Identification *in silico* and *in vitro* of novel trypanosomicidal drug-like compounds

Juan Alberto Castillo-Garit,[1,2,3,4,*] Oremia del Toro-Cortés,[1] Vladimir V. Kouznetsov,[5] Crisitan Ochoa Puentes,[5] Arnold R. Romero Bohorquez,[5] Maria C. Vega,[6] Miriam Rolón,[6] José A. Escario,[7] Alicia Gómez-Barrio,[7] Yovani Marrero-Ponce,[2,4] Francisco Torrens[4] and Concepción Abad[3]

[1] *Applied Chemistry Research Center, Faculty of Chemistry-Pharmacy, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, 54830, Villa Clara, Cuba. e-mail: jacgarit@yahoo.es, juancg.22@gmail.com or juancg@uclv.edu.cu*

[2] *Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy. Universidad Central "Marta Abreu" de Las Villas, Santa Clara, 54830, Villa Clara, Cuba.*

[3] *Departament de Bioquímica i Biologia Molecular, Universitat de València, E-46100 Burjassot, Spain.*

[4] *Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, P.O. Box 22085, E-46071, València, Spain.*

[5] *Laboratorio de Química Orgánica y Biomolecular, Escuela de Química, Universidad Industrial de Santander, Bucaramanga, Colombia.*

[6] *Centro para el Desarrollo de la Investigacion Cientifica (CEDIC), Pai Perez 265 casi Mariscal Estigarribia. Asuncion-Paraguay.*

[7] *Departamento de Parasitología, Facultad de Farmacia, UCM, Pza. Ramón y Cajal s/n, 28040 Madrid.*

Abstract

Atom-based bilinear indices and linear discriminant analysis (LDA) are used to discover novel trypanosomicidal compounds. The obtained LDA-based quantitative structure-activity relationship (QSAR) models, using non-stochastic and stochastic indices, provide accuracies of 89.02% (85.11%) and 89.60% (88.30%) of the chemicals in the training (test) sets, respectively. Later, both models were applied to the virtual screening of 18 *in house* synthesized compounds to find new pro-lead antitrypanosomal agents. The *in vitro* antitrypanosomal activity of this set against epimastigote forms of *Trypanosoma cruzi* is assayed. Predictions agree with experimental results to a great extent (16/18) of the chemicals. Sixteen compounds show more than 70% of epimastigote inhibition at a concentration 100 µg/mL. In addition, three compounds (CRIS 112, CRIS 140 and CRIS 147) present more than 70% of epimastigote inhibition at a concentration of 10 µg/mL (79.95%, 73.97% and 78.13%, respectively) with low values of cytotoxicity (19.7%, 7.44% and 20.63%, correspondingly).Taking into account all these results, we could say that these three compounds could be optimized in forthcoming works. Even though none of them resulted more active than nifurtimox, the current results constitute a step forward in the search for efficient ways to discover new lead antitrypanosomals.

Keywords: Atom-based bilinear indices, Anti-epimastigote elimination, Cytotoxicity, *Trypanosoma cruzi*, Trypanosomicidal, *virtual* screening.

1. Introduction

Chagas disease is an autochthonous illness that affects to 22 countries in the continental Western Hemisphere (1), caused by the protozoa *Trypanosoma cruzi*. The parasite is found in the vector as an epimastigote and in the human host as an intracellular amastigote (2). It is estimated about 15 million people infected with *T. cruzi*, almost 28 million in risk of being infected and 41 200 new cases reported each year (3). It is also estimated that up to 5.4 million people will develop chronic Chagas heart disease (4, 5), while 900 000 will develop megaesophagus and megacolon (5). Although this disease is typically related to poor and/or rural populations, recent trends in migration have brought *T. cruzi* infection to Latin-American cities and far beyond the borders of Latin America(6, 7).

Human infection is primarily transmitted by domestic and sylvatic insects of the subfamily Triatominae (Hemiptera, Reduviidae), the kissing bug, whose habitat in the Americas ranges from the US and Mexico in the North to Argentina and Chile in the South (1, 8). Vectorial transmission of *T. cruzi* occurs only in endemic countries in the Western Hemisphere. The haematophagous triatomine deposits containing the parasite are excreted on the host while taking a blood meal; inoculation of the parasite into the bite wound, conjunctivae or mucus membranes can result in *T. cruzi* infection. Recently, oral transmission has been reported in several outbreaks (9). *T. cruzi* infection may be also transmitted to humans congenitally, by blood transfusion and organ transplant in non-endemic countries as well as in Latin America (1, 8). The acute phase of Chagas' disease lasts 6–8 weeks; some patients have fever, lymphadenopathy, splenomegaly and/or oedema, but most cases are asymptomatic or oligosymptomatic. Rarely, patients may develop severe disease, with myocarditis or encephalomyelitis; but without treatment, around 5–10% of these patients die (7, 10). When the acute phase ends, *T. cruzi* infection passes into a clinically silent chronic phase designated the indeterminate form (10). Infected individuals may remain asymptomatic for life. However, over a period of ten to thirty years, 20–35% of patients develop symptomatic chronic Chagas' disease, characterized by cardiac and/or gastrointestinal disorders. The gravest complications include high-grade conduction blocks, ventricular arrhythmias, ventricular aneurysm, thrombo-embolic complications, heart failure, and sudden death. Optimum management may require expensive procedures, such as subspecialist examinations, pacemakers, defibrillators, and even heart transplant. A smaller proportion of patients develop digestive system disease, especially megaesophagus or megacolon (7).

On the other hand, the chemotherapy of this parasitic infection remains undeveloped. The treatment is based on old and quite unspecific drugs that have significant activity only in the acute phase of the disease and, when associated with long-term treatments, give rise to severe side effects (11). The currently available chemotherapy for Chagas disease is based on two agents introduced in the market in the 1970s: nifurtimox (a nitrofuran derivative) and benznidazole, (a nitroimidazole derivative). They show limited efficacy to the diseases' acute phase and only against some pathogen strains; they are also associated to severe side effects, including cardiac and/or renal toxicity (12). Their efficacy also varies according to geographical areas, mainly because of differences in drug susceptibility of different *T. cruzi* strains (13). Benznidazole efficacy and tolerance are inversely related to the age of the patient, while its side effects are more frequent in elderly patients (14). Furthermore, medication is expensive, for example, nifurtimox regimen requiring 10 mg/kg in three or four doses per day over a 60- to 120-day period (15). Once the disease has progressed to later stages, no medication has consistently proved to be effective (12).

As mentioned above it is necessary to search for new effective and less toxic chemotherapeutic and chemo-prophylactic agents against *T. cruzi*. However, the great costs associated with the development of new drugs and the small economic size of the market for antiprotozoals, make this development slow (16). In this context, our research group has recently developed a novel scheme to generate molecular fingerprints based on the application of discrete mathematics and linear algebra theory. The approach [known as ***TOMOCOMD*** acronym of *TOpological MOlecular COMputer Design*] (17-19) allows us to perform rational *in silico* molecular design (selection/identification) and Quantitative Structure-Activity/Property Relationship (QSAR/ QSPR) studies. In fact, this scheme has been successfully applied to the prediction of several physical, physicochemical, chemical, pharmacokinetical, toxicological as well as biological properties (20-25). Furthermore, these molecular descriptors (MDs) have been extended to consider three-dimensional (3D) features of small/medium-sized molecules based on the trigonometric-3D-chirality-correction factor approach (26-31).

In the present report, atom-based non-stochastic and stochastic bilinear indices are used to find classification models that allow the discrimination of antitrypanosomal compounds. This approach permits the rational identification of those candidates to be evaluated, which have the highest probabilities of being active ones. Following this idea, 18 already-synthesized compounds were then *in silico* evaluated and, after that, *in vitro* assayed against epimastigote forms of *Trypanosoma cruzi*. Cytotoxic studies were also conducted, as a selection criterion of compounds with good activity vs, lowest toxicity.

2. Results and Discussion

2.1. Development and validation of the discriminant functions.

The linear discriminant analysis (LDA) has become an important tool for the prediction of chemical properties. Because of the simplicity of this method, many useful discriminant models have been developed and presented by different authors in the literature (21, 23, 32-35). It was the technique used in the generation of a discriminant function in the present work. Also, the principle of maximal parsimony (Occam's razor) was taken into account as the strategy for model selection (36). The general dataset was randomly divided into two subsets, training and test set (which have 346 and 94 compounds, respectively), both of them containing active and inactive compounds. The best models obtained using atom-based non-stochastic and stochastic bilinear indices as molecular descriptors, together with their statistical parameters are given below, respectively:

$$\boldsymbol{Class} = -5.103 - 2.9\text{x}10^{-8}\, {}^{\text{MK}}\boldsymbol{b}_{13\text{L}}(\overline{w}_E) + 6.62\text{x}10^{-9}\, {}^{\text{MK}}\boldsymbol{b}_{14\text{L}}(\overline{w}_E) + 1.52\text{x}10^{-5}\, {}^{\text{MP}}\boldsymbol{b}_{8\text{L}}{}^{\text{H}}(\overline{w}_E)$$
$$-8.70\text{x}10^{-6}\, {}^{\text{MP}}\boldsymbol{b}_{9\text{L}}{}^{\text{H}}(\overline{w}_E) + 5.30\text{x}10^{-7}\, {}^{\text{MV}}\boldsymbol{b}_{9\text{L}}{}^{\text{H}}(\overline{w}_E) - 3.93\text{x}10^{-3}\, {}^{\text{VK}}\boldsymbol{b}_{1\text{L}}{}^{\text{H}}(\overline{w}_E)$$
$$+7.93\text{x}10^{-6}\, {}^{\text{VP}}\boldsymbol{b}_{6}(\overline{w}) \qquad \textbf{(1)}$$

N = 346 $\lambda = 0.42$ Q_{Total} = 89.02 % *MCC* = 0.76

$D^2 = 5.975$ F = 65.73 $p < 0.001$

$$\boldsymbol{Class} = -4.531 + 9\text{x}10^{-2}\, {}^{\text{VPs}}\boldsymbol{b}_{0}{}^{\text{H}}(\overline{w}) + 12.29\text{x}10^{-2}\, {}^{\text{VKs}}\boldsymbol{b}_{1}{}^{\text{H}}(\overline{w}) - 4.22\text{x}10^{-2}\, {}^{\text{VKs}}\boldsymbol{b}_{7}(\overline{w})$$
$$-4.69\text{x}10^{-2}\, {}^{\text{VKs}}\boldsymbol{b}_{1\text{L}}{}^{\text{H}}(\overline{w}_E) - 1.63\, {}^{\text{PKs}}\boldsymbol{b}_{9}{}^{\text{H}}(\overline{w}) + 9.09\text{x}10^{-1}\, {}^{\text{PKs}}\boldsymbol{b}_{6\text{L}}{}^{\text{H}}(\overline{w}_E)$$
$$-2.47\text{x}10^{-2}\, {}^{\text{MPs}}\boldsymbol{b}_{2\text{L}}{}^{\text{H}}(\overline{w}_E) \qquad \textbf{(2)}$$

N = 346 $\lambda = 0.45$ Q_{Total} = 89.60 % *MCC* = 0.77

$D^2 = 5.357$ F = 58.95 $p < 0.001$

where, ***Class*** refers to antitrypanosomal activity, N is the number of compounds, λ is the Wilks' statistic, Q_{Total} is the accuracy of the model for the training set, MCC is the Matthews' correlation coefficient, D^2 is the square Mahalanobis distance, F is the Fisher ratio and *p*-value is its significance level.

Both equations appeared statistically significant at $p<0.001$. The best non-stochastic model (Eq. **1**), which includes non-stochastic indices, present a good overall accuracy of 89.02% for the training set (see Table 1). In addition, this model showed an adequate Matthews' correlation coefficient of 0.76; MCC quantifies the strength of the linear relation between the molecular descriptors and the classifications and, usually, it may provide a much more balanced evaluation of the prediction than, for instance, the percentages (accuracies). Together with the accuracy other parameters such as sensitivity, specificity, and false-positive rate (also known as 'false-alarm rate'), are among the most commonly used parameters in medical statistics. While the sensitivity is the probability of correctly predicting a positive case, the specificity (also known as 'hit rate') is the probability that a positive prediction be correct (37). The non-stochastic model shows, for the training set, a good value of sensitivity of 85.83%, a specificity value of 83.06% and a false-positive rate of only 9.29% (See Table 1). Nevertheless, the most important criterion, for the acceptance or not of a discriminant model, is based on statistics for the external prediction set, which is known as *the predictive power* of the model. For the test set, the non-stochastic model showed an accuracy of 85.11%, MCC of 0.67, a good value of sensitivity of 91.30% and a specificity value of 63.64%, with a 16.90% of false-positive rate.

Table 1. Prediction performances for LDA-based QSAR models for training and test sets.

Models	**Matthews Corr. Coefficient (*C*)**	**Accuracy 'Q_{Total}' (%)**	**Specificity (%)**	**Sensitivity 'hit rate' (%)**	**False positive rate (%)**
			Training set		
Eq. 1	0.76	89.02	83.06	85.83	9.29
Eq. 2	0.77	89.60	82.81	88.33	9.73
			Test set		
Eq. 1	0.67	85.11	63.64	91.30	16.90
Eq. 2	0.74	88.30	68.75	95.65	14.08

On the other hand, the best stochastic model (Eq. **2**) presents a good overall accuracy of 89.60%, with a good MCC value of 0.77 for the training set. These values are slightly better than those obtained with the non-stochastic model. The achieved values for sensitivity and specificity were 88.33% and 82.81%, respectively, as well as a false-positive rate of only 9.73%. For the test set the results of the stochastic model were an accuracy of 88.30%, MCC of 0.74, sensitivity of 95.65%, and specificity of 68.75%; these values are acceptable. All the values are reported in Table 1. The results of the classification for compounds in both, training and test, sets achieved with Eqs. **1** and **2** can be seen in the Supporting Information (Tables S1-S4).

Therefore, the *robustness* of the model refers to the stability of its parameters (predictor coefficients) and, consequently, to the stability of its predictions when a perturbation is applied to the training set and the model is regenerated from the "perturbed" training set. Here, we develop the leave-group-out (LGO) and *Y-scrambling* procedures (3, 38) as very important tools in order to detect what is sometimes referred to as "internal predictivity" and possible chance correlation in the models obtained, respectively (For details, see section 1 of Supporting Information). First, a LGO strategy was performed and the calculation of accuracies in the new training sets and test set compounds permitted us to carry out the assessment of the models. The results of this validation process are illustrated in Figure S1 (see Supporting Information). It can be observed from this plot that the models present a high stability to disturbances within the database. The results of the stochastic model were better than those obtained with the non-stochastic model. After that, the *Y-scrambling* test was carried out. The results of our randomization experiments are shown in Figure S2 (see Supporting Information) and indicate that, when the random group size is increased, the globally good accuracy of the model decreased gradually. This outcome indicates that the values of good overall classification are not because of chance correlation or structural redundancy in the training set.

2.2 *In silico* and experimental identification of novel antitrypanosomals.

The entire algorithm, described in the sections above, was made up with the main objective of exploring the applicability of the QSAR models, obtained with the atom-based bilinear indices, for the identification of 'hits' (pro-lead compounds) from large databases. Therefore, an *in silico* screening of novel compounds was performed, looking for the biological activity concerning this work. In order to carry out this, a pool of approximately 200 compounds available from our academic collaborators never described in the literature as antitrypanosomal agents was chosen. Later, the *in silico* assays were performed by using all

Table 2. Compounds evaluated in the present study, their classification (ΔP%) according to the obtained models, their antitrypanosomal activity and cytotoxicity at three different concentrations (100, 10, and 1 μg/mL) and antitrypanosomal activity of nifurtimox (reference).

		ΔP	**ΔP**	**%AE (SD)[d]**			**%CI(SD)[e]**		
Compound	**Exp.[a]**	**Eq. 1[b]**	**Eq. 2[c]**	100μg/mL	10μg/mL	1μg/mL	100μg/mL	10μg/mL	1μg/mL
CRIS 105	A	94.5	97.5	**72.10 ±0.28**	38.20 ±2.61	14.83 ±5.16	27.58 ±1.45	0.00 ±4.35	0.00 ±2.18
CRIS 109	A	96.0	97.3	**84.21 ±0.75**	56.20 ±1.39	0.00 ±2.05	49.21 ±0.60	10.88 ±1.36	11.66 ±1.70
CRIS 110	A	96.3	97.3	**82.14 ±0.72**	54.15 ±0.89	8.56 ±0.47	65.85 ±1.68	33.48 ±4.61	7.14 ±2.05
CRIS 111	A	96.2	97.8	**83.80 ±1.47**	41.73 ±1.25	23.94 ±1.02	42.91 ±0.47	8.68 ±0.72	0.00 ±1.64
CRIS 112	A	96.4	97.8	**87.24 ±0.29**	**79.95 ±2.17**	15.42 ±1.34	57.99 ±4.88	19.70 ±0.85	0.00 ±1.15
CRIS 116	A	97.9	98.5	**70.84 ±2.38**	53.18 ±1.88	6.98 ±4.25	24.31 ±1.52	9.71 ±1.57	7.85 ±1.30
CRIS 119	A	98.0	98.6	**73.77 ±1.66**	30.71 ±0.88	19.65 ±2.57	63.22 ±1.32	25.69 ±1.32	11.22 ±2.28
CRIS 130	A	97.9	98.6	**76.45 ±2.31**	46.09 ±2.53	0.00 ±2.68	50.21 ±0.82	12.60 ±1.18	0.00 ±2.14
CRIS 131	I	99.8	99.3	35.56 ±2.35	21.71 ±1.81	4.24 ±0.82	20.54 ±1.63	27.56 ±1.45	7.14 ±1.20
CRIS 135	A	94.6	97.6	**81.13 ±2.55**	35.48 ±4.16	10.69 ±1.35	35.18±1.54	11.71±1.33	0.00±0.85
CRIS 140	A	96.1	97.3	**77.46 ±2.69**	**73.97 ±1.79**	33.25 ±1.78	64.19 ±1.10	**7.44 ±1.47**	0.00 ±1.97
CRIS 141	A	96.2	97.8	**75.64 ±0.80**	54.38 ±0.55	8.27 ±1.05	99.46 ±0.21	99.90 ±0.07	34.66 ±1.91
CRIS 142	A	99.8	99.0	**74.82 ±1.65**	22.23 ±5.23	2.51 ±1.67	31.41 ±4.48	19.24 ±1.72	5.72 ±0.65
CRIS 143	A	99.8	99.1	**80.35 ±3.25**	39.01 ±2.11	7.80 ±3.28	71.14 ±3.60	23.14 ±4.10	4.67 ±0.80
CRIS 147	A	99.8	99.1	**99.29 ±0.74**	**78.13 ±0.78**	23.44 ±2.00	37.23 ±0.79	20.63 ±2.12	6.28 ±2.62
CRIS 148	A	99.8	98.9	**82.26 ±1.32**	31.77 ±0.78	12.56 ±4.04	26.79 ±2.42	26.74 ±5.06	6.71 ±1.06
CRIS 149	A	99.8	99.0	**75.00 ±2.96**	48.56 ±0.87	14.34 ±1.95	41.32 ±2.76	10.10 ±1.32	0.00 ±1.93
CRIS 153	I	99.9	99.5	20.31 ±0.56	18.75 ±0.54	21.41 ±0.52	20.63 ±1.20	20.70 ±0.56	3.50 ±1.63
Nifurtimox	A	99.98	98.39	**100±1.49**	**85.45±2.43**	38.21±2.17	11.68	0.6	0.32

[a]Observed activity: A (active), I (inactive)

[b]Results of the classification of compounds obtained from Model 1, ΔP% = [P(active) _ P(inactive)] · 100

[c]Results of the classification of compounds obtained from Model 2, ΔP% = [P(active) _ P(inactive)] · 100

[d]Anti-epimastigotes percentage and standard deviation (SD)

[e]Cytotoxicity percentage and standard deviation (SD)

the models developed inside this report, in order to identify bioactive chemicals that present trypanocidal activity.

Here, 18 new organic compounds were selected as putative antitrypanosomal by the LDA-based QSAR models. However, it is generally acknowledged that QSARs are valid only within the same domain for which they were developed. In fact, even if the models are developed on the same chemicals, the applicability domain (AD) for new chemicals can differ from model to model, depending on the specific molecular descriptors. Therefore, the leverage values (*h*) and standardized residuals related to these 18 compounds were calculated, the *leverage* values of these new compounds and were lower than the value of *warning leverage* (h^* = 0.06); the corresponding leverage plot is shown in Fig. S3 (For details, see Section 2 of supporting information). According to this, these chemicals lie in the applicability domain of the model, consequently their predictions are reliable. This proves the good valuation for the classification of this set of compounds as new antitrypanosomal, and so, this model can be used with high accuracy for the prediction of new compounds within its AD.

After that, the *in vitro* assays of the previously synthesized compounds (Figure 1) were carried out to corroborate the *in silico* predictions. We proceeded to test the compounds in an epimastigote inhibition (*in vitro*) assay (39). The ΔP% values of the compounds in the dataset, using all the discriminant functions and the chemical structures are depicted in Table 2 and Figure 1, respectively. A good agreement (16/18) is observed between the experimental antitrypanosomal activity and theoretical predictions for this set of compounds. Sixteen compounds showed more than 70% of epimastigote inhibition at a concentration of 100µg/mL (see Table 2). Also, three compounds (CRIS 112, CRIS 140 and CRIS 147) demonstrated more than 70% of epimastigote inhibition at a concentration of 10µg/mL (79.95%, 73.97% and 78.13%, respectively). Even though none of them resulted more active than nifurtimox, the current results constitute a step forward in the search for efficient ways to discover new lead antitrypanosomals.

After this preliminary *in vitro* test, the unspecific cytotoxicity was determined against macrophages at the concentrations that were used in the previous assay (39, 40). At this time, three compounds (CRIS 105, CRIS 116 and CRIS 148) that showed more than 70% of epimastigote inhibition, at a concentration of 100µg/mL (Table 2), also presented acceptable values of cytotoxicity (27.58%, 24.31% and 26.79%, respectively). The three compounds with more than 70% activity at a concentration of 10µg/mL (CRIS 112, CRIS 140 and CRIS 147) showed low values of cytotoxicity (19.7%, 7.44% and 20.63%, correspondingly). Taking into account all these results, we can say that some compounds of this group can be optimized in forthcoming works, but we consider that compound CRIS 140 is the best candidate (see Figure 1).

Here we would like to give a brief consideration about the possible structure-activity relationship for this set of compounds. According with the experimental results if we select for example compound CRIS140 with CRIS149 and CRIS153 we can see that the hybridization sp3 of the carbon which the pyridyl ring is attached seem to be better than sp2 hybridization for the trypanosomicidal action. Similar situation can be seen if we compare compounds CRIS112 and CRIS131; in both cases carbons with sp3 hybridization present more % of AE than those which have sp2 hybridization in the same position.

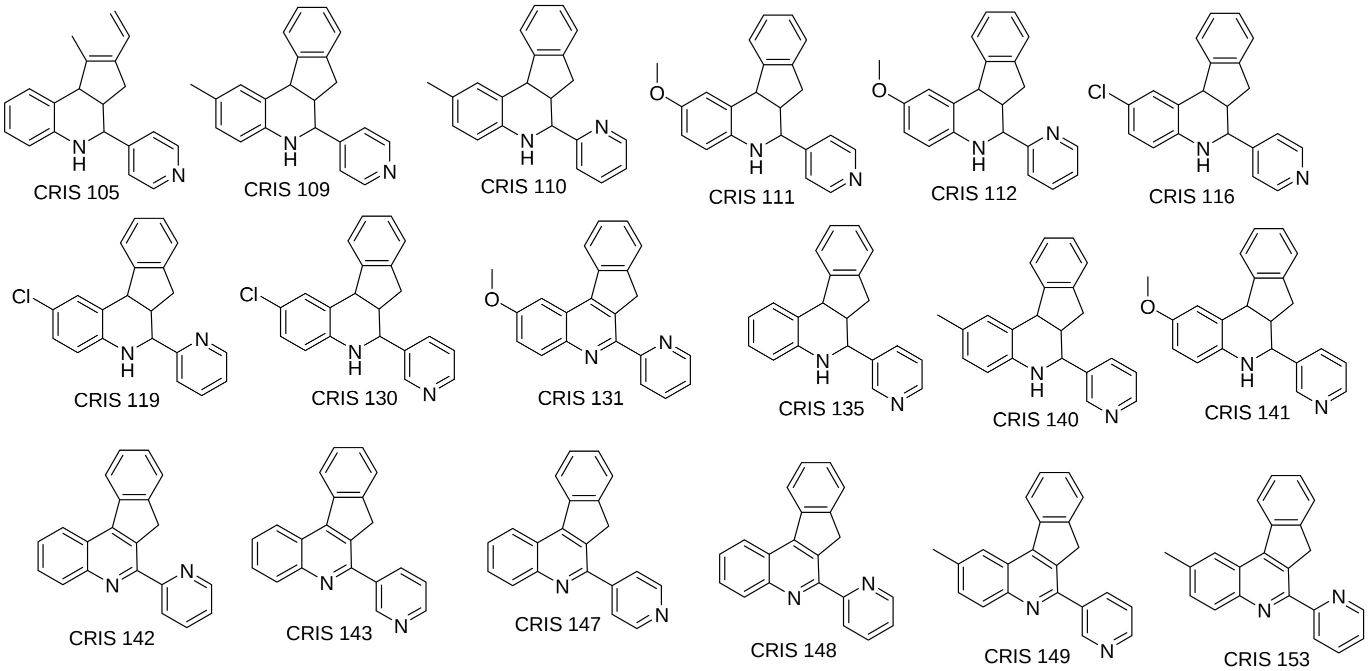

Figure 1. Molecular structures of experimentally evaluated compounds.

On the other hand, the same group of chemicals used in this work was recently tested against other protozoan parasite, *Trichomonas vaginalis*, and all compounds were found inactive at all assayed concentrations, with exception of compound CRIS 148 (41). Therefore, we can say that the antitrypanosomal activity, predicted and experimentally corroborated in this work, is quite specific for this group of compounds. However, a *T. cruzi* amastigote susceptibility assay and other tests of activity against other protozoa parasites are needed, in particular with other protozoa that also belong to the trypanosomatida family like *Leishmania* and *Trypanosoma brucei*.

3. Conclusions

The obtained models, developed using atom-based non-stochastic and stochastic bilinear indices, permit us to classify new "physical" or "virtual" chemicals as active or inactive ones, in the chemotherapy of trypanosomiasis, and they will contribute to a more rational discovery of new lead compounds with antitrypanosomal activity. The usage of this method permits a good prediction of the biological property under consideration, thus increasing the likelihood of an *in silico* discovery of new candidate lead compounds and minimizing the use of resources. In the present report, 16 new compounds, subjected to *in silico* screening, were recognized with antitrypanosomal activity. Afterward, several *in vitro* experiments are performed to corroborate the reliability of the classification functions developed in this work and permit us to select the candidates with the best "activity against epimastigote forms/unspecific cytotoxicity" rate. Finally, we can say that the present algorithm constitutes a step forward in the search for efficient ways of discovering new antitrypanosomal compounds, and constitutes an example of how this rational computer-aided method can help to reduce cost and to increase the rate in which novel chemical entities progress through the pipeline.

4. Experimental Section

4.1. Data-set for QSAR Study

The general data-set used in this study was the same that we utilize in previous works (24, 25) and it consists of 440 compounds of great structural variation, 143 of which are actives and 297 are inactive against trypanosome. For active compounds, it is remarkable that the wide variability of drugs and mechanisms of action in the training and prediction sets assures adequate extrapolation power (For details about the data set please see Section 3 of supporting information).

4.2. Computational approach

The theory of the atom-based bilinear indices used in this study was discussed in detail in earlier publications (31, 35). Specifically, the *CARDD* (*C*omputed-*A*ided *R*ational *D*rug *D*esign) module implemented in the ***TOMOCOMD*** Software (42) was used in the calculation of atom-based non-stochastic and stochastic bilinear indices. In this study, the properties used to differentiate the molecular atoms are those previously proposed for the calculation of the DRAGON descriptors (43-45) i.e., atomic mass (M), atomic polarizability (P), atomic Mullinken electronegativity (K), van der Waals atomic volume (V), plus the atomic electronegativity in Pauling scale (G) (46).

The following descriptors were calculated in this work:

(I) the k^{th} non-stochastic total bilinear indices, not considering and considering H atoms in the molecular pseudograph (G) [$\boldsymbol{b_k}(\bar{x}, \bar{y})$ and $\boldsymbol{b_k}^{H}(\bar{x}, \bar{y})$, respectively].

(II) the k^{th} non-stochastic local (atomic group = heteroatoms: S, N, O) bilinear indices, not considering and considering H atoms in the molecular pseudograph (G) [$\boldsymbol{b_k}_{L}(\bar{x}_{E}, \bar{y}_{E})$ and $\boldsymbol{b_k}_{L}^{H}(\bar{x}_{E}, \bar{y}_{E})$, correspondingly]. These local descriptors denote putative H-bonding acceptors; in addition, they represent charge as well as dipole moment.

(III) the k^{th} non-stochastic local (atomic group = H-atoms bonding to heteroatoms: S, N, O) bilinear indices, considering H atoms in the molecular pseudograph (G) [$\boldsymbol{b}_{kL}^{H}(\bar{x}_{E\text{-}H}, \bar{y}_{E\text{-}H})$]. These local descriptors denote putative H-bonding donors.

The k^{th} stochastic total [${}^{s}\boldsymbol{b}_{k}(\bar{x}, \bar{y})$ and ${}^{s}\boldsymbol{b}_{k}^{H}(\bar{x}, \bar{y})$] and local [${}^{s}\boldsymbol{b}_{kL}(\bar{x}_{E}, \bar{y}_{E})$, ${}^{s}\boldsymbol{b}_{kL}^{H}(\bar{x}_{E}, \bar{y}_{E})$ and ${}^{s}\boldsymbol{b}_{kL}^{H}(\bar{x}_{E\text{-}H}, \bar{y}_{E\text{-}H})$] bilinear indices were also computed.

4.3. Chemometric method

4.3.1 Linear discriminant analysis

The LDA was performed with software package STATISTICA (47). Forward stepwise was fixed as the strategy for variable selection. The quality of the models was determined by examining Wilk´s λ parameter (U-statistic), square Mahalanobis distance (D^2), Fisher ratio (F) and the corresponding p-level [p(F)], as well as the percentage in training and test sets of global good classification, Matthews' correlation coefficient, sensitivity, specificity, negative predictive value (sensitivity of the negative category) and false positive rate (false alarm rate) (37). Models with a proportion between the number of cases and variables in the equation lower than 4 were rejected.

Validation external process is necessary to ensure the quality and predictive power of the QSAR models to predict the activity of compounds that were not used for model development. In this study, the original data are divided into two series, the training and test sets. The training set is used to build the QSAR models, and these discriminant functions (DFs) are used to predict the activities of compounds in the test set. The predictivity of a model is estimated by comparing the predicted and observed classes of a sufficiently large and representative test of compounds.

4.4 Biological assay: Determination of 'in vitro' tripanosomicidals activity and cytotoxicity

4.4.1 Parasites and culture procedure

The strain-Y of *T. cruzi* (48) was originally isolated from an acute human case coming from Marília (São Paulo, Brazil) in 1950. Epimastigotes were grown at 28° C in liver infusion tryptose (LIT) broth with 10% fetal bovine serum (FBS), penicillin and streptomycin as previously described (49).

4.4.2 Epimastigotes susceptibility assay

The activity was evaluated with resazurin by a colorimetric method previously described (39). The screening assay was performed in 96-well microplates with cultures in LIT with 10% FBS, which had not reached the stationary phase. Epimastigotes were seeded at 3 x 10^6 per milliliter in culture tubes. Following a 24 h incubation to allow homogeneous growth, 200 μL volumes were seeded in the plates in the presence of serial dilutions of reference drugs (concentration range as above) at 28° C for 48 hours, at which time 20 μL of resazurin solution 3mM was added and returned to the incubator for another 5h. The solution of resazurin was prepared in 1% phosphate buffer solution (PBS) pH 7, and filter-sterilized before use. Growth controls were also included. The oxidation-reduction was quantified at 490 and 595 nm. Each concentration was assayed in triplicate. In order to avoid drawbacks, medium and drug controls were used in each test. The anti-epimastigotes percentage (%AE) was calculated as follows:

%AE = [(ALW–(AHW×RO) test well)/(ALW–(AHW×RO) positive growth control)] ×100

where, ALW and AHW represents the absorbances at the lower and the higher wavelength respectively (milieu was subtracted) and RO represents the correction factor (RO=ALW/AHW for resazurin in the milieu).

4.4.3 Cell culture

The cell lines used were National Collection of Type Cultures (NCTC) clone 929 and murine J774 macrophages. The NCTC clone 929 cells were grown in Minimal Essential Medium (Sigma) and J774 macrophages were grown in RPMI 1640 medium (Sigma). Both media were supplemented with 10% heat-inactivated FBS (30 minutes at 56ºC), penicillin G (100 U/mL) and streptomycin (100 μg/mL). For the experiments, cells in the pre-confluence phase were harvested with trypsin. Cell cultures were maintained at 37ºC in a humidified 5% CO_2 atmosphere.

4.4.4 Cytotoxicity assays

The procedure for cell viability measurement was evaluated with resazurin by a colorimetric method described previously (39, 40). The macrophages J774 were seeded (5 ×10^4 cells/well) in 96-well flat-bottom microplates with 100 μL of RPMI 1640 medium. The cells were allowed to attach for 24 h at 37ºC, 5% CO_2 and the medium was replaced by different concentrations of the drugs in 200 μL of medium, and exposed for another 24 h. Growth controls were also included. Afterwards, a volume 20 μL the 2mM resazurin solution was added and plates were returned to incubator for another 3h to evaluate cell viability. The reduction of resazurin was determined by dual wavelength absorbance measurement at 490 nm and 595 nm. Background was subtracted. Each concentration was assayed in triplicate. Medium and drug controls were used as blanks in each test.

References

1. (2002) Control of Chagas' disease. Second Report of the WHO Expert Committee. . *Control of Chagas' disease. Second Report of the WHO Expert Committee.* . Geneva: W.H.O. Tech. Rep. Ser; p. 1-109.
2. Gilbert I.H. (2002) Inhibitors of dihydrofolate reductase in leishmania and trypanosomes. *Biochim Biophys Acta;* **1587**: 249-57.
3. World_Health_Organization. (TDR/GTC/09 2007) Reporte sobre la enfermedad de Chagas. Grupo de trabajo científico. *Reporte sobre la enfermedad de Chagas. Grupo de trabajo científico*. Ginebra: World Health Organization.
4. Marin-Neto J.A., Cunha-Neto E., Maciel B.C., Simoes M.V. (2007) Pathogenesis of chronic Chagas heart disease. *Circulation* **115**: 1109-23.
5. PAHO. (2007) Health in the Americas 2007. In: Regional s.a.t.p., editor. *Health in the Americas 2007*. Washington (D.C.): Pan American Health Organization.
6. Gascón J. (2005) Diagnóstico y tratamiento de la Enfermedad de Chagas importada. *Med Clín;* **125**: 230-5.
7. Gascon J., Bern C., Pinazo M.-J. (2010) Chagas disease in Spain, the United States and other non-endemic countries. *Acta Tropica;* **115**: 22-7.
8. Schmunis G.A., Yadon Z.E. (2010) Chagas disease: A Latin American health problem becoming a world health problem. *Acta Tropica;* **115**: 14-21.
9. Dias J.P., Bastos C., Araujo E., Mascarenhas A.V., MartinsNetto E., Grassi F., et al. (2008) Acute Chagas disease outbreak associated with oral transmission. *Rev Soc Bras MedTrop;* **41**: 296-300.
10. Prata A. (2001) Clinical and epidemiological aspects of Chagas disease. *Lancet Infect Dis;* **1**: 92-100.
11. Cerecetto H., Gonzalez M. (2002) Chemotherapy of Chagas' Disease: Status and New Developments. *Curr Top Med Chem;* **2**: 1187-213.
12. Roldos V., Nakayama H., Rolón M., Montero-Torres A., Trucco F., Torres S., et al. (2008) Activity of a hydroxybibenzyl bryophyte constituent against Leishmania spp. and Trypanosoma cruzi: In silico, in vitro and in vivo activity studies. *Europ J Med Chem;* **43**: 1797-807.
13. Urbina J.A. (2002) Chemotherapy of Chagas Disease. *Current Pharm Design;* **8**: 287-95.
14. Viotti R., Vigliano C., Lococo B., Alvarez M.G., Petti M., Bertocchi G., et al. (2009) Side effects of benznidazole as treatment in chronic Chagas disease: fears and realities. *Expert Rev of Anti-Infect Ther;* **7**: 157-63.
15. Cavalli A., Bolognesi M.L. (2009) Neglected Tropical Diseases: Multi-Target-Directed Ligands in the Search for Novel Lead Candidates against Trypanosoma and Leishmania. *J Med Chem;* **52**: 7339-59.
16. DiMasi J.A., Hansen R.W., Grabowski H.G. (2003) The price of innovation: new estimates of drug development costs. *J Health Econom;* **22**: 151-85.
17. Marrero-Ponce Y., Castillo-Garit J.A., Torrens F., Romero-Zaldivar V., Castro E. (2004) Atom, Atom-Type, and Total Linear Indices of the "Molecular Pseudograph's Atom

Adjacency Matrix": Application to QSPR/QSAR Studies of Organic Compounds. *Molecules* **9**: 1100-23.

18. Marrero-Ponce Y., Meneses-Marcel A., Rivera-Borroto O., García-Domenech R., de Julián-Ortiz J., Montero A., et al. (2008) Bond-based linear indices in QSAR: computational discovery of novel anti-trichomonal compounds. *J Comput Aided Mol Des;* **22**: 523-40.
19. Marrero-Ponce Y., Torrens F., Alvarado Y.J., Rotondo R. (2006) Bond-Based Global and Local (Bond, Group and Bond-Type) Quadratic Indices and Their Applications to Computer-Aided Molecular Design. 1. QSPR Studies of Diverse Sets of Organic Chemicals. *J Comput Aided Mol Des* **20**: 685–701.
20. Marrero-Ponce Y., Khan M.T.H., Casañola-Martín G.M., Ather A., Sultankhodzhaev M.N., Torrens F., et al. (2007) Prediction of Tyrosinase Inhibition Spectra for Chemicals Using Novel Atom-Based Bilinear Indices. *ChemMedChem;* **2**: 449 – 78.
21. Castillo Garit J.A., Martinez-Santiago O., Marrero Ponce Y., Casañola-Martin G.M., Torrens F. (2008) Atom-based non-stochastic and stochastic bilinear indices: Application to QSPR/QSAR studies of organic compounds. *Chem Phys Lett;* **464**: 107-12.
22. Castillo-Garit J.A., Marrero-Ponce Y., Escobar J., Torrens F., Rotondo R. (2008) A novel approach to predict aquatic toxicity from molecular structure. *Chemosphere;* **73**: 415-27.
23. Castillo-Garit J.A., Marrero-Ponce Y., Torrens F., García-Domenech R. (2008) Estimation of ADME Properties in Drug Discovery: Predicting Caco-2 Cell Permeability Using Atom-Based Stochastic and Non-Stochastic Linear Indices. *J Pharm Sci;* **97**: 1946-76.
24. Castillo-Garit J.A., Vega M.C., Rolon M., Marrero-Ponce Y., Kouznetsov V., Torres D.F., et al. (2010) Computational discovery of novel trypanosomicidal drug-like chemicals by using bond-based non-stochastic and stochastic quadratic maps and linear discriminant analysis. *Eur J Pharm Sci;* **39**: 30-6.
25. Castillo-Garit J.A., Vega M.C., Rolón M., Marrero-Ponce Y., Gómez-Barrio A., Escario J.A., et al. (2011) Ligand-based discovery of novel trypanosomicidal drug-like compounds: In silico identification and experimental support. *Europ J Med Chem;* **46**: 3324-30.
26. Marrero-Ponce Y., Castillo-Garit J.A. (2005) 3D-chiral Atom, Atom-type, and Total Non-Stochastic and Stochastic Molecular Linear Indices and Their Applications to Central Chirality Codification. *J Comput Aided Mol Des;* **19**: 369-83.
27. Marrero-Ponce Y., Castillo-Garit J.A., Castro E.A., Torrens F., Rotondo R. (2008) 3D-Chiral (2.5) Atom-Based TOMOCOMD-CARDD Descriptors: Theory and QSAR Applications to Central Chirality Codification. *J Math Chem;* **44**: 755–86.
28. Castillo-Garit J.A., Marrero-Ponce Y., Torrens F. (2006) Atom-based 3D-chiral quadratic indices. Part 2: prediction of the corticosteroid-binding globulinbinding affinity of the 31 benchmark steroids data set. *Bioorg Med Chem;* **14**: 2398-408.
29. Castillo-Garit J.A., Marrero-Ponce Y., Torrens F., García-Domenech R., Rodríguez-Borges J.E. (2009) Applications of Bond-Based 3D-Chiral Quadratic Indices in QSAR Studies Related to Central Chirality Codification. *QSAR & Comb Sci;* **28**: 1465-77.
30. Castillo-Garit J.A., Marrero-Ponce Y., Torrens F., García-Domenech R., Romero-Zaldivar V. (2008) Bond-Based 3D-Chiral Linear Indices: Theory and QSAR Applications to Central Chirality Codification. *J Comput Chem;* **29**: 2500 - 12.
31. Castillo-Garit J.A., Marrero-Ponce Y., Torrens F., Rotondo R. (2007) Atom-based Stochastic and non-Stochastic 3D-Chiral Bilinear Indices and their Applications to Central Chirality Codification. *J Mol Graphics Model;* **26**: 32-47.
32. Estrada E., Peña A. (2000) In Silico Studies for the Rational Discovery of Anticonvulsant Compounds. *Bioorg Med Chem;* **8**: 2755-70.

33. Estrada E., Uriarte E., Montero A., Teijeira M., Santana L., De Clercq E. (2000) A Novel Approach for the Virtual Screening and Rational Design of Anticancer Compounds. *J Med Chem;* **43**: 1975-85.
34. Marrero-Ponce Y., Castillo-Garit J.A., Olazabal E., Serrano H.S., Morales A., Castanedo N., et al. (2005) Atom, Atom-Type and Total Molecular Linear Indices as a Promising Approach for Bioorganic and Medicinal Chemistry: Theoretical and Experimental Assessment of a Novel Method for Virtual Screening and Rational Design of New Lead Anthelmintic. *Bioorg Med Chem;* **13**: 1005-20.
35. Marrero-Ponce Y., Meneses-Marcel A., Castillo-Garit J.A., Machado-Tugores Y., Escario J.A., Gómez-Barrio A., et al. (2006) Predicting Antitrichomonal Activity: A Computational Screening Using Atom-Based Bilinear Indices and Experimental Proofs. *Bioorg Med Chem* **14**: 6502–24.
36. Estrada E. (1999) Novel Strategies in the Search of Topological Indices. In: Devillers J., Balaban A.T., editors *Novel Strategies in the Search of Topological Indices*. Amsterdam: Gordon and Breach: p. 403–53.
37. Baldi P., Brunak S., Chauvin Y., Andersen C.A., Nielsen H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics;* **16**: 412-24.
38. Tropsha A., Gramatica P., Gombar V.K. (2003) The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR & Comb Sci;* **22**: 69-77.
39. Rolón M., Vega C., Escario J., Gómez-Barrio A. (2006) Development of resazurin microtiter assay for drug sensibility testing of Trypanosoma cruzi epimastigotes. *Parasitol Res;* **99**: 103-7.
40. Rolón M., Seco E.M., Vega C., Nogal J.J., Escario J.A., Gómez-Barrio A., et al. (2006) Selective activity of polyene macrolides produced by genetically modified Streptomyces on Trypanosoma cruzi. *Int J Antimicrob Agents;* **28**: 104-9.
41. Meneses-Marcel A., Rivera-Borroto O.M., Marrero-Ponce Y., Montero A., Machado Tugores Y., Escario J.A., et al. (2008) New Antitrichomonal Drug-like Chemicals Selected by Bond (Edge)-Based TOMOCOMD-CARDD Descriptors. *J Biomol Screen;* **13**: 785-94.
42. Marrero-Ponce Y., Romero V. (2002) TOMOCOMD-CARDD software. TOMOCOMD (TOpological MOlecular COMputer Design) for Windows, version 1.0 is a preliminary experimental version; in future a professional version can be obtained upon request to Y. Marrero: yovanimp@uclv.edu.cu or ymarrero77@yahoo.es
43. Kier L.B., Hall L.H. (1986) *Molecular Connectivity in Structure–Activity Analysis*. Letchworth, U. K: Research Studies Press.
44. Todeschini R., Gramatica P. (1998) New 3D Molecular Descriptors: the WHIM Theory and QSAR Applications. *Perspect Drug Disc Des;* **9-11**: 355–80.
45. Consonni V., Todeschini R., Pavan M. (2002) Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. *J Chem Inf Comput Sci;* **42**: 682-92.
46. Pauling L. (1939) *The Nature of Chemical Bond*. Ithaca (New York): Cornell University Press
47. STATISTICA version. 6.0 (2001) **StatSoft,**: Tulsa.
48. Silva L.H., Nussensweig V. (1953) Sobre uma cepa Trypanosoma cruzi virulenta para o camundongo branco. *Folia of Clinical Biology;* **20**: 191-207.
49. Gómez-Barrio A., Martínez-Díaz R.A., Atienza J., Escario J.A., Diego C., Avendaño C. (1997) New derivatives of gentian violet as trypanocides: In vitro and in vivo assays on Trypanosoma cruz. *Res Rev Parasitol;* **57**: 25-31.