A Steroids QSAR Approach Based on Approximate Similarity Measurements

Manuel Urbano Cuadrado,[†] Irene Luque Ruiz,*[,‡] and Miguel Ángel Gómez-Nieto[‡]

Institute of Chemical Research of Catalonia ICIQ, Avinguda Països Catalans 16, E-43007 Tarragona, Spain, and Department of Computing and Numerical Analysis, University of Córdoba, Campus Universitario de Rabanales, Albert Einstein Building, E-14071 Córdoba, Spain

Received February 15, 2006

A new QSAR method based on *approximate similarity* measurements is described in this paper. Approximate similarity is calculated using both the classical similarity based on the graph isomorphism and a distance computation between nonisomorphic subgraphs. The latter is carried out through a parametric function where different topological invariants can be considered. After optimizing the contribution of nonisomorphic distance to the new graph similarity, predictive models built with approximate similarity matrixes show higher predictive ability than those using traditional similarity matrixes. The new method has been applied to the prediction of steroids binding to the corticosteroid globulin receptor. The proposed model allows us to obtain valuable external predictions ($r = 0.82$ and SEP $= 0.30$) after training the model by cross-validation ($Q^2 = 0.84$ and SECV $= 0.47$). Slope and bias parameters are also given.

1. INTRODUCTION

The biological activity of a drug usually results from a noncovalent interaction between a macromolecular biological receptor and a compound, usually small, that acts as the drug. Calculation of the receptor−drug interaction strength constitutes one of the stages that compose drug design methods, and it is directly related to efficiency. Rational drug design seeks mathematical relationships between molecular structures (2D or 3D conformations) and biological activities. This methodology, known as QSAR (quantitative structure−activity relationship), is employed to establish models able to determine new compounds' activities.[1−3]

Molecular descriptors, which constitute pieces of information from chemical structures, are the variables employed as both modelers and predictors.[4,5] These descriptors account for characteristics of either a molecule part or all of the structure (local or global descriptors, respectively). Descriptors can be also classified into 2D and 3D descriptors depending on the space dimensions that generate the chemical information. Because good predictions are not often achieved by a single descriptor, equations that use arrays of descriptors and multivariate training techniques—multiple linear regression, partial least-squares regression (PLSR), artificial neural networks, and so forth[6−8]—are, most times, the summarized approaches in recent QSAR literature.

Three-dimensional methodologies—comparative molecular field analysis (CoMFA) and related approaches,[9−13] comparative molecular moment analysis,[14,15] grid-weighted holistic invariant molecular descriptors,[16,17] and so forth—have become very popular in rational drug design because of their ability to describe how biological receptors feel electrostatic and steric characteristics of molecules. Despite this fact, some shortcomings are involved in these methodologies. Structural information of the receptor is necessary when an alignment step is presented in CoMFA-like methods. In addition, if the data set is composed by very flexible molecules, the possibility of achieving a single structural alignment is impossible. Another shortcoming can be considered the 3D optimal conformation search because of the complexity of the optimization methods based on molecular mechanics, molecular dynamics, density functional theory, and so forth.

Two-dimensional topological descriptors imply neither the geometrical optimization nor the structural alignment. Thus, the development of 2D-QSAR approaches is a less complex process and less subjective than when 3D methodologies are involved. Using the graph theory, topological organic chemistry has been developed by Schultz et al. in a series of related papers.[18] Since then, a large series of topological descriptors has been developed and employed in structure−activity relationships.[19−21]

Also derived from graphs, structural similarity measurements have been widely used in computational chemistry. Similarity calculation algorithms support many approaches for both screening chemical databases and predicting physical−chemical properties.[22] In recent years, methods that correlate the structural similarity with their properties have been proposed on the basis of the following chemical principle: "*structurally similar molecules show similar properties and biological activities*".[23,24]

From the point of view of multivariate calibration, a similarity matrix **S** composed by $N \times N$ structural similarity calculations can be considered a predictive space—consisting of n objects (molecular graphs) with n variables (also molecular graphs) that represent the similarity value between the object graph i and the variable graph j. However, if structural isomers are considered, classical 2D similarity calculations yield inconsistencies, as can be seen in Figure 1. When we calculate structural similarity between the *A*, *B*, and *C* molecular graphs, the isomorphism consisting of 20 vertexes and 23 edges is equal for the three graphs. The

* Corresponding author phone: +34-957-21-2082; fax: +34-957-21-8630; e-mail: ma1lurui@uco.es.
[†] Institute of Chemical Research of Catalonia ICIQ.
[‡] University of Córdoba.

10.1021/ci0600511 CCC: $33.50
Published on Web 05/19/2006

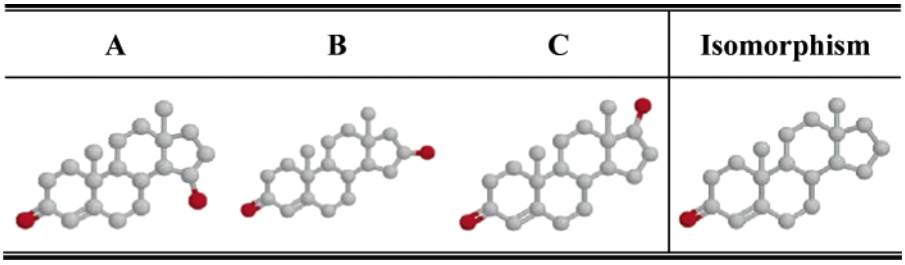

Figure 1. Example of graphs showing equal isomorphism. However, molecules A, B, and C show different properties.

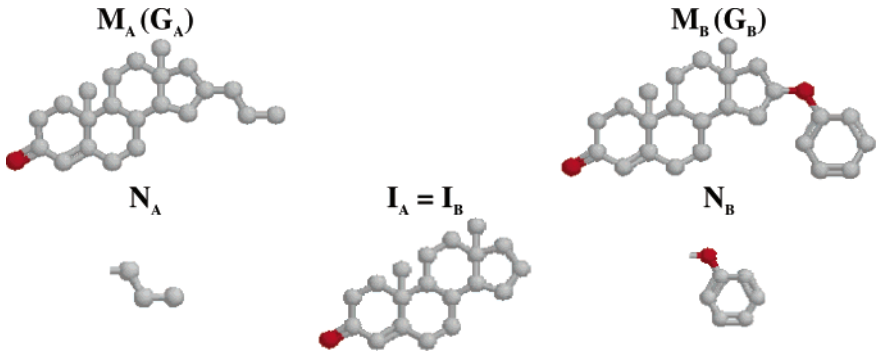

Figure 2. Graphical description of molecules M_A and M_B with sizes A and B. These compounds show the isomorphism $I_{A,B} = I_A = I_B$. N_A and N_B represent the subgraphs not present in the isomorphism.

similarity between any pair of the three molecular graphs is therefore equal ($S_{A,B} = S_{A,C} = S_{B,C}$). Nevertheless, properties of the molecules represented by the *A*, *B*, and *C* molecular graphs are different, thus leading to low correlations.

To overcome the shortcoming commented on above, we proposed a new way[25] based on differences between nonisomorphic subgraphs to obtain a comparison index between two molecular graphs. With this aim, Wiener, Hyper Wiener, and Valence Overall Wiener (VOW) descriptors[4,22] were employed for describing the nonisomorphic subgraphs, and Euclidean distance was the metric technique which quantified the topological differences. This method was the basis of a quantitative structure−property relationship model for predicting sublimation enthalpies of polychlorinated biphenyls (PCBs). Despite achieving good correlation, this method is not fine enough because of the nonconsideration of similarity—the main reason for its correct functioning for PCBs is the low differences between their similarity values. Thus, a similarity index must be considered in the new graph comparison.

In this article, we define a new topological index, *approximate similarity* (AS), with the aim of refining the similarity predictive space in structure−activity relationships using both similarity and distance measurements. The method proposed is validated through the prediction of activities of a set of steroids that has classically been the testing benchmark in new QSAR developments.[13,14,26]

2. DEFINING THE APPROXIMATE SIMILARITY (AS) INDEX

2.1. Meaning of the Similarity and Distance Concepts in Graph Comparison. Given two graphs G_A and G_B of sizes (number of nodes and edges) A and B, respectively, which represent, as shown in Figure 2, the molecules M_A and M_B, we define $I_{A,B}$ as the isomorphism present in these graphs and N_A and N_B as the G_A and G_B of the noncommon parts (nonisomorphic subgraphs). The structural similarity can be calculated as follows:

$$S_{A,B} = f(I_{A,B},A,B) \quad (1)$$

where f is a function which matches $S_{A,B}$ and $I_{A,B}$ taking into account the sizes of graphs A and B. The similarity $S_{A,B}$ is a value within the range [0,1] that gives the similarity between graphs G_A and G_B; thus, the closer to 1 the $S_{A,B}$ becomes, the higher the similarity between the molecules M_A and M_B is.

Different similarity values are obtained depending on both the method employed for calculating the isomorphism and the $f(x)$ similarity function (similarity index considered). Regarding the isomorphism calculation, the MCES (maximum common edges subgraph), MCS (maximum common subgraph), and AMCS (all maximum common subgraphs) approaches, in addition to the methods based on transforming graphs into fingerprints, are the most common methodologies.[27,28] And in regard to the similarity function, there are several similarity indexes summarized in the literature whose difference lies in the function, namely, Tanimoto, Cosine, Simpson, Raymond, and so forth.[28]

Because our proposal is also to consider distances between the subgraphs that do not form the isomorphism $I_{A,B}$, the structural difference $\Gamma_{A,B}$ (dissimilarity or distance) between two molecular graphs G_A and G_B is calculated as follows:

$$\Gamma_{A,B} = g[\mathrm{td}(G_A,I_{A,B}),\mathrm{td}(G_B,I_{A,B})] = g[\mathrm{td}(N_A),\mathrm{td}(N_B)] \quad (2)$$

where $I_{A,B}$ has the same meaning as that in expression 1, $N_A = G_A - I_{A,B}$ and $N_B = G_B - I_{A,B}$ represent the subgraphs of G_A and G_B, respectively, that do not form the isomorphism $I_{A,B}$, $g(x)$ is a function aimed to obtain a distance value (e.g., Euclidean, Mahalanobis, etc.) between td($G_A - I_{A,B}$) and td($G_B - I_{A,B}$), and td is a topological descriptor which describes the noncommon subgraphs, namely, Wiener, Hyper Wiener, VOW, and so forth indexes.[4,22,25]

Contrary to similarity, the higher the $\Gamma_{A,B}$ is, the higher the dissimilarity between the molecules M_A and M_B is. In addition, $\Gamma_{A,B}$ shows a value much higher than 1.

2.2. Refinement of the Structural Similarity: The Approximate Similarity (AS) Index. As stated above, refinement of the similarity measurement is necessary in order to augment the correlation between molecular topologies and activities. Aimed at removing the inconsistencies shown by similarity values, we have defined a new index AS, which takes into account both the graph isomorphism and the subgraph dissimilarity for two compounds. On the basis of simple mathematical combinations, the approximate structural similarity is calculated as follows:

$$\mathrm{AS}_{A,B} = S_{A,B} - w_\Gamma \bar{\Gamma}_{A,B} \tag{3}$$

where $S_{A,B}$ is the structural similarity defined in expression 1, $\bar{\Gamma}_{A,B}$ is a scaled value of the structural dissimilarity $\Gamma_{A,B}$ defined in expression 2, and w_Γ is a weighting factor that gives a different significance to the dissimilarity values.

The scaling of the Γ values is necessary to provide dissimilarities within the range [0,1]. As stated above, Γ shows values much higher than the similarity, which is always within the range [0,1] by definition, and this makes it necessary to equalize both scales. Several normalization methods have been tested, leading the maximum method to the best results. Data normalization is carried out by the following expression:

$$\forall n, \bar{\Gamma}(i,j) = \frac{\Gamma(i,j)}{\max[\Gamma(n,j)]} \tag{4}$$

The similarity concept is enhanced by expression 3: the fact of considering noncommon subgraphs distances in similarity calculation leads to more precise values than when only classical structural similarity is taken into account. Thus, differences in molecular topologies refine the structural similarity by means of subtracting a distance quantity to the isomorphism measurement. The influence of distances on the new similarity calculation depends on the w_Γ weight choice. To achieve a good predictive ability, it is necessary to set an appropriate weight, which depends on the chemical space we are modeling. For example, a set of compounds which show very different similarity values will require low values for w_Γ because of the high descriptive power available using only similarity. Contrarily, if compounds show similar common parts, the significance of the differences of the noncommon subgraphs should be higher.

3. BUILDING A MULTIVARIATE PREDICTIVE SPACE: APPROXIMATE SIMILARITY MATRIXES AND PARTIAL LEAST-SQUARES REGRESSION

3.1. Approximate Similarity Matrixes: A Multivariate Descriptor Space. In the introductory section, it was stated that using arrays of pieces of chemical information usually gives QSAR models with better predictive and modeling characteristics than those shown by single-variable approaches. The AS values are scalars, which give information about both the global similarity and dissimilarity between two compounds. If a set of compounds is compared with a single reference (the same compound), the QSAR approaches can only be considered as single-descriptor models.

If an $N \times N$ structural AS matrix is built using N compounds, this AS matrix can be employed to develop multivariate QSAR approaches. Each element $\mathrm{AS}_{i,j}$ provides the approximate similarity between the compounds i and j and shows the same value as the element $\mathrm{AS}_{j,i}$. The greater the differences between molecules are, the closer to 0 the $\mathrm{AS}_{i,j}$ value is. The diagonal of the matrix (elements $\mathrm{AS}_{i,i}$) is equal to 1. From the point of view of multivariate regression, the AS matrix is considered a set of N objects (rows) characterized by N variables (columns). Thus, an object is a given compound described by a series of global variables which account for the similarity between the compound and a reference compound.

3.2. PLS Regression for Reducing the Similarity Predictive Space. PLSR[7] allows us to reduce the original data space. In the reduced space, it is most times easier to visualize both trends and influences of the original variables on properties than when the original space is considered. Thus, the study of number, type, and characteristics of the PLS factors provides scientists with structured information of their multivariate systems.

In addition, PLSR considers the variance of both predictors and properties in the building of the reduced space. Thus, this construction leads to better correlations between AS data and properties. Other techniques also based on the reduction of the variables only take into account the predictor variance. For example, principal components regression retains the relevant factors which explain the predictor set.

4. APPLICATION: THEORETICAL STUDY OF STEROID BINDING TO THE CORTICOSTEROID-BINDING GLOBULIN RECEPTOR

The steroids studied in the first original CoMFA application have provided a benchmark for testing numerous subsequent structure−activity relationships,[13,14,26] and therefore, their theoretical study using the approximate similarity measurement can be considered as a valid test because of the high number of standards. The 31 steroids shown in Figure 3 were divided into training and test sets, **1**−**21** and **22**−**30**, respectively. Compound **31** has been traditionally considered as an outlier, and it was also removed from the test set. The steroid nomenclature and the observed log K values for the corticosteroid-binding affinity are shown in Table 1.

Tables 2 and 3 show the similarity (S) and distance (Γ) matrixes, respectively, calculated for compounds **1**−**30**. Similarities were computed using both the MCES graph isomorphism algorithm and the cosine index, and dissimilarities were calculated using both the Euclidean distance and the VOW topological descriptor. As described in section 2, dissimilarity values are much higher than similarity values, and this is in addition to the contrary sense of their scales. Thus, the diagonals of the similarity and distance matrixes (each molecule compared with itself) are 1 and 0, respectively.

Software employed for computing similarity, distance, and approximate similarity was developed by Cerruela García et al.[27] in the C programming language.

4.1. Model Training: Evaluation of Precision and Weights Settings. Cross-validation, also known as internal validation, was employed for the training stage. Final models

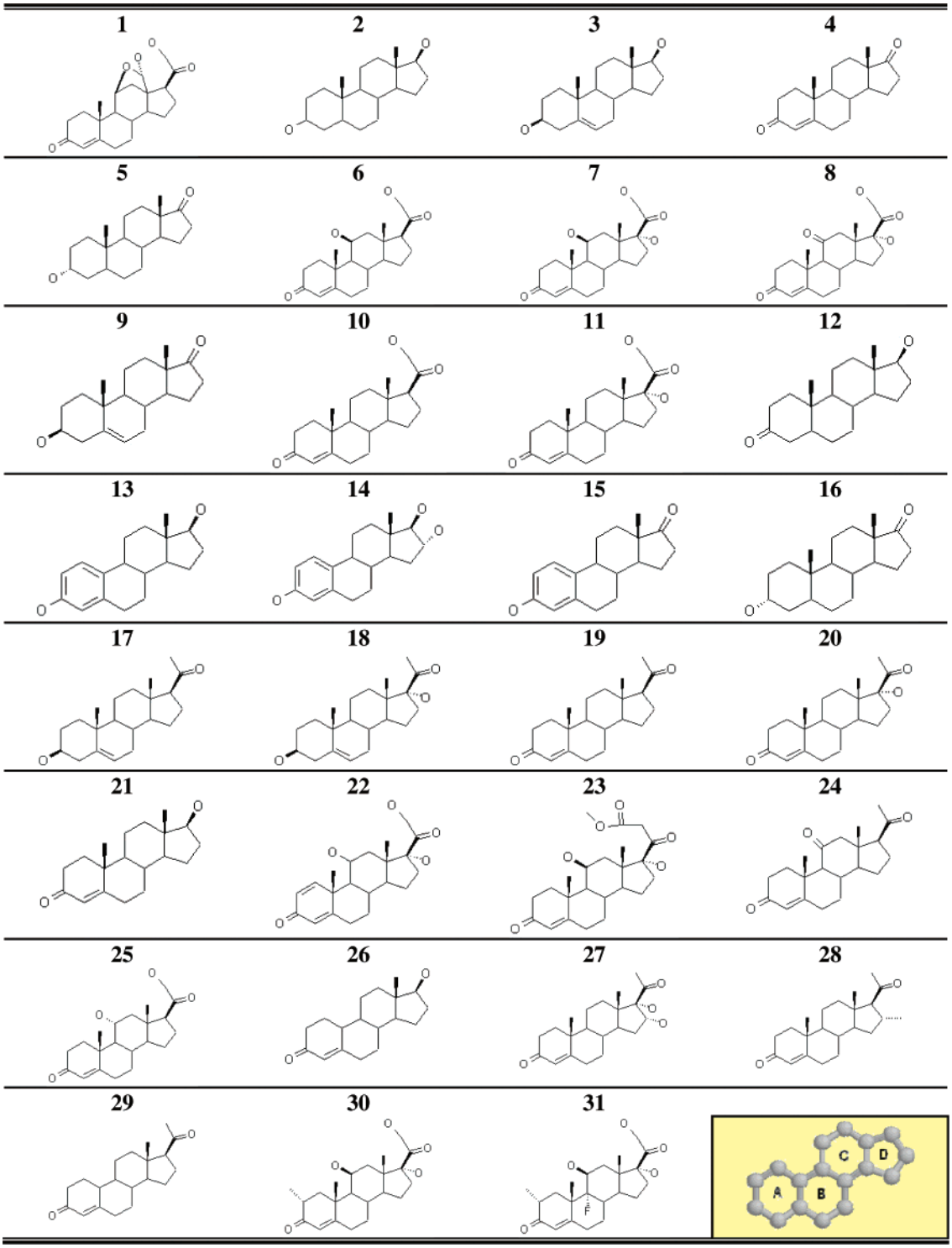

Figure 3. Steroids employed in this QSAR study. Ring nomenclature is also shown in the last picture.

were the average of the six individual models (cross-validation segments), which assured both 4:1 training−internal test ratios and the inclusion of all the samples at least once in the internal test sets. Table 4 shows the statistical parameters—Q^2 (determination coefficient in cross-validation), SECV (standard error in cross-validation), slope, and bias—obtained in this stage for the similarity, distance, and approximate similarity matrixes. All of the parameters refer to the predictive ability and not to fitting. It is stated in QSAR literature[29] that a statistically meaningful model is achieved when $Q^2 > 0.5$. Then, all of the models summarized in Table 4 can be considered as predictive. Despite this criterion, other more restrictive criteria have been proposed in chemometrics.

Shenk and Westerhaus[30] proposed that $Q^2 > 0.90$ indicates excellent precision. If Q^2 values are between 0.7 and 0.90, that would mean good precision. On the other hand, $Q^2 < 0.70$ indicates that the equation can only be used for screening purposes, which enable distinction between low, medium, and high values for the measured parameter. Finally, if $Q^2 < 0.50$, the equation only discriminates between high and low values. According to this criterion, all of the models

Table 1. Set of the 31 Steroids' Binding Data to the Corticosteroid-Binding Globulin Receptor

number	steroid	observed log *K*
1	aldosterone	6.28
2	androstanediol	5.00
3	5-androstanediol	5.00
4	4-androstenedione	5.76
5	androsterone	5.61
6	corticosterone	7.88
7	cortisol	7.88
8	cortisone	6.89
9	dehydrotestosterone	5.00
10	11-deoxycorticosterone	7.65
11	11-deoxycortisol	7.88
12	dihydrotestosterone	5.92
13	estradiol	5.00
14	estriol	5.00
15	estrone	5.00
16	etiocholanolone	5.23
17	pregnenolone	5.23
18	17a-hydroxypregnenolone	5.00
19	progesterone	7.38
20	17a-hydroxyprogesterone	7.74
21	testosterone	6.72
22	prednisolone	7.51
23	cortisolacetat	7.55
24	4-pregenene-3,11,20-trione	6.78
25	epicorticosterone	7.20
26	19-nortestosterone	6.14
27	16a,17a-dihydroxyprogesterone	6.25
28	17a-methylprogesterone	7.12
29	19-norprogesterone	6.82
30	2a-methylcortisol	7.69
31	2a,9a-methylfluorocortisol	5.80

show good precision with the exception of that built using distances. This fact justifies the search of an optimal weight w_Γ for combining structural similarities and distances. An excessive significance of distances in AS calculations can add random information and, then, no deterministic part to the total predictive ability. In this study, the factor w_Γ was moved from 1.0 to 0.1 and important variations were obtained.

When w_Γ was set to 0.3, the predictive ability in internal validation was maximal. It was also observed that compound **1** was considered as an object−activity outlier—student parameter (activity residual/SECV) cutoff set to 2.5—in the training stage for the AS models with w_Γ 0.3 and 0.5. As can be seen in Figure 1, its structure shows a significant difference in comparison to the rest of compounds: an epoxy substructure. Therefore, this characteristic was not modeled by the similarity space, thus, not allowing us to correlate correctly with the activity. The subsequent validation stage confirmed that the weight w_Γ 0.3 led to the most robust AS model.

4.2. Model Validation: Study of the Robustness. The test of the predictive ability of equations, also known as validation, is a crucial step in the QSAR methodology. External validation, based on the use of new objects, is the most recommendable method if the split of the compounds into training and test subsets does not involve the use of such a low number of objects that the predictive ability cannot be modeled.

Here, the applicability of models built using the **1**−**21** steroids subset was studied. Table 5 shows the statistical parameters—*r* (determination coefficient in prediction), SEP (standard error in prediction), slope, and bias—obtained in the external validation stage. The best predictive ability was achieved when approximate similarity with $w_\Gamma = 0.3$ was considered. In this case, a significant improvement in the coefficient of determination value was observed. Table 6 shows the activities for the test set predicted with the $w_\Gamma = 0.3$ AS model.

Taking into account the SEP value, it is also accepted by the scientific community that the limit for considering equations as robust tools is 1.5 × SECV. All of the models were within this range, and therefore, robust equations were achieved. It has to be remarked that SEP < SECV was obtained for almost all of the equations. This is not logical and could point to overfitting problems. But, both the low number of PLS factors (see Table 4) and the method for setting the optimal number of these factors (selection of the factor showing the minimum SECV value) do not indicate that overfitting is the reason for obtaining SEP < SECV. Likely, this can be due to both SECV being an error parameter obtained in internal predictions (with chemometrically appropriate sizes for test segments) and the low number of compounds employed in the external test set (**21**−**30**).

4.3. Interpretation of Equations Coefficients. Figure 4 shows the regression coefficients for the model built with approximate similarity indexes ($w_\Gamma = 0.3$). The minimum value was obtained for the variable **18**, whose reference compound is 17a-hydroxypregnenolone (also number **18** in Figure 3). The observed log *K* for this compound is 5.00, the lowest activity value summarized in Table 1. The higher the similarity between 17a-hydroxypregnenolone and a new compound is, the lower the activity that will be shown by the new compound. Thus, this logical fact has been corroborated by the mathematical equation. Other compounds showing activities equal to 5.0 are also references of the variables with the lowest coefficients (e.g., androstanediol, 5-androstanediol, dehydrotestosterone, estradiol, estriol, and estrone). The common characteristic is the hydroxyl group present in the ring A (see Figure 3) of the steroid structure. Therefore, the design of efficient steroids has to avoid the presence of this group.

On the other hand, compounds that show the highest activity values (corticosterone, cortisol, 11-deoxycorticosterone, and 11-deoxycortisol) correspond to coefficients with high positive values. The common substructures present in these compounds are the carboxylic acid and the ketonic group in rings D and A, respectively.

4.4. Comparison with Recent QSAR Methods Employed for the Study of Steroids. As commented upon above, the steroids studied in this paper comprise the traditional set that new QSAR approaches have employed for their test stages. Our results compare reasonably well with other QSAR models. Coats[26] summarized 14 3D-QSAR methods with values of Q^2_{LOO} (full cross-validation) ranging between 0.23 and 0.93 (average $Q^2_{LOO} = 0.71$). Our best Q^2 value was 0.84 for the training set, and the use of internal test sets composed of more than 1 could be considered as more robust than LOO.

Regarding external validation, if the AS model is compared with other recent methods[13,14] that have employed the external set (compounds **22**−**30**), reasonable results (better and similar statistical values, in addition to not considering outliers in the external set) are also obtained. It should be

Table 2. Similarity Values Using the Cosine Index for Steroids **1**−**30**

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
1	1.00	0.80	0.78	0.86	0.80	0.97	0.95	0.92	0.78	0.95	0.94	0.84	0.59	0.62	0.59	0.80	0.86	0.84	0.94	0.92	0.86	0.94	0.87	0.92	0.97	0.84	0.90	0.92	0.92	0.94
2	0.80	1.00	0.98	0.89	0.96	0.82	0.84	0.84	0.93	0.83	0.86	0.96	0.64	0.62	0.64	0.96	0.89	0.92	0.85	0.88	0.93	0.70	0.80	0.83	0.82	0.91	0.86	0.83	0.83	0.83
3	0.78	0.98	1.00	0.87	0.93	0.80	0.82	0.82	0.96	0.81	0.84	0.93	0.77	0.76	0.73	0.93	0.92	0.94	0.83	0.86	0.91	0.80	0.78	0.81	0.80	0.89	0.84	0.81	0.80	0.81
4	0.86	0.89	0.87	1.00	0.93	0.88	0.86	0.86	0.91	0.90	0.88	0.93	0.66	0.64	0.70	0.93	0.83	0.81	0.92	0.90	0.96	0.84	0.82	0.90	0.88	0.93	0.88	0.90	0.89	0.85
5	0.80	0.96	0.93	0.93	1.00	0.82	0.80	0.80	0.98	0.83	0.82	0.91	0.64	0.62	0.64	1.00	0.89	0.88	0.85	0.83	0.89	0.70	0.76	0.83	0.82	0.86	0.82	0.83	0.83	0.79
6	0.97	0.82	0.80	0.88	0.82	1.00	0.98	0.94	0.80	0.98	0.96	0.86	0.61	0.63	0.61	0.82	0.88	0.87	0.96	0.94	0.88	0.96	0.90	0.94	1.00	0.86	0.92	0.94	0.94	0.96
7	0.95	0.84	0.82	0.86	0.80	0.98	1.00	0.96	0.78	0.96	0.98	0.88	0.64	0.62	0.60	0.80	0.87	0.89	0.94	0.96	0.90	0.98	0.92	0.93	0.98	0.88	0.94	0.93	0.92	0.98
8	0.92	0.84	0.82	0.86	0.80	0.94	0.96	1.00	0.78	0.96	0.98	0.88	0.64	0.62	0.60	0.80	0.87	0.89	0.94	0.96	0.90	0.95	0.88	0.96	0.94	0.88	0.94	0.93	0.92	0.95
9	0.78	0.93	0.96	0.91	0.98	0.80	0.78	0.78	1.00	0.81	0.80	0.89	0.73	0.71	0.77	0.98	0.92	0.90	0.83	0.81	0.87	0.76	0.74	0.81	0.80	0.84	0.80	0.81	0.80	0.77
10	0.95	0.83	0.81	0.9	0.83	0.98	0.96	0.96	0.81	1.00	0.98	0.88	0.62	0.61	0.62	0.83	0.90	0.88	0.98	0.96	0.90	0.94	0.88	0.96	0.98	0.88	0.94	0.96	0.96	0.95
11	0.94	0.86	0.84	0.88	0.82	0.96	0.98	0.98	0.80	0.98	1.00	0.90	0.65	0.63	0.61	0.82	0.88	0.90	0.96	0.98	0.92	0.96	0.90	0.94	0.96	0.90	0.96	0.94	0.94	0.96
12	0.84	0.96	0.93	0.93	0.91	0.86	0.88	0.88	0.89	0.88	0.90	1.00	0.64	0.62	0.64	0.91	0.85	0.88	0.89	0.92	0.98	0.66	0.84	0.88	0.86	0.95	0.90	0.88	0.87	0.87
13	0.63	0.64	0.77	0.68	0.64	0.65	0.64	0.64	0.73	0.64	0.65	0.61	1.00	0.98	0.95	0.64	0.70	0.73	0.65	0.66	0.70	0.72	0.61	0.64	0.65	0.72	0.65	0.64	0.67	0.61
14	0.62	0.64	0.76	0.67	0.64	0.63	0.62	0.62	0.71	0.63	0.63	0.60	0.98	1.00	0.93	0.64	0.68	0.71	0.64	0.65	0.69	0.70	0.59	0.63	0.63	0.70	0.63	0.63	0.65	0.61
15	0.61	0.61	0.73	0.70	0.61	0.63	0.62	0.64	0.77	0.64	0.63	0.64	0.95	0.93	1.00	0.61	0.70	0.68	0.65	0.64	0.68	0.68	0.59	0.66	0.63	0.70	0.63	0.64	0.67	0.61
16	0.80	0.96	0.93	0.93	1.00	0.82	0.80	0.80	0.98	0.83	0.82	0.91	0.64	0.62	0.64	1.00	0.89	0.88	0.85	0.83	0.89	0.70	0.76	0.83	0.82	0.86	0.82	0.83	0.83	0.79
17	0.86	0.89	0.92	0.83	0.89	0.88	0.87	0.87	0.92	0.90	0.88	0.85	0.70	0.68	0.70	0.89	1.00	0.98	0.92	0.90	0.83	0.85	0.82	0.90	0.88	0.81	0.88	0.90	0.90	0.85
18	0.84	0.92	0.94	0.81	0.88	0.87	0.89	0.89	0.90	0.88	0.90	0.88	0.73	0.71	0.68	0.88	0.98	1.00	0.90	0.92	0.86	0.87	0.84	0.88	0.87	0.83	0.90	0.88	0.88	0.87
19	0.94	0.85	0.83	0.92	0.85	0.96	0.94	0.94	0.83	0.98	0.96	0.89	0.63	0.62	0.63	0.85	0.92	0.90	1.00	0.98	0.92	0.92	0.90	0.98	0.96	0.89	0.96	0.98	0.98	0.93
20	0.92	0.88	0.86	0.90	0.83	0.94	0.96	0.96	0.81	0.96	0.98	0.92	0.66	0.65	0.62	0.83	0.90	0.92	0.98	1.00	0.94	0.94	0.91	0.96	0.94	0.92	0.98	0.96	0.96	0.95
21	0.86	0.93	0.91	0.96	0.89	0.88	0.90	0.90	0.87	0.90	0.92	0.98	0.70	0.69	0.66	0.89	0.83	0.86	0.92	0.94	1.00	0.88	0.86	0.90	0.88	0.98	0.92	0.90	0.89	0.89
22	0.94	0.66	0.80	0.84	0.66	0.96	0.98	0.95	0.76	0.94	0.96	0.64	0.72	0.70	0.68	0.66	0.85	0.87	0.92	0.94	0.88	1.00	0.90	0.91	0.96	0.86	0.93	0.91	0.90	0.96
23	0.87	0.80	0.78	0.82	0.76	0.90	0.92	0.88	0.74	0.88	0.90	0.84	0.61	0.59	0.57	0.76	0.82	0.84	0.90	0.91	0.86	0.90	1.00	0.88	0.9	0.84	0.90	0.88	0.88	0.90
24	0.92	0.83	0.81	0.9	0.83	0.94	0.93	0.96	0.81	0.96	0.94	0.88	0.62	0.61	0.62	0.83	0.90	0.88	0.98	0.96	0.90	0.91	0.88	1.00	0.94	0.88	0.94	0.96	0.96	0.91
25	0.97	0.82	0.80	0.88	0.82	1.00	0.98	0.94	0.80	0.98	0.96	0.86	0.61	0.63	0.61	0.82	0.88	0.87	0.96	0.94	0.88	0.96	0.90	0.94	1.00	0.86	0.92	0.94	0.94	0.96
26	0.84	0.91	0.89	0.93	0.86	0.86	0.88	0.88	0.84	0.88	0.90	0.95	0.72	0.70	0.67	0.86	0.81	0.83	0.89	0.92	0.98	0.86	0.84	0.88	0.86	1.00	0.90	0.88	0.91	0.87
27	0.90	0.86	0.84	0.88	0.82	0.92	0.94	0.94	0.80	0.94	0.96	0.90	0.65	0.63	0.61	0.82	0.88	0.90	0.96	0.98	0.92	0.93	0.90	0.94	0.92	0.90	1.00	0.94	0.94	0.93
28	0.92	0.83	0.81	0.90	0.83	0.94	0.93	0.93	0.81	0.96	0.94	0.88	0.62	0.61	0.62	0.83	0.90	0.88	0.98	0.96	0.90	0.91	0.88	0.96	0.94	0.88	0.94	1.00	0.96	0.91
29	0.92	0.83	0.80	0.89	0.83	0.94	0.92	0.92	0.8	0.96	0.94	0.87	0.65	0.63	0.65	0.83	0.90	0.88	0.98	0.96	0.89	0.90	0.88	0.96	0.94	0.91	0.94	0.96	1.00	0.91
30	0.94	0.83	0.81	0.85	0.79	0.96	0.98	0.95	0.77	0.95	0.96	0.87	0.61	0.59	0.61	0.79	0.85	0.87	0.93	0.95	0.89	0.96	0.90	0.91	0.96	0.87	0.93	0.91	0.91	1.00

Table 3. Γ Dissimilarities Using the Euclidean Distance and the VOW Descriptor for the Noncommon Parts of Steroids **1**−**30**

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
1	0	753	730	692	752	90	148	288	729	192	222	632	982	911	981	752	399	429	338	355	692	91	882	350	90	761	407	358	431	283
2	753	0	50	219	141	664	672	668	92	567	566	156	595	662	593	141	316	317	428	417	89	816	1247	514	664	153	533	538	437	822
3	730	50	0	189	92	641	652	648	141	543	546	114	354	420	447	92	361	354	403	396	58	528	1228	489	641	126	513	513	411	802
4	692	219	189	0	89	602	718	714	58	502	613	69	451	518	328	89	404	507	360	465	141	682	1292	448	602	193	581	473	365	868
5	752	141	92	89	0	664	778	774	50	566	674	283	593	660	591	0	315	422	427	529	219	815	1348	513	664	267	642	537	435	926
6	90	664	641	602	664	0	117	238	641	101	151	542	896	826	895	664	310	348	248	271	603	85	861	264	0	671	335	274	341	268
7	148	672	652	718	778	117	0	148	756	219	106	556	900	926	1007	778	426	327	365	257	611	60	751	376	117	688	282	383	458	151
8	288	668	648	714	774	238	148	0	752	215	103	552	897	923	1004	774	422	324	361	253	607	90	871	270	238	684	279	380	455	277
9	729	92	141	58	50	641	756	752	0	542	651	242	446	513	354	50	360	465	402	505	189	627	1327	488	641	239	619	513	410	905
10	192	567	543	502	566	101	219	215	542	0	112	442	800	840	799	566	211	263	147	182	503	182	953	173	101	571	269	187	240	370
11	222	566	546	613	674	151	106	103	651	112	0	450	797	826	905	674	321	224	259	150	504	75	847	274	151	582	191	284	352	257
12	632	156	114	69	283	542	556	552	242	442	450	0	593	660	591	283	451	441	297	301	76	935	1132	387	542	62	418	412	299	706
13	824	589	354	510	587	735	900	897	446	854	797	743	0	100	130	587	661	675	720	652	329	939	1469	803	735	269	765	826	638	1222
14	847	725	420	577	723	761	926	923	513	896	826	811	100	0	213	723	708	712	770	687	394	967	1485	847	761	346	795	869	693	985
15	1034	744	447	328	741	948	1059	848	354	853	959	588	130	213	0	741	660	760	719	818	510	1023	1619	584	948	452	927	825	637	1222
16	752	141	92	89	0	664	778	774	50	566	674	283	593	660	591	0	315	422	427	529	219	815	1348	513	664	267	642	537	435	926
17	399	316	361	404	315	310	426	422	360	211	321	451	661	708	660	315	0	109	82	175	405	284	1000	159	310	469	289	183	162	576
18	429	317	354	507	422	348	327	324	465	263	224	441	675	712	760	422	109	0	177	92	399	183	900	224	348	475	193	241	267	476
19	338	428	403	360	427	248	365	361	402	147	259	297	664	712	663	427	82	177	0	109	361	329	942	91	248	427	226	116	93	516
20	355	417	396	465	529	271	257	253	505	182	150	301	652	687	763	529	175	92	109	0	354	220	833	142	271	431	117	159	202	408
21	692	89	58	141	219	603	611	607	189	503	504	76	329	394	451	219	405	399	361	354	0	577	1187	449	603	77	471	473	366	761
22	91	640	528	682	599	85	60	90	627	182	75	955	939	967	1023	599	284	183	329	220	577	0	704	331	85	656	231	337	423	118
23	882	1247	1228	1292	1348	861	751	871	1327	953	847	1132	1469	1485	1570	1348	1000	900	942	833	1187	704	0	946	861	1265	842	949	1035	796
24	350	514	489	448	513	264	376	270	488	173	274	387	748	790	747	513	159	224	91	142	449	331	946	0	264	516	244	148	184	524
25	90	664	641	602	664	0	117	238	641	101	151	542	896	826	895	664	310	348	248	271	603	85	861	264	0	671	335	274	341	268
26	761	153	126	193	267	671	688	684	239	571	582	62	269	346	393	267	469	475	427	431	77	656	1265	516	671	0	549	541	337	839
27	407	533	513	581	642	335	282	279	619	269	191	418	765	795	874	642	289	193	226	117	471	231	842	244	335	549	0	254	319	424
28	358	538	513	473	537	274	383	380	513	187	284	412	772	813	771	537	183	241	116	159	473	337	949	148	274	541	254	0	209	529
29	431	437	411	365	435	341	458	455	410	240	352	299	581	634	580	435	162	267	93	202	366	423	1035	184	341	337	319	209	0	609
30	283	822	802	868	926	268	151	277	905	370	257	706	1222	1253	1222	926	576	476	516	408	761	118	796	524	268	839	424	529	609	0

Table 4. Statistical Parameters Obtained in the Training Stage (Compounds **1**−**21**)

descriptor	Q^2	SECV	slope	bias	PLS factors	outliers
similarity[a]	0.80	0.53	0.97	0.15	2	no
distance[b]	0.60	0.74	0.98	0.14	1	no
AS ($w_\Gamma = 1.0$)	0.71	0.66	0.85	0.90	4	no
AS ($w_\Gamma = 0.5$)	0.82	0.51	0.97	0.17	3	steroid 1
AS ($w_\Gamma = 0.3$)	0.84	0.47	0.97	0.15	3	steroid 1
AS ($w_\Gamma = 0.1$)	0.77	0.57	0.97	0.17	2	no

[a] Similarity calculated using the cosine index. [b] Dissimilarity (for noncommon subgraphs) calculated using the Euclidean distance and the VOW descriptor.

Table 5. Statistical Parameters Obtained in the Validation Stage (Compounds **21**−**30**)

descriptor	r	SEP	slope	bias
similarity[a]	0.68	0.45	0.95	0.20
distance[b]	0.61	0.55	0.54	3.29
AS ($w_\Gamma = 1.0$)	0.28	0.78	−0.39	9.70
AS ($w_\Gamma = 0.5$)	0.77	0.39	0.90	0.58
AS ($w_\Gamma = 0.3$)	0.82	0.39	1.03	−0.45
AS ($w_\Gamma = 0.1$)	0.65	0.44	0.79	1.38

[a] Similarity calculated using the cosine index. [b] Dissimilarity (for noncommon subgraphs) calculated using the Euclidean distance and the VOW descriptor.

Table 6. Predicted and Observed log *K* Values for the Test Set (Steroids **22**−**30**)

number	predicted log *K*	observed log *K*
22	7.62	7.51
23	7.30	7.55
24	7.05	6.78
25	7.75	7.20
26	6.46	6.14
27	7.08	6.25
28	7.05	7.12
29	6.95	6.82
30	7.83	7.69

stressed that our model is based on topological measurements, and therefore, it is a simpler method than other approaches.

5. CONCLUSIONS

The method proposed in this paper has achieved the aim of refining the classical similarity measure by means of considering differences or dissimilarities (the distance concept) between the subgraphs which do not form the isomorphism of two molecular graphs. Approximate similarity (AS), the new index we defined for the first time, provides us with a more reliable comparison between chemical compounds. This fact has been proven in the development of a QSAR model for predicting the corticosteroid-binding affinity of steroids.

Moreover, when our results were compared with other more complex QSAR approaches summarized in the literature, better and similar statistical values (here, the determination coefficient, the standard error in prediction, and the slope and bias of the correlation analysis have been given) were obtained. The robustness of the method has also been tested by external validation, showing acceptable results.

The influence of the noncommon subgraphs on correcting the similarity values has to be studied in order to optimize the predictive ability of models. This weight setting is a simple process aimed at searching the dissimilarity proportion to be taken into account in the similarity calculation. New approaches aimed at searching the appropriate influence of the dissimilarity values are being studied. With this purpose, an automatic and individual setting of the weighting factor—carried out for each comparison—is being developed taking into account, on one hand, the similarity and distance values between the two compounds to be compared and, on the other hand, the similarity and distance values of the rest of the data set.

The steroids studied in this work show a rigid common structure, which characterizes this set of compounds. Therefore, the application of the AS measurement to other data sets composed of more flexible molecules has to be carried out in future work, as well as the study of the feasibility of using AS in chemical database screenings. In addition, new topological descriptors that account for electronic and steric effects are being developed and tested in order to provide AS with greater chemical information.

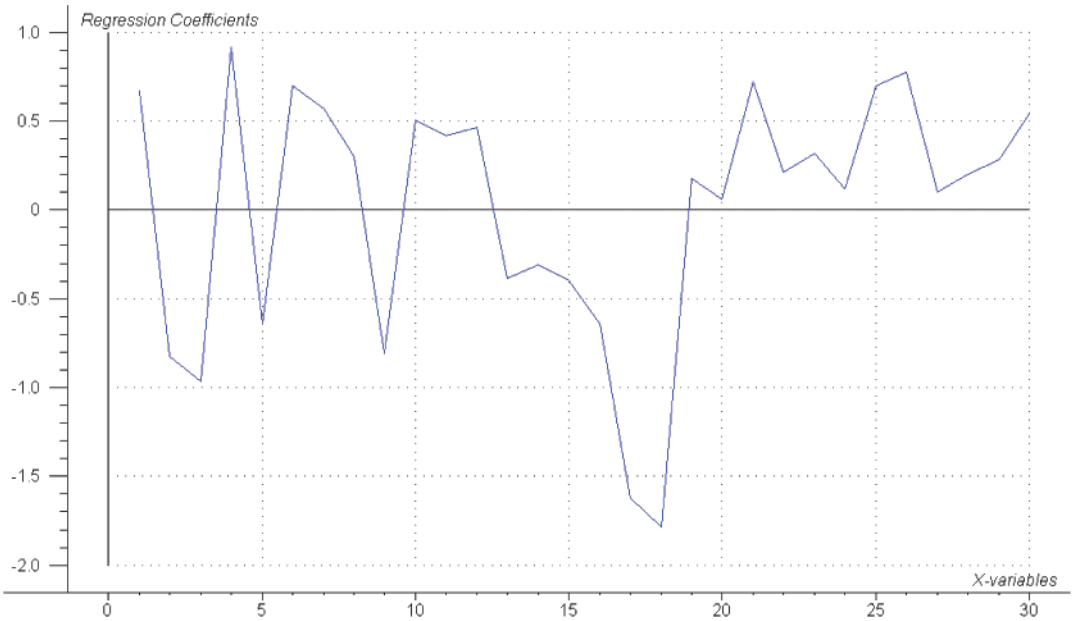

Figure 4. Regression coefficients for the AS variables of the model built with w_Γ 0.3.

Simplicity of the method is assured by the topological calculations carried out, which are open to different kinds of isomorphisms, similarity matching functions, topological descriptors, and distance measures.

REFERENCES AND NOTES

(1) *Structure−Property Correlations in Drug Research*; Van de Waterbeemd, H., Ed.; Academic Press: Austin, TX, 1996.

(2) Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; VCH: Weinheim, Germany, 1993.

(3) Hansch, C. A Quantitative Approach to Biochemical Structure−Activity Relationships. *Acc. Chem. Res*. **1969**, *2*, 232−239.

(4) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.

(5) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; John Wiley & Sons: New York, 2000.

(6) Esbensen, K. H. *Multivariate Data Analysis−in Practice*; Camo Process AS: Oslo, 2002.

(7) Wold, S.; Sjostrom, M.; Eriksson, L. PLS−Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(8) *8. Nature-Inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks*; Leardi, R., Ed.; Elsevier: New York, 2003.

(9) Cramer, R. D., III; Paterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(10) Kubinyi, H. Comparative Molecular Field Analysis (CoMFA). In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: New York, 1998.

(11) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules To Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130−4146.

(12) Good, A. C.; So, S.-S.; Richards, W. G. Structure−Activity Relationships from Molecular Similarity Matrices. *J. Med. Chem.* **1993**, *36*, 433−438.

(13) Dixon, S.; Merz, K. M., Jr.; Lauri, G.; Ianni, J. C. QMQSAR: Utilization of a Semiempirical Probe Potencial in a Field-Based QSAR Method. *J. Comput. Chem*. **2004**, *26*, 23−34.

(14) Silverman, B. D. The Thirty-one Benchmark Steroids Revisited: Comparative Molecular Moment Analysis (CoMMA) with Principal Component Regression. *Quant. Struct.-Act. Relat.* **2000**, *19*, 237−246.

(15) Silverman, B. D.; Platt, D. E. Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR without Molecular Superposition. *J. Med. Chem.* **1996**, *39*, 2129−2140.

(16) Todeschini, R.; Gramatica, P. New 3D Molecular Descriptors: The WHIM Theory and QSAR Applications. *Perspect. Drug Discovery Des.* **1998**, *9/10/11*, 335−380.

(17) Todeschini, R.; Moro, G.; Boggia, R.; Bonati, L.; Cosentino, U.; Lasagni, M.; Pitea, D. Modeling and Prediction of Molecular Properties. Theory of Grid-Weighted Holistic Invariant Molecular (G-WHIM) Descriptors. *Chemom. Intell. Lab. Syst.* **1997**, *36*, 65−73.

(18) (a) Schultz, H. P. Topological Organic Chemistry. 1. Graph Theory and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 227−228. (b) Schultz, H. P.; Schultz, E. B.; Schultz, T. P. Topological Organic Chemistry. 2. Graph Theory, Matrix Determinants and Eigenvalues, and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 27−29. (c) Schultz, H. P.; Schultz, T. P. Topological Organic Chemistry. 3. Graph Theory, Binary and Decimal Adjacency Matrices, and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci*. **1991**, *31*, 144−147. (d) Schultz, H. P.; Schultz, E. B.; Schultz, T. P. Topological Organic Chemistry. 4. Graph Theory, Matrix Permanents, and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 69−72. (e) Schultz, H. P.; Schultz, T. P. Topological Organic Chemistry. 5. Graph Theory, Matrix Hafnians and Pfanians, and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 364−368. (f) Schultz, H. P.; Schultz, T. P. Topological Organic Chemistry. 6. Graph Theory and Topological Indices of Cycloalkanes. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 240−244. (g) Schultz, H. P.; Schultz, E. B.; Schultz, T. P. Topological Organic Chemistry. 7. Graph Theory and Molecular Topological Indices of Unsatured and Aromatic Hydrocarbons. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 863−867. (h) Schultz, H. P. Topological Organic Chemistry. 8. Graph Theory and Topological Indices of Heteronuclear Systems. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 1151−1157. (i) Schultz, H. P. Topological Organic Chemistry. 9. Graph Theory and Topological Indices of Stereoisomeric Organic Compounds. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 864−870.

(19) Golbraikh, A.; Tropsha, A. QSAR Modeling Using Chirality Descriptors Derived from Molecular Topology. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 144−154.

(20) Golbraikh, A.; Bonchev, D. Novel ZE-Isomerism Descriptors Derived from Molecular Topology and Their Application to QSAR Analysis. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 769−787.

(21) Yao, Y.-Y.; Xy, L. Study on Structure−Activity Relationships of Organic Compounds: Three New Topological Indices and Their Applications. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 590−594.

(22) Downs, G. M.; Barnard, J. M. Clustering and Their Uses in Computational Chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B.; Eds.; Wiley-VCH: New York, 2003; pp 1−39.

(23) Ivanciuc, O.; Balaban, A. T. The Graph Description of Chemical Structures. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: The Netherlands, 1999; pp 59−167.

(24) Rouvray, D. H.; Balaban, A. T. *Chemical Applications of Graph Theory. Applications of Graph Theory*; Wilson, R. J., Beineke, L. W., Eds.; Academic Press: New York, 1979; pp 177−221.

(25) Urbano Cuadrado, M.; Luque Ruiz, I.; Gómez-Nieto, M. A. A New Quantitative Structure−Property Relationship (QSPR) Approach using Dissimilarity Measurements based on Topological Distances of Non-Isomorphic Subgraphs. *MATCH*. Paper submitted.

(26) Coats, E. A. The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. *In 3D QSAR in Drug Design*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/Escom: Dordrecht, The Netherlands, 1998; pp 199−213.

(27) Cerruela García, G.; Luque Ruiz, I.; Gómez-Nieto, M. A. Step-by-Step Calculation of All Maximum Common Substructures through a Constraint Satisfaction Based Algorithm. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 30−41.

(28) Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(29) Lipkowitz, K. B.; Pradham, M. Computational Studies of Chiral Catalysts: A Comparative Molecular Field Analysis of an Asymmetric Diels−Alder Reaction with Catalysts Containing Bisoxazoline or Phosphinooxazoline Ligands. *J. Org. Chem*. **2003**, *68*, 4648−4656.

(30) Shenk, J. S.; Westerhaus M. O. Calibration the ISI way. In *Near Infrared Spectroscopy: The Future Waves*; NIR Publications: Chichester, U. K., 1996; pp 198−202.

CI0600511