

Bond-based linear indices in QSAR: computational discovery of novel anti-trichomonal compounds

Yovani Marrero-Ponce · Alfredo Meneses-Marcel · Oscar M. Rivera-Borroto · Ramón García-Domenech · Jesus Vicente De Julián-Ortiz · Alina Montero · José Antonio Escario · Alicia Gómez Barrio · David Montero Pereira · Juan José Nogal · Ricardo Grau · Francisco Torrens · Christian Vogel · Vicente J. Arán

Received: 6 November 2006 / Accepted: 5 January 2008 / Published online: 16 May 2008
© Springer Science+Business Media B.V. 2008

Abstract *Trichomonas vaginalis* (Tv) is the causative agent of the most common, non-viral, sexually transmitted disease in women and men worldwide. Since 1959, metronidazole (MTZ) has been the drug of choice in the systemic treatment of trichomoniasis. However, resistance to MTZ in some patients and the great cost associated with the development of new trichomonacids make necessary the development of computational methods that shorten the drug discovery pipeline. Toward this end, bond-based linear indices, new **TOMOCOMD-CARDD** molecular descriptors, and linear discriminant analysis were used to discover novel trichomonacidal chemicals. The obtained models, using non-stochastic and stochastic indices, are able to classify correctly 89.01% (87.50%) and 82.42% (84.38%) of the chemicals in the training (test) sets, respectively. These results validate the models for their use

in the ligand-based *virtual* screening. In addition, they show large Matthews' correlation coefficients (*C*) of 0.78 (0.71) and 0.65 (0.65) for the training (test) sets, correspondingly. The result of predictions on the 10% *full-out* cross-validation test also evidences the robustness of the obtained models. Later, both models are applied to the *virtual* screening of 12 compounds already proved against Tv. As a result, they correctly classify 10 out of 12 (83.33%) and 9 out of 12 (75.00%) of the chemicals, respectively; which is the most important criterion for validating the models. Besides, these classification functions are applied to a library of seven chemicals in order to find novel antitrichomonal agents. These compounds are synthesized and tested for in vitro activity against Tv. As a result, experimental observations approached to theoretical predictions, since it was obtained a correct classification of 85.71% (6 out of 7) of the chemicals. Moreover, out of the

Y. Marrero-Ponce · A. Meneses-Marcel ·
O. M. Rivera-Borroto · A. Montero
Faculty of Chemistry-Pharmacy, Unit of Computer-Aided
Molecular "Biosilico" Discovery and Bioinformatic Research
(CAMD-BIR Unit), Central University of Las Villas, Santa Clara
54830, Villa Clara, Cuba

Y. Marrero-Ponce · F. Torrens
Institut Universitari de Ciència Molecular, Universitat de
València, Edifici d'Instituts de Paterna, P.O. Box 22085,
Valencia 46071, Spain

Y. Marrero-Ponce (✉) · R. García-Domenech ·
J. V. De Julián-Ortiz
Departamento de Química Física, Facultad de Farmacia, Unidad
de Investigación de Diseño de Fármacos y Conectividad
Molecular, Universitat de València, Valencia, Spain
e-mail: ymarrero77@yahoo.es; ymponce@gmail.com;
yovanimp@uclv.edu.cu
URL: <http://www.uv.es/yoma/>

A. Meneses-Marcel · J. A. Escario · A. G. Barrio ·
D. M. Pereira · J. J. Nogal
Departamento de Parasitología, Facultad de Farmacia, UCM,
Pza. Ramón y Cajal s/n, Madrid 28040, Spain

O. M. Rivera-Borroto · R. Grau
Faculty of Mathematics, Physics & Computer Science, Center
of Studies on Informatics, Central University of Las Villas,
Santa Clara 54830, Villa Clara, Cuba

C. Vogel
Institut für Chemie, Universität Rostock, Abteilung für
Organische Chemie, Albert-Einstein-Straße 3a 18059 Rostock,
Germany

V. J. Arán
Instituto de Química Médica, CSIC, c/ Juan de la Cierva 3,
Madrid 28006, Spain

seven compounds that are screened, synthesized and biologically assayed, six compounds (VA7-34, VA7-35, VA7-37, VA7-38, VA7-68, VA7-70) show pronounced cytotoxic activity at the concentration of 100 µg/ml at 24 h (48 h) within the range of 98.66%–100% (99.40%–100%), while only two molecules (chemicals VA7-37 and VA7-38) show high cytotoxic activity at the concentration of 10 µg/ml at 24 h (48 h): 98.38% (94.23%) and 97.59% (98.10%), correspondingly. The LDA-assisted QSAR models presented here could significantly reduce the number of synthesized and tested compounds and could increase the chance of finding new chemical entities with antitrichomonal activity.

Keywords TOMOCOMD-CARDD software · Bond-based linear indices · LDA-assisted QSAR model · Virtual screening · Trichomonacidal · In vitro cytostatic and cytotoxic activities

Background

Trichomonas vaginalis (Tv) is a common sexually transmitted infection, which is increasingly recognized as an important disease in both women and men [1, 2]. Few years ago, the World Health Organization estimated the number of adults with trichomoniasis at 170 million worldwide, more than the combined numbers of gonorrhea, syphilis, and chlamydia [3].

In 1959, a nitroimidazole derivative of a *Streptomyces* antibiotic, azomycin, was found to be highly effective in the systemic treatment of trichomoniasis [4]. This derivative was α,β -hydroxyethyl-2-methyl-5-nitroimidazole, commonly referred as metronidazole (MTZ) and marketed under the trade name Flagyl.

The recommended MTZ regimen results in cure rates of approximately 95% [5]. In addition, it is remarkably safe compared to the most toxic antiprotozoan products [6]. However, resistance to MTZ has been proven to be geographically widely distributed, and no clustering or temporal trends in patients have been observed [7]. A good alternative to palliate this problem could be the clinical treatment with other nitroimidazoles, but unfortunately all of them have modes of antibacterial activity similar to that of MTZ [8] and therefore, the resistance to MTZ often includes resistance to the other nitroimidazoles [9].

Also, in patients who do not respond to high-dose MTZ therapy, a variety of regimens have been evaluated for possible effectiveness, with rare or only occasional success. These include zinc sulfate, povidone-iodine douche, arsenicals, nonoxynol-9 cream, mebendazole, albendazole, furazolidone, and rifabutin [10–15]. These agents, although

they demonstrate considerable in vitro activity, have been clinically disappointing. Paromomycin was previously reported to be useful in the management of resistant trichomoniasis. It was used fairly effectively (cure rate, 58%) in 12 patients and remains an important option; however, local side effects were considerable and can be quite severe [16, 17].

Currently, it is clear that new trichomonacidal agents are needed to treat resistant organisms. However, the great cost associated to the development of new compounds and the small economic size of the market for antiprotozoan drugs makes this development slow. Therefore, it is necessary to develop computational methods permitting theoretical—in silico—evaluations of trichomonacidal activity for virtual libraries of chemicals, before these compounds are synthesized in the laboratory [18, 19].

In this context, our research group has recently introduced a novel scheme to perform rational—in silico—molecular design (or selection/identification of lead drug-like chemicals) and QSAR/QSPR studies, known as **TOMOCOMD-CARDD** (acronym of *To* pological *MO*lecular *COM*puter *DES*ign-Computer Aided “*R*ational” *DRUG* Design) [20]. This method has been developed to generate 2D (topological), 2.5 (3D-chiral) and 3D (topographical and geometrical) molecular descriptors based on the application of discrete mathematics and linear algebra theory to chemistry. This in silico method has been successfully applied to the prediction of several physical, physicochemical, chemical and biochemical properties of organic compounds [21–38].

Recently, some of the present authors have proposed a new extended local (bond and bond-type) as well as total molecular descriptors (MDs) based on the adjacency of edges as well as on quadratic, bilinear and linear maps, which are similar to those typically defined by mathematicians in linear algebra. These researchers also proposed a new matrix representation of the molecule on the “stochastic” adjacency of edges and quadratic (linear) indices derived from there. These descriptors, called bond-based quadratic, bilinear and linear indices, encode topological information given by the associated molecular graph, weighted by chemical information encoded in selected bond weightings. Finally, the correlation ability of the new descriptors is tested in QSPR and QSAR studies [39, 40].

The main objective of the present report was to use non-stochastic and stochastic bond-type linear indices to generate predictive LDA (linear discriminant analysis)-assisted QSAR models, enabling the selection of novel drug-like compounds with antitrichomonal activity. The in vitro evaluation of a new series of heterocyclic compounds with antitrichomonal activity is also presented.

Theoretical framework

Background in graph-theoretical edge-adjacency matrix

Let $G = \langle V, E \rangle$ be a simple graph, with $V = (v_1, v_2, \dots, v_n)$ and $E = (e_1, e_2, \dots, e_m)$ being the vertex- and edge-sets of G , respectively. Then G represents a molecular graph having n vertices and m edges (bonds). The edge-adjacency matrix \mathbf{E} of G (likewise called bond-adjacency matrix, \mathbf{B}) is a square and symmetric matrix whose elements e_{ij} are 1 if and only if edge i is adjacent to edge j [41–43]. Two edges are adjacent if they are incidental to a common vertex. This matrix corresponds to the vertex-adjacency matrix of the associated line graph. Finally, the sum of the i th row (or column) of \mathbf{E} is named the edge degree of bond i , δ_i [41, 44–46].

New edge-relations: stochastic edge-adjacency matrix

By using the edge (bond)-adjacency relationships we can find another new relation for a molecular graph, which will be introduced here. The k th stochastic edge-adjacency matrix, \mathbf{ES}^k , can be obtained directly from \mathbf{E}^k (obtained by inner or scalar product of \mathbf{E} by itself k times). Here, $\mathbf{ES}^k = [es_{ij}^k]$ is a square table of order m (m = number of bonds), and the elements es_{ij}^k are defined as it follows:

$$es_{ij}^k = \frac{e_{ij}^k}{\sum_i e_{ij}^k} = \frac{e_{ij}^k}{\delta_i^k} \quad (1)$$

where, e_{ij}^k are the elements of the k th power of \mathbf{E} and the sum $\left(\sum_i e_{ij}^k\right)$ of the i th row of \mathbf{E}^k is named the k -order edge degree of bond i (δ_i^k). Notice that the matrix \mathbf{ES}^k in Eq. 1 has the property that the sum of the elements in each row is 1. Such an $m \times m$ matrix, with nonnegative entries having this property, is called a “stochastic matrix” [47].

Chemical information and bond-based molecular vector

The atom-based molecular vector (\mathbf{x}), used to represent small-to-medium-sized organic chemicals, has been explained in some detail elsewhere [24, 29, 32, 34, 48]. In a way parallel to the development of \mathbf{x} , we present the extension to the bond-based molecular vector \mathbf{w} . The components (w_i) of \mathbf{w} are numerical values, which represent a certain standard bond property (bond-label). Namely, these weights correspond to different bond properties for organic molecules. Thus, a molecule having 5, 10, 15, ..., m bonds can be represented by means of vectors, with 5, 10, 15, ..., m components, belonging to the spaces \mathbb{R}^5 , \mathbb{R}^{10} , \mathbb{R}^{15} , ..., \mathbb{R}^m , respectively, where m is the dimension of the real vector space \mathbb{R}^m .

Former properties characterize each kind of bond (and bond-type) within the molecule. Diverse kinds of bond weights (w_i) can be used in order to codify information related to each bond in the molecule. These bond labels are chemically meaningful numbers such as standard bond distance, standard bond dipole [19, 49–51] or even mathematical expressions involving atomic weights such as atomic partition coefficients (log P) [52], surface area contributions of polar atoms [53], atomic molar refractivities [54], atomic hybrid polarizabilities [55], as well as Gasteiger-Marsilli atomic net charges [56], atomic electronegativities in Pauling scale [57], and so on. Here, we characterized each bond (between atomic nuclei i and j) with the following parameter:

$$w_{ij} = \frac{x_i}{\delta_i} + \frac{x_j}{\delta_j} \quad (2)$$

In this expression x_i can be any standard weight of the atom i bonded with atom j . The δ_i is the vertex (atom) degree of atom i . The use of each scale (bond property) defines alternative molecular vectors \mathbf{w}

Calculation of linear indices

If a molecule consists of m bonds (vectors of \mathbb{R}^m), then the k th linear indices for such a molecule are calculated from linear maps on \mathbb{R}^m (endomorphism on \mathbb{R}^m) in canonical basis set. Specifically, the k th linear maps, $\mathbf{f}_k(\mathbf{w})$ and ${}^s\mathbf{f}_k(\mathbf{w})$, are computed from the k th non-stochastic and stochastic edge-adjacency matrices, \mathbf{E}^k and \mathbf{ES}^k , as shown in Eqs. 3 and 4, respectively:

$$\mathbf{f}_k(\mathbf{w}) = \left[\sum_{j=1}^m e_{ij}^k w_j \right] = \mathbf{E}^k \cdot \mathbf{w} \quad (3)$$

$${}^s\mathbf{f}_k(\mathbf{w}) = \left[\sum_{j=1}^m es_{ij}^k w_j \right] = \mathbf{ES}^k \cdot \mathbf{w} \quad (4)$$

where, m is the number of bonds in the molecule, and w_j are the coordinates of the bond-based molecular vector (\mathbf{w}) in the so-called canonical (‘natural’) basis set. In this basis system, the coordinates of any vector \mathbf{w} coincide with the components of this vector [47, 58, 59]. Therefore, those coordinates can be considered as weights (bond-labels) of the edge of the molecular graph. The coefficients e_{ij}^k and es_{ij}^k are the elements of the k th power of the matrices $\mathbf{E}(G)$ and $\mathbf{ES}(G)$, correspondingly, of the molecular graph. The symbol “ \cdot ” denotes the scalar product between matrices.

Notice that both linear maps are defined as a linear transformation $\mathbf{f}_k(\mathbf{w})$ on molecular vector space \mathbb{R}^m . This map is a correspondence that assigns a vector ${}^s\mathbf{f}_k(\mathbf{w})$ to a vector \mathbf{w} in \mathbb{R}^m so that:

$$\mathbf{f}(\lambda_1 \mathbf{w}_1 + \lambda_2 \mathbf{w}_2) = \lambda_1 \mathbf{f}(\mathbf{w}_1) + \lambda_2 \mathbf{f}(\mathbf{w}_2) \quad (5)$$

For any scalars λ_1, λ_2 and any vectors $\mathbf{w}_1, \mathbf{w}_2$ in \mathbb{R}^m .

Total (whole-molecule) bond-based non-stochastic and stochastic linear indices, $f_k(\mathbf{w})$ and $^s f_k(\mathbf{w})$, are calculated from local (bond) linear indices as shown in Eqs. 6 and 7, correspondingly:

$$f_k(\mathbf{w}) = \sum_{i=1}^m f_{ki}(\mathbf{w}) = \sum_{i=1}^m \sum_{j=1}^m e_{ij}^k w_j = \sum_{i=1}^m \mathbf{u}_i^t \cdot \mathbf{E}^k \cdot \mathbf{w} \quad (6)$$

$$^s f_k(\mathbf{w}) = \sum_{i=1}^m ^s f_{ki}(\mathbf{w}) = \sum_{i=1}^m \sum_{j=1}^m e_{ij}^k w_j = \sum_{i=1}^m \mathbf{u}_i^t \cdot \mathbf{E}^k \cdot \mathbf{w} \quad (7)$$

where, m is the number of bonds, $f_{ki}(\mathbf{w})$ and $^s f_{ki}(\mathbf{w})$ are the local non-stochastic and stochastic linear indices obtained by Eqs. 3 and 4 as the coordinates of $\mathbf{f}_k(\mathbf{w})$ and $^s \mathbf{f}_k(\mathbf{w})$ referred to the canonical basis, respectively. This means that, whenever we calculate total bond-based non-stochastic and stochastic linear indices in fact we are calculating the 1-norm associated to the well known linear maps $\mathbf{f}_k(\mathbf{w})$ and $^s \mathbf{f}_k(\mathbf{w})$, defined in mathematical analysis for such vector spaces as \mathbb{R}^m , \mathbf{u}_i^t is some m -dimensional canonical row vector [58].

Finally, in addition to total (whole molecule) and local (bond) linear indices computed for each bond in the molecule, a fragments formalism (bond-type) can be developed. *The k th bond-type linear index of the edge-adjacency matrix is calculated by summing up the k th local linear indices of all bonds of the same bond type in the molecule.* Consequently, if abstractly a molecule (its corresponding graph) is partitioned into X molecular fragments, the total linear indices can be partitioned into T fragment (bond-type) linear indices, $t = 1, 2, \dots, X$. Then, the total linear indices of order k can be expressed as the sum of the local linear indices of the X fragments for the same order:

$$f_k(\mathbf{w}) = \sum_{i=1}^m f_{ki}(\mathbf{w}) = \sum_{t=1}^X f_{kt}(\mathbf{w}) \quad X, m \in \mathbb{N} \wedge X \leq m \quad (8)$$

$$^s f_k(\mathbf{w}) = \sum_{i=1}^m ^s f_{ki}(\mathbf{w}) = \sum_{t=1}^X ^s f_{kt}(\mathbf{w}) \quad X, m \in \mathbb{N} \wedge X \leq m \quad (9)$$

In the bond-type linear indices formalism, each bond in the molecule is classified into a bond-type (fragment). To this effect, bonds may be classified into bond types in terms of the characteristics of the two atoms that define the bond. For all data sets, including those with a common molecular scaffold as well as those with very diverse structure, the k th fragment (bond-type) linear indices provide much useful information. Thus, the development of the bond-type linear indices description provides a basis for the application to a wider range of biological problems in which the local formalism is appropriated without the need for superposition of a closely related set of structures.

It should be highlighted that our approach is analogous to the **LCBO-MO** (Linear Combination of Bond Orbitals-Molecular Orbitals) method (e.g., for $k = 1$) [60]. **LCBO-MO** is another way of forming molecular orbitals by taking linear combinations of functions associated with the different bonds in the molecule. To this effect, MOs are made up as LCBO of bonds composing the system, i.e. they are written in the form

$$\varphi_i = \sum_{j=1}^m c_{ij} \psi_j \quad (10)$$

where, i is the number of the **MO** φ (in our case, $f_i(\mathbf{w})$), j is the number of bond ψ —orbitals (in our case, w_j) and c_{ij} (in our case, e_{ij}^1 or e_{ij}^1 for non-stochastic and stochastic indices, respectively) are the numerical coefficients defining the contributions of individual **BOs** to the given **MO**. Although the **LCAO** (Linear Combination of Atom Orbitals) approximation has been particularly useful for the study of conjugated hydrocarbons, the **LCBO** method has been chiefly applied to the calculation of the properties of saturated hydrocarbons. As a saturated molecule can be considered as made up of localized bonds, it is reasonable to associate an orbital to each of the corresponding regions [60].

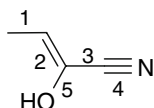
In addition, both \mathbf{E}^k and \mathbf{E}^k can be seen as graph-theoretical electronic-structure models [61]. In fact, quantum chemistry starts from the fact that a molecule is made up of both electrons and nuclei. The distinction here between bonded and non-bonded atoms is difficult to justify. Any two nuclei in a molecule interact directly and indirectly through the electrons present in the molecule. Only the intensity of this interaction varies on going from one pair of nuclei to another. To this effect, an electron in an arbitrary bond i can move (step-by-step) to other bonds at different discrete time periods t_k ($k = 0, 1, 2, \dots$) through the chemical-bonding network. That is to say, both \mathbf{E}^1 and \mathbf{E}^1 matrices consider the valence-bond electrons in one step; their powers ($k = 0, 1, 2, 3, \dots$) can be considered as an interacting-electron chemical-network model in k steps. Thus, this model can be seen as an intermediate between the quantitative quantum-mechanical Schrödinger equation and classical chemical bonding ideas [61]. Taking into account the former interpretations on our method, we propose the name of **LCTBW-LF** or *Linear Combination of Topochemical Bond Weights-Local Fingerprints* in analogy to LCBO-MO method.

Sample calculation

The linear indices of the bond matrix are calculated in the following way:

Let's consider the molecule of 2-hydroxybut-2-enenitrile as a simple example; from here we obtain the following

labeled molecular graph as well as both bond-based adjacency matrices \mathbf{E} and \mathbf{ES} . The second ($k = 2$) and third ($k = 3$) powers of these matrices and the bond-based molecular vector, $\bar{\mathbf{w}}$ are also given:



$$\begin{aligned} f_{01} &= 1 \cdot 3.4000 + 0 \cdot 1.4875 + 0 \cdot 1.2750 + 0 \cdot 1.6508 \\ &\quad + 0 \cdot 4.0775 = 3.4000 \\ f_{02} &= 0 \cdot 3.4000 + 1 \cdot 1.4875 + 0 \cdot 1.2750 + 0 \cdot 1.6508 \\ &\quad + 0 \cdot 4.0775 = 1.4875 \\ f_{03} &= 0 \cdot 3.4000 + 0 \cdot 1.4875 + 1 \cdot 1.2750 + 0 \cdot 1.6508 \\ &\quad + 0 \cdot 4.0775 = 1.2750 \end{aligned}$$

$$\begin{aligned} \mathbf{E}^0 = \mathbf{ES}^0 &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} & \mathbf{E}^1 &= \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} & \mathbf{E}^2 &= \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 3 & 1 & 1 & 1 \\ 1 & 1 & 3 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 2 \end{bmatrix} & \mathbf{E}^3 &= \begin{bmatrix} 0 & 3 & 1 & 1 & 1 \\ 3 & 2 & 5 & 1 & 4 \\ 1 & 5 & 2 & 3 & 4 \\ 1 & 1 & 3 & 0 & 1 \\ 1 & 4 & 4 & 1 & 2 \end{bmatrix} \\ \mathbf{ES}^1 &= \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0.33 & 0 & 0.33 & 0 & 0.33 \\ 0 & 0.33 & 0 & 0.33 & 0.33 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 \end{bmatrix} & \mathbf{ES}^2 &= \begin{bmatrix} 0.33 & 0 & 0.33 & 0 & 0.33 \\ 0 & 0.5 & 0.16 & 0.16 & 0.16 \\ 0.16 & 0.16 & 0.5 & 0 & 0.16 \\ 0 & 0.33 & 0 & 0.33 & 0.33 \\ 0.16 & 0.16 & 0.16 & 0.16 & 0.33 \end{bmatrix} & \mathbf{ES}^3 &= \begin{bmatrix} 0 & 0.5 & 0.16 & 0.16 & 0.16 \\ 0.2 & 0.13 & 0.33 & 0.06 & 0.26 \\ 0.06 & 0.33 & 0.13 & 0.2 & 0.26 \\ 0.16 & 0.16 & 0.5 & 0 & 0.16 \\ 0.083 & 0.33 & 0.33 & 0.083 & 0.16 \end{bmatrix} \end{aligned}$$

The molecule contains five localized bonds (corresponding to five edges in the H-suppressed molecular graph). To these we will associate the five “bond orbitals” w_1, w_2, w_3, w_4 , and w_5 . Thus, $\mathbf{w} = [w_1, w_2, w_3, w_4, w_5] = [w_{(C-C)}, w_{(C=C)}, w_{(C-C)}, w_{(C \equiv N)}, w_{(C-O)}]$ and each “bond orbital” can be computed by Eq. 2 by using, for instance, the atomic electronegativity in Pauling scale (x_i) [57] as atomic weight (atom-label):

$$\begin{aligned} w_1 &= x_C/1 + x_C/3 = 2.55/1 + 2.55/3 = 3.4000 \\ w_2 &= x_C/3 + x_C/4 = 2.55/3 + 2.55/4 = 1.4875 \\ w_3 &= x_C/4 + x_C/4 = 2.55/4 + 2.55/4 = 1.2750 \\ w_4 &= x_C/4 + x_N/3 = 2.55/4 + 3.04/3 = 1.6508 \\ w_5 &= x_C/4 + x_O/1 = 2.55/4 + 3.44/1 = 4.0775 \end{aligned}$$

And, therefore, $\mathbf{w} = [3.4000, 1.4875, 1.2750, 1.6508, 4.0775]$.

K th ($k = \overline{0, 3}$) non-stochastic “molecular orbitals” (bond linear indices) can be calculated for this molecule as it follows:

$$\begin{aligned} f_{0i}(\mathbf{w}) &= \mathbf{u}_i^t \cdot \mathbf{E}^0 \cdot \mathbf{w} = \sum_{j=1}^5 e_{ij}^0 w_j \\ &= e_{i1}^0 w_1 + e_{i2}^0 w_2 + e_{i3}^0 w_3 + e_{i4}^0 w_4 + e_{i5}^0 w_5 \end{aligned}$$

$$\begin{aligned} f_{04} &= 0 \cdot 3.4000 + 0 \cdot 1.4875 + 0 \cdot 1.2750 + 1 \cdot 1.6508 \\ &\quad + 0 \cdot 4.0775 = 1.6508 \end{aligned}$$

$$\begin{aligned} f_{05} &= 0 \cdot 3.4000 + 0 \cdot 1.4875 + 0 \cdot 1.2750 + 0 \cdot 1.6508 \\ &\quad + 1 \cdot 4.0775 = 4.0775 \end{aligned}$$

$$\begin{aligned} f_{1i}(\mathbf{w}) &= \mathbf{u}_i^t \cdot \mathbf{E}^1 \cdot \mathbf{w} = \sum_{j=1}^5 e_{ij}^1 w_j \\ &= e_{i1}^1 w_1 + e_{i2}^1 w_2 + e_{i3}^1 w_3 + e_{i4}^1 w_4 + e_{i5}^1 w_5 \end{aligned}$$

$$\begin{aligned} f_{11} &= 0 \cdot 3.4000 + 1 \cdot 1.4875 + 0 \cdot 1.2750 + 0 \cdot 1.6508 \\ &\quad + 0 \cdot 4.0775 = 1.4875 \end{aligned}$$

$$\begin{aligned} f_{12} &= 1 \cdot 3.4000 + 0 \cdot 1.4875 + 1 \cdot 1.2750 + 0 \cdot 1.6508 \\ &\quad + 1 \cdot 4.0775 = 8.7525 \end{aligned}$$

$$\begin{aligned} f_{13} &= 0 \cdot 3.4000 + 1 \cdot 1.4875 + 0 \cdot 1.2750 + 1 \cdot 1.6508 \\ &\quad + 1 \cdot 4.0775 = 7.2158 \end{aligned}$$

$$\begin{aligned} f_{14} &= 0 \cdot 3.4000 + 0 \cdot 1.4875 + 1 \cdot 1.2750 + 0 \cdot 1.6508 \\ &\quad + 0 \cdot 4.0775 = 1.2750 \end{aligned}$$

$$\begin{aligned} f_{15} &= 0 \cdot 3.4000 + 1 \cdot 1.4875 + 1 \cdot 1.2750 + 0 \cdot 1.6508 \\ &\quad + 0 \cdot 4.0775 = 2.7625 \end{aligned}$$

$$\begin{aligned}
 f_{2i}(\mathbf{w}) &= \mathbf{u}_i^t \cdot \mathbf{E}^2 \cdot \mathbf{w} = \sum_{j=1}^5 e_{ij}^2 w_j \\
 &= e_{i1}^2 w_1 + e_{i2}^2 w_2 + e_{i3}^2 w_3 + e_{i4}^2 w_4 + e_{i5}^2 w_5 \\
 f_{21} &= 1*3.4000 + 0*1.4875 + 1*1.2750 + 0*1.6508 \\
 &\quad + 1*4.0775 = 8.7525 \\
 f_{22} &= 0*3.4000 + 3*1.4875 + 1*1.2750 + 1*1.6508 \\
 &\quad + 1*4.0775 = 11.4658 \\
 f_{23} &= 1*3.4000 + 1*1.4875 + 3*1.2750 + 0*1.6508 \\
 &\quad + 1*4.0775 = 12.7900 \\
 f_{24} &= 0*3.4000 + 1*1.4875 + 0*1.2750 + 1*1.6508 \\
 &\quad + 1*4.0775 = 7.2158 \\
 f_{25} &= 1*3.4000 + 1*1.4875 + 1*1.2750 + 1*1.6508 \\
 &\quad + 2*4.0775 = 15.9683
 \end{aligned}$$

$$\begin{aligned}
 f_{3i}(\mathbf{w}) &= \mathbf{u}_i^t \cdot \mathbf{E}^3 \cdot \mathbf{w} = \sum_{j=1}^5 e_{ij}^3 w_j \\
 &= e_{i1}^3 w_1 + e_{i2}^3 w_2 + e_{i3}^3 w_3 + e_{i4}^3 w_4 + e_{i5}^3 w_5 \\
 f_{31} &= 0*3.4000 + 3*1.4875 + 1*1.2750 + 1*1.6508 \\
 &\quad + 1*4.0775 = 11.4658 \\
 f_{32} &= 3*3.4000 + 2*1.4875 + 5*1.2750 + 1*1.6508 \\
 &\quad + 4*4.0775 = 37.5108 \\
 f_{33} &= 1*3.4000 + 5*1.4875 + 2*1.2750 + 3*1.6508 \\
 &\quad + 4*4.0775 = 34.6499 \\
 f_{34} &= 1*3.4000 + 1*1.4875 + 3*1.2750 + 0*1.6508 \\
 &\quad + 1*4.0775 = 12.7900 \\
 f_{35} &= 1*3.4000 + 4*1.4875 + 4*1.2750 + 1*1.6508 \\
 &\quad + 2*4.0775 = 24.2558
 \end{aligned}$$

Also, the stochastic “molecular orbitals” (linear indices for each bond i) can be computed for this molecule in a similar form:

$$\begin{aligned}
 {}^s f_{0i}(\mathbf{w}) &= \mathbf{u}_i^t \cdot \mathbf{E} \mathbf{S}^0 \cdot \mathbf{w} = \sum_{j=1}^5 e_{ij}^0 w_j \\
 &= e_{i1}^0 w_1 + e_{i2}^0 w_2 + e_{i3}^0 w_3 + e_{i4}^0 w_4 + e_{i5}^0 w_5 \\
 {}^s f_{01} &= 1*3.4000 + 0*1.4875 + 0*1.2750 + 0*1.6508 \\
 &\quad + 0*4.0775 = 3.4000 \\
 {}^s f_{02} &= 0*3.4000 + 1*1.4875 + 0*1.2750 + 0*1.6508 \\
 &\quad + 0*4.0775 = 1.4875 \\
 {}^s f_{03} &= 0*3.4000 + 0*1.4875 + 1*1.2750 + 0*1.6508 \\
 &\quad + 0*4.0775 = 1.2750 \\
 {}^s f_{04} &= 0*3.4000 + 0*1.4875 + 0*1.2750 + 1*1.6508 \\
 &\quad + 0*4.0775 = 1.6508 \\
 {}^s f_{05} &= 0*3.4000 + 0*1.4875 + 0*1.2750 + 0*1.6508 \\
 &\quad + 1*4.0775 = 4.0775
 \end{aligned}$$

$$\begin{aligned}
 {}^s f_{0i}(\mathbf{w}) &= \mathbf{u}_i^t \cdot \mathbf{E} \mathbf{S}^1 \cdot \mathbf{w} = \sum_{j=1}^5 e_{ij}^1 w_j \\
 &= e_{i1}^1 w_1 + e_{i2}^1 w_2 + e_{i3}^1 w_3 + e_{i4}^1 w_4 + e_{i5}^1 w_5 \\
 {}^s f_{11} &= 0*3.4000 + 1*1.4875 + 0*1.2750 + 0*1.6508 \\
 &\quad + 0*4.0775 = 1.4875 \\
 {}^s f_{12} &= 0.33*3.4000 + 0*1.4875 + 0.33*1.2750 + 0*1.6508 \\
 &\quad + 0.33*4.0775 = 2.8883 \\
 {}^s f_{13} &= 0*3.4000 + 0.33*1.4875 + 0*1.2750 + 0.33*1.6508 \\
 &\quad + 0.33*4.0775 = 2.3812 \\
 {}^s f_{14} &= 0*3.4000 + 0*1.4875 + 1*1.2750 + 0*1.6508 \\
 &\quad + 0*4.0775 = 1.2750 \\
 {}^s f_{15} &= 0*3.4000 + 0.5*1.4875 + 0.5*1.2750 + 0*1.6508 \\
 &\quad + 0*4.0775 = 1.3813
 \end{aligned}$$

$$\begin{aligned}
 {}^s f_{0i}(\mathbf{w}) &= \mathbf{u}_i^t \cdot \mathbf{E} \mathbf{S}^2 \cdot \mathbf{w} = \sum_{j=1}^5 e_{ij}^2 w_j \\
 &= e_{i1}^2 w_1 + e_{i2}^2 w_2 + e_{i3}^2 w_3 + e_{i4}^2 w_4 + e_{i5}^2 w_5 \\
 {}^s f_{21} &= 0.33*3.4000 + 0*1.4875 + 0.33*1.2750 + 0*1.6508 \\
 &\quad + 0.33*4.0775 = 2.8883 \\
 {}^s f_{22} &= 0*3.4000 + 0.5*1.4875 + 0.16*1.2750 + 0.16*1.6508 \\
 &\quad + 0.16*4.0775 = 1.8643 \\
 {}^s f_{23} &= 0.16*3.4000 + 0.16*1.4875 + 0.5*1.2750 + 0*1.6508 \\
 &\quad + 0.16*4.0775 = 2.0719 \\
 {}^s f_{24} &= 0*3.4000 + 0.33*1.4875 + 0*1.2750 + 0.33*1.6508 \\
 &\quad + 0.33*4.0775 = 2.3812 \\
 {}^s f_{25} &= 0.16*3.4000 + 0.16*1.4875 + 0.16*1.2750 \\
 &\quad + 0.16*1.6508 + 0.33*4.0775 = 2.5957
 \end{aligned}$$

$$\begin{aligned}
 {}^s f_{0i}(\mathbf{w}) &= \mathbf{u}_i^t \cdot \mathbf{E} \mathbf{S}^3 \cdot \mathbf{w} = \sum_{j=1}^5 e_{ij}^3 w_j \\
 &= e_{i1}^3 w_1 + e_{i2}^3 w_2 + e_{i3}^3 w_3 + e_{i4}^3 w_4 + e_{i5}^3 w_5 \\
 {}^s f_{31} &= 0*3.4000 + 0.5*1.4875 + 0.16*1.2750 + 0.16*1.6508 \\
 &\quad + 0.16*4.0775 = 1.8643 \\
 {}^s f_{32} &= 0.2*3.4000 + 0.13*1.4875 + 0.33*1.2750 \\
 &\quad + 0.06*1.6508 + 0.26*4.0775 = 2.4533 \\
 {}^s f_{33} &= 0.06*3.4000 + 0.33*1.4875 + 0.13*1.2750 \\
 &\quad + 0.2*1.6508 + 0.26*4.0775 = 2.2509 \\
 {}^s f_{34} &= 0.16*3.4000 + 0.16*1.4875 + 0.5*1.2750 + 0*1.6508 \\
 &\quad + 0.16*4.0775 = 2.0719 \\
 {}^s f_{35} &= 0.083*3.4000 + 0.33*1.4875 + 0.33*1.2750 \\
 &\quad + 0.083*1.6508 + 0.16*4.0775 = 1.9832
 \end{aligned}$$

The total non-stochastic linear indices [1-norm of $\mathbf{f}_k(\mathbf{w})$] can be expressed as the sum of the local (bond) linear indices for this molecule as it follows:

$$\begin{aligned}
 f_0(\mathbf{w}) &= \sum_{i=1}^5 \mathbf{u}_i^t \cdot \mathbf{E}^0 \cdot \mathbf{w} = \sum_{i=1}^5 f_{0i}(\mathbf{w}) = f_{01}(\mathbf{w}) + f_{02}(\mathbf{w}) \\
 &\quad + f_{03}(\mathbf{w}) + f_{04}(\mathbf{w}) + f_{05}(\mathbf{w}) \\
 &= 3.4000 + 1.4875 + 1.2750 + 1.6508 \\
 &\quad + 4.0775 = 11.8908 \\
 f_1(\mathbf{w}) &= \sum_{i=1}^5 \mathbf{u}_i^t \cdot \mathbf{E}^1 \cdot \mathbf{w} = \sum_{i=1}^5 f_{1i}(\mathbf{w}) = f_{11}(\mathbf{w}) + f_{12}(\mathbf{w}) \\
 &\quad + f_{13}(\mathbf{w}) + f_{14}(\mathbf{w}) + f_{15}(\mathbf{w}) \\
 &= 1.4875 + 8.7525 + 7.2158 + 1.2750 \\
 &\quad + 2.7625 = 21.49333 \\
 f_2(\mathbf{w}) &= \sum_{i=1}^5 \mathbf{u}_i^t \cdot \mathbf{E}^2 \cdot \mathbf{w} = \sum_{i=1}^5 f_{2i}(\mathbf{w}) = f_{21}(\mathbf{w}) + f_{22}(\mathbf{w}) \\
 &\quad + f_{23}(\mathbf{w}) + f_{24}(\mathbf{w}) + f_{25}(\mathbf{w}) \\
 &= 8.7525 + 11.4658 + 12.7900 + 7.2158 \\
 &\quad + 15.9683 = 56.1925 \\
 f_3(\mathbf{w}) &= \sum_{i=1}^5 \mathbf{u}_i^t \cdot \mathbf{E}^3 \cdot \mathbf{w} = \sum_{i=1}^5 f_{3i}(\mathbf{w}) = f_{31}(\mathbf{w}) + f_{32}(\mathbf{w}) \\
 &\quad + f_{33}(\mathbf{w}) + f_{34}(\mathbf{w}) + f_{35}(\mathbf{w}) \\
 &= 11.4658 + 37.5108 + 34.6499 + 12.7900 \\
 &\quad + 24.2558 = 120.6723
 \end{aligned}$$

The terms in the sums for calculating the total linear indices are the so-called local (bond) linear indices and, in this case, they coincide with the fragment (bond-type) linear indices. We have written these terms in the consecutive order of the bond labels in the graph. For instance, the non-stochastic local (bond) linear indices of orders 0, 1, 2 and 3 for the bond labeled as 1 are 3.4000, 1.4875, 8.7525, and 11.4658, correspondingly.

The values of the k th total stochastic linear indices [1 - norm of ${}^s\mathbf{f}_k(\mathbf{w})$] are also the sum of the k th local (bond) stochastic linear indices values for all bonds in the molecule:

$$\begin{aligned}
 {}^s f_0(\mathbf{w}) &= \sum_{i=1}^5 \mathbf{u}_i^t \cdot \mathbf{E}^0 \cdot \mathbf{w} = \sum_{i=1}^5 {}^s f_{0i}(\mathbf{w}) = {}^s f_{01}(\mathbf{w}) \\
 &\quad + {}^s f_{02}(\mathbf{w}) + {}^s f_{03}(\mathbf{w}) + {}^s f_{04}(\mathbf{w}) + {}^s f_{05}(\mathbf{w}) \\
 &= 3.4000 + 1.4875 + 1.2750 + 1.6508 \\
 &\quad + 4.0775 = 11.8908 \\
 {}^s f_1(\mathbf{w}) &= \sum_{i=1}^5 \mathbf{u}_i^t \cdot \mathbf{E}^1 \cdot \mathbf{w} = \sum_{i=1}^5 {}^s f_{1i}(\mathbf{w}) = {}^s f_{11}(\mathbf{w}) \\
 &\quad + {}^s f_{12}(\mathbf{w}) + {}^s f_{13}(\mathbf{w}) + {}^s f_{14}(\mathbf{w}) + {}^s f_{15}(\mathbf{w}) \\
 &= 1.4875 + 2.8883 + 2.3812 + 1.2750 \\
 &\quad + 1.3813 = 9.4133
 \end{aligned}$$

$$\begin{aligned}
 {}^s f_2(\mathbf{w}) &= \sum_{i=1}^5 \mathbf{u}_i^t \cdot \mathbf{E}^2 \cdot \mathbf{w} = \sum_{i=1}^5 {}^s f_{2i}(\mathbf{w}) = {}^s f_{21}(\mathbf{w}) \\
 &\quad + {}^s f_{22}(\mathbf{w}) + {}^s f_{23}(\mathbf{w}) + {}^s f_{24}(\mathbf{w}) + {}^s f_{25}(\mathbf{w}) \\
 &= 2.8883 + 1.8643 + 2.0719 + 2.3812 \\
 &\quad + 2.5957 = 11.8014
 \end{aligned}$$

$$\begin{aligned}
 {}^s f_3(\mathbf{w}) &= \sum_{i=1}^5 \mathbf{u}_i^t \cdot \mathbf{E}^3 \cdot \mathbf{w} = \sum_{i=1}^5 {}^s f_{3i}(\mathbf{w}) = {}^s f_{31}(\mathbf{w}) \\
 &\quad + {}^s f_{32}(\mathbf{w}) + {}^s f_{33}(\mathbf{w}) + {}^s f_{34}(\mathbf{w}) + {}^s f_{35}(\mathbf{w}) \\
 &= 1.8643 + 2.4533 + 2.2509 + 2.0719 \\
 &\quad + 1.9832 = 10.6236
 \end{aligned}$$

Methods

TOMOCOMD-CARDD approach

TOMOCOMD is an interactive program for molecular design and bioinformatic research [20]. It consists of four subprograms; each one allows drawing the structures (drawing mode) and calculating molecular 2D/3D (calculation mode) descriptors. The modules are named **CARDD** (Computed-Aided ‘Rational’ Drug Design), **CAMPS** (Computed-Aided Modeling in Protein Science), **CANAR** (Computed-Aided Nucleic Acid Research) and **CABPD** (Computed-Aided Bio-Polymers Docking). In the present report, we outline salient features concerned with only one of these subprograms, **CARDD**, and with the calculation of non-stochastic and stochastic 2D bond-based linear indices.

Computational strategies

The main steps for the application of the present method to QSAR/QSPR and drug design can be summarized briefly in the following algorithm: (1) draw the molecular pseudo-graphs for each molecule of the data set by using the software drawing mode. This procedure is performed by a selection of the active atomic symbol belonging to the different groups in the periodic table of the elements, (2) use appropriated atomic properties in order to weight and differentiate the molecular bonds. In this study, the used properties are those previously proposed for the calculation of the DRAGON descriptors [57, 62, 63], i.e., atomic mass (M), atomic polarizability (P), atomic Mulliken electronegativity (K), van der Waals atomic volume (V), plus the atomic electronegativity in Pauling scale (G) [64]. The values of these atomic labels are shown in Table 1. In order to calculate the

Table 1 Values of the atomic weights used for bond-based linear indices calculation

ID	Atomic mass	VdW (\AA^3)	volume	Mulliken electronegativity	Polarizability (\AA^3)	Pauling electronegativity
H	1.01	6.709		2.592	0.667	2.2
B	10.81	17.875		2.275	3.030	2.04
C	12.01	22.449		2.746	1.760	2.55
N	14.01	15.599		3.194	1.100	3.04
O	16.00	11.494		3.654	0.802	3.44
F	19.00	9.203		4.000	0.557	3.98
Al	26.98	36.511		1.714	6.800	1.61
Si	28.09	31.976		2.138	5.380	1.9
P	30.97	26.522		2.515	3.630	2.19
S	32.07	24.429		2.957	2.900	2.58
Cl	35.45	23.228		3.475	2.180	3.16
Fe	55.85	41.052		2.000	8.400	1.83
Co	58.93	35.041		2.000	7.500	1.88
Ni	58.69	17.157		2.000	6.800	1.91
Cu	63.55	11.494		2.033	6.100	1.9
Zn	65.39	38.351		2.223	7.100	1.65
Br	79.90	31.059		3.219	3.050	2.96
Sn	118.71	45.830		2.298	7.700	1.96
I	126.90	38.792		2.778	5.350	2.66

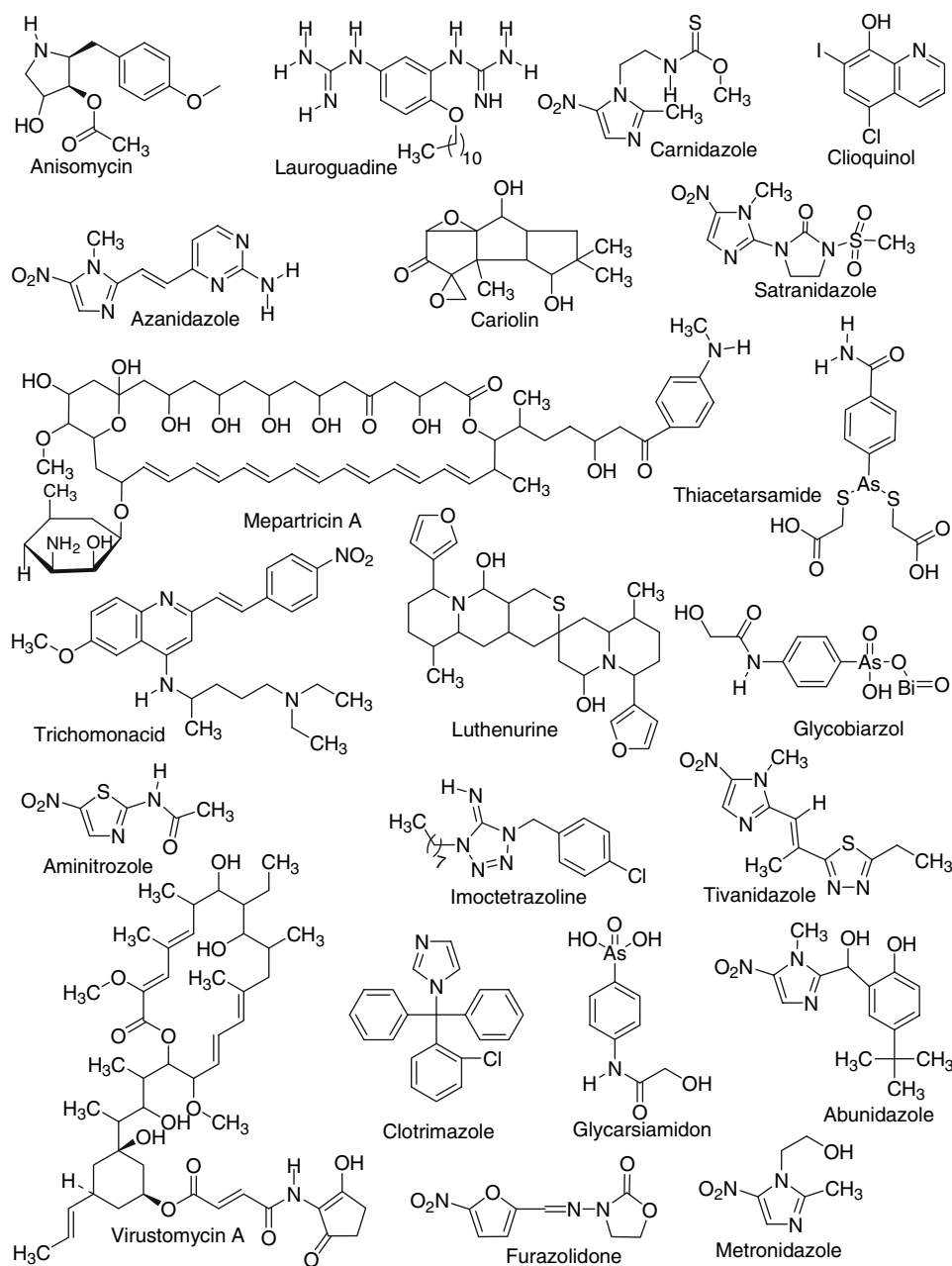
required weights, we used the mathematical expression given by Eq. 2, which involves atomic weights, (3) compute the total and local (bond and bond-type) non-stochastic and stochastic linear indices. The estimate can be carried out in the software calculation mode, where the user can previously select the atomic properties and the descriptor family to calculate the molecular indices. This software generates a table in which the rows correspond to the compounds, as well as the columns correspond to the total and local bond-based linear indices or other molecular descriptors family implemented in this program, (4) Find an QSPR/QSAR equation by using several multivariate analytical techniques, such as multilinear regression analysis (MRA), neural networks (NN), linear discrimination analysis (LDA), and so on. Therefore, the user can find a quantitative relation between an activity A and the linear indices having, for instance, the following appearance, $A = a_0f_0(\mathbf{w}) + a_1f_1(\mathbf{w}) + a_2f_2(\mathbf{w}) + \dots + a_kf_k(\mathbf{w}) + c$, where A is the measured activity, $f_k(\mathbf{w})$ are the k th total bond-based linear indices, and the a_k 's are the coefficients obtained by the linear regression analysis, (5) test the robustness and predictive power of the QSPR/QSAR equation by using internal (cross-validation) and external (with a test set and an external predicting set) validation techniques, and (6) apply the obtained LDA-based QSAR models as a cheminformatic tool for identifying and/or discovering novel drugs through the ligand-based *virtual* screening procedure.

The bond-based **TOMOCOMD-CARDD** descriptors computed in this study were the following:

- (1) $kth(k = \overline{0, 15})$ total non-stochastic bond-based linear indices, not considering and considering H-atoms in the molecular graph (G) [$f_k(\mathbf{w})$ and $f_k^H(\mathbf{w})$, respectively].
- (2) $kth(k = \overline{0, 15})$ total stochastic bond-based linear indices, not considering and considering H-atoms in the molecular graph (G) [$f_k^s(\mathbf{w})$ and $f_k^{sH}(\mathbf{w})$, respectively].
- (3) $kth(k = \overline{0, 15})$ bond-type (group = heteroatoms: S, N, O) non-stochastic linear indices, not considering and considering H-atoms in the molecular graph (G) [$f_{kt}^E(\mathbf{w})$ and $f_{kt}^{EH}(\mathbf{w})$, correspondingly]. These fragment descriptors are putative molecular charge, dipole moment, and H-bonding acceptors.
- (4) $kth(k = \overline{0, 15})$ bond-type (group = heteroatoms: S, N, O) stochastic linear indices not considering and considering H-atoms in the molecular graph (G) [$f_{kt}^{Es}(\mathbf{w})$, and $f_{kt}^{EsH}(\mathbf{w})$, correspondingly]. These fragment descriptors are also putative molecular charge, dipole moment, and H-bonding acceptors.

The total number of computed linear molecular descriptors was 640 for each molecule. However, redundant and highly-correlated descriptors were eliminated before model development. Finally, forward stepwise (for each family and type of descriptors) and best subset was fixed as the strategy for variable selection.

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



Data set for QSAR study

In order to obtain linear discriminant functions capable of discriminating between active and inactive compounds, the chemical information contained in a greater 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, through a random selection, for the set of inactive compounds, guaranteeing also a great structural variability. All these chemicals were taken from the Negwer Handbook [65] and Merck Index [66], where their names, synonyms and structural formulas can be found.

From these 123 chemicals, 91 were chosen at random to form the training set, being 40 of them active and 51 inactive ones. The great structural variability of the selected training data makes possible the discovery of lead compounds, not only with determined mechanisms of

antitrichomonal activity, but also with novel modes of action. This will be illustrated well in this report in a *virtual* experiment for lead compounds generation.

The remaining subseries, consisting of 10 trichomonacids and 22 non-trichomonacids, were prepared as test sets for the external validation of the models (32 chemicals). These compounds were never used in the development of the classification models.

Statistical analysis

The discriminant functions were obtained by using the Linear Discriminant Analysis (LDA) [67], as implemented in STATISTICA [68]. The default parameters of this program were used in the development of the model. Forward stepwise and best subset were fixed as the strategy for variable selection. The principle of maximal parsimony (Occam's razor) was taken into account as the strategy for model selection [69]. In this connection, we selected the model with higher statistical significance but having as few parameters (a_k) as possible.

The quality of the models were determined by examining Wilks' λ parameter (U -statistic), square Mahalanobis' distance (D^2), Fisher's ratio (F) and the corresponding p -level [$p(F)$] as well as the percentage of good classification in the training and test sets [67]. Models with a proportion lower than 5 between the number of cases and variables in the equation were rejected [70].

The Wilks' λ for the overall discrimination can take values in the range from 0 (perfect discrimination) to 1 (no discrimination). The D^2 statistic indicates the separation of the respective groups, showing whether the model possesses an appropriate discriminatory power for differentiating between the two respective groups.

By using the models, one compound can then be classified as either active, if $\Delta P\% > 0$, being $\Delta P\% = [P(\text{active}) - P(\text{inactive})] \times 100$, or inactive otherwise. $P(\text{active})$ and $P(\text{inactive})$ are the probabilities with which the equations classify a compound as active and inactive, respectively.

The statistical robustness and predictive power of the obtained model were assessed using a prediction (test) set [71]. In addition, a leave-group-out (LGO) cross-validation strategy was carried out. In this case, 10% of the data was used as group size, i.e., groups including 10% of the training data were left out and predicted by the model based on the remaining 90%. This process was carried out 10 times on 10 unique subsets. In this way, every observation was predicted once (in its group of left-out observations). The overall mean for this process (10% *full* leave-out cross-validation) was used as a good indication of robustness, stability and predictive power of the obtained models [71].

Finally, the calculation of percentages of global good classification (accuracy), sensibility, specificity (also known as 'hit rate'), false positive rate (also known as 'false alarm rate') and Matthews' correlation coefficient (C) in the training and test (predicting) sets permitted the assessment of the model [72].

Biological assay: determination of in vitro trichomonadal activity

The biological activity was assayed on Tv JH31A #4 Ref. No. 30326 (ATCC, MD, USA) in modified Diamond medium supplemented with equine serum and grown at 37°C (5% CO₂). 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. MTZ (Sigma-Aldrich SA, Spain) was used as reference drug at concentrations of 2, 1, 0.5 µg/ml. Cytocidal and cytostatic activities were determined by calculation of percentages of cytocidal (%C) and cytostatic (%CA) activities in relation to controls, as previously reported [73, 74].

Results and discussion

Development and validation of the discriminant functions

Although the number of existing statistical methods to get classification functions is relatively extensive, we select linear discriminant analysis (LDA) given the simplicity of the method [67]. The use of LDA in drug design has been extensively reported by different authors [19, 23–25, 27–35, 75, 76]. 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 [68], the following linear models were obtained, in which the total as well as local non-stochastic and stochastic bond-based linear indices were used as independent variables:

$$\begin{aligned} \text{Class} = & -5.53 - 2.96 \times 10^{-5} E_{f_{6t}}^H(\mathbf{w}_M) - 0.07 E_{f_{0t}}^H(\mathbf{w}_M) \\ & - 0.05 E_{f_{0t}}(\mathbf{w}_M) \\ & - 5.29 \times 10^{-4} E_{f_{7t}}(\mathbf{w}_P) + 4.73 \times 10^{-5} E_{f_{7t}}(\mathbf{w}_V) \\ & + 0.36 E_{f_{0t}}^H(\mathbf{w}_K) \end{aligned} \quad (11)$$

$$N = 91 \quad \lambda = 0.46 \quad D^2 = 4.54 \quad F(6,84) = 15.99 \\ p < 0.01$$

Table 2 Names and classification of active compounds into training and test sets according to both TOMOCOMD-CARDD models developed in this work

Name	$\Delta P\%^a$	$\Delta P\%^b$	Name	$\Delta P\%^a$	$\Delta P\%^b$
<i>Active training set</i>					
Anisomycin	−90.67	−68.54	Abunidazole	73.86	83.90
Virustomycin A	98.26	99.72	Imoctetrazoline	63.30	93.47
Azanidazole	96.39	94.97	Forminitrazole	80.64	45.86
Carnidazole	93.99	90.28	Chlomizol	85.64	84.98
Propenidazole	98.57	94.48	Acinitrazole	80.64	71.72
Lauroguadine	−43.58	58.70	Moxnidazole	99.81	97.26
Mepartricin A	82.28	91.89	Isometronidazole	45.23	46.05
Metronidazole	50.39	42.97	Mertronidazole phosphate	73.62	67.93
Nifuratel	98.23	96.99	Benzoylmetronidazole	93.98	92.71
Nifuroxime	66.68	60.69	Bamnidazole	94.34	82.87
Nimorazole	51.98	−20.24	Glycarsiamidon	−51.91	−50.87
Secnidazole	46.88	46.04	Fexinidazole	95.09	82.59
Cariolin	50.23	−20.26	Piperanitrozole	79.08	87.64
2-Amino-5-nitrotiazole	30.56	−42.27	Gynotabs	68.15	81.93
Glycobiartzol	58.90	35.45	Pirinidazole	93.22	93.15
Clioquinol	38.35	−21.20	Metronidazole hydrogen succinate	98.18	90.43
Diiodohydroxy quinoline	60.86	−73.55	Tolamizol	81.98	90.04
Ornidazol	88.14	85.11	Thiacetarsamide	32.50	2.59
Trichomonacid	77.09	68.11	Tivanidazole	98.42	99.58
Lutenurine	−86.14	71.53	Policresulen	−50.52	31.32
<i>Active test set</i>					
Acertarsone	−28.77	−22.46	Pentamycin	−97.52	−66.85
Furazolidone	98.27	96.87	Azomycin	12.44	11.91
Mepartricin B	76.62	92.54	Ternidazole	50.88	53.30
Aminitrozole	80.64	71.72	Misonidazole	61.56	29.13
Clotrimazol	14.60	39.75	Satranidazole	97.86	97.50

^{a,b} Antitrichomonal activity predicted by Eqs. 11 and 12, respectively: $\Delta P\% = [P(\text{active}) - P(\text{inactive})] \times 100$

Bold values represent the misclassification's cases by obtained models

$$\text{Class} = -4.93 - 0.12 f_{9t}^{\text{Es}}(\mathbf{w}_M) + 0.10 f_{0t}^{\text{Es}}(\mathbf{w}_V) + 1.20 f_{2t}^{\text{Es}}(\mathbf{w}_K) - 0.77 f_{5t}^{\text{Es}}(\mathbf{w}_K) \quad (12)$$

$N = 91$ $\lambda = 0.48$ $D^2 = 4.28$ $F(6.84) = 123,18$ $p < 0.01$ where, N is the number of compounds, λ is Wilks' statistics, D^2 is the square of the Mahalanobis' distance, F is the Fisher's ratio and p is the significance level.

Model 11 correctly classifies 87.50% of active and 90.20% of inactive compounds in the training set for a global good classification (accuracy) of 89.01%. Model 12 correctly classifies 82.42% of the compounds in training set. Specifically, the model correctly classifies 33 out of 40 (82.50%) trichomonacidal compounds and 42 out of 51 (82.35%) inactive chemicals in the training series. On the other hand, Eqs. 11 and 12 show an 87.50% (28/32) and 84.38% (27/32), respectively, of global predictability in the prediction series. These results validate the models for use in the ligand-based *virtual* screening, taking into consideration that 85.0% is considered as an acceptable threshold limit for this kind of analysis [77].

In Tables 2 and 3 we give the names of all the compounds in the training and test, active and inactive sets together with their posterior probabilities calculated from the Mahalanobis distance using both equations. The same information of all compounds in the training and test sets appears in Table 4, which summarizes the results of the classifications for both models in the training and test groups.

A more serious analysis was carried out by calculating most of the parameters commonly used in medical statistics (accuracy, sensitivity, specificity and false positive rate) and the Matthews correlation coefficient (C). Table 4 also lists these parameters for both obtained models [72, 78]. While the sensitivity is the probability of correctly predicting a positive example, the specificity is the probability that a positive prediction be correct. On the other hand, C quantifies the strength of the linear relation between the molecular descriptors and the classifications, and it may often provide a much more balanced evaluation of the prediction than, for instance, the percentages [72, 78]. The obtained models, Eqs. 11 and 12, showed considerably

Table 3 Names and classification of inactive compounds into training and test sets according to both TOMOCOMD-CARDD models developed in this work

Name	$\Delta P\%^a$	$\Delta P\%^b$	Name	$\Delta P\%^a$	$\Delta P\%^b$
<i>Inactive training set</i>					
Amantadine	−99.63	−96.58	Nonaferone	−89.35	−91.95
Thiacetazone	−42.85	−50.91	Rolipram	−69.90	−74.67
Cloral betaine	−96.14	−98.86	N-hydroxymethyl-N-methylurea	−95.44	−96.97
Carbavin	−80.39	−82.77	4 chlorobenzoic acid	−71.36	−6.54
Norantoin	−70.13	−36.88	Acetanilide	−95.13	−91.81
Orotonsan Fe	−3.32	51.99	Guanazole	−99.82	−98.67
Picosulfate	78.57	0.89	Tetramin	−99.37	−99.09
Naftazone	−59.18	−35.79	Mecysteine	−97.80	−98.27
Besunide	−65.47	22.10	Cirazoline	−90.59	−91.89
Acetazolamide	−48.00	−45.29	Methocarbamol	−10.45	−83.39
Propamine”soviet	−99.72	−99.19	Lysergide	−89.11	−73.32
RMI 11894	−98.27	−94.42	Dopamine	−98.24	−81.00
Ag 307	−52.03	−78.33	Bufeniodo	−19.12	−18.70
Barbismethylii iodide	−10.56	−98.62	Celiprolol	−21.41	−41.31
Pancuronium bromide	−83.53	−97.69	Erysimin	12.24	39.02
Vinyl ether	−94.55	−98.59	Peruvoside	8.24	−7.89
Basedol	−48.72	11.14	Amitraz	−67.45	−74.39
Carbimazole	46.85	55.40	Proclonol	−42.09	59.74
Didym levulinate	−91.58	−95.22	Asame	−90.29	−95.41
Perchloroethane	−96.30	−82.82	KC-8973	−82.47	−81.26
Pyrantel tartrate	−80.54	−82.55	Ethydine	64.58	36.30
Fentanyl	−93.52	−94.42	Magnesii metioglicas	−46.45	−97.37
Petidina	−87.83	−91.89	Alibendol	−71.94	−37.14
Tenalidine tartrate	−99.78	−98.92	Diponium Bromide	−96.50	−97.93
Bamipine	−98.72	−98.82	Streptomycin	−90.64	69.92
Colestipol	−99.71	−99.76			
<i>Inactive test set</i>					
Citenazone	−19.35	−28.82	Metriponate	−99.97	−95.87
Methenamine	−99.66	−99.43	Ciclopramine	−97.18	−87.45
Pentrichloral	−90.52	25.87	Litracen	−99.31	−98.95
Calcium sodium ferriclate	−100.00	−100.00	Trimetilsulfonium hydroxyde	−99.48	−99.97
Ferrocron	−99.06	−95.58	Norgamem	−98.23	−94.55
Emodin	−89.49	−94.53	Emylcamate	5.07	21.19
Butanolum	−98.18	−99.67	Acetylcholine	−96.79	−99.97
Spironolactone	−99.70	−99.91	Carazolol	−94.86	−86.37
Bromcholine	−57.67	−99.43	Cefazolin	−14.78	−82.60
Imekhin	68.54	−31.90	Penicillin I	−91.40	42.04
Diphenadione	−85.12	−85.68	Aziromycin	−90.04	−82.75

^{a,b} Antitrichomonal activity predicted by Eqs. 11 and 12, respectively: $\Delta P\% = [P(\text{active}) - P(\text{inactive})] \times 100$
 Bold values represent the misclassification's cases by obtained models

high *C* of 0.78 (0.71) and 0.65 (0.65) in training (test) sets, correspondingly.

Internal validation of the discriminant functions. Cross-validation methods

In the recent years, exhaustive validation of mathematical models constitutes a main key of current QSAR theory

[71]. To this effect, 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 a 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 [79]. A more exhaustive cross-validation method can be used, in which a fraction of

Table 4 Prediction performances for two LDA-based QSAR models (using non-stochastic and stochastic bond-type linear indices) in the training and test sets

	Matthews' Corr. Coeff. (C)	Accuracy 'Q _{Total} ' (%)	Sensitivity 'hit rate' (%)	Specificity (%)	False positive rate 'false alarm rate' (%)
<i>Non-stochastic bond-type linear indices (Eq. 11)</i>					
Training set	0.78	89.01	87.50	87.50	9.80
Predicting set	0.71	87.50	80.00	80.00	9.09
<i>Stochastic bond-type linear indices (Eq. 12)</i>					
Training set	0.65	82.42	82.50	78.57	17.65
Predicting set	0.65	84.38	80.00	72.73	13.64

Table 5 Results of the 10-fold full cross-validation procedure

Groups	Q% ^a	λ	D ²	F	Q% ^b	Q% ^a	λ	D ²	F	Q% ^b
	Eq. 11 (non-stochastic bond-based linear indices)					Eq. 12 (stochastic bond-based linear indices)				
1	88.89	0.45	4.92	15.35	80.00	85.19	0.46	4.73	22.74	70.00
2	89.02	0.48	4.38	13.81	77.78	82.93	0.49	4.19	20.34	77.78
3	87.80	0.49	4.16	13.14	100.00	82.93	0.49	4.18	20.30	77.78
4	87.80	0.48	4.35	13.73	88.89	80.49	0.49	4.10	19.94	88.89
5	87.80	0.49	4.16	13.14	100.00	81.71	0.50	4.03	19.56	77.78
6	89.02	0.46	4.68	14.77	88.89	81.71	0.47	4.39	21.34	77.78
7	90.24	0.43	5.20	16.39	88.89	81.71	0.49	4.16	20.20	77.78
8	90.24	0.44	5.13	16.19	77.78	80.49	0.48	4.25	20.66	88.89
9	89.02	0.46	4.59	14.47	88.89	82.93	0.48	4.22	20.52	77.78
10	89.16	0.46	4.65	14.93	87.50	81.93	0.46	4.63	22.86	87.50
Mean	88.90	0.46	4.62	14.59	87.86	82.20	0.48	4.29	20.85	80.19
SD	0.90	0.02	0.37	1.16	7.92	1.38	0.01	0.23	1.13	6.18

^{a,b} Global good classification from both models in training (90% of the data) and test (10% of the data) sets, respectively

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 [79, 80].

In this work, we carried out a leave-10-fold full-out (LGO) cross-validation procedure. For each group of observations left out (10% of the whole data set, 9 compounds), a model was developed from the remaining 90% of the data (81 compounds). This process was carried out ten times on ten unique independent subsets. The statistical results are depicted in Table 5. The overall mean of the correct classification in the training (test) set for this process for Eqs. 11 and 12 was 88.90% (87.86%) and 82.20% (80.19%), correspondingly. The result for predictions on the 10% full cross-validation test evidenced the quality (robustness, stability and predictive power) of the obtained models. This validation process is important, if we take into consideration that the predictive ability of an QSAR model can be estimated by using only an external test set of compounds that were not utilized for building the model [71, 79, 80].

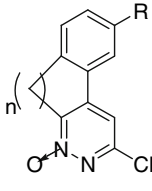
Identification of reported chemicals through a simulated ligand-based virtual screening experiment

One of the main features that any theoretical approach to drug discovery needs is the identification of active compounds from databases of chemicals. Given that the absence of a receptor 3D structure is the main reason for the application of ligand-based methods [19, 81, 82], ligand-based virtual screening becomes our work philosophy.

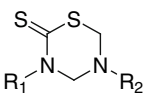
In order to prove the possibilities of the **TOMOCMD-CARDD** approach for the ligand-based virtual screening of anti-trichomonal compounds, we have selected a series of 12 compounds whose activities against *Tv* have been already proved by several researchers [83–85]. They all were evaluated with models 11 and 12 as active/inactive ones. Their structures as well as the results of the classification are shown in Table 6.

As can be seen, both models correctly classify most of the 12 selected compounds. The first model (Eq. 11) classifies incorrectly only two compounds (both as false

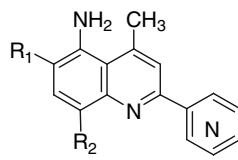
Table 6 Identification of chemicals extracted from literature as active or inactive toward the anti-trichomonal activity, by using LDA-based QSAR models in a simulated ligand-based *virtual*-screening experiment



1: $n = \text{CH}_2$
 $R = \text{NH}_2$
2: $n = \text{CH}_2\text{-CH}_2$
 $R = \text{NH}_2$
3: $n = \text{CH}_2=\text{CH}_2$
 $R = \text{H}$



	R ₁	R ₂
4	furfuril	CH(CH ₂ COOH)-COOH
5	furfuril	CH(CH ₃)-COOH
6	furfuril	CH[(CH ₂) ₂ SCH ₃] -COOH
7	furfuril	CH[(CH ₃) ₂] -COOH
8	ciclohexil	CH ₂ -COOH
9	ciclohexil	CH ₂ -CONH-CH ₂ -COOH



	R ₁	R ₂	N Position
10	CH ₃	H	β
11	CH ₃	H	γ
12	H	CH(CH ₃) ₂	β

Comp. ^a	Ref. ^b	ΔP% ^c	ΔP% ^d	Antitrichomonal activity
1	Gavini et al. (2000)	−22.12	72.87	Inactive
2		−23.35	76.98	Inactive
3		−21.91	41.24	Inactive
4	Ochoa et al. (1999)	79.96	98.60	100 μg/ml = 100 ^d 10 μg/ml = (100) ^e 1 μg/ml = (100) ^f
5		47.53	94.28	100 μg/ml = 100 ^e 10 μg/ml = (100) ^f 1 μg/ml = (77) ^f
6		56.49	96.15	100 μg/ml = 100 ^e 10 μg/ml = (100) ^f 1 μg/ml = (73) ^f
7		84.64	96.83	100 μg/ml = 100 ^e 10 μg/ml = (13) ^f 1 μg/ml = (66) ^f
8		− 90.77	3.29	100 μg/ml = 100 ^e 10 μg/ml = (67) ^f 1 μg/ml = (93) ^f
9		− 79.79	78.68	100 μg/ml = 100 ^e 10 μg/ml = (74) ^f 1 μg/ml = (94) ^f
10	Kouznetsov et al. (2004)	−78.68	−25.29	100 μg/ml = (58.3) ^f 10 μg/ml = (29.1) ^f 1 μg/ml = (18.1) ^f
11		−78.87	−35.17	100 μg/ml = (66.7) ^f 10 μg/ml = (33.9) ^f 1 μg/ml = (25.2) ^f
12		−77.55	−25.34	100 μg/ml = (65.4) ^f 10 μg/ml = (56.7) ^f 1 μg/ml = (40.1) ^f

^a The molecular structures of the compounds represented with bold figures are shown at the top of this table^b Bibliographic references from where molecules together with its in vitro activities were taken^{c,d} Antitrichomonal activity predicted by Eqs. 11 and 12; ΔP% = [P(active) − P(inactive)] × 100^e Percentage of reduction of *Tv* or cytotoxic activity at the indicated doses at 24 h^f Specific activity against *Tv* (in parenthesis) expressed as percentages of growth inhibition or cytostatic activity at 24 h

Bold values represent the misclassification's cases by obtained models

Fig. 2 Structures of quinoxaline derivatives for novel trichomonacids discovery by ligand-based virtual screening LDA-assisted models

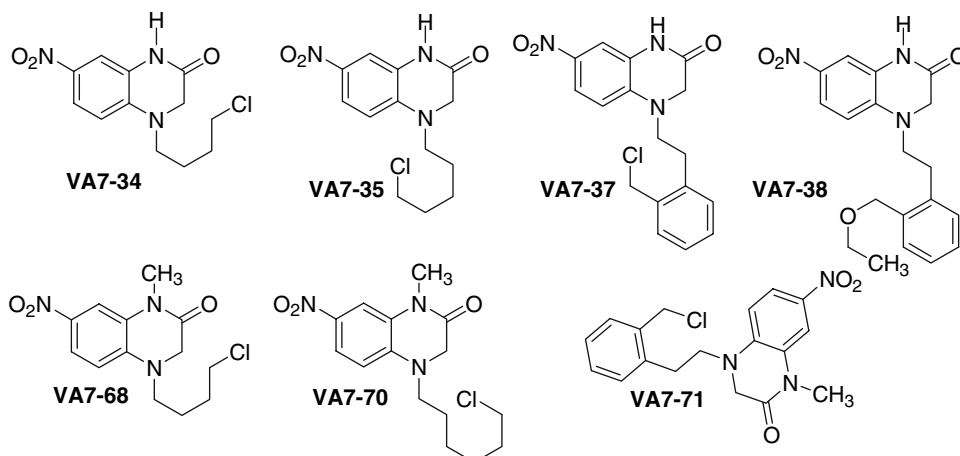


Table 7 Results of the computational evaluation using LDA-assisted QSAR models and percentages of cytostatic and/or cytotoxic activity [brackets] for the three concentrations assayed in vitro against *Tv*

Compound ^a	Theoretical results				Class ^f	In vitro activity (μg/ml) ^g					
	Class ^b	ΔP% ^c	Class ^d	ΔP% ^e		%CA _{24 h} [%C _{24 h}]			%CA _{48 h} [%C _{48 h}]		
						100	10	1	100	10	1
VA7-34	+	77.57	+	95.43	+	[100]	87.13	15.63	[100]	35.17	0
VA7-35	+	78.75	+	95.71	+	[98.66]	88.92	2.27	[99.40]	56.93	0
VA7-37	+	80.82	+	97.66	+	[100]	[98.38]	5.11	[100]	[94.23]	11.11
VA7-38	+	71.88	+	83.80	+	[100]	[97.59]	1.99	[100]	[98.10]	0
VA7-68	+	89.54	+	91.82	+	[100]	82.84	22.73	[100]	39.29	0
VA7-70	+	90.24	+	92.44	+	[99.83]	[94.38]	22.73	[100]	83.64	6.99
VA7-71	+	91.14	+	95.76	—	87.16	51.28	18.47	56.93	17.98	4.70
MTZ	+	50.39	+	42.97	+	[100]	[100]	87.89	[100]	[100]	71.25

^a The molecular structures of the compounds represented with codes are shown in Fig. 2

^{b,d} In silico classification obtained from models Eqs. 11 and 12 using non-stochastic and stochastic bond-type linear indices, respectively

^{c,e} Results for the classification of compounds obtained from models Eqs. 11 and 12, correspondingly: ΔP% = [P(active) – P(inactive)] × 100

^f Observed (experimental activity) classification against *Tv*

^g 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 or 48 h); [%C_#], cytotoxic activity (24 or 48 h); MTZ, metronidazole (concentrations for MTZ were 2, 1 and 0.5 μg/ml, respectively)

Bold values represent the misclassification's cases by obtained models

negative), thus achieving 83.33% of correct classification, while the second model (Eq. 12) classifies incorrectly three compounds (all of them as false positive) for yielding 75.00% of correct classification. This result is the most important criterion for the validation of the models developed here, since they have been able to detect series of compounds from literature as active/inactive and these chemicals have shown, in general, the predicted activity.

The next step in this approach would be the inclusion of these 'novel' compounds in the training set and the development of new discrimination models. In this sense, the derivation of the models 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.

In addition, no previous reports dealing with to the application of pattern recognition techniques to the selection of trichomonacidal compounds from a heterogeneous series of compounds were found in the literature. Therefore, the present algorithm constitutes a step forward in the search of efficient ways to discover new drugs bioactive against *Tv*.

Discovery of novel antitrichomonal compounds via ligand-based virtual screening LDA-assisted models as a rational search procedure. Experimental 'in vitro' corroboration

The massive cost involved in the development of new drugs, together with the low effectiveness of traditional

assays in drug discovery, highlights the need for a ‘sea change’ in the drug-discovery paradigm. Predictive in silico models could be used for the desired-property identification, accelerating the selection process of leads and predicting their modes of action [86]. One of the most important features of any QSAR model is its ability to predict the desired property for new compounds, from databases of chemicals [18]. Computational in silico screening of large databases considering the use of such models, has emerged as an interesting alternative to high-throughput screening (HTS) and as an important drug-discovery tool [87, 88]

In order to test the potential of the **TOMOCOMD-CARDD** method and LDA, for detecting novel antiprotozoan compounds, we predicted the biological activity of all the chemicals contained in our ‘in-house’ collection of quinoxaline derivatives that were provided by one of our synthesis research teams from IQM, CSIC, Spain [89]. The structures of these compounds are depicted in Fig. 2.

All these compounds were initially screened (evaluated) with the QSAR models **11** and **12**, also they were assayed in vitro in order to corroborate the predictions against *Tv*. Table 7 summarizes these theoretical and biological achievements.

In general, it was observed a pretty good concordance between the theoretical predictions and the observed activity for both active and inactive compounds. Our trained LDA-assisted QSAR models (Eqs. 11 and 12) were able of successfully classify 6 out of 7 compounds, yielding (both) an accuracy of 85.71%.

As for the in vitro experiments, it should be highlighted that almost all compounds (VA7-34, VA7-37, VA7-38, VA7-68) exhibited pronounced cytotoxic activities of 100% at the concentration of 100 µg/ml and at 24 h (48 h) but VA7-35 and VA7-70: 98.66% (99.40%), and 99.83% (100%), respectively. It is remarkable that these compounds did not showed toxic activity in macrophages cultivations at this concentration. In addition, as observed in Table 7, compounds VA7-37, VA7-38 and VA7-70 maintained a high trichomonacidal activity (98.38%, 97.59% and 94.38%, respectively) and low non-specific cytotoxicity at concentrations of 10 µg/ml at 24 h. However, only VA7-37 and VA7-38 remained with high levels of percentage of reduction of *Tv* (94.23% and 98.10%, respectively), at 48h at this concentration.

These last results can be considered as a promising starting point for the future design and refinement of novel compounds with higher anti-trichomonal activity with low toxicity. Although compounds VA7-37 and VA7-38 were active at higher doses than metronidazole, MTZ (reference drug), this result leaves a door open to a *virtual* variational study of the structure of these compounds in order to improve their activity. Besides our current results are

significant because they demonstrate the straightforward way in which the **TOMOCOMD-CARDD** method can identify new trichomonacidal agents.

Concluding remarks

The bioinformatic tools **TOMOCOMD-CARDD** and STATISTICA 6.0 and, therefore, the underlying work philosophy were successfully applied to the discovery of novel antitrichomonals. Combined features of bond-based linear non-stochastic and stochastic MDs joined to LDA technique allowed us to generate robust *biosilico* models capable of discriminating among active and inactive chemicals. The models’ predictive power was assessed in a simulated experiment, where these screening functions identified chemical agents already proved against *Tv*, non-stochastic model outperformed the stochastic one. Later, our approach permitted us the generation of novel drug-like compounds, which were in vitro assayed achieving promissory results as possible alternatives to MTZ treatment of trichomoniasis.

Acknowledgements The authors wish to express their gratitude to Prof. Dr. Jorge Gálvez for his attention to this work and valuable suggestions. Yovani Marrero-Ponce (M.-P. Y) acknowledges the Valencia University for kind hospitality during the second semester of 2007. M.-P. Y thanks are given to the international relationships of Valencia University, (Spain) for partial financial support as well as the program ‘Estades Temporals per an Investigadors Convitats’ for a fellowship to work at Valencia University. Some authors’ thanks support from Spanish MEC (Project Reference: SAF2006-04698). Finally, 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). Last but not least, Yovani Marrero-Ponce would like to express thanks for the partial support received from the project entitled Strengthening postgraduate education and research in Pharmaceutical Sciences. This project is funded by the Flemish Interuniversity Council (VLIR) of Belgium.

References

- Krieger JN (2000) Sex Transm Dis 27:241
- Petrin D, Delgaty K, Bhatt R, Garber G (1998) Clin Microbiol Rev 11:300
- World-Health-Organization (1995) An overview of selected curable sexually transmitted diseases. World Health Organization, Geneva Switzerland, p 2
- Cosar C, Julou L (1959) Ann Inst Pasteur 96:238
- Centers for Disease Control and Prevention (1993) Morb Mortal Wkly Rep 42(RR-14) 70
- Knight R (1980) J Antimicrob Chemother 6:577
- Gillette H, Schmid GP, Moswe D (1985) Metronidazole-resistant *Trichomonas vaginalis*, a case series, Denver, 1999
- Lumsden WHR, Robertson DHH, Heyworth R, Harrison C (1988) Genitourin Med 64:217
- Narcisi EM, Secor WE (1996) Antimicrob Agents Chemother 40:1121
- Narcisi EM, Secor WE (1996) Antimicrob Agents Chemother 40:1121

11. Houang ET, Ahmet Z, Lawrence AG (1997) Sex Transm Dis 24:116
12. Pattman RS, Sprott MS, Kerns AM, Earnshaw M (1989) Genitourin Med 65:274
13. Wong CA, Wilson PD, Chew TA (1990) Aust N Z J Obstet Gynaecol 30:169
14. Livengood CHI, Lossick JG (1991) Obstet Gynecol 78:954
15. Watson PG, Pattman RS (1996) Int J STD AIDS 7:296
16. Nyirjesy P, Sobel JD, Weitz MV (1998) Clin Infect Dis 26:986
17. Nyirjesy P, Weitz MV, Gelone SP, Fekete T (1995) Lancet 346:1110
18. Estrada E, Peña A (2000) Bioorg Med Chem 8:2755
19. Estrada E, Uriarte E, Montero A, Teijeira M, Santana L, De Clercq E (2000) J Med Chem 43:1975
20. Marrero-Ponce Y, Romero V, 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 can be obtained upon request to Y. Marrero: yovanimp@uclv.edu.cu or ymarrero77@yahoo.es.
21. Marrero-Ponce Y (2003) Molecules 8:687
22. Marrero-Ponce Y (2004) J Chem Inf Comput Sci 44:2010
23. Marrero-Ponce Y (2004) Bioorg Med Chem 12:6351
24. Marrero-Ponce Y, Castillo-Garit JA, Torrens F, Romero-Zaldivar V, Castro E (2004) Molecules 9:1100
25. Marrero-Ponce Y, Díaz HG, Romero V, Torrens F, Castro EA (2004) Bioorg Med Chem 12:5331
26. Marrero-Ponce Y, Cabrera MA, Romero V, Ofori E, Montero LA (2003) Int J Mol Sci 4:512
27. Marrero-Ponce Y, Cabrera MA, Romero V, González DH, Torrens F (2004) J Pharm Pharmaceut Sci 7:186
28. Marrero-Ponce Y, Cabrera MA, Romero-Zaldivar V, Bermejo M, Siverio D, Torrens F (2005) Internet Electrón J Mol Des 4:124
29. Marrero-Ponce Y, Castillo-Garit JA, Olazabal E, Serrano HS, Morales A, Castanedo N, Ibarra-Velarde F, Huesca-Guillen A, Sanchez AM, Torrens F, Castro EA (2005) Bioorg Med Chem 13:1005
30. Marrero-Ponce Y, Castillo-Garit JA, Olazabal E, Serrano HS, Morales A, Castanedo N, Ibarra-Velarde F, Huesca-Guillen A, Jorge E, del Valle A, Torrens F, Castro EA (2004) J Comput Aided Mol Des 18:615
31. Marrero-Ponce Y, Huesca-Guillen A, Ibarra-Velarde F (2005) J Mol Struct (Theochem) 717:67
32. Marrero-Ponce Y, Montero-Torres A, Zaldivar CR, Veitia MI, Perez MM, Sanchez RN (2005) Bioorg Med Chem 13:1293
33. Marrero-Ponce Y, Medina-Marrero R, Torrens F, Martinez Y, Romero-Zaldivar V, Castro EA (2005) Bioorg Med Chem 13:2881
34. Marrero-Ponce Y, Medina-Marrero R, Martinez Y, Torrens F, Romero-Zaldivar V, Castro EA (2006) J Mol Mod 12:255
35. Marrero-Ponce Y, Nodarse D, González HD, Ramos de Armas R, Romero-Zaldivar V, Torrens F, Castro E (2004) Int J Mol Sci 5:276
36. Marrero-Ponce Y, Castillo-Garit JA, Nodarse D (2005) Bioorg Med Chem 13:3397
37. Marrero-Ponce Y, Medina R, Castro EA, de Armas R, González H, Romero V, Torrens F (2004) Molecules 9:1124
38. Marrero-Ponce Y, Medina-Marrero R, Castillo-Garit JA, Romero-Zaldivar V, Torrens F, Castro EA (2005) Bioorg Med Chem 13:3003
39. Marrero-Ponce Y, Torrens F (2006) J Comp-Aided Mol Des 20:685
40. Casañola-Martin GM, Khan MTH, Marrero-Ponce Y, Ather A, Sultan S, Torrens F, Rotondo R (2007) Bioorg Med Chem 15:1483
41. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Germany
42. Estrada E (1996) J Chem Inf Comput Sci 36:844
43. Estrada E, Molina E (2001) J Mol Graph Model 20:54
44. Estrada E (1995) J Chem Inf Comput Sci 35:31
45. Estrada E, Guevara N, Gutman I (1998) J Chem Inf Comput Sci 38:428
46. Estrada E (1999) J Chem Inf Comput Sci 39:1042
47. Edwards CH, Penney DE (1988) Elementary linear algebra. Prentice-Hall, Englewood Cliffs, New Jersey, USA
48. Marrero Ponce Y (2004) J Chem Inf Comput Sci 44:2010
49. Estrada E, Vilar S, Uriarte E, Gutierrez Y (2002) J Chem Inf Comput Sci 42:1194
50. Estrada E, Peña A, Garcia-Domenech R (1998) J Comput Aided Mol Des 12:583
51. Potapov VM (1978) Stereochemistry. Mir, Moscow
52. Wang R, Gao Y, Lai L (2000) Perspect Drug Dis Des 19:47
53. Ertl P, Rohde B, Selzer P (2000) J Med Chem 43:3714
54. Ghose AK, Crippen GM (1987) J Chem Inf Comput Sci 27:21
55. Miller KJ (1990) J Am Chem Soc 112:8533
56. Gasteiger J, Marsili M (1978) Tetrahedron Lett 19:3181
57. Pauling L (1939) The nature of chemical bond. Cornell University Press, Ithaca (New York)
58. Browder A (1996) Mathematical analysis. An introduction. Springer-Verlag, New York
59. Axler S (1996) Linear algebra done right. Springer-Verlag, New York
60. Daudel R, Lefebvre R, Moser C (1984) Quantum chemistry: methods and applications. Wiley, New York
61. Klein DJ (2003) Internet Electron J Mol Des 2:814
62. Todeschini R, Gramatica P (1998) Perspect Drug Dis Des 9–11:355
63. Consonni V, Todeschini R, Pavan M (2002) J Chem Inf Comput Sci 42:682
64. Kier LB, Hall LH (1986) Molecular connectivity in structure–activity analysis. Research Studies Press, Letchworth, UK
65. Negwer M (1987) Organic-chemical drugs and their synonyms. Akademie-Verlag, Berlin
66. Budavari S, O'Neil M, Ann Smith, Heckelman P, Obenchain J (1999) The Merck Index on CD-ROM. Chapman & Hall and Merck & Co., Inc
67. van de Waterbeemd H (1995) In: van Waterbeemd H (ed) Chemometric methods in molecular design. VCH Publishers, Weinheim, p 265
68. STATISTICA (data analysis software system) vs 6.0.
69. Estrada E, Patlewicz G (2004) Croat Chim Acta 77:203
70. Topliss JG, Edwards RP (1979) J Med Chem 22:1238
71. Wold S, Erikson L (1995) In: van de Waterbeemd H (ed) VCH Publishers, New York, p 309
72. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H (2000) Bioinformatics 16:412
73. Kouznetsov VV, Rivero CJ, Ochoa PC, Stashenko E, Martínez JR, Montero PD, Nogal RJJ, Fernández PC, Muelas SS, Gómez BA, Bahsas A, Amaro L (2005) J Arch Pharm 1:338
74. Kouznetsov VV, Vargas MLY, Tibaduiza B, Ochoa C, Montero PD, Nogal RJJ, Fernández C, Muelas S, Gómez A, Bahsas A, Amaro-Luis J (2004) J Arch Pharm 337:127
75. Gálvez J, Garcia-Domenech R, de Julián-Ortiz JV, Soler R (1995) J Chem Inf Comput Sci 35:272
76. Cercos-del-Pozo RA, Pérez-Giménez F, Salabert-Salvador MT, Garcia-March FJ (2000) J Chem Inf Comput Sci 40:178
77. Gálvez J, García R, Salabert MT, Soler R (1994) J Chem Inf Comput Sci 34:520
78. Johnson RA, Wichern DW (1988) Applied multivariate statistical analysis. Prentice-Hall, New Jersey

79. Golbraikh A, Tropsha A (2002) *J Mol Graph Model* 20:269
80. Rose K, Hall LH, Kier LB (2002) *J Chem Inf Comput Sci* 42:651
81. Mc Farland JW, Gans DJ (1995) In: Waterbeemd H (ed) *Chemometric methods in molecular design*. VCH Publishers, New York, p 295
82. Estrada E, Uriarte E (2001) *Curr Med Chem* 8:1573
83. Gavini E, Juliano C, Mulé A, Pirisino G, Murineddu G, Pinna A (2000) *Arch Pharm (Weinheim)* 333:341
84. Ochoa A, Pérez E, Pérez R, Suárez M, Ochoa E, Rodríguez H, Gómez A, Muelas S, Nogal RJJ, Martínez RA (1999) *Arzneim Forsch* 49:764
85. Kouznetsov V, Rodríguez W, Stashenko E, Ochoa C, Vega C, Rolón M, Montero PD, Escario JA, Gómez BA (2004) *J Heterocyclic Chem* 41:1
86. Watson C (2003) *Biosilico* 1:83
87. Lajiness MS (1990) In: Rouvray DH (ed) *Computational chemical graph theory*. Nova Science, New York, p 299
88. Walters WP, Stahl MT, Murcko MA (1998) *Drug Discov Today* 3:160
89. Castro S, Chicharro R, Arán VJ (2002) *J Chem Soc, Perkin Trans* 1:790