Quick Access to Potential Trichomonacidals through Bond Linear Indices-Trained Ligand-Based *virtual* Screening Models

Yovani Marrero-Ponce,[a,b*] Alfredo Meneses-Marcel,[a,b] Oscar M. Rivera-Borroto,[a,d] Alina Montero,[a] José Antonio Escario,[c] Alicia Gómez Barrio,[c] David Montero Pereira,[c] Juan José Nogal,[c] Ricardo Grau,[d] Francisco Torrens,[b] Froylán Ibarra-Velarde,[e] Richard Rotondo,[f] Ysaias J. Alvarado,[g] Christian Vogel,[h] and Lizet Rodriguez-Machin,[a]

[a]*Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy and Department of Drug Design, Chemical Bioactive Center. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.*
[b]*Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, P.O. Box 22085, E-46071, València, Spain.*
[c]*Departamento de Parasitología, Facultad de Farmacia, UCM, Pza. Ramón y Cajal s/n, 28040 Madrid.*
[d]*Bioinformatics Group, Center of Studies on Informatics (CEI), Faculty of Mathematics, Physics and Computer Science. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.*
[e]*Department of Parasitology, Faculty of Veterinarian Medicinal and Zootecnic, UNAM, Mexico, D.F. 04510, Mexico*
[f]*Mediscovery, Inc. Suite 1050, 601 Carlson Parkway, Minnetonka, MN 55305, USA.*
[g]*Laboratorio de Electrónica Molecular, Departamento de Química, Modulo II, grano de Oro, Facultad Experimental de Ciencias, La Universidad del Zulia (LUZ), Venezuela.*
[h]*Universität Rostock, Institut für Chemie, Abteilung für Organische Chemie, Albert-Einstein-Straße 3a, 18059 Rostock.*

To whom correspondence should be addressed:

Fax: 53-42-281130 [or 53-42-281455] (Cuba) and 963543156 (València)
Phone: 53-42-281192 [or 53-42-281473] (Cuba) and 963543156 (València)
***e*-mail**: ymarrero77@yahoo.es; yovani.marrero@uv.es; ymponce@gmail.com
or yovanimp@qf.uclv.edu.cu
URL: http://www.uv.es/yoma/

Running head: *Potential Antitrichomonacidals Discovery through QSAR Studies*

Abstract: *Trichomonas vaginalis* (Tv) is the causative agent of the most common, non-viral, sexually transmitted disease in women and men world-wide. 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 to the development of new trichomonacidals 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 (LDA) were used to discover novel trichomonacidal chemicals. The obtained models, using non-stochastic and stochastic indices, were able to classify correctly 89.01% (87.50%) and 82.42% (84.38%) of the chemicals in training (test) sets, respectively. These results validate the models for use in the ligand-based *virtual* screening. Also they showed 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 evidenced the robustness of the obtained models. Later, both models were applied to the *virtual* screening of 12 compounds already proved against Tv. As a result, they correctly classified 10 out of 12 (83.33%) and 9 out of 12 (75.00%) of the chemicals, respectively; which is a more important criterion for validating the models. In addition, these classification functions were applied to a library of seven chemicals in order to find novel antitrichomonal agents. These compounds were 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. Besides, out of the seven compounds that were screened, synthesized and biologically assayed, six compounds (VA7-34, VA7-35, VA7-37, VA7-38, VA7-68, VA7-70) showed pronounced cytocidal activity at the concentration of 100μg/ml at 24h (48h) within the range of 98.66%-100% (99.40%-100%) while only two molecules (chemicals VA7-37 and VA7-38) showed high cytocidal activity at the concentration of 10μg/ml at 24h (48h): 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 increase the chance of finding new chemical entities with trichomonacidal activity.

Keywords: *TOMOCOMD-CARDD* Software, Bond-based Linear Indices, LDA-assisted QSAR Models, *virtual* Screening, Trichomonacidal, *in vitro* Cytostatic and Cytocidal Activity.

"You know my methods. Apply them."

Conan Doyle

1. Background

Trichomonas vaginalis (Tv) is a common sexually transmitted infection that is increasingly recognized as an important infection in women and men [1,2]. Recent estimates have suggested that Tv infections account for nearly one-third of the 15.4 million cases of sexually transmitted diseases in the United States [3]. In 1995, the World Health Organization estimated the number of adults with trichomoniasis at 170 million worldwide, more than the numbers for gonorrhea, syphilis, and chlamydia combined [4].

This parasite is also known to be the main cause of vaginitis, cervicitis and urethritis in women and may be responsible for prostatitis and other genito-urinary syndromes in men [5,6]. Infection with this organism has been linked to various additional pathologic manifestations, including cervical neoplasia [7-10], atypical pelvic inflammatory disease [11], and tubal infertility [12], and has been reported to be a risk factor in the development of posthysterectomy cuff cellulites [13]. Infection with Tv has also been related to premature rupture of placental membranes, and low birth weight [14,15]. Intrauterine transmission of cytomegalovirus has been reported to be increased by Tv infection [16]. As similar, this infection can elevate the risk of acquiring human immunodeficiency virus [17,18].

Although Tv was first described by Donné in 1836, research on this organism did not begin until the 20th century. The research has been a progression of phases throughout the last 60 years and has gone from developing axenic culture and defining nutritional requirements to finding an effective treatment. In the 1960s and 1970s, research focused on biochemical tests and microscopic examination to understand the growth characteristics and behavior of the organism. It was not until the 1980s that immunologic methods and molecular biological techniques became available and were applied to study the pathogenesis and immunology of this organism [2]

Until 1959, topical vaginal preparations available against trichomoniasis provided some symptomatic relief but were ineffective as cures [19]. In 1959, a nitroimidazole derivative of a *Streptomyces* antibiotic, azomycin, was found to be highly effective in the systemic treatment of trichomoniasis [20]. This derivative was a,b-hydroxyethyl-2-methyl-5-nitroimidazole, commonly referred to as metronidazole (MTZ) and marketed under the trade name Flagyl. Other nitroimidazoles, although unavailable in North America, are also approved for clinical use in other parts of the world. These include tinidazole [21], ornidazole [22], secnidazole [23], flunidazole [24], nimorazole [25], and carnidazole [26]. These nitroimidazoles are not themselves cytocidal against Tv, but their metabolic products are [27].

MTZ enters the cell through diffusion[4] and is activated in the hydrogenosomes of Tv [28]. Here, the nitro group of the drug is anaerobically reduced by pyruvate-ferredoxin oxidoreductase [28]. This results in cytotoxic nitro radical-ion intermediates that break the DNA strands [29]. The response is rapid: cell division and motility cease within 1h and cell death occurs within 8h as seen in cell culture [30].

The recommended MTZ regimen results in cure rates of approximately 95% [31]. In fact, MTZ is the drug now most widely used in the treatment of anaerobic protozoan parasitic infections caused by Tv, *Giardia duodenalis*, and *Entamoeba histolytica* [20,32-35]**.** In addition, it is remarkably safe compared to the most toxic antiprotozoal products [36].

Although there are clinical reports [37-44] that document the refractoriness of infections with Tv to treatment with MTZ, susceptibility tests have failed to demonstrate conclusively that the parasites isolated from such cases after treatment were resistant to this drug [45,46]. Thus, the resistance of Tv has not been generally accepted as the factor responsible for failure of MTZ therapy [47], since reinfection, irregular medication, poor absorption of the drug, and its inactivation by the vaginal flora have not been excluded [46,48,49]. However, a strain of Tv, unequivocally resistant to MTZ, 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 MTZ and other 5-nitroimidazole derivatives [50].

Although MTZ resistance has been considered rare, treatment of these rare patients who do not respond to treatment is extremely problematic for physicians and is associated with enormous patient suffering [51]. A good alternative to palliate this problem could be clinical treatment with other nitroimidazoles but unfortunately all of them have similar modes of antibacterial activity to MTZ [52], and therefore resistance to MTZ often includes resistance to the other nitroimidazoles [53].

Currently, 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 antiprotozoal drugs makes this development slow. For this reason, 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. This '*in silico'* world of data, analysis, hypothesis, and models that reside inside a computer is alternative to the ‘real’ world of synthesis and screening of compounds in the laboratory [54,55].

At present, many large pharmaceutical industries have reoriented their research strategies seeking to solve the problem of generation/selection of novel chemical entities (NCEs), one of the major bottlenecks in the drug discovery pipeline. In fact, currently most integration projects include efforts to integrate the data associated with NCE generation [56]. Alternatively, several approaches to the computer-aided molecular design and high-throughput *in silico* screening (or *virtual* high-throughput screening) have been introduced in the literature [57]. Nevertheless, novel computational methods and strategies are required to deliver a system that significantly reduces the time-to-market and research and development (R&D) spendings, and increase the rate at which NCEs progress through the pipeline. Such studies if they are implemented successfully can deliver substantial benefits and act as the bedrock for NCE selection [56].

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 ***D***esign-***C***omputer ***A***ided “***R***ational” ***D***rug ***D***esign) [58]. This method has been developed to generate 2D (topologic), 2.5 (3D-chiral) and 3D (topographic and geometric) molecular descriptors based on the application of the discrete mathematics and linear algebra theory to chemistry. In this sense, atomic, atom-type, atom-group and total linear, bilinear and quadratic molecular fingerprints have been defined in analogy to the linear, bilinear and quadratic mathematical maps [59,60]. This *in silico* method has been successfully applied to the prediction of several physical, physicochemical and chemical properties of organic compounds [59,61-63]. In addition, ***TOMOCOMD-CARDD*** has been extended to consider three-dimensional features of small/medium-sized molecules based on the trigonometric-3D-chirality-correction factor approach [64]. This strategy has also been useful for the prediction of the pharmacokinetic properties of organic compounds [65-67]**,** and the selection of novel subsystems of compounds having a desired property/activity [68-73].

Later, promising results have been found in the modeling of the interaction between drugs and HIV-1 RNA packaging region in the field of bioinformatics using the ***TOMOCOMD-CANAR*** (***C***omputed-***A***ided ***N***ucleic ***A***cid ***R***esearch) approach [74,75].

Finally, an alternative formulation of our approach for structural characterization of proteins was carried out recently [76,77]. This extended method ***TOMOCOMD-CAMPS*** (***C***omputed-***A***ided ***M***odeling in ***P***rotein ***S***cience) was used to encompass protein stability studies by means of a combination of protein linear or quadratic indices (macromolecular fingerprints) and statistical (linear and nonlinear model) methods [76,77].

Recently, some of present authors have proposed a new extended local (bond and bond-type) and total (whole) molecular descriptors based on the adjacency of edges and based on quadratic and linear maps 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 (linear) indices, encode topological information given by the molecular graph, weighted by chemical information encoded in selected bond weightings. Finally, the correlation ability of the new descriptors is tested in a QSPR and QSAR studies [78,79].

The main objective of this work 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.

2. Theoretical Framework

The basis of the extension of linear indices that will be given here is the edge-adjacency matrix considered and explicitly defined in the chemical graph-theory literature [80,81], and rediscovered by Estrada as an important source of new molecular descriptors [82-87]. In this section, we first will define the nomenclature to be used in this work, then the atom-based molecular vector ($\bar{x}$) will be redefined for bond characterization using the same approach as previously reported, and finally some new definition of bond-based non-stochastic and stochastic linear indices with its peculiar mathematical properties will be given.

2.1. Background in Graph-Theoretical Edge-Adjacency Matrix

Let $G = (V, E)$ be a simple graph, with $V = \{v_1, v_2, ..., v_n\}$ and $E = \{e_1, e_2, ...e_m\}$ being the vertex- and edge-sets of G, respectively. Then G represents a molecular graph having n vertices and m edge (bonds). The edge-adjacency matrix E of G (likewise called bond adjacency matrix, B) is a square and symmetric matrix whose elements e_{ij} are 1 if and only if edge i is adjacent to edge j [84,87,88]. 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 E is named the edge degree of bond *i*, δ(e_i) [82,85,86,88].

2.1.1. New Edge-Relations: Stochastic Edge-Adjacency Matrix

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

$$ {}^k es_{ij} = \frac{{}^k e_{ij}}{{}^k SUM(E^k)_i} = \frac{{}^k e_{ij}}{{}^k \delta(e)_i} \qquad \textbf{(1)} $$

where, ${}^k e_{ij}$ are the elements of the k^{th} power of E and the SUM of the *i*th row of E^k are named the *k*-order edge degree of bond *i*, ${}^k\delta(e_i)$. Note that the matrix ES^k in Eq. 1 has the property that *the sum of the elements in each row* is 1. Such an *m*x*m* matrix with nonnegative entries having this property is called a “stochastic matrix” [89].

2.2. Chemical Information and Bond-based Molecular Vector

The atom-based molecular vector ($\bar{x}$) used to represent small-to-medium size organic chemicals has been explained in some detail elsewhere [60,63,68,71,73]. In a parallel manner to the development of$\bar{x}$, we present the extension to the bond-based molecular vector ($\bar{w}$). The components (w_i) of $\bar{w}$ are numeric values, which represent a certain standard bond property (bond-label). That is to say, 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 $\Re^5$, $\Re^{10}$, $\Re^{15}$,..., $\Re^m$, respectively; where *m* is the dimension of the real sets ($\Re^m$). This approach allows us encoding organic molecules such as 2-hydroxybut-2-

enenitrile through the molecular vector $\overline{w}$ = [$w_{\text{Csp3-Csp2}}$, $w_{\text{Csp2=Csp2}}$, $w_{\text{Csp2-Osp3}}$, $w_{\text{H-Osp3}}$, $w_{\text{Csp2-Csp}}$, $w_{\text{Csp≡Nsp}}$]. This vector belongs to the product space $\Re^6$.

These properties characterize each kind of bond (and bond-types) 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 [55,90-92], standard bond dipole [55,90-92] or even mathematical expressions involving atomic weights such as atomic Log P [93], surface contributions of polar atoms [94], atomic molar refractivity [95], atomic hybrid polarizabilities [96], and Gasteiger-Marsilli atomic charge [97], atomic electronegativity in Pauling scale [98] and so on. Here, we characterized each bond with the following parameter:

$$w_i = x_i/\delta_i + x_j/\delta_j \qquad \textbf{(2)}$$

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

2.3. Calculation of Linear Indices for Bonds, Bond-Types and the Whole Molecule

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

$$\bar{f}_k(\overline{w}) = [\ \sum_{j=1}^{m} {}^k e_{ij} w^j\ ] = E^k\ \overline{w} \qquad \textbf{(3)}$$

$$^{s}\bar{f}_k(\overline{w}) = [\sum_{j=1}^{m} {}^{k}es_{ij} w^j] = ES^k \; \overline{w} \tag{4}$$

where m is the number of bonds of the molecule and w^j are the coordinates of the bond-based molecular vector ($\overline{w}$) in the so-called canonical ('natural') basis. In this basis system, the coordinates of any vector $\overline{w}$ coincide with the components of this vector [89,99,100]. For that reason, those coordinates can be considered as weights (bond-labels) of the edge of the molecular graph. The coefficients $^{k}e_{ij}$ and $^{k}es_{ij}$ are the elements of the k^{th} power of the matrix E(G) and ES(G), correspondingly, of the molecular graph. The defining equations (**3**) and (**4**) for $\bar{f}_k(\overline{w})$ and $^{s}\bar{f}_k(\overline{w})$, respectively, may be also written as the single matrix equation (see Eqs. **3** and **4**), where $\overline{w}$ is a column vector (an mx1 matrix) of the coordinates of $\overline{w}$ in the canonical basis of $\Re^m$. Here, $E^{\boldsymbol{k}}$ and $ES^{\boldsymbol{k}}$ denote the matrices of linear maps with respect to the natural basis set.

Note that both linear maps are defined as a linear transformation $\bar{f}_k(\overline{w})$ on molecular vector space $\Re^m$. This map is a correspondence that assigns a vector $\bar{f}(\overline{w})$ to a vector $\overline{w}$ in $\Re^m$ in such a way that:

$$\bar{f}(\lambda_1 \overline{w}_1 + \lambda_2 \overline{w}_2) = \lambda_1 \bar{f}(\overline{w}_1) + \lambda_2 \bar{f}(\overline{w}_2) \tag{5}$$

for any scalar λ_1, λ_2 and any vector $\overline{w}_1$, $\overline{w}_2$ in $\Re^m$.

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

$$f_k(\overline{w}) = \sum_{i=1}^{m} f_{ki}(\overline{w}) = \sum_{i=1}^{m} \sum_{j=1}^{m} {}^{k}e_{ij} w^j = \sum_{i=1}^{m} \overline{u}_i^{\,t} \; E^k \; \overline{w} \tag{6}$$

$$ {}^{s}f_{k}(\overline{w}) = \sum_{i=1}^{m} {}^{s}f_{ki}(\overline{w}) = \sum_{i=1}^{m}\sum_{j=1}^{m} {}^{k}es_{ij}\, w^{j} = \sum_{i=1}^{m} \overline{u}_{i}^{\,t}\, ES^{k}\, \overline{w} \qquad \textbf{(7)} $$

where *m* is the number of bonds, $f_{ki}(\overline{w})$ and ${}^{s}f_{ki}(\overline{w})$ are the local non-stochastic and stochastic linear indices obtained by Eqs. **3** and **4** as the coordinates of $\bar{f}_{k}(\overline{w})$ and ${}^{s}\bar{f}_{k}(\overline{w})$ referred to the canonical basis, respectively. It means, whenever we calculate total bond-based non-stochastic and stochastic linear indices in fact we are calculating the 1-norm associated to the linear maps ($\bar{f}_{k}(\overline{w})$ and ${}^{s}\bar{f}_{k}(\overline{w})$) well known and defined in mathematical analysis for such vector spaces as $\Re^{m}$ [99]. Then, both total linear forms, $f_{k}(\overline{w})$ and ${}^{s}f_{k}(\overline{w})$, can also be written in matrix form for each molecular vector $\overline{w} \in \Re^{m}$, where $\overline{u}_{i}^{\,t}$ is an *m*-dimensional unitary row vector (see Eqs. **3** and **4**). As it can be seen, the *k*th total linear indices (both non-stochastic and stochastic) are calculated by summing the local (bond) linear indices of all bonds in the molecule.

Finally, in addition to total and bond linear indices computed for each bond in the molecule, local-fragment (bond-type) formalism can be developed. The *k*th bond-type linear index of the edge-adjacency matrix is calculated by summing up the *k*th bond linear indices of all bonds of the same bond type in the molecule. That is to say, this extension of the bond linear index is similar to the group additive schemes, in which an index appears for each bond type in the molecule together with its contribution based on the bond linear index. Consequently, if a molecule is partitioned into Z molecular fragments, the total non-stochastic (or stochastic) linear indices can be partitioned into Z local non-stochastic (or stochastic) linear indices $f_{kL}(\overline{w})$ (or ${}^{s}f_{kL}(\overline{w}_{i})$), L = 1, …, Z. That is to say, the total (both non-stochastic and stochastic) linear indices of order *k* can be expressed as the sum of the local linear indices of the Z fragments of the same order:

$$f_k(\overline{w}) = \sum_{L=1}^{Z} f_{kL}(\overline{w}) \qquad \textbf{(8)}$$

$$^{s}f_k(\overline{w}) = \sum_{L=1}^{Z} {}^{s}f_{kL}(\overline{w}) \qquad \textbf{(9)}$$

In the bond-type linear indices formalism, each bond in the molecule is classified into a bond-type (fragment). In this sense, 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 the basis for application to a wider range of biological problems in which the local formalism is applicable without the need for superposition of a closely related set of structures.

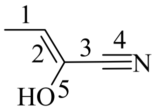
It is useful to perform a calculation on a molecule to illustrate the steps in the procedure. For this, in the next section the calculus of the non-stochastic and stochastic linear indices of the bond matrix (both total and local) using a simple chemical example is depicted. In that section, it will also stand out that our approach is rather similar to the **LCBO-MO** (**L**inear **C**ombination of **B**ond **O**rbitals-**M**olecular **O**rbitals) method (e.g., for *k* = 1)[101]. **LCBO-MO** is another way of forming molecular orbitals by taking linear combinations of functions associated with the different bonds in the molecule. In this sense, MOs are made up as LCBO of bonds composing the system, i.e. are written in the form,

$$\varphi_i = \sum_{j=1}^{n} c_{ij} \psi_j \qquad \textbf{(10)}$$

where i is the number of the MO, φ (in our case, $f_i(\overline{w})$); j are the numbers of bond ψ–orbitals (in our case, w^j); c_{ij} (in our case, ${}^1e_{ij}$ or ${}^1es_{ij}$ for non-stochastic and stochastic indices, respectively) are the numerical coefficients defining the contributions of individuals BOs to the given MO. Although the **LCAO** (**L**inear **C**ombination of **A**tom **O**rbitals) approximation has been particularly useful for the study of conjugated hydrocarbons, the **LCBO** method has been particularly applied to the calculation of 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 [101].

2.4. Sample Calculation

The linear indices of the bond matrix are calculated in the following way. Considering the molecule of 2-hydroxybut-2-enenitrile as a simple example, we have the following labeled molecular graph and bond-based adjacency matrices (E and ES). The second (k = 2) and third (k = 3) power of these matrices and bond-based molecular vector, $\overline{w}$ are also given:



$$E^0 = ES^0 = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix} \quad E^1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \quad E^2 = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 3 & 1 & 1 & 1 \\ 1 & 1 & 3 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 2 \end{bmatrix} \quad E^3 = \begin{bmatrix} 0 & 3 & 1 & 1 & 1 \\ 3 & 2 & 5 & 1 & 4 \\ 1 & 5 & 2 & 3 & 4 \\ 1 & 1 & 3 & 0 & 1 \\ 1 & 4 & 4 & 1 & 2 \end{bmatrix}$$

$$ES^1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0.33 & 0 & 0.33 & 0 & 0.33 \\ 0 & 0.33 & 0 & 0.33 & 0.33 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 \end{bmatrix} ES^2 = \begin{bmatrix} 0.33 & 0 & 0.33 & 0 & 0.33 \\ 0 & 0.5 & 0.16 & 0.16 & 0.16 \\ 0.16 & 0.16 & 0.5 & 0 & 0.16 \\ 0 & 0.33 & 0 & 0.33 & 0.33 \\ 0.16 & 0.16 & 0.16 & 0.16 & 0.33 \end{bmatrix} \quad ES^3 = \begin{bmatrix} 0 & 0.5 & 0.16 & 0.16 & 0.16 \\ 0.2 & 0.13 & 0.33 & 0.06 & 0.26 \\ 0.06 & 0.33 & 0.13 & 0.2 & 0.26 \\ 0.16 & 0.16 & 0.5 & 0 & 0.16 \\ 0.083 & 0.33 & 0.33 & 0.083 & 0.16 \end{bmatrix}$$

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, $\overline{w}$ = [w_1, w_2, w_3, w_4, w_5] = [$w_{(\text{C-C})}$, $w_{(\text{C=C})}$, $w_{(\text{C-C})}$, $w_{(\text{C≡N})}$, $w_{(\text{C-O})}$] and each "bond orbital" can be computed by Eq. **2** using, for instance, the atomic electronegativity in Pauling scale (x_i) [98] as atomic weight (atom-label):

$w_1 = x_C/1 + x_C/3 = 2.55/1 + 2.55/3 = 3.4$

$w_2 = x_C/3 + x_C/4 = 2.55/3 + 2.55/4 = 1.4875$

$w_3 = x_C/4 + x_C/4 = 2.55/4 + 2.55/4 = 1.275$

$w_4 = x_C/4 + x_N/3 = 2.55/4 + 3.04/3 = 1.650833$

$w_5 = x_C/4 + x_O/1 = 2.55/4 + 3.44/1 = 4.0775$

and therefore, $\overline{w}$ = [3.4, 1.4875, 1.275, 1.650833, 4.0775]

Each non-stochastic and stochastic "molecular orbital" will have the form:

$$f_{ki}(\overline{w}) = {}^k e_{i1} w^1 + {}^k e_{i2} w^2 + {}^k e_{i3} w^3 + {}^k e_{i4} w^4 + {}^k e_{i5} w^5 \qquad \textbf{(11)}$$

$${}^s f_{ki}(\overline{w}) = {}^k es_{i1} w^1 + {}^k es_{i2} w^2 + {}^k es_{i3} w^3 + {}^k es_{i4} w^4 + {}^k es_{i5} w^5 \qquad \textbf{(12)}$$

The ${}^k e_{ii}$'s and ${}^k es_{ii}$'s can be considered as a measure of the attraction of an electron for a bond in the k step. The ${}^k e_{ij}$'s and ${}^k es_{ij}$'s are the terms of interaction between two bonds in the k-step. The ${}^{k}e_{ij} = {}^{k}e_{ji}$ are equal by symmetry (non-oriented molecular graph). However, ${}^{k}es_{ij}$'s $\neq {}^{k}es_{ji}$'s. This is a logical result because the k^{th} es_{ij} elements are the transition probabilities with the 'electrons' moving from bond i to j at the discrete time periods t_k and it should be different in both senses. This result is in total agreement if the electronegativity of the two atom types in the bonds are taken into account.

In this way, $E^{\boldsymbol{k}}$ and ES^{k} can be seen as graph–theoretic electronic–structure models[102]. In fact, quantum chemistry starts from the fact that a molecule is made up of electrons and nuclei. The distinction here between bonded and non-bonded atoms is difficult to justify. Any two nuclei of 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. In this sense, the 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, 3,…) through the chemical-bonding network. That is to say, the E^1 and ES^1 matrices consider the valence-bond electrons in one step and their power (k = 0, 1, 2, 3…) can be considering as an interacting–electron chemical–network model in k step. This model can be seen as an intermediate between the quantitative quantum-mechanical Schrödinger equation and classical chemical bonding ideas [102].

On the other hand, the k^{th} (k = 0-3) non-stochastic bond linear indices can be calculated for this molecule as follows:

$$f_{0i}(\overline{w}) = \sum_{j=1}^{5} {}^{0}e_{ij}w^{j} = \overline{u}_i^{\,t}\, E^{0}\, \overline{w} = {}^{0}e_{i1}w^{1} + {}^{0}e_{i2}w^{2} + {}^{0}e_{i3}w^{3} + {}^{0}e_{i4}w^{4} + {}^{0}e_{i5}w^{5}$$

$$f_{1i}(\overline{w}) = \sum_{j=1}^{5} {}^{1}e_{ij}w^{j} = \overline{u}_i^{\,t}\, E^{1}\, \overline{w} = {}^{1}e_{i1}w^{1} + {}^{1}e_{i2}w^{2} + {}^{1}e_{i3}w^{3} + {}^{1}e_{i4}w^{4} + {}^{1}e_{i5}w^{5}$$

$$f_{2i}(\overline{w}) = \sum_{j=1}^{5} {}^{2}e_{ij}\,w^{j} = \overline{u}_i^{\,t}\; E^{2}\; \overline{w} = {}^{2}e_{i1}w^{1} + {}^{2}e_{i2}w^{2} + {}^{2}e_{i3}w^{3} + {}^{2}e_{i4}w^{4} + {}^{2}e_{i5}w^{5}$$

$$f_{3i}(\overline{w}) = \sum_{j=1}^{5} {}^{3}e_{ij}\,w^{j} = \overline{u}_i^{\,t}\; E^{3}\; \overline{w} = {}^{3}e_{i1}w^{1} + {}^{3}e_{i2}w^{2} + {}^{3}e_{i3}w^{3} + {}^{3}e_{i4}w^{4} + {}^{3}e_{i5}w^{5}$$

and stochastic linear indices for each bond *i* can be computed for this molecule in a similar form:

$${}^{s}f_{0i}(\overline{w}) = \sum_{j=1}^{5} {}^{0}es_{ij}\,w^{j} = \overline{u}_i^{\,t}\; ES^{0}\; \overline{w} = {}^{0}es_{i1}w^{1} + {}^{0}es_{i2}w^{2} + {}^{0}es_{i3}w^{3} + {}^{0}es_{i4}w^{4} + {}^{0}es_{i5}w^{5}$$

$${}^{s}f_{1i}(\overline{w}) = \sum_{j=1}^{5} {}^{1}es_{ij}\,w^{j} = \overline{u}_i^{\,t}\; ES^{1}\; \overline{w} = {}^{1}es_{i1}w^{1} + {}^{1}es_{i2}w^{2} + {}^{1}es_{i3}w^{3} + {}^{1}es_{i4}w^{4} + {}^{1}es_{i5}w^{5}$$

$${}^{s}f_{2i}(\overline{w}) = \sum_{j=1}^{5} {}^{2}es_{ij}\,w^{j} = \overline{u}_i^{\,t}\; ES^{2}\; \overline{w} = {}^{2}es_{i1}w^{1} + {}^{2}es_{i2}w^{2} + {}^{2}es_{i3}w^{3} + {}^{2}es_{i4}w^{4} + {}^{2}es_{i5}w^{5}$$

$${}^{s}f_{3i}(\overline{w}) = \sum_{j=1}^{5} {}^{3}es_{ij}\,w^{j} = \overline{u}_i^{\,t}\; ES^{3}\; \overline{w} = {}^{3}es_{i1}w^{1} + {}^{3}es_{i2}w^{2} + {}^{3}es_{i3}w^{3} + {}^{3}es_{i4}w^{4} + {}^{3}es_{i5}w^{5}$$

The total non-stochastic linear indices can be expressed as the sum of the local (bond) linear indices for this molecule as follows:

$$f_{0}(\overline{w}) = \sum_{i=1}^{5} f_{0i}(\overline{w}) = \sum_{i=1}^{5} \overline{u}_i^{\,t}\; E^{0}\; \overline{w} = f_{01}(\overline{w}) + f_{02}(\overline{w}) + f_{03}(\overline{w}) + f_{04}(\overline{w}) + f_{05}(\overline{w})$$

$$= 3.4 + 1.4875 + 1.275 + 1.650833 + 4.0775 = 11.89083$$

$$f_{1}(\overline{w}) = \sum_{i=1}^{5} f_{1i}(\overline{w}) = \sum_{i=1}^{5} \overline{u}_i^{\,t}\; E^{1}\; \overline{w} = f_{11}(\overline{w}) + f_{12}(\overline{w}) + f_{13}(\overline{w}) + f_{14}(\overline{w}) + f_{15}(\overline{w})$$

$$= 1.4875 + 8.7525 + 7.215833 + 1.275 + 2.7625 = 21.49333$$

$$f_{2}(\overline{w}) = \sum_{i=1}^{5} f_{2i}(\overline{w}) = \sum_{i=1}^{5} \overline{u}_i^{\,t}\; E^{2}\; \overline{w} = f_{21}(\overline{w}) + f_{22}(\overline{w}) + f_{23}(\overline{w}) + f_{24}(\overline{w}) + f_{25}(\overline{w})$$

$$= 8.7525 + 11.46583 + 12.79 + 7.215833 + 15.96833 = 56.1925$$

$$f_{3}(\overline{w}) = \sum_{i=1}^{5} f_{3i}(\overline{w}) = \sum_{i=1}^{5} \overline{u}_i^{\,t}\; E^{3}\; \overline{w} = f_{31}(\overline{w}) + f_{32}(\overline{w}) + f_{33}(\overline{w}) + f_{34}(\overline{w}) + f_{35}(\overline{w})$$

$$= 11.46583 + 37.51083 + 34.65 + 12.79 + 24.25583 = 120.6725$$

The terms in the summations for calculating the total linear indices are the so-called bond linear indices. We have written these terms in the consecutive order of the bond labels in the graph. For instance, the non-stochastic bond linear indices of order 0, 1, 2 and 3 for the bond labeled as 1 are 3.4, 1.4875, 8.7525, and 11.46583, respectively.

The *k*th total stochastic linear indices values are also the sum of the *k*th local (bond) stochastic linear indices values for all bonds in the molecule:

$${}^{s}f_0(\overline{w}) = \sum_{i=1}^{5} {}^{s}f_{0i}(\overline{w}) = \sum_{i=1}^{5} \overline{u}_i^{\,t}\; ES^{0}\; \overline{w} = {}^{s}f_{01}(\overline{w}) + {}^{s}f_{02}(\overline{w}) + {}^{s}f_{03}(\overline{w}) + {}^{s}f_{04}(\overline{w}) + {}^{s}f_{05}(\overline{w})$$

$$= 3.4 + 1.4875 + 1.275 + 1.650833 + 4.0775 = 11.89083$$

$${}^{s}f_1(\overline{w}) = \sum_{i=1}^{5} {}^{s}f_{1i}(\overline{w}) = \sum_{i=1}^{5} \overline{u}_i^{\,t}\; ES^{1}\; \overline{w} = {}^{s}f_{11}(\overline{w}) + {}^{s}f_{12}(\overline{w}) + {}^{s}f_{13}(\overline{w}) + {}^{s}f_{14}(\overline{w}) + {}^{s}f_{15}(\overline{w})$$

$$= 1.4875 + 2.9175 + 2.405278 + 1.275 + 1.38125 = 9.466528$$

$${}^{s}f_2(\overline{w}) = \sum_{i=1}^{5} {}^{s}f_{2i}(\overline{w}) = \sum_{i=1}^{5} \overline{u}_i^{\,t}\; ES^{2}\; \overline{w} = {}^{s}f_{21}(\overline{w}) + {}^{s}f_{22}(\overline{w}) + {}^{s}f_{23}(\overline{w}) + {}^{s}f_{24}(\overline{w}) + {}^{s}f_{25}(\overline{w})$$

$$= 2.9175 + 1.910972 + 2.131667 + 2.405278 + 2.661389 = 12.02681$$

$${}^{s}f_3(\overline{w}) = \sum_{i=1}^{5} {}^{s}f_{3i}(\overline{w}) = \sum_{i=1}^{5} \overline{u}_i^{\,t}\; ES^{3}\; \overline{w} = {}^{s}f_{31}(\overline{w}) + {}^{s}f_{32}(\overline{w}) + {}^{s}f_{33}(\overline{w}) + {}^{s}f_{34}(\overline{w}) + {}^{s}f_{35}(\overline{w})$$

$$= 1.910972 + 2.500722 + 2.31 + 2.131667 + 2.021319 = 10.87468$$

3. Methods

3.1. TOMOCOMD-CARDD Approach

TOMOCOMD is an interactive program for molecular design and bioinformatic research [58]. It is composed of four subprograms; each one of them 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.

3.1.1. Computational Strategies

The main steps for the application of present method in QSAR/QSPR and drug design can be briefly summarized in the following set of steps: 1) Draw the molecular pseudographs for each molecule of the data set, 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 properties used are those previously proposed for the calculation of the DRAGON descriptors [98,103,104] i.e., atomic mass (M), atomic polarizability (P), atomic Mullinken electronegativity (K), van der Waals atomic volume (V), plus the atomic electronegativity in Pauling scale (G) [105].The values of these atomic labels are shown in Table 1. In order to calculate the required weights, we used the mathematical expression given by Eq. 2, which involve atomic weights, 3) Compute the total and local (bond and bond-type) non-stochastic and stochastic linear indices. It can be carried out in the software calculation mode, where you 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, and columns correspond to the total and local bond-based linear indices or other molecular descriptors family implemented in this program, 4) Find a 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. That is to say, we can find a quantitative relation between an activity **A** and the linear indices having, for instance, the following appearance, $\mathbf{A} = a_0\boldsymbol{f}_0(\overline{w}) + a_1\boldsymbol{f}_1(\overline{w}) + a_2\boldsymbol{f}_2(\overline{w}) +....+ a_k\boldsymbol{f}_k(\overline{w}) + \mathrm{c}$, where **A** is the measured activity, $\boldsymbol{f}_k(\overline{w})$ are the k^{th} total bond-based linear indices, and the $\boldsymbol{a}_k$*'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 (using a test set and an external predicting set) validation techniques, and 6) Apply the obtained LDA-based QSAR models as 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) k^{th} $\left(k = \overline{0,15}\right)$ total non-stochastic bond-based linear indices not considering and considering H-atoms in the molecular graph (G) [$\boldsymbol{f_k}(\overline{w})$ and $\boldsymbol{f_k}^{\mathrm{H}}(\overline{w})$, respectively].

2) k^{th} $\left(k = \overline{0,15}\right)$ total stochastic bond-based linear indices not considering and considering H-atoms in the molecular graph (G) [${}^{\mathbf{s}}\boldsymbol{f_k}(\overline{w})$ and ${}^{\mathbf{s}}\boldsymbol{f_k}^{\mathrm{H}}(\overline{w})$, respectively].

3) k^{th} $\left(k=\overline{0,15}\right)$ bond-type (group = heteroatoms: S, N, O) non-stochastic linear indices not considering and considering H-atoms in the molecular graph (G) [$\boldsymbol{f}_{k\mathrm{L}}(\overline{w}_{\mathrm{E}})$ and $\boldsymbol{f}_{k\mathrm{L}}{}^{\mathrm{H}}(\overline{w}_{\mathrm{E}})$, correspondingly]. These local descriptors are putative molecular charge, dipole moment, and H-bonding acceptors.

4) k^{th} $\left(k=\overline{0,15}\right)$ bond-type (group = heteroatoms: S, N, O) stochastic linear indices not considering and considering H-atoms in the molecular graph (G) [${}^{\mathbf{s}}\boldsymbol{f}_{k\mathrm{L}}(\overline{w}_{\mathrm{E}})$, and ${}^{\mathbf{s}}\boldsymbol{f}_{k\mathrm{L}}{}^{\mathrm{H}}(\overline{w}_{\mathrm{E}})$, correspondingly]. These local descriptors are putative molecular charge, dipole moment, and H-bonding acceptors.

Table 1. Values of the Atom Weights Used for Linear Indices Calculation.

ID	Atomic Mass	VdW Volume	Mulliken Electronegativity	Polarizability	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

3.2.Data Set for QSAR Study

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 [106], and Merck Index [107], 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 set makes possible the discovery of lead compounds, not only with determined mechanisms of antitrichomonal activity, but also with novel modes of action (which will be illustrated well in this paper in a *virtual* experiment for lead compounds generation) The remaining subseries consisting of 10 trichomonacidals and 22 non-trichomonacidals 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.

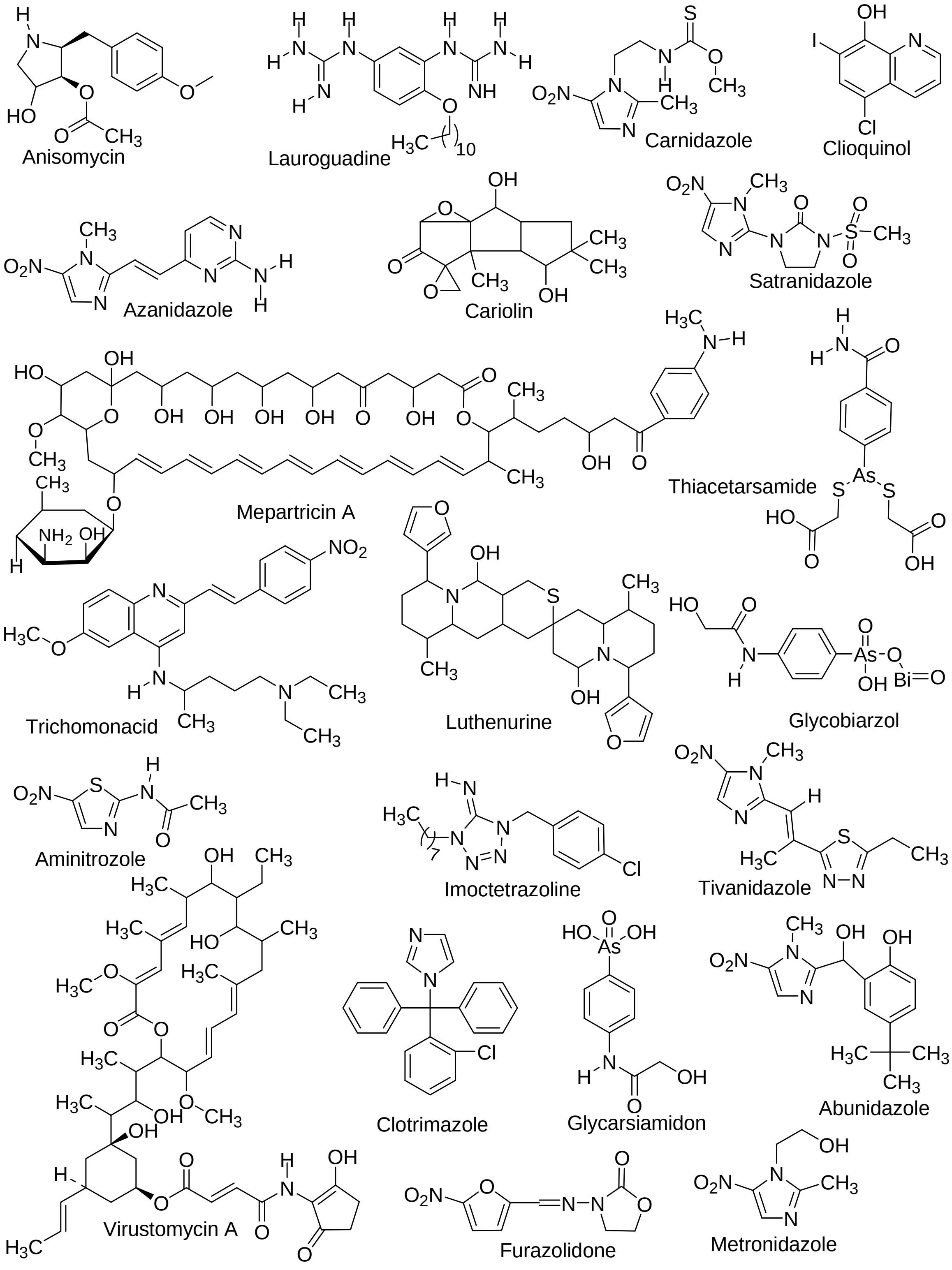


Figure 1. Random sample of the molecular families of trichomonacidal agents studied here.

3.2.1. Statistical Analysis

The discriminant functions were obtained by using the Linear Discriminant Analysis (LDA) [108] as implemented in the STATISTICA [109]. The default parameters of this program were used in the development of the model. Forward stepwise was fixed as the strategy for variable selection. The principle of parsimony (Occam's razor) was taken into account as they strategy for model selection. In its original form, the Occam's razor states that »*Numquam ponenda est pluritas sin necesitate*«, which can be translated as »Entities should not be multiplied beyond necessity« [110]. In this case, simplicity is loosely equated with the number of parameters in the model. If we understand the predictive error to be the error rate for unseen examples, the Occam's razor can be stated for the selection of QSAR/QSPR models as ("*QSAR/QSPR Occam's Razor*"): Given two QSAR/QSPR models with the same predictive error, the simplest one should be preferred because simplicity is desirable in itself [110]. In this connection, we select the model with higher statistical signification but having as few parameters (a_k) as possible.

The quality of the models were determined by examining Wilks' λ parameter (*U*-statistic), squared Mahalanobis distance (D^2), Fisher ratio (F) and the corresponding *p*-level (*p*(F)) as well as the percentage of good classification in the training and test sets [108]. Models with a proportion between the number of cases and variables in the equation lower than 5 were rejected.

The Wilks' λ for the overall discrimination can take values in the range of 0 (perfect discrimination) to 1 (no discrimination). The D^2 statistics 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 ΔP% > 0, being ΔP% = [P(Active) - P(Inactive)]x100 or inactive otherwise. P(Active) and P(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 [111]. Also a leave-group-out (LGO) cross-validation strategy was carried out. In this case, 10% of the data set was used as group size, i.e. groups including 10% of the training data set 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 powers of the obtained models [111].

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 [112].

3.3. Biological Assay: Determination of in vitro Trichomonacidal 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% 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. 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 activities (%CA), in relation to controls as previously reported [113,114].

4. Results and Discussion

4.1. 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 [108]. The use of LDA in rational drug design has been extensively reported by different authors [55,62-64,66-74]. 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 [109], the following linear models were obtained; in which total as well as local non-stochastic and stochastic bond-based linear indices were used as independent variables:

$$\textbf{\textit{Class}} = -5{,}53 - 2{,}96\text{x}10^{-5}\,{}^{M}\boldsymbol{f}_{6L}{}^{H}(\overline{w}_{E}) - 0{,}07\,{}^{M}\boldsymbol{f}_{0L}{}^{H}(\overline{w}_{E}) - 0{,}05\,{}^{M}\boldsymbol{f}_{0L}(\overline{w}_{E}) - 5{,}29\text{x}10^{-4}\,{}^{P}\boldsymbol{f}_{7L}(\overline{w}_{E}) + 4{,}73\text{x}10^{-5}\,{}^{V}\boldsymbol{f}_{7L}(\overline{w}_{E}) + 0{,}36\,{}^{E}\boldsymbol{f}_{0L}{}^{H}(\overline{w}_{E}) \qquad \textbf{(13)}$$

$$N = 91 \qquad \lambda = 0{,}46 \qquad D^{2} = 4{,}54 \qquad F(6{,}84) = 15{,}99 \qquad p<0{,}0000$$

$$\textbf{\textit{Class}} = -4{,}93 - 0{,}12\,{}^{Ms}\boldsymbol{f}_{9L}{}^{H}(\overline{w}_{E}) + 0{,}10\,{}^{Vs}\boldsymbol{f}_{0L}{}^{H}(\overline{w}_{E}) + 1{,}20\,{}^{Es}\boldsymbol{f}_{2L}{}^{H}(\overline{w}_{E}) - 0{,}77\,{}^{Es}\boldsymbol{f}_{5L}(\overline{w}_{E}) \qquad \textbf{(14)}$$

$$N = 91 \qquad \lambda = 0{,}48 \qquad D^{2} = 4{,}28 \qquad F(6{,}84) = 123{,}18 \qquad p<0{,}0000$$

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.

Model **13** classifies correctly 87.50 % of active and 90.20% of inactive compounds in the training set for a global good classification (accuracy) of 89.01%. Model **14** classifies correctly 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. **13** and **14** show an 87.50% (30/32) and 84.38% (27/32) of global predictability in the prediction series, respectively. 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 [115].

In Tables 2 and 3 we give the names of all 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 inactive set 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 [112,116]. While the sensitivity is the probability of correctly predicting a positive example, the specificity is the probability that a positive prediction is 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 [112,116]. The obtained models, Eqs. **13** and **14**, showed a high ***C*** of 0.78 (0.71) and 0.65 (0.65) in training (test) sets, correspondingly.

Table 2. 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]**	**ΔP%[b]**	**Name**	**ΔP%[a]**	**ΔP%[b]**
			Active training set		
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 -nitrotiazola	30.56	**-42.27**	Gynotabs	68.15	81.93
Glycobiarzol	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
			Active test set		
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 **13** and **14**, respectively: ΔP% = [P(Active) - P(Inactive)]x100.

Table 3. 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]**	**ΔP%[b]**	**Name**	**ΔP%[a]**	**ΔP%[b]**
			Inactive training set		
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	Bufeniode	-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			
			Inactive test set		
Citenazone	-19.35	-28.82	Metriponate	-99.97	-95.87
Methenamine	-99.66	-99.43	Ciclopramine	-97.18	-87.45
Penthrichloral	-90.52	**25.87**	Litracen	-99.31	-98.95
Calcium Sodium ferriclate	-100.00	-100.00	Trimetilsulfonium hidroxide	-99.48	-99.97
Ferroceron	-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 **13** and **14**, respectively: $\Delta P\% = [P(Active) - P(Inactive)] \times 100$.

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. Coefficient (C)	**Accuracy 'Q_{Total}' (%)**	**Sensitivity 'hit rate' (%) (%)**	**Specificity (%)**	**False positive rate 'false alarm rate' (%)**
Non-Stochastic Bond-Type Linear Indices (Eq. 13)					
Training set	0.78	89.01	87.50	87.50	9.80
Predicting set	0.71	87.50	80.00	80.00	9.09
Stochastic Bond-Type Linear Indices (Eq. 14)					
Training set	0.65	82.42	82.50	78.57	17.65
Predicting set	0.65	84.38	80.00	72.73	13.64

4.1.1. Internal Validation of the Descriminant Functions. Cross-Validation Methods

In recent years, exhaustive validation of mathematical models constitutes a main key of current QSAR theory [111]. 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 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 [117]. 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 [117,118].

In this report, 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 subsets. The statistical results are depicted in Table 5. The overall mean of the correct classification in training (test) set for this process for Eq. **13** and **14** was 88.90% (87.86%) and 82.20% (80.19%), correspondingly. The result of predictions on the 10% *full* cross-validation test evidenced the quality (robustness, stability and predictive power) of the obtained models.

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. 13 (Non-Stochastic Bond-based Linear Indices)						**Eq. 14** (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.

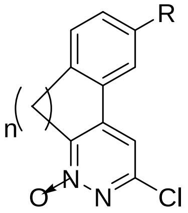
The former model validation process is important, if we take into consideration that the predictive ability of a QSAR model can be estimated using only an external test set of compounds that were not used for building the model [111,117,118].

3.1.1. Identification of Reported Chemicals Through a Simulated Ligand-Based Virtual Screening Experiment

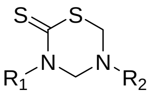
In this approach, instead of essaying a large number of chemicals in a series of biological tests we 'virtually essay' these compounds by evaluating their activities by the models developed to this effect; this process is known today as computational (*virtual* or in silico) screening [57,119,120]. Virtual screening techniques may be classified according to their particular modeling of molecular recognition and the type of algorithm used in database searching [56,57,119]. If the target (or at least its active site) 3D structure is known, one of the structure-based *virtual* screening methods can be applied. By contrast, ligand-based methods are founded on the principle of similarity, that is, similar compounds are assumed to produce similar effects. Nevertheless, the absence of a receptor 3D structure is the main reason for the application of ligand-based methods [55,121,122]. Due to these last fundamental facts, ligand-based *virtual* screening becomes our work philosophy.

In order to prove the possibilities of the ***TOMOCOMD-CARDD*** approach for the ligand-based *virtual* screening of antitrichomonal compounds, we have selected a series of 12 compounds whose activities against Tv have been already proved by several researchers [114,123,124]. They all were evaluated with models **13** and **14** as active/inactive ones. Its structures as well as the results of the classification are shown in Table 6.

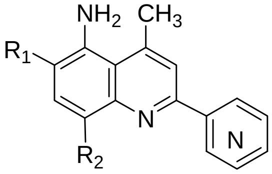
Table 6. Identification of chemicals extracted from literature as active or inactive toward the antitrichomonal activity by using LDA-based QSAR models in a simulated ligand-based *virtual* screening experiment.



1: n = CH_2
R = NH_2
2: n = CH_2-CH_2
R = NH_2
3: n = CH_2=CH_2
R = H



	R_1	R_2
4	furfuril	$CH(CH_2COOH)$-COOH
5	furfuril	$CH(CH_3)$-COOH
6	furfuril	$CH[(CH_2)_2SCH_3]$-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 Posición
10	CH_3	H	ß
11	CH_3	H	γ
12	H	$CH(CH_3)_2$	ß

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^{e} 10 μg/ml = (100)^{f} 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 numbers are shown at the top of this table. [b]Bibliographical references from where molecules together with its *in vitro* activities were taken. [c,d]Antitrichomonal activity predicted by Eq **13** and Eq **14**; ΔP% = [P(Active) - P(Inactive)]x100. [e]Percentage of reduction of *Tv* or cytocidal activity at the indicated doses at 24h. [f]Specific activity against *Tv* (in brackets) expressed as percentages of growth inhibition or cytostatic activity at 24h.

As can be seen, both models classify correctly most of the 12 selected compounds. The first model (Eq. **13**) classifies only two compounds incorrectly (both as false negative) thus achieving 83.33% of correct classification, while the second model (Eq. **14**) classifies three compounds incorrectly (all of them as false positive) for yielding 75.00% of correct classification. This result is a more 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 terms, 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 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.

Since tests above simulated the situation of *virtual* screening, the particular ability to select compounds from a never used dataset demonstrates the effectiveness of this approach for the computational high throughput *virtual* or/and *in silico* screening of trichomonacidal agents. No previous reports related 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.

4.2. 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 [120]. 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 [54]. 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 an important drug-discovery tool [125,126]

In order to test the potential of ***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 which were provided by one of our synthesis research teams from IQM, CSIC, Spain [127]. The structures of these compounds are depicted in Figure 2.

All these compounds were initially screened (evaluated) with the QSAR models **13** and **14** and then 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 coincidence between the theoretical predictions and the observed activity for both active and inactive compounds. Our trained LDA-based QSAR models (Eq. **13** and Eq.**14**) were capable of successfully classify 6 out of 7 compounds yielding (both) an accuracy of the 85.71%.

As for the *in vitro* experiments, should be highlighted that almost all compounds (VA7-34, VA7-37, VA7-38, VA7-68) exhibited pronounced cytocidal activities of 100% at the concentration of 100 μg/ml and at 24h (48h) but VA7-35 and VA7-70: 98,66% (99,40%), 99,83% (100%) respectively. It is remarkable that these compounds did not showed toxic activity in macrophages cultivations at this concentration. Also, 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 10μg/ml at 24h. 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.

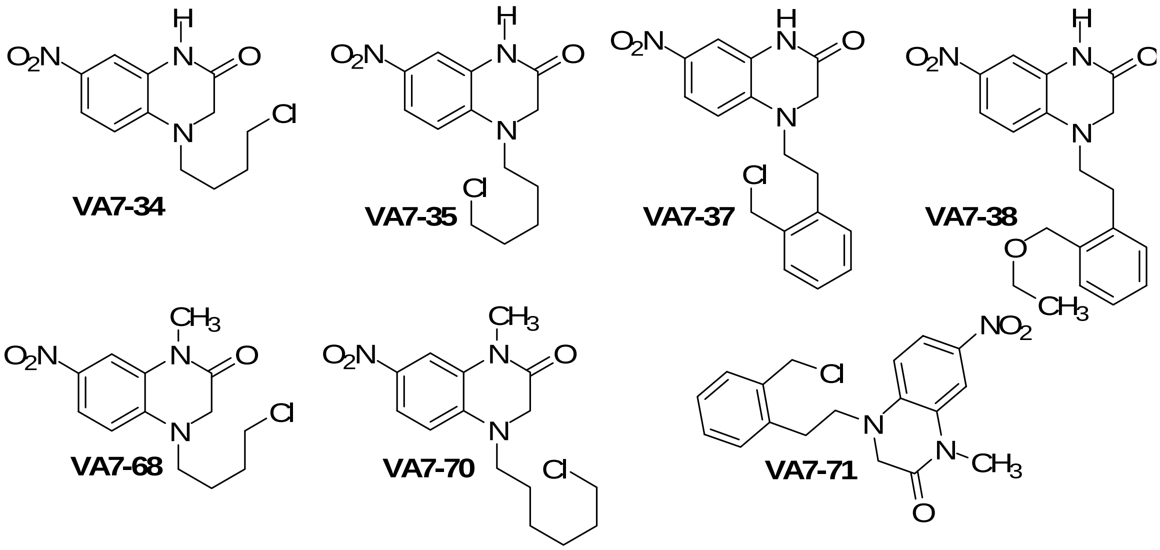


Figure 2. Structures of quinoxaline derivatives for novel trichomonacidals discovery by ligand-based *virtual* screening LDA-assisted models.

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

Compound[*]	**Theoretical results**					***in vitro* activity (μg/ml)**[f]					
	Class[a]	**ΔP%**[b]	**Class**[c]	**ΔP%**[d]	**Class**[e]	**$\%CA_{24h}$ [$\%C_{24h}$]**			**$\%CA_{48h}$ [$\%C_{48h}$]**		
						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

[*] The molecular structures of the compounds represented with codes are shown in Figure 2. [a,c] *In silico* classification obtained from models Eq. **13** and Eq. **14** using non-stochastic and stochastic bond-type linear indices, respectively. [b,d] Results for the classification of compounds obtained from models Eq. **13** and Eq. **14**, correspondingly: ΔP% = [P(Active) - P(Inactive)]x100. [e] Observed (experimental activity) classification against Tv. [f] Pharmacological activity of each tested compound, which was added to the cultures at doses of 100, 10 and 1μg/ml: $\%CA_{\#}$ = Cytostatic activity $_{(24 \text{ or } 48 \text{ hours})}$ and **[$\%C_{\#}$]** = Cytocidal activity $_{(24 \text{ or } 48 \text{ hours})}$. MTZ = Metronidazole (concentrations for MTZ were 2, 1 and 0.5 μg/ml, respectively).

These last 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. 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 ***TOMOCOMD-CARDD*** method can identify new trichomonacidal agents.

5. Concluding Remarks

The bioinformatic tools ***TOMOCOMD-CARDD*** & STATISTICA 6.0, and therefore, the underlying work philosophy, were successfully applied to the discovery of novel antitrichomonals. Combine features of bond-based linear stochastic and non-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. Finally, 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.

6. References and Notes

[1] Krieger, J. N.;Sex. Transm. Dis. 27 (2000) 241.

[2] Petrin, D.;Delgaty, K.;Bhatt, R.; Garber, G.; Clin. Microbiol. Rev. 11 (1998) 300.

[3] Cates, W. J.;Sex. Transm. Dis. 26(Suppl.) (1999) S2.

[4] World-Health-Organization, An overview of selected curable sexually transmitted diseases, World Health Organization, Geneva, Switzerland, 1995, p. 2.

[5] García, S.; Bruckner, D. A.; Diagnostic Medical Parasitology, American Society for Microbiology Washington (D.C), 1993.

[6] Rein, M. F.; Trichomoniasis, Goldsmith R. & Heyneman D., Santafé de Bogotá, 1995.

[7] Gram, I. T.;Macaluso, M.;Churchill, J.; Stalsberg, H.;Cancer Causes and Control 3 (1992) 231.

[8] Zhang, Z. F.; Begg, C. B.;Int. J. Epidemiol. 23 (1994) 682.

[9] Viikki, M.;Pukkala, E.;Nieminen, P.; Hakama, M.;Acta Oncol. 39 (2000) 71.

[10] Kharsany, B. M.;Hoosen, A. A.;Moodley, J.;Bagaratee, J.; Gouws, E.;Genitourin. Med. 69 (1993) 357.

[11] Cates, W.;Joesoef, M. R.; Goldman, M. B.;Am. J. Obstet. Gynecol. 169 (1993) 341.

[12] Grodstein, F.;Goldman, M. B.; Cramer, D. W.;Am. J. Epidemiol. 137 (1993) 577.

[13] Soper, D. E.;Bump, R. C.; Hurt, W. G.;Am. J. Obstet. Gynecol. 163 (1990) 1016.

[14] Cotch, M. F., Vaginal infections and prematurity study group. Carriage of Trichomonas vaginalis (Tv) is associated with adverse pregnancy outcome, American Society for Microbiology Washington D.C., 1990.

[15] Minkoff, H.;Grunebaum, A. N.;Schwarz, R. H.;Feldman, J.;Cummings, M.;Crombleholme, W.;Clark, L.;Pringle , G.; McCormack, W. M.;Am. J. Obstet. Gynecol. 150 (1984) 965.

[16] Fowler, K. B.; Pass, R. F.;J. Infect. Dis. 164 (1991) 259.

[17] Laga, M.;Manoka, A.;Kivuvu, M.;Malele, B.;Tuliza, M.;Nzila, N.;Goeman, J.;Behets, F.;Batter, V.;Alary, M.;Heyward, W. L.;Ryder, R. W.; Piot, P.;AIDS. 7 (1993) 95.

[18] Sorvillo, F.; Kerndt, P.;Lancet. 351 (1998) 213.

[19] Lossick, J. G.; Kent, H. L.;Am. J. Obstet. Gynecol. 165 (1991) 1217.

[20] Cosar, C.; Julou, L.;Ann. Inst. Pasteur 96 (1959) 238.

[21] Sucharit, P.;Uthaischant, A.;Chintana, T.;Suphadtanapongs, W.;Eamsobhana, P.; Prasomsitti, P.;S. E. Asian J. Trop. Med. Public Health 10 (1979) 556.

[22] Fugere, P.;Verschelden, G.; Caron, M.;Obstet. Gynecol. 62 (1983) 502.

[23] Videau, D.;Niel, G.;Siboulet, A.; Catalan, F.;Br. J. Vener. Dis. 54 (1978) 77.

[24] Pereyra, A. J.;Nelson, R. M.; Ludders, D. J.;Am. J. Obstet. Gynecol. 112 (1972) 963.

[25] Hayward, M. J.; Roy, R. B.;Br. J. Vener. Dis. 52 (1976) 63.

[26] Chaudhuri, P.; Drogendijk, A. C.;Eur. J. Obstet. Gynecol. Rep. Biol. 10 (1980) 325.

[27] Heine, P.; McGregor, J. A.;Clin. Obstet. Gynecol. 36 (1993) 137.

[28] Yarlett, N.;Yarlett, N. C.; Lloyd, D.;Biochem. Pharmacol. 35 (1986) 1703.

[29] Tocher, J. H.; Edwards, D. I.;Biochem. Pharmacol. 48 (1994) 1089.

[30] Nielsen, M. H.;Acta Pathol. Microbiol. Scand. Sect B. 84 (1976) 93.

[31] Morb. Mortal. Wkly. Rep. 42(RR-14) (1993) 70.

[32] Garcia-Leverde, A.; de Bonila, L.;Am. J. Trop. Med. Hyg. 24 (1975) 781.

[33] Powell, S. J.;Macleod, L.;Wilmot, A. J.; Elsdon-Dew, R.;Lancet. ii (1966) 1329.

[34] Scheider, J.; Bull. Soc. Pathol. Exot. 54 (1961) 84.

[35] Townson, S. M.;Boreham, P. F. L.;Upcroft, P.; Upcroft, J. A.;Acta Trop. 56 (1994) 173.

[36] Knight, R.;J Antimicrob Chemother. 6 (1980) 577.

[37] Arnold, M.;Ther Umsch. 23 (1966) 356.

[38] Aure, J. C.; Gjonnaess, H.;Acta Obstet. Gynecol. Scand. 48 (1969) 440.

[39] de Carneri, I., Pergamon Press, New York, 1966, p. 366.

[40] de Carneri, I.;Baldi, G. F.;Giannone, R.; Passalia, S.;Arch. Ostet. Ginecol. 68 (1963) 422.

[41] Diddle, A. W.;Am. J. Obstet. Gynecol. 98 (1967) 583.

[42] Giannone, T.;Minerva Ginecol. 24 (1972) 354.

[43] Kurnatowska, A.;Donné. Wiad Parazytol. 15 (1969) 399.

[44] Robinson, S. C.;Can. Med. Assoc. J. 86 (1962) 665.

[45] Korner, B.; Jensen, H. K.;Br J Vener Dis. 52 (1976) 404.

[46] McFadzean, J. A.;Pugh, L. M.;Squires, S. L.; Whelan, J. P.;Br. J. Vener. Dis. 45 (1969) 161.

[47] Roe, F. J.;J. Antimicrob. Chemother. 3 (1977) 205.

[48] Kane, P. O.;McFadzean, J. A.; Squires, S.;Br J Vener Dis. 37 (1961) 276.

[49] Nicol, C. S.;Evans, A. J.;McFadzean, J. A.; Squires, S. L.;Lancet ii (1966) 441.

[50] Meingassner, J. G.; Thurner, J.;Antimicrob. Agents Chemother. 15 (1979) 254.

[51] Sobel, J. D.;Nyirjesy, P.; Brown, W.;Clin. Inf. Dis. 33 (2001) 1341.

[52] Lumsden, W. H. R.;Robertson, D. H. H.;Heyworth, R.; Harrison, C.;Genitourin. Med. 64 (1988) 217.

[53] Narcisi, E. M.; Secor, W. E.;Antimicrob. Agents Chemother. 40 (1996) 1121.

[54] Estrada, E.; Peña, A.;Bioorg Med Chem 8 (2000) 2755.

[55] Estrada, E.;Uriarte, E.;Montero, A.;Teijeira, M.;Santana, L.; De Clercq, E.;J Med Chem 43 (2000) 1975.

[56] Scott, R. K.;Biosilico. 1 (2003) 14.

[57] Seifert, M. H. J.;Wolf, K.; Vitt, D.;Biosilico. 1 (2003) 143.

[58] 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@qf.uclv.edu.cu or ymarrero77@yahoo.es, 2002.

[59] Marrero-Ponce, Y.;Molecules 8 (2003) 687.

[60] Marrero Ponce, Y.;J Chem Inf Comput Sci 44 (2004) 2010.

[61] Marrero-Ponce, Y.;J Chem Inf Comput Sci 44 (2004) 2010.

[62] Marrero-Ponce, Y.;Bioorg. Med. Chem. 12 (2004) 6351.

[63] Marrero-Ponce, Y.;Castillo-Garit, J. A.;Torrens, F.;Romero-Zaldivar, V.; Castro, E.;Molecules 9 (2004) 1100.

[64] Marrero-Ponce, Y.;Díaz, H. G.;Romero, V.;Torrens, F.; Castro, E. A.;Bioorg. Med. Chem. 12 (2004) 5331.

[65] Marrero-Ponce, Y.;Cabrera, M. A.;Romero, V.;Ofori, E.; Montero, L. A.;Int. J. Mol. Sci. 4 (2003) 512.

[66] Marrero-Ponce, Y.;Cabrera, M. A.;Romero, V.;González, D. H.; Torrens, F.;J. Pharm. Pharmaceut. Sci. 7 (2004) 186.

[67] Marrero-Ponce, Y.;Cabrera, M. A.;Romero-Zaldivar, V.;Bermejo, M.;Siverio, D.; Torrens, F.;Internet Electrón. J. Mol. Des. 4 (2005) 124.

[68] Marrero-Ponce, Y.;Castillo-Garit, J. A.;Olazabal, E.;Serrano, H. S.;Morales, A.;Castanedo, N.;Ibarra-Velarde, F.;Huesca-Guillen, A.;Sanchez, A. M.;Torrens, F.; Castro, E. A.;Bioorg Med Chem 13 (2005) 1005.

[69] Marrero-Ponce, Y.;Castillo-Garit, J. A.;Olazabal, E.;Serrano, H. S.;Morales, A.;Castanedo, N.;Ibarra-Velarde, F.;Huesca-Guillen, A.;Jorge, E.;del Valle, A.;Torrens, F.; Castro, E. A.;J Comput Aided Mol Des 18 (2004) 615.

[70] Marrero-Ponce, Y.;Huesca-Guillen, A.; Ibarra-Velarde, F.;J. Mol. Struct. (Theochem) 717 (2005) 67.

[71] Marrero-Ponce, Y.;Montero-Torres, A.;Zaldivar, C. R.;Veitia, M. I.;Perez, M. M.; Sanchez, R. N.;Bioorg Med Chem 13 (2005) 1293.

[72] Marrero-Ponce, Y.;Medina-Marrero, R.;Torrens, F.;Martinez, Y.;Romero-Zaldivar, V.; Castro, E. A.;Bioorg Med Chem 13 (2005) 2881.

[73] Marrero-Ponce, Y.;Medina-Marrero, R.;Martinez, Y.;Torrens, F.;Romero-Zaldivar, V.; Castro, E. A.;J. Mol. Mod. 12 (2006) 255.

[74] Marrero-Ponce, Y.;Nodarse, D.;González, H. D.;Ramos de Armas, R.;Romero-Zaldivar, V.;Torrens, F.; Castro, E.;Int. J. Mol. Sci. 5 (2004) 276.

[75] Marrero-Ponce, Y.;Castillo-Garit, J. A.; Nodarse, D.;Bioorg. Med. Chem. 13 (2005) 3397

[76] Marrero-Ponce, Y.;Medina, R.;Castro, E. A.;de Armas, R.;González, H.;Romero, V.; Torrens, F.;Molecules 9 (2004) 1124.

[77] Marrero-Ponce, Y.;Medina-Marrero, R.;Castillo-Garit, J. A.;Romero-Zaldivar, V.;Torrens, F.; Castro, E. A.;Bioorg Med Chem 13 (2005) 3003.

[78] Marrero-Ponce, Y.; Torrens, F.; Alvarado, Y.; Rotondo, R.; J. Comp-Aided Mol. Des. Accepted for Publication (minor revision).

[79] Marrero-Ponce, Y.; Torrens, F. J. Mol. Graph. Mod. Submitted for Publication. (see also http://www.usc.es/congresos/ecsoc/9/ECSOC9.HTM)

[80] Rouvray, D. H.; in Balaban, A. T. (Editor), Chemical Applications of Graph Theory Academic Press, London, 1976, p. 180.

[81] Trinajstić, N.; Chemical Graph Theory, CRC Press, Boca Raton, FL, 1992.

[82] Estrada, E.;J Chem Inf Comput Sci 35 (1995) 31.

[83] Estrada, E.; Ramirez, A.;J Chem Inf Comput Sci 36 (1996) 837.

[84] Estrada, E.;J Chem Inf Comput Sci 36 (1996) 844.

[85] Estrada, E.;Guevara, N.; Gutman, I.;J Chem Inf Comput Sci 38 (1998) 428.

[86] Estrada, E.;J Chem Inf Comput Sci 39 (1999) 1042.

[87] Estrada, E.; Molina, E.;J Mol Graph Model 20 (2001) 54.

[88] Todeschini, R.; Consonni, V.; Handbook of Molecular Descriptors, Wiley-VCH, Germany, 2000.

[89] Edwards, C. H.; Penney, D. E.; Elementary Linear Algebra, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1988.

[90] Estrada, E.;Vilar, S.;Uriarte, E.; Gutierrez, Y.;J Chem Inf Comput Sci 42 (2002) 1194.

[91] Estrada, E.;Peña, A.; Garcia-Domenech, R.;J Comput Aided Mol Des 12 (1998) 583.

[92] Potapov, V. M.; Stereochemistry, Mir, Moscow, 1978.

[93] Wang, R.;Gao, Y.; Lai, L.;Perspect. Drug Dis. Des. 19 (2000) 47.

[94] Ertl, P.;Rohde, B.; Selzer, P.;J Med Chem 43 (2000) 3714.

[95] Ghose, A. K.; Crippen, G. M.;J Chem Inf Comput Sci 27 (1987) 21.

[96] Miller, K. J.;J. Am. Chem. Soc. 112 (1990) 8533.

[97] Gasteiger, J.; Marsili, M.;Tetrahedron Lett. 19 (1978) 3181.

[98] Pauling, L.; The Nature of Chemical Bond, Cornell University Press, Ithaca (New York), 1939.

[99] Browder, A.; Mathematical Analysis. An Introduction, Springer-Verlag, New York, 1996.

[100] Axler, S.; Linear Algebra Done Right, Springer-Verlag, New York, 1996.

[101] Daudel, R.;Lefebre, R.; Moser, C.; Quantum Chemistry: Methods and Applications, Wiley, New York, 1984.

[102] Klein, D. J.;Internet Electron. J. Mol. Des. 2 (2003) 814.

[103] Todeschini, R.; Gramatica, P.;Perspect. Drug Dis. Des. 9-11 (1998) 355–380.

[104] Consonni, V.;Todeschini, R.; Pavan, M.;J Chem Inf Comput Sci 42 (2002) 682.

[105] Kier, L. B.; Hall, L. H.; Molecular Connectivity in Structure–Activity Analysis, Research Studies Press, Letchworth, U. K, 1986.

[106] Negwer, M.; Organic-Chemical Drugs and their Synonyms, Akademie-Verlag, Berlin, 1987.

[107] The Merck Index, Chapman & Hall, 1999.

[108] van de Waterbeemd, H.; in van Waterbeemd, H. (Editor), Chemometric Methods in Molecular Design VCH Publishers, Weinheim, 1995, p. 265.

[109] STATISTICA (data analysis software system) vs 6.0, StatSoft Inc, 2001.

[110] Estrada, E.; Patlewicz, G.;Croat. Chim. Acta 77 (2004) 203.

[111] Wold, S.; Erikson, L.; in van de Waterbeemd, H. (Editor), VCH Publishers, New York, 1995, p. 309.

[112] Baldi, P.;Brunak, S.;Chauvin, Y.;Andersen, C. A.; Nielsen, H.;Bioinformatics 16 (2000) 412.

[113] 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. 1 (2005) 338.
[114] 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.;J. Arch. Pharm. 337 (2004) 127.
[115] Gálvez, J.;García, R.;Salabert, M. T.; Soler, R.;J Chem Inf Comput Sci 34 (1994) 520.
[116] Johnson, R. A.; Wichern, D. W.; Applied Multivariate Statistical Analysis, Prentice-Hall, New Jersey, 1988.
[117] Golbraikh, A.; Tropsha, A.;J Mol Graph Model 20 (2002) 269.
[118] Rose, K.;Hall, L. H.; Kier, L. B.;J Chem Inf Comput Sci 42 (2002) 651.
[119] Xu, J.; Hagler, A.;Molecules 7 (2002) 566.
[120] Watson, C.;Biosilico 1 (2003) 83.
[121] Mc Farland, J. W.; Gans, D. J.; in Waterbeemd, H. (Editor), Chemometric Methods in Molecular Design, VCH Publishers, New York, 1995, p. 295–307.
[122] Estrada, E.; Uriarte, E.;Curr Med Chem 8 (2001) 1573.
[123] Gavini, E.;Juliano, C.;Mulé, A.;Pirisino, G.;Murineddu, G.; Pinna, A.;Arch. Pharm. (Weinheim) 333 (2000) 341.
[124] 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. 49 (1999) 764.
[125] Lajiness, M. S.; in Rouvray, D. H. (Editor), Computational Chemical Graph Theory, Nova Science, New York, 1990, p. 299.
[126] Walters, W. P.;Stahl, M. T.; Murcko, M. A.;Drug Discov. Today 3 (1998) 160.
[127] Castro, S.;Chicharro, R.; Arán, V. J.;J.Chem.Soc., Perkin Trans. 1 (2002) 790.