Available online at www.sciencedirect.com

ScienceDirect

Bioorganic & Medicinal Chemistry 14 (2006) 6502–6524

Bioorganic & Medicinal Chemistry

Predicting antitrichomonal activity: A computational screening using atom-based bilinear indices and experimental proofs

Yovani Marrero-Ponce,[a,b,c,*] Alfredo Meneses-Marcel,[d,e] Juan A. Castillo-Garit,[b,c,f] Yanetsy Machado-Tugores,[b,c,d] José Antonio Escario,[e] Alicia Gómez Barrio,[e] David Montero Pereira,[e] Juan José Nogal-Ruiz,[e] Vicente J. Arán,[g] Antonio R. Martínez-Fernández,[e] Francisco Torrens,[a] Richard Rotondo,[h] Froylán Ibarra-Velarde[i] and Ysaias J. Alvarado[j]

[a]*Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, Poligon la Coma s/n (detras de Canal Nou), PO Box 22085, E-46071 Valencia, Spain*

[b]*Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba*

[c]*Department of Drug Design, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba*

[d]*Department of Parasitology, Chemical Bioactive Center, Central University of Las Villas, 54830 Villa Clara, Cuba*

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

[f]*Applied Chemistry Research Center, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba*

[g]*Instituto de Química Médica, CSIC, c/ Juan de la Cierva 3, 28006 Madrid, Spain*

[h]*Mediscovery, Inc. Suite 1050, 601 Carlson Parkway, Minnetonka, MN 55305, USA*

[i]*Department of Parasitology, Faculty of Veterinarian Medicinal and Zootecnic, UNAM, Mexico DF 04510, Mexico*

[j]*Laboratorio de Electrónica Molecular, Departamento de Química, Modulo II, grano de Oro, Facultad Experimental de Ciencias, La Universidad del Zulia (LUZ), Venezuela*

Received 10 January 2006; revised 6 June 2006; accepted 8 June 2006
Available online 27 July 2006

Abstract—Existing *Trichomonas vaginalis* therapies are out of reach for most trichomoniasis people in developing countries and, where available, they are limited by their toxicity (mainly in pregnant women) and their cost. New antitrichomonal agents are needed to combat emerging metronidazole-resistant trichomoniasis and reduce the side effects associated with currently available drugs. Toward this end, atom-based bilinear indices, a new TOMOCOMD-CARDD molecular descriptor, and linear discriminant analysis (LDA) were used to discover novel, potent, and non-toxic lead trichomonacidal chemicals. Two discriminant functions were obtained with the use of non-stochastic and stochastic atom-type bilinear indices for heteroatoms and H-bonding of heteroatoms. These atomic-level molecular descriptors were calculated using a weighting scheme that includes four atomic labels, namely atomic masses, van der Waals volumes, atomic polarizabilities, and atomic electronegativities in Pauling scale. The obtained LDA-based QSAR models, using non-stochastic and stochastic indices, were able to classify correctly 94.51% (90.63%) and 93.41% (93.75%) of the chemicals in training (test) sets, respectively. They showed large Matthews' correlation coefficients (***C***); 0.89 (0.79) and 0.87 (0.85), for the training (test) sets, correspondingly. The result of predictions on the 15% full-out cross-validation test also evidenced the robustness and predictive power of the obtained models. In addition, canonical regression analyses corroborated the statistical quality of these models (R_{can} of 0.749 and of 0.845, correspondingly); they were also used to compute biological activity canonical scores for each compound. On the other hand, a close inspection of the molecular descriptors included in both equations showed that several of these molecular fingerprints are strongly interrelated with each other. Therefore, these models were orthogonalized using the Randić orthogonalization procedure. These classification functions were then applied to find new lead antitrichomonal agents and six compounds were selected as possible active compounds by computational screening. The designed compounds were synthesized and tested for in vitro activity against *T. vaginalis*. Out of the six compounds that were designed, and synthesized, three molecules (chemicals VA5-5a, VA5-5c, and VA5-12b) showed high to moderate cytocidal activity at the

Keywords: TOMOCOMD-CARDD software; Atom-based bilinear index; LDA-based QSAR model; Computational screening; Lead generation; Trichomonacidal; Cytostatic and cytocidal activity.

* Corresponding author. Tel.: +53 42 281192; fax: +53 42 281130; e-mail: ymarrero77@yahoo.es
URL: http://www.uv.es/yoma.

doi:10.1016/j.bmc.2006.06.016

concentration of 10 μg/ml, other two compounds (VA5-8pre and VA5-8) showed high cytocidal and cytostatic activity at the concentration of 100 μg/ml and 10 μg/ml, correspondingly, and the remaining chemical (compound VA5-5e) was inactive at these assayed concentrations. Nonetheless, these compounds possess structural features not seen in known trichomonacidal compounds and thus can serve as excellent leads for further optimization of antitrichomonal activity. The LDA-based QSAR models presented here can be considered as a computer-assisted system that could potentially significantly reduce the number of synthesized and tested compounds and increase the chance of finding new chemical entities with antitrichomonal activity.

1. Introduction

Trichomonas vaginalis (Tv) is a common sexually transmitted infection that is increasingly recognized as an important infection in women and men.[1,2] Recent data have shown that the annual incidence of trichomoniasis is more than 170 million cases worldwide.[3] In North America alone, more than eight million new cases are reported yearly,[3] with an estimated rate of asymptomatic cases as high as 50%.[4,5]

Tv is recognized as a common cause of vaginitis[6] as well as a factor contributing to preterm birth and low birth weight.[7] Tv infections have also been linked with increased human immunodeficiency virus transmission.[8–11]

Metronidazole has been the drug of choice for treating trichomoniasis since 1959 and is currently the only drug licensed for this purpose in the United States. The recommended metronidazole regimen results in cure rates of approximately 95%.[12] Metronidazole enters the cell through diffusion[13] and is activated in the hydrogenosomes of Tv.[14] Here, the nitro group of the drug is anaerobically reduced by pyruvate-ferredoxin oxidoreductase.[14] This results in cytotoxic nitro radical–ion intermediates that break the DNA strands.[15] The response is rapid: cell division and motility cease within 1 h and cell death occurs within 8 h as seen in cell culture.[16]

Although there are clinical reports[17–24] that document the refractoriness of infections with Tv to treatment with metronidazole, susceptibility tests have failed to demonstrate conclusively that the parasites isolated from such cases after treatment were resistant to this drug.[25,26] Thus, the resistance of Tv has not been generally accepted as the factor responsible for failure of metronidazole therapy,[27] since reinfection, irregular medication, poor absorption of the drug, and its inactivation by the vaginal flora have not been excluded.[26,28,29] However, a strain of Tv, unequivocally resistant to metronidazole, was recently isolated from a female patient who had not responded to two courses of treatment with this agent. The current report is concerned with the isolation of this strain and its in vitro and in vivo susceptibilities to metronidazole and other 5-nitroimidazole derivatives.[30] Therefore, new antitrichomonal agents are needed to combat emerging metronidazole-resistant trichomoniasis; they reduce the side effects associated with currently available drugs. However, the great cost associated with the development of new drugs and the small economic size of the market for this type of antiprotozoal agents make this development slow.

On the other hand, cheminformatics has become an independent discipline by itself. For pharmaceutical research and development (R&D), this discipline provides the tools for the identification/selection and design/optimization of compounds with improved drug (and/or lead)-like qualities—often reducing the number of tested compounds, compared with conventional trial-and-error methods.[31–34] Although pharmaceutical companies are highly motivated to reduce the discovery-to-market time and cost, an increase in R&D dollars dedicated to the business of discovering new therapeutics has not resulted in a correspondingly increased number of successful drugs on the market.[31–34] Therefore, the development of a novel computational method is currently required to deliver a system that significantly reduces the time-to-market and R&D overheads, and increases the rate at which novel chemical entities (NCEs) progress through the pipeline.[31–34] Such studies, if they are successfully implemented, deliver substantial benefits and act as the bedrock for NCE selection.

Our research group has recently developed simple non-stochastic and stochastic atom- and bond-based molecular descriptors (MDs) based on algebraic theory. They have been defined by analogy with the quadratic, linear, and bilinear mathematical maps.[35–40] Applications included the prediction of several physical, physicochemical, chemical, pharmacokinetical, and pharmacological properties of organic compounds.[40–49]

Taking into consideration that mentioned above, the aims of the present paper were: (1) to use a new molecular descriptor family, atom-based non-stochastic and stochastic bilinear indices, in the generation of discriminant functions by linear discriminant analysis (LDA) that permits the classification of chemicals (antitrichomonal and non-antitrichomonal drug-like compounds) on a data set drawn from the literature, (2) to assess the 'biosilico' models by the use of different validation tests, (3) to develop a virtual screening of some libraries in order to identify potential novel chemical entities (NCEs) and, (4) to evaluate the in vitro antitrichomonal activity of the best candidates selected from thousands of chemicals in the virtual-computational-screening process.

2. Method

In earlier publications, we outlined outstanding features concerned with the theory of 2D atom-based TOMOCOMD-CARDD MDs.[36–49] This method codifies molecular structure by means of mathematical quadratic, linear and bilinear transformations. In order to calculate these algebraic maps for a molecule, the atom-based molecular vector, $\bar{x}$ (vector representation), as well as kth 'non-stochastic and stochastic graph–theoretical electronic-density matrices' $\mathbf{M}^k$ and $\mathbf{S}^k$ (matrix representation), correspondingly, are constructed.[35–40] Such atom-adjacency relationships and chemical-information codification will be applied in the present study to generate a series of atom-based MDs, atom, group, and atom-type as well as total bilinear indices, to be used in drug design and chemoinformatic studies.

Therefore, the structure of this section will be as follows: (1) a background in atom-based molecular vector as well as non-stochastic and stochastic graph–theoretical electronic-density matrices will be described in the next subsections (2.1 and 2.2, respectively), and (2) an outline of the mathematical definition of bilinear maps and a definition of our procedures will be developed in Sections 2.3 and 2.4, correspondingly.

2.1. Chemical information and atom-based molecular vector

The atom-based molecular vector ($\bar{x}$), used to represent small-to-medium size organic chemicals, has been explained elsewhere in some detail.[35–40] The components (x) of $\bar{x}$ are numerical values, which represent a certain standard atomic property (atomic labels). Therefore, these weights correspond to different atomic properties for organic molecules. Thus, a molecule having $5, 10, 15, \ldots, n$ atomic nuclei can be represented by means of vectors with $5, 10, 15, \ldots, n$ components, respectively, belonging to the spaces $\mathfrak{R}^5, \mathfrak{R}^{10}, \mathfrak{R}^{15}, \ldots, \mathfrak{R}^n$, where n is the dimension of the real set ($\mathfrak{R}^n$). Therefore, $\bar{x}$ is the n-dimensional vector property of the atoms (atomic nuclei) in a molecule.

This approach allows us to encode organic molecules such as 3-mercaptopyridine-4-carbaldehyde through the molecular vector $\bar{x} = [x_{N1}, x_{C2}, x_{C3}, x_{C4}, x_{C5}, x_{C6}, x_{C7}, x_{O8}, x_{S9}]$ (see also Table 1 for molecular structure). This vector belongs to the product space $\mathfrak{R}^9$. However, diverse kinds of atomic weights (x) can be used for codifying information related to each atomic nucleus in the molecule. These atomic labels are chemically meaningful numbers or their contributions derived by atom-to-atom analysis such as atomic log P,[50] surface contributions of polar atoms,[51] atomic molar refractivities,[52] atomic hybrid polarizabilities,[53] Gasteiger–Marsilli atomic charges,[54] atomic masses (M),[55] van der Waals volumes (V),[55] atomic polarizabilities (P),[55] atomic electronegativities (K) in Pauling scale,[56] and so on.

Now, if we are interested in codifying the chemical information by means of two different molecular vectors, for instance, $\bar{x} = [x_1, \ldots, x_n]$ and $\bar{y} = [y_1, \ldots, y_n]$, then different combinations of molecular vectors ($\bar{x} \neq \bar{y}$) are possible when a weighting scheme is used. In the present report, we characterized each atomic nucleus with the following parameters: atomic masses (M),[55] van der Waals volumes (V),[55] atomic polarizabilities (P),[55] and atomic Pauling electronegativities (E).[56] The values of these atomic labels are shown in Table 2. From this weighting scheme, six (or twelve if $\bar{x}_M - \bar{y}_V \neq \bar{x}_V - \bar{y}_M$) combinations (pairs) of molecular vectors ($\bar{x}, \bar{y}; \bar{x} \neq \bar{y}$) can be computed, $\bar{x}_M - \bar{y}_V$, $\bar{x}_M - \bar{y}_P$, $\bar{x}_M - \bar{y}_K$, $\bar{x}_V - \bar{y}_P$, $\bar{x}_V - \bar{y}_K$, and $\bar{x}_P - \bar{y}_K$. Here, we used the symbols $\bar{x}_W - \bar{y}_Z$, where the subscripts $_W$ and $_Z$ mean two different atomic properties from our weighting scheme and a hyphen (-) expresses the combination (pair) of two selected atom-label chemical properties. In order to illustrate this, let us consider the same organic molecule as in the example above (3-mercaptopyridine-4-carbaldehyde) and the following weighting scheme: M and V ($\bar{x}_M - \bar{y}_V = \bar{x}_V - \bar{y}_M$). The following molecular vectors, $\bar{x} = [14.01, 12.01, 12.01, 12.01, 12.01, 12.01, 12.01, 16.0, 32.07]$ and $\bar{y} = [15.599, 22.449, 22.449, 22.449, 22.449, 22.449, 22.449, 11.494, 24.429]$, are obtained when we use M and V as chemical weights for codifying each atom in the example molecule in $\bar{x}$ and $\bar{y}$ vectors, respectively.

2.2. Background in non-stochastic and stochastic graph–theoretical electronic-density matrices

In molecular topology, molecular structure is expressed, generally, by the hydrogen-suppressed graph. Therefore, a molecule is represented by a graph. Informally, a graph G is a collection of vertices (points) and edges (lines or bonds) connecting these vertices.[35–40] In more formal terms, a simple graph G is defined as an ordered pair $[V(G), E(G)]$, which consists of a nonempty set of vertices $V(G)$ and a set $E(G)$ of unordered pairs of elements of $V(G)$, called edges.[57–59] In this particular case, we are not dealing with a simple graph but with a so-called pseudograph (G). Informally, a pseudograph is a graph with multiple edges or loops between the same vertices or the same vertex. Formally, a pseudograph is a set V of vertices along a set E of edges, and a function f from E to $\{\{u,v\} \mid u,v \text{ in V}\}$ (The f function shows which pair of vertices is connected by which edge). An edge is a loop if $f(e) = \{u\}$ for some vertex u in V.[35,36,60]

In earlier reports we have introduced new molecular matrices that describe changes for some time in the electronic distribution throughout the molecular backbone. The $n \times n$ kth non-stochastic graph–theoretical electronic-density matrix of the molecular pseudograph (G), $\mathbf{M}^k$, is a symmetric square matrix, where n is the number of atoms (atomic nuclei) in the molecule.[35–40] The coefficients $^k m_{ij}$ are the elements of the kth power of $\mathbf{M}(G)$ and are defined as follows:

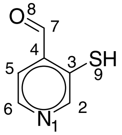
$$\begin{aligned} m_{ij} &= P_{ij} \text{ if } i \neq j \text{ and } \exists e_k \in E(G) \\ &= L_{ii} \text{ if } i = j \\ &= 0 \text{ otherwise} \end{aligned} \qquad (1)$$

where $E(G)$ represents the set of edges of G. P_{ij} is the number of edges (bonds) between vertices (atomic nuclei) v_i and v_j; L_{ii} is the number of loops in v_i.

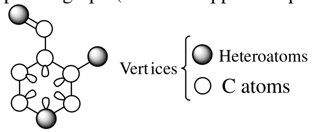
The elements $m_{ij} = P_{ij}$ of such a matrix represent the number of chemical bonds between an atomic nucleus i and other j. The matrix $\mathbf{M}^k$ provides the numbers of walks of length k that links every pair of vertices v_i and v_j. For this reason, each edge in $\mathbf{M}^1$ represents 2

Table 1. (A) Chemical structure of 3-mercapto-pyridine-4-carbaldehyde and its labeled molecular pseudograph, *G*; (B and C) the zero (k = 0), first (k = 1), second (k = 2), and third (k = 3) powers of the non-stochastic and stochastic graph–theoretical electronic-density matrices of *G*, respectively

A) molecular structure



molecular pseudograph (H-atom-suppressed pseudograph)[a]



B) k^{th} non-stochastic graph–theoretical electronic-density matrices, $\mathbf{M}^k$ (k = 0-3)

zero order (k = 0)

$$\begin{bmatrix}1&0&0&0&0&0&0&0&0\\0&1&0&0&0&0&0&0&0\\0&0&1&0&0&0&0&0&0\\0&0&0&1&0&0&0&0&0\\0&0&0&0&1&0&0&0&0\\0&0&0&0&0&1&0&0&0\\0&0&0&0&0&0&1&0&0\\0&0&0&0&0&0&0&1&0\\0&0&0&0&0&0&0&0&1\end{bmatrix}$$

first order (k = 1)

$$\begin{bmatrix}1&1&0&0&0&1&0&0&0\\1&1&1&0&0&0&0&0&0\\0&1&1&1&0&0&0&0&1\\0&0&1&1&1&0&1&0&0\\0&0&0&1&1&1&0&0&0\\1&0&0&0&1&1&0&0&0\\0&0&0&1&0&0&0&2&0\\0&0&0&0&0&0&2&0&0\\0&0&1&0&0&0&0&0&0\end{bmatrix}$$

second order (k = 2)

$$\begin{bmatrix}3&2&1&0&1&2&0&0&0\\2&3&2&1&0&1&0&0&1\\1&2&4&2&1&0&1&0&1\\0&1&2&4&2&1&1&2&1\\1&0&1&2&3&2&1&0&0\\2&1&0&1&2&3&0&0&0\\0&0&1&1&1&0&5&0&0\\0&0&0&2&0&0&0&4&0\\0&1&1&1&0&0&0&0&1\end{bmatrix}$$

third order (k = 3)

$$\begin{bmatrix}7&6&3&2&3&6&0&0&1\\6&7&7&3&2&3&1&0&2\\3&7&9&8&3&2&2&2&4\\2&3&8&9&7&3&8&2&2\\3&2&3&7&7&6&2&2&1\\6&3&2&3&6&7&1&0&0\\0&1&2&8&2&1&1&10&1\\0&0&2&2&2&0&10&0&0\\1&2&4&2&1&0&1&0&1\end{bmatrix}$$

C) stochastic graph–theoretic electronic-density matrices, $\mathbf{S}^k$ (k = 1-3)[b,c]

first order (k = 1)

$$\begin{bmatrix}0.3333&0.3333&0&0&0&0.3333&0&0&0\\0.3333&0.3333&0.3333&0&0&0&0&0&0\\0&0.25&0.25&0.25&0&0&0&0&0.25\\0&0&0.25&0.25&0.25&0&0.25&0&0\\0&0&0&0.3333&0.3333&0.3333&0&0&0\\0.3333&0&0&0&0.3333&0.3333&0&0&0\\0&0&0&0.3333&0&0&0&0.6666&0\\0&0&0&0&0&0&1&0&0\\0&0&1&0&0&0&0&0&0\end{bmatrix}$$

second order (k = 2)

$$\begin{bmatrix}0.3333&0.2222&0.1111&0&0.1111&0.2222&0&0&0\\0.2&0.3&0.2&0.1&0&0.1&0&0&0.1\\0.0833&0.166&0.3333&0.1666&0.0833&0&0.0833&0&0.0833\\0&0.0714&0.1429&0.2857&0.1429&0.0714&0.0714&0.1429&0.0714\\0.1&0&0.1&0.2&0.3&0.2&0.1&0&0\\0.2222&0.1111&0&0.1111&0.2222&0.3333&0&0&0\\0&0&0.125&0.125&0.125&0&0.625&0&0\\0&0&0&0.3333&0&0&0&0.6666&0\\0&0.25&0.25&0.25&0&0&0&0&0.25\end{bmatrix}$$

third order (k = 3)

$$\begin{bmatrix}0.25&0.2142&0.1071&0.0714&0.1071&0.2143&0&0&0.0357\\0.1935&0.2258&0.2258&0.0967&0.0645&0.0967&0.0323&0&0.0645\\0.075&0.175&0.225&0.2&0.075&0.05&0.05&0.05&0.1\\0.0455&0.0682&0.1818&0.2045&0.1591&0.0682&0.1818&0.0455&0.0455\\0.0909&0.0606&0.0909&0.2121&0.2121&0.1818&0.0606&0.0606&0.0303\\0.2143&0.1071&0.0714&0.1971&0.2143&0.25&0.0357&0&0\\0&0.0385&0.0769&0.3076&0.0769&0.0385&0.0385&0.3846&0.0385\\0&0&0.125&0.125&0.125&0&0.625&0&0\\0.0833&0.1666&0.333&0.1666&0.0833&0&0.0833&0&0.0833\end{bmatrix}$$

[a] Each edge in the pseudograph of $\mathbf{M_1}$ represents 2 electrons belonging to the covalent bond between atoms (vertices) v_i and v_j, for example, the inputs of $\mathbf{M_1}$ are equal to 1, 2 or 3 when single, double or triple bonds, correspondingly, appear between vertices v_i and v_j. The presence of pi (π) electrons in aromatic systems such as benzene is accounted by means of loops in each atom of the aromatic ring. Therefore, the $\mathbf{M_1}$ matrix considers all valence-bond electrons (σ- and π-networks) in one step and their powers ($k = 0, 1, 2, 3, \ldots$) can be considered as an interacting-electron chemical-network model in the k step.

[b] The zero power (k = 0) of the stochastic graph-theoretical electronic-density matrix, $\mathbf{S}^0$, coincides with the non-stochastic matrix one ($\mathbf{M}^0 = \mathbf{S}^0$).

[c] The values of the elements of kth matrices $\mathbf{S}^k$ (${}^{k}s_{ij}$) have been rounded.

Table 2. Values of the atomic weights used for bilinear indices calculation

ID	Atomic Mass	VdW[a] Volume ($Å^3$)	Polarizability ($Å^3$)	Pauling Electronegativity
H	1.01	6.709	0.667	2.2
B	10.81	17.875	3.030	2.04
C	12.01	22.449	1.760	2.55
N	14.01	15.599	1.100	3.04
O	16.00	11.494	0.802	3.44
F	19.00	9.203	0.557	3.98
Al	26.98	36.511	6.800	1.61
Si	28.09	31.976	5.380	1.9
P	30.97	26.522	3.630	2.19
S	32.07	24.429	2.900	2.58
Cl	35.45	23.228	2.180	3.16
Fe	55.85	41.052	8.400	1.83
Co	58.93	35.041	7.500	1.88
Ni	58.69	17.157	6.800	1.91
Cu	63.55	11.494	6.100	1.9
Zn	65.39	38.351	7.100	1.65
Br	79.90	31.059	3.050	2.96
Sn	118.71	45.830	7.700	1.96
I	126.90	38.792	5.350	2.66

[a] VdW: van der Waals.

electrons belonging to the covalent bond between atomic nuclei i and j; for example, the inputs of $\mathbf{M}^1$ are equal to 1, 2 or 3 when single, double or triple bonds, correspondingly, appear between vertices v_i and v_j. On the other hand, molecules containing aromatic rings with more than one canonical structure are represented by pseudographs. It happens for substituted aromatic compounds such as pyridine, naphthalene, quinoline, and so on, where the presence of pi (π) electrons is accounted by means of loops in each atomic nucleus of the aromatic ring. Conversely, aromatic rings having only one canonical structure, such as furan, thiophene, and pyrrole, are represented by a multigraph. In order to illustrate the calculation of these matrices, let us consider the same molecule selected in the previous section. Table 1 depicts the molecular structure of this compound and its labeled molecular pseudograph. The zero ($k = 0$), first ($k = 1$), second ($k = 2$), and third ($k = 3$) powers of the non-stochastic graph–theoretical electronic-density matrices are also given in this table.

As it can be seen, $\mathbf{M}^k$ is graph–theoretical electronic-structural model, like an 'extended Hückel theory (EHT) model.' The $\mathbf{M}^1$ matrix considers all valence-bond electrons (σ- and π-networks) in one step and its powers ($k = 0, 1, 2, 3, \ldots$) can be considered as interacting-electron chemical-network models in the k step. The complete model can be seen as an intermediate between the quantitative quantum-mechanical Schrödinger equation and classical chemical-bonding ideas.[61]

The present approach is based on a simple model for the intramolecular movement of all outer-shell electrons. Let us consider a hypothetical situation in which a set of atoms is free in space at an arbitrary initial time (t_0). At this time, the electrons are distributed around the atomic nuclei. Alternatively, these electrons can be distributed around cores at discrete intervals of time t_k. In this sense, the electron in an arbitrary atom i can move (step-by-step) to other atoms at different discrete time periods t_k ($k = 0, 1, 2, 3, \ldots$) throughout the chemical-bonding network.

On the other hand, the kth stochastic graph–theoretical electronic-density matrix of G, $\mathbf{S}^k$, can be directly obtained from $\mathbf{M}^k$. Here, $\mathbf{S}^k = [{}^k\mathbf{s}_{ij}]$ is a square matrix of order n (n = number of atomic nuclei) and the elements ${}^k\mathbf{s}_{ij}$ are defined as follows:[35–40]

$$ {}^k s_{ij} = \frac{{}^k m_{ij}}{{}^k\mathrm{SUM}_i} = \frac{{}^k m_{ij}}{{}^k\delta_i} \tag{2}$$

where ${}^k m_{ij}$ are the elements of the kth power of $\mathbf{M}$ and the SUM of the ith row of $\mathbf{M}^k$ are named the k-order vertex degree of atom i, ${}^k\delta_i$. It should be remarked that the matrix $\mathbf{S}^k$ in Eq. 2 has the property that the sum of the elements in each row is 1. An $n \times n$ matrix with non-negative entries having this property is called a 'stochastic matrix.'[62] The kth s_{ij} elements are the transition probabilities with the electrons moving from atom i to j in the discrete time periods t_k. It should be also pointed out that the kth element s_{ij} takes into consideration the molecular topology in the k step throughout the chemical-bonding (σ- and π-) network. In this sense, the ${}^2s_{ij}$ values can distinguish between hybrid states of atoms in bonds. For instance, the self-return probability of second order (${}^2s_{ii}$) [i.e., the probability with which an electron returns to its original atom at t_2] varies regularly according to the different hybrid states of atom i in the molecule, for example, an electron will have a higher probability of returning to the sp C atom than to the sp^2 (or sp^3) C atom in $t_2[p(\mathrm{C_{sp}}) > p(\mathrm{C_{sp^2}}) > p(\mathrm{C\ _{sp^2arom}}) > p(\mathrm{C_{sp^3}})]$ (see Table 1 for more details).

This is a logical result if the electronegativity scale of these hybrid states is taken into account.

2.3. Mathematical bilinear forms: a theoretical framework

In mathematics, a bilinear form in a real vector space is a mapping $b : VxV \rightarrow \mathfrak{R}$, which is linear in both arguments.[63–68] Therefore, this function satisfies the following axioms for any scalar α and any choice of vectors $\bar{v}, \bar{w}, \bar{v}_1, \bar{v}_2, \bar{w}_1$, and $\bar{w}_2$.

(i) $b(\alpha\bar{v}, \bar{w}) = b(\bar{v}, \alpha\bar{w}) = \alpha b(\bar{v}, \bar{w})$
(ii) $b(\bar{v}_1 + \bar{v}_2, \bar{w}) = b(\bar{v}_1, \bar{w}) + b(\bar{v}_2, \bar{w})$
(iii) $b(\bar{v}, \bar{w}_1 + \bar{w}_2) = b(\bar{v}, \bar{w}_1) + b(\bar{v}, \bar{w}_2)$

That is, b is *bilinear* if it is linear in each parameter, taken separately.

Let V be a real vector space in $\mathfrak{R}^n (V \in \mathfrak{R}^n)$ and consider that the following vector set, $\{\bar{e}_1, \bar{e}_2, \ldots, \bar{e}_n\}$, is a basis set of $\mathfrak{R}^n$. This basis set permits us to write in unambiguous form any vectors $\bar{x}$ and $\bar{y}$ of V, where $(x^1, x^2, \ldots, x^n) \in \mathfrak{R}^n$ and $(y^1, y^2, \ldots, y^n) \in \mathfrak{R}^n$ are the coordinates of the vectors $\bar{x}$ and $\bar{y}$, respectively. Therefore,

$$\bar{x} = \sum_{i=1}^{n} x^i \bar{e}_i \tag{3}$$

and,

$$\bar{y} = \sum_{j=1}^{n} y^j \bar{e}_j \tag{4}$$

Subsequently,

$$b(\bar{x}, \bar{y}) = b(x^i \bar{e}_i, y^j \bar{e}_j) = x^i y^j b(\bar{e}_i, \bar{e}_j) \tag{5}$$

If we take a_{ij} as the $n \times n$ scalars $b(\bar{e}_i, \bar{e}_j)$, that is,

$$a_{ij} = b(\bar{e}_i, \bar{e}_j), \text{ to } i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, n \tag{6}$$

Then,

$$b(\bar{x}, \bar{y}) = \sum_{i,j}^{n} a_{ij} x^i y^j = [X]^{\mathrm{T}} A [Y] = [x^1 \ldots x^n] \begin{bmatrix} a_{11} & \ldots & a_{jn} \\ \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{nn} \end{bmatrix} \begin{bmatrix} y^1 \\ \vdots \\ y^n \end{bmatrix} \tag{7}$$

As it can be seen, the defined system of equations for b may be written in matrix form (see Eq. 7), where $[Y]$ is a column vector (an $nx1$ matrix) of the coordinates of $\bar{y}$ in a basis set of $\mathfrak{R}^n$, and $[X]^{\mathrm{T}}$ (a $1xn$ matrix) is the transpose of $[X]$, where $[X]$ is a column vector (an $nx1$ matrix) of the coordinates of $\bar{x}$in the same basis set of $\mathfrak{R}^n$.

Finally, we introduce the formal definition of symmetric bilinear form. Let V be a real vector space and b be a bilinear function in $V \times V$. The bilinear function b is called symmetric if $b(\bar{x}, \bar{y}) = b(\bar{y}, \bar{x}), \forall \bar{x}, \bar{y} \in V$.[74,75,32,76–78] Then,

$$b(\bar{x}, \bar{y}) = \sum_{i,j}^{n} a_{ij} x^i y^j = \sum_{i,j}^{n} a_{ji} x^j y^i = b(\bar{y}, \bar{x}) \tag{8}$$

2.4. Non-stochastic and stochastic atom-based bilinear indices: total definition

The kth non-stochastic and stochastic bilinear indices for a molecule, $b_k(\bar{x}, \bar{y})$ and ${}^{s}b_k(\bar{x}, \bar{y})$, respectively, are computed from these kth non-stochastic and stochastic graph–theoretical electronic-density matrices, $\mathbf{M}^{\boldsymbol{k}}$ and $\mathbf{S}^{\boldsymbol{k}}$, as shown in Eqs. 9 and 10:

$$b_k(\bar{x}, \bar{y}) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^{k}m_{ij} x^i y^j \tag{9}$$

$${}^{s}b_k(\bar{x}, \bar{y}) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^{k}s_{ij} x^i y^j \tag{10}$$

where n is the number of atoms in the molecule, and $x^1, \ldots, x^n$ and $y^1, \ldots, y^n$ are the coordinates or components of the molecular vectors $\bar{x}$ and $\bar{y}$ in a canonical basis set of $\mathfrak{R}^n$.

The defined Eqs. 9 and 10 for $b_k(\bar{x}, \bar{y})$ and ${}^{s}b_k(\bar{x}, \bar{y})$ may also be written as the single matrix equations:

$$b(\bar{x}, \bar{y}) = [X]^{\mathrm{T}} \mathbf{M}^{\boldsymbol{k}} [Y] \tag{11}$$

$${}^{s}b(\bar{x}, \bar{y}) = [X]^{\mathrm{T}} \mathbf{S}^{\boldsymbol{k}} [Y] \tag{12}$$

where $[Y]$ is a column vector (an $nx1$ matrix) of the coordinates of $\bar{y}$ in the canonical basis set of $\mathfrak{R}^n$, and $[X]^{\mathrm{T}}$ is the transpose of $[X]$, where $[X]$ is a column vector (an $nx1$ matrix) of the coordinates of $\bar{x}$ in the canonical basis set of $\mathfrak{R}^n$. Therefore, if we use the canonical basis set, the coordinates $[(x^1, \ldots, x^n)$ and $(y^1, \ldots, y^n)]$ of any molecular vectors ($\bar{x}$ and $\bar{y}$) coincide with the components of those vectors $[(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)]$. Therefore, those coordinates can be considered as the weights (atomic labels) of the vertices of the molecular pseudograph, due to the fact that the components of the molecular vectors are values of some atomic property that characterizes each kind of atomic nuclei in the molecule.

It should be remarked that non-stochastic and stochastic bilinear indices are symmetric and non-symmetric bilinear forms, respectively. Therefore, if in the following weighting scheme, M and V are used as atomic weights to compute these MDs, two different sets of stochastic bilinear indices, ${}^{M\text{-}V s}\boldsymbol{b}_{\boldsymbol{k}}^{\mathbf{H}}(x, y)$ and ${}^{V\text{-}M s}\boldsymbol{b}_{\boldsymbol{k}}^{\mathbf{H}}(x, y)$ [because $\bar{x}_M - \bar{y}_V \neq \bar{x}_V - \bar{y}_M$], can be obtained and only one group of non-stochastic bilinear indices $[{}^{M\text{-}V s}\boldsymbol{b}_{\boldsymbol{k}}^{\mathbf{H}}(x, y) = {}^{V\text{-}\ \mathrm{M}s}\boldsymbol{b}_{\boldsymbol{k}}^{\mathbf{H}}(x, y)$ because in this case $\bar{x}_M - \bar{y}_V = \bar{x}_V - \bar{y}_M]$ can be calculated.

2.5. Non-stochastic and stochastic atom-based bilinear indices: local (atomic, group, and atom-type) definition

In the last decade, Randić[69] proposed a list of desirable attributes for a MD. Therefore, this list can be considered as a methodological guide for the development of new topological indices (TIs). One of the most important criteria is the possibility of defining the MDs locally. This attribute refers to the fact that the index could be calculated not only for the molecule as a whole, but also over certain fragments of the structure itself.

Sometimes, the properties of a group of molecules are more related to a certain zone or fragment than to the molecule as a whole. Thereafter, the global definition never satisfies the structural requirements needed to obtain a good correlation in QSAR and QSPR studies. Furthermore, the local indices can be used in certain problems such as:

- Research on drugs, toxics or, generally, any organic molecules with a common skeleton, which is responsible for the activity or property under study.
- Study on the reactivity of specific sites in a series of molecules, which can undergo a chemical reaction or enzymatic metabolism.
- In the study of molecular properties such as spectroscopic measurements, which are obtained experimentally in a local way (chromophore).
- In any general case where it is necessary to study not only the molecule as a whole, but also some local properties of certain fragments, the definition of local descriptors could then be necessary.

Therefore, in addition to total bilinear indices computed for the whole molecule, local-fragment (atomic, group or atom-type) formalism can be developed. These MDs are termed local non-stochastic and stochastic bilinear indices, $b_{kL}(\bar{x},\bar{y})$ and ${}^{s}b_{kL}(\bar{x},\bar{y})$, respectively. The definition of these descriptors is as follows:

$$b_{kL}(\bar{x},\bar{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^{k}m_{ijL}x^{i}y^{j} \tag{13}$$

$${}^{s}b_{kL}(\bar{x},\bar{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^{k}s_{ijL}x^{i}y^{j} \tag{14}$$

where ${}^{k}m_{ijL}$ $[{}^{k}s_{ijL}]$ is the kth element of the 'i' row and 'j' column of the local matrix $\mathbf{M}_{L}^{k}[\mathbf{S}_{L}^{k}]$. This matrix is extracted from the $\mathbf{M}^{k}$ $[\mathbf{S}^{k}]$ matrix and contains information referred to the pairs of vertices (atomic nuclei) of the specific molecular fragments and also of the molecular environment in the kth step. The matrix $\mathbf{M}_{L}^{k}[\mathbf{S}_{L}^{k}]$ with elements ${}^{k}m_{ijL}$ $[{}^{k}\boldsymbol{s}_{ijL}]$ is defined as follows:

$$\begin{aligned}{}^{k}m_{ijL}[{}^{k}\boldsymbol{s}_{ijL}] &= {}^{k}m_{ij}[{}^{k}\boldsymbol{s}_{ijL}] \text{ if both } v_i \text{ and } v_j \text{ are atomic nuclei}\\ &\quad\text{contained within the molecular fragment}\\ &= 1/2\,{}^{k}m_{ij}[{}^{k}\boldsymbol{s}_{ijL}] \text{ if either } v_i \text{ or } v_j \text{ is an atomic}\\ &\quad\text{nucleus contained within the molecular fragment}\\ &= 0 \text{ otherwise}\end{aligned} \tag{15}$$

These local analogues can also be expressed in matrix form by the expressions:

$$b_{L}(\bar{x},\bar{y}) = [X]^{\mathrm{T}}\mathbf{M}_{L}^{k}[Y] \tag{16}$$

$${}^{s}b_{L}(\bar{x},\bar{y}) = [X]^{\mathrm{T}}\mathbf{S}_{L}^{k}[Y] \tag{17}$$

It should be remarked that the scheme above follows the spirit of a Mulliken population analysis of atomic net charges.[70] It should be also pointed out that for each partitioning of a molecule into Z molecular fragments; there will be Z local fragmental matrices. In this case, if a molecule is partitioned into Z fragments, the matrix $\mathbf{M}^{k}[\mathbf{S}^{k}]$ can be correspondingly partitioned into Z local matrices $\mathbf{M}_{L}^{k}[\mathbf{S}_{L}^{k}]$, $L = 1,\ldots,Z$, and the kth power of matrix $\mathbf{M}[\mathbf{S}]$ is exactly the sum of the kth powers of the local Z matrices. Therefore, the total non-stochastic and stochastic bilinear indices are the sum of the non-stochastic and stochastic bilinear indices, respectively, of the Z molecular fragments:

$$b(\bar{x},\bar{y}) = \sum_{L=1}^{Z} b_{kL}(\bar{x},\bar{y}) \tag{18}$$

$${}^{s}b(\bar{x},\bar{y}) = \sum_{L=1}^{Z} {}^{s}b_{kL}(\bar{x},\bar{y}) \tag{19}$$

Atomic, group, and atom-type bilinear fingerprints are specific cases of local bilinear indices. Atomic bilinear indices, $b_{kL}(\bar{x}_i,\bar{y}_i)$ and ${}^{s}b_{kL}(\bar{x}_i,\bar{y}_i)$, can be computed for each atom i in the molecule and contain electronic as well as topological structural information from all other atoms within the structure. The values of atom-level bilinear indices for the common scaffold atoms can be directly used as variables in seeking a QSPR/QSAR model, as long as these atoms are numbered in the same way in all the molecules in the database.

In addition, the atom-type bilinear indices can also be calculated. In the same way as atom-type E-state values,[71] for all data sets (including those with a common skeletal core as well as those with diverse structures), these novel local MDs provide much useful information. Therefore, this approach provides the basis for application to a wider range of problems to which the atomic bilinear indices formalism is applicable without the need for superposition.[72,73] For this reason, the present method represents a significant advantage over traditional QSAR methods. The atom-type bilinear descriptors are calculated by adding the kth atomic bilinear indices for all the atoms of the same type in the molecule. This atom type index allows a group additive-type scheme in which an index appears for each atom type in the molecule. In the atom-type bilinear indices formalism, each atom in the molecule is classified into an atom-type (fragment), such as –F, –OH, =O, $-CH_3$, and so on.[71–73] Therefore, each atom in the molecule is categorized according to a valence-state classification scheme including the number of attached H-atoms.[71] The atom-type descriptors combine three important aspects of structural information: (1) collective electron and topological accessibility to the atoms of the same type (for each structural feature: either atom or hybrid group, either such as –Cl, =O, $-CH_2-$, etc.), (2) presence/absence of the atom type (structural features), and (3) count of the atoms in the atom-type sets.

Finally, these local MDs can be calculated by a chemical (or functional) group in the molecule, such as heteroatoms (O, N, and S in all valence states as well as including the number of attached H-atoms), hydrogen bonding (H-bonding) to heteroatoms (O, N, and S in all valence states), halogen atoms (F, Cl, Br, and I), all aliphatic carbon chains (several atom types), all aromatic atoms (aromatic rings), and so on. The group-level bilinear indices are the sum of the individual atom-level bilinear indices for a particular group of atoms. For all data set structures, the *k*th group-based bilinear indices provide important information for QSAR/QSPR studies.

2.6. Sample calculation

It is useful to perform a calculation on a molecule to illustrate the effect of structure on the values of atomic and global bilinear indices. Thus, we use the 3-mercaptopyridine-4-carbaldehyde molecule. The labeled (atom numbering) molecular structure of this chemical as well as the non-stochastic and stochastic (atom-level, group, and atom-type as well as total) atom-based bilinear indices are shown in Table 3.

The following descriptors, using a weighting scheme of different combinations of four atomic properties (see Section 5.1), were calculated in this study:

(i) $\boldsymbol{b}_{\boldsymbol{k}}(x,y)$ and $\boldsymbol{b}_{\boldsymbol{k}}^{\mathrm{H}}(x,y)$ are the *k*th total bilinear indices not considering and considering the H atoms in the molecule, respectively.

(ii) $\boldsymbol{b}_{\boldsymbol{k}L}(x_E,y_E)$ and $\boldsymbol{b}_{kL}^{\mathrm{H}}(x_E,y_E)$ are the *k*th local (group = heteroatoms: S, N, O) bilinear indices not considering and considering H atoms in the molecule, correspondingly. These local descriptors are putative molecular charge, dipole moment, and H-bonding acceptors.

(iii) $\boldsymbol{b}_{\mathbf{k}L}^{\mathrm{H}}(x_{\mathrm{E-H}},y_{\mathrm{E-H}})$ are the *k*th local (group = H atoms bonding to heteroatoms: S, N, O) bilinear indices considering H atoms in the molecule. These local descriptors are putative H-bonding donors (hydrogen bonding capacity), lipophilicity, and so on.

(iv) $\boldsymbol{b}_{\boldsymbol{k}}(x_{\mathrm{Hal}}, y_{\mathrm{Hal}})$ and $\boldsymbol{b}_{\boldsymbol{k}}^{\mathrm{H}}(x_{\mathrm{Hal}},y_{\mathrm{Hal}})$ are the *k*th local (group = halogens: F, Cl, Br, I) bilinear indices not considering and considering the H atoms in the molecule, correspondingly.

(v) The *k*th total [$^{s}\boldsymbol{b}_{\boldsymbol{k}}(x,y)$ and $^{s}\boldsymbol{b}_{\boldsymbol{k}}^{\mathrm{H}}(x,y)$]€, as well as atom-type [$^{s}\boldsymbol{b}_{\boldsymbol{k}}(x_{\mathrm{E}},y_{\mathrm{E}})$, $^{s}\boldsymbol{b}_{\boldsymbol{k}}{}^{\mathrm{H}}(x_{\mathrm{E}},y_{\mathrm{E}})$ and $^{s}\boldsymbol{b}_{\boldsymbol{k}}{}^{\mathrm{H}}(x_{\mathrm{E-H}},y_{\mathrm{E-H}})$] stochastic bilinear indices, were also computed.

3. Results and discussion

3.1. Clustering and 'rational' design of training and test sets

In order to obtain mathematical expressions capable of discriminating between active and inactive compounds, the chemical information contained in a great number of compounds with and without the desired biological activity must be statistically processed. Taking into account that the most critical aspect in the construction of a training data set is the molecular diversity of the included compounds, we selected a group of 123 organic chemicals having as much structural variability as possible. The 50 antitrichomonals considered in this study are representative of families with diverse structural patterns and action modes. Figure 1 shows a representative sample of such active compounds. On the other hand, 73 compounds having different clinical uses were selected for the set of inactive compounds, through a random selection, guaranteeing also a great structural variability. All these chemicals were taken from the Negwer Handbook,[74] and Merck Index,[75] where their names, synonyms, and structural formulas can be found.

We performed a hierarchical cluster analysis of the active and inactive series using statistical software package STATISTICA.[76] This procedure permits to select compounds for the training and test sets, in a representative way. The main idea of this procedure consists in making a partition of either active or inactive series of chemicals in several statistically representative classes of compounds. It ensures that any chemical class (as determined by the clusters) will be represented in both compounds' series. This 'rational' design of the training and predicting series allowed us to design both sets that are representative of the whole 'experimental universe.' A detailed discussion of this procedure can be seen as Supplementary Data.

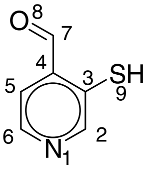
3.2. Discriminant models

3.2.1. Development. Linear discriminant analysis (LDA) has become an important tool for the prediction of biological properties.[42–49,77–81] On the basis of the simplicity of this method many useful discriminant models have been developed and presented by different authors in the literature.[42–49,77–81] Therefore, LDA was also the technique used in the generation of discriminant functions in the current work. Making use of the LDA technique implemented in the STATISTICA software,[76] the following linear models were obtained; in which total as well as local non-stochastic and stochastic bilinear indices were used as independent variables:

$$\begin{aligned}\text{Class} = {} & -3.37 - 0.07^{MP}\boldsymbol{b}_{1L}^{\mathrm{H}}(x_E,y_E) \\ & + 0.04^{ME}\boldsymbol{b}_{1L}(x_E,y_E) \\ & + 0.10^{MV}\boldsymbol{b}_{0L}^{\mathrm{H}}(x_{E-\mathrm{H}},y_{E-\mathrm{H}}) \\ & + 1.39 \times 10^{-10MV}\boldsymbol{b}_{15L}(x_{\mathrm{hal}},y_{\mathrm{hal}}) \\ & N = 91,\ \lambda = 0.43,\ D^2 = 5.07, \\ & F(4,86) = 27.48,\ p < 0.0001.\end{aligned} \tag{20}$$

$$\begin{aligned}\text{Class} = {} & -6.83 + 0.21^{ME\mathbf{s}}\boldsymbol{b}_{1L}^{\mathrm{H}}(x_E,y_E) \\ & - 0.20^{MP\mathbf{s}}\boldsymbol{b}_{1L}^{\mathrm{H}}(x_E,y_E) - 0.08^{ME\mathbf{s}}\boldsymbol{b}_{4L}(x_E,y_E) \\ & + 0.05^{MP\mathbf{s}}\boldsymbol{b}_{4L}(x_E,y_E) \\ & N = 91,\ \lambda = 0.28,\ D^2 = 9.90, \\ & F(4,86) = 53.6,\ p < 0.0001.\end{aligned} \tag{21}$$

Table 3. Values of atom-based bilinear indices for 3-mercapto-pyridine-4-carbaldehyde



	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 8$	$k = 15$
***k*th non-stochastic MDs, $b_{kL}(\bar{x}_i,\bar{y}_i)$[a]**						
Atom (*i*)	Atom-level bilinear indices					
N_1	17.1589	69.30704	207.92112	677.6482	217716.1986	711876800.9
C_2	39.51024	105.09455	382.76839	1173.27708	380330.8902	1242200889
C_3	39.51024	172.57929	475.22504	1557.467468	498904.5116	1625489636
C_4	39.51024	158.04096	526.894748	1699.867938	545797.7789	1784733323
C_5	39.51024	118.53072	381.66623	1237.280798	406509.5989	1327324524
C_6	39.51024	105.09455	328.71982	1025.6697	337449.5327	1105114810
C_7	39.51024	77.743778	316.08192	837.86986	288840.4172	930027508.4
O_8	9.218188	38.233538	75.10629	305.868304	85899.64937	284074052.7
S_9	70.8441	54.04857	232.98981	647.3843	213952.7799	698815935.8
Total (sum)	334.282628	898.672996	2927.373368	9162.333648	2975401.357	9709657480
Group (*i*)	Group bilinear indices					
Heteroatoms (sum of local-index values for N, O, and S atoms)	97.221188	161.589148	516.01722	1630.900804	517568.6278	1694766789
Atom-type (*i*)	Atom-type bilinear indices					
CH-arom (sum of local-index values for C-atoms 2, 5, and 6)	118.53072	328.71982	1093.15444	3436.227578	1124290.022	3674640223
***k*th stochastic MDs, ${}^{s}b_{kL}(\bar{x}_i,\bar{y}_i)$[a]**						
Atom (*i*)	Atom-level bilinear indices					
N_1	17.1589	23.10234667	22.34781083	23.66748208	23.89914186	24.01863153
C_2	39.51024	33.38525667	40.80307944	39.00990209	38.9415613	38.95862823
C_3	39.51024	60.9142225	44.90684156	47.06658972	45.29450781	45.13795589
C_4	39.51024	42.80276	47.40669467	47.62823907	47.72612389	47.70941996
C_5	39.51024	37.86398	37.79391122	39.18662953	40.21791136	40.32910683
C_6	39.51024	35.03151667	35.08198711	34.64609483	36.44270378	36.7396616
C_7	39.51024	27.639906	37.134922	32.10412486	34.42791207	33.95447149
O_8	9.218188	16.116086	10.803039	14.98191176	13.79473544	14.04362415
S_9	70.8441	29.6352825	42.1269325	34.8188677	34.3736933	34.24319914
Total (sum)	334.282628	306.491357	318.4052183	313.1098416	315.1182908	315.1346988
Group (*i*)	Group bilinear indices					
Heteroatoms (Sum of local-index values for N, O, and S atom)	97.221188	68.85371517	75.27778233	73.46826154	72.0675706	72.30545482
Atom-type (*i*)	Atom-type bilinear indices					
CH-arom (sum of local-index values for C-atoms 2, 5, and 6)	118.53072	106.2807533	113.6789778	112.8426264	115.6021765	116.0273967

[a] Calculation development using VdW volume (*V*) and polarizability (*P*) (see Table 2) as combination (pairs) of two atom-label chemical properties from our weighting schemes.

where *N* is the number of compounds, λ is Wilks' statistics, D^2 is the square of the Mahalanobis distance, *F* is the Fisher ratio, and *p* is the significance level.

It should be remarked that these equations have shown to be statistically significant at *p*-level ($p < 0.0001$). The results of global good classification of chemicals, in both training and test sets, are shown in Table 4. As it can also be computed from the results shown in Tables 5 and 6, models **20** and **21** showed appropriate overall accuracies (94.51% and 93.41%, respectively) as well as a high Matthews correlation coefficient (***C*** of 0.89 and of 0.87, correspondingly).[82] Statistic ***C*** quantifies the strength of the linear relation between the molecular descriptors and the classifications, as well as it may often provide a much more balanced evaluation of the prediction than, for instance, the percentages (accuracy).[82] In Table 4 we also list most of the parameters commonly used in medical statistics [sensitivity, specificity, and false positive rate (also known as 'false alarm rate')] for the whole set of developed models.

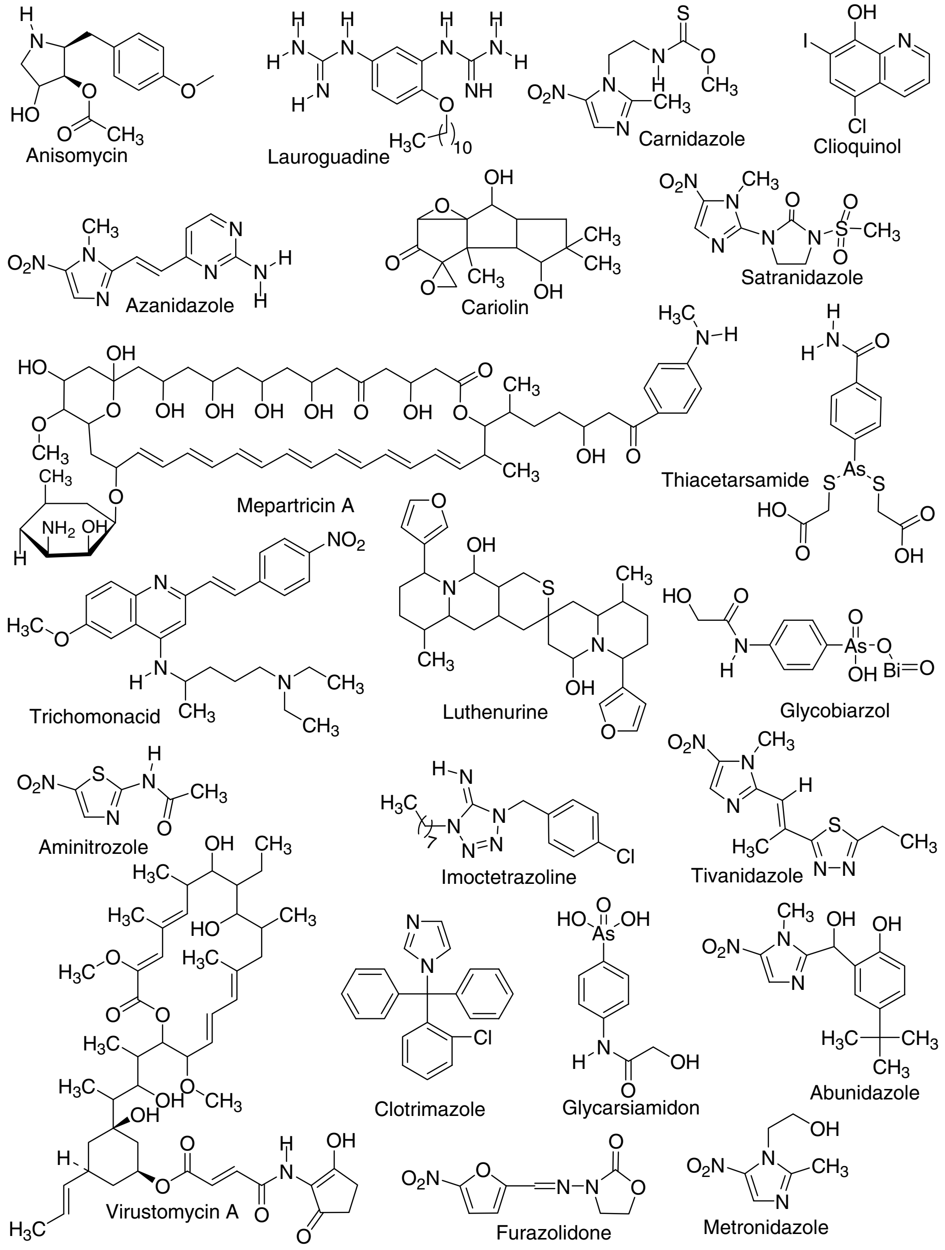


Figure 1. Random sample of the molecular families of trichomonacidal agents studied here.

Table 4. Prediction performances for two LDA-based QSAR models in the learning and predicting sets

	Matthews Corr. Coefficient (C)	Accuracy ‘Q_{Total}’ (%)	Sensitivity ‘hit rate’ (%)	Specificity (%)	False positive rate ‘false alarm rate’ (%)
Non-Stochastic Atom-Type Bilinear Indices (Eq. 20)					
Learning set	0.89	94.51	0.9	0.97	0.02
Predicting set	0.79	90.63	0.9	0.82	0.09
Stochastic Atom-Type Bilinear Indices (Eq. 21)					
Learning set	0.87	93.41	0.90	0.95	0.04
Predicting set	0.85	93.75	0.90	0.90	0.05

While the sensitivity is the probability of correctly predicting a positive example, the specificity (also known as ‘hit rate’) is the probability that a positive prediction is correct.[82] The classification from Eqs. 20 and 21 of all active and inactive compounds appears in Tables 8 and 9, respectively. A plot of the $\Delta P\%$

Table 5. Names and classification of active compounds in training and test series according to the two TOMOCOMD-CARDD models developed in this work

Name	ΔP%[a]	Score[b]	ΔP%[c]	Score[d]	Name	ΔP%[a]	Score[b]	ΔP%[c]	Score[d]
Active training set									
Anisomycin	-47.33	0.21	-97.43	-1.28	Abunidazole	91.72	-1.64	99.98	3.05
Virustomycin A	93.52	-1.75	67.64	0.62	Imoctetrazoline	71.08	-1.03	92.44	1.13
Azanidazole	96.69	-2.06	100.00	3.60	Forminitrazole	32.42	-0.54	98.62	1.68
Carnidazole	38.54	-0.60	99.36	1.92	Chlomizol	81.88	-1.27	98.90	1.75
Propenidazole	94.86	-1.86	99.59	2.06	Acinitrazole	32.42	-0.54	98.13	1.58
Lauroguadine	-7.45	-0.18	-80.17	-0.60	Moxnidazole	98.77	-2.50	99.99	3.33
Mepartricin A	99.81	-3.33	99.59	2.07	Isometronidazole	86.81	-1.42	99.76	2.24
Metronidazole	86.81	-1.42	99.76	2.23	Mertronidazole phosphate	29.96	-0.52	59.30	0.53
Nifuratel	84.15	-1.33	99.95	2.71	Benzoylmetronidazole	91.73	-1.64	99.62	2.09
Nifuroxime	97.68	-2.22	100.00	3.59	Bamnidazole	95.43	-1.91	98.48	1.65
Nimorazole	88.44	-1.48	99.86	2.40	Glycarsiamidon	97.91	-2.26	97.33	1.47
Secnidazole	86.81	-1.42	99.72	2.19	Fexinidazole	21.48	-0.44	99.92	2.57
Cariolin	-42.76	0.16	-81.68	-0.63	Piperanitrozole	28.04	-0.50	98.85	1.74
2 -Amino -5 -nitrotiazola	20.56	-0.43	99.31	1.90	Gynotabs	14.71	-0.38	99.54	2.03
Glycobiarzol	99.94	-3.87	99.97	2.89	Pirinidazole	29.16	-0.51	99.95	2.72
Clioquinol	73.30	-1.07	1.41	0.11	Metronidazole hydrogen succinate	96.82	-2.08	98.48	1.65
Diiodohydroxyquinoline	97.05	-2.11	96.77	1.41	Tolamizol	32.42	-0.54	99.38	1.94
Ornidazol	75.77	-1.12	99.76	2.24	Thiacetarsamide	94.51	-1.83	99.69	2.16
Trichomonacid	94.38	-1.82	99.99	3.13	Tivanidazole	55.69	-0.80	99.96	2.83
Lutenurine	-93.16	1.24	-80.97	-0.62	Policresulen	92.49	-1.68	46.69	0.42
Active test set									
Acertarsone	97.91	-2.26	97.07	1.44	Pentamycin	85.17	-1.36	68.05	0.63
Furazolidone	98.14	-2.32	99.93	2.62	Azomycin	83.30	-1.31	99.50	2.00
Mepartricin B	99.85	-3.44	99.64	2.11	Ternidazole	86.81	-1.42	99.62	2.09
Aminitrozole	32.42	-0.54	98.13	1.58	Misonidazole	91.72	-1.64	99.84	2.37
Clotrimazol	-76.05	0.64	-97.67	-1.31	Satranidazole	97.72	-2.23	99.94	2.70

[a,c] Antitrichomonal activity predicted by Eqs. 20 and 21, respectively, Δ*P*% = [*P*(active) − *P*(inactive)] × 100.

[b,d] Canonical scores obtained from canonical analysis, Eqs. 20a and 21a, correspondingly.

(see Supplementary data) from both models, Eqs. 20 and 21, for each compound in the training and test sets is illustrated in Figures 2 and 3, respectively, where the good classification results obtained with the current approach can be observed.

Later, in both cases we also developed the linear discriminant canonical analysis by checking the following statistics: canonical correlation coefficient ($R_{\text{canonical}}$), Chi-squared and its p-level [$p(\chi^2)$].[83] In this sense, the canonical transformation of the LDA results with non-stochastic (Eq. 20) and stochastic (Eq. 21) bilinear fingerprints gives rise to canonical roots with good canonical correlation coefficients of 0.75 and 0.84, correspondingly. The chi-squared test permits us to assess the statistical significance of this analysis as having a p-level <0.0001.[83] Atom-type non-stochastic and stochastic linear indices, as well as LDA antitrichomonal activity canonical analysis principal root, are given below:

$$\begin{aligned}\text{Classroot} = {} & 1.25 + 0.03\,{}^{MP}\boldsymbol{b}^{\mathrm{H}}_{1L}(x_E, y_E) \\ & - 0.02\,{}^{ME}\boldsymbol{b}_{1L}(x_E, y_E) \\ & - 0.04\,{}^{MV}\boldsymbol{b}^{\mathrm{H}}_{0L}(x_{E-\mathrm{H}}, y_{E-\mathrm{H}}) \\ & - 6.19 \times 10^{-11}\,{}^{MV}\boldsymbol{b}_{15L}(x_{\mathrm{hal}}, y_{\mathrm{hal}}) \\ & N = 91,\ \lambda = 0.43,\ R_{\mathrm{can}} = 0.75, \\ & \chi^2 = 71.63,\ mean(+) = -1.26, \\ & mean(-) = 0.99,\ p < 0.0001.\end{aligned} \tag{20a}$$

$$\begin{aligned}\text{Classroot} = {} & -2.07 + 0.07\,{}^{ME\mathbf{s}}\boldsymbol{b}^{\mathrm{H}}_{1L}(x_E, y_E) \\ & - 0.06\,{}^{MP\mathbf{s}}\boldsymbol{b}^{\mathrm{H}}_{1L}(x_E, y_E) \\ & - 0.02\,{}^{ME\mathbf{s}}\boldsymbol{b}_{4L}(x_E, y_E) + 0.02\,{}^{MP\mathbf{s}}\boldsymbol{b}_{4L}(x_E, y_E) \\ & N = 91,\ \lambda = 0.28,\ R_{\mathrm{can}} = 0.84, \\ & \chi^2 = 108.42,\ mean(+) = 1.76, \\ & mean(-) = -1.38,\ p < 0.0001.\end{aligned} \tag{21a}$$

Table 6. Names and classification of inactive compounds in training and test series according to the two TOMOCOMD-CARDD models developed in this work

Name	ΔP%^{a}	Score^{b}	ΔP%^{c}	Score^{d}	Name	ΔP%^{a}	Score^{b}	ΔP%^{c}	Score^{d}
Inactive training set									
Amantadine	-88.09	0.98	-99.23	-1.67	Nonaferone	-73.42	0.59	-98.81	-1.53
Thiacetazone	-85.62	0.89	-83.27	-0.66	Rolipram	-77.92	0.68	-98.10	-1.38
Cloral betaine	-97.74	1.74	-99.95	-2.51	N-hydroxymethyl-N-methylurea	-99.53	2.44	-99.41	-1.75
Carbavin	-71.03	0.54	-99.86	-2.22	4 chlorobenzoic acid	-88.49	1.00	-99.02	-1.59
Norantoin	-73.34	0.59	-99.43	-1.76	Acetanilide	-91.13	1.12	-99.67	-1.94
Orotonsan Fe	-31.51	0.05	-98.93	-1.56	Guanazole	-96.22	1.51	-87.22	-0.75
Picosulfate	99.83	-3.39	-47.26	-0.23	Tetramin	-92.94	1.22	-98.54	-1.46
Naftazone	-11.30	-0.14	-88.74	-0.80	Mecysteine	-98.38	1.89	-99.40	-1.75
Besunide	-7.50	-0.18	-78.98	-0.58	Cirazoline	-89.86	1.06	-98.50	-1.45
Acetazolamide	-55.85	0.32	-75.20	-0.52	Methocarbamol	-76.15	0.64	-98.79	-1.52
Propamine"soviet	-96.72	1.57	-95.79	-1.12	Lysergide	-88.43	0.99	-98.02	-1.36
RMI 11894	-89.12	1.02	-99.67	-1.93	Dopamine	-96.55	1.55	-86.50	-0.73
Ag 307	-99.53	2.45	-99.12	-1.62	Bufeniode	-74.49	0.61	97.42	1.48
Barbismethylii iodide	-55.21	0.31	-98.63	-1.48	Celiprolol	-80.26	0.74	-97.56	-1.30
Pancuronium bromide	-47.42	0.21	-99.64	-1.90	Erysimin	-54.17	0.29	-75.64	-0.53
Vinyl ether	-89.37	1.03	-99.72	-1.99	Peruvoside	-31.22	0.04	-79.67	-0.59
Basedol	-98.42	1.90	-99.52	-1.82	Amitraz	-90.07	1.07	-97.80	-1.33
Carbimazole	-97.38	1.68	-99.93	-2.43	Proclonol	-89.86	1.06	-96.32	-1.16
Didym levulinate	-73.91	0.60	-99.90	-2.32	Asame	-49.07	0.23	-97.97	-1.36
Perchloroethane	-99.89	3.07	-100.00	-3.34	KC-8973	-68.81	0.51	-99.32	-1.70
Pyrantel tartrate	-99.01	2.11	-98.87	-1.54	Ethydine	-35.76	0.09	-99.19	-1.65
Fentanyl	-86.18	0.91	-99.44	-1.77	Magnesii metioglicas	-99.97	3.65	-99.63	-1.90
Petidina	-80.90	0.75	-99.59	-1.87	Alibendol	-87.74	0.97	-89.15	-0.81
Tenalidine tartrate	-99.67	2.60	-99.17	-1.64	Diponium Bromide	-80.08	0.73	-99.85	-2.19
Bamipine	-91.31	1.13	-98.89	-1.55	Streptomycin	-93.72	1.28	98.61	1.68
Colestipol	-98.01	1.80	-96.05	-1.14					
Inactive test set									
Citenazone	-99.19	2.20	-70.13	-0.45	Metriponate	-99.90	3.12	-100.00	-3.89
Methenamine	-88.66	1.00	-98.78	-1.52	Ciclopramine	-89.12	1.02	-97.90	-1.34
Penthrichloral	-93.91	1.29	-99.82	-2.12	Litracen	-90.48	1.09	-99.26	-1.68
Ferdomus	-100.00	5.36	-100.00	-9.98	Trimethylsulfonium hydroxyde	-99.75	2.73	-99.60	-1.87
Phenoltetrachlorophthalein	99.88	-3.53	-41.97	-0.18	Norgamem	-99.06	2.14	-99.74	-2.01
Bisoxatin acetate	-2.21	-0.22	-98.52	-1.46	Zoxazolamine	-88.01	0.98	-97.42	-1.28
Glicondamide	53.05	-0.77	-51.19	-0.26	Acetylcholine	-80.08	0.73	-99.76	-2.04
Bromcholine	-98.57	1.94	-97.24	-1.26	Carazolol	-93.13	1.24	-91.98	-0.91
Imekhin	-92.06	1.17	-99.68	-1.95	Cefazolin	-86.99	0.94	99.31	1.90
Frigen 113	-96.68	1.57	-100.00	-4.71	Penicillin I	-97.15	1.64	-99.48	-1.79
Eticoumarolum	-20.79	-0.06	-91.89	-0.90	Aziromycin	-88.70	1.01	-99.57	-1.85

[a,c] Antitrichomonal activity predicted by Eqs. 20 and 21, respectively: ΔP% = [P(active) − P(inactive)] × 100.

[b,d] Canonical scores obtained from canonical analysis, Eqs. 20a and 21a, correspondingly.

When the LDA is applied to solve the two-group classification problem, we always find two classification functions. However, we cannot use these two classification functions to evaluate all the compounds and obtain a bivariate activity map because they are not orthogonal.[83] In order to solve this problem we used canonical analysis, in this case the dimensional reduction caused by canonical analysis makes possible to obtain a 1-dimensional activity map. Therefore, we can sort all compounds taking into account its canon-

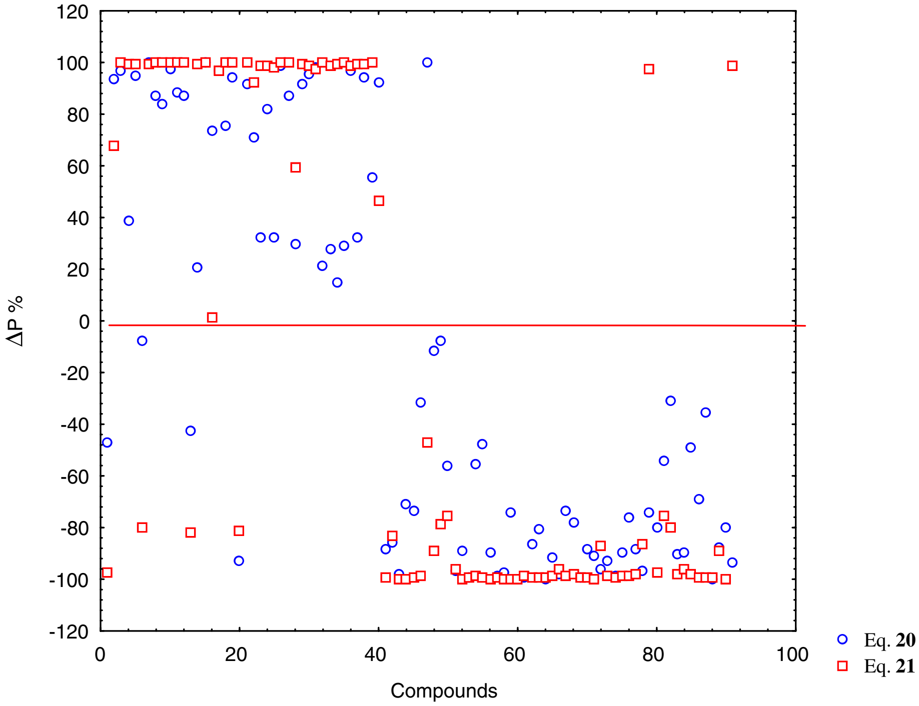


Figure 2. Plot of the Δ*P*% from Eq. 20 (blue) and Eq. 21 (red) [using non-stochastic and stochastic bilinear indices, respectively] for each compound in the training sets. Compounds **1–40** are active and chemicals 41–91 are inactive.

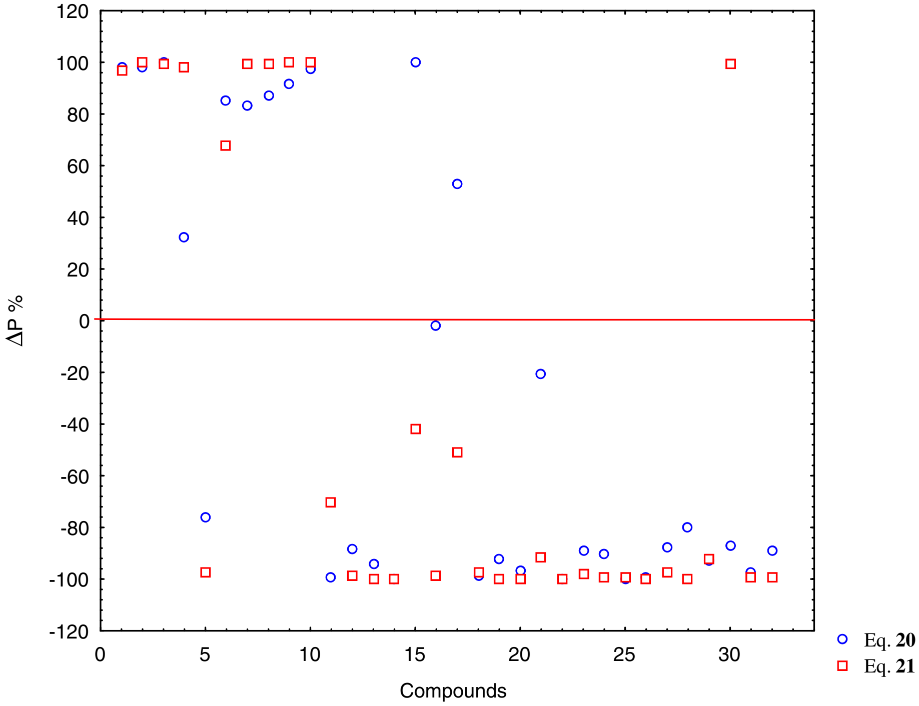


Figure 3. Plot of the Δ*P*% from Eq. 20 (blue) and Eq. 21 (red) [using non-stochastic and stochastic bilinear indices, respectively] for each compound in the test sets. Compounds **1–10** are active and chemicals 11–32 are inactive.

ical scores. The canonical scores of all the active and inactive compounds appear in Tables 8 and 9, correspondingly.

3.2.2. Intercorrelation study and orthogonalization process. A close inspection of the molecular descriptors included in both LDA-based QSAR models showed that several of these molecular fingerprints are strongly interrelated with each other. In Table 7 we give the correlation coefficients of the molecular descriptors in Eqs. 20 and 21.

The interrelation among the molecular descriptors makes difficult the interpretation of the QSAR model. That is to say, it is well known that the interrelated-

Table 7. Intercorrelation of the molecular descriptors included in the LDA-based QSAR models and results of Randić's orthogonalization analysis

Non-orthogonal atom-type nonstochastic bilinear indices				Non-orthogonal atom-type nonstochastic bilinear indices			
$^{ME}\boldsymbol{b}_{1L}(x_{E},y_{E})$	$^{MP}\boldsymbol{b}_{1L}{}^{H}(x_{E},y_{E})$	$^{MV}\boldsymbol{b}_{15L}(x_{hal},y_{hal})$	$^{MV}\boldsymbol{b}_{0L}{}^{H}(x_{E\text{-}H},y_{E\text{-}H})$	$^{MEs}\boldsymbol{b}_{1L}{}^{H}(x_{E},y_{E})$	$^{MPs}\boldsymbol{b}_{1L}{}^{H}(x_{E},y_{E})$	$^{MEs}\boldsymbol{b}_{4L}(x_{E},y_{E})$	$^{MPs}\boldsymbol{b}_{4L}(x_{E},y_{E})$
1.00	0.97	-0.06	0.28	1.00	0.94	0.96	0.83
	1.00	0.04	0.32		1.00	0.91	0.91
		1.00	-0.06			1.00	0.87
			1.00				1.00
Orthogonal atom-type non-stochastic bilinear indices				Orthogonal atom-type stochastic bilinear indices			
$^{1}O(^{ME}\boldsymbol{b}_{1L}(x_{E},y_{E}))$	$^{2}O(^{MP}\boldsymbol{b}_{1L}{}^{H}(x_{E},y_{E}))$	$^{3}O(^{MV}\boldsymbol{b}_{15L}(x_{hal},y_{hal}))$	$^{4}O(^{MV}\boldsymbol{b}_{0L}{}^{H}(x_{E\text{-}H},y_{E\text{-}H}))$	$^{1}O(^{MEs}\boldsymbol{b}_{1L}{}^{H}(x_{E},y_{E}))$	$^{2}O(^{MPs}\boldsymbol{b}_{1L}{}^{H}(x_{E},y_{E}))$	$^{3}O(^{MEs}\boldsymbol{b}_{4L}(x_{E},y_{E}))$	$^{4}O(^{MPs}\boldsymbol{b}_{4L}(x_{E},y_{E}))$
1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
	1.00	0.00	0.00		1.00	0.00	0.00
		1.00	0.00			1.00	0.00
			1.00				1.00
LDA-based model derived with orthogonal atom-type non-stochastic bilinear indices				LDA-based model derived with orthogonal atom-type stochastic bilinear indices			
$\boldsymbol{Class} = -1.61 + 1.07\,{}^{1}O(^{ME}\boldsymbol{b}_{1L}(x_{E},y_{E})) - 4.68\,{}^{2}O(^{MP}\boldsymbol{b}_{1L}{}^{H}(x_{E},y_{E})) + 0.94\,{}^{3}O(^{MV}\boldsymbol{b}_{15L}(x_{hal},y_{hal})) + 0.72\,{}^{4}O(^{MV}\boldsymbol{b}_{0L}{}^{H}(x_{E\text{-}H},y_{E\text{-}H}))$ (**20b**)				$\boldsymbol{Class} = -0.32 - 11.38\,{}^{1}O(^{MEs}\boldsymbol{b}_{1L}{}^{H}(x_{E},y_{E})) + 3.53\,{}^{2}O(^{MPs}\boldsymbol{b}_{1L}{}^{H}(x_{E},y_{E})) - 8.84\,{}^{3}O(^{MEs}\boldsymbol{b}_{4L}(x_{E},y_{E})) + 4.3\,{}^{4}O(^{MPs}\boldsymbol{b}_{4L}(x_{E},y_{E}))$ (**21b**)			
N = 91 $\lambda = 0.43$ $D^2 = 5.07$ F(4. 86) = 27.48 $\boldsymbol{C} = 0.89$ $Q_{total} = 94.51$ $p<0.0001$				N = 91 $\lambda = 0.28$ $D^2 = 9.90$ F(4. 86) = 53.6 $\boldsymbol{C} = 0.87$ $Q_{total} = 93.41$ $p<0.0001$			

ness among the different descriptors can result in highly unstable correlation coefficients, which make impossible to know the relative importance of an index and underestimate the utility of the correlation coefficient in a model.[75,32,77–79,84,85] To overcome this difficulty, an approach based on the orthogonalization of the descriptors has been introduced in the literature.[86–88] The main philosophy of this approach is to avoid the exclusion of descriptors on the basis of its collinearity with other variables included in the model. However, in some cases, strongly interrelated descriptors can enhance the quality of a model, because the small fraction of a descriptor that is not reproduced by its strongly interrelated pair can provide positive contributions to the modeling.

This process is an approach in which molecular descriptors are transformed in such a way that they do not mutually correlate. Both, the non-orthogonal descriptors and derived orthogonal descriptors, contain the same information. In this sense, the same statistical parameters of the QSAR models are obtained.[86–88] In addition, the coefficients of the QSAR model based on orthogonal descriptors are stable to the inclusion of novel descriptors, which permit interpreting the correlation coefficient and evaluating the role of individual fingerprints in the QSAR model.

In Table 7, we also resume the results of the orthogonalization of molecular descriptors included in both equations. In this case, the models **20a** and **21a** correspond to the final models with the orthogonalized bilinear indices. Here, we used the symbols $^{m}O(\boldsymbol{b}_k(x,y))$, where the superscript m expresses the order of importance of the variable $(\boldsymbol{b}_k(x,y))$ after a preliminary forward stepwise analysis, and O means orthogonal.

It has to be highlighted here that the orthogonal descriptor-based models coincide with the collinear (i.e., ordinary) bilinear descriptor-based models in all statistical parameters. The statistical coefficients of LDA-QSARs λ, D^2, F, $\boldsymbol{C}$, accuracy (Q_{total}) are the same whether we use either a set of non-orthogonal descriptors or the corresponding set of orthogonal indices (see Table 7).[86–88] This is not surprising, because the latter models are derived as linear combinations of the former ones and, therefore, the latter cannot have more information content than the former have.

This fact also makes possible the interpretation of the coefficients in the LDA-QSAR equations. Therefore, $^{m}O(\boldsymbol{b}_k(x))$ may be classified according to the distance k into short- (0–5), mid- (6–10), and long-range non-stochastic and stochastic bilinear indices. The information in Table 7 clearly shows that the major contribution to antitrichomonal activity is provided by short-range atom-type (heteroatoms and H atoms bonding to heteroatoms) bilinear indices. These short-range local descriptors are putative molecular charge, dipole moment, as well as H-bonding acceptors, and H-bonding donors.

In general way, for non-stochastic descriptors the variables weighted with the combination of atomic masses and van der Waals volumes and the parameters weighted with the combination atomic masses and Pauling electronegativities have a positive contribution to the antitrichomonal activity. While, the variable weighted with the combination of atomic masses and Pauling electronegativities has a negative contribution to the antitrichomonal activity. On the other hand, for stochastic descriptors the variables weighted with the combination of atomic masses and atomic polarizabilities have a positive contribution to the antitrichomonal activity. While, variables weighted with the combination of atomic masses and Pauling electronegativities have a negative contribution to the antitrichomonal activity.

3.2.3. Internal and external validation of the discriminant functions.

In recent years, exhaustive validation of mathematical models constitutes a main key of current QSAR theory.[89] In this sense, internal validation methods (e.g., cross-validation) are considered by many authors as an indicator or even as the ultimate proof of the stability and high-predictive power of an QSAR model. However, Golbraikh and Tropsha demonstrated that high values of leave-one-out square correlation coefficient q^2 appear to be a necessary, but not the sufficient, condition for the model to have a high predictive power.[90] A more exhaustive cross-validation method can be used in which a fraction of the data (10–20%) is left out and predicted from a model based on the remaining data. This process (leave-group-out, LGO) is repeated until each observation has been left out at least once.[90,91] In this report, each investigated data set was split randomly into seven groups of approximately the same size (15%). Each group was left out (LGO) and that group was then predicted by a model developed from the remaining observations (85% of the data). This process was carried out seven times on seven unique subsets. In this way, each observation was left out once, in groups of 15%, and its value predicted. The mean of the accuracies for the seven groups will be used as the significant criterion for assessing model quality. The level of overall (average) accuracy (for a 15% full leave-out) of 7-fold cross-validation procedure can be taken as good confirmation of the predictive quality of the model. In addition, to assess the robustness and predictive power of the found models, external prediction (test) sets were also used. This type of model validation is important, if we take into consideration that the predictive ability of a QSAR model can be estimated using only an external test set of compounds that were not used for building the model.[89–91] Therefore, it is important to ensure that the prediction algorithms are able to perform well on novel data from the same data domain.

The statistical results of a 7-fold cross-validation procedure are depicted in Table 8. The overall mean of the correct classification in the training (test) set for this process for Eqs. 20 and 21 was 94.51% (93.20%) and 92.67% (90.84%), correspondingly. The result of predictions on the 15% full cross-validation test evidenced the quality (robustness, stability, and predictive power) of the obtained models.

Table 8. Results of the 7-fold full cross-validation procedure

Groups	Q%[a]	λ	D^2	F	Q%[b]	Q%[a]	λ	D^2	F	Q%[b]
1	96.10	0.44	4.93	22.45	85.71	93.51	0.29	9.86	44.94	92.31
2	94.81	0.46	4.61	20.99	100.00	93.51	0.26	11.43	52.07	92.31
3	92.31	0.49	4.15	19.10	100.00	89.74	0.31	8.95	41.23	100.00
4	93.59	0.45	4.88	22.48	100.00	93.59	0.28	10.23	47.11	92.31
5	94.87	0.47	4.44	20.44	100.00	91.03	0.31	8.80	40.55	92.31
6	97.47	0.38	6.89	30.36	83.33	93.67	0.26	11.10	51.97	83.33
7	92.41	0.33	8.12	38.04	83.33	93.67	0.27	10.53	49.34	83.33
Mean	94.51	0.43	5.43	24.84	93.20	92.67	0.28	10.13	46.74	90.84
SD	1.90	0.06	1.49	6.87	8.52	1.61	0.02	1.00	4.73	5.85

[a,b]Global good classification from both models in training (85% of the data) and test (15% of the data) sets, respectively.

In this study, an external prediction data set was also evaluated as the second validation experiment. The computation of some performance measures permitted us to carry out the assessment of the model. The results for this validation process are summarized in Table 4. The classification's results using both obtained equations for active and inactive compounds in the selected test set are shown in Tables 8 and 9, respectively. Eqs. 20 and 21 showed a high ***C*** (*Q*%) of 0.79 (90.63) and 0.85 (93.75) in the predicting series, correspondingly. These results validate the models for their use in the ligand-based virtual screening.

3.3. 'Virtual' and 'in silico' screening as promissory alternative for drug discovery

The great cost associated with the development of new compounds and the small economic size of the market for most of the drugs make this development slow. For this reason, it is necessary to develop computational methods permitting theoretical in silico evaluations of antitrichomonal activity for virtual libraries of chemicals before these compounds are synthesized in the laboratory. This in silico world of data, analyses, hypotheses, and models that resides inside a computer is alternative to the 'real' world of syntheses and screening of compounds in the laboratory.[92]

One of the main features that any theoretical approach to drug discovery needs is the identification of active compounds from databases of chemicals. This search can be understood as an alternative to the screening approaches to drug discovery. In this approach, instead of essaying a large number of chemicals in a series of biological tests we 'virtually essay' these compounds by evaluating their activities by the models developed to this effect; this process is known today as computational (virtual or in silico) screening.[32–34,93] Both computational screening, 'virtual' or in silico, were used in this study in order to show the potentialities of the TOMOCOMD-CARDD for drug discovery with antitrichomonal activity.

3.3.1. Selection/identification of new trichomonacidals via ligand-based virtual screening. In order to prove the possibilities of the present approach for the ligand-based virtual screening of antitrichomonal compounds, we have selected a series of 18 compounds whose structures are given in Table 9. They have been selected from the medicinal chemistry literature reporting them as active/inactive compounds.

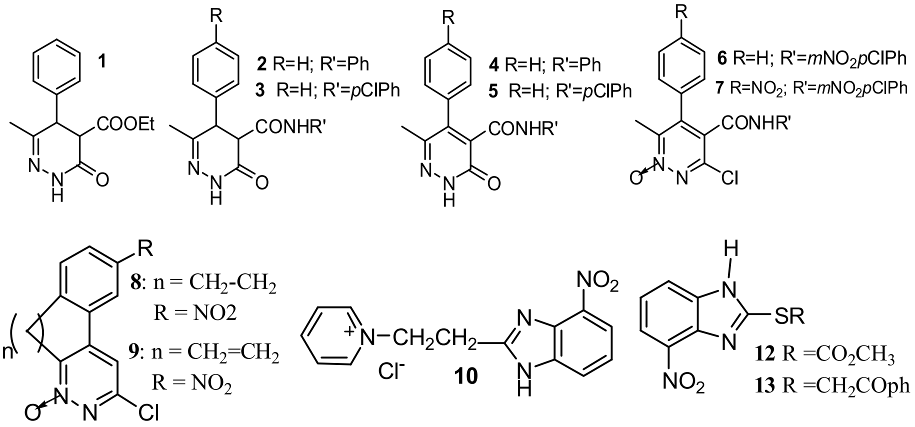
By these means, the present study is conducted to test the possibilities of the classification model developed here, in detecting trichomonacidals with diverse chemical structures. The verification of the predictions carried out by all the obtained models comes from the recent reports in the literature, from where these compounds were selected.

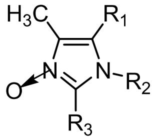
A hierarchical CA was first realized to observe the molecular variability of this data set. As it can be seen in the dendrogram, a great number of different subsets can be differentiated, which prove the molecular diversity of the selected chemicals in this set (see Figure 4).

The results of the classification of the compounds in these external test sets are also summarized in Table 9. Most of the chemicals included in this 'simulated' virtual screening experiment were well classified as active/inactive for the entire obtained model. For instance, the 100% of the screened chemicals were well classified by both LDA-based QSAR models developed with non-stochastic and stochastic bilinear indices, (Eqs. 20 and 21, respectively).

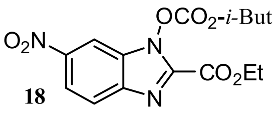
The next step in this approach would be the inclusion of these 'novel' compounds in the training set and the developing of a new discrimination model. This new model can be significantly different from the previous one, due to the inclusion of a new structural pattern, but it should be able to recognize a greater number of such compounds as trichomonacidals. By these ways, the derivation of the classifier model is considered as an iterative process, in which novel compounds with novel structural features are incorporated into the training set for improving the quality of the models so developed (Fig. 5).

Table 9. Lead identified as trichomonacidal from literature by using LDA-based QSAR models in simulate virtual screening





	R_1	R_2	R_3
14	CH_3	*p*Meph	H
15	H	*p*Meph	H
16	CH_3	*p*MeOph	H
17	CH_3	furylCH$_2$	CH_3



18

Compound[a]	Ref.[b]	$\Delta P\%$[c]	Score[d]	$\Delta P\%$[e]	Score[f]	Antitrichomonal activity
1	104	−17.22	−0.09	−97.16	−1.25	Inactive
2	104	−22.76	−0.04	−84.30	−0.68	Inactive
3	104	−18.25	−0.08	−69.10	−0.44	Inactive
4	104	−22.76	−0.04	−65.23	−0.39	Inactive
5	104	−17.02	−0.09	−37.80	−0.15	Inactive
6	104	100.00	−5.49	100.00	5.19	MIC = 31.5 μg/ml[g]
7	104	100.00	−7.84	100.00	9.05	MIC = 12.5 μg/ml[g]
8	105	99.99	−4.48	100.00	4.69	MIC = 31.3 μg/ml[g]
9	105	99.99	−4.61	100.00	4.76	MIC = 3.9 μg/ml[g]
10	106	91.37	−1.62	99.97	2.87	MLC = 50 μg/ml[h] LD_{50} = 50 μg/ml[h]
11	106	13.03	−0.36	99.82	2.33	MLC = 50 μg/ml[h] LD_{50} = 10–50 μg/ml[h]
12	106	52.43	−0.76	98.70	1.70	MLC = 50 μg/ml[h] LD_{50} = 10–50 μg/ml[h]
13	106	32.42	−0.54	99.78	2.26	MLC = 50 μg/ml[h] LD_{50} = 10–50 μg/ml[h]
14	107	−59.06	0.36	31.89	0.31	Inactive
15	107	−59.06	0.36	37.25	0.35	Inactive
16	107	−40.75	0.14	70.88	0.66	Inactive
17	107	−40.75	0.14	57.78	0.52	Inactive
18	107	99.53	−2.93	99.96	2.80	100 μg/ml = [87.5][i] 10 μg/ml = 17.3[i] 1 μg/ml = 9.6[i]

[a] The molecular structures of the compounds represented with numbers are shown at the top of this table.

[b] Bibliographical references from where they were taken the molecules together with in vitro activity.

[c] Antitrichomonal activity predicted by Eq. 20; $\Delta P\% = [P(\text{active}) - P(\text{inactive})] \times 100$.

[d] Canonical scores obtained from canonical analysis (Eq. 20a).

[e] Antitrichomonal activity predicted by Eq. 21; $\Delta P\% = [P(\text{active}) - P(\text{inactive})] \times 100$.

[f] Canonical scores obtained from canonical analysis (Eq. 21a).

[g] MIC: minimum inhibitory concentration. Most of these compounds showed better activity against *T. vaginalis* than metronidazole (MIC = 25 μg/ml).

[h] MLC: minimum lethal concentration that killed all the parasites by 24 h. LD_{50}: minimum concentration that reduced the number of parasites at least 50%. This compound showed a weak inhibitory activity compared with metronidazole (MLC = 1 μg/ml and LD_{50} = 1 μg/ml).

[i] Percentage of inhibition of *T. vaginalis* growth at the indicated doses at 24 h. In brackets, percentage of reduction.

Several drugs were identified by the discrimination functions as possible active lead; among them, we can find known drugs with other pharmacological properties and several natural products with different uses. At present, several experimental tests have been conducted in order to identify new leads and obtained an experimental corroboration of the obtained models. This result is the most important validation for the models developed here, because we have demonstrated that they are able to detect a series of compounds as active from a data-

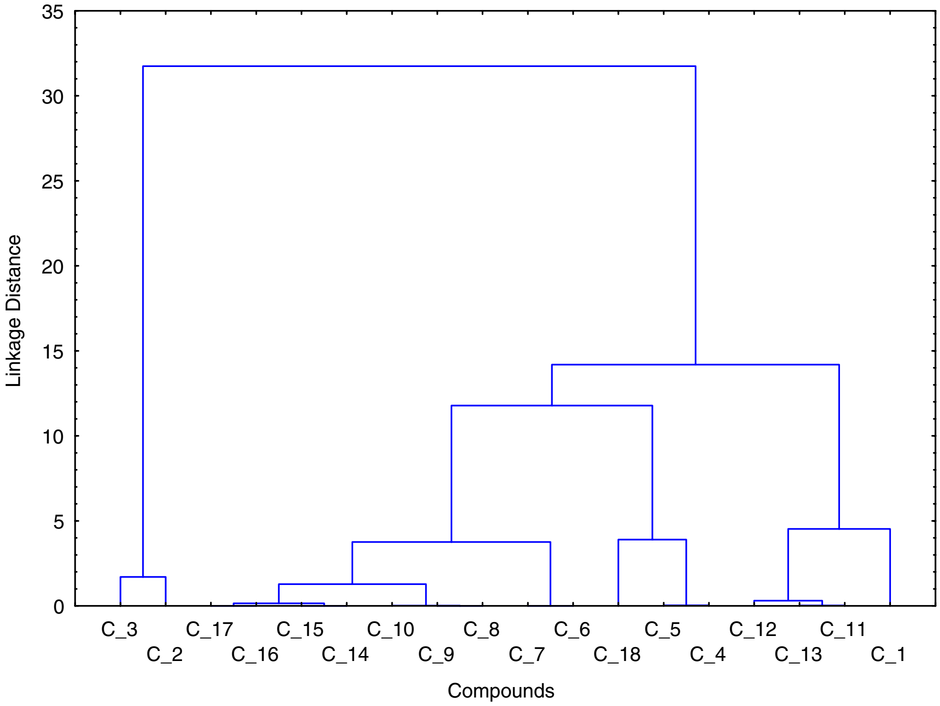


Figure 4. A dendrogram illustrating the results for the hierarchical *k*-NNCA of the set of active/inactive chemicals used for evaluating the predictive ability of the QSAR models for ligand-based virtual screening.

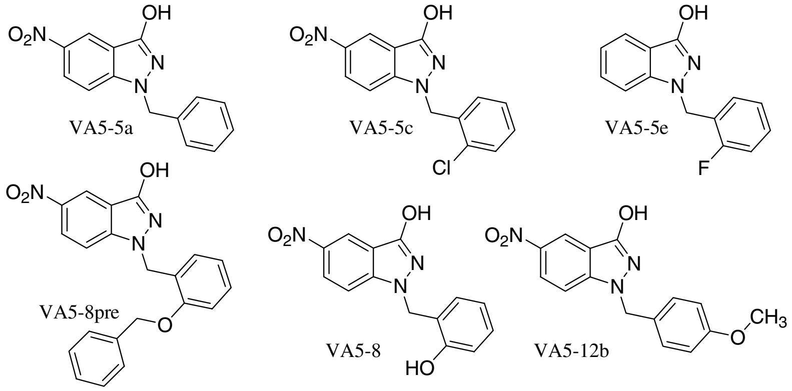


Figure 5. Chemical structures of the selected compounds.

base consisting of thousands of chemical compounds, and these chemicals have shown the predicted activity. However, even important, it is the fact that most of these compounds selected from the ligand-based virtual screening can have well-established methods of synthesis and, in many cases (for drugs), their toxicological, pharmacodynamical, and pharmaceutical properties are well known.

3.3.2. Lead discovery by ligand-based in silico screening: from dry selection to wet evaluation. One of the main objectives of the approach developed here is the selection of subsystems from a large group of chemical-organic compounds. A subsystem is understood, in general, as a number of compounds formed by a significant variation in a given parent structure, which is referred to as the lead compound.[94] The QSAR equations found by using the TOMOCOMD-CARDD approach recognize some structural patterns which are not related to the common patterns that appear in active compounds.

For this reason, in the present section we describe an experiment of lead generation through ligand-based in silico screening. This strategy permits the design of novel lead compounds having the desired activity, but first they need to be synthesized, then tested for the pharmacological activity and they finally need to pass the toxicological, pharmacodynamical, and pharmaceutical tests. As previously indicated, one of our research teams has worked mainly on trial-error searching for new antiprotozoan.[95–99]

In order to test the potential of TOMOCOMD-CARDD method and LDA for detecting novel antitrichomonal leads, we predicted the biological activity of all the chemicals contained in our ‘in-house’ collections of indazole, indazolols, indole, cinnoline, and quinoxaline derivatives, which have been recently obtained by our chemical synthesis team. On the basis of computer-aided predictions we selected potential trichomonacidal compounds (virtual hits). The following criteria were used for the hits’ selection: (1) compounds were selected as hits if the value of posterior probability of possessing antitrichomonal activity exceeded 95% ($\Delta P = 95\%$) by both LDA-based QSAR models, and (2) If, among the compounds designed by our chemical team, too many similar compounds satisfied criterion 1, then only several representative structures were selected. Here we perform in silico mining of a library of indazolols from synthetic sources,[100] for novel trichomonacidals, by using the discriminant functions obtained through the TOMOCOMD-CARDD method and LDA technique. Seven of these compounds (see Fig. 5) were initially evaluated with QSAR models as well as, in order to corroborate the predictions, prepared with excellent yields by economic and simple methods,[100] as well as in vitro evaluated against Tv. The results for the classification of the compounds in this set are summarized in Table 10. At the same time, this Table also depicts the $\Delta P\%$ values of the compounds in these series, as well as their canonical scores using all the developed models. The results shown in Table 10 exemplify how the present approach could be used for the selection/identification of novel leads, which may be used to treat the Tv.

In general, a good coincidence between the theoretical predictions and the observed activity for both active and inactive compounds was observed. The antitrichomonal activity of the seven compounds on Tv was studied (see Table 10). In these experiments, compounds VA5-5a, VA5-5c, and VA5-12b exhibited pronounced cytocidal activities at the concentrations of 100 and 10 μg/ml. In addition, compounds VA5-8pre and VA5-8 showed cytocidal and cytostatic activities at the concentrations of 100 μg/ml and 10 μg/ml, correspondingly (see Table 10). On the contrary, chemical VA5-5e resulted to be inactive at all assayed concentrations; coinciding with model predictions.

These results can be considered as a promising starting point for the future design and refinement of novel compounds with higher antitrichomonal activity with low toxicity. Therefore, compounds VA5-5a, VA5-5c, and VA5-12b were active at higher doses than Metronidazole (reference drug), but this result leaves a door open to a virtual variational study of the structure of these compounds in order to improve their activity. Therefore, these chemicals can be taken as hits, which are amenable for further chemistry optimization in order to derive the appropriate combination of potency, pharmacokinetic properties, toxicity, etc., as well as good activity in animal models. However, it is important to recall that the aim of this study is not to validate the model, but to provide an experimental example of how to use the model for potentially earlier drug discovery.

Table 10. Results of the computational evaluation using LDA-based QSAR models and percentages of cytostatic and/or cytocidal activity [brackets] for the three concentrations assayed in vitro against *Trichomonas vaginalis*

Compound	Theoretical results						*in vitro* activity (μg/ml)[h]						
	Class[a]	ΔP%[b]	Score[c]	Clas s[d]	ΔP%[e]	Score[f]	Class[g]	%C A_{24h} [% C_{24h}]			%C A_{48h} [% C_{48h}]		
								100	10	1	100	10	1
VA5-5a	+	95.09	-1.88	+	99.99	3.26	+	[100]	[97.66]	5.23	[100]	[99.18]	20.97
VA5-5c	+	96.53	-2.04	+	100.00	3.47	+	[99.6]	[94.13]	17.48	[100]	[98.91]	0.3
VA5-5e	-	-30.34	0.03	-	-71.93	-0.48	-	0.54	0.18	0	10.33	0	0
VA5-8pre	+	96.97	-2.10	+	100.00	3.59	+	[100]	[89.33]	0	[100]	83.43	6.99
VA5-8	+	96.97	-2.10	+	100.00	3.70	+	[100]	[90.05]	0	[100]	87.96	7.9
VA5-12b	+	96.97	-2.10	+	100.00	3.61	+	[100]	[94.77]	10.63	[100]	[92.64]	2.43

*The molecular structures of the compounds represented with codes are shown in Figure 2.

[a,d] In silico classification obtained from models **20** and **21** using non-stochastic and stochastic atom-type bilinear indices, respectively.

[b,e] Results for the classification of compounds obtained from models **20** and **21**, correspondingly: $\Delta P\% = [P(\text{active}) - P(\text{inactive})] \times 100$.

[c,f] Canonical scores obtained from models **20a** and **21a**, correspondingly.

[g] Observed (experimental activity) classification against *T. vaginalis*.

[h] Pharmacological activity of each tested compound, which was added to the cultures at doses of 100, 10, and 1 μg/ml: $\%CA_{\#}$ = cytostatic activity$_{(24\text{ or }48\text{ h})}$ and [$\%C_{\#}$] = cytocidal activity$_{(24\text{ or }48\text{ h})}$. MTZ = metronidazole (concentrations for metronidazole were 2, 1, and 0.5 mg/ml, respectively).

4. Conclusions

In conclusions, the combination of LDA and TOMOCOMD-CARDD MDs can contribute to the design of new compounds with antitrichomonal activity and possibly identify drugs with a broader spectrum of antiprotozoan activity. Computational screening of thousands of virtual and in silico molecules using this method in search for optimal substitutions is readily feasible and is far less costly than combinatorial chemistry and in vitro screening. The main conclusion of this report is that it has been able to develop QSAR models for the main step of drug discovery: lead identification (also known as lead generation step). These in silico models permit one to classify new 'physical' or 'virtual' chemicals as active or inactive ones in the chemotherapy of the trichomoniasis, and they will permit a more rational discovery of new lead compounds with antitrichomonal activity. In fact, this report showed the five new chemicals with potentialities in antitrichomonal therapeutics. These compounds possess structural features not seen in known trichomonacidal compounds and thus can serve as excellent leads for further optimization of antitrichomonal activity. The identification of this new family, making use of the TOMOCOMD-CARDD approach, constitutes an example of how this rational computer-aided design method can help to reduce cost, and to increase the rate in which NCEs progress through the pipeline.

5. Experimental

5.1. Computational strategies

Molecular fingerprints were generated by means of the interactive program for molecular design and bioinformatic research TOMOCOMD.[101] It consists of four subprograms; each one of them allows both drawing the structures (drawing mode) and calculating molecular 2D/3D descriptors (calculation mode). The modules are named computed-aided 'rational' drug design (CARDD), computed-aided modeling in protein science (CAMPS), computed-aided nucleic acid research (CANAR) and computed-aided bio-polymers docking (CABPD). The CARDD module was selected for drawing all the structures, as well as for the computation of non-stochastic and stochastic bilinear indices. The main steps for the application of this method in QSAR/QSPR and for drug design can be briefly summarized as follows:

1. Drawing of the molecular pseudographs for each molecule in the data set, using the drawing mode.
2. Use appropriate weights in order to differentiate the molecular atoms. The weights used in this study are those previously proposed for the calculation of the DRAGON descriptors,[55,56,102,103] that is, atomic mass (*M*), atomic polarizability (*P*), atomic electronegativity in Pauling scale (*E*), plus van der Waals atomic volume (*V*).The values of these atomic labels are shown in Table 2.
3. Computation of the total and local (atomic, group, and atom-type) bilinear indices can be carried out in the software calculation mode, where one can select the atomic properties and the descriptor family before calculating the MDs. This software generates a table in which the rows correspond to the compounds, and the columns correspond to the atom-based (both total and local) bilinear maps or other MD family implemented in this program.
4. Development of a QSPR/QSAR equation by using several multivariate analytical techniques, for instance, linear discriminant analysis. Therefore, one can find a quantitative relationship between an activity **A** and the bilinear fingerprints having, for instance, the following appearance:
$$\mathbf{A} = a_0\boldsymbol{b}_0(x,y) + a_1\boldsymbol{b}_1(x,y) + a_2\boldsymbol{b}_2(x,y) + \ldots . + a_k\boldsymbol{b}_k(x,y) + c \quad (22)$$
where **A** is the measured activity, $\boldsymbol{b}_k(x,y)$ are the *k*th non-stochastic total bilinear indices, and the $\boldsymbol{a_k}$*'s* are the coefficients obtained by the linear regression analysis.
5. Test of the robustness and predictive power of the QSPR/QSAR equation by using internal [leave-*group*-out (LGO)] and external (using a test set and an external predicting set) validation techniques.

5.2. Database selection

A data set of 123 organic-chemicals having a great structural variability was collected from the literature for the present study.[74,75] The data set of active compounds (50 chemicals used as trichomonacidal in clinic) was chosen, considering a representation of most of the different structural patterns and action modes for the case of compounds with antitrichomonal activity. In this sense, it is remarkable that the wide variability of drugs and mechanisms of action of active compounds in the training and prediction sets assures an adequate extrapolation power, and increases the possibilities of the discovery of new lead compounds with novel mechanisms of antitrichomonal activity, one of the most critical aspects in the construction of non-congeneric data. Therefore, this data set provides a useful tool for scientific research in synthesis, natural-product chemistry, theoretical chemistry, and other areas related to the field of antiprotozoal drugs.

In addition, the set of inactive compounds was obtained by selecting at random 73 drugs with different pharmacological uses. These drugs include, for instance, antibiotic, antivirals, sedatives/hypnotics, diuretics, anticonvulsants, hemostatics, oral hypoglycemics, antihypertensives, antihelminthics, anticancer, antifungal, etc, guaranteeing also a great structural variability. However, the declaration of these compounds as 'inactive' antitrichomonal *per se* does not guarantee that they do not exhibit trichomonacidal side-effects for some of those organic-chemical drugs that have been left undetected so far. This problem can be reflected in the results of classification for the series of inactive chemicals. Therefore, the developed LDA-based QSAR models can classify some of these compounds as active against the Tv, helping with the identification of new chemicals, among drugs from large data sets with other pharmacological uses.

Finally, two kinds of cluster analyses (CA) were performed for active and inactive series of compounds, in order to split (design) the data set into training and predicting series in a 'rational' way.

5.3. Data analysis and processing

Statistical analyses were performed using STATISTICA[77] software package. Cluster analysis and linear discriminant analysis modules were used to perform both series of compounds (training and test) and to get classification functions, correspondingly. After that, as a way to improve the statistical interpretation of the models, the Randić's method of orthogonalization was carried out. For a detailed description of all these methods please see the Supplementary data.

5.4. Chemical methods

The synthesis and characterization of seven indazolols, and cross references have been reported by other of our research teams.[100]

5.5. Determination of in vitro trichomonacidal activity

The biological activity was assayed on *T. vaginalis* JH31A#4 Ref. No. 30326 (ATCC, MD, USA) in modified Diamond medium supplemented with equine serum and grown at 37 °C (5% CO_2). The compounds were added to the cultures at several concentrations (100, 10, and 1 μg/ml) after 6 h of the seeding (0 h). Viable protozoa were assessed at 24 and 48 h after incubation at 37 °C by using the Neubauer chamber. Metronidazole (Sigma–Aldrich, SA, Spain) was used as reference drug at concentrations of 2, 1, and 0.5 μg/ml. Cytocidal and cytostatic activities were determined by calculation of percentages of cytocidal (%C) and cytostatic activities (%CA), in relation to controls as previously reported.[96,97]

Acknowledgments

One of the authors (M.-P.Y.) thanks the program 'Estades Temporals per a Investigadors Convidats' for a fellowship to work at Valencia University (2006–2007). F.T. thanks support from Spanish MEC DGI (Project No. CTQ2004-07768-C02-01/BQU) and Generalitat Valenciana (DGEUI INF01-051 and INFRA03-047, and OCYT GRUPOS03-173).

Supplementary data

A detailed description of cluster analysis, linear discriminant analysis, orthogonalization of descriptors, and 'virtual' and 'in silico' screening is available via the Internet at http://www.sciencedirect.com. Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2006.06.016.

References and notes

1. Petrin, D.; Delgaty, K.; Bhatt, R.; Garber, G. *Clin. Microbiol.* **1998**, *11*, 300.
2. Krieger, J. N. *Sex. Transm. Dis.* **2000**, *27*, 241.
3. World Health Organization. In *Global program on AIDS*. World Health Organization, Geneva, Switzerland, 1995, pp 2–27.
4. Wisdom, A. R.; Dunlop, E. M. C. *Br. J. Vener. Dis.* **1965**, *41*, 90.
5. Fouts, A. C.; Kraus, S. J. *J. Infect. Dis.* **1980**, *141*, 137.
6. Quan, M. *Clin. Cornerstone* **2000**, *3*, 36.
7. Cotch, M. F.; Pastorek, J. G., II; Nugent, R. P.; Hillier, S. L.; Gibbs, R. S.; Martin, D. H.; Eschenbach, D. A.; Edelman, R.; Carey, J. C.; Regan, J. A.; Krohn, M. A.; Klebanoff, M. A.; Rao, A. V.; Rhoads, G. G. *Sex. Transm. Dis.* **1997**, *24*, 353.
8. Bowden, F. J.; Garnett, G. P. *Sex. Transm. Infect.* **1999**, *75*, 372.
9. Hook, E. W., III *Sex. Transm. Dis.* **1999**, *26*, 388.
10. Sorvillo, F.; Kovacs, A.; Kerndt, P.; Stek, A.; Muderspach, L.; Sanchez-Keeland, L. *Am. J. Trop. Med. Hyg.* **1998**, *58*, 495.
11. Laga, M.; Manoka, A.; Kivuvu, M.; Malele, V.; Tuliza, M.; Nzila, N.; Goeman, J.; Behets, F.; Batter, V.; Alary, M.; Heyward, W. L.; Ryder, R. W.; Piot, P. *AIDS. London* **1993**, *7*, 95.
12. Centers for Disease Control and Prevention. Morbid. Mortal. Weekly Report, 1993, *42*, p 70.
13. Müller, M.; Lindmark, R. G. *Antimicrob. Agents Chemother.* **1976**, *9*, 696.
14. Müller, M. *Biochem. Pharmacol.* **1986**, *35*, 37.
15. Tocher, J. H.; Edwards, D. I. *Biochem. Pharmacol.* **1994**, *48*, 1089.
16. Nielsen, M. H. *Acta Pathol. Microbiol. Scand. Sect., B* **1976**, *84*, 93.
17. Arnold, M. *Ther. Umsch.* **1966**, *23*, 356.
18. Aure, J. C.; Gjonnaess, H. *Acta Obstet. Gynecol. Scand.* **1969**, *48*, 440.
19. de Carneri, I. In A. Corradetti, Ed.; In *Proceedings of the First International Congress of Parasitology*, Pergamon Press: New York, 1966;Vol. 1, pp 366–367.
20. de Carneri, I.; Baldi, G. F.; Giannone, R.; Passalia, S. *Arch. Ostet. Gynecol.* **1963**, *68*, 422.
21. Diddle, A. W. *Am. J. Obstet. Gynecol.* **1967**, *98*, 583.
22. Giannone, R. M. *Gynecology* **1972**, *24*, 354.
23. Kurnatowska, A. *Wiad Parazytol.* **1969**, *15*, 399.
24. Robinson, S. C. *Can. Med. Assoc. J.* **1962**, *86*, 665.
25. Korner, B.; Jensen, H. K. *Br. J. Vener. Dis.* **1976**, *52*, 404.
26. McFadzean, J. A.; Pugh, L. M.; Squires, S. L.; Whelan, J. P. F. *Br. J. Vener. Dis.* **1969**, *45*, 161.
27. Roe, F. J. C. *J. Antimicrob. Chemother.* **1977**, *3*, 205.
28. Kane, P. O.; McFadzean, J. A.; Squires, S. *Br. J. Vener. Dis.* **1961**, *37*, 276.
29. Nicol, C. S.; Evans, A. J.; McFadzean, J. A.; Squires, S. L. *Lancet* **1966**, 441.
30. Meingassner, J. G.; Thurner *J. Antimicrob. Agents Chemother.* **1979**, *15*, 254.
31. Apweiler, R. *Biosilico* **2003**, *I*, 5.
32. Xu, J.; Hagler, A. *Molecules* **2002**, *7*, 566.
33. Seifert, H. J. M.; Wolf, K.; Vitt, D. *Biosilico* **2003**, *1*, 143.
34. Dixit, K. S.; Mitra, S. N. *CRIPS* **2002**, *3*, 1.
35. Marrero-Ponce, Y. *J. Chem. Inf. Comput. Sci* **2004**, *44*, 2010.
36. Marrero-Ponce, Y. *Molecules* **2003**, *8*, 687.
37. Marrero-Ponce, Y.; Torrens, F. *Molecules*, submitted for publication (see also ECSOC-9, Conference Hall G, G-014).

38. Marrero-Ponce, Y.; Torrens, F. *J. Comput.-Aided Mol. Des.*, accepted for publication.
39. Marrero-Ponce, Y.; Torrens, F. *J. Phys. Chem. A.*, submitted for publication (see also ECSOC-9, Conference Hall G, G-015).
40. Marrero-Ponce, Y.; Torrens, F. *J. Mol. Struct.* (*THEOCHEM*), submitted for publication.
41. Marrero-Ponce, Y.; Castillo-Garit, J. A. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 369.
42. Marrero-Ponce, Y.; Huesca, G. A.; Ibarra, V. F. *J. Mol. Struct. (THEOCHEM)* **2005**, *717*, 67.
43. Marrero-Ponce, Y.; Montero, T. A.; Romero, Z. C.; Iyarreta, V. M.; Mayón, P. M.; García, S. R. *Bioorg. Med. Chem.* **2005**, *13*, 1293.
44. Marrero-Ponce, Y.; Iyarreta, V. M.; Montero, T. A.; Romero, Z. C.; Brandt, C. A.; Ávila, P. E.; Kirchgatter, K.; Machado, Y. *J. Chem. Inf. Model.* **2005**, *45*, 1082.
45. Marrero-Ponce, Y.; Medina, M. R.; Martinez, Y.; Torrens, F.; Romero, Z. V.; Castro, E. A.; Abalo, R. G. *J. Mol. Model.* **2006**, *12*, 255–271.
46. Meneses, M. A.; Marrero-Ponce, Y.; Machado-Tugores, Y.; Montero-Torres, A.; Montero Pereira, D.; Escario, J. A.; Nogal-Ruiz, J. J.; Ochoa, C.; Arán, V. J.; Martínez-Fernández, A. R.; García Sánchez, R. N. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3838.
47. Castillo-Garit, J. A.; Marrero-Ponce, Y.; Torrens, F. *Bioorg. Med. Chem.* **2006**, *14*, 2398–2408.
48. Marrero-Ponce, Y.; Cabrera, M. A.; Romero, Z. V.; Bermejo, M.; Siverio, D.; Torrens, F. *Internet Electron. J. Mol. Des.* **2005**, *4*, 124.
49. Marrero-Ponce, Y.; Meneses, M. A.; Machado, T. Y.; Montero, P. D.; Escario, J. A.; Nogal, R. J. J.; Ochoa, C.; Arán, V. J.; Martínez, F. A. R.; García, S. R. N.; Montero, T. A.; Torrens, F. *Curr. Drug Discov. Technol.* **2005**, *2*, 245–265.
50. Wang, R.; Gao, Y.; Lai, L. *Perspect. Drug Discov. Des.* **2000**, *19*, 47.
51. Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.* **2000**, *43*, 3714.
52. Ghose, A. K.; Crippen, G. M. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21.
53. Millar, K. J. *J. Am. Chem. Soc.* **1990**, *112*, 8533.
54. Gasteiger, J.; Marsilli, M. *Tetrahedron Lett.* **1978**, *34*, 3181–3184.
55. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity análisis*; Research Studies: Letchworth, UK, 1986.
56. Pauling, L. In *The Nature of Chemical Bond*; Cornell University: Ithaca, NY, 1939; pp 2–60.
57. Rouvray, D. H. In *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic: London, 1976; pp 180–181.
58. Trinajstić, N. *Chemical Graph Theory*; CRC: Boca Raton, FL, 1983; 2nd ed.; 1992; pp 32–33.
59. Gutman, I.; Polansky, O. E. In *Mathematical Concepts in Organic Chemistry*; Springer: Berlin, 1986.
60. Estrada, E.; Patlewicz, G. *Croat. Chim. Acta* **2004**, *77*, 203.
61. Klein, D. J. *Internet Electron. J. Mol. Des.* **2003**, *2*, 814.
62. Edwards, C. H.; Penney, D. E. In *Elementary Linear Algebra*; Prentice-Hall: Englewood Cliffs, NY, USA, 1988.
63. Jacobson, N. In *Basic Algebra I*, 2nd ed.; W.H. Freeman and Company: New York, 1985; pp 343–361.
64. Riley, K. F.; Hobson, M. P.; Vence, S. J. In *Mathematical Methods for Physics and Engineering*; Cambridge University: Cambridge, 1998; pp 228–236.
65. Hernández, E. *Álgebra y Geometría*. Universidad Autónoma de Madrid: Madrid, 1987; pp 521–544.
66. de Burgos-Román, J. *Álgebra y Geometría Cartesíana*. McGraw-Hill Interamericana de España, Madrid; 2^{da} Edición, 2000; pp 208–246.
67. de Burgos-Román, J. *Curso de Álgebra y Geometría*. Alambra Longman, Ed.; Madrid, **1994**; pp. 638-684.
68. Werner, G. In *Linear Algebra*, 4th ed.; Springer: New York, 1981; pp 261–288.
69. Randić, M. *J. Math. Chem* **1991**, *7*, 155.
70. Walker, P. D.; Mezey, P. G. *J. Am. Chem. Soc.* **1993**, *115*, 12423.
71. Kier, L. B.; Hall, L. H. In *Molecular Structure Description. The Electrotopological State*; Academic: New York, 1999.
72. Hall, L. H.; Story, C. T. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1004.
73. Gough, J. D.; Hall, L. H. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 356–361.
74. Negwer, M., Ed.; *Organic-Chemical Drugs and their Synonyms*, Akademie: Berlin, 1987.
75. Chapman & Hall. The Merck Index, ver. 12:3, 1999.
76. STATISTICA Software (data analysis software system), vs 6.0; StatSoft Inc., **2001**. <www.statsoft.com/>.
77. Estrada, E.; Peña, A.; García, D. R. J. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 583–595.
78. González-Díaz, H.; Marrero-Ponce, Y.; Hernández, I.; Bastida, I.; Tenorio, E.; Nasco, O.; Uriarte, U.; Castañedo, N.; Cabrera, M. A.; Aguila, E.; Marrero, O.; Morales, A.; Pérez, M. *Chem. Res. Toxicol.* **2003**, *16*, 1318.
79. de Julián-Ortiz, J. V.; de Alapont, C. G.; Ríos-Santamarina, I.; García-Doménech, R.; Gálvez, J. *J. Mol. Graphics Modell.* **1998**, *16*, 14.
80. Estrada, E.; Uriarte, E. *Curr. Med. Chem.* **2001**, *8*, 1573.
81. van de Waterbeemd, H. Discriminant Analysis for Activity Prediction. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; pp 265–288.
82. Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. *Bioinformatics* **2000**, *16*, 412.
83. Ford, M.-G.; Salt, D.-W. The Use of Canonical Correlation Analysis. In *Chemometric methods in molecular design*; van de Waterbeemd, H., Ed.; VCH: Weinhein, 1995; pp 283–292.
84. Mc Farland, J. W.; Gans, D. J. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; pp 295–307.
85. Johnson, R. A.; Wichern, D. W. *Applied Multivariate Statistical Analysis*; Prentice-Hall: Englewood Cliffs (NJ), 1988.
86. Randić, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311.
87. Randić, M. *New J. Chem.* **1991**, *15*, 517.
88. Randić, M. *J. Mol. Struct. (THEOCHEM)* **1991**, *233*, 45.
89. Wold, S.; Erikson, L. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: New York, 1995; pp 309–318.
90. Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell.* **2002**, *20*, 269.
91. Rose, K.; Hall, L. H.; Kier, L. B. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 651.
92. Estrada, E.; Peña, A. *Bioorg. Med. Chem.* **2000**, *8*, 2755.
93. Watson, C. *Biosilico* **2003**, *1*, 83.
94. Estrada, E.; Uriarte, E.; Montero, A.; Teijeira, M.; Santana, L.; De Clercq, E. *J. Med. Chem.* **2000**, *10*, 1975.
95. Aguirre, G.; Boiani, M.; Cerecetto, H.; Gerpe, A.; González, M.; Fernández, S. Y.; Denicola, A.; Ochoa de Ocáriz, C.; Nogal, R. J. J.; Montero, D.; Escario, J. A. *Arch. Pharm. (Weinheim, Ger.)* **2004**, *337*, 259.

96. Kouznetsov, V. V.; Rivero, C. J.; Ochoa, P. C.; Stashenko, E.; Martínez, J. R.; Montero, P. D.; Nogal, R. J. J.; Fernández, P. C.; Muelas, S. S.; Gómez, B. A.; Bahsas, A.; Amaro, L. J. *Arch. Pharm. (Weinheim, Ger.)* **2005**, *338*, 1.

97. Kouznetsov, V. V.; Vargas, M. L. Y.; Tibaduiza, B.; Ochoa, C.; Montero, P. D.; Nogal, R. J. J.; Fernández, C.; Muelas, S.; Gómez, A.; Bahsas, A.; Amaro-Luis, J. *Arch. Pharm. (Weinheim, Ger.)* **2004**, *337*, 127.

98. Ochoa, A.; Pérez, E.; Pérez, R.; Suárez, M.; Ochoa, E.; Rodríguez, H.; Gómez, A.; Muelas, S.; Nogal, R. J. J.; Martínez, R. A. *Arzneim.-Forsch./Drug Res.* **1999**, *49*, 764.

99. Meneses, M. A.; Montero, D.; Escario, J.A;, Nogal, R.J.J.; Ochoa, C.; Arán, V.J.; Martínez, F.A.R. IX European Multicolloquium of Parasitology, Valencia, July 18-23, 2004; Spain.

100. Arán, V. J.; Flores, M.; Muñoz, P.; Páez, J. A.; Sánchez, V. P.; Stud, M. *Liebigs Ann. Chem.* **1996**, 683.

101. Marrero-Ponce Y.; Romero, V. 2002; TOMOCOMD software. Central University of Las Villas. TOMOCOMD (topological molecular computer design) for Windows, version 1.0 is a preliminary experimental version; in future a professional version will be obtained upon request to Y. Marrero: yovanimp@qf.uclv.edu.cu or ymarrero77@yahoo.es.

102. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682.

103. Todeschini, R.; Gramatica, P. *Perspect. Drug Discovery Des.* **1998**, *9*, 355.

104. Gavini, E.; Juliano, C.; Mulé, A.; Pirisino, G. *Fàrmaco* **1997**, *52*, 67.

105. Gavini, E.; Juliano, C.; Mulé, A.; Pirisino, G.; Murineddu, G.; Pinna, A. *Arch. Pharm. (Weinheim, Ger.)* **2000**, *333*, 341.

106. Alcalde, E.; Pérez, L.; Dinarés, I.; Frigola *J. Chem. Pharm. Bull.* **1995**, *43*, 493.

107. Aguirre, G.; Boiani, M.; Cerecetto, H.; Gerpe, A.; González, M.; Fernandez, S. Y.; Denicola, A.; Ochoa de Ocáriz, C.; Nogal, R. J. J.; Montero, D.; Escario, J. A. *Arch. Pharm. (Weinheim, Ger.)* **2004**, *337*, 259.