chem 2011; 18: 4513-37.

[17] Castro HC, Abreu PA, Geraldo RB, Martins RC, dos Santos R, Loureiro NI, et al. Looking at the proteases from a simple perspective. J Mol Recognit 2011; 24: 165-81.

[18] Wu B, Chien EY, Mol CD, Fenalti G, Liu W, Katritch V, et al. Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. Science 2010; 330: 1066-71.

[19] Tan Q, Zhu Y, Li J, Chen Z, Han GW, Kufareva I, et al. Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex. Science 2013; 341: 1387-90.

[20] Lee WG, Gallardo-Macias R, Frey KM, Spasov KA, Bollini M, Anderson KS, et al. Picomolar Inhibitors of HIV Reverse Transcriptase Featuring Bicyclic Replacement of a Cyanovinylphenyl Group. J Am Chem Soc 2013; 135: 16705-13.

[21] Wielens J, Headey SJ, Rhodes DI, Mulder RJ, Dolezal O, Deadman JJ, et al. Parallel screening of low molecular weight fragment libraries: do differences in methodology affect hit identification? J Biomol Screen 2013; 18: 147-59.

[22] Zhang H, Wang YF, Shen CH, Agniswamy J, Rao KV, Xu CX, et al. Novel P2 tris-tetrahydrofuran group in antiviral compound 1 (GRL-0519) fills the S2 binding pocket of selected mutants of HIV-1 protease. J Med Chem 2013; 56: 1074-83.

[23] Qian K, Morris-Natschke SL, Lee KH. HIV entry inhibitors and their potential in HIV therapy. Med Res Rev 2009; 29: 369-93.

[24] Perno CF. The discovery and development of HIV therapy: the new challenges. Ann Ist Super Sanita 2011; 47: 41-3.

[25] Saag MS. New and investigational antiretroviral drugs for HIV infection: mechanisms of action and early research findings. Top Antivir Med 2012; 20: 162-7.

[26] Arts EJ, Hazuda DJ. HIV-1 antiretroviral drug therapy. Cold Spring Harb Perspect Med 2012; 2: a007161.

[27] Powderly WG. Integrase inhibitors in the treatment of HIV-1 infection. J Antimicrob Chemother 2010; 65: 2485-8.

[28] Adams JL, Greener BN, Kashuba AD. Pharmacology of HIV integrase inhibitors. Curr Opin HIV AIDS 2012; 7: 390-400.

[29] Chougrani I, Luton D, Matheron S, Mandelbrot L, Azria E. In: HIV AIDS (Auckl). V: 2013; 5: pp. 253-62.

[30] Guha R. On exploring structure-activity relationships. Methods Mol Biol 2013; 993: 81-94.

[31] Burbidge R, Trotter M, Buxton B, Holden S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. Comput Chem 2001; 26: 5-14.

[32] Manallack DT, Livingstone DJ. Neural networks in drug discovery: have they lived up to their promise? Eur J Med Chem 1999; 34: 195-208.

[33] Si H, Yuan S, Zhang K, Fu A, Duan Y-B, Zhide H. Quantitative structure activity relationship study on EC50 of anti-HIV drugs. Chemometr Intell Lab 2008; 90: 15-24.

[34] Magalhaes Ude O, Souza AM, Albuquerque MG, Brito MA, Bello ML, Cabral LM, et al. Hologram quantitative structure-activity relationship and comparative molecular field analysis studies within a series of tricyclic phthalimide HIV-1 integrase inhibitors. Drug Des Devel Ther 2013; 7: 953-61.

[35] Saranya N, Selvaraj S. QSAR studies on HIV-1 protease inhibitors using non-linearly transformed descriptors. Curr Comput Aided Drug Des 2012; 8: 10-49.

[36] Debnath AK. Application of 3D-QSAR techniques in anti-HIV-1 drug design--an overview. Curr Pharm Des 2005; 11: 3091-110.

[37] Debnath U, Verma S, Jain S, Katti SB, Prabhakar YS. Pyridones as NNRTIs against HIV-1 mutants: 3D-QSAR and protein informatics. J Comput Aided Mol Des 2013; 27: 637-54.

[38] Sun XH, Guan JQ, Tan JJ, Liu C, Wang CX. 3D-QSAR studies of quinoline ring derivatives as HIV-1 integrase inhibitors. SAR QSAR Environ Res 2012; 23: 683-703.

[39] Swiderek K, Marti S, Moliner V. Theoretical studies of HIV-1 reverse transcriptase inhibition. Phys Chem Chem Phys 2012; 14: 12614-24.

[40] Hu Y, Bajorath J. Molecular scaffolds with high propensity to form multi-target activity cliffs. J Chem Inf Model 2010; 50: 500-10.

[41] Erhan D, L'Heureux P J, Yue SY, Bengio Y. Collaborative filtering on a family of biological targets. J Chem Inf Model 2006; 46: 626-35.

[42] Namasivayam V, Hu Y, Balfer J, Bajorath J. Classification of compounds with distinct or overlapping multi-target activities and diverse molecular mechanisms using emerging chemical patterns. J Chem Inf Model 2013; 53: 1272-81.

[43] Cruz-Monteagudo M, Cordeiro MN, Tejera E, Dominguez ER, Borges F. Desirability-based multi-objective QSAR in drug discovery. Mini Rev Med Chem 2012; 12: 920-35.

[44] Machado A, Tejera E, Cruz-Monteagudo M, Rebelo I. 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 2009; 44: 5045-54.

[45] Saiz-Urra L, Bustillo Perez AJ, Cruz-Monteagudo M, Pinedo-Rivilla C, Aleu J, Hernandez-Galan R, et al. Global antifungal profile optimization of chlorophenyl derivatives against Botrytis cinerea and Colletotrichum gloeosporioides. J Agric Food Chem 2009; 57: 4838-43.

[46] Cruz-Monteagudo M, Borges F, Cordeiro MN, Cagide Fajin JL, Morell C, Ruiz RM, et al. Desirability-based methods of multiobjective optimization and ranking for global QSAR studies. Filtering safe and potent drug candidates from combinatorial libraries. J Comb Chem 2008; 10: 897-913.

[47] Nicolaou CA, Brown N, Pattichis CS. Molecular optimization using computational multi-objective methods. Curr Opin Drug Discov Devel 2007; 10: 316-24.

[48] Shannon CE. A Mathematical Theory of Communication. Bell Syst Tech J 1948; 27: 379-423.

[49] Godden JW, Stahura FL, Bajorath J. Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. J Chem Inf Comput Sci 2000; 40: 796-800.

[50] Stahura FL, Godden JW, Bajorath J. Differential Shannon entropy analysis identifies molecular property descriptors that predict aqueous solubility of synthetic compounds with high accuracy in binary QSAR calculations. J Chem Inf Comput Sci 2002; 42: 550-8.

[51] Stahura FL, Godden JW, Xue L, Bajorath J. Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations. J Chem Inf Comput Sci 2000; 40: 1245-52.

[52] Acharya UR, Faust O, Sree V, Swapna G, Martis RJ, Kadri NA, et al. Linear and nonlinear analysis of normal and CAD-affected heart rate signals. Comput Methods Programs Biomed 2014; 113: 55-68.

[53] Roy K, Saha A. Comparative QSPR studies with molecular connectivity, molecular negentropy and TAU indicesPart I: Molecular thermochemical properties of diverse functional acyclic compounds. J Mol Model (Online) 2003; 9: 259-70.

[54] Agrawal VK, Khadikar PV. Modelling of carbonic anhydrase inhibitory activity of sulfonamides using molecular negentropy. Bioorg Med Chem Lett 2003; 13: 447-53.

[55] Katritzky AR, Lomaka A, Petrukhin R, Jain R, Karelson M, Visser AE, et al. QSPR correlation of the melting point for pyridinium bromides, potential ionic liquids. J Chem Inf Comput Sci 2002; 42: 71-4.

[56] Katritzky AR, Perumal S, Petrukhin R, Kleinpeter E. Codessa-based theoretical QSPR model for hydantoin HPLC-RT lipophilicities. J Chem Inf Comput Sci 2001; 41: 569-74.

[57] Graham DJ, Schacht D. Base Information Content in Organic Molecular Formulae. J Chem Inf Comput Sci 2000; 40: 942.

[58] Graham DJ. Information Content in Organic Molecules: Structure Considerations Based on Integer Statistics. J Chem Inf Comput Sci 2002; 42: 215.

[59] Graham DJ, Malarkey C, Schulmerich MV. Information Content in Organic Molecules: Quantification and Statistical Structure via Brownian Processing. . J Chem Inf Comput Sci 2004; 44.

[60] Graham DJ, Schulmerich MV. Information Content in Organic Molecules: Reaction Pathway Analysis via Brownian Processing. J Chem Inf Comput Sci 2004; 44.

[61] Graham DJ. Information Content and Organic Molecules: Aggregation States and Solvent Effects. J Chem Inf Model 2005; 45.

[62] Graham DJ. Information Content in Organic Molecules: Brownian Processing at Low Levels. J Chem Inf Model 2007; 47: 376-89.

[63] Todeschini R, Consonni V. Handbook of Molecular Descriptors: Weinheim, Germany 2000.

[64] Sharma BK, Pilania P, Singh P, Prabhakar YS. Combinatorial protocol in multiple linear regression/partial least-squares directed rationale for the caspase-3 inhibition activity of isoquinoline-1,3,4-trione derivatives. SAR QSAR Environ Res 2010; 21: 169-85.

[65] Singh P, Kumar R, Sharma BK, Prabhakar YS. Topological descriptors in modeling malonyl coenzyme A decarboxylase inhibitory activity: N-Alkyl-N-(1,1,1,3,3,3-hexafluoro-2-hydroxypropylphenyl)amide derivatives. J Enzyme Inhib Med Chem 2009; 24: 77-85.

[66] Prabhakar YS, Rawal RK, Gupta MK, Solomon VR, Katti SB. Topological descriptors in modeling the HIV inhibitory activity of 2-aryl-3-pyridyl-thiazolidin-4-ones. Comb Chem High Throughput Screen 2005; 8: 431-7.

[67] Singh P, Shekhawat N. Chemometric descriptors in the rationale of antimalarial activity of natural and synthetic prodiginines. J Curr Chem Pharm Sc 2012; 2: 244-60.

[68] Magnuson VR, Harriss DK, Basak SC. In: Studies in Physical and Theoretical Chemistry; King, RB. Elsevier. V: Amsterdam (The Netherlands) 1983; pp. 178-91.

[69] Todeschini R, Consonni V, Mauri A, Pavan M. DRAGON [computer program]. Milano, Italy: Talete srl; 2005.

[70] Riera-Fernandez P, Munteanu CR, Escobar M, Prado-Prado F, Martin-Romalde R, Pereira D, et al. 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 2012; 293: 174-88.

[71] Gonzalez-Diaz H, Aguero G, Cabrera MA, Molina R, Santana L, Uriarte E, et al. Unified Markov thermodynamics based on stochastic forms to classify drugs considering molecular structure, partition system, and biological species: distribution of the antimicrobial G1 on rat tissues. Bioorg Med Chem Lett 2005; 15: 551-7.

[72] Gonzalez-Diaz H, Cruz-Monteagudo M, Molina R, Tenorio E, Uriarte E. Predicting multiple drugs side effects with a general drug-target interaction thermodynamic Markov model. Bioorg Med Chem 2005; 13: 1119-29.

[73] Van Waterbeemd H. In: Chemometric methods in molecular design. VCH. V: New York, NY 1995; 2: pp. 265-82.

[74] Pabayo R, Kawachi I, Gilman SE. Income inequality among American states and the incidence of major depression. J Epidemiol Community Health 2013.

[75] Burns JK, Tomita A, Kapadia AS. Income inequality and schizophrenia: Increased schizophrenia incidence in countries with high levels of income inequality. Int J Soc Psychiatry 2013.

[76] Green C, Yu BN, Marrie RA. Exploring the implications of small-area variation in the incidence of multiple sclerosis. Am J Epidemiol 2013; 178: 1059-66.

[77] Haidich AB, Ioannidis JP. The Gini coefficient as a measure for understanding accrual inequalities in multicenter clinical studies. J Clin Epidemiol 2004; 57: 341-8.

[78] González-Díaz H, Prado-Prado FJ, Santana L, Uriarte E. Unify QSAR approach to antimicrobials. Part 1: Predicting antifungal activity against different species. Bioorg Med Chem 2006; 14 5973–80.

[79] Tenorio-Borroto E, Garcia-Mera X, Penuelas-Rivas CG, Vasquez-Chagoyan JC, Prado-Prado FJ, Castanedo N, et al. Entropy model for multiplex drug-target interaction endpoints of drug immunotoxicity. Curr Top Med Chem 2013; 13: 1636-49.

[80] Box GEP, Jenkins GM. Time series analysis: Forecasting and control. Holden-Day: San Francisco, California 1970.

[81] Langenfeld MC, Cipani E, Borckardt JJ. Hypnosis for the control of HIV/AIDS-related pain. Int J Clin Exp Hypn 2002; 50: 170-88.

[82] Gupta S, Singh M, Madan AK. Predicting anti-HIV activity: computational approach using a novel topological descriptor. J Comput Aided Mol Des 2001; 15: 671-8.

[83] Chen MP, Shang N, Winston CA, Becerra JE. A Bayesian analysis of the 2009 decline in tuberculosis morbidity in the United States. Stat Med 2012; 31: 3278-84.

[84] Gupta M, Madan AK. Diverse models for the prediction of HIV integrase inhibitory activity of substituted quinolone carboxylic acids. Arch Pharm (Weinheim) 2012; 345: 989-1000.

[85] Botella-Rocamora P, Lopez-Quilez A, Martinez-Beneito MA. Spatial moving average risk smoothing. Stat Med 2013; 32: 2595-612.

[86] Alonso N, Caamano O, Romero-Duran FJ, Luan F, Dias Soeiro Cordeiro MN, Yanez M, et al. Model for High-Throughput Screening of Multi-Target Drugs in Chemical Neurosciences; Synthesis, Assay and Theoretic Study of Rasagiline Carbamates. ACS Chem Neurosci 2013.

[87] Luan F, Cordeiro MN, Alonso N, Garcia-Mera X, Caamano O, Romero-Duran FJ, et al. TOPS-MODE model of multiplexing neuroprotective effects of drugs and experimental-theoretic study of new 1,3-rasagiline derivatives potentially useful in neurodegenerative diseases. Bioorg Med Chem 2013; 21: 1870-9.

[88] Speck-Planche A, Kleandrova VV, Cordeiro MN. Chemoinformatics for rational discovery of safe antibacterial drugs: simultaneous predictions of biological activity against streptococci and toxicological profiles in laboratory animals. Bioorg Med Chem 2013; 21: 2727-32.

[89] Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Chemoinformatics in multi-target drug discovery for anti-cancer therapy: in silico design of potent and versatile anti-brain tumor agents. Anticancer Agents Med Chem 2012; 12: 678-85.

[90] Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Chemoinformatics in anti-cancer chemotherapy: multi-target QSAR model for the in silico discovery of anti-breast cancer agents. Eur J Pharm Sci 2012; 47: 273-9.

[91] Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Rational drug design for anti-cancer chemotherapy: multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents. Bioorg Med Chem 2012; 20: 4848-55.

[92] Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Multi-target drug discovery in anti-cancer therapy: fragment-based approach toward the design of potent and versatile anti-prostate cancer agents. Bioorg Med Chem 2011; 19: 6239-44.

[93] Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Multi-target inhibitors for proteins associated with Alzheimer: in silico discovery using fragment-based descriptors. Curr Alzheimer Res 2013; 10: 117-24.

[94] Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. In silico discovery and virtual screening of multi-target inhibitors for proteins in Mycobacterium tuberculosis. Comb Chem High Throughput Screen 2012; 15: 666-73.

[95] Speck-Planche A, Kleandrova VV. In silico design of multi-target inhibitors for C-C chemokine receptors using substructural descriptors. Mol Divers 2012; 16: 183-91.

[96] González-Díaz H, Herrera-Ibatá DM, Duardo-Sanchez A, Munteanu CR, Orbegozo-Medina RA, Pazos A. 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 2014; Submitted.

[97] StatSoft.Inc. STATISTICA (data analysis software system), version 6.0, www.statsoft.com.Statsoft, Inc. 6.0 ed2002.

[98] Hill T, Lewicki P. STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining: Tulsa 2006

[99] Rosenblatt F. Principles of neurodynamics; perceptrons and the theory of brain mechanisms: Washington D.C. 1962.

[100] Hecker N, Ahmed J, von Eichborn J, Dunkel M, Macha K, Eckert A, et al. SuperTarget goes quantitative: update on drug-target interactions. Nucleic Acids Res 2012; 40: D1113-7.

[101] Prado-Prado F, Garcia-Mera X, Escobar M, Alonso N, Caamano O, Yanez M, et al. 3D MI-DRAGON: new model for the reconstruction of US FDA drug- target network and theoretical-experimental studies of inhibitors of rasagiline derivatives for AChE. Curr Top Med Chem 2012; 12: 1843-65.

[102] Prado-Prado F, Garcia-Mera X, Abeijon P, Alonso N, Caamano O, Yanez M, et al. Using entropy of drug and protein graphs to predict FDA drug-target network: theoretic-experimental study of MAO inhibitors and hemoglobin peptides from Fasciola hepatica. Eur J Med Chem 2011; 46: 1074-94.

[103] Vina D, Uriarte E, Orallo F, Gonzalez-Diaz H. Alignment-free prediction of a drug-target complex network based on parameters of drug connectivity and protein sequence of receptors. Mol Pharm 2009; 6: 825-35.

[104] Araujo RP, Liotta LA, Petricoin EF. Proteins, drug targets and the mechanisms they control: the simple truth about complex networks. Nat Rev Drug Discov 2007; 6: 871-80.

[105] Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet 2011; 12: 56-68.

[106] Shey MS, Kongnyuy EJ, Alobwede SM, Wiysonge CS. Co-formulated abacavir-lamivudine-zidovudine for initial treatment of HIV infection and AIDS. Cochrane Database Syst Rev 2013; 3: CD005481.

7.3 Publicación 3

Herrera-Ibatá DM, Pazos A, Orbegozo-Medina RA, González-Díaz H. Mapping networks of anti-HIV drug cocktails vs. AIDS epidemiology in US Counties. *Chemometrics and Intelligent Laboratory Systems*, 138 (2014) 161-170.

Chemometrics and Intelligent Laboratory Systems 138 (2014) 161–170

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

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

ARTICLE INFO

Article history:
Received 26 February 2014
Received in revised form 3 July 2014
Accepted 11 August 2014
Available online 20 August 2014

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

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.

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 [4,5]. 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

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; η_i, Atom eccentricity; σ_i,, i^{th} vertex distance degree.

* Corresponding author.

** Correspondence to: H. González-Díaz, Department of Organic Chemistry II, Faculty of Science and Technology, University of the Basque Country UPV/EHU, 48940, Leioa, Spain.

E-mail addresses: diana.herrera@udc.es (D.M. Herrera-Ibatá), humberto.gonzalezdiaz@ehu.es (H. González-Díaz).

http://dx.doi.org/10.1016/j.chemolab.2014.08.006

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 [14,15]. 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 [17,18].

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 [28–30]. 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 [34–37] 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 [38–40]. Multi-objective optimization problem (MOOP) [41–45] 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/) [46–48], 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) [53,54], and neighborhood symmetry indices (IC_k) [52,55]. 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, η_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 σ_i, which is the ith 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 [52,56–61].

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].

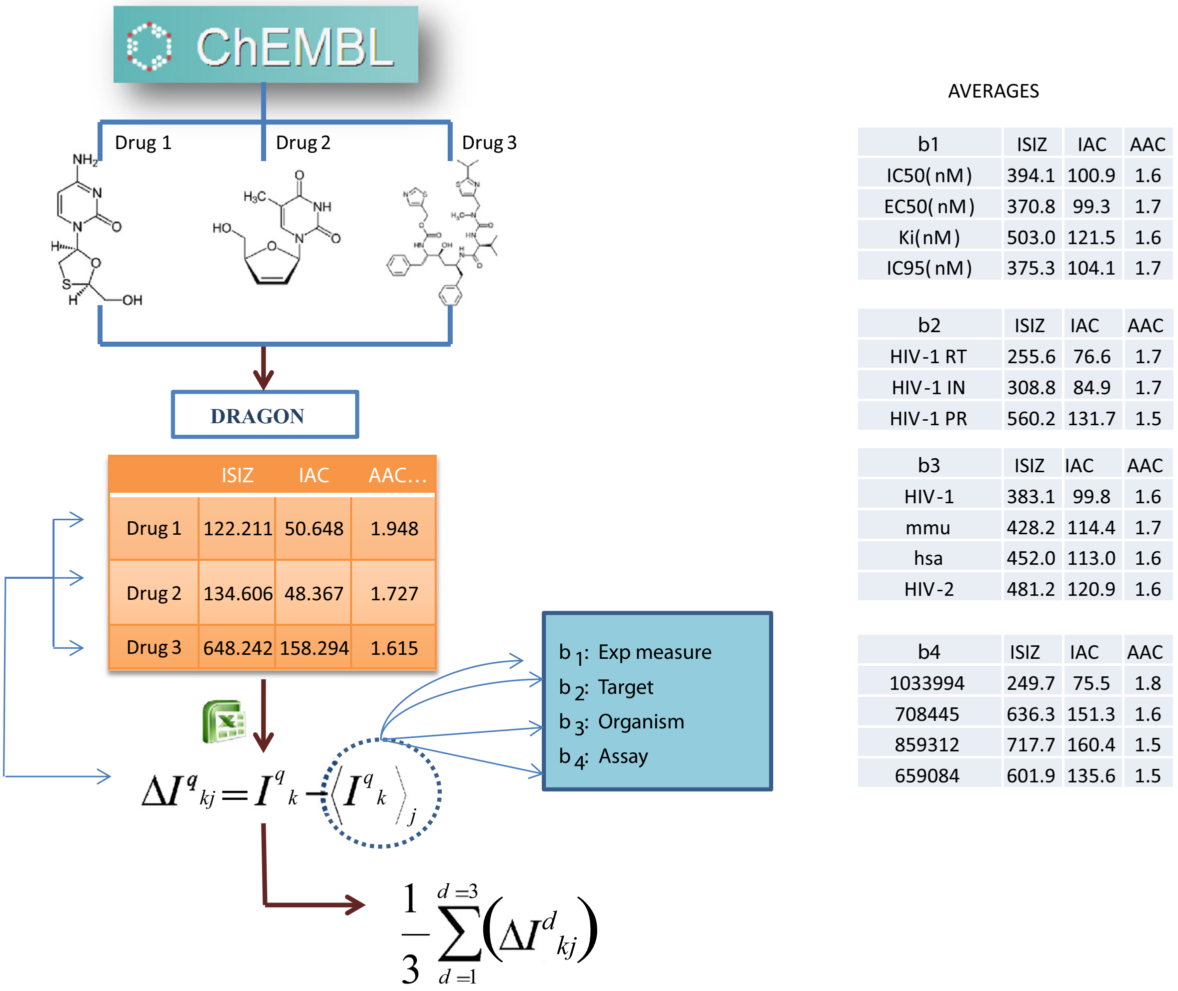
We used as inputs of the model the MI_k of a given drug (dth) and the Shannon information indices ($I^a{}_0$) for the population, i.e., the US County (ath). This model may predict the formation of links ($L_{ac} = 1$) or not ($L_{ac} = 0$) in a complex network of AIDS pharmacoepidemiology in the US. 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 (ath) and the second type of node characterizes the drugs (dth). The value is $L_{ac}(b_j)_{obs} = 1$ when the cocktail–disease ratio $= CDR_{ac}(b_j) > \text{cutoff} = 0.001$ 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 (dth) present in the cocktail assayed in the sets of conditions (b_j).The term D_a is the AIDS prevalence rate for the county (ath). 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 δ_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:

$$\begin{aligned}
S_{ac} &= \sum_{k=1}^{k=13} e_k \cdot \left(\frac{1}{3}\sum_{d=1}^{d=3} I^d{}_k\right) + \sum_{k=1}^{k=13}\sum_{j=1}^{j=4} e_{kj} \cdot \left[\frac{1}{3}\sum_{d=1}^{d=3}\left(\Delta I^d{}_{kj}\right)\right] + e_a \cdot I^a{}_0 + e_0 \\
&= \sum_{k=1}^{k=4} e_k \cdot \left(\frac{1}{3}\sum_{d=1}^{d=3} I^d{}_k\right) + \sum_{j=1}^{j=4} e_{kj} \cdot \left[\frac{1}{3}\sum_{d=1}^{d=3}\left(I^d{}_k - \left\langle I^d{}_k\right\rangle_j\right)\right] + e_a \cdot I^a{}_0 + e_0 \qquad (12) \\
&= \sum_{k=d=1}^{k=13,c,d=3} {}'e_k \cdot I^d{}_k + \sum_{k=j=d=1}^{k=13,j=4,d=3} {}'e_{kj} \cdot \left(I^d{}_k - \left\langle I^d{}_k\right\rangle_j\right) + e_a \cdot I^a{}_0 + e_0
\end{aligned}$$

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

Table 1
Names, symbols, and formula for the calculation of different MI_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	$ISIZ = nAT.\ \log_2 nAT$	[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	$HVcpx = \frac{1}{nSK} \cdot \sum_{i=1}^{nSK}\left(-\sum_{g=0}^{\eta_i} {}^{g}f\,nSK.\ \log_2 {}^{g}f\,nSK\right)$	[60]
I_{hdcpx}	HDcpx	Graph distance complexity index	$HDcpx = \sum_{i=1}^{nSK} \frac{\sigma_i}{2W} \cdot \left(-\sum_{j=1}^{nSK} \frac{d_{ij}}{\sigma_i}.\ \log_2 \frac{d_{ij}}{\sigma_i}\right)$	[60,61]



	ISIZ	IAC	AAC...
Drug 1	122.211	50.648	1.948
Drug 2	134.606	48.367	1.727
Drug 3	648.242	158.294	1.615

$$\Delta I^{q}{}_{kj} = I^{q}{}_{k} - \left\langle I^{q}{}_{k} \right\rangle_{j}$$

$$\frac{1}{3}\sum_{d=1}^{d=3}\left(\Delta I^{d}{}_{kj}\right)$$

AVERAGES

b1	ISIZ	IAC	AAC
IC50(nM)	394.1	100.9	1.6
EC50(nM)	370.8	99.3	1.7
Ki(nM)	503.0	121.5	1.6
IC95(nM)	375.3	104.1	1.7

b2	ISIZ	IAC	AAC
HIV-1 RT	255.6	76.6	1.7
HIV-1 IN	308.8	84.9	1.7
HIV-1 PR	560.2	131.7	1.5

b3	ISIZ	IAC	AAC
HIV-1	383.1	99.8	1.6
mmu	428.2	114.4	1.7
hsa	452.0	113.0	1.6
HIV-2	481.2	120.9	1.6

b4	ISIZ	IAC	AAC
1033994	249.7	75.5	1.8
708445	636.3	151.3	1.6
859312	717.7	160.4	1.5
659084	601.9	135.6	1.5

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

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 Tables SM1, SM2, and SM3 of the supplementary material, respectively.

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

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

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].

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

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,

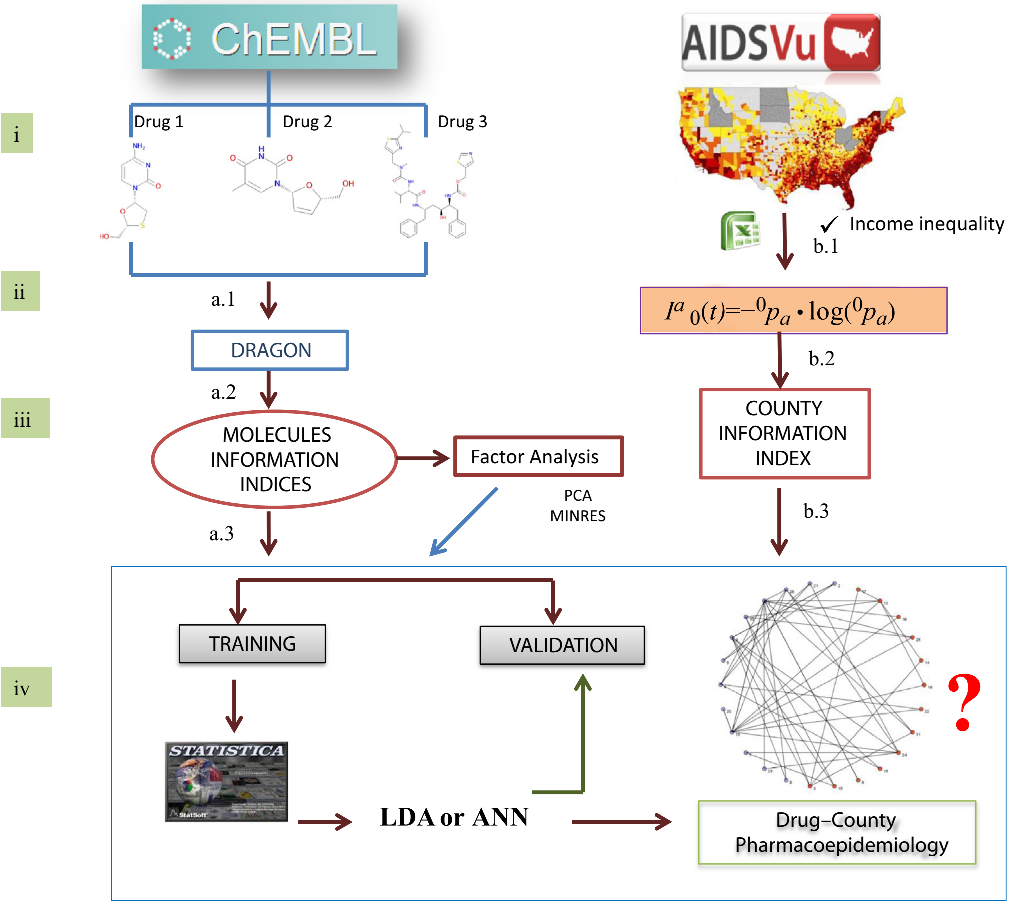


Fig. 2. Flowchart to construct the ML methods for the AIDS pharmacoepidemiology model in the US.

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

Table 2
Machine learning classifiers based on MI_k information indices.

Models	Model		Training		Selection		Validation	
	profile [a]	Observed	$L_{ac} = 0$	$L_{ac} = 1$	$L_{ac} = 0$	$L_{ac} = 1$	$L_{ac} = 0$	$L_{ac} = 1$
LDA	66-23-1	Parameter [a]	Sn	Sp	Sn	Sp	Sn	Sp
		Predicted	83.64	77.15	–	–	83.69	77.37
		$L_{ac} = 0$	67971	2356	–	–	45183	1599
		$L_{ac} = 1$	13292	7959	–	–	8801	5467
MLP	66-26-1	Parameter[a]	Sn	Sp	Sn	Sp	Sn	Sp
		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
LDA-MLP	19-10-1	Parameter [a]	Sn	Sp	Sn	Sp	Sn	Sp
		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
LNN	66-1	Parameter [a]	Sn	Sp	Sn	Sp	Sn	Sp
		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
PCA-LDA	8-7-1	Parameter [a]	Sn	Sp	Sn	Sp	Sn	Sp
		Predicted	50.98	70.66	–	–	50.94	70.93
		$L_{ac} = 0$	41434	3026	–	–	27504	2054
		$L_{ac} = 1$	39829	7289	–	–	26480	5012
MINRES-LDA	8-5-1	Parameter [a]	Sn	Sp	Sn	Sp	Sn	Sp
		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
ΔAAC(c_1)	−97.34	MA for AAC of drugs with the same experimental measure
ΔIDE(c_1)	−1111.82	MA for IDE of drugs with the same experimental measure
ΔIDM(c_1)	−955.75	MA for IDM of drugs with the same experimental measure
ΔIVDM(c_1)	1472.82	MA for IVDM of drugs with the same experimental measure
ΔHVcpx(c_1)	1028.83	MA for HVcpx of drugs with the same experimental measure
ΔHDcpx(c_1)	3011.83	MA for HDcpx of drugs with the same experimental measure
ΔHVcpx(c_2)	0.45	MA for HDcpx of drugs with the same protein
ΔAAC(c_3)	23.96	MA for AAC of drugs with the same organism
ΔIDE(c_3)	−519.39	MA for IDE of drugs with the same organism
ΔIDM(c_3)	954.21	MA for IDM of drugs with the same organism
ΔIVDM(c_3)	−1901.33	MA for IVDM of drugs with the same organism
ΔHDcpx(c_3)	955.72	MA for HDcpx of drugs with the same organism
ΔHDcpx(c_3)	−2256.14	MA for HDcpx of drugs with the same organism
ΔAAC(c_4)	−1.46	MA for AAC of drugs with the same assay protocol
ΔIDE(c_4)	−8.43	MA for IDE of drugs with the same assay protocol
ΔIVDE(c_4)	1.28	MA for IVDE of drugs with the same assay protocol
ΔHVcpx(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

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 [67,68] software package.

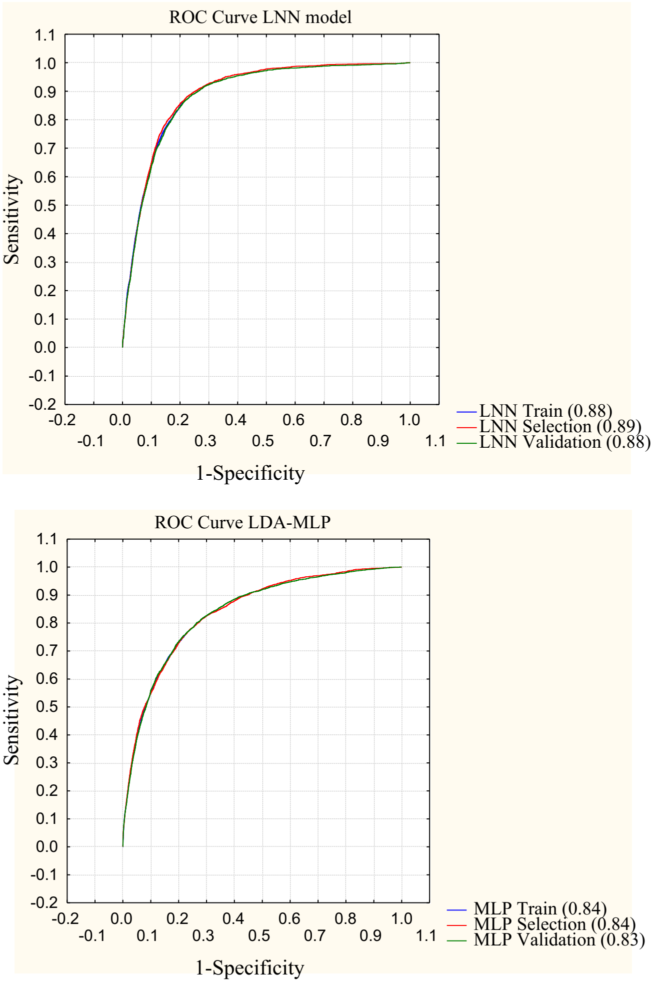


Fig. 3. AUROC curve values for the ANNs.

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

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 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.

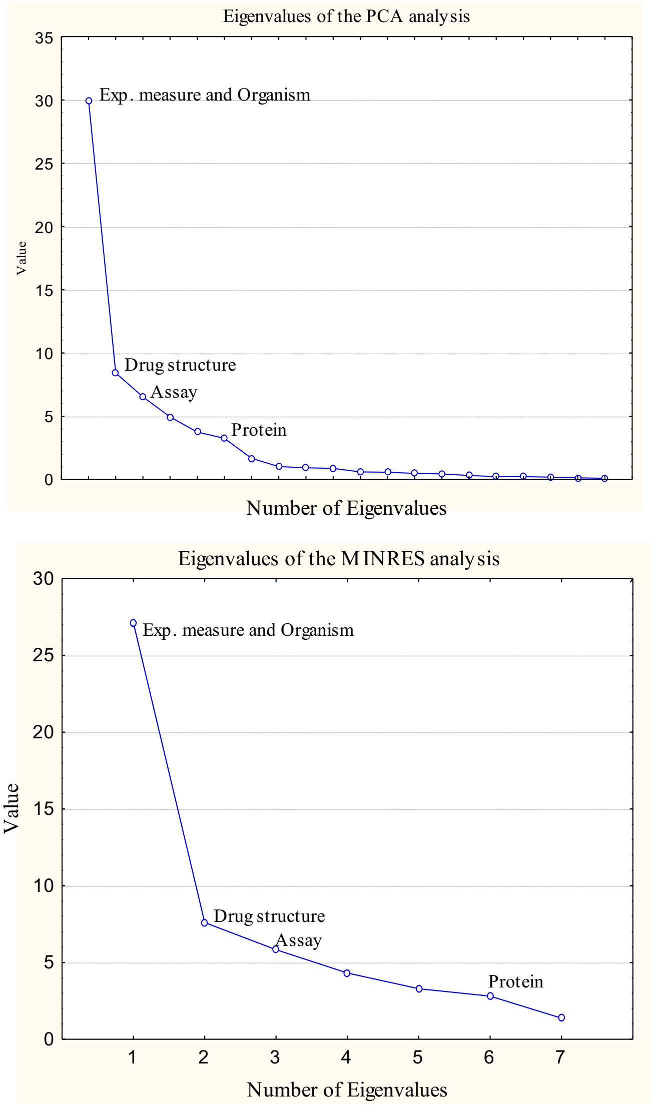


Fig. 4. Plot of bio-molecular eigenvalues for PCA and MINRES.

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.

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 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

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).

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,

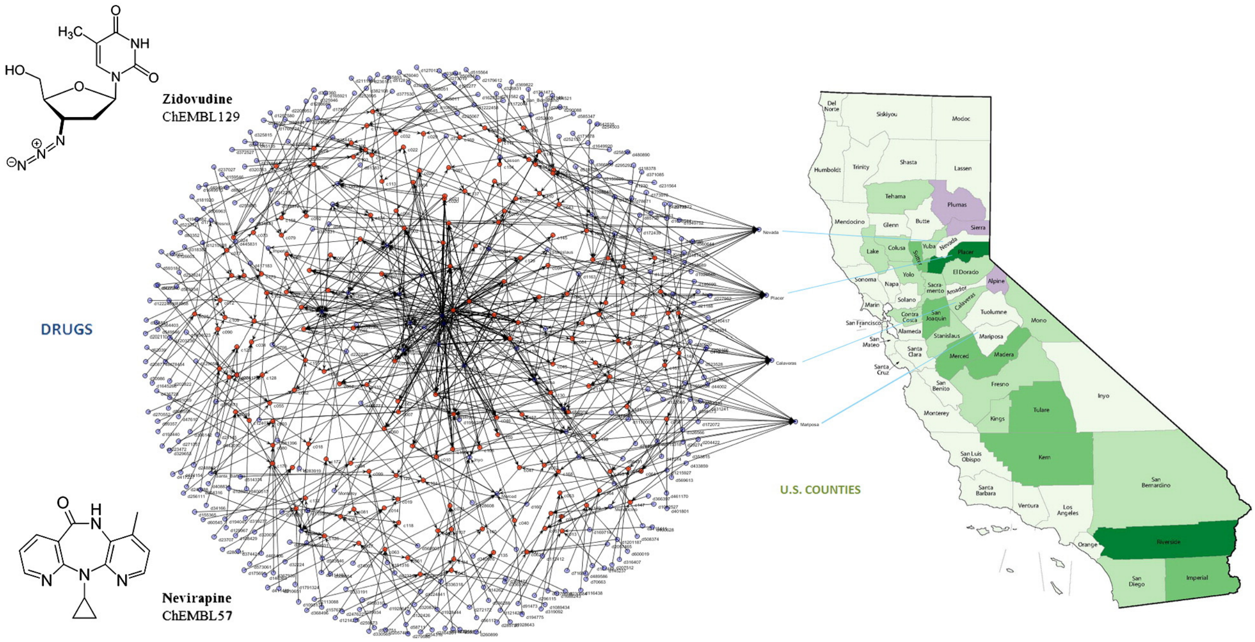


Fig. 5. Sub-network of anti-HIV drug cocktails vs. AIDS prevalence for the US state of California (CA).

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.

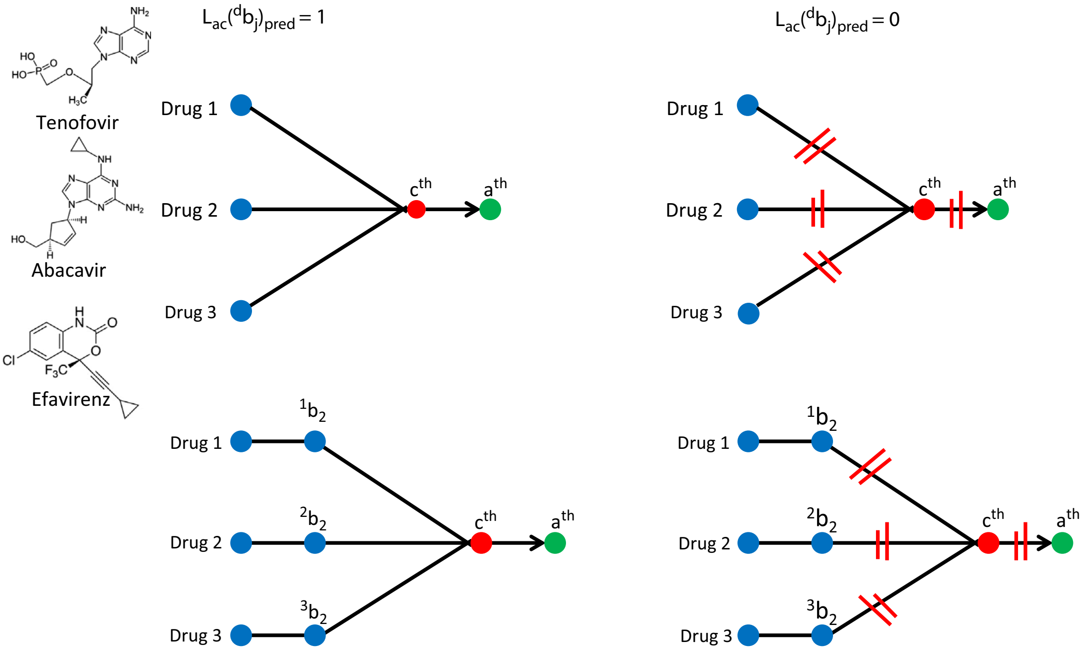


Fig. 6. Sub-network node connection $L_{ac}({}^{d}b_j)_{pred} = 1$ and non-connection $L_{ac}({}^{d}b_j)_{pred} = 0$. The b_2 represents the drug target, c^{th} represents the drug cocktails, a^{th} represents the US counties.

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).

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.

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.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.chemolab.2014.08.006.

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) 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) 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) 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) 165–181.

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

[7] C. Blanpain, F. Libert, G. Vassart, M. Parmentier, CCR5 and HIV infection, Recept. Channels 8 (2002) 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) 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) 369–393.

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

[11] C.F. Perno, The discovery and development of HIV therapy: the new challenges, Ann. Ist. Super. Sanita 47 (2011) 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) 75–90.

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

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

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

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

[17] E.J. Arts, D.J. Hazuda, HIV-1 antiretroviral drug therapy, Cold Spring Harb. Perspect. Med. 2 (2012) 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) 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) e130–e132.

[20] R. O'Neal, Rilpivirine and complera: new first-line treatment options, BETA 23 (2011) 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) 75–97.

[22] S.D. Portsmouth, C.J. Scott, The renaissance of fixed dose combinations: Combivir, Ther. Clin. Risk Manag. 3 (2007) 579–583.
[23] B. Coutinho, R. Prasad, Emtricitabine/tenofovir (Truvada) for HIV prophylaxis, Am. Fam. Physician 88 (2013) 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) 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) 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) 2230–2240.
[27] R. Guha, On exploring structure-activity relationships, Methods Mol. Biol. 993 (2013) 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) 5–14.
[29] J. Patel, Science of the science, drug discovery and artificial neural networks, Curr. Drug Discov. Technol. 10 (2013) 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) 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) 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) 70–82.
[33] A.K. Debnath, Application of 3D-QSAR techniques in anti-HIV-1 drug design–an overview, Curr. Pharm. Des. 11 (2005) 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) 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) 683–703.
[36] K. Swiderek, S. Marti, V. Moliner, Theoretical studies of HIV-1 reverse transcriptase inhibition, Phys. Chem. Chem. Phys. 14 (2012) 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) 2010–2026.
[38] Y. Hu, J. Bajorath, Molecular scaffolds with high propensity to form multi-target activity cliffs, J. Chem. Inf. Model. 50 (2010) 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) 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) 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) 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) 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) 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) 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) 316–324.
[46] K. Heikamp, J. Bajorath, Large-scale similarity search profiling of ChEMBL compound data sets, J. Chem. Inf. Model. 51 (2011) 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) 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) D1083–D1090.
[49] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423.
[50] J.S. Gonzalez-Garcia, J. Diaz, Information theory and the ethylene genetic network, Plant Signal. Behav. 6 (2011) 1483–1498.
[51] C. Waltermann, E. Klipp, Information theory based approaches to cellular signaling, Biochim. Biophys. Acta 1810 (2011) 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) 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) 21–33.
[55] V.R. Magnuson, D.K. Harriss, S.C. Basak, Studies in Physical and Theoretical Chemistry, in: 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) 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) 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) 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) 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) 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) 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) 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) 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.