Modeling Antibacterial Activity with Machine Learning and Fusion of Chemical Structure Information with Microorganism Metabolic Networks

Deyani Nocedo-Mena, Carlos Cornelio, María del Rayo Camacho-Corona, Elvira Garza-González, Noemi Waksman de Torres, Sonia Arrasate, Nuria Sotomayor, Esther Lete, and Humbert González-Día

Modeling Antibacterial Activity with Machine Learning and Fusion of Chemical Structure Information with Microorganism Metabolic Networks

Deyani Nocedo-Mena[1,2], Carlos Cornelio[1], María del Rayo Camacho-Corona[2,*],

Elvira Garza-González[3], Noemi Waksman de Torres[4], Sonia Arrasate[1],

Nuria Sotomayor,[1] Esther Lete,[1] and Humbert González-Díaz[1,5,*]

[1]Department of Organic Chemistry II, University of the Basque Country UPV/EHU, 48940, Leioa, Spain.

[2]Universidad Autónoma de Nuevo León, Facultad de Ciencias Químicas, CP 66455, San Nicolás de los Garza, Nuevo León, México.

[3]Universidad Autónoma de Nuevo León, Servicio de Gastroenterología, Hospital Universitario, Dr. Eleuterio González, CP 64460, Monterrey, Nuevo León, México.

[4]Universidad Autónoma de Nuevo León, Facultad de Medicina, CP 64460, Monterrey, Nuevo León, México.

[5]IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Biscay, Spain.

ABSTRACT. Predicting the activity of new chemical compounds over pathogenic microorganisms with different Metabolic Reaction Networks (MRN$_s$) is an important goal due to the different susceptibility to antibiotics. ChEMBL database contains >160 000 outcomes of preclinical assays of antimicrobial activity for 55931 compounds with >365 parameters of activity (MIC, IC$_{50}$, *etc.*) and >90 bacteria strains of >25 bacterial species. In addition, Leong & Barabàsi data set includes >40 MRNs of microorganisms. However, there are no models able to predict antibacterial activity for multiple assays considering both drug and MRN structures at the same time. In this work, we combined Perturbation-Theory, Machine Learning, and Information Fusion techniques to develop the first PTMLIF model. The best linear model found presented values of Specificity = 90.31/90.40 and Sensitivity = 88.14/88.07 in training/validation series. We carried out a comparison to non-linear Artificial Neural Network (ANN) techniques and previous models from literature. Next, we illustrated the practical use of the model with an experimental case of study. We reported for the first time the isolation and characterization of terpenes from the plant *Cissus incisa*. The antibacterial activity of the terpenes was experimentally determined. The more active compounds were phytol and α-amyrin, with MIC = 100 μg/mL for Vancomycin-resistant *Enterococcus faecium* and *Acinetobacter baumannii* resistant to carbapenems. These compounds are already known from other sources. However, they have been isolated and evaluated for the first time here against several strains of multidrug-resistant bacteria included World Health Organization (WHO) priority pathogens. Last, we used the model to predict the activity of these compounds *vs.* other microorganisms with different MRNs in order to find other potential targets.

1. INTRODUCTION

The current situation of bacterial resistance according to World Health Organization (WHO) is alarming.[1] Bacterial resistance to conventional antibiotics has risen dramatically over the past decade, depleting treatment options and fundamentally altering the approach to infection prevention and treatment.[2] The unabated rise in antibiotic resistance, coupled to collateral damage to normal flora by overuse of broad-spectrum antibiotics, requires the development of new antibiotics that are specifically active against multidrug-resistant and extensively drug-resistant Gram-negative bacteria.[3] The global threat of antibiotic resistant bacteria has led to development of different strategies to address this problem. In this sense, understanding the metabolism of pathogens plays an important role, although little known, in the development of antibiotic resistance. Several approaches that integrate experimental data at the level of systems with metabolic networks (for example, genomic scale)[4] have recently been applied to elucidate the metabolic dependencies of resistance, as well as to identify pharmacological targets and possible antibacterials.[5] In fact, Barabási´s group and other authors have demonstrated the influence of the changes in Metabolic Reaction Networks (MRN_s) over the capacity of survival of different microorganisms.[6] In this context, due to the increase in the incidence of antibiotic resistant infections, natural products from plants become interesting alternatives. Therefore, the search for new antibacterial agents derived from plants should be directed to the discovery of natural sources of structurally diverse compounds, whose mechanisms of action were different from those of commercial drugs.[7] As part of the extensive exploration of the endemic flora of Mexico, studies on species that have not validated their medicinal uses are underway. Among these species, *Cissus incisa* is included, which has been traditionally used to treat respiratory and skin infections, as well as abscesses. However, to the best of our knowledge there are not previous reports on phytochemical studies of this plant.

On the other hand, the use of computational models may become a very useful tool in the discovery and development of drugs. First, these models can lead to savings in terms of resources and research time. Additionally, it is possible to analyze hundreds of data at the same time and to get valuable conclusions when establishing relationships between them. Many Cheminformatics models have been developed for the discovery of antimicrobial compounds against different microbes, but they are limited to the prediction of their biological activity in a given strain under certain conditions.[8] In this sense, the Perturbation Theory model combined with ML methods (PT + ML = PTML models) can overcome these limitations. The PTML models developed by our group have been used in Medicinal Chemistry, Proteomics, Materials Chemistry, *etc*., to model large data sets with Big Data characteristics.[9,10] Speck-Planche and Cordeiro *et al*. have also developed some PTML models for different biological activities.[11]To the best of our knowledge, there are no reports on PTML models for the prediction of antimicrobial compounds against several types of bacterial strains, analyzing at the same time, modifications in their MRN_s involved in this biological activity.

In this work, we report for the first time a new PTML model for the prediction of antibacterial compounds taking into account the structure of the compound, the conditions of assay (different activity parameters or bacterial strains), and variations on the MRN of the bacteria. For that purpose, we downloaded a large database from ChEMBL with >83000 preclinical assays of compounds *vs*. different bacterial strains. We also compiled

the structural information for >40 MRN$_s$ of different microorganisms reported by Barabási´s group.[6] Then, we applied an Information Fusion (IF) to merge both ChEMBL and MRN$_s$ datasets. The information of both datasets was pre-processed and all values transformed into a Shannon's entropy scale previously to fusion.[12] Next, we applied a ML technique to find the best PTMLIF (PTML + IF) predictive model. On the other hand, we also carried out for the first time a phytochemical study of *C. incisa*, which allowed us to identify several compounds, among them: phytol, α-amyrin, β-amyrin, and β-sitosterol. Additionally, their antibacterial properties against multi-resistant strains were evaluated experimentally. At last, we used the PTMLIF model to predict the antibacterial activity of the more active compound to exemplify the use of the model in the practice. In **Figure 1**, we illustrate the general workflow for this research.
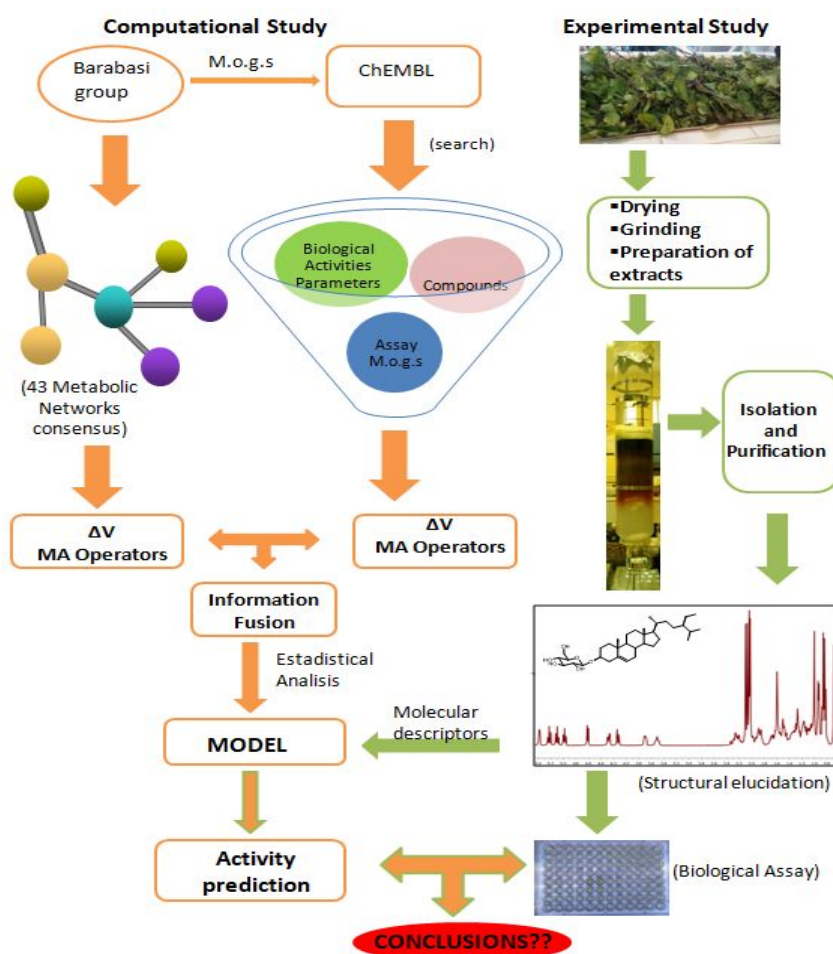

Figure 1. General workflow used in this paper

2. MATERIALS AND METHODS

2.1 Computational Section.

ChEMBL-MRN Data Pre-processing and Information Fusion. The data analysis procedure used here implies three main steps: (1) data acquisition and pre-processing, (2) the IF procedure to fusion both ChEMBL and MRNs datasets, and (3) obtaining the PTMLIF model *per se*. In the data acquisition and pre-processing process, we started obtaining the outcomes of many preclinical assays from ChEMBL database. The result of each assay was expressed by one experimental parameter ε_{ij} used to quantify the biological activity of the ith molecule (m_i) over the jth target. The values of ε_{ij} depend on the structure of the drug and also on a series of

boundary conditions that delimit the characteristics of the assay $\mathbf{c}_j = (c_0, c_1, c_2, \ldots c_n)$. The conditions c_j are $c_1 =$ the biological activity and $c_2 =$ bacteria strain used in the preclinical assay. In the data set we found many different biological parameters v_{ij}; for instance Minimal Inhibitory Concentration (MIC ($\mu g \cdot mL^{-1}$)) or Minimal Bactericide Concentration (MBC ($\mu g \cdot mL^{-1}$)), *etc*. The values v_{ij} compiled are not exact numbers in many cases; they report for instance MIC ($\mu g \cdot mL^{-1}$) < 100. In addition, we considered that the properties may have a positive or negative desirability $d(c_1)$. Positive desirability was set $d(c_1) = 1$ when we want to maximize the value v_{ij} of the biological activity parameter to obtain an optimal drug (this is the case of Selectivity ratio). Conversely, negative desirability $d(c_1) = -1$ means that we want to minimize the value v_{ij} of the biological activity parameter (for instance MIC ($\mu g \cdot mL^{-1}$). These facts difficult the development of a regression model and consequently we discretized the values to seek a classification model. Discretization was as follow: $f(v_{ij})_{obs} = 1$ when $v_{ij} >$ cutoff and $d(c_1) = 1$. The value is also $f(v_{ij})_{obs} = 1$ when $v_{ij} <$ cutoff and desirability $d(c_1) = -1$, $f(v_{ij})_{obs} = 0$ otherwise. The value $f(v_{ij})_{obs} = 1$ points to and strong effect of the compound over the target.[10]

In order to carry out the IF process ChEMBL and MRNs data fusion we decided to express all the information from ChEMBL and from MRNs (metabolic information) in the same scale. Consequently, the information of both datasets was transformed into a Shannon`s entropy scale previously to fusion. The information obtained from ChEMBL (chemical structure) have been scaled using the following formula to calculate the Shannon's entropy value.[12]

$$Sh_k(Drug_i) = -p(D_{ki}) \cdot \log(p(D_{ki})) = -\frac{(D_{ki} - D_{kimin} + 0.001)}{(D_{kimax} - D_{kimin} + 0.001)} \cdot \log\left[\frac{(D_{ki} - D_{kimin} + 0.001)}{(D_{kimax} - D_{kimin} + 0.001)}\right] \quad (1)$$

In this formula, D_{ki} is the value of the molecular descriptor of the drug (LogP or PSA). The value $p(D_{ki})$ is the result of scaling the molecular descriptors to a probability scale ranging from 0 to 1. The values D_{kimin} and D_{kimax} are the minimum and maximum value of the molecular descriptor D_{ki} throughout the data set. The value 0.001 was added as a scaling value to avoid $p(D_{ki}) = 0$; forbidden for the logarithmic entropy function.[12]

The previous $Sh_k(Drug_i)$ values were used to quantifying the structure of the chemical compounds. However, in this IF procedure we have to quantify also the structural information of the MRNs for the different species. Barabàsi's group[6] kindly released MRN_s data upon author's request. The files are in gzipped ASCII files, where each number represents a substrate in the metabolic network. The data format is from $\rightarrow$ to (directed link). This information was published originally by Jeong *et al.*[6] We also quantified the structural information of the MRNs with Shannon's entropy information scale. The formula used to calculate these values is the following; please see details on the literature (see details in SI00.pdf):[12]

$$Sh_k(MRN_s) = -\sum_{q=1}^{q=qmax} p(m_q) \cdot \log p(m_q) \quad (2)$$

PTMLIF linear model. PTML Cheminformatics method ideas have been extended here to find the new PTMLIF model. The output of the PTMLIF model are the scoring function values $f(v_{ij})_{calc}$ for biological activity of the ith compound assayed in the jth preclinical assay with conditions $\mathbf{c}_j = (c_0, c_1)$ against the sth bacteria specie with MRN_s. PTMLIF linear models have the following general equation:

$$f\langle v_{ijs}\rangle_{calc} = a_0 + a_1 \cdot f\langle v_{ij}\rangle_{expt} + \sum_{k=1}^{k=2} a_k \cdot Sh_k(Drug_i) \qquad (3)$$

$$+ \sum_{k=1,j=1}^{k=2,j=2} a_{k,j} \cdot \Delta Sh_k\langle Assay_j\rangle_{c_j}$$

$$+ \sum_{k=1,j=1}^{k=2,j=2} a_{k,j} \cdot \Delta Sh_k(MRN_s)_{c_j}$$

The PTMLIF model starts with the expected value of biological activity $f(v_{ij})_{expt}$ and sums the effect of the chemical information related to the structure of the drug and the accumulated effects due to changes or perturbations (PT operators) in the conditions of assay or the bacteria strain used. The PT operators used here are similar to Box-Jenkins Moving Average (MA) operators used in previous works.[13] The other PT operators included in this model are Moving Averages (MA) calculated for one condition at time. In the PTMLIF model we have two types of PT operators due to the IF process. One type of PT operators is the terms $\Delta Sh_k(Assay_j)_{cj}$ and the other type are the terms $\Delta Sh_k(MRN_s)_{cj}$. We calculated these variables as $\Delta Sh_k(Assay_j)_{cj} = Sh_k(Drug_i) - \langle Sh_k(Assay_j)_{cj}\rangle$ or $\Delta Sh_k(MRN_s) = Sh_k(MRN_s) - \langle Sh_k(MRN_s)_{cj}\rangle$, respectively. Consequently, $\Delta Sh_k(Assay_j)_{cj}$ terms account for the deviation of the chemical information of the compound $Sh_k(Drug_i)$ from the expected value of $\langle Sh_k(c_j)\rangle$ (average value) for all compounds assayed under the same conditions c_j in ChEMBL[12] dataset. By analogy, $\Delta Sh_k(MRN_s)_{cj}$ terms quantify the deviation of the metabolic information of the bacteria $Sh_k(MRN_s)$ from the expected value of $\langle Sh_k(MRN_s)_{cj}\rangle$ for all bacteria strains used in assays with the same conditions c_j in ChEMBL data.

2.2. Experimental section.

Chemicals and equipment. Reagents and solvents were purchased from Sigma-Aldrich and used without further treatment. The following compounds: phytol (**1**), α-amyrin (**2**), β-amyrin (**3**), were purchased from Sigma-Aldrich. β-sitosterol acetate (**5**)[14] was synthesized from β-sitosterol (**4**). Thin layer chromatography (TLC) was carried out on 0.2 mm thick silica gel plates (Merck 60 F254). Visualization was accomplished by UV light and/or ceric sulfate solution in sulfuric acid. NMR spectra were recorded on a Bruker NMR 400 spectrometer at 20-25 ˚C, at 400 MHz for ^{1}H and 100 MHz for ^{13}C in CDCl$_3$ solutions using tetramethylsilane (TMS) as internal reference.

GC/MS analysis. The sample was processed on an Agilent 6890 Gas Chromatograph coupled to an Agilent model 5973 Selective Mass Detector. The column used was a HP-5MS column (30 m x 0.250 mm x 0.25 microM). Helium was used as a carrier gas at a constant flow of 1 mL per min. Injector temperature of 250 ^{0}C, temperature of the ion source 230 ^{0}C. The temperature of the oven was programmed from 50^{0}C, with an increase of 2^{0}C / min to 285^{0}C. The total execution time of the GC was 35 minutes. An MSD detector was used. Mass spectra were recorded under electron impact (EI) at 70 eV. The results that are reported are given with reference to the NIST library database version 1.7^a.

Vegetal material. *Cissus incisa* leaves were collected in Rayones, Nuevo Leon, Mexico, in October 2016. A reference sample was deposited in the herbarium of the Faculty of Biology of the Autonomous University of Nuevo Leon obtaining the voucher number: 027499. Leaves were dried in the shade for 2 weeks and then ground in a knife mill, obtaining 809 g of plant material.

Preparation of the extract. Dry and ground material was macerated 24 h with 1000 mL of hexane. Then, the organic extract was filtered by gravity, then under vacuum and finally concentrated in a rotary evaporator, yielding 0.748 g of the dried extract.

Synthesis. β-sitosterol acetate (**5**).[14] To a solution of β-sitosterol (**4**) (25 mg, 0.060 mmol) in pyridine (0.5 mL, 6.2 mmol), acetic anhydride (0.5 mL, 5.3 mmol) was added slowly. The reaction mixture was stirred overnight. Then ethyl acetate (10 mL) and the organic layer was washed with 10% HCl (4 × 10 mL), dried over sodium sulfate and the solvent concentrated under reduced pressure to give acetylated β-sitosterol as white needles (13,8 mg, 61%). ^{1}H NMR (400 MHz, CDCl$_3$) δ (ppm): 0.70 (3H, s, CH$_3$), 0.84 (d, J = 7.6 Hz, 6H, (CH_3)$_2$CH), 0.87 (t, J = 7.8 Hz, 3H, CH_3CH$_2$), 0.94 (d, J = 6.2 Hz, 3H, CH$_3$), 1.04 (s, 3H, CH$_3$), 1.08-2.05 (m, 27H), 2.06 (3H, s, CH$_3$CO), 2.33-2.39 (2H, m), 4.57-4.68 (m, 1H, H-3), 5.40 (1H, bd, J=4.2Hz, H-6). ^{13}C NMR (100 MHz, CDCl$_3$) δ (ppm): 11.87 (CH$_3$), 12.00 (CH$_3$), 18.74 (CH$_3$), 19.04 (CH$_3$), 19.32 (CH$_3$), 19.83 (CH$_3$), 21.04 (C11), 21.46 (CH$_3$), 23.08 (C28), 24.3 (C16), 26.08 (C23), 27.79 (C15), 28.26 (C2), 29.16 (C25), 31.87 (C8), 31.91 (C7), 33.95 (C22), 36.17 (C20), 36.61 (C10), 37.00 (C1), 38.13 (C4), 39.73 (C12), 42.33 (C13), 45.85 (C24), 50.04 (C9), 56.04 (C17), 56.7 (C14), 74.00 (C3), 122.66 (C6), 139.67 (C5), 170.56 (CH$_3$CO).

2.3 Antibacterial activity assays.

Bacteria and inoculum preparation. Strains of drug-resistant clinical isolates of Gram-negative and Gram-positive bacteria were used, from the University Hospital Dr. Eleuterio González of the Autonomous University of Nuevo Leon, four of which are included in the list of priority pathogens issued by the WHO.[1] The bacteria Gram positive tested were Methicillin-Resistant *Staphylococcus aureus* (MRSA) (14-2095), Linezolid- resistant *Staphylococcus epidermidis* (LRSE) (14-583), Vancomycin-resistant *Enterococcus faecium* (VREF) (10-984). Gram-negative: *Acinetobacter baumannii* resistant to carbapenems(ABRC),*Escherichia coli* producing Extended-spectrum beta lactamase (ECPE)(14-2081), *Pseudomona aeruginosa* resistant to carbapenems (PARC) (13-1391), *Klebsiella pneumonia* NDM-1+ resistant to carbapenems and broad-spectrum cephalosporins (KPNDM-1+)(14-3335), *Klebsiella pneumonia* producer of ESBL (KPPE)(14-2081) and *Klebsiella pneumonia*(OXA-48) resistant to oxacillins (KPRO). Strains were inoculated on plates prepared with 5% blood agar and cultured for 24 h at 37°C. The *inoculum* was prepared by transferring three to five colonies of each culture to tubes with sterile saline, and the turbidity was adjusted to 0.5 of the McFarland standard (1.5 X 10^8 CFU/ml). Then 10 μL in 11 ml of Mueller Hinton broth were transferred to reach 5 x 10^{5}CFU / ml.[15]

Determination of the Minimum Inhibitory Concentration (MIC). The antibacterial activity was developed by microdilution method previously reported by *Zgoda et al.*[15] Levofloxacin was used as reference standard. Experiments were conducted in triplicate. The MIC was determined as the Minimum Concentration of the compound that inhibits the growth of the bacteria.

3. RESULTS AND DISCUSSION

PTMLIF linear model. The projected PTMLIF model is the combination of PTML modeling and Information Fusion (IF) procedures. The model begins with the expected value of biological activity and incorporates the effect of different perturbations in the system. These perturbations are expressed in terms of PT operators. The selected operators are of different type $f(v_{ij})_{expt}$, $Sh_k(Drug_i)$, $\Delta Sh_k(Assay_j)$, $\Delta Sh_k(MRN_s)$. A detailed explanation about all the input variables analyzed is shown in **Table 1**. The equation of the best model found is the following:

$$f\langle v_{ij}\rangle_{calc} = -5.683 + 14.434 \cdot f\langle v_{ij}\rangle_{expt} - 16.426 \cdot Sh_1(Drug) \quad (4)$$
$$+ 24.818 \cdot DSh_1(Assay)_{c1} + 0.211 \cdot DSh_2(Assay)_{c1}$$
$$+ 1.882 \cdot DSh_1(Assay)_{c0} - 107.050 \cdot Sh_1(MRN)_{c2}$$
$$+ 155.395 \cdot Sh_2(MRN)_{c2}$$

$$n = 126848 \qquad \chi^2 = 122496.8 \qquad p < 0.05$$

Table 1. Input variables

c_j	Condition	Symbol	Operator Formula	Operator Information
c_0	Biological activity	$f(v_{ij})_{expt}$	$n(f(v_{ij})_{obs}=1)/n_j$	Expected value of probability $p(f(v_{ij})=1)_{expt}$ for a given type of activity (v_{ij})
-	Drugs Chemical structure	$Sh_k(Drug_i)$	-	Accounts for variability on chemical structure information of the drugs in terms of lipophilicity expressed as LogP ($k=1$) or surface area expressed as PSA ($k=2$)
c_0	Drug structure *vs*. Biological activity	$\Delta Sh_k(Assay_j)_{c0}$	$Sh_k(Drug_i) - \langle Sh_k(Assay_j)_{c0}\rangle$	Accounts for variability on chemical structure information with respect to the structure of the drugs with the same biological parameter measured (c_0)
c_1	Drug structure *vs*. Assay organism	$\Delta Sh_k(Assay_j)_{c1}$	$Sh_k(Drug_i) - \langle Sh_k(Assay_j)_{c1}\rangle$	Accounts for variability on chemical structure information with respect to the structure of the drugs assayed against the same bacterial strain (c_1)
c_2	MRNs structure	$Sh_k(MRN_s)$	-	Accounts for the variability on the information about MRN_s structure

The first input variable $f(v_{ij})_{expt}$ is the expected value of biological activity for one compound or a given type of activity $c_j = (c_1, c_2)$. The specific molecular descriptors were the min-max scaled Shannon entropies used to measure hydrophobicity and polar surface area features of the drug. In the materials and methods section, we show how these entropies can be obtained from the original values of LogP (*n*-Octanol/Water Partition Coefficient) and PSA (Polar Surface Area).These values were taken directly from ChEMBL data set. The output of the model $f(v_{ij})_{calc}$ is a scoring function of the value v_{ij} of biological activity of the ith drug in the different combinations of conditions of assay c_j. For an LDA model $f(v_{ij})_{calc}$ is not in the range 0-1 and it is not a probability. Nevertheless, for a given value of $f(v_{ij})_{calc}$ the LDA algorithm can calculate the respective values of posterior probabilities $p(f(v_{ij})= 1)_{pred}$. The LDA algorithm uses the Mahalanobis's distance metric to calculate these probabilities.[16] Calculating $p(f(v_{ij}) = 1)_{pred}$, we can decide whether the compound is active with $f(v_{ij})_{pred}= 1$ (when $p(f(v_{ij}) = 1)_{pred} > 0.5$) or not. Counting the number of cases with $f(v_{ij})_{pred} = f(v_{ij})_{obs}= 1$ or -1 (correct classifications) *vs*. $f(v_{ij})_{pred} \neq f(v_{ij})_{obs}$ (incorrect classification), we can determine the Sn and Sp of the model.[16] This model is useful to discover the activity of any compound for different combinations of experimental conditions. First, we have to substitute the expected probability of activity $p(f(v_{ij})_{obs}=1)_{expt}$ on the equation, see **Table 2**. It should be noted that these values change for different activities, like Zone of inhibition (mM), MIC_{50} (ug.mL^{-1}), MBC (ug.mL^{-1}), LD_{50} (uM), Activity (%), *etc.* As consequence, the model can predict several activity parameters for a given compound. Then, we have to substitute the values of ALOGP for a new compound (taken from ChEMBL and/or calculated with software).

Table 2.One-condition averages, cutoff, desirability $d(c_0)$,*etc.*, for selected biological parameters

Activity $c_0{}^a$	$<Sh_k(Drug_i)>$		$<Sh_k(MRNs)>$		Parameters used to specify $c_0{}^b$				
	k = 1	k = 2	k = 1	k = 2	$n_j(c_1)$	n	p	cutoff	$d(c_0)$
Zone of inhibition (mM)	0.160	2.950	0.014	0.0126	70	37	0.529	19.6	1
MIC_{50} (ug.mL^{-1})	0.142	2.318	0.011	0.012	2930	2442	0.833	21.5	-1
MIC_{90} (ug.mL^{-1})	0.144	2.483	0.012	0.012	4670	3607	0.772	34.9	-1
MIC (ug.mL^{-1})	0.153	1.935	0.011	0.012	92674	92064	0.993	3825.5	-1
MBC	0.153	2.648	0.013	0.013				117.2	-1

(ug.mL^{-1})					1349	1004	0.744		
LD_{50} (uM)	0.150	2.556	0.012	0.014	24	19	0.792	28.3	-1
Activity (%)	0.152	2.732	0.011	0.012	4880	2373	0.486	47.2	1

[b] $n = n_j(f(v_{ij}) = 1)_{obs}$ and $p = p(f(v_{ij})=1)_{expt}$

Regarding the computation of these expected values of probability, we must evaluate the formula $p(f(v_{ij})_{obs}=1)_{expt} = n(f(v_{ij})=1)_{obs}/n_j$. This is the ratio between the number of drugs $n(f(v_{ij})=1)_{obs}$ with a desired level of activity for the condition c_j and the number of drugs n_j assayed for the same condition c_j. We assume that a compound has a desired level of activity $f(v_{ij})_{obs}=1$ when the value of activity $v_{ij}>$ cutoff for those activities with desirability $d(c_0) = 1$. A compound also has a desired level of activity $f(v_{ij})_{obs}= 1$ when the value of activity $v_{ij}<$ cutoff for activities with desirability $d(c_0) = -1$. On the other hand, when the compound is considered not to have a desired level of activity, $f(v_{ij})_{obs}= 0$. Otherwise, the desirability $d(c_0) = 1$ for properties of the compound that we want to maximize and $d(c_0) = -1$. The cutoff $= 100$ for properties with units in nM. If not, cutoff $= <v_{ij}>$ expected value (average) of the value of activity v_{ij}.

In order to predict the activity of a new compound, we also have to substitute in the model the expected values of the molecular descriptors $\Delta DSh(c_j)$ for different conditions. **Table 3** shows selected values of the averages $\Delta DSh(c_j)$. We can appreciate that these values change depending on the bacterial strain, so the model provides a different result for one compound if you change this condition. For example, $<Sh_0(MRN)>=0.0036$ for *P. aeruginosa* and $<Sh_0(MRN)> = 0.0047$ for *H. influenzae*. This means that the model is able to predict a different activity in different microorganisms for the same drug. The full list of the values of entropy for MRN_s of selected organisms is included in the supplementary material.

Table 3. Values of entropy for MRN_s of selected organisms

Condition c_2[a]	Parameters used to specify c_2		
MRN_s	$<Sh_0(MRN)>$	$<Sh_1(MRN)>$	$n_j(c_2)$
P. aeruginosa	0.0036	0.0191	15457
A. thaliana	0.0088	0.0116	177
B. subtilis	0.0032	0.0159	20547
C. acetobutylicum	0.0051	0.0066	15
C. elegans	0.0056	0.0079	441
C. jejuni	0.0068	0.0100	598
C. trachomatis	0.0129	0.0170	323

E. coli	0.0031	0.0081	16799
E. faecalis	0.0064	0.0079	15449
E. nidulans	0.0070	0.0101	112
H. influenzae	0.0047	0.0159	7171
M. tuberculosis H37Rv	0.0045	0.0155	23048

[a]The full names of the species are: *Pseudomona aeruginosa, Arabidopsis thaliana, Bacillus subtilis, Clostridium acetobutylicum, Caenorhabditis elegans, Campylobacter jejuni, Chlamydia trachomatis, Escherichia coli, Enterococcus faecalis, Emericella nidulans, Haemophilus influenzae, Mycobacterium tuberculosis* H37Rv.

This model shows high values of Specificity Sp = 90.31, Sensitivity Sn = 88.14, and overall Accuracy Ac = 88.65 in training series, taking into account the high number of experimental conditions (see **Table 4).**In addition, the model displays very similar values of Sn, Sp, and Ac in external validation series, see also **Table 4**. As reported by Hill and Lewicki, [16] we used the forward-stepwise strategy of variable selection to detect the more important perturbations on different conditions. It is important to mention that the obtained values are in the range considered as useful for classification models with application in Medicinal Chemistry.[17] The data points (Drug-Assay pair) used in validation series have not been used to train the model. In Supporting Information (SI) file SI00.pdf we give details about the model, in SI01.xls we give the average values, and in SI02.xls we depict details of the classification and probability for each case.

Table 4.Results of the model and input variables analyzed

Obs. Sets[a]	Stat. Param.[b]	Pred. Stat.	Predicted sets	
			$f(v_{ij})_{pred} = -1$	$f(v_{ij})_{pred} = 1$
Training series				
$f(v_{ij})_{obs} = -1$	Sp	90.3	27248	2933
$f(v_{ij})_{obs} = 1$	Sn	88.1	11464	85203
Total	Ac	88.7		
Validation series				
$f(v_{ij})_{obs} = -1$	Sp	90.3	9062	968
$f(v_{ij})_{obs} = 1$	Sn	88.1	3842	28410
Total	Ac	88.6		

PTMLIF ANN models. We also trained other type of PTMLIF models using a different class of ML algorithms. Specifically, we used linear and non-linear Artificial Neural Network (ANN) algorithms to train alternative models. Almost all found PTMLIF-ANN models reached values of Sp and Sn ≈ 88% and AUROC> 0.9 in training and validation series. In **Figure 2**, the curves for all ANN models (almost overlapped in many cases)

are shown. In **Table 5**, the values of Sp and Sn are depicted. In any case, none of them outperformed the PMTLIF-LDA linear model reported in the previous section with Sp and Sn $\approx$ 88.1 - 90.3%. The Linear Neural Network (LNN) model, very similar to the LDA technique, gives almost the same results. The addition of one or two hidden layers of neurons in the Multilayer Perceptron (MLP) does not improve the Sp and Sn during training times above 1h. In addition, the Radial Basis Function (RBF) topology presented a decrease in these values with Sp and Sn $\approx$ 73 − 74%, see **Table 5**, which is in agreement with the previous hypothesis that there is a linear relationship between the classification of the compound and the used PT operators.
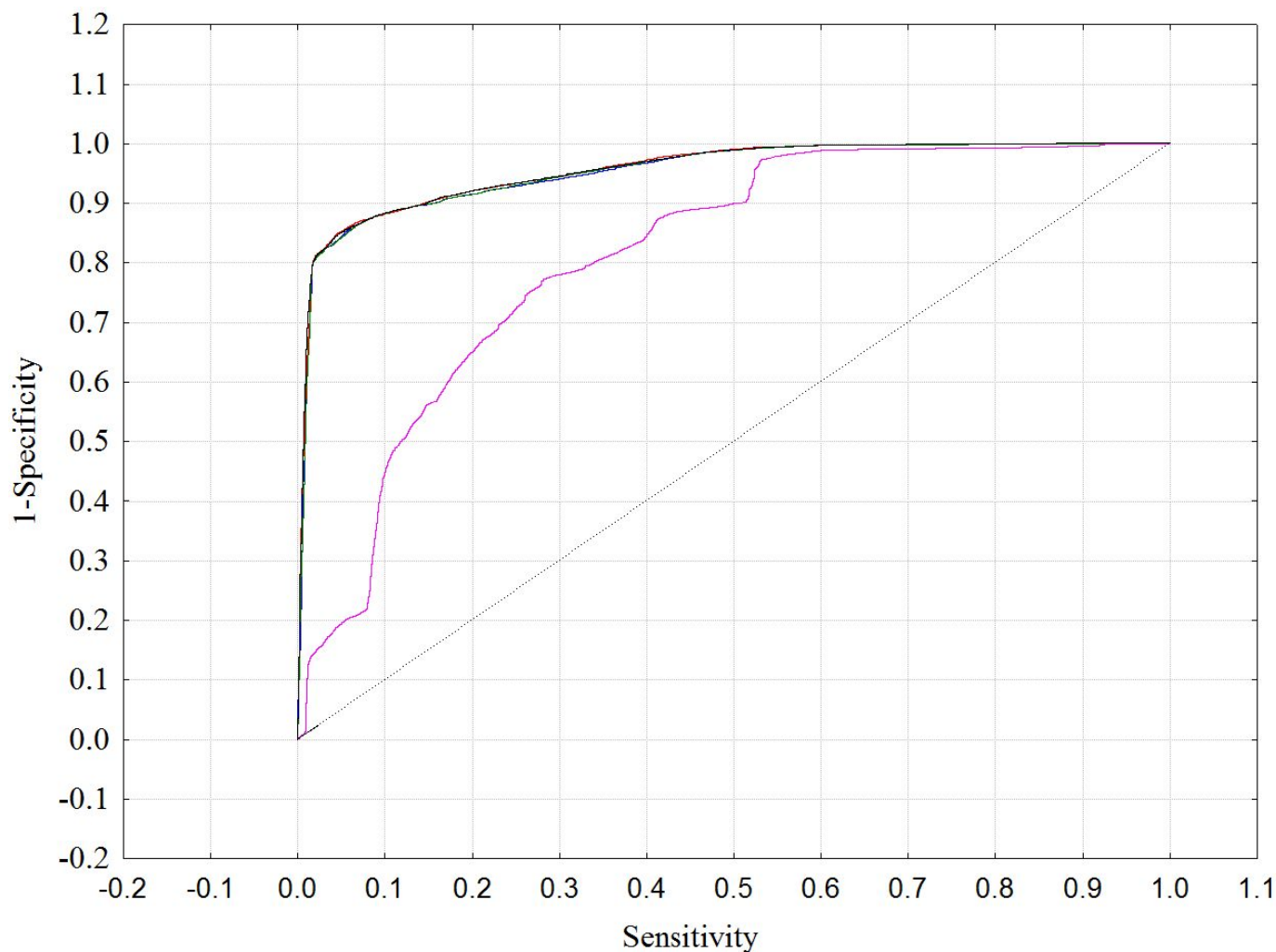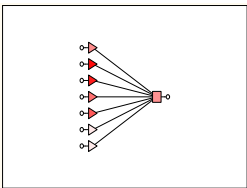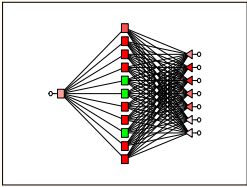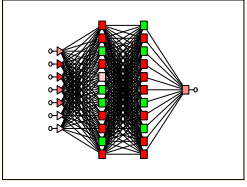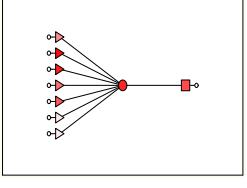

Figure 2. ROC curve analysis

Table 5. PTMLIF-ANN models

Profile	Set	-1	1	(%)	Parm.	(%)	L_{ij}	-1	1
LNN 7:7-1:1	-1	26784	10818	88.74	Sp	88.66	8893	3634	26784
	1	3397	85849	88.81	Sn	88.73	1137	28618	3397
MLP 7:7-11-1:1	-1	26742	10963	88.61	Sp	88.74	8901	3668	26742
	1	3439	85704	88.66	Sn	88.63	1129	28584	3439
MLP 7:7-11-11-1:1	-1	26783	10798	88.74	Sp	88.70	8897	3631	26783
	1	3398	85869	88.83	Sn	88.74	1133	28621	3398
RBF 7:7-1-1:1	-1	22245	25230	73.71	Sp	74.54	7476	8362	22245
	1	7936	71437	73.90	Sn	74.07	2554	23890	7936

Comparison with other models. Various PTML models for the discovery of antibacterial compounds have been previously reported. In **Table 6**, a comparison between the present model and some of these models is shown. In this comparative study, we included 20 models [8,9,11,18-30] most of which (80%) are based on heterogeneous series of compounds (model 4[18], model 5[19], model 7[22], model 8[8], models 10-20[23-30]. However, two models were based on peptides[11,20], one on nanoparticles[21] and another one on antituberculosis drugs.[9] Regarding the number of cases, we can see twelve models that include hundreds of cases, which represent 60 %, while the rest (40%) include much larger amounts. We should note that the model reported in this paper fits a very complex and notably larger data set of n> 83000 cases as compared to the other models. Regarding the complexity of the models, most of them are small models, including between 4 and 7 variables, except model 13, which included 62 variables. The LDA predominates among the techniques used in the realization of the models. This technique was used in 15 out of 20 models, representing 75% (models: 1-4,6-7,9-11,14-15,17-20). It was followed by ANN in four models (5,8,13,16), which represents the 20% and BLR in only one (5%).[24] Regarding the accuracy, it should be noted that all compared models have precision values higher than 85%.The predominant validation technique was the external predicting series, which was used in 17 out of 20 models, including this one. This shows that we used a proven validation technique. As shown in **Table 6** (entries 10-

20), the models are not able of predicting multiple species, that is, they only predict a single type of microorganism.

Table 6. Comparison to other PTML models of antibacterial compounds

m^a	Cmpd. Type[a]	n^b	Var.[b]	Tech.[c]	Acc (%)	Val.[d]	Multi Species[e]	Drug Family[f]	MO[g]	Net.[h]	Ref.
1	HSC	83605	6	LDA	88.6	i	MBS	>10	Yes	Yes	This work
2	Peptide	3592	4	LDA	96.0	i	MBS	>10	Yes	No	11
3	Peptide	2488	6	LDA	90.0	i	*Gram + bacteria*	>10	Yes	No	20
4	HSC	30181	6	LDA	90.0	i	*F.necrophorum P. intermedia*	>10	Yes	No	18
5	HSC	54000	6	ANN	90.0	i	*Pseudomonas* spp	>10	Yes	No	19
6	Nano.	300	7	LDA	77.7	i	MBS	>10	Yes	No	21
7	HSC	37800	5	LDA	95.0	i	No	>10	Yes	No	22
8	HSC	11576	4	ANN	97.0	i	*Streptococcus* spp	>10	Yes	No	8
9	ATD	12000	4	LDA	90.0	i	*Mycobacterium* spp	>10	Yes	No	9
10	HSC	667	7	LDA	92.9	i	No	>10	No	No	23
11	HSC	661	6	LDA	92.6	ii	No	8	No	No	24
12	HSC	661	6	BLR	94.7	ii	No	8	No	No	24
13	HSC	661	62	ANN	-	iii	No	8	No	No	24
14	HSC	352	7	LDA	91.0	i	No	9	No	No	25
15	HSC	111	7	LDA	94.0	i	No	3	No	No	26
16	HSC	111	7	ANN	89.0	i	No	3	No	No	26
17	HSC	-	8	LDA	> 90	i	No	-	No	No	27
18	HSC	972	8	LDA	86.8	i	No	> 5	No	No	28
19	HSC	458	2	LDA	~ 85	i	No	-	No	No	29
20	HSC	433	6	LDA	~ 85	i	No	> 8	No	No	30

[a]Compound type: HSC = Heterogeneous Series of compounds, anti-TB drug = antituberculosis drugs. [b]Total number of cases in training and/or validation series and Vars. = Variables in the model. [c]Technique: LDA = Linear discriminant analysis, ANN= artificial neural network, BLR=binary logistic regression. [d]Validation methods: i) external predicting series, ii) leave-30%-out cross validation, and iii) 100-times-averaged re-substitution technique. Furthermore, note that methods ii and iii are cross-validation methods. [e]Multi Species: Multiple bacterial strain (MBS), *Fusobacterium necrophorum, Prevotella intermedia.* [f]Drug Family: Only largely represented families were considered. [g]MO = Multi Output: multi-output models are those able to predict more than one type of

biological activity (MIC, IC_{50}, MBC, *etc*.). [h]Net. =MRN_s: Models able to account for changes in the MRN_s of different microorganisms.

Multispecies models appeared recently, however, some of them predict biological activity only for the same genus or within a subgroup of bacteria (models 1 to 9). Similarly, models from 10 to 20 are not multi-output, while the rest are. Therefore, we have presented two generations of models. Those between one and nine correspond to the more contemporary models, representing the 45% of the analyzed total. The present PTML model is able to predict the antibacterial activity of any compound against different bacteria strains. However, the principal contribution is to include the MRN_s. With the present model, a determined reaction on the interior of a bacterium can be varied, which consequently brings changes in its metabolic pathway. Thus, key points are identified, which can be targeted for the action of the drug. In addition, the search for drugs is addressed. The application of the model can reduce the number of candidates, with the subsequent saving in time and resources. These results are in consonance with the application of other ML techniques in drug discovery.[31-36]

Phytochemical study. To our knowledge, until now, there are no reports on chemical composition of extracts from *C. incisa*. Hence, the present study is focused on the analysis of terpenoid compounds and its antimicrobial activity. Using chromatographic and spectral analyses and reported data, four known compounds, which had not been previously isolated from the leaves of *C. incisa*, were identified. The hexane extract was analyzed by GC/MS. The chromatogram (**Figure 3)** showed the following compounds: phytol **(1)** (71.91 min; 4.09%), α-amyrin **(2)** (116.18 min; 3.18%), β-amyrin **(3)** (114.91; 8.43%), and β-sitosterol **(4)** (114.53 min; 19.44%), the latter one being the most abundant. Mass spectra of compounds **1-4** are included in the supporting information section. These results are consistent with those reported by Pathomwichaiwat *et al*.,[37] who identified triterpenes, phytols, steroids, and their derivatives from a hexane extract of *Cissus quadrangularis*. Besides, species within this genus, such as: *C. quadrangularis*, *C. aralioides*, *C. polyantha,* and *C. cornifolia* have also been previously studied.[37-40] Different compounds, such as: fatty acids, phenolic compounds, pyrogallics, polysaccharides, flavonoids, sterols (β-sitosterol, stigmasterol), terpenes (α-amyrin, β-amyrin, oleanolic acid, and lupeol), stilbenes, and glycosides, among others, have been identified and/or isolated.
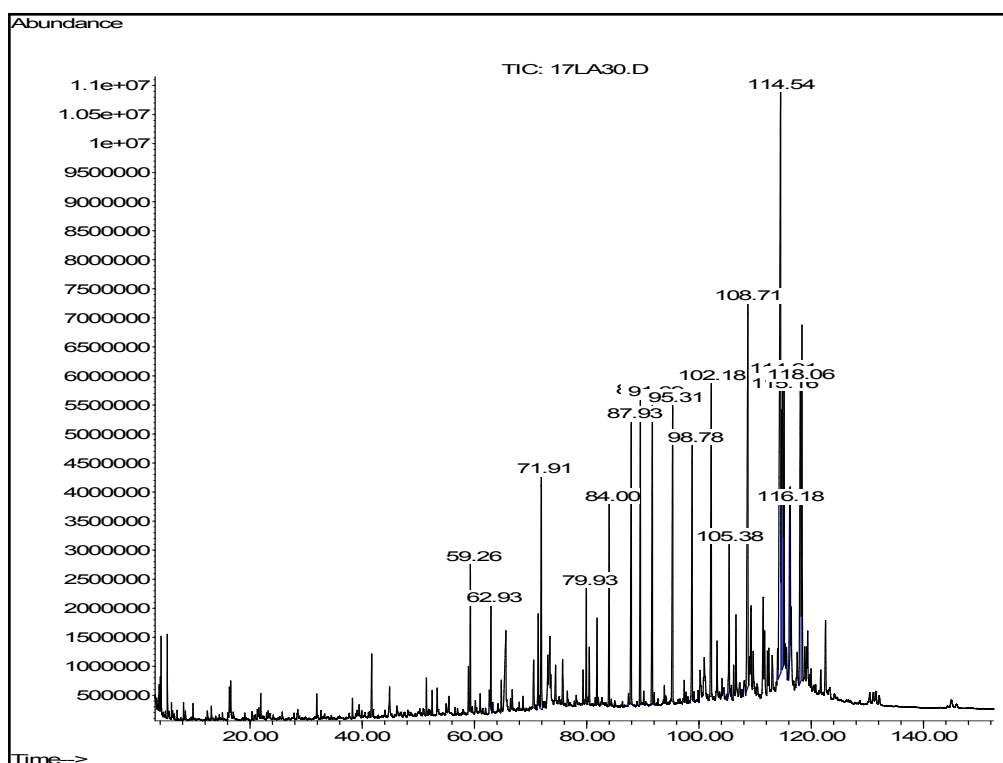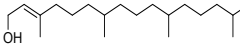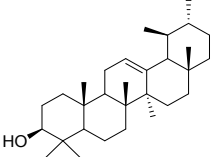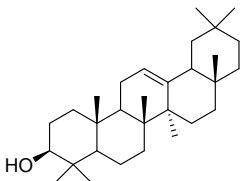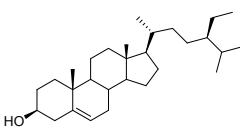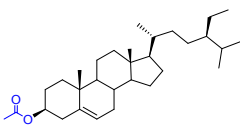
Figure 3.CG chromatogram of hexane extract of *C. incisa* leaves

In the plant world, it is common to find these compounds, because they fulfill important roles in plants. The phytol is a well-known diterpene. It is found in the wax layer of the leaf and is an essential component of chlorophyll. Sterols are very abundant and their function is to maintain the structure and functioning of cell membranes. On the other hand, amyrins are found in various plants and plant materials, such as leaves, bark, wood and resins. These compounds, which provide protection against herbivores, favor germination of seeds and inhibit root growth. It is promising for our study the presence of these metabolites. Specially, those whose biological properties have been previously reported (antibacterial, antifungal, and anticancer activity).[41-45] In addition, acetylation of β-sitosterol was performed obtaining the corresponding acetylated derivative, whose spectroscopic data are in agreement with those previously reported for β-sitosterol acetate **(5)**,[14] which allowed us to confirm its structure.

Antibacterial Activity. Results in **Table 7** showed that compounds **1, 2,** and **5** were active against different strains. Among them, phytol developed major antibacterial activity with MIC of 100µg / mL against two strains (Vancomycin-resistant *E. faecium* and *A. baumannii* resistant to carbapenems). This result is in agreement with the previous outcome reported for the antimicrobial activity of pentacyclic terpenes by Hernández-Vázquez *et al.*[46] They obtained a MIC range of 64-1088 µg/mL against ATCC strains of *S. aureus, E. faecium, P. aeruginosa, E. coli,* and *K. pneumoniae*. It is important to point out that the acetylated derivative was more active than the natural product. In fact, β-sitosterol acetate achieved MIC of 100 µg/mL for *A. baumannii* resistant to carbapenems. To obtain semisynthetic derivatives is a strategy used to improve multiple characteristics of natural antibiotics, such as their power and antimicrobial spectrum, decreasing the toxicity and other unwanted effects.[47]

Table 7. Antibacterial activity of isolated compounds from *C. incisa* leaves

Compounds	MRSA	LRSE	VREF	ABRC	EPEC	PARC	KPNMD-1	KPPE	KPRO	Ref
(1) Phytol	>200	200	100	100	200	200	200	200	200	This work
						ATCC 20				43
(2) α−amyrin	>200	200	200	100	200	200	>200	200	200	This work
	ATCC >1024					ATCC 200				48,49
(3) β−amyrin	>200	>200	>200	200	>200	200	>200	>200	>200	This work
	STM >800	ATCC 500	ATCC 250		ATCC 120	ATCC 1000	ATCC 500			50
(4) β−sitosterol	>200	200	200	200	>200	200	200	>200	200	This work
	ATCC >500				ATCC >500	ATCC 200	STM >800			49,50,51
(5) β−sitosterol acetate	>200	200	200	100	200	200	>200	200	200	This work
	>200	>200	>200	>200	>200	100	>200	>200	>200	
Levofloxacin	12.5	6.25	12.5	12.5	25	0.78	>50	12.5	>50	

[a](MRSA)*Methicillin*-resistant *Staphylococcus aureus,*(LRSE)Linezolid- resistant *Staphylococcus epidermidis*, (VREF) Vancomycin–resistant *Enterococcus faecium,* (ABRC) *Acinetobacter baumannii* resistant to carbapenems, (EPEC) ESBL-producing *Escherichia coli*, (PARC) *Pseudomona aeruginosa* resistant to carbapenems, (KPNDM-1 +) *Klebsiella pneumoniae* NDM-1 +, (KPPE) *Klebsiella pneumoniae* producer of ESBL, (KPRO) *Klebsiella pneumoniae* resistant to oxacillins, (STM)*Streptococcus mutants* ATCC.

The tested compounds were more effective against Gram-negatives bacteria: *A. baumannii* resistant to carbapenems*,* and *P. aeruginosa* resistant to carbapenems (MIC=100 µg/mL). Although the antibacterial activity of our compounds was not comparable to that of standard Levofloxacin, the fact that they were active against Gram-negatives bacteria is valuable. Making them good candidates to be used in combination with established antimicrobials or became platforms for future antibiotics. In this sense, natural products and those optimized by synthesis will become the next generation of antibacterial agents.[47,52]

Predictive study. In this section, we are going to illustrate the practical use of the model with one case of study. Based on earlier results, we selected phytol and α-amyrin, compounds with most interesting biological activity tested in the experimental part. Both compounds are widely distributed in nature: phytol is a diterpene, which is a component of chlorophyll and α-amyrin is related to plant protection. As we mentioned above, the

antibacterial properties have been previously determined for these compounds, but in sensitive strains.[43] In ChEMBL, 149 assays for phytol, *e.g.* against *Mycobacterium tuberculosis H37Rv, E. coli, S. aureus, Aspergillus flavus* have been reported.[53-56] However, we did not found reports of assays of phytol against other species of bacteria with different MRN_s. In the case of α-amyrin, we found in ChEMBL 33 assays of biological activity as anti-cancer, anti-viral and anti-parasitic, but there are no antibacterial reports.

After applying the model, phytol was predicted to be active against all the tested bacteria obtaining values of $p(f(v_{ij}) = 1)_{pred}$ equal to 1 for all cases. These results are in agreement with those obtained experimentally (see **Table 7**). In our reported experimental results, phytol was active against another strain of *Enterococcus* (*faecium)*. On the other hand, selected results for the predictive study of phytol and α-amyrin are shown in **Table 8**. Like phytol, α-amyrin would be active for all strains. However, its probability to reach this level of activity would be variable for different species of bacteria taking into account the values of $p(f(v_{ij}) = 1)_{pred}$. The highest value was obtained for *E. faecalis*, which repeats as the most sensitive strain. The analysis of $p(f(v_{ij}) = 1)_{min-max}$ totally matches the predicted activity. Consequently, we can conclude that the model applied for the predictions for the antibacterial activity is correct, because of the concordance between computational and experimental assays.

Table 8. Selected results for the prediction of the antibacterial activity for selected compounds

Organism of Assay	ChEMBL n_j	Assay		$f(v_{ij})$ calc	$p(f(v_{ij})=1)$ pred	MRN			
		$<Sh_1>$	$<Sh_2>$			N	L_{out}	$<Sh_1>$	$<Sh_2>$
Phytol									
Bacillus subtilis	20547	0.141	2.442	67.25	1.0000	785	2741	0.016	0.014
Escherichia coli	16259	0.148	2.606	66.99	1.0000	778	2859	0.008	0.008
Enterococcus faecalis	15006	0.146	2.543	67.52	1.0000	386	1218	0.008	0.011
Haemophilus influenzae	7164	0.150	2.774	66.86	1.0000	526	1746	0.016	0.013
Pseudomonas aeruginosa	14968	0.147	2.582	67.08	1.0000	587	1823	0.015	0.014
α-Amyrin									
Bacillus subtilis	20547	0.141	2.442	2.79	0.8313	785	2741	0.016	0.014
Escherichia coli	16259	0.148	2.606	2.53	0.8064	778	2859	0.008	0.008
Enterococcus faecalis	15006	0.146	2.543	3.06	0.8632	386	1218	0.008	0.011
Haemophilus influenzae	7164	0.150	2.774	2.40	0.7899	526	1746	0.016	0.013
Pseudomonas aeruginosa	3283	0.141	2.360	2.90	0.8432	587	1823	0.015	0.014

4. CONCLUSIONS

We have developed the first computational model able to predict the antibacterial activity taking into account the structure of MRN$_s$. We demonstrated that entropies could be used to measure the structure of the drug, the different assays, and the metabolic network. The neural networks showed no improvement over the linear model. On the other hand, we report the first phytochemical study of the leaves of *Cissus incisa*. Regarding the antibacterial activity of the identified compounds, phytol was the compound with the best antibacterial activity (MIC = 100μg/mL) against Vancomycin-resistant *E. faecium* and *A. baumannii* resistant to carbapenems. Finally, the predictive study showed that predictions of other compounds against different bacterial strains can be made using the developed computational model. It was also shown that phytol is active for measured biological activity, just like amyrin, but with greater variability. Finally, our model is superior to others in relation with the number of cases and the incorporation of complex networks.

■ SUPPORTING INFORMATION

This section includes mass spectra of compounds **1-4** of the hexane extract**.** Also includes a more detailed explanation of the computational section. In addition, we released the dataset used: compound code, molecular descriptors, assay conditions, values of entropy for MRN$_s$, and observed vs. predicted classification of each compounds.

■ AUTHORS INFORMATION

Corresponding Author

*E-mail: humberto.gonzalezdiaz@ehu.es (H.G.-D.); phone: +34 (94) 6013547

*E-mail: camacho.corona@uanl.mx.cu (M.D.R.C.C.); phone: +52 (81) 8329-4000 (3414)

ORCID:

Deyani Nocedo-Mena:0000-0001-8061-8609

Sonia Arrasate:0000-0003-2601-5959

Nuria Sotomayor: 0000-0003-3079-6380

Esther Lete: 0000-0001-8624-6842

Humbert González-Díaz:0000-0002-9392-2797

Authors Contributions

D.N.M. and M.D.R.C.C., extraction, isolation, and characterization of compounds, writing and discussion of results D.N.M., E.G. and M.D.R.C.C., biological assays. M.D.R.C.C and E.G., supervision of D.N.M experimental work. E.L. and N.S., compound characterization, writing, and discussion of results. C.C., S.A., and H.G.D., ChEMBL dataset download, data pre-processing, obtained and discussed computational models. S.A. and H.G.D., supervision of C.C. and D.N.M computational work. NWT registration of NMR spectra.

■ ACKNOWLEDGMENTS

■ ABBREVIATIONS

ChEMBL, Chemical European Molecular Biology Laboratory database; AUROC, Area Under Receiver Operating Curve; ML, machine learning; ROC, Receiver Operating Curve; MIC, Minimal inhibitory concentration; ESBL, Extended-spectrum beta lactamase; NDM, New Delhi metallo-β-lactamase; ATCC, American Type Culture Collection.

■ CONFLICT OF INTEREST

The authors declare that they have no competing interests.

■ REFERENCES

(1) Tacconelli, E., Carrara, E., Savoldi, A., Harbarth, S., Mendelson, M., Monnet, DL., Pulcini, C., Kahlmeter, G., Kluytmans, J., Carmeli, Y., Ouellette, M., Outterson, K., Patel, J., Cavaleri, M., Cox, E.M., Houchens, C.R., Grayson, M.L., Hansen, P., Singh, N., Theuretzbacher, U., Magrini, N. WHO Pathogens Priority List Working Group. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect. Dis*. **2018**, *18,* 318-327.

(2) Zaengle-Barone, J. M., Jackson, A.C., Besse, D. M., Becken, B. A., Mehreen, S., Patrick, C., Franz, K. J. Copper Influences the Antibacterial Outcomes of a β-Lactamase-Activated Prochelator against Drug-Resistant Bacteria. *ACS Infect. Dis*. **2018**, *4*, 1019–1029.

(3) Tehrani,K. H. M., Ebrahim, M., Nathaniel, I. Thiol-Containing Metallo-β-Lactamase Inhibitors Resensitize Resistant Gram-Negative Bacteria to Meropenem. *ACS Infect. Dis*. **2017**, *3*, 711–717.

(4) Dunphy, L. J., Papin, J. A. Biomedical applications of genome-scale metabolic reconstructions network of human pathogens. *Curr. Opin. Biotechnol*. **201**8, *51*, 70–79.

(5) Lupoli, T. J., Vaubourgeix, J., Burns-Huang, K., Gold, B. Targeting the Proteostasis Network for Mycobacterial Drug Discovery. *ACS Infect. Dis*. **2018,** *4*, 478–498.

(6) Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabási, A.L. The large-scale organization of metabolic networks. *Nature*. **2000,** *760*, 651–654.

(7) Camacho-Corona, M.R., García, A., Mata-Cárdenas, B.D., Garza-González, E., Ibarra-Alvarado, C., Rojas-Molina, A., Rojas-Molina, I., Bah, M., Sánchez, M.Á., Gutiérrez, S.P. Screening for antibacterial and antiprotozoal activities of crude extracts derived from mexican medicinal plants. *African J. Tradit. Complementary Altern. Med*. **2015**, *12*, 104–112.

(8) Speck-Planche, A., Kleandrova, V. V., Cordeiro, M. N. D. S. Cheminformatics 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.

(9) Speck-Planche, A., Kleandrova, V. V., Cordeiro, M. N. D. S. New insights toward the discovery of

antibacterial agents: Multi-tasking QSBER model for the simultaneous prediction of anti-tuberculosis activity and toxicological profiles of drugs. *Eur. J. Pharm. Sci*. **2013,** *48*, 812–818.

(10) Bediaga, H., Arrasate, S., González-Díaz, H. PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS Comb Sci*. **2018,** *20*, 621–632.

(11) Kleandrova, V. V., Ruso, J. M., Speck-Planche, A., Cordeiro, M. N. D. S. Enabling the Discovery and Virtual Screening of Potent and Safe Antimicrobial Peptides. Simultaneous Prediction of Antibacterial Activity and Cytotoxicity. *ACS Comb. Sci*. **2016,** *18*, 490–498.

(12) González-Díaz, H., Herrera-Ibatá, D. M., Duardo-Sánchez, A., Munteanu, C.R., Orbegozo-Medina, R. A., 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. *J. Chem. Inf. Model.* **2014**, *54*, 744–755.

(13) Martínez, S. G., Tenorio-Borroto, E., Pliego, A. B., Díaz-Albiter, H., Váazquez-Chagoyan, J.C., Gonzalez-Díaz, H. PTML Model for Proteome Mining of B-cell Epitopes and Theoretic- Experimental Study of Bm86 Protein Sequences from Colima Mexico. *J. Proteome Res.* **2017**, *16*, 4093-4103.

(14) McCarthy, F. O., Chopra, J., Ford, A., Hogan, S.A., Kerry, J. P. B., O'Brien, N.M., Ryanb, E., Maguirea, A. R. Synthesis, isolation and characterization of β-sitosterol and β-sitosterol oxide derivatives. *Org.Biomol.Chem.* **2005**, *3*, 3059–3965.

(15) Zgoda, J. R., Porter, J. R. A Convenient Microdilution Method for Screening Natural Products against Bacteria and Fungi. *Pharm. Biol*. **2001**, *39*, 221–225.

(16) Hill, T., Lewicki, P. *Statistics : Methods and Applications : A Comprehensive Reference for Science, Industry, and Data Mining*, 1st ed; StatSoft, Inc.: Tulsa, Oklahoma**, 2006**.

(17) Marrero-Ponce, Y., Siverio-Mota, D., Gálvez-Llompart, M., Recio, R. M., Giner, M.C. Discovery of novel anti-inflammatory drug-like compounds by aligning in silico and in vivo screening: the nitroindazolinone chemotype. *Eur. J. Med. Chem*. **2011,** *46*, 5736–5753.

(18) Speck-Planche, A., Cordeiro, M. N. D. S. Enabling virtual screening of potent and safer antimicrobial agents against noma: mtk-QSBER model for simultaneous prediction of antibacterial activities and ADMET properties. *Mini-Rev. Med. Chem*. **2015,** *15*, 194–202.

(19) Speck-Planche, A., Cordeiro, M. N. D. S. Computer-Aided Discovery in Antimicrobial Research: In Silico Model for Virtual Screening of Potent and Safe Anti-Pseudomonas Agents. *Comb. Chem. High Throughput Screening*. **2015,** *18*, 305–314.

(20) Speck-Planche, A., Kleandrova, V. V., Ruso, J. M., Cordeiro, M. N. D. S. First Multitarget Chemo-Bioinformatic Model to Enable the Discovery of Antibacterial Peptides against Multiple Gram-Positive Pathogens. *J. Chem. Inf. Model*. **2016,** *56,* 588–598.

(21) Speck-Planche, A., Kleandrova, V. V., Luan, F., Cordeiro, M. N. D. S. Computational modeling in nanomedicine: Prediction of multiple antibacterial profiles of nanoparticles using a quantitative structure-activity relationship perturbation model. *Nanomedicine.* **2015,** *10*, 193–204.

(22) Speck-Planche, A., Cordeiro, M. N. D. S. Simultaneous virtual prediction of anti- *Escherichia coli*

activities and admet profiles: A chemoinformatic complementary approach for high-throughput screening. *ACS Comb. Sci*. **2014,** *16*, 78–84.

(23) González-Díaz, H., Torres-Gómez, L.A., Guevara, Y.,Almeida, M. S., Molina, R., Castañedo, N. Markovian chemicals 'in silico' design (MARCH-INSIDE), a promising approach for computer-aided molecular design III: 2.5D indices for the discovery of antibacterials. *J. Mol. Model*. **2005,** *11*, 116–123.

(24) Cronin, M.T.D.,Aptula, A.O., Dearden, J.C., Duffy, J.C., Netzeva, T.I., Patel, H..Structure-based classification of antibacterial activity. *J. Chem. Inf. Comput. Sci*. **2002,** *42*, 869–878.

(25) Molina, E., Díaz, H. G., González, M. P., Rodríguez, E., Uriarte, E. Designing antibacterial compounds through a topological substructural approach. *J. Chem. Inf. Comput. Sci*. **2004,** *44*, 515–521.

(26) García-Domenech, R., de Julián-Ortiz, J. V. Antimicrobial activity characterization in a heterogeneous group of compounds. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 445–449.

(27) Mut-Ronda, S., Salabert-Salvador, M. T., Duart, M. J., Antón-Fos, G. M. Search compounds with antimicrobial activity by applying molecular topology to selected quinolones. *Bioorg. Med. Chem. Lett*. **2003,** *13*, 2699–2702.

(28) Murcia-Soler, M., Pérez-Giménez, F., García-March, F.J., Salabert-Salvador,M. T., Díaz-Villanueva, W., Medina-Casamayor, P. Discrimination and selection of new potential antibacterial compounds using simple topological descriptors. *J. Mol. Graph. Model*. **2003,** *21*, 375–390.

(29) Murcia-Soler, M. Pérez-Giménez, F., García-March, F.J., Salabert-Salvador, M. T., Díaz-Villanueva, W., Castro-Bleda, M. J. Artificial neural networks and linear discriminant analysis: A valuable combination in the selection of new antibacterial compounds. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 1031–1041.

(30) Mishra, R. K., García-Domenech, R., Gálvez, J. Getting discriminant functions of antibacterial activity from physicochemical and topological parameters. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 387–393.

(31) Pérez-Sánchez, H.; Cano, G.; García, J. Improving Drug Discovery using Hybrid Softcomputing Methods. *Appl. Soft Comput.* **2013**, *20*, 119-126 .

(32) Cano, G.; García-Rodríguez, J.; García-García, A.; Pérez-Sánchez, H.; Benediktsson, J. A.; Thapa, A. ; Barre, A. Automatic Selection of Molecular Descriptors using Random Forest: Application to Drug Discovery, *Expert Syst. Appl.* **2016**,*71*, 151-159.

(33) Ghasemi, F.; Mehridehnavi, A.; Pérez-Garrido, A.; Pérez-Sánchez, H. Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks, *Drug Discovery Today* **2018**, *23*, 1784-1790.

(34) Ghasemi, F.; Mehridehnavi, A. R.; Fassihi, A.; Pérez-Sánchez, H. Deep Neural Network in QSAR studies using Deep Belief Network, Appl. Soft Comput. **2017**, *62*, 251-258.

(35) Jiménez, F.; Pérez-Sánchez, H.; Palma, J.; Sánchez, G.; Martínez, C. A methodology for evaluating multi-objective evolutionary feature selection for classification in the context of virtual screening, *Soft Computing* **2018***, 1-26.

(36) Ghasemi, F.; Fassihi, A.; Pérez-Sánchez, H.; Mehridehnavi, A. The role of Different Sampling Methods in Improving Biological Activity Prediction Using Deep Belief Network, *J. Comput. Chem.* **2017**, *38*, 195-203.

(37) Pathomwichaiwat, T., Ochareon, P., Soonthornchareonnon, N., Ali, Z., Khan, I. A., Prathanturarug, S. Alkaline phosphatase activity-guided isolation of active compounds and new dammarane-type triterpenes from *Cissus quadrangularis* hexane extract. *J. Ethnopharmacol.* **2015**, *160*, 52–60.

(38) Vijayalakshmi, G., Aysha, O. S., Valli, S. Antibacterial, and phytochemical analysis of *Cissus quadrangularis* on selected UTI pathogens and molecular characterization for phylogenetic analysis of *Klebsiella Pneumoniae.* *World J. Pharm. Pharm. Sci.* **2015**, *4*, 1702–1713.

(39) Olaoye S., B., Ibrahim A., O., Zhiqiang, L. Chemical compositions and radical scavenging potentials of essential oils from *Tragia benthamii* (Baker) and *Cissus aralioides* (Welw). *J. Biol. Act. Prod. Nat.* **2016**, *6*, 59–64.

(40) Sani, Y.M, Musa, A. M., Tajuddeen, N., Abdullahi, S. M., Abdullahi, M. I., Pateh, U. U., Idris, A. Y. Isoliquiritigenin and β-sitosterol from *Cissus polyantha* Tuber Glig and Brandt. *J. Med. Plants Res.* **2015**, *9 , 918–921.*

(41) Musa, A. M., Tajuddeen, N., Idris, A. Y, Rafindadi, A.Y., Abdullahi, M.I., Aliyu, A., B., Abdullahi, M. S., Ibrahim, M.A. A New Antimicrobial Prenylated Benzo-lactone from the Rhizome of *Cissus cornifolia*. *Pharmacogn. Res*. **2015**, *7*, 363–366.

(42) Tetali S. D. Terpenes and isoprenoids: a wealth of compounds for global use. *Plant*. **2019,** *249*, 1-8.

(43) Lee, W., Woo, E. R., Lee, D. G. Phytol has antibacterial property by inducing oxidative stress response in *Pseudomona aeruginosa*. *Free Radical Res*. **2016**, *50*, 1309–1318.

(44) Yessoufou, K., Elansary, H. O., Mahmoud, E. A., Skalicka-wo, K. Antifungal, antibacterial and anticancer activities of *Ficus drupacea* L. stem bark extract and biologically active isolated compounds. *Ind. Crops Prod.* **2015**, *74*, 752–758.

(45) Suttiarporn, P., Chumpolsri, W., Mahatheeranont, S., Luangkamin, S., Teepsawang, S., Leardkamolkarn, V. Structures of phytosterols and triterpenoids with potential anti-cancer activity in bran of black non-glutinous rice. *Nutrients* **2015**, *7*, 1672–1687.

(46) Hernández-Vázquez, L., Palazon, J., Navarro-Ocaña, A. The Pentacyclic Triterpenes α, β-amyrins: A Review of Sources and Biological Activities. In *Phytochemicals: A Global Perspective of their Role in Nutrition and Health* (Rao, V., Ed.), In Tech, Shangai, **2012,** 487-502.

(47) Brown, P., Dawson, M. J. A perspective on the next generation of antibacterial agents derived by manipulation of natural products. *Progr. Med. Chem.* **2015**, *54*, 135-184.

(48) Abreu, V. G. C, Takahashi, J. A., Duarte, L.P., Piló-Veloso, D., Junior, P. A. S, Alves, R.O. Evaluation of the bactericidal and trypanocidal activities of triterpenes isolated from the leaves, stems, and flowers of *Lychnophora pinaster*. *Brazilian J. Pharmacogn*. **2011,** *21*, 615–621.

(49) Rasamiravaka, T., Ngezahayo, J., Pottier, L., Ribeiro, S. O., Souard, F., Hari, L., Stévigny, C., El Jaziri, M., Duez, P. Terpenoids from *Platostoma rotundifolium* (Briq.). A. J. Paton alters the expression of
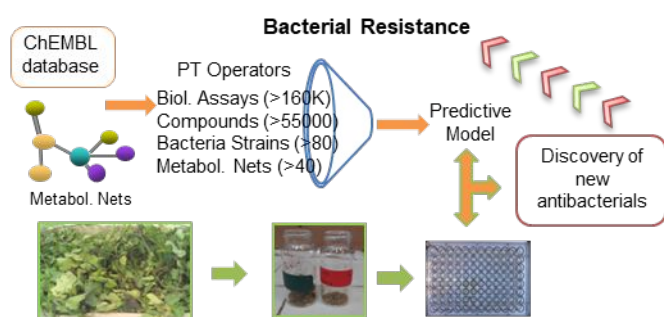
quorum sensing-related virulence factors and the formation of biofilm in *Pseudomona aeruginosa*
PAO1. *Int. J. Mol. Sci*. **2017**, *18*, 1-22.

(50) Kubo, I., Muroi, H., Kubo, A. Antibacterial activity of long-chain alcohols against *Streptococcus mutans*. *J. Agric. Food Chem*. **1993**, *41*, 2447–2450.

(51) Hess, S.C., Brum, R. L., Honda, N. K., Cruz, A.B., Moretto, E., Cruz, R. B., Messana, I., Ferrari, F., Filho, V. C., Yunes, R. A. Antibacterial activity and phytochemical analysis of *Vochysia divergens* (Vochysiaceae*). J. Ethnopharmacol*. **1995**, *47*, 97–100.

(52) Miller, S.I. Antibiotic resistance and regulation of the gram-negative bacteria outer membrane barrier by host innate immune molecules. *mBio.* **2016**, *7*, 1-3.

(53) Saikia, D., Parihar, S.,Chanda, D. Antitubercular potential of some semisynthetic analogues of phytol. *Bioorg. Med. Chem. Lett*. **2019**, *20*, 508–512.

(54) Chen, L., Liang, Y. Synthesis and bioactivity of tripolinolate A from *Tripolium vulgare* and its analogs. *Bioorg. Med. Chem. Lett*. **2015**, *25*, 2629–2633.

(55) Kim, Y.S., Shin, D.H. Volatile Constituents from the leaves of *Callicarpa japonica* Thumb and their Antibacterial Activities. *J Agric. Food Chem.* **2004**, *52*, 781–787.

(56) Escalante, A. Gattuso, M. Evidence for the Mechanism of Action of the Antifungal Phytolaccoside B Isolated from *Phytolacca tetramera* Hauman. *J. Nat. Prod*. **2008**, *71*, 1720–1725.
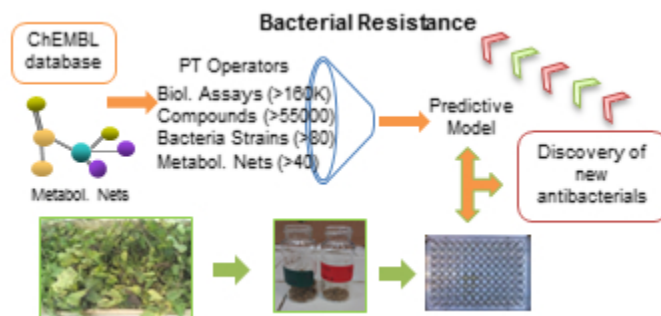
For Table of Contents use only

Modeling Antibacterial Activity with Machine Learning and Fusion of Chemical Structure Information with

Microorganism Metabolic Networks

Deyani Nocedo-Mena, Carlos Cornelio, María del Rayo Camacho-Corona, Elvira Garza-González,

Noemi Waksman de Torres, Sonia Arrasate, Nuria Sotomayor, Esther Lete, and Humbert González-Díaz

Graphical Abstract

24

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

TOC

88x43mm (96 x 96 DPI)