2013

Predicting Dyspnea Inducers by Molecular Topology

María Gálvez-Llompart,
University of Valencia

Jorge Gálvez
University of Valencia

Ramón García-Domenech
University of Valencia

Lemont B. Kier
Virginia Commonwealth University, lbkier@vcu.edu

Hindawi Publishing Corporation
Journal of Chemistry
Volume 2013, Article ID 798508, 11 pages
http://dx.doi.org/10.1155/2013/798508

Research Article

Predicting Dyspnea Inducers by Molecular Topology

María Gálvez-Llompart,[1] Jorge Gálvez,[1] Ramón García-Domenech,[1] and Lemont B. Kier[2]

[1] *Molecular Connectivity and Drug Design Research Unit, Department of Physical Chemistry, Faculty of Pharmacy, University of Valencia Avd, V.A. Estellés, Burjassot, 46100 Valencia, Spain*
[2] *Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284-2030, USA*

Correspondence should be addressed to María Gálvez-Llompart; galloma@postal.uv.es

Received 22 June 2012; Accepted 25 July 2012

Academic Editor: M. Natália D. S. Cordeiro

QSAR based on molecular topology (MT) is an excellent methodology used in predicting physicochemical and biological properties of compounds. This approach is applied here for the development of a mathematical model capable to recognize drugs showing dyspnea as a side effect. Using linear discriminant analysis, it was found a four-variable regression equations enabling a predictive rate of about 81% and 73% in the training and test sets of compounds, respectively. These results demonstrate that QSAR-MT is an efficient tool to predict the appearance of dyspnea associated with drug consumption.

1. Introduction

1.1. Dyspnea as Side Effect. A side effect can be defined as an expected and known effect of a drug that is not the intended therapeutic outcome [1]. The US has been forced to remove from the market 75 drugs and combination drugs products since 1969 for safety reasons [2]. Although it is almost imperceptible if referred to all marketed drugs (less than 1%), safety-related regulatory actions (e.g. labeling changes, such as the addition of precautions, contraindications, or black box warnings) are much more common and less widespread. From 1969 to 2002, the Adverse Event Reporting System (AERS) of the Food and Drug Administration (FDA), as Nebeker et al. remarked, received approximately 2.3 million reports of adverse events on more than 6,000 drug products [2].

Adverse drug effects (ADEs), are a cause of injury or death to about 770,000 people each year, which may cost up to $5.6 million each year per hospital depending on the hospital size [3]. Therefore, anticipating the side effect profile for drugs is becoming more and more important in current drug discovery, development, and marketing. This strategy can lead to millions of dollars of savings in health care.

One side-effect having notable repercussions in national health care is dyspnea, defined by the American Thoracic Society as "a term used to characterize a subjective experience of breathing discomfort that is comprised of qualitatively distinct sensations that vary in intensity" [4]. Moreover, this symptom, characterized by the difficulty of getting sufficient air past the larynx, is associated with secondary physiological and behavioral responses [4].

Dyspnea is largely linked to people suffering from advanced cancer and cardiac, respiratory, and certain neurological diseases [5]. Patients mostly respond to breathlessness by adopting a sedentary lifestyle in order to relieve their symptomatology. This leads to social isolation, depression, fatigue, and dissatisfaction with life and significant emotional distress apart from skeletal muscle deconditioning [5, 6].

Dyspnea is one of the most distressing symptoms suffered by 30–75% terminal cancer patients [5]. The care of patients with dyspnea requires multidisciplinary resources such as palliative care, physiotherapy, respiratory medicine, and nursing [5].

There are different types of drugs causing dyspnea as a part of their side effects, such as chemotherapeutic agents (Capecitabine, Imatinib, Irinotecan…) [7], and other

therapeutic drugs (Amphotericin Bs, Nicotine, Dofetilide, Cyclosporine…) [7].

Breathlessness has marked significant impact on the quality of life of the patients but has also an economic impact in our society. An example of that can be seen in a study quanti fying the direct medical costs of dyspnea patients in 2008 or 2009 with a history of acute coronary syndrome (ACS) [8].

ACS patients included in the study were required to have six months of continuous medical enrollment prior to an emergency room (ER) visit. A total of 8433 emergency room (ER) visits for dyspnea were identified during 2008 to 2009 from these databases of approximately 74 million beneficiaries as reported by Bonafede et al. [8].

The average cost per dyspnea episode was $6958, associated with the ER visit, physician services, and diagnosis techniques such as electrocardiogram (71.3%) and chest radiograph (75.9%) as Bonafede et al. reported [8]. This is not all, more than one-fourth (25.8%) of dyspnea ER visits preceded an inpatient stay, with an average cost of $20 693 per patient. As the authors remarked, dyspnea is a significant event associated with high medical resource utilization and hospital costs.

1.2. Molecular Topology. Molecular topology (MT) is a discipline based in the topological description of molecules by using numerical invariants, called topological indices (TIs). These descriptors are able to characterize the most important features of molecular structure: molecular size, binding, cycles and branching.

Topological indices have the advantage of being true structural invariants, so that they are independent of the spatial and temporal position of the atoms in the molecule. However, TI's extensions that give account of three-dimensional structure have been also devised [9–11].

Many physicochemical and biological properties have been predicted by MT up to date, including groups of compounds showing considerable structural diversity [12–16]. Among the properties modeled stand several pharmacological activities such as anticonvulsant [17], antimalarial [18, 19], antimicrobial [20, 21], antifungal [22], antineoplastic [23], antihistaminic [24], bronchodilator [25], cytostatic [26], and anti-inflammatory compounds [27, 28] just to mention some examples.

No matter the application field, MT's strength lies in the reliable prediction of specific activities or properties of molecules. This way it is possible to select or design new compounds, particularly new drugs, thereby producing a high social and economic impact.

In a very recent paper [29], we demonstrated MT's effectiveness in predicting drug-induced anorexia. As a follow-up study, in the present work it is sought to raise one more step in the complex world of adverse effects to drugs, analyzing another side effect, namely, dyspnea.

2. Material and Methods

The application of the MT-model, involves the following steps (Figure 1).

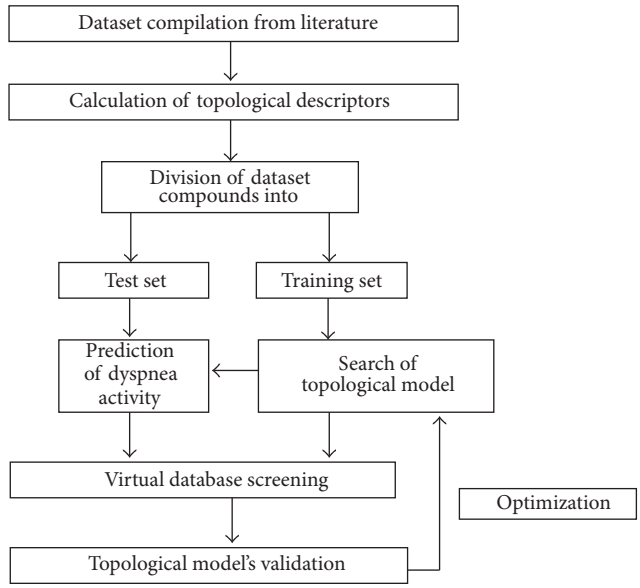

Figure 1: General scheme of the methodology followed for building up the dyspnea topological-mathematical model.

Step 1. Selection of dataset from the literature: the data were comprised of both drugs inducing (active) and not inducing (inactive) dyspnea.

Step 2. Calculation of topological descriptors: for that purpose, we used Dragon software, version 5.4. [30].

Step 3. Splitting of the dataset in two groups training set and test set: the criteria applied for such a splitting was based on the degree of dyspnea induction by the drugs. A level of 3% was used as threshold to distinguish dypsneagenic (above 3%) from non-dypsneagenic (below 3%) drugs.

Step 4. Application of linear discriminant analysis (LDA) to the training set.

Step 5. Validation of the LDA through an external test set.

Step 6. Application of the topological model to the identification of potential dyspnea-inducers: (not carried out here).

2.1. Selection of Dataset and Indices Calculation. A model to distinguish the dypsneagenic (active), from the nondypsneagenic (inactive) drugs, was built up with a dataset of 176 drugs. The drugs were from the database named SIDER [7] and from the internet site "drugs.com" [31].

The dataset was split into two, namely, a training and a test set. The first set was made up of 33 compounds causing dyspnea as side-effect with an incidence rate greater than 3% (active set) and 68 compounds showing an incidence rate less than or equal to 3% (inactive set). The second was made of 30 compounds causing dyspnea (test active set) and 45 compounds not showing dyspnea as a side effect (test inactive set).

The chemical structure of each drug was depicted by using ChemBioDraw Ultra version 12.0 (CambridgeSoft Corporation, Cambridge, MA, USA).

2.2. Molecular Descriptors. Each compound was charact erized by a set of 444 topological indices (TIs) obtained by Dragon software, version 5.4. Among the graph-theoretical descriptors calculated, the 2D autocorrelation indices demonstrated to be the most representative, and hence they were selected.

The chemical structures of the drugs studied were very heterogeneous. To guarantee that all groups were balanced, a study of molecular similarity using TIs was performed. Among the parameters calculated were the Tanimoto coefficient, TC, and the Euclidean distance, ED as follows:

Euclidean distance:

$$\mathrm{ED} = \sqrt{\sum \left(x_i - x_j\right)^2}. \tag{1}$$

Tanimoto coefficient:

$$\mathrm{TC} = \frac{\sum x_i x_j}{\sum x_i^2 + \sum x_j^2 - \sum x_i x_j}, \tag{2}$$

where x_i and x_j correspond to the topological indices of molecules i and j.

The topological similarity between two given compounds, i and j, will be higher when TC is closer to the unit and ED closer to zero [32].

Figure 2 shows the pairs of compounds with greater topological similarity (Adenosine, Lenalidomide and Clonazepam, Olanzapine for the active and the inactive training group, resp.) and with lesser similarity (Cyclosporine, Nicotine and Nitric oxide, and Ritonavir). The average values of the parameters TC and ED for all compounds studied (TC = 0.49; ED = 3.01×10^3) are analog to those obtained for the training set (TC = 0.46; ED = 5.3×10^3 and TC = 0.53; ED = 1.17×10^3 for the active and inactive group, resp.). The results for the test set were similar to those from the training set, what indicates that the groups were well arranged.

2.3. Modeling Techniques. Linear discriminant analysis (LDA) [33] is a statistical technique providing a classification based on the combination of variables that best predict the category or group to which a given object—a compound in our case-belongs. The compounds in the training set were allocated to active or inactive groups, according to their capability to produce dyspnea. Hence, the discriminant property was the capability of producing dyspnea as a side effect, and the independent variables were the TIs. The LDA final outcome is a discriminant function (DF), that is, an equation relating the activity, expressed in disjunctive terms in a Boolean way (1 = active; 2 = inactive), with the set of TIs. To get the LDA, the software Statistica version 9.0 (StatSoft, Inc., Tulsa) was used.

The discriminant capability was assessed as the percentage of correct classifications in each set of compounds. The classification criterion was the minimal Mahalanobis distance [34] (distance of each case to the mean of all the cases in a category), and the quality of the discriminant function was evaluated by using the Wilks parameter [35, 36], λ, which was obtained by multivariate analysis of variance that tests the equality of the group means for the variable in the discriminant model. The method used to select the descriptors were based on the Fisher-Snedecor parameter (F) [37], which determines the relative importance of candidate variables. The topological variables input is chosen in a stepwise manner; at each step, the variable that makes the largest contribution to the separation of the groups is entered into the discriminant equation (or the variable that makes the smallest contribution is removed).

The validation of the selected function was done using an external test set. Compounds that comprise the test set were randomly selected from approximately 20% of the data.

Another important parameter that usually provides a balanced evaluation of the model's prediction is the Matthews correlation coefficient (MCC) [38]. This coefficient is based on the fact that in any prediction process there can be four different possibilities to account for:

TP: true positive, a drug-induced dyspnea correctly classified or predicted.

FP: False positive, a drug not inducing dyspnea predicted as induced or when there was none to predict.

TN: True negative, a drug not inducing dyspnea correctly classified.

FN: False negative, a drug-induced dyspnea predicted as not inducing or when there was none to predict.

It is clear therefore, that any single number that represents the predictive power of the method must account for all of the possibilities listed above. One such factor is the Matthews correlation coefficient, which is given by:

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN})}}. \tag{3}$$

The Matthews correlation coefficient ranges from $-1 \leq \mathrm{MCC} \leq 1$. A value of MCC = 1 indicates the best possible prediction, in that every drug and all drug inducting dyspnea are correctly predicted. A MCC = −1 indicates the worst possible prediction (or anticorrelation), where no one dyspneagenic drug is detected, whereas all the nondyspneagenic drugs are erroneously predicted as dypsneagenic. Finally, a Matthews correlation coefficient of MCC = 0 indicates a random prediction.

Furthermore, a receiver operating characteristic curve (ROC) was drawn to evaluate the accuracy of the selected DF through the sensitivity (true positive fraction) and specificity (true negative fraction) for different DF thresholds. ROC curve is the representation of sensitivity versus (1-specificity) (false positive fraction). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. Accuracy is measured by the area under the ROC curve, AUC

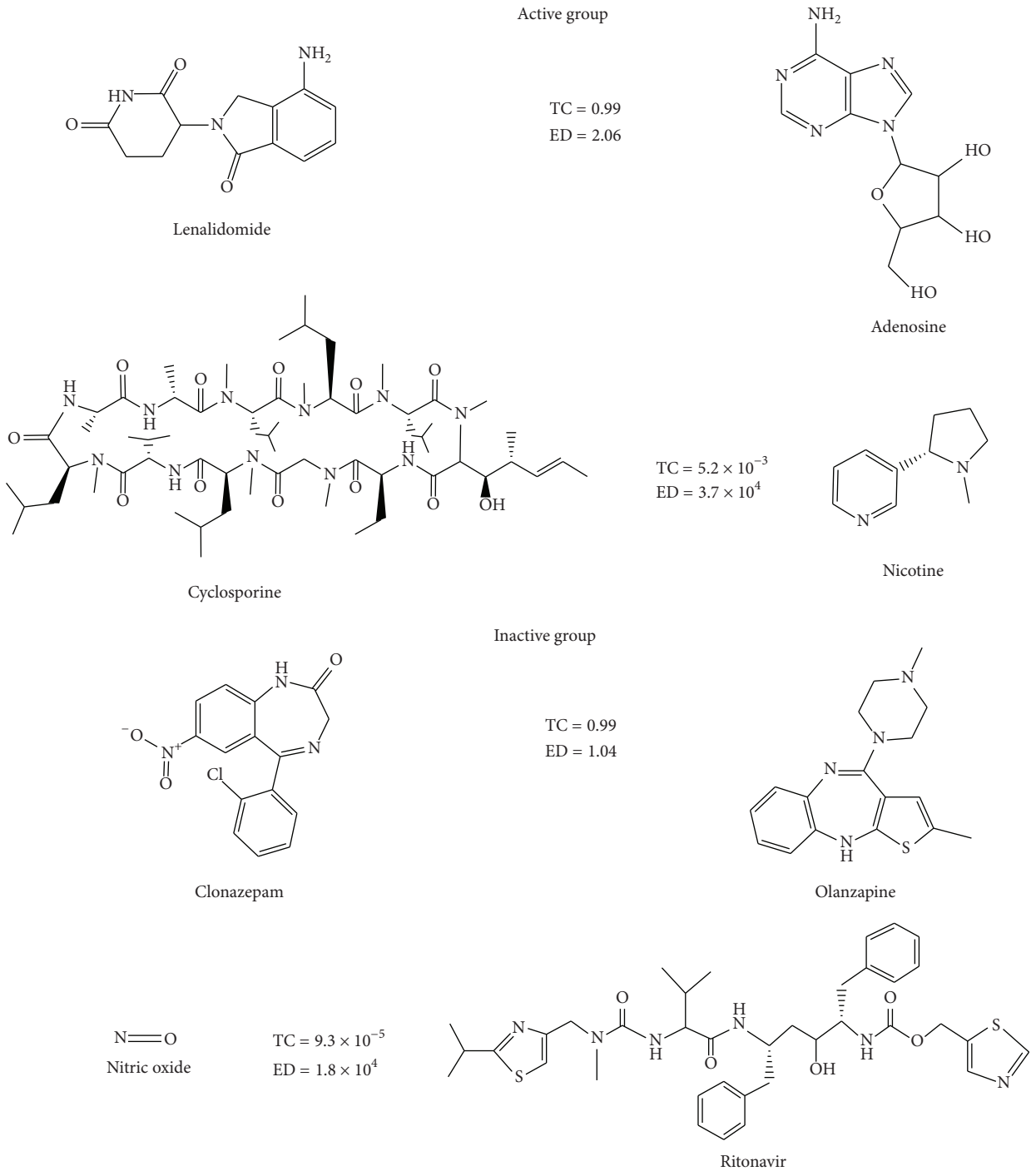

Figure 2: Results obtained from the molecular similarity study. TC: Tanimoto coefficient and ED: Euclidean distance.

[39, 40]. An AUC value = 1 represents a perfect test, whereas AUC of 0.5 represents a worthless test.

The model's predictive power was also assessed by using internal and external validation tests. A cross-validation, as an internal validation, was carried out by changing the roles of randomly 15–20% of active and inactive compounds from the training to the test set. Later on, it is checked if the model continues to show a good classification rate of the remaining compounds in the training and test set or not.

The equation obtained for the training set is used to predict the corresponding values of the test set.

2.4. Arrangement of the Pharmacological Distribution Diagram. The corresponding distribution diagram PDD [41]

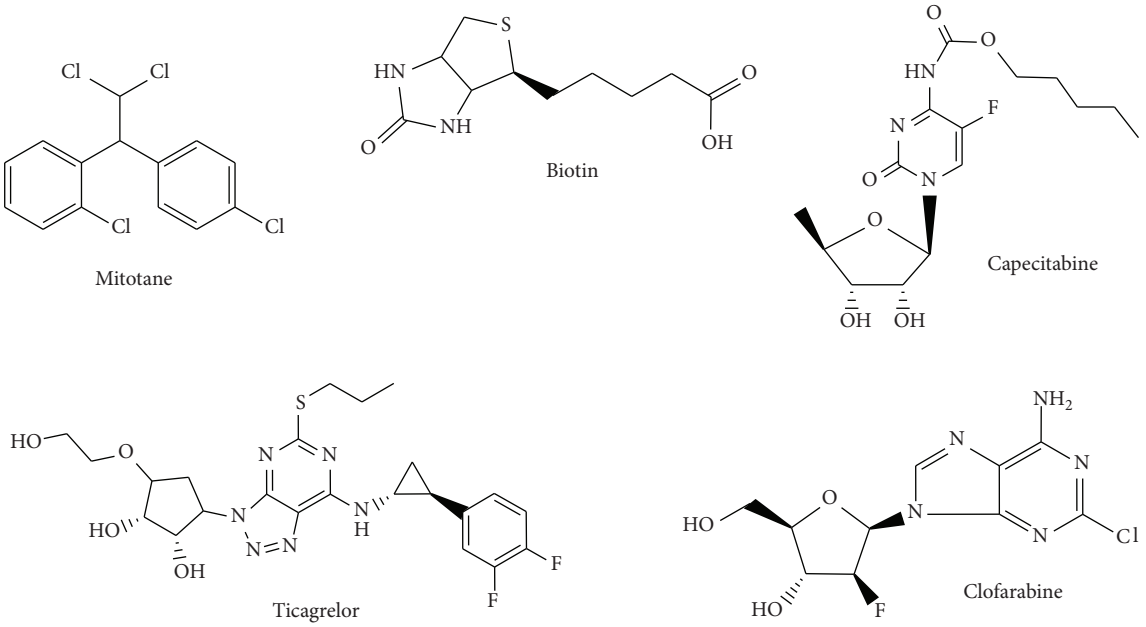

Figure 3: Example of molecules conforming training and test set having differences in polarizable heteroatoms and electronegativity and, hence, showing or not dyspnea as side effect.

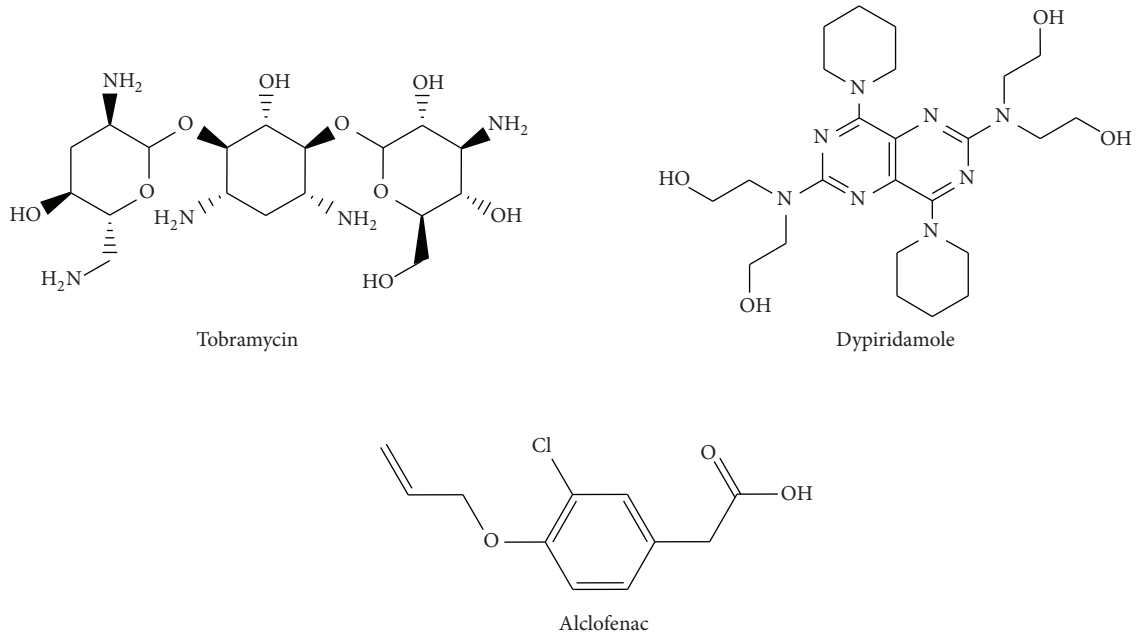

Figure 4: Example of molecules conforming training and test set, having differences in the presence of saturated alicyclic or aromatic rings and, hence, determining dyspnea side effect.

was drawn for dyspnea. Such diagrams are graphic representations providing a straightforward way to visualize the regions of minimum overlap between the active and inactive compounds, as well as the regions in which the probability of finding active compounds is maximum. Actually, a PDD is a frequency distribution diagram of dependent variables in which the ordinate represents the expectancy, E, (probability of activity) and the abscissa represents the DF values in the range. For an arbitrary range of values of a given function, an expectancy of activity can be defined as $E_a = a/(i + 1)c$, where "a" is the number of active compounds in the range divided by the total number of active compounds, and "i" is the number of inactive compounds in the interval divided by the total number of inactive compounds. The expectancy of inactivity is defined in a symmetrical way, as $E_i = i/(a + 1)$. Upon these diagrams, it is easy to visualize the intervals in

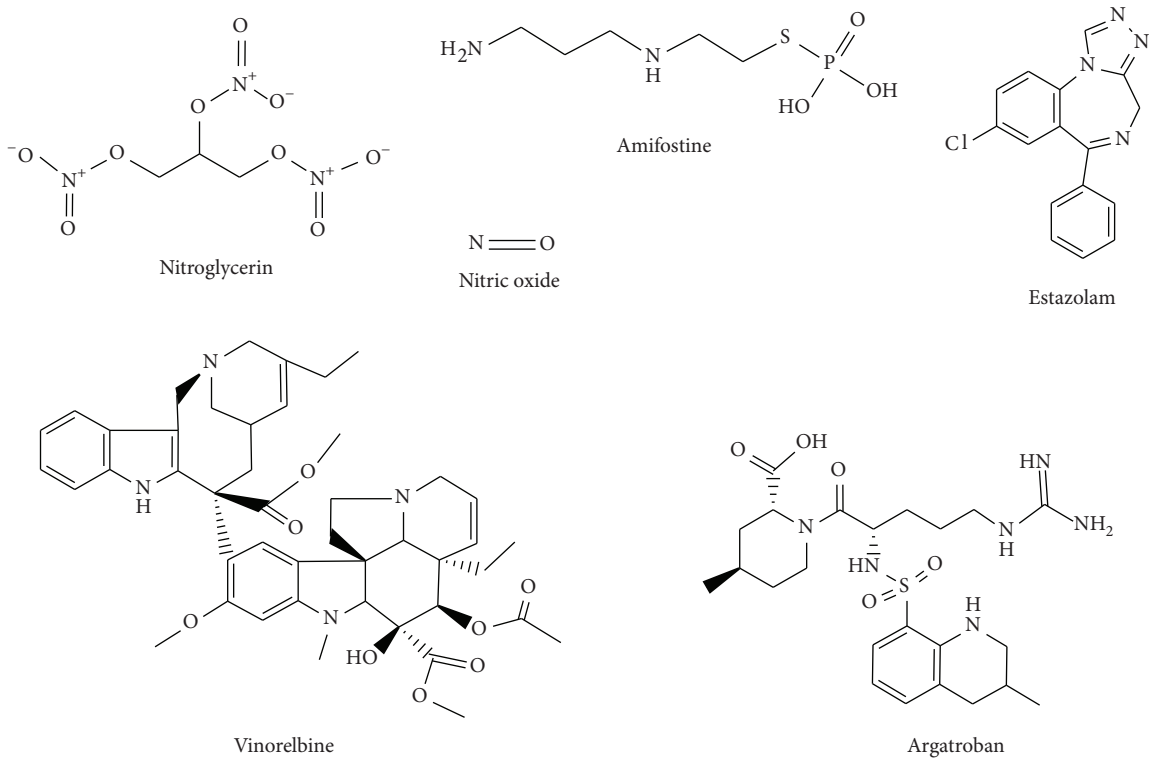

Figure 5: Example of differences in branching degree and, hence, in dyspnea side effect the degree of molecules conforming to the training and test sets.

which there is a maximum probability of finding new active compounds and a minimum probability of finding inactive compounds.

3. Results and Discussion

Appendix A in *Supplementary material* shows the values of the indices for every compound conforming the training and the test sets. The discriminant function selected as the best one was

$$\begin{aligned} \mathrm{DF} &= 2.892x\mathrm{ATS2e} - 7.159x\mathrm{MATS2v} \\ &\quad + 2.281x\mathrm{MATS8e} - 7.437x\mathrm{MATS1p} - 10.954, \quad (4) \\ N &= 101, \quad \lambda = 0.573, \quad F = 17.92, \quad P < 0.00001, \end{aligned}$$

where DF is discriminant function, ATS2e is Broto-Moreau autocorrelation of a topological structure at lag 2 weighted by atomic Sanderson electronegativities, MATS2v is Moran autocorrelation at lag 2 weighted by atomic van der Waals volumes, MATS8e is Moran autocorrelation at lag 8 weighted by atomic Sanderson electronegativities, MATS1p is Moran autocorrelation at lag 1 weighted by atomic polarizabilities, N is Number of data compounds, λ is Wilks' lambda, F is Fisher-Snedecor parameter, and P is Statistic significance.

From this equation, a given compound will be selected as active, that is, as a potential producer of dyspnea, if DF > 0; otherwise, it is classified as inactive. The classification matrix for DF is very significant for the training set: 82% of correct prediction for the active group, that is, 27 out of 33 compounds, and 55 out of 68 (81%) for the inactive group (see Table 1).

As pointed above, an additional way to check the model's predictive capability is through the Matthews' coefficient, which returns a range of values between −1 and +1. The higher its value the more reliable is the model. However, we calculated the Matthews' coefficient in a slightly different way, that is, just adding +1 to each threshold value, so that the outcome is expressed as % accuracy. In other words, 0 would mean no correlation at all, 1 represents 50%, and 2 accounts for 100% correlation. By doing so, the output was 83% of accuracy (modified Matthews' coefficient = 1.66).

An internal cross-validation (CV) analysis was also carried out to check the DF quality. Table 2 shows the values of λ (Wilks' lambda) and the classification matrix for compounds in the training and test sets. The values of λ for both sets are very close to each other. In fact, the values for the selected model and the average of the cross-validation model were $\lambda = 0.573$, and $\lambda = 0.609$, respectively. Moreover, the average percentage of correctness classification is also similar in both models (78% for the average CV and 77 %, for the selected model).

As can be seen in Table 2, the DF percentage of correct in the test set was 77%, what means that 23 out of 30 active compounds were correctly classified, while 31 out of 45 (69%) were correct in the inactive group (see Table 2). Table 3 outlines the results of the prediction for every compound of the external test.

Table 1: Classification of compounds showing and not showing dyspnea, obtained by linear discriminant analysis on the training set.

Compound	DF	Prob(*A*)	Class.	Compound	DF	Prob(*A*)	Class.
Active group							
Adenosine	2.259	0.905	A	Mycophenolic acid	−1.514	0.18	I
Amphotericin	3.866	0.979	A	Nelarabine	3.177	0.96	A
Anagrelide	5.07	0.994	A	Nicotine	0.575	0.64	A
Anastrozole	−2.044	0.115	I	Nilutamide	−2.939	0.05	I
Argatroban	2.23	0.903	A	Pentostatin	2.321	0.911	A
Atovaquone	−0.122	0.47	I	Pergolide	0.822	0.695	A
Busulfan	2.502	0.924	A	Pindolol	2.179	0.898	A
Capecitabine	2.054	0.886	A	Porfimersodium	1.69	0.844	A
Cidofovir	3.462	0.97	A	Propafenone	1.127	0.755	A
Cyclosporine	4.788	0.992	A	Ribavirin	2.928	0.949	A
Daunorubicin	2.559	0.928	A	Risperidone	1.021	0.735	A
Dofetilide	3.228	0.962	A	Sprycel	0.917	0.715	A
Hycamtin	1.16	0.762	A	Tiagabine	−2.125	0.107	I
Imatinib	1.837	0.863	A	Tobramycin	4.012	0.982	A
Irinotecan	1.287	0.784	A	Vinorelbine	3.295	0.964	A
Lenalidomide	−1.628	0.164	I	Zoledronic acid	2.792	0.942	A
Mitoxantrone	2.082	0.889	A				
Inactive group							
Acitretin	−2.385	0.084	I	Lamotrigine	−3.637	0.026	I
Allopurinol	−1.102	0.249	I	Latanoprost	−0.038	0.491	I
Alprazolam	−2.209	0.099	I	Leflunomide	−1.467	0.187	I
Altretamine	−1.527	0.178	I	Loratadine	−2.907	0.052	I
Amlodipine	−0.004	0.499	I	Lorazepam	−2.892	0.053	I
Bexarotene	−2.322	0.089	I	Mesalamine	−3.105	0.043	I
Buspirone	0.36	0.589	A	Misoprostol	−0.17	0.458	I
Carbamazepine	−4.109	0.016	I	Nabilone	1.131	0.756	A
Carmustine	−4.441	0.012	I	Nabumetone	−1.483	0.185	I
Celecoxib	−1.132	0.244	I	Naproxen	−4.037	0.017	I
Chantix	0.949	0.721	A	Nicardipine	0.558	0.636	A
Clonazepam	−1.456	0.189	I	Nitric oxide	−3.517	0.029	I
Clozapine	0.327	0.581	A	Nitroglycerin	−3.265	0.037	I
Cysteamine	−6.532	0.001	I	Olanzapine	−0.006	0.499	I
Deferasirox	0.062	0.516	A	Oxcarbazepine	−2.257	0.095	I
Desloratadine	−2.321	0.089	I	Oxybutynin	0.131	0.533	A
Desvenlafaxine	−1.86	0.135	I	Phentermine	−1.992	0.12	I
Diclofenac	−0.314	0.422	I	Propofol	−2.077	0.111	I
Didanosine	−0.853	0.299	I	Riluzole	−5.256	0.005	I
Entacapone	0.009	0.502	A	Ritonavir	−0.438	0.392	I
Estazolam	−3.203	0.039	I	Selegiline	−0.388	0.404	I
Femara	−2.381	0.085	I	Tacrine	−0.196	0.451	I
Fenofibrate	−0.079	0.48	I	Telbivudine	−0.065	0.484	I
Flurbiprofen	−3.028	0.046	I	Temazepam	−2.125	0.107	I
Fluvoxamine	−2.285	0.092	I	Temozolomide	0.312	0.577	A
Gabapentin	−4.73	0.009	I	Terbinafine	0.013	0.503	A
Gadoversetamide	0.613	0.649	A	Thalidomide	−1.799	0.142	I
Guanfacine	−6.447	0.002	I	Theophylline anhydrous	0.306	0.576	A
Indomethacin	−1.603	0.168	I	Valsartan	−3.183	0.04	I
Iopromide	−0.36	0.411	I	Venlafaxine	−0.794	0.311	I
Isocarboxazid	−2.376	0.085	I	Ziagen	−0.435	0.393	I
Isradipine	−1.116	0.247	I	Zidovudine	−1.208	0.23	I

Table 1: Continued.

Compound	DF	Prob(*A*)	Class.	Compound	DF	Prob(*A*)	Class.
Ketoprofen	−3.277	0.036	I	Zolmitriptan	−2.13	0.106	I
Lamivudine	−3.59	0.027	I	Zolpidem	0.006	0.502	A

Prob(*A*): probability of activity. DF: discriminant function. Class.: clasiffication as active or inactive.

Table 2: Classification matrix obtained with the selected discriminant function (DF) and internal cross-validation analysis.

DF	λ (Wilks)	Training set		Test set		Total
		Active (%)	Inactive (%)	Active (%)	Inactive (%)	%
Model selected	0.573	81.8	80.8	76.6	68.8	77.0
CV1	0.596	75.7	77.9	86.6	69.6	77.4
CV2	0.576	84.8	86.8	76.6	76.1	81.1
CV3	0.657	75.8	77.9	70.0	76.1	75.0
CV4	0.598	84.8	83.8	76.6	73.9	79.8
CV5	0.617	81.8	76.5	83.3	54.3	74.0
CV average	0.609	80.6	80.6	78.6	70	77.5

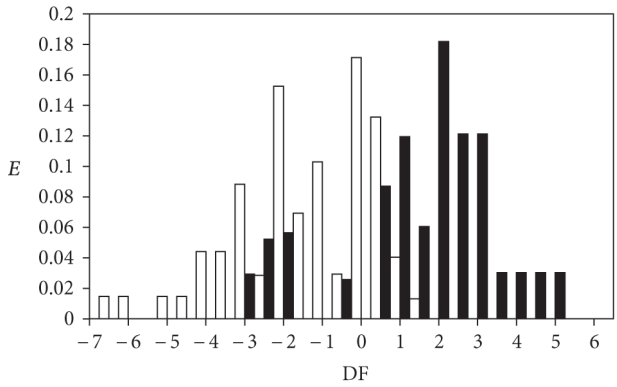

Figure 6: Pharmacological distribution diagram for compounds causing dyspnea as a side effect. The plot represents expectancy (*E*) versus DF function (the black and white bars are the active and inactive compounds, resp.).

Regarding the interpretation of results, it is noteworthy the presence of the autocorrelation indices in the DF (4). Autocorrelation indices enable the representation of the molecule as a vector that can be compared with the vectors of other molecules at a given position or lag (lag = 1,2,3…). This singular molecular frame has a further desirable property; it is independent of the orientation of the molecule; in other words, it is a real topological index or graph invariant. This approach has also the advantage that there are many different ways to define a particular alignment of molecules (structure, shape, and function) although, contrary to other topological indices, they do not enable the interconnection between the vector values and the original molecules; that is, they are not bijective. In addition, autocorrelation indices allow weighting the graph vertices by different parameters such as electronegativity and atomic mass.

In our case, the indexes at (4), namely, ATS2e, MATS2v, MATS8e, and MATS1p, are weighted by polarizabilities, van der Waals volumes, and electronegativity. Moreover, the first and third indices (ATS2e and MATS8e) contribute positively to dyspneagenesis, while the second and forth (MATS2v and MATS1p) contribute negatively. This indicates that, roughly speaking, molecules containing big and highly polarizable heteroatoms (such as S and Cl) as, for example, Mitotane and Biotin (see Figure 3) would be in general less dypsneagenic, whereas molecules containing highly electronegative atoms, such as fluorine, would be more dypsneagenic, as, for instance, Capecitabine, Clofarabine, and Ticagrelor.

It can be noticed that molecules showing dyspnea as side effect usually have saturated allicyclic rings (such as Piperazine or Morpholine), as, for example, Irinotecan, Tobramycin, Dipyridamole, and Sildenafil (see Figure 4), which are isolated from each other or separated from aromatic rings if exist. On the contrary, the inactive molecules exhibit in general many aromatic rings and show high conjugation, as, for instance, Alclofenac and Mitotane (see Figures 3 and 4).

A very basic structural factor, the number of branches (vertices with valence 3 or 4 in the simple graph) play also a significant role. There seem to be a threshold of about 9 branches separating the active from the inactive compounds. Indeed, the low branching molecules (number of branches below 9) are, in more than 80% of cases, inactive, as, for instance, Amifostine, Estazolam, Nitric Oxide, and Nitroglycerin… (see Figure 5). On the contrary, highly branched compounds (number of branches above 9) may be either active or not, as for example, Tobramycin, Vinorelbine, and Atovaquone (see Figures 4 and 5). In other words, a branching threshold above 9 is a necessary but not a sufficient condition.

In summary, we find the following features regarding the active-inactive compounds, related to TIs in the model and visual analysis of the structures in the data set.

Table 3: Results of prediction of dyspneagenicity through the discriminant function DF for the test set.

Compound	DF	Prob(*A*)	Class.	Compound	DF	Prob(*A*)	Class.
Active group							
Acamprosate	0.778	0.685	A	Dirithromycin	5.284	0.995	A
Acyclovir	−0.704	0.331	I	Dobutamine	0.347	0.586	A
Alimta	−1.216	0.229	I	Eszopiclone	1.195	0.768	A
Ambrisentan	0.207	0.552	A	Flecainide	1.192	0.767	A
Amiodarone	0.76	0.682	A	Fluoxetine	−3.256	0.037	I
Atenolol	0.427	0.605	A	Ipratropiumbromide	0.519	0.627	A
ATP	2.634	0.933	A	Labetalol	−0.665	0.34	I
Bortezomib	2.13	0.894	A	Methamphetamine	−2.018	0.117	I
Clarithromycin	4.786	0.992	A	Mirtazapine	1.098	0.75	A
Clofarabine	2.441	0.92	A	Nimodipine	0.44	0.608	A
Clopidogrel	−2.056	0.114	I	Oxycodone	0.447	0.61	A
Cyanocobalamin	5.267	0.995	A	Propranolol	0.477	0.617	A
Cytarabine	0.946	0.72	A	Sildenafil	3.672	0.975	A
Dexrazoxane	−0.056	0.486	I	Ticagrelor	2.593	0.93	A
Dipyridamole	3.748	0.977	A	Trovafloxacin	1.574	0.828	A
Inactive group							
Acecainide	1.21	0.77	A	Cibenzoline	−1.924	0.127	I
Aceclofenac	−0.952	0.279	I	Cocaine	−0.577	0.36	I
Acipimox	−0.692	0.334	I	Disulfiram	0.386	0.595	A
Acivicin	−0.901	0.289	I	Ergocalciferol	0.107	0.527	A
Acrivastine	−0.907	0.288	I	Flolan	−0.493	0.379	I
Adefovir	3.246	0.963	A	Folic acid	0.54	0.632	A
Ademetionine	1.008	0.733	A	Gonadorelin	4.554	0.99	A
Adiphenine	−1.527	0.179	I	Lisinopril	−1.382	0.201	I
Ajmaline	0.635	0.654	A	Lisuride	−0.846	0.3	I
Albendazole	−5.145	0.006	I	Loxoprofen	−3.178	0.04	I
Alclofenac	−3.074	0.044	I	Metaproterenol	−1.194	0.233	I
Alfentanil	0.078	0.52	A	Methyl phenidate	−1.979	0.121	I
Amifostine	−3.844	0.021	I	Mitotane	−7.45	0.001	I
Aminophenazone	0.998	0.731	A	Pantothenic	−3.674	0.025	I
Aminorex	−4.244	0.014	I	Penicillamine	−4.915	0.007	I
ASA	−2.311	0.09	I	Procainamide	−0.23	0.443	I
Ascorbic acid	1.378	0.799	A	Pyridoxine	−1.065	0.256	I
Balsalazide	−1.582	0.171	I	Repaglinide	1.438	0.808	A
Bezafibrate	−0.573	0.361	I	Thiamine	−1.881	0.132	I
Biotin	−5.031	0.006	I	Tretinoin	−3.889	0.02	I
Bucillamine	−3.581	0.027	I	Troglitazone	2.314	0.91	A
Carnidazole	0.668	0.661	A	Valproic acid	−6.145	0.002	I
Cefpodoxime	−0.139	0.466	I				

Prob(*A*): probability of activity. DF: discriminant function. Class.: classification as active or inactive.

(a) Above 80% of compounds not showing dypsnea as a side effect show less than 9 branches in their simple graphs.

(b) The dypsneagenic molecules usually have saturated heterocyclic rings (such as piperazine or morpholine) while compounds not exhibiting dyspnea in general exhibit many aromatic rings with a high level of conjugation.

(c) The presence of highly polarizable (MATS1p) and large Van der Waals volume atoms (such as S or Cl) (MATS2v) diminish the dyspnea effect.

One of the most interesting consequences of the QSAR analysis we have just described is that it can be applied as a filter to avoid selecting dypsnea-inducing drugs. A good way to proceed is to use the pharmacological distribution diagrams (PDDs). Figure 6 shows the PDD obtained for our

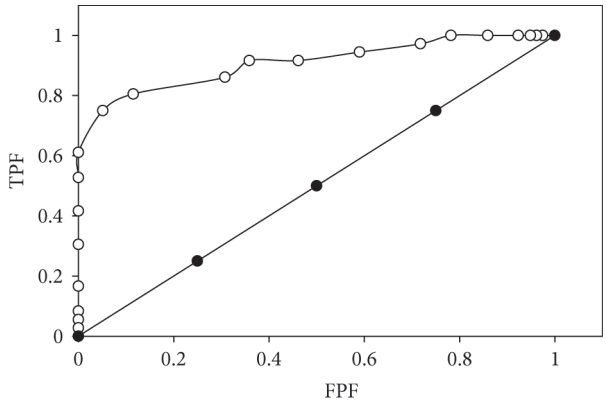

Figure 7: Receiver operating characteristic curve (ROC) for (4). Training set is represented by white points and random classifier by black points. TPF (true positive fraction)= sensitivity and FPF (false positive fraction) = 1-specificity, for different thresholds of class function (between −1.0 and +1.0).

data set. As can be seen from the diagram, those compounds with DF values between 0.5 and 5.5 are clearly dyspnea inducers, those between 0.5 and −7 are generally dyspnea not inducers. Finally, are those compounds located between −3 and −2.5 are uncertain (not classified). Of course, if we are trying to identify possible compounds showing dyspnea as side-effect, we must search in the DF ranges between −3 to −2.5 and 0.5 to 5.5. The drugs outside these ranges are not dyspnea inducers. As it is arranged, the model enables not missing any potential dypsneagenic; that is, in this case sensitivity was preferred over specificity, just to prevent the risk of dyspnea.

Receiver operating characteristic curve for the training set is shown in Figure 7 for (4). The area under the curve (AUC) is 0.9046, which accounts for the high accuracy of model.

4. Conclusions

The results outlined in this work clearly point toward the efficacy of molecular topology in the prediction of a very important drug side effect: the induction of dyspnea. As far as authors' knowledge, this is the second time that molecular topology has been applied to the identification of drug side effects [29]. Furthermore, it is the first time that dyspnea, as a side effect, has been so accurately predicted by a mathematical model, what opens the pathway to the design of drugs free from this undesirable side effect.

Acknowledgment

The authors thank the Ministerio de Economía y Competitividad, Spain (project no. SAF2009-13059-C03-02) for support of this work. M. Gálvez-Llompart acknowledges the "V Segles" Fellowship provided by the University of Valencia to carry out this study.

References

[1] D. K. Wysowski and L. Swartz, "Adverse drug event surveillance and drug withdrawals in the United States, 1969–2002: the importance of reporting suspected reactions," *Archives of Internal Medicine*, vol. 165, no. 12, pp. 1363–1369, 2005.

[2] J. R. Nebeker, P. Barach, and M. H. Samore, "Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting," *Annals of Internal Medicine*, vol. 140, no. 10, pp. 795–801, 2004.

[3] Agency for Healthcare Research Quality, Ed., *Reducing and Preventing Adverse Drug Events to Decrease Hospital Costs. Research in Action, Issue 1*, AHRQ, Rockville, Md, USA, 2001, http://www.ahrq.gov/qual/aderia/aderia.htm.

[4] N. Raghavan, K. Webb, N. Amornputtisathaporn, and D. E. O'Donnell, "Recent advances in pharmacotherapy for dyspnea in COPD," *Current Opinion in Pharmacology*, vol. 11, no. 3, pp. 204–210, 2011.

[5] S. Booth and D. Dudgeon, *Dyspnoea in Advanced Disease: A Guide to Clinical Management*, Oxford University Press, New York, NY, USA, 2006.

[6] A. W. Sheel, G. E. Foster, and L. M. Romer, "Exercise and its impact on dyspnea," *Current Opinion in Pharmacology*, vol. 11, no. 3, pp. 195–203, 2011.

[7] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs," *Molecular Systems Biology*, vol. 6, article 343, 2010.

[8] M. Bonafede, Y. Jing, J. Gdovin Bergeson et al., "Impact of dyspnea on medical utilization and affiliated costs in patients with acute coronary syndrome," *Hospital Practice*, vol. 39, no. 3, pp. 16–22, 2011.

[9] J. V. de Julián-Ortiz, C. de Gregorio Alapont, I. Ríos-Santamarina, R. García-Doménech, and J. Gálvez, "Prediction of properties of chiral compounds by molecular topology," *Journal of Molecular Graphics and Modelling*, vol. 16, no. 1, pp. 14–18, 1998.

[10] A. Golbraikh, D. Bonchev, and A. Tropsha, "Novel ZE-isomerism descriptors derived from molecular topology and their application to QSAR analysis," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 4, pp. 769–787, 2002.

[11] R. García-Domenech, J. Gálvez, J. V. de Julián-Ortiz, and L. Pogliani, "Some new trends in chemical graph theory," *Chemical Reviews*, vol. 108, no. 3, pp. 1127–1169, 2008.

[12] L. B. . Kier and L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, NY, USA, 1976.

[13] L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, John Wiley and Sons, Chichester, UK, 1986.

[14] L. B. Kier and L. H. Hall K Hall, *Molecular Structure Description: The Electrotopological State*, Academic Press, San Diego, Calif, USA, 1999.

[15] J. Galvez, "Prediction of molecular volume and surface of alkanes by molecular topology," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 4, pp. 1231–1239, 2003.

[16] J. Pla-Franco, M. Gálvez-Llompart, J. Gálvez, and R. García-Domenech, "Application of molecular topology for the prediction of reaction yields and anti-inflammatory activity of heterocyclic amidine derivatives," *International Journal of Molecular Sciences*, vol. 12, no. 2, pp. 1281–1292, 2011.

[17] L. Bruno-Blanch, J. Gálvez, and R. García-Domenech, "Topological virtual screening: a way to find new anticonvulsant drugs from chemical diversity," *Bioorganic and Medicinal Chemistry Letters*, vol. 13, no. 16, pp. 2749–2754, 2003.

[18] N. Mahmoudi, J. V. de Julián-Ortiz, L. Ciceron et al., "Identification of new antimalarial drugs by linear discriminant analysis and topological virtual screening," *Journal of Antimicrobial Chemotherapy*, vol. 57, no. 3, pp. 489–497, 2006.

[19] N. Mahmoudi, R. Garcia-Domenech, J. Galvez et al., "New active drugs against liver stages of Plasmodium predicted by molecular topology," *Antimicrobial Agents and Chemotherapy*, vol. 52, no. 4, pp. 1215–1220, 2008.

[20] C. de Gregorio Alapont, R. García-Domenech, J. Gálvez, M. J. Ros, S. Wolski, and M. D. García, "Molecular topology: a useful tool for the search of new antibacterials," *Bioorganic and Medicinal Chemistry Letters*, vol. 10, no. 17, pp. 2033–2036, 2000.

[21] R. K. Mishra, R. Garcia-Domenech, and J. Galvez, "Getting discriminant functions of antibacterial activity from physicochemical and topological parameters," *Journal of Chemical Information and Computer Sciences*, vol. 41, no. 2, pp. 387–393, 2001.

[22] L. Pastor, R. García-Domenech, J. Gálvez, S. Wolski, and M. D. García, "New antifungals selected by molecular topology," *Bioorganic and Medicinal Chemistry Letters*, vol. 8, no. 18, pp. 2577–2582, 1998.

[23] P. Jasinski, B. Welsh, J. Galvez et al., "A novel quinoline, MT477: suppresses cell signaling through Ras molecular pathway, inhibits PKC activity, and demonstrates in vivo anti-tumor activity against human carcinoma cell lines," *Investigational New Drugs*, vol. 26, no. 3, pp. 223–232, 2008.

[24] M. J. Duart, R. García-Domenech, J. Gálvez, P. A. Alemán, R. V. Martín-Algarra, and G. M. Antón-Fos, "Application of a mathematical topological pattern of antihistaminic activity for the selection of new drug candidates and pharmacology assays," *Journal of Medicinal Chemistry*, vol. 49, no. 12, pp. 3667–3673, 2006.

[25] I. Ríos-Santamarina, R. García-Domenech, J. Gálvez, J. Cortijo, P. Santamaria, and E. Morcillo, "New bronchodilators selected by molecular topology," *Bioorganic and Medicinal Chemistry Letters*, vol. 8, no. 5, pp. 477–482, 1998.

[26] J. Gálvez, R. García-Domenech, M. J. Gómez-Lechón, and J. V. Castell, "Use of molecular topology in the selection of new cytostatic drugs," *Journal of Molecular Structure*, vol. 504, pp. 241–248, 2000.

[27] M. Galvez-Llompart, R. M. Giner, M. C. Recio, S. Candeletti, and R. Garcia-Domenech, "Application of molecular topology to the search of novel NSAIDs: experimental validation of activity," *Letters in Drug Design and Discovery*, vol. 7, no. 6, pp. 438–445, 2010.

[28] M. Galvez-Llompart, R. Zanni, and R. García-Domenech, "Modeling natural anti-inflammatory compounds by molecular topology," *International Journal of Molecular Sciences*, vol. 12, no. 12, pp. 9481–9503, 2011.

[29] M. Gálvez-Llompart, J. Gálvez, R. García-Domenech, and L. B. Kier, "Modeling drug-induced anorexia by molecular topology," *Journal of Chemical Information and Modeling*, vol. 52, no. 5, pp. 1337–1344, 2012.

[30] Talete srl, Dragon for windows (Software for molecular Descriptor Calculations), version 5.4, 2006, http://www.talete.mi.it/.

[31] Drugs side effect, http://www.drugs.com/sfx.

[32] P. Willett and V. Winterman, "A comparison of some measures for the determination of inter-molecular structural similarity measures of inter-molecular structural similarity," *Quantitative Structure-Activity Relationships*, vol. 5, no. 1, pp. 18–25, 1986.

[33] T. Hastie and R. Tibshirani, "Discriminant analysis by Gaussian mixtures," *Journal of the Royal Statistical Society*, vol. 58, pp. 155–176, 1996.

[34] R. de Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The Mahalanobis distance," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 1, pp. 1–18, 2000.

[35] B. Everitt, *Applied Multivariate Data Analysis*, Arnold, London, UK, 2001.

[36] W. T. . Eadie, D. Drijard, F. E. . James, M. Roos, and B. Sadoulet, *Statistical Methods in Experimental Physics*, North-Holland, Amsterdam, The Netherlands, 1971.

[37] G. M. . Furnival, "All possible regressions with less computation," *Technometrics*, vol. 13, no. 2, pp. 403–408, 1971.

[38] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, 1975.

[39] D. Katzman McClish, "Analyzing a portion of the ROC curve," *Medical Decision Making*, vol. 9, no. 3, pp. 190–195, 1989.

[40] M. J. Pencina, R. B. D'Agostino, R. B. D'Agostino, and R. S. Vasan, "Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond," *Statistics in Medicine*, vol. 27, no. 2, pp. 157–172, 2008.

[41] J. Gálvez, R. García-Domenech, C. de Gregorio Alapont, J. V. de Julián-Ortiz, and L. Popa, "Pharmacological distribution diagrams: a tool for de novo drug design," *Journal of Molecular Graphics*, vol. 14, no. 5, pp. 272–276, 1997.