Biochimica et Biophysica Acta 1848 (2015) 2448–2459

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbamem

Molecular properties of steroids involved in their effects on the biophysical state of membranes

Jorge J. Wenz

Instituto de Investigaciones Bioquímicas de Bahía Blanca, Camino "La Carrindanga" Km 7, B8000FWB Bahía Blanca, Argentina

ARTICLE INFO

Article history:
Received 22 January 2015
Received in revised form 28 July 2015
Accepted 29 July 2015
Available online 31 July 2015

Keywords:
Steroid molecular property
Membrane biophysical property
Multivariate analysis

ABSTRACT

The activity of steroids on membranes was studied in relation to their ordering, rigidifying, condensing and/or raft promoting ability. The structures of 82 steroids were modeled by a semi-empirical procedure (AM1) and 245 molecular descriptors were next computed on the optimized energy conformations. Principal component analysis, mean contrasting and logistic regression were used to correlate the molecular properties with 212 cases of documented activities. It was possible to group steroids based on their properties and activities, indicating that steroids having similar molecular properties have similar activities on membranes. Steroids having high values of area, partition coefficient, volume, number of rotatable bonds, molar refractivity, polarizability or mass displayed ordering, rigidifying, condensing and/or raft promoting activity on membranes higher than those steroids having low values in such molecular properties. After a variable selection procedure circumventing correlation problems among descriptors, area and log P were found as the most relevant properties in governing and predicting the activity of steroids on membranes. A logistic regression model as a function of the area and log P of the steroids is proposed, which is able to predict correctly 92.5% of the cases. A rationale of the findings is discussed.

1. Introduction

Several functions of biological membranes, such as the sorting of membrane components, membrane signaling, viral budding, biosynthetic and endocytotic trafficking, *etc.*, have been linked to the physical state of the bilayer [1–9]. The pioneering depiction of a homogenous phospholipid matrix [10] in the early seventies, was gradually modified by introducing the raft hypothesis in the nineties, which proposed a laterally segregated and heterogeneous distribution of lipid molecules with different biophysical properties [1,11,12]. Typically, the ordered gel phase bilayer displays a tight phospholipid molecular packing in which the lipid molecules also have restricted lateral motion. In the liquid-crystalline phase, a more disordered structure exists with a faster lateral molecular motion. A third phase, the liquid-ordered phase, exhibits a well-packed and ordered arrangement of lipids, together with a relatively fast lateral diffusion [13–15]. Even though the bilayer architecture depends primarily on the physicochemical properties of phospholipids, their differential interactions and so the miscibility of the components, further molecules incorporated to the bilayer, such as steroids, can modify its properties [16].

The literature on the non-genomic effects of steroids generally deals with their ability to modify membrane properties such as permeability, lateral diffusion, the ordering/packing of lipids, and formation/stabilization of lateral-segregated lipid domains [17–30]. Unfortunately, there is a significant inter-study variability and even conflicting results, where certain steroids have been reported as having opposite activities on membranes in different studies [29]. At least partially, this can be attributed to the varying experimental conditions (type and concentration of phospholipids, steroid concentration, *etc.*) and the wide range of methods that have been employed in different works (detergent solubility, fluorescence techniques, optical and electron microscopy, differential scanning calorimetry, nuclear magnetic resonance, electron paramagnetic resonance, Fourier transform infrared spectroscopy, wide- and small-angle X-ray scattering, differential scanning densitometry, dilatometry and ultrasound velocimetry, freeze-fracture electron microscopy, atomistic molecular dynamics simulations, *etc.*). Most studies involve a small number of structurally related steroids (usually less than ten) with similar structural motifs in either the steroid ring or alkyl side-chain, but which may include differences in the number and position of ring and side-chain double bonds and methyl branches. Thus, inferences from a limited number of steroid/lipid mixtures cannot be easily generalized or extrapolated to other steroid/lipid systems. A wider set of steroids should be evaluated concurrently to solve this matter in order to establish relationships between steroid molecular properties and their activity on membranes.

Abbreviations: PCA, principal component analysis; PC, principal component; LR, logistic regression; ULR, univariate logistic regression; MLR, multivariate logistic regression; OR, odds ratio; CI, confidence interval.

E-mail address: jwenz@criba.edu.ar.

http://dx.doi.org/10.1016/j.bbamem.2015.07.017
0005-2736/

In 2012 I presented a study [30] regarding the relation between the chemical structure of steroids and its effect on steroid-containing membranes, attempting to reveal the effect of different substituents on the biophysical state of bilayers. The study was carried out using combined multivariate approaches based on principal coordinates analysis and logistic regression on a library of steroids with a documented effect on membrane properties. It was found that the presence (or the absence) of certain structural traits at specific positions in fused rings of steroids (an 8 to 10 isoalkyl side-chain at C17, an hydroxyl group at C3 and a double bond between C5 and C6) are the most influencing factors in determining the physical properties of steroid-containing bilayers. In the present study, the molecular properties of steroids are correlated with their effects on the biophysical state of steroid-containing membranes through a comprehensive analysis of a similar library of steroids, whose impact on different lipid mixtures has been documented (Supplementary Table 1). The effect of 82 steroids (Supplementary Fig. 1) on membranes was quantified by defining a binary variable that encodes and categorizes the reported membrane activity. This encoding procedure, stated in the previous work [30], was required seeing that the membrane activity of steroids has been typically reported in qualitative or semi-quantitative terms, and there is not an absolute unit of measurement. The effect of a steroid on membranes has been mainly expressed relative to a control cholesterol-containing or to a steroid-free bilayer. After a computer-assisted construction and geometric optimization of the molecules, several molecular descriptors were computed. Principal component analysis (PCA) and logistic regression (LR) were subsequently performed to assess the influence of the descriptors on the membranes and to develop a simple model to estimate and predict the steroid activity as a function of its molecular properties. Given that no distinctions were made regarding aspects other than the molecular properties and the membrane activity of steroids, this work attempts to ascertain a broad picture of the steroid property–activity relationship, without concern for the composition of the phospholipid matrix, methods and/or experimental conditions associated to the measurements.

2. Methods

2.1. Codifying the steroid membrane activity

The effects of steroids on phospholipid membranes, mostly documented in qualitative terms with reference to those of cholesterol, were analyzed by means of a quantitative analysis. The documented activity of steroids was translated into quantitative data by coding the reported information into categorical values, as described in a previous work [30]. Briefly, a categorical variable, designated as "activity", was 1 for those steroids reported as having rigidifying, molecular ordering, condensing effect, and/or raft promoting/stabilizing ability on membranes, relative to that of steroid-free membranes; these steroids were designated as promoters. When a steroid displayed the opposite membrane activity the variable activity was 0; this refers to those steroids reported as having fluidifying, disordering, and/or raft disrupting/destabilizing effect on membranes, and they were designated as "disrupters". It should be mentioned that, as usually occurs throughout a discretization process, some information is lost after the conversion of the original data into categories, as different magnitudes of the same activity were assumed as equals (*e.g.*, steroids documented as having different degrees of condensing effects on membranes will have an equal activity of 1). Those steroids reported as having no effect on the membrane were named as "neutral", a third category with a value of 0.5.

It is worth noticing that in this work the relationship between steroid properties and their action on membranes is evaluated in relation to a control membrane containing no steroid, and not relative to a cholesterol-containing lipid mixture, as is frequently employed.

As mentioned, some steroids have been reported as having different membrane activities. Hereafter, the term "case" will be employed to refer to each time an activity (equal to or different from) of a given steroid was reported. For that reason, the activity of some steroids is quoted more than once and the number of cases is higher (212) than the number of steroids (82). Finally, each steroid has an encoded value (0, 0.5 or 1) that reflects its reported membrane activity.

2.2. Geometry optimization and calculation of molecular descriptors

In order to compute the molecular descriptors on reasonable low-energy conformations that simulate that in the bilayer, the structure of steroids was modeled using MOPAC [31]. Particular care was taken when constructing the molecular structures (Supplementary Fig. 1), in order to respect the spatial isomers informed by the authors, such as chirality (R,S), epimers (alpha, beta) and conformers of every atom and ring in the molecules. Standard bond lengths, and planar and dihedral angles from the program database were firstly assigned to the structure. The geometry optimization of each steroid was accomplished by using the quantum chemical semi-empirical method AM1. The modeling was performed using the conjugate gradient Polak–Ribiere algorithm and a gradient limit of 0.01 kcal/Å mol. The energy of the 82 molecules were minimized using the previous procedure and conditions, in order to find realistic stable conformations that simulate the one in the bilayer. In spite of the approximations that all semi-empirical methods utilized in their calculus, AM1 is a common choice suitable for a basic modeling of medium to large complexity systems or molecules [32]. More accurate modeling requires the application of complex, time-consuming *ab initio* quantum chemistry software.

Molecular descriptors (variables that inform on some feature of a molecule) encoding information about the molecules were computed on the optimized structures using the Dragon software [33]. A total of 245 descriptors were calculated and distributed in groups as follows: 48 constitutional descriptors, 154 functional group counts, 14 charge descriptors and 29 molecular properties descriptors. In a first step, 152 descriptors were discarded from the analysis for being constant along the steroids. Finally, 93 descriptors remained for further analysis, distributed as follows (Supplementary Table 1): 29 constitutional descriptors (31%), 21 functional group counts (23%), 15 charge descriptors (16%) and 28 molecular properties (30%). Constitutional descriptors are the most simple and commonly used descriptors, informing on the chemical composition of a molecule without any information about its molecular geometry or atom connectivity. Many of them are well explained by their definition such as the molecular weight, number of atoms, number of rotatable bonds, number of specific atoms (hydrogens, carbons, nitrogens, oxygens, *etc.*). Functional group counts inform on the number of chemical functional groups in the molecule (number of different types of carbons, ketones, hydroxyls, alcohols, ethers, *etc.*). Charge descriptors informs on the electronic characteristics of molecules, based on the atomic charges of atoms and the 3D geometry of the molecule. The molecular properties block includes a set of heterogeneous molecular descriptors describing physicochemical and biological properties. Detailed information on descriptors can be found elsewhere [34].

2.3. Principal component analysis

This procedure was employed to search for similarities among steroids regarding their molecular properties (93 descriptors) and group steroids accordingly. Then, by determining the prevailing membrane activity of the steroids within each cluster, this would inform on the link between membrane activity and molecular properties. The dimension reduction method known as principal component analysis (PCA) is a mathematical manipulation of data where the goal is to represent the variation present in many independent variables using

a small number of factors or principal components (PC), and it is able to explain as much variability as possible using a reduced number of variables [35,36]. The method is useful to explore and visualize similarities (and dissimilarities) between samples by means of a graphical representation of multidimensional distances between samples. PCA assigns to each sample a location in a low-dimensional space (usually as a 2D or 3D graphic), where individual and/or inter-group differences can be visualized and samples can be classified or grouped according to their nearness. It also permits the analysis of the importance and interdependence between variables [35,36]. In this study, PCA was used [37] to classify and organize steroids according to their similarities in the properties represented by the molecular descriptors. The resulting organization and grouping of steroids was correlated next with the predominant activity within clusters. As will be discussed later, PCA was also used to ascertain which descriptors are better correlated with the steroid membrane activity.

Since PCA is sensitive to the units of measurements and places more influence on variables with larger scales, all descriptors were autoscaled prior to the analysis to level their weight. Some multivariate analysis, as PCA, place more weight on variables with higher values, which occurs even if one modifies the units of the variable (*e.g.*, from Dalton to kiloDalton). The autoscaling process places the variables in similar ranges although maintaining their inner variability which is proportionally adjusted within the range [35,36]. The average of each descriptor was subtracted from each value and then divided by the standard deviation of the descriptor along steroids. After this procedure the autoscaled variables fluctuate around zero with a SD = 1. PCA was next applied over the autoscaled matrix of 82 steroids and 93 descriptors.

To assess their performance, PCA models were cross-validated by predicting steroids that were not included in the construction of the model. A randomly chosen subset of steroids was excluded from the model building and next predicted with the model constructed with the remaining samples; this process was repeated 18 times excluding 4 different steroids and 2 times excluding 5 steroids. In this manner, each steroid was excluded once from the calculation and predicted with the respective model, among the 20 constructed models. This validation procedure might be optional when performing a PCA, but it is strongly recommended to ensure that the model is able to correctly predict unknown samples, and not only the samples used in the model build. This model ability will be reflected in the percentage of the data variability that it can explain.

In order to preserve the inherent variability of the information, the term "case" accounts here for any quote of a steroid (with same or distinct activity) that was found. Steroids having all their cases reporting the same activity (*i.e.*, activity 0, 0.5 or 1) obviously retained such activity (*i.e.*, disrupters, neutrals or promoters, respectively. See Supplementary Table 1) regardless of the number of cases involved. In circumstances of disagreement in the documented activity a conservative criterion was taken to assign the activity: steroids having at least one case different from the rest (*e.g.*, one promoter and one o more neutrals and/or disrupters) were designated as neutral; these situations entailed an average activity between 0 and 1 for the steroid. This is precisely, between 0.25 and 0.94 for the present sample set. This can be better understood by observing the columns "membrane activity" on the right end of Supplementary Table 1, where steroids are organized according to their activity: disrupting (top), neutral (middle) and promoters (bottom). There are no contradictory reports concerning the activity of steroids 1 to 24 since all works reported the same activity; consequently, all steroids have a "0" in both columns. No conflicting results were found either for steroids 45 to 82, and all works reported the same activity (promoting). Among the steroids 25 to 44, those having only one reference (see references on the right) retained their reported neutral activity (0.5). For those having two or more references, an average was calculated from the disagreeing membrane activities (which can be 0, 0.5 or 1) and they were then considered as neutral steroids. Following this procedure in order to weigh the activity of each of the 82 steroids according to the type and number of activities reported, 24 emerged as disrupters, 20 as neutrals and 38 as promoters of membrane activity.

2.4. Logistic regression

Given that the reported membrane activity is now expressed with a categorical variable (Section 2.1) and that the molecular descriptors are expressed in their proper units (continuous variables), this mix of variables requires a mathematical tool able to handle both categorical and continuous data. Logistic regression is an explanatory and predictive tool which analyzes the relationship between a dependent binary variable (0 or 1) and the independent variables, which may be of any type, categorical or continuous [38,39]. It may be used to determine the importance or the weight of the independents over the dependent variable, and to estimate a dependent variable as a function of one or more independent variables. A dependent variable is a variable whose value is sensitive to the value of other variables, the independents. In the present work, the dependent variable is the membrane activity of steroids, which will be estimated from the values of several independent variables, the molecular descriptors.

The general logistic regression equation is:

$$p = \frac{1}{1 + \exp -(b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n)} \quad (1)$$

where p is the probability of an event to occur; b_0 is the constant of the model, and b_n is the regression coefficient of each n independent variable (X). In this work, the event to be predicted is that a steroid displays membrane promoting activity as a function of the molecular descriptors. Thus, the probability p will range within the 0–1 interval, and it is expected to be near 1 for membrane promoting steroids, and near 0 for membrane disrupting steroids. The cut value was established at 0.65 and thus steroids with a p higher (or lower) than this value were classified as promoters (or disrupters).

The weight or importance of each descriptor is represented by the regression coefficient b_n and/or the corresponding odds ratio (OR), related to both b and p of the event to occur as:

$$OR = \exp(b) = \frac{\left[\frac{p}{1-p}\right]_{x=1}}{\left[\frac{p}{1-p}\right]_{x=0}}. \quad (2)$$

Eq. (2) represents the increase in the odds of a steroid to have membrane promoting activity if the descriptor increases in one unit (from $x = 0$ to $x = 1$), as long as the rest of the descriptors remain equal. Molecular descriptors positively correlated with a promoting activity are expected to exhibit regression coefficients $b > 0$, and ORs > 1. In contrast, descriptors positively correlated with a disrupting activity (or negatively with a promoting one) are expected to have negative regression coefficients b and ORs close to 0 (*i.e.*, $b < 0$, and $0 < OR < 1$). Finally, descriptors having no effect on the activity are expected to have coefficients $b \sim 0$ and OR ~ 1. OR can range from 0 to, theoretically, infinite.

Since interpretation of logistic regression is based on the increment in one unit in the dependent variable (descriptor), the outcome parameters are not directly comparable among descriptors whether they are expressed in different units or scales. For instance, a similar coefficient b obtained for descriptors log P and molecular weight expressed in their original units would not represent an equal influence on the steroid activity, as one-unit increment in log P (which range from 2.4 to 8.7 in the present steroid library) is relatively more significant than a one-unit increment in mass (which range from 272 to 487 g/mol) (see units and scales in Fig. 3). In view of that, each value of the descriptors (a_i) were converted into normalized values (a_N) by dividing by the

average of the descriptor (a_m) along the 212 cases, and then multiplied by 100:

$$a_N = \left({}^{a_i}/_{a_m}\right)100. \quad (3)$$

After this transformation, all descriptors were expressed in equivalent units and an increment of one-unit can be promptly compared.

This multivariate tool can work either with continuous and/or binary independent variables, but only a binary variable (with two categories) is allowed as the dependent. In line with the intention of capturing the variability of the information, steroids reported as having no action on membranes (*i.e.*, neutral, activity = 0.5) were not discarded but redistributed among the disrupter and promoter categories under a defined criterion. Steroids with all of its cases reported as neutral were shared out equitably and randomly among the two categories, generating roughly an equal quantity of the opposite 0's and 1's. If more than one case having a different activity was found for a given steroid, the guideline for the redistribution between categories was the maintenance of the overall activity of such cases. Thus, neutral cases were redistributed between the disrupter and promoter categories in such a way to keep the average as constant as possible. For an even number of neutral cases this was accomplished by sharing out a half and a half to each category. For example, in a situation where two promoter and two neutral cases were observed for a given steroid [average activity = (1 + 1 + 0.5 + 0.5) / 4 = 0.75], one neutral was assigned to the promoter and one to the disrupter category [resulting average activity = (1 + 1 + 1 + 0) / 4 = 0.75]. In this manner, the intrinsic variability of data encoded with three categories is maintained after the redistribution of cases into two categories. Following this procedure, the thirteen neutral cases were shared out between the disrupter and promoter categories.

3. Results

3.1. Grouping of steroids based on the similarities in their molecular properties

PCA was performed to reduce the dimensionality of the data (93 variables) and next construct plots of two or three dimensions (components), where steroids are expected to arrange in groups according to the similarities in their molecular properties. Only the molecular descriptors (autoscaled) were employed for the PCA and no information on steroid activity was employed in this step. From the 93 descriptors encoding information, the first two orthogonal components (PCs) were able to explain 63% of the entire steroid data set variability. After the incorporation of the third PC, the explained variance increases to 71%. Although more PCs might be included attempting to increase the explained variance, three PCs are adequate for the actual purposes without making the interpretation difficult. The consideration of a fourth (or higher) dimension makes the visual inspection of clustering patterns difficult, complicating the interpretation of the outcome.

The coordinates of steroids in the new axes (PC1, PC2 and PC3) are shown in Fig. 1. This score plot reveals patterns, where the distance between points is proportional to differences (or similarities) in the molecular properties of this steroid set. To help in the view of this three-component scattering plot, values of PC3 are highlighted as projections of the points (dotted lines) onto the plane PC1/PC2. On the basis of the 93 molecular properties examined, an obvious clustering pattern is observed, with the majority of the steroids spread out into one of two clusters. In the context of this analysis, the steroids within cluster 1 are more similar between them, *i.e.*, are in closer proximity, than those in cluster 2. Thus, the arrangement in two dimensions (PCs) of steroids in Fig. 1 reflects the similarities regarding all the molecular properties. The next step was to ascertain the prevailing membrane activity within each cluster.

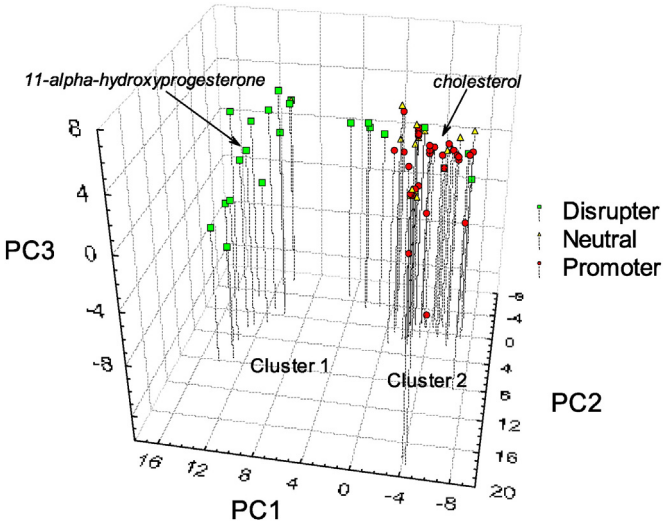

Fig. 1. The grouping of a range of steroid chemical structures based on a PCA analysis (with 3 principal components) of 93 molecular properties (see Supplementary Table 1) described in Sections 2.2 and 3.1, and categorized according to Section 2.3. The dotted lines are projections of PC3 values onto the plane PC1/PC2. Symbols denote the three categories of activities: disrupters (green squares), neutrals (yellow triangles) and promoters (red circles). A disrupter (11-alpha-hydroxyprogesterone) and a promoter (cholesterol) steroid are shown as examples.

3.2. Association between molecular properties and membrane activity of steroids

The membrane activity of steroids in each cluster was counted to search for correlations with the molecular properties. The type of activity (promoter, disrupter and neutral) is shown in Fig. 1 as different symbols for each category. As mentioned for the PCA (Section 2.3), neutral steroids are those documented as such, but also those for which at least one case was different from the rest. Accordingly, three categories emerge from the 82 steroids regarding their membrane activity: 24 disrupters (average activity = 0), 20 neutrals (average activity ranging between 0 and 1) and 38 promoters (average activity = 1). The steroids included in each of these categories can be observed in the column "Membrane activity for principal component analysis", on the right end of Supplementary Table 1.

Panel A in Table 1 shows the distribution of steroids according to their activity and cluster membership calculated using the PCA.

Table 1
The distribution of steroid clusters according to their effect on membrane activity. Steroid counts are shown on the left block, and percentages on the two right blocks of the table. Steroid categories were determined from both the score plots and from the score sheets of the PCA results. PCA models were constructed from the initial 93 molecular descriptors (panel A) and with the 7 molecular descriptors selected subsequently (panel B). See Section 2.3 for details about the procedure to establish the activity of steroids into one of the three categories.

	Steroid counts			% in cluster		% in category		
	Cluster 1	Cluster 2	Total	Cluster 1	Cluster 2	Cluster 1	Cluster 2	
Panel A								
Disrupters	15	9	24	93.8	13.6	62.5	37.5	100
Neutrals	1	19	20	6.2	28.8	5.0	95.0	100
Promoters	0	38	38	0	57.6	0	100	100
Total	16	66	82	100	100			
Panel B								
Disrupters	19	5	24	95.0	8.1	79.2	20.8	100
Neutrals	1	19	20	5.0	30.6	5.0	95.0	100
Promoters	0	38	38	0	61.3	0	100	100
Total	20	62	82	100	100			

Memberships to each cluster were determined from the score plot and score sheet of the PCA results. The majority cluster 1 members (around 94%) were steroids with a disrupting activity. No promoter steroids were found in this cluster, and neutrals represent only the 6%. On the other hand, the cluster 2 population is composed of roughly by 14%, 29% and 58% of disrupter, neutral and promoter steroids, respectively. In other words, cluster 2 encloses 100% of the promoters, 95% of the neutral and roughly 38% of the disrupter steroids (last column in panel A, Table 1). This finding indicates that the activity of steroids in membranes is linked to some of their molecular properties, and that steroids having similarities in some molecular properties are expected to have a similar effect on membranes.

3.3. Molecular properties that determine steroid grouping

3.3.1. Searching for relevant descriptors by the comparison of PCA scores and loadings

After an initial determination of the similarities between steroids based on the PCA scores computed from 93 descriptors (Supplementary Table 1), a combination of approaches was utilized to determine the descriptors having more influence in determining the segregation of steroids into clusters, and consequently linked to their activity.

Besides the scores, the PCA outcome includes the loadings, parameters that describe the data structure in terms of the weight of the descriptors and the correlations between them. As each steroid has a score on each PC, each descriptor has a loading on each PC, which fluctuates between −1 and 1. Loading reflects both how much the descriptor contributed to that PC, and how well that PC takes into account the variation of that descriptor over the data. A descriptor having small loadings, whatever the sign of such loadings, is not useful for interpretation since that descriptor is poorly accounted for by the PC [35,36]. Following this rationale, those descriptors with small loadings, *i.e.*, located close to the intersection between PC1, PC2 and PC3 axes (point 0,0,0) in the loading plot (not shown), were the first descriptors discarded from the analysis.

The sign of a descriptor loading can be compared with the sign of a sample score to look for correlations between variables and steroids. If a descriptor has a positive loading, it means that all steroids with positive scores have higher than average values for that descriptor; on the contrary, steroids with negative scores have lower than average values for that descriptor. If a descriptor has a negative loading, it means just the opposite. Thus, steroids with positive scores have lower than average values for that descriptor and steroids with negative scores have higher than average values for that descriptor. If the number of samples plus variables are not so large (which is not the case), this can be observed by placing scores and loading in the same plot. The nearness between a steroid (scores) and a descriptor (loadings) will mean that such steroid have values higher than average for that descriptor. The rationale is valid in quantitative terms as well, as the higher the positive score of a steroid, the larger its values for descriptors with positive loadings, and *vice-versa* [35,36]. In view of this observation, the search and selection of the most relevant descriptors that govern the segregation and clustering of steroids was focused on those having the highest loadings and that were located in regions equivalent to that of the clusters in the PC1/PC2/PC3 space (see Fig. 1). The descriptors that fulfill these two requirements and concurrently have a straightforward interpretation were: the partition coefficient (log P), area, volume, mass, refractivity, number of rotatable bonds and polarizability. The loading of these descriptors have negative values on PC1 and values scattered around zero on PC2 and PC3, which correspond to a spatial region equivalent to that of cluster 2 (see Fig. 1). According to the previous rationale, steroids in cluster 2 are expected to have the highest values for log P, area, volume, mass, refractivity, number of rotatable bonds and polarizability, in agreement with the next findings from the *t*-test (Section 3.4) and from the logistic regression analysis (Section 3.5).

3.3.2. A simplified PCA model constructed with the most relevant descriptors

To reinforce these statements, a new PCA model was then constructed considering only the 7 selected descriptors. Using a smaller number of variables both the goodness of fit and the discriminating power of the model improved. Only two PCs suffice to explain 98% of the total data variability, in contrast with the 63% when employing the 93 descriptors. Given that the third PC would add only 1% to the explained variance, its incorporation is not worthy and thus the model can be easily represented in a two-dimension score plot (following the principle of the most simple, parsimonious model). With respect to the training step, the explained variance of the data in the validation step (*i.e.*, prediction of samples that were not employed in the model building) decreased only from 98% to 97%, indicating a good stability and performance of the model. The score plot for the 82 steroids in the new PC1/PC2 axes is shown in Fig. 2. Again, two clusters appear segregated from each other, mirroring similarities (intra-cluster) and differences (inter-cluster) of its members on the basis of the seven molecular properties considered. These findings agree with those achieved when analyzing the 93 descriptors.

The distribution of steroids according to the activity and cluster membership (*i.e.*, degree of similarity) obtained with this simple model is shown Table 1, panel B. The number of steroids having each type of activity (disrupter, neutral or promoter) remains the same as in the previous PCA model (Table 1, panel A). Only one difference was found in the steroid counts. Comparing both panels, four disrupter steroids classified as members of cluster 2 by the former model (constructed from 93 descriptors, 3 PCs) were next classified as members of cluster 1 by the second model (7 selected descriptors, 2 PCs). This rearrangement of cluster memberships led to a more accurate classification of steroids, as the percentage of disrupter steroids in cluster 1 slightly increased to 95%, with only one steroid belonging to the neutral category. The population of promoters in cluster 2 also increased to 61% and that of disrupters decreased to around 8%, whereas the percentage of the neutral steroids remains almost the same (roughly 29% and 30%, respectively). In other words, almost 80% of the disrupter steroids were placed in cluster 1, whereas the totality of promoters (100%) was located in cluster 2 (Table 1, panel A, right block). These findings reinforce the previous findings concerning the link between the activity of steroids and some molecular properties, deduced from the

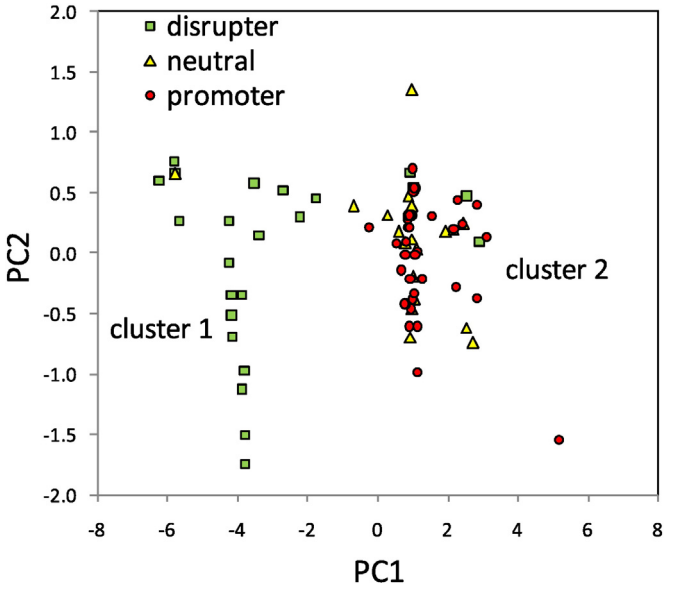

Fig. 2. Grouping of steroids according to similarities/differences as deduced from the first two PCA components assessed from the seven selected molecular descriptors (log P, area, volume, mass, refractivity, number of rotatable bonds and polarizability). The symbols are the same as those used in Fig. 1.

correlation between scores and loading. The selected descriptors log P, area, volume, mass, refractivity, number of rotatable bonds and polarizability were found to be straightforward related to the effect of steroids on the biophysical state of membranes. Promoter steroids exhibit values of such molecular descriptors higher than disrupters.

3.4. Differences in the molecular properties of promoter and disrupter steroids

The differences among the promoter and disrupter populations regarding the 7 selected molecular properties found in the preceding analysis were checked by means of a *t*-test. Instead of steroids, the 212 cases were employed in the trial as each one is an individual and independent report on the activity of a steroid. In addition, by including all cases the inter-study variation is again captured. Since one descriptor was compared each time, no problems with units or scale arise; no pre-processing was required and the original values were then employed. Fig. 3 shows the average value of each descriptor for promoter (n = 151) and disrupter (n = 61) cases. The seven descriptors displayed higher values in the promoter than in the disrupter population at a high degree of significance ($p < 0.001$), in agreement with the previous findings from the correlation between scores and loading.

The correlation between descriptors was examined by means of the squared correlation coefficients (R^2) for each pair (Table 2). All descriptors are positively correlated, indicating that an increase in one descriptor is linked to an increase in other descriptors, or *vice-versa*. The correlation between some descriptors can be easily presumed. For instance, an increase in the molecular volume is expected to be associated with an increase in the molecular area or in the molecular weight. On the basis of these correlations between descriptors, it is not possible at this stage to dissect the relative importance of these seven descriptors in affecting the physical state of membranes.

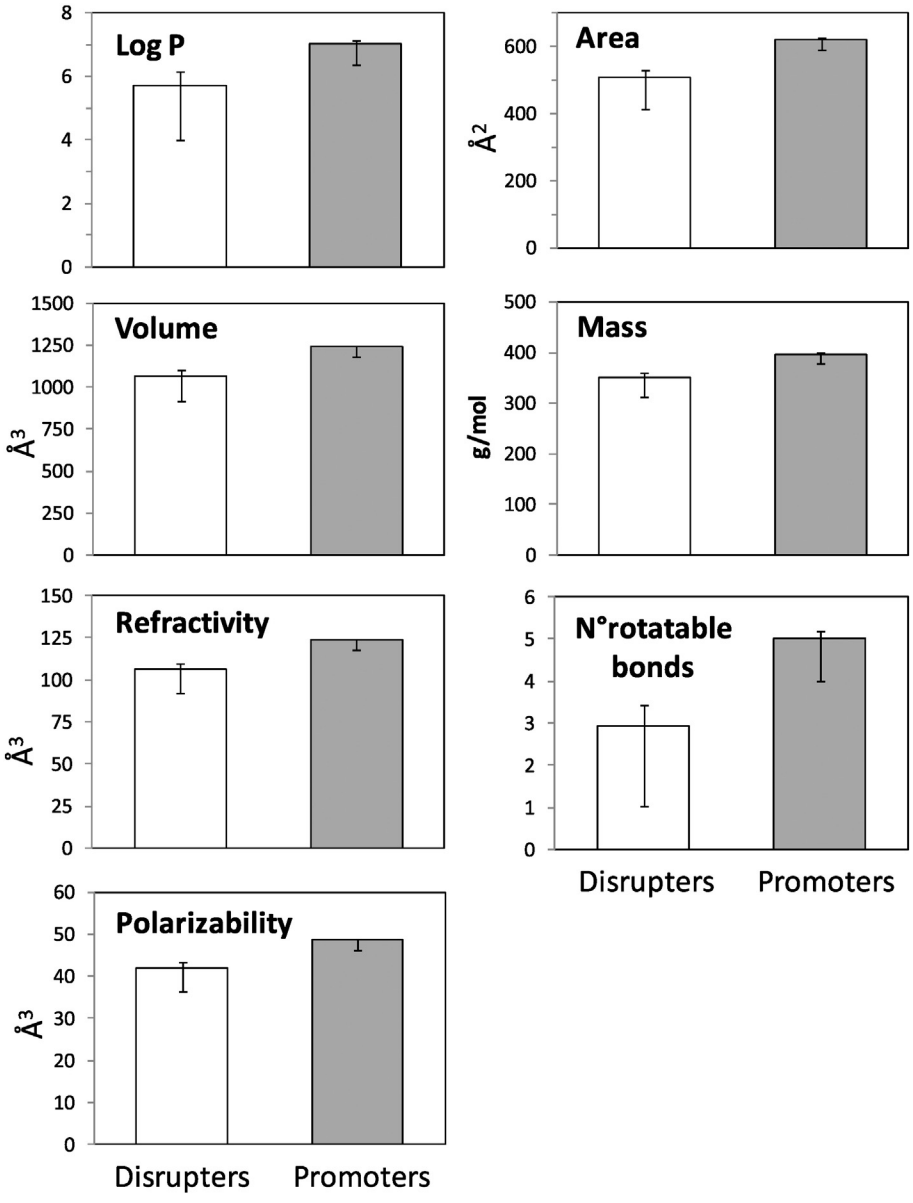

Fig. 3. Comparison of molecular descriptors between known disrupter cases and known promoter cases. Error bars: lower, SD; upper, CI 95% (average ± 1.96 * S.E., where S.E. = SD/√N). The differences between the disrupter and promoter cases are significant (at $p < 0.001$) for all the examined descriptors. The total number of cases studied was 212 (61 disrupters and 151 promoters). The general average of each descriptor, *i.e.*, considering both disrupter and promoter cases, are: log P: 6.6; area: 588.8 Å^2; volume: 1189.2 Å^3; mass: 383.8 g/mol; refractivity: 118.7 Å^3; number of rotatable bonds: 4.4; polarizability: 46.7 Å^3.

Table 2
The squared correlation coefficients (R^2) between the seven selected molecular descriptors.

	Log P	Area	Volume	Mass	Refractivity	N° rotatable bonds	Polarizability
Log P	1						
Area	0.73	1					
Volume	0.74	0.99	1				
Mass	0.53	0.86	0.89	1			
Refractivity	0.72	0.93	0.97	0.88	1		
N° rotatable bonds	0.58	0.86	0.80	0.70	0.66	1	
Polarizability	0.75	0.95	0.98	0.88	0.98	0.70	1

3.5. Molecular properties governing the activity of steroids on membranes

3.5.1. Univariate logistic regression (ULR)

In order to identify the relative influence of each of the 7 most important variables determining the steroid's activity descriptors found as the most important in determining the membrane activity, and independently of the others, ULR was performed [40] firstly between the activity and each of the 7 descriptors independently, disregarding their possible interactions. The weight of each molecular property determining the steroid membrane activity is represented by the *b* coefficient (Fig. 4). A value of *b* above or below 0 implies that an increase in the descriptor is associated with an increase or a decrease in the probability *p*, respectively (Eq. (1)). As *b* approaches zero, the smaller is the influence of the descriptor on steroid membrane activity. If the 95% confidence interval for *b* ($CI = b \pm 1.96 * S.E.$) includes zero, the sterol membrane activity is not significantly dependent (at $p < 0.05$) on the molecular property. Fig. 4 shows that the seven descriptors displayed positive *b* values and their confidence intervals did not contain zero. This indicates that independently increasing any of the 7 descriptors increases *p* and promoter steroids have higher values in those molecular descriptors than do disrupters, in agreement with the results in the preceding Sections 3.3 and 3.4.

As with any statistical method, there are some parameters in the logistic regression results that must be examined to ensure their reliability. The Wald chi-square test in a logistic regression proves the null hypothesis that the parameter (*b* coefficient and the constant) equals zero, and so it tests the statistical significance of coefficient *b* and the constant of the model. The hypothesis is rejected if the associated *p*-value is smaller than a given critical value (*e.g.*, 0.05) and the conclusion is that the parameter is not zero. As shown in Table 3, all descriptors were highly significant (high Wald values and associated $p < 0.001$ for both the coefficients and the constants), indicating that when taken independently, all seven molecular descriptors are strongly correlated with steroid activity in model membranes.

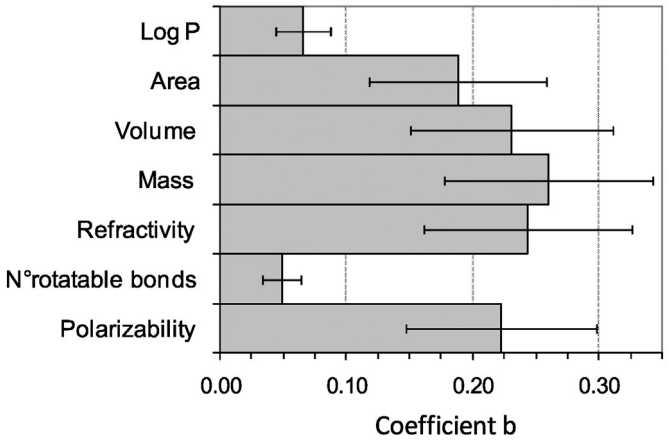

Fig. 4. The importance (expressed as coefficients *b*) of the seven selected molecular properties in determining steroid activity on membranes, based on the univariate logistic regression with each descriptor taken singly. The error bars indicate a 95% confidence interval for the mean ($b \pm 1.96 * S.E.$).

OR is a function of coefficient *b* (Eq. (2)) and it represents the increase in the odds of a steroid to have membrane promoting activity (*i.e.*, probability $p \approx 1$) if the descriptor increases in one unit (*e.g.*, from 1 to 2, or 3 to 4). An OR higher than unity is expected for descriptors positively associated with a promoting activity in the steroid; *i.e.*, an increase in the descriptor is correlated with an increase in the promoting activity. On the other hand, OR values within the limits 0–1 denote a steroid disrupting membrane activity. The 95% confidence interval for OR is calculated as $\exp(b \pm 1.96 * S.E.)$ and is asymmetric due to this mode of calculation. Thus, the mean value of OR is not centered between these limits (Table 3), but closest to the lower limit. If 1 is contained in the interval, it should be concluded that there are no significant differences (at $p < 0.05$) in the OR after a one-unit increment in the descriptor and thus there is no reason to assume that the variation in the OR is not due to a random error. Analyzed isolated, all descriptors had ORs higher than 1 and 95% CIs that exclude the unity (Table 3), indicating that the likelihood of a steroid being a promoter increases with an increase in any of the 7 selected descriptors. For example, the odds of a steroid being a promoter increase 1.07 times (which means a 7%) for each integer increase in the log P (expressed in the normalized units employed in the building of the logistic regression model). Similarly, it can be said that if two steroids differ in one normalized unit of mass, the one with the highest value has 30% (*i.e.*, OR = 1.30) more chances to be a promoter than the other with the lower value (see Table 3). It should be reminded that descriptors were mean normalized and raised to 100 prior to performing the logistic regression. Hence, the variation in the OR associated with an increment in one normalized unit is equivalent and comparable among the descriptors, independently of the original units and/or ranges. Conversion between the normalized and original units can be accomplished using Eq. (3). Thus, an increase of one normalized unit corresponds to the following increments in the original units of each descriptor: area = 5.9 $Å^2$, volume = 11.9 $Å^3$, log P = 0.07, refractivity = 1.19 $Å^3$, polarizability = 0.47 $Å^3$, mass = 3.84 g/mol and the number of rotatable bonds = 0.04. From previous examples, it is evident that the probability of a steroid being a promoter of membrane activity increases by 7% for each 0.07 unit increase in log P. However, such probability increases 30% for each time the mass increases by 3.84 g/mol.

Statistical parameters describing goodness of fit, such as −2 Log Likelihood and the Nagelkerke R^2, are useful to compare LR models. The latter is a pseudo R^2 and is analogous to the R^2 in standard multiple regression, but it does not represent the extent of variance in the dependent variable accounted for by the independent variables. Best models have low values of −2 Log Likelihood and high values of the Nagelkerke R^2 (within the limit 0–1), and are usually inversely related [38,39,41] (Table 3). The model of molecular area displays the best fit to data among the 7 univariate models, as deduced from both the smallest −2 Log Likelihood and the highest Nagelkerke R^2. As the goodness of fit does not depend on the value of OR, no relation should be expected between these parameters. For instance, the −2 Log Likelihood and the Nagelkerke R^2 for descriptor area are 149.4 and 0.56, respectively (Table 3), which are better parameters than those for descriptor mass (161.4 and 0.51, respectively). Nevertheless, for equivalent increases

Table 3
The logistic regression models. The last four columns on the right describe the goodness of fit between the observed and predicted values for disrupting (D) and promoting (P) indicators of membrane activity for a range of steroid chemical structures. Column "Obs.": number of known disrupter (D) and promoter (P) cases in the data set. Columns "Predicted": number of cases predicted as disrupting (D) or promoting (P) by the model; "% Corr.": percentage of cases correctly predicted into each category (represented in rows named D and P within each model); "Overall": overall percentage of cases correctly predicted by the model, including both promoters and disrupters. The meaning of the statistical parameters is explained in Sections 2.4 and 3.5.

Dependent variable: activity									Predicted			
Independent variables (descriptors)	Wald	Sig. (p)	exp(b) [OR]	95% CI lower	95% CI upper	−2 Log Likelihood	Nagelkerke R^2	Obs.	D	P	% Corr.	Overall
Univariate												
Log P	33.7	0.000	**1.07**	1.04	1.09	205.0	0.30	D	36	25	59.0	**82.5**
Constant = −5.4	24.1	0.000						P	12	139	92.1	
Area	27.1	0.000	**1.21**	1.12	1.30	149.4	0.56	D	36	25	59.0	**87.3**
Constant = −17.8	23.2	0.000						P	2	149	98.7	
Volume	31.4	0.000	**1.26**	1.16	1.36	154.3	0.54	D	37	24	60.7	**86.3**
Constant = −22.0	27.9	0.000						P	5	146	96.7	
Mass	38.0	0.000	**1.30**	1.19	1.41	161.4	0.51	D	39	22	63.9	**87.3**
Constant = −24.7	34.9	0.000						P	5	146	96.7	
Refractivity	33.5	0.000	**1.28**	1.17	1.39	150.5	0.55	D	37	24	60.7	**86.3**
Constant = −23.2	30.2	0.000						P	5	146	96.7	
N° rotatable bonds	39.6	0.000	**1.05**	1.03	1.07	175.8	0.44	D	34	27	55.7	**86.8**
Constant = −3.7	22.8	0.000						P	1	150	99.3	
Polarizability	33.3	0.000	**1.25**	1.16	1.35	158.2	0.52	D	36	25	59.0	**86.8**
Constant = −21.1	29.7	0.000						P	3	148	98.0	
Multivariate with variable selection												
Area	33.4	0.000	**1.43**	1.27	1.62	129.7	0.64	D	49	12	80.3	**92.5**
Log P	15.7	0.000	**0.89**	0.83	0.94			P	4	147	97.4	
Constant = −22.55	34.2	0.000										

in each descriptor (*i.e.*, one normalized unit in both, or a 5.9 Å^2 increase in the area and 3.84 g/mol in the mass) the increase in the probability of being a promoter is lower for the area (OR = 1.21; 21%) than for the mass (OR = 1.30; 30%).

The results of this statistical method also reveal the LR model's ability to correctly classify cases (see the last five columns of Table 3), *i.e.*, the percentage of citations with a known membrane activity that is correctly predicted by the LR model. A case is correctly classified when the membrane activity deduced from the computed probability p (Eq. (1)) coincides with the known activity for the case. If $p > 0.65$ or <0.65 the steroid was classified as a promoter or disrupter, respectively. A LR model with no predicting capability at all should correctly classify approximately 50% of the cases because of simple rules of probability, whereas a satisfactory model should correctly classify more than 75% of cases [38]. As shown in Table 3, all of the ULR models presented here predict at least 82.5% of the cases into the correct class as a function of one descriptor. The descriptor "area" displays the best prediction capability, with 87.3% of cases classified into the correct class, in agreement with the preceding finding from other goodness of fit parameters. It is worth noting that all ULR models classified promoter steroids more accurately than disrupters. This difference may depend on the degree of correlation between descriptors and the activity of steroids, and/or on the unbalanced proportions between disrupter and promoter cases in the data set. The accuracy of the predictions is sensitive to the total number of cases (212) but also to the proportion of cases belonging to each category employed in the model build [38,39]. Note that in this work, there are more promoters (n = 151) than disrupters (n = 61).

These observations obtained using the ULR method considering each descriptor individually provide an initial picture of the relative importance of the molecular properties that affect steroid membrane activity. Any of these 7 models corresponding to each descriptor classifies more than 82.5% of the cases into the correct class (last column in Table 3). Replacing the corresponding coefficient b (expressed as exp(b) in Table 3) in Eq. (1), each of the 7 models can be employed to predict the activity of a steroid on membranes by calculating the probability (p) as a function of only one descriptor. The accuracy in the predictions of each model is shown in the last four columns of Table 3, as absolute counts and percentages of cases for which their membrane activity was correctly predicted. Concerning descriptor volume, for example, of the 61 known disrupter cases (sum of 37 + 24, columns "D" and "P", in Table 3) the model predicts 37 cases as disrupters and 24 as promoters, which means 60.7% of cases (37 ∗ 100 / 61) were correctly predicted (see Table 3, column "% Corr."). For the 151 known disrupters (5 + 146) the same model calculation using the volume descriptor predicts 5 cases as disrupters and 146 cases as promoters, *i.e.*, 96.7% of the cases (146 ∗ 100 / 151) were correctly predicted. The general predicting accuracy of the model, *i.e.*, considering both types of membrane activities for steroids at the same time, was 86.3% [(37 + 146) / (61 + 151)], as shown in the column "Overall" of Table 3. The same rationale is valid for evaluating the predicting capability of the remaining ULR models considering each descriptor individually. A more satisfactory model is, however, proposed in the next section by means of a multivariate analysis that takes into account a possible interaction between descriptors.

3.5.2. Multivariate logistic regression (MLR)

Attempting to develop a model with a balanced combination between predicting capability and amount of variables, a MLR with a forward variable selection was performed [40], taking into account concurrently the 7 selected descriptors mentioned above. The forward variable selection process involves starting with no descriptor in the model, testing the addition of each descriptor and adding the next one (if any) that improves the MLR model for the best predictability of steroid membrane activity. The process is repeated for all the descriptors until the addition of the last descriptor does not improve the MLR model significantly; thus, it stops when the optimum number of descriptors is reached. Area and log P were found to be the most effective descriptors in a MLR model combining both simplicity and predicting ability, with a Wald's associated $p < 0.001$ and a 95% CI that excludes the unit. The remaining descriptors were not included in the model as they do not add extra information nor improve the model performance substantially. However, this does not necessarily imply the absence of correlation with steroid membrane activity.

It should be noted that the coefficients of descriptors differ slightly from those obtained previously analyzed using the ULR. Comparing the univariate with the multivariate approaches, the OR for area (and also the 95% CI) shifts from 1.21 to 1.43, and for log P from 1.07 to 0.89 for log P, respectively (Table 3). The positive association between activity and area remains, but it is more noticeable when area and log

P are included in the model. By means of this MLR model adjusted for both descriptors, an increment in one normalized area unit (a calculated molecular area of 5.9 Å^2 in the original units) signifies an increment of 43% in the odds of a steroid being a promoter, instead of a 21% in the ULR. On the other hand, a slight negative correlation was found between membrane activity and log P in the two-variable model, where OR (OR = exp *b*) shifts from 1.07 to 0.89 (Table 3), *i.e.*, an increase in log P of the steroid implies a lower probability of being a promoter (or in a higher chance of being disrupter). As discussed (see Section 3.4), the 7 selected descriptors were correlated between themselves. For the pair area–log P the square correlation coefficient (R^2) was 0.73 (Table 2), meaning that a fraction of the variability of data is concurrently accounted for by both descriptors. The difference in the OR of log P between the univariate to the multivariate logistic regression does not disagree with the preceding finding concerning the significance of the descriptor in the activity of steroids. Area and log P exhibit both a positive relation with the steroid activity when analyzed isolated in the univariate regression (OR = 1.21 and 1.07, respectively), as previously found. When both are present in the multivariate model, descriptor area (OR = 1.43) is still more relevant in describing the activity than log P (OR = 0.89), but the latter displays an inverse relation with membrane activity to what it does in the ULR (*i.e.*, 0.89 < 1) (Table 2). This apparent contradiction is not surprising considering that a goal of multivariate approaches is precisely to disclose possible correlations and interactions between variables by adjusting their weight (coefficients) accordingly, which are not revealed in a univariate analysis [35,36]. The magnitude of the changes will depend on the degree of correlation and interaction between variables, where the effect of an independent variable (log P) over the dependent variable (membrane activity) may be different (even opposite) because of the presence of another independent variable (area) in the model, with which it interacts [38,39,41]. Due to the reciprocal interaction, a change in the value of one descriptor modifies the effect of the other descriptor on the membrane activity.

It is apparent that the positive correlation of area and log P with steroid membrane promoting activity is not the reason why they have been found as the most relevant. Their importance is dictated by their influence or weight in the MLR model compared to the other descriptors examined, not by the positive or negative correlation with the membrane activity. The variable selection process selects those that better explain the membrane activity. As long as it fits the data adequately, the fewer the number of descriptors the better the model. However, the exclusion of a descriptor from the model does not necessarily imply the absence of correlation between the molecular property and the steroid activity, as shown when descriptors volume, mass, refractivity, number of rotatable bonds and polarizability were excluded when building the area–log P logistic regression model. In the same way, the existence of relevant molecular properties different from those examined here cannot be dismissed, and further studies should be conducted to explore within the thousands of available descriptors that were not examined in this study.

The present model of area and log P displays a better prediction ability than the previous ULR method, as determined by the smaller −2 Log Likelihood (129.7), the higher Nagelkerke R^2 (0.64), and its ability to correctly predict 92.5% of the cases (Table 3). The prediction capability of the two-descriptor model is shown in Fig. 5 as a comparison between the predicted and observed activity. The number of cases (frequency, Y-axis) is plotted against the probability *p* calculated with the model (X-axis), disclosed by known disrupter cases and known promoter cases. Cases predicted as disrupters are those situated on the left side of the cut value of 0.65, and cases predicted as promoters are situated on the right side. The width of each bar on the X-axis is 0.05, and represents the cases contained in each interval of *p*. Some bars are divided into known disrupter cases (dotted green pattern) and known promoter cases (red fill pattern). For example, a bar located on the right side of the cut value corresponds to cases predicted as promoters (*i.e.*, $p > 0.65$).

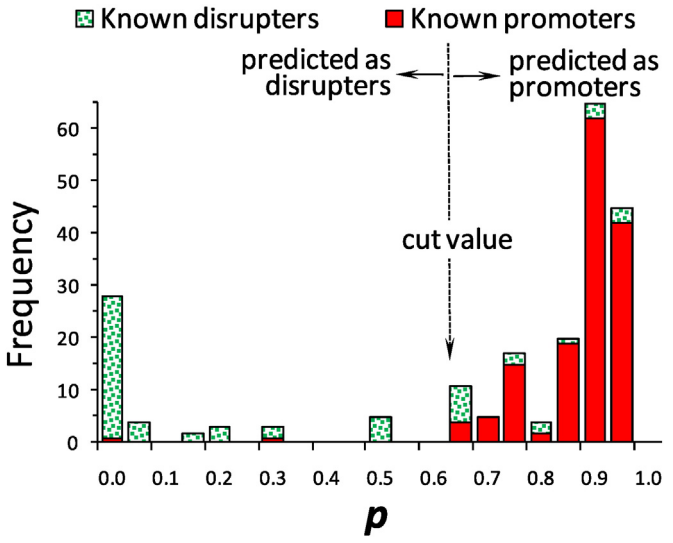

Fig. 5. Known and predicted steroid membrane activity according to the two-descriptor logistic regression model described in Section 3.5. Each bar represents the number of cases (Y-axis) with a probability *p* (X-axis) calculated with the regression model, expressed in intervals of $p = 0.05$. The dotted green pattern within a bar corresponds to the known disrupter cases, whereas the red fill pattern corresponds to the known promoter cases. Thus, a dotted green pattern within those bars situated the right side of the graph (right of $p = 0.65$) represents the misclassified known disrupter cases, whereas a red fill pattern within those bars located onto the left represents misclassified known promoter cases.

Among such cases, the red fill pattern inside the bar represents the known promoters (thus they are correctly predicted) whereas those represented by the dotted green pattern correspond to known disrupter cases (so they are incorrectly predicted). A similar rationale can be followed to interpret those bars on the left side of the cut value, which correspond to cases predicted as disrupter cases. A red pattern in these bars represents known promoter cases incorrectly predicted as disrupter cases. The sum of all the dotted green bars in the graph represents the totality of the known disrupter cases, whereas the sum of the red filled bars corresponds to the known disrupter cases. As expected, the plot has a U-shape, where the predicted number of disrupters increases toward the left, and the predicted number of promoters increases toward the right side. With an ideal theoretical model with perfect discriminating power, no cases should be found in the middle zone of the plot seeing that an ideal model is expected have its predictions (*p*) really close to 1 or 0, *i.e.*, far from 0.5. Additionally, no known cases belonging to the opposite class are expected to be found on each side of the plot, *i.e.*, no known disrupters (dotted green pattern) on the right and no known promoter (red fill pattern) on the left. As can be seen on Fig. 5, the model is real (not ideal) and a few known promoter cases were wrongly predicted as disrupters and thus they are located on the left; similarly, a few known disrupter cases were predicted as promoters and they are located on the right, reflecting the cases for which the model fails in the prediction.

3.5.3. Model validation

To assess the reliability of the multivariate model as a function of the area and log P and the confidence that can be placed on results and predictions, an external validation process was utilized. This process is based on the prediction of the activity of steroids that were not employed in the construction of the model. It is useful to detect bias and inaccuracy of models that are difficult to detect when the model is merely evaluated by predicting steroids that were employed in the model construction. From the 212 cases, a randomly chosen test set of 62 cases was excluded from the data set and a model was constructed with the remaining 150 cases (training set). The activity of steroids in

the test set was then predicted by using Eq. (1) and the parameters obtained with the corresponding model. The test set was then returned to the data set and a new test set was similarly excluded from the model building, repeating the complete process ten times. At each step, the ORs for descriptors area and log P were computed and the activity of steroids in the training set were predicted with each model. Then, the averages and standard deviations of the main parameters of the logistic regressions are shown in Table 4. The average ORs for area (1.47) and log P (0.89) in these ten training sets showed no significant differences with respect to those obtained by employing the entire data set (1.43 and 0.89, respectively) (see Table 3, block "Multivariate with variable selection"). The percentages of cases correctly predicted in both the training set (91.0%) and in the test set (90.2%) were also very close to that obtained from the entire data set (92.5%). These findings indicate the absence of anomalous samples and/or bias in the model, as cases were predicted with models constructed with different (*i.e.*, others) cases, and it refers to a robust model with a scope for predicting the membrane activity of steroids that is not restricted to cases employed in its construction, but also to any typical steroid structure.

3.6. Prediction of the steroid activity as a function of area and log P

Another utility of the logistic regression is that the model allows the calculation of the probability of a given steroid to possess membrane promoting ($p \approx 1$) or disrupting ($p \approx 0$) activity based only on its area and log P. By replacing coefficients b (shown in Table 3 as $\exp(b)$) in Eq. (1), the probability p of a steroid of having membrane disrupting or promoting activity can be calculated as a function of the steroid's area ($Area_N$) and log P ($Log\ P_N$), expressed in normalized units:

$$p = \frac{1}{1 + \exp-(-22.55 + 0.359\ Area_N - 0.121\ Log\ P_N)}. \quad (4)$$

By arrangement of Eq. (3) and combining with Eq. (4), the probability p can be calculated also by introducing the descriptors in their regular units ($Area_i$ and $Log\ P_i$), Å^2 and unitless, respectively:

$$p = \frac{1}{1 + \exp-\left[-22.55 + 0.359\left(\frac{Area_i}{588.8}100\right) - 0.121\left(\frac{Log\ P_i}{6.6}100\right)\right]} \quad (5)$$

where 588.8 and 6.6 (shown in caption of Fig. 3) are the average of area (Å^2) and log P (unitless) in the present data set, respectively. Considering the broad range of cases covered (212), these averages can be assumed as truthful values for the majority of steroids. As an example, for the well-known membrane promoting steroid, cholesterol [23,42,43], which has a molecular area of approximately 616 Å^2 and a log P of 7.2, Eq. (5) gives a value of $p = 0.88$. As the cut value for p was established at 0.65, cholesterol is then classified as a promoter steroid. For an opposite example, the known membrane disrupting steroid 11-alpha-hydroxyprogesterone [27] having a smaller area (422 Å^2) and log P (3.6), the model calculates a $p = 0.03$, and so it is classified as a disrupter ($p < 0.65$). Following this procedure, the two-descriptor model classified into the correct class around 80% and 97% of the disrupter and promoter steroids, respectively, and 92.5% of the totality of steroids (promoters and disrupters) (Table 3). Accordingly, the activity of a steroid in membranes can be predicted if its molecular area and log P are known.

4. Discussion

It was found that an increase in any of the seven molecular descriptors alone increases the rigidifying, molecular ordering, condensing effect, and/or raft promoting/stabilizing ability of steroids on membranes in relation to that of steroid-free membranes. These findings match the general depiction for the fit of steroids in bilayers. A substantial molecular

Table 4
Validation of the two-descriptor logistic regression model. A training set of 150 cases randomly chosen from the entire data set (212 cases) was used to construct a model. A case corresponds to each time a membrane activity, the same or different, of a given steroid found in the literature. The logistic regression model was next used to predict the activity of the remaining 62 samples (test set), which were not employed in the model building, as described in Section 3.5.3, Model validation. This process was repeated ten times and thus ten different logistic regression models were constructed from ten different training sets. The membrane activity of steroids in each training set (upper panel) and test set (lower panel) was predicted. Values in the table represent the average and the standard deviation of OR and the percentage of cases correctly predicted for the logistic regression in these ten stages. Similarly as in Table 3, "% Corr." refers to the percentage of cases with known membrane activity correctly predicted by the logistic regression model, discriminated among disrupters ("D") and ("P"). "Overall" represents the overall percentage of cases correctly predicted by the model, including both promoters and disrupters.

		Average	SD
Training set (N = 150)			
OR	Area	**1.47**	0.14
	Log P	**0.89**	0.03
% corr.	D	78.3	5.7
	P	96.1	2.2
	Overall	**91.0**	**1.4**
Test set (N = 62)			
% corr.	D	80.6	11.5
	P	93.8	4.8
	Overall	**90.2**	**2.7**

area seems to be required for a proper interaction with the neighboring phospholipids through Van der Waals forces, as it increases with the size of the non-polar area of an amphiphilic molecule. The same rationale can be used to explain the positive correlation between the promoting activity of steroids and volume and mass, seeing that both properties are expected to increase concurrently with the increase in the molecular area. The molar refractivity, which depends on the molar volume and density of the molecule, is a measure of the steric volume. It often shows a high association with binding and interaction phenomena, and higher values are linked to large London forces and dipole–dipole interaction. The positive influence of log P on the promoting activity of steroids is not surprising either, since it is well-known that molecules with augmented hydrophobicity often fit better in phospholipid bilayers, as it was previously reported for some particular steroids [27]. Since rotatable bonds are infrequent in the four-ring system of steroids, its number informs mainly on the length and branching of the alkyl side-chain at C17. Linear alkyl chains possess more rotatable bonds than branched alkyl chains. Rotatable bonds increase the flexibility of the chain and may thus fit better in the phospholipid matrix. The positive correlation observed between the promoting activity of steroids and the number of rotatable bounds agrees with a previous work [30], which claims that an 8–10 carbon isoalkyl side-chain at C17 is a very common structural trait in those steroids having rigidifying, molecular ordering/condensing effects and/or a raft promoting ability in membranes. Most of the rotatable bonds of a steroid are those belonging to the side chain at C17. The augmented polarizability (a measure of the ease with which the electron cloud of the molecule can be distorted by an applied electric field) found for steroids with a high degree of interaction with membrane phospholipids (promoters) concurs with the classic steroid anchoring, *i.e.*, with its polar headgroup toward the outer polar surroundings and its non-polar tail toward the inner side of the bilayer.

The relationship between steroid molecular properties and their effects on the physical properties of membranes has been investigated by means of multivariate approaches. In order to ascertain the inter-study variability, this work has contemplated the discrepancies found in the literature by including all cases in spite of the type of reported activity. Consequently, findings represent average tendencies in the complex property–activity relationship of steroids in membranes. Conclusions should be assumed as a general overview of the phenomenon, and may not agree with some reported cases, including some of those employed here (Supplementary Table 1). Besides the molecular properties of steroids, several additional factors may affect the physical state of bilayers, as steroid concentration in the bilayer, solubility, phospholipid composition, steroid depth and tilt in the bilayer, *etc.* [44]. The variation associated with these changing factors was also captured in the present work; it was modeled and subsequently reflected in the outcome of the analysis by means of the parameters of goodness of fit. From the examination of 245 molecular properties only a few were required for an acceptable estimation and prediction of the activity of steroids on membranes. The exclusion of the rest of the descriptors from the models means that they do not add additional information in relation to that provided by the selected ones, nor do they improve the model performance substantially.

The picture emerging from this study is that the area and log P are useful to estimate the activity of steroids on membranes, and that both molecular properties were found as truthful predictors of such activity. The other descriptors inspected, such as volume, mass, refractivity, number of rotatable bonds and polarizability were also found to be closely correlated with such activity. However, and due to a significant degree of correlation among them, they are not strictly necessary for predicting purposes. On the basis of the current definition of membrane activity, an increase in any of the mentioned molecular properties of steroids can be translated to an increase in the rigidity, molecular ordering, packing, and/or raft formation/stabilization of the steroid-containing bilayers.

As well as the contribution to the understanding of the molecular property–membrane activity relationship of steroids in membranes, knowing the influence of the molecular properties could be useful in those fields where the activity of a number of molecules want to be examined. It can help in experimental and computer-assisted design projects aimed at synthesizing or modeling a prototype with a required effect on the membrane biophysical state. The knowledge of the influence of a molecular property could also be useful as a guide for possible structural modification in the lab. By knowing the area and the log P of a steroid, a first depiction of the membrane activity could be accomplished with the proposed model, and the amount of confirmatory experiments can then be minimized to that of the promising candidates.

Much work remains to be done regarding the steroid–phospholipid interaction itself, considering steric issues, H-bondings, van der Waals forces, *etc.* As they are the underlying factors ruling the phenomenon, our knowledge of these factors can supply a detailed insight into the complex steroid/lipid interactions.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.bbamem.2015.07.017.

Transparency document

The Transparency document associated with this article can be found, in the online version.

Financial and competing interest disclosure

No financial support, grants or writing assistance was utilized in the production of this work.

Acknowledgment

I thank Daniel Enriz for helpful comments on the revised manuscript.

References

[1] A. Rietveld, K. Simons, The differential miscibility of lipids as the basis for the formation of functional membrane rafts, Biochim. Biophys. Acta 1376 (1998) 467–479.
[2] D.A. Brown, E. London, Functions of lipid rafts in biological membranes, Annu. Rev. Cell Dev. Biol. 14 (1998) 111–136.
[3] K. Simons, E. Ikonen, How cells handle cholesterol, Science 290 (2000) 1721–1726.
[4] S. Heino, S. Lusa, P. Somerharju, C. Ehnholm, V.M. Olkkonen, E. Ikonen, Dissecting the role of the Golgi complex and lipid rafts in biosynthetic transport of cholesterol to the cell surface, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 8375–8380.
[5] J. Herreros, T. Ng, G. Schiavo, Lipid rafts act as specialized domains for tetanus toxin binding and internalization into neurons, Mol. Biol. Cell 12 (2001) 2947–2960.
[6] D.H. Nguyen, D. Taub, CXCR4 function requires membrane cholesterol: implications for HIV infection, J. Immunol. 168 (2002) 4121–4126.
[7] W. Popik, T.M. Alce, W.C. Au, Human immunodeficiency virus type 1 uses lipid raft-colocalized CD4 and chemokine receptors for productive entry into CD4(+) T cells, J. Virol. 76 (2002) 4709–4722.
[8] F.R. Maxfield, I. Tabas, Role of cholesterol and lipid organization in disease, Nature 438 (2005) 612–621.
[9] T. Rog, M. Pasenkiewicz-Gierula, I. Vattulainen, M. Karttunen, Ordering effects of cholesterol and its analogues, Biochim. Biophys. Acta 1788 (2009) 97–121.
[10] S.J. Singer, G.L. Nicolson, The fluid mosaic model of the structure of cell membranes, Science 175 (1972) 720–731.
[11] D.A. Brown, J.K. Rose, Sorting of GPI-anchored proteins to glycolipid-enriched membrane subdomains during transport to the apical cell surface 3, Cell 68 (1992) 533–544.
[12] K. Simons, E. Ikonen, Functional rafts in cell membranes, Nature 387 (1997) 569–572.
[13] D.J. Recktenwald, H.M. McConnell, Phase equilibria in binary mixtures of phosphatidylcholine and cholesterol, Biochemistry 20 (1981) 4505–4510.
[14] J.H. Ipsen, G. Karlstrom, O.G. Mouritsen, H. Wennerstrom, M.J. Zuckermann, Phase equilibria in the phosphatidylcholine–cholesterol system, Biochim. Biophys. Acta 905 (1987) 162–172.
[15] M.B. Sankaram, T.E. Thompson, Interaction of cholesterol with various glycerophospholipids and sphingomyelin, Biochemistry 29 (1990) 10670–10675.
[16] Y. Barenholz, Cholesterol and other membrane active sterols: from membrane evolution to "rafts", Prog. Lipid Res. 41 (2002) 1–5.
[17] R.A. Demel, K.R. Bruckdorfer, L.L. van Deenen, Structural requirements of sterols for the interaction with lecithin at the air–water interface, Biochim. Biophys. Acta 255 (1972) 311–320.

[18] P.L. Yeagle, R.B. Martin, A.K. Lala, H.K. Lin, K. Bloch, Differential effects of cholesterol and lanosterol on artificial membranes, Proc. Natl. Acad. Sci. U. S. A. 74 (1977) 4924–4926.
[19] K.W. Butler, I.C. Smith, Sterol ordering effects and permeability regulation in phosphatidylcholine bilayers. A comparison of ESR spin-probe data from oriented multilamellae and dispersions, Can. J. Biochem. 56 (1978) 117–122.
[20] J. Rogers, A.G. Lee, D.C. Wilton, The organisation of cholesterol and ergosterol in lipid bilayers based on studies using non-perturbing fluorescent sterol probes, Biochim. Biophys. Acta 552 (1979) 23–37.
[21] K.E. Bloch, Sterol structure and membrane function 1, CRC Crit Rev. Biochem. 14 (1983) 47–92.
[22] J.A. Urbina, S. Pekerar, H.B. Le, J. Patterson, B. Montez, E. Oldfield, Molecular order and dynamics of phosphatidylcholine bilayer membranes in the presence of cholesterol, ergosterol and lanosterol: a comparative study using 2H-, 13C- and 31P-NMR spectroscopy, Biochim. Biophys. Acta 1238 (1995) 163–176.
[23] X. Xu, E. London, The effect of sterol structure on membrane lipid domains reveals how cholesterol can induce lipid domain formation, Biochemistry 39 (2000) 843–849.
[24] A.B. Serfis, S. Brancato, S.J. Fliesler, Comparative behavior of sterols in phosphatidylcholine-sterol monolayer films, Biochim. Biophys. Acta 1511 (2001) 341–348.
[25] G.V. Martinez, E.M. Dykstra, S. Lope-Piedrafita, M.F. Brown, Lanosterol and cholesterol-induced variations in bilayer elasticity probed by 2H NMR relaxation, Langmuir 20 (2004) 1043–1046.
[26] J. Wang, Megha, E. London, Relationship between sterol/steroid structure and participation in ordered lipid domains (lipid rafts): implications for lipid raft structure and function, Biochemistry 43 (2004) 1010–1018.
[27] J.J. Wenz, F.J. Barrantes, Steroid structural requirements for stabilizing or disrupting lipid domains, Biochemistry 42 (2003) 14267–14276.
[28] G. Oradd, V. Shahedi, G. Lindblom, Effect of sterol structure on the bending rigidity of lipid membranes: a (2)H NMR transverse relaxation study, Biochim. Biophys. Acta 1788 (2009) 1762–1771.
[29] D.A. Mannock, R.N. Lewis, T.P. McMullen, R.N. McElhaney, The effect of variations in phospholipid and sterol structure on the nature of lipid–sterol interactions in lipid bilayer model membranes, Chem. Phys. Lipids 163 (2010) 403–448.
[30] J.J. Wenz, Predicting the effect of steroids on membrane biophysical properties based on the molecular structure, Biochim. Biophys. Acta 1818 (2012) 896–906.
[31] MOPAC2009, J.P. Stewart, Stewart Computational Chemistry, Colorado Springs, CO, USA, 2008.
[32] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.P. Stewart, Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model, J. Am. Chem. Soc. 107 (1985) 3902–3909.
[33] Dragon, Talete SRL, Milano Chemometrics and QSAR Research Group, Milano, Italy, in, 2007.
[34] R. Todeschini, V. Consonni, Handbook of molecular descriptors, in: R. Mannhold, H. Kubinyi, H. Timmerman (Eds.), Methods and Principles in Medicinal Chemistry, Wiley-VCH, Weinheim, Germany, 2000.
[35] K.R. Beeve, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, John Wiley & Sons, New York, 1998.
[36] R.G. Brereton, Chemometrics: Data Analysis for the Laboratory and Chemical Plant, John Wiley & Sons, England, 2003.
[37] The Unscrambler, CAMO Process AS, Oslo, Norway, http://www.camo.no2007.
[38] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, John, Wiley & Sons Inc., New York, 1989.
[39] D.G. Kleinbaum, Logistic Regression: A Self-learning Text, Springer-Verlag, New York, 1994.
[40] SPSS Statistic, IBM Corporation, Somers, NY, USA, 2008.
[41] C.J. Peng, K.L. Lee, G.M. Ingersoll, An introduction to logistic regression analysis and reporting, J. Educ. Res. 96 (2002) 3–13.
[42] V. Ben-Yashar, Y. Barenholz, The interaction of cholesterol and cholest-4-en-3-one with dipalmitoylphosphatidylcholine. Comparison based on the use of three fluorophores 1, Biochim. Biophys. Acta 985 (1989) 271–278.
[43] M.E. Beattie, S.L. Veatch, B.L. Stottrup, S.L. Keller, Sterol structure determines miscibility versus melting transitions in lipid vesicles 1, Biophys. J. 89 (2005) 1760–1768.
[44] G. Khelashvili, D. Harries, How cholesterol tilt modulates the mechanical properties of saturated and unsaturated lipid membranes, J. Phys. Chem. B 117 (2013) 2411–2421.