Three-Dimensional Quantitative Structure−Activity Relationships from Tuned Molecular Quantum Similarity Measures: Prediction of the Corticosteroid-Binding Globulin Binding Affinity for a Steroid Family

David Robert, Lluís Amat, and Ramon Carbó-Dorca*

Institute of Computational Chemistry, University of Girona, 17071 Girona, Catalonia, Spain

Received September 16, 1998

Predictive models based on tuned molecular quantum similarity measures and their application to obtain quantitative structure−activity relationships (QSAR) are described. In the present paper, the corticosteroid-binding globulin binding affinity of a 31 steroid family is studied by means of a multilinear regression using molecular descriptors derived from mixed steric−electrostatic quantum similarity matrixes as parameters, obtaining satisfactory predictions. A systematic procedure to treat outliers by using triple-density quantum similarity measures is also presented. This method depicts an alternative to the grid-based QSAR techniques, providing a consistent approach that avoids problematic result dependency on the grid parameters.

INTRODUCTION

One of the most progressing subjects in present-day chemistry consists of the establishment of quantitative relationships between biological or pharmacological properties and molecular structure. This topic has become a solid branch in chemistry, usually known as quantitative structure−activity relationships (QSAR). The pioneering studies on QSAR were performed by Hansch and Fujita,[1] and since then the advances in this matter have not ceased. The predictive capabilities of the earliest models were substantially improved when 3D structural descriptors were introduced, providing a powerful alternative to the use of semiempirical extrathermodynamic parameters in QSAR studies.[2] In addition, the definition of different quantitative similarity measures between two molecules, first introduced by Carbó et al.,[3] proved a great aid in order to provide molecular descriptors as a source of 3D QSAR parameters.

Many QSAR methods have been proposed in the last years, all of them employing different molecular descriptors and dealing with different statistical tools.[4−10] In the present paper, a 3D QSAR model using tuned molecular quantum similarity measures and a multilinear regression will be described. This work is included in a series of QSAR studies developed in our laboratory,[11−13] where different molecular descriptors, namely, molecular quantum similarity measures (MQSM) and indices (MQSI), are used to relate biological activities to quantum similarity general theory. The underlying premise in all the similarity−property studies is that "similar" molecules will possess "similar" properties. Quantum similarity theory provides an adequate way to quantify the resemblance between a set of compounds, making possible to use the derived similarity measures to make the previous concept practical.

The molecular set chosen to validate this new methodology is a well-known set for QSAR researchers, the so-called Tripos or Cramer steroid data set.[14] It is made up of 31 steroids that bind to the corticosteroid-binding globulin (CBG) receptor. This set has already been studied by several authors, and it has become a standard training set for the majority of the novel QSAR methods. In previous works different methodologies have been used to examine this system, achieving satisfactory predictions.[12,14−31]

THEORETICAL FRAMEWORK

In the present paper, the steroid family has been studied by a new methodological procedure derived from the connection between MQSM and the general convex sets theory.[13] Before the presentation of the results, some theoretical details on the method used and its fundamental magnitudes will be exposed.

Molecular Quantum Similarity Measures. Molecular quantum similarity general theory was developed in order to obtain a quantitative measure of the similarity between two molecules by using their electronic density functions as quantum mechanical descriptors.[3,32−35] The quantum mechanical postulates assume that the wave function contains all the information of a system. Löwdin[36] and McWeeny[37] developed the density function formalism, which is the basis of quantum similarity. Later, Hohenberg and Kohn[38] demonstrated the existence of one-to-one correspondence between the wave function and the density function of a quantum system, concluding that densities are as robust a source of quantum object information as the wave functions. Starting from this framework, a MQSM can be constructed from the scalar product between two first-order molecular electronic density functions, defined as the integral

$$Z_{AB}(\Omega) = \int\int \rho_A(\mathbf{r}_1)\Omega(\mathbf{r}_1,\mathbf{r}_2)\rho_B(\mathbf{r}_2)\, d\mathbf{r}_1\, d\mathbf{r}_2 \qquad (1)$$

where $\{\rho_A(\mathbf{r}_1),\rho_B(\mathbf{r}_2)\}$ are the density functions of each molecule and $\Omega(\mathbf{r}_1,\mathbf{r}_2)$ is an arbitrary positive definite operator. In relation to this operator, different types of MQSM can be defined. Three classes of MQSM will be used

* To whom correspondence should be addressed: E-mail director@iqc.udg.es; Fax 34 972 418356.

10.1021/ci980410v CCC: $18.00
Published on Web 02/09/1999

in this paper: (a) *Overlap-like MQSM*: this type of MQSM is defined by a Dirac δ distribution, i.e., $\Omega(\mathbf{r}_1,\mathbf{r}_2) = \delta(\mathbf{r}_1 - \mathbf{r}_2)$. (b) *Coulomb-like MQSM*: $\Omega(\mathbf{r}_1,\mathbf{r}_2) = |\mathbf{r}_1 - \mathbf{r}_2|^{-1}$. (c) *Triple-Density* (TD) *MQSM*: the operator Ω is substituted by another first-order molecular density function,[39] $\Omega(\mathbf{r}_1,\mathbf{r}_2) = \rho_C(\mathbf{r}_1)\delta(\mathbf{r}_1 - \mathbf{r}_2)$.

Computation of Molecular Density Functions and MQSM. In this work, approximate first-order molecular density functions have been used in order to avoid extensive ab initio calculations. An atomic shell approximation (ASA) algorithm[40−42] has been employed to obtain accurate fitted molecular density functions. In addition, a promolecular approximation[42] has been chosen to construct the first-order molecular densities, considered as a sum of atomic density functions:

$$\rho_M(\mathbf{r}) \approx \frac{1}{Z_M}\sum_{a\in M} Z_a\rho_a(\mathbf{r} - \mathbf{R}_a) \tag{2}$$

where Z_a is the number of electrons of the atom a and $Z_M = \sum_{a\in M} Z_a$ is the total number of electrons of the molecule M. Considering the atomic densities normalized to 1, this yields to a normalization of the molecular density:

$$\int \rho_M(\mathbf{r})\, d\mathbf{r} = 1 \tag{3}$$

The atomic density functions ρ_a are constructed as a linear combination of squared 1s Gaussian-type orbitals:

$$\rho_a(\mathbf{r} - \mathbf{R}_a) = \sum_{i\in a} w_i |g_i(\mathbf{r} - \mathbf{R}_a,\zeta_i^a)|^2 = \sum_{i\in a} w_i \left|\left(\frac{2\zeta_i^a}{\pi}\right)^{3/4} \exp(-\zeta_i^a|\mathbf{r} - \mathbf{R}_a|^2)\right|^2 \tag{4}$$

with $\{\zeta_i^a\}$ being the exponents of the basis Gaussian functions centered on atom a. The coefficients $\{w_i\}$ satisfy the convex constraint: that is, they are positive definite and sum 1, and they have been obtained by minimizing the quadratic error integral function[42] between ASA and ab initio atomic first-order densities. The rule used in this work to construct the promolecular ASA density functions is one function centered on H atoms and three functions on C, N, O, and F atoms.[43]

Further, the molecular superpositions have been optimized to obtain maximal MQSM values, since MQSM depends on the relative position of the molecules involved. For two-body operators, the superposition search procedure encompasses two stages.[44] In a first stage, a systematic search is performed until the optimal superposition of three pairs of atoms of the involved molecules is obtained. After this, and starting from the best molecular alignment found, a local search is implemented in order to refine the solution and obtain the maximal value of the similarity measure, using a Newton or Simplex technique depending on whether the first or second derivatives of a given MQSM are known, respectively. On the other hand, TD MQSM need a particular procedure based on bimolecular superpositions, recently developed.[13] The relative rigidity of the system analyzed allows the conformational variable to be neglected.

Once the overall quantum similarity measures for all the pairs of molecules are calculated, they can be expressed in a matrix form, yielding the quantum similarity matrix $Z(\Omega) = \{Z_{AB}(\Omega)\}$, depending on which operator is chosen.

Tuned QSAR Formalism. Convex sets theory[45,46] is a mathematical formalism related to fuzzy sets and Boolean tagged sets. The term convex set can be interpreted as a collection of vectors in a vector semispace, fulfilling some well-defined constraints. In the present model, convex sets have been associated to the manipulation of the molecular information contained in the different similarity matrixes $\mathbf{Z}(\Omega_\alpha)$ by means of a positive linear combination:[13]

$$\mathbf{Z} = \sum_{\alpha=1}^{v} c_\alpha \mathbf{Z}(\Omega_\alpha) \tag{5}$$

where the real valued set of weights $\{c_\alpha\}$ fulfills the convex conditions $\{c_\alpha > 0;\ \forall\alpha\} \wedge \{\sum_\alpha c_\alpha = 1\}$. With this linear combination of ν matrixes, the final similarity matrix $\mathbf{Z}$ can be taken as a possible representation of the molecular set and used as a source of parameters for the QSAR study. This procedure is similar to the usual combination of several descriptors, namely, similarity measures or topological indices, encompassing different molecular aspects such as shape,[47] topology,[12] electrostatic potentials or electric fields,[48] molecular moments,[21] or even descriptors associated to substituent information;[49] not to mention the traditional 2D descriptors: hydrophobicity constants,[50] molar refractivities, Hammett electronic constants,[51] Sterimol parameters[52] or Kier and Hall's topological indices.[53] In the present case, each type of MQSM, and subsequently, each associated operator, represents different aspects of the whole molecular information. Thus, overlaplike MQSM brings forward steric information on the shape and charge concentration[54] of the molecules compared, and Coulomb-like MQSM is, except for a numerical factor, the electrostatic repulsion energy between two molecular charge distributions. As will be shown later, TD MQSM will be introduced to correct possible deviations arising in the adjustment of the data. The use of MQSM to describe other relevant molecular properties, such as hydrophobicity, has been recently introduced.[55]

The weights $\{c_\alpha\}$ for the different similarity matrixes are calculated in such a way that they optimize the adjustment to the experimental observations, to obtain the optimal QSAR model. This methodology has been called tuned QSAR (TQSAR).

Before computation of the set of $\{c_\alpha\}$ coefficients, all similarity matrixes are scaled as follows:

$$Z_{IJ}^{(n)} = Z_{IJ}\left(\sum_{KL} Z_{KL}\right)^{-1} \tag{6}$$

The $\{c_\alpha\}$ optimized weights will then represent the correctly scaled contribution that every type of MQSM possesses in the TQSAR model.

Dimensionality Reduction and Selection of Variables. Once the tuned similarity matrix is constructed, a transformation is necessary to reduce the dimension of the data and eliminate redundant information. An appropriate method for this purpose is the classical scaling,[57,58] also known as multidimensional scaling or principal coordinate analysis. Classical scaling considers the objects as points in a multidimensional Euclidean space and finds coordinates for these points in such a way that the interpoint distances fit as

well as possible the original similarities. The classical metric scaling solution $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_p\}$ for a general matrix **S**; **S**, being square, symmetric, and positive semidefinite, is obtained by manipulating the matrix product

$$\mathbf{B} = \mathbf{HSH} = \mathbf{XX}^{\mathrm{T}} \tag{7}$$

where **H** is the $(n \times n)$ centring matrix:

$$\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1} \tag{8}$$

where **I** is the identity matrix and the *unity matrix* **1** is the $(n \times n)$ matrix of ones. The double centering of **S** sets the origin of coordinates at the centroid of the data. To recover the coordinates of **B**, its spectral decomposition will be used:

$$\mathbf{B} = \mathbf{V}\Lambda\mathbf{V}^{\mathrm{T}} \tag{9}$$

where $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_p)$ is the diagonal matrix of the **B** eigenvalues, and **V** is the corresponding eigenvector matrix. It must be noted that the rank of the configuration matrix **X** is p, where $p \leq n - 1$.[59] Hence it can be straightforwardly proved that the coordinate matrix **X** is

$$\mathbf{X} = \mathbf{V}\Lambda^{1/2} \tag{10}$$

This method, according to the Eckart and Young theorem,[60] finds the optimal p-dimensional subspace such that $\mathbf{S}_{(p)}$ is best fitted to **HSH** in the least-squares sense:

$$\mathrm{trace}(\mathbf{HSH} - \mathbf{S}_{(p)})^2 = \mathrm{minimum} \tag{11}$$

Quantum similarity matrixes **Z** satisfy the demanded prerequisites and therefore can be dealt with by the classical scaling technique.

The axes in the multidimensional Euclidean space are called *principal axes*, and the point coordinates in this space are called the principal coordinates (PCs) of the system. From all the possible PCs, those ones used in the models have been chosen by means of the most predictive variables method.[61,62] This method selects the first k columns ($k \leq p$) of the configuration matrix **X** arranged in descending order of absolute correlation with the studied data, given by

$$\chi^2(\mathbf{y},\mathbf{x}_i) = \frac{(\mathbf{y}^T\mathbf{x}_i)^2}{\sum_j (y_j - \bar{y})^2 \lambda_j} \tag{12}$$

where λ_j is the eigenvalue corresponding to the jth axis. Principal axes accounting for less than an arbitrary threshold have been neglected in order to avoid possible noise parametrization. One percent of variance has been chosen here to be the threshold. The justification of this variable selection method in terms of explained variation and the proof that the risk of chance correlations is not improved substantially will be discussed later.

The adopted selection criterion is better than the usual one based on the arrangement of the eigenvectors in descending order of the corresponding eigenvalues, since there can exist cases where the properties are highly correlated with one or several variables with small variances.[63]

General Connection between MQSM and QSAR. In a previous paper,[64] a formal connection between the molecular properties and the structural descriptors generated by the MQSM was proposed. This is based on one of the most fundamental principles of quantum mechanics, which relates the density function to any physical observable through a Hermitian operator, by means of

$$y = \langle\omega\rangle = \int \omega\, \rho(\mathbf{r})\, \mathrm{d}\mathbf{r} = \langle\omega|\rho\rangle \tag{13}$$

where y is the observable studied and $\boldsymbol{\omega}$ is an unknown quantum operator, considered of multiplicative type. The equations reported in that paper, and also used here, are a discretization of this expectation value rule, where a transformation of the corresponding similarity vector plays the role of the density function:

$$y_I = \langle\omega\rangle \approx \beta^{\mathrm{T}}\mathrm{f}(\mathbf{z}_I) \tag{14}$$

As in all QSAR techniques, an adjustment model is built from a set of compounds of known activity, the training set, and this model is then used to predict the property for a series of molecules of unknown activity, the test set. The relationship is presented here in the form of a multilinear regression[65] (MLR), with the transformed similarity vectors as parameters:

$$\mathbf{y} = \beta_0\mathbf{1} + \mathbf{x}_{(k)}\beta_{(k)} + \epsilon_{(k)} \tag{15}$$

where $\mathbf{y} = (y_1, ..., y_n)^{\mathrm{T}}$ is the n-dimensional vector containing the activities of the system studied, **1** is a $(n \times 1)$ unity vector, $\mathbf{X}_{(k)}$ denotes the reduced $(n \times k)$ configuration matrix, and $\epsilon_{(k)}$ is a normally distributed error term having expected value 0 and dispersion matrix $\mathbf{I}\sigma^2$.[65] The $\boldsymbol{\beta}$ estimators are computed by an ordinary least-squares technique, and usually no more than a small number of PCs will be necessary to describe the property desired.

The statistical parameters computed to assess the model accuracy are the conventional squared regression coefficient (r^2) and the cross-validated coefficient for prediction (q^2), defined as

$$q^2 = (\mathrm{SD} - \mathrm{PRESS})/\mathrm{SD} \tag{16}$$

where SD is the sum of squared deviations of each property value from their mean and PRESS, the predictive residual sum of squares,[66,67] is the sum of the squared differences between the actual and the cross-validated property values, found either by the leave-one-out procedure or by manipulations of the so-called hat matrix.[65]

The standard deviation of the errors of prediction, the SDEP coefficient, will be used to give a measure of the predictive power of the model. It is defined as

$$\mathrm{SDEP} = [\sum (y_{\mathrm{pred}} - y_{\mathrm{obs}})^2/n]^{1/2} \tag{17}$$

The use of similarity matrixes as QSAR descriptors was first proposed by Good *et al.*,[15,68] and the method shown utilizes several aspects of this work.

TQSAR Computational Procedure. Once all the MQSM matrixes are obtained by the optimal superposition of the molecules under comparison, the TQSAR model is constructed following the indications given in the previous section.

The optimization procedure for the linear combination weights is performed in the following way. First, a random

Table 1. CBG Binding Affinity Data[a]

compound	no.	CBG (pK_a)	compound	no.	CBG (pK_a)
aldosterone	**1**	−6.279	etiocholanolone	**16**	−5.225
androstanediol	**2**	−5.000	pregnenolone	**17**	−5.225
androstenediol	**3**	−5.000	hydroxypregnenolone	**18**	−5.000
androstenedion	**4**	−5.763	progesterone	**19**	−7.380
androsterone	**5**	−5.613	hydroxyprogesterone	**20**	−7.740
corticosterone	**6**	−7.881	testosterone	**21**	−6.724
cortisol	**7**	−7.881	prednisolone	**22**	−7.512
cortisone	**8**	−6.892	cortisolacetate	**23**	−7.553
dehydroepiandrosterone	**9**	−5.000	4-pregnene-3,11,20-trione	**24**	−6.779
deoxycorticosterone	**10**	−7.653	epicorticosterone	**25**	−7.200
deoxycortisol	**11**	−7.881	19-nortestosterone	**26**	−6.144
dihydrotestosterone	**12**	−5.919	16a,17a-dihydroxyprogesterone	**27**	−6.247
estradiol	**13**	−5.000	17a-methylprogesterone	**28**	−7.120
estriol	**14**	−5.000	19-norprogesterone	**29**	−6.817
estrone	**15**	−5.000	2a-methylcortisol	**30**	−7.688
			2a-methyl-9a-fluorocortisol	**31**	−5.797

[a] From ref 72.

sequence of $\{c_\alpha\}$ coefficients is generated by the Monte Carlo method. Next, and starting from this sequence, a one-dimensional Fibonacci search is carried out for each weight separately. This cycle is repeated until a satisfactory TQSAR model is produced with a maximum value of the predictive statistical parameter q^2. This process is then repeated for all the possible combinations of ν similarity matrixes in order to find the optimal operators for the description of the studied property. The number of existing possibilities is given by the combinatorial factor

$$\binom{\mu}{\nu}$$

where μ is the total number of similarity matrixes considered. A more detailed description of the computational aspects, such as the use of nested summation symbol parallelizable algorithms[69,70] for the generation of combinations, is presented in ref 13. All the calculations reported in this work have been performed with the TQSAR−SIM program developed in our laboratory.[71] One hundred cycles for the Monte Carlo optimization procedure have been employed when two similarity matrixes are combined ($\nu = 2$), and 300 cycles were employed when three or more matrixes ($\nu \geq 3$) are used, to avoid finding undesirable local maxima.

The degrees of freedom in the model are the number of matrixes involved in the linear combination (ν), the number of principal axes used in the MLR (k), and more accurately, also the number of optimization cycles in the Monte Carlo method. Considering this, the total number of parameters of the model is $k + \nu - 1$, where this last subtraction is due to the existence of the convex constraint.

RESULTS AND DISCUSSION

The training set used to validate the TQSAR model is made up of 31 molecules, a steroid set that binds to the CBG receptor.[14] Table 1 gathers the entire studied set with the actual binding affinities, taken from ref 72. Steroid geometries used in this work have been kindly provided by the Gasteiger group[73] and can also be found in our Website.[74] Because the studied steroid structures have been already depicted in several papers, they will not be included here. For more details see, for example, Figure 1 of ref 12.

Table 2. MLR Using One up to Five Principal Axes[a]

operator	no. of PCs (k)	selected PCs	q^2	r^2
overlap	1	1	0.603	0.643
	2	1, 3	0.692	0.737
	3	1, 3, 2	0.730	0.781
	4	1, 3, 2, 4	0.743	0.795
	5	1, 3, 2, 4, 9	0.751	0.817
coulomb	1	1	0.632	0.669
	2	1, 4	0.689	0.767
	3	1, 4, 5	0.651	0.803
	4	1, 4, 5, 7	0.689	0.819
	5	1, 4, 5, 7, 8	0.759	0.833

[a] Overlaplike and Coulomb-like MQSM have been used. The dimensions have been selected by the most predictive variables method with a threshold of 1% of variance.

As indicated previously, the steroid set has been employed repeatedly by QSAR researchers to validate their models, since it has traditionally been, at least until the CoMFA appearance, a complicated set to correlate the affinities successfully.[75] The TQSAR study will be structured in the following way: first, a simple QSAR analysis using only one similarity matrix at a time will be carried out. Next, linear combinations of two-body operators (overlap and Coulomb) will be permitted. TD MQSM will then be introduced as corrective descriptors. Finally, a statistical validation of the model using the randomization test[76] and a comparison between TQSAR results and those attained by other QSAR studies will be performed.

QSAR Study Using Single Quantum Similarity Matrixes. The first step in the prediction of the CBG binding affinities for the steroid family set is the construction of a QSAR model using only one quantum similarity matrix. This matrix will be orthogonally transformed by means of the classical scaling analysis and the steroid activities will be adjusted through an MLR using a reduced number of PCs. The results are shown in Table 2. The MLR statistics indicate that a reasonably good correlation between the experimental CBG binding affinities and the molecular structure descriptors, represented by the similarity measures derived from the first-order molecular densities, exists, as was found in a previous paper.[12] The best QSAR model is obtained when Coulomb-like MQSM and five PCs are used, yielding $r^2 = 0.833$ and $q^2 = 0.759$. Note that a clear consistency in the PCs chosen appears. The cumulated explained variance for

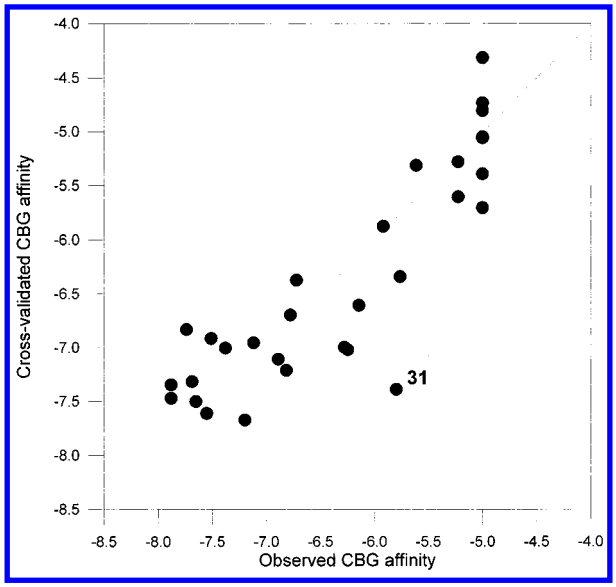

Figure 1. Cross-validated versus experimental CBG binding affinities for the 31 steroids. QSAR model: Coulomb operator, five PCs. Molecule **31** has been explicitly marked.

the five selected PCs is still rather high, 54%, in relation to the first five, 71%, and this justifies the choice. This is a general behavior for all the cases discussed, and it will not be discussed anymore. The assessment that the variable selection method does not produce chance correlations in the model will be discussed later. These results agree with other QSAR studies[28,29] where the electrostatic descriptors were found to be better than the steric ones in describing steroid CBG binding affinity.

A representation of the experimental activities versus the predicted ones for the optimal model is shown in Figure 1. In this plot, there is one molecule relatively poorly located. Compound **31** presents 27% relative error and can be considered an outlier. This behavior is also obtained for the rest of calculated QSAR models, and it is a feature stated in many QSAR studies.[14−28,30,31] This can be attributed to the existence of a fluorine atom, the only non-oxygen functionality present in the data set. As indicated by Jain et al.,[17] this fact clearly determines its CBG binding affinity, since otherwise this compound is identical to molecule **30**, which is much more active. Up to a point, molecules **1**, **4**, **9**, **18**, **20**, and **27** also present small deviations, with approximately 11−14% relative errors.

Construction of a TQSAR Model Using a Mixture of Steric and Electrostatic Descriptors. To improve the above description, a TQSAR model is constructed with a mixture of two-body operators, namely, overlap and Coulomb MQSM, representing the steric and electrostatic molecular contributions, respectively. Thus, the linear combination of these two similarity matrixes will be permitted. Table 3 gathers the results obtained with this first tuned model.

As can be seen, a considerably better adjustment is obtained, reflected in the r^2 and q^2 statistical parameter values. A high coherence with the previous models is also achieved. Thus, the basis of the TQSAR models is the Coulomb-like MQSM, which was found to be the optimal single descriptor in the above subsection, and minimal corrections are introduced in the form of overlap contributions (about 0.1−6%). As in the CoMFA study, the predictive

Table 3. TQSAR Model for Linear Combinations of Two Matrixes[a]

ν	k	selected PCs	TQSAR model	q^2	r^2
2	1	1	0.06329 × OVE + 0.93671 × COU	0.650	0.687
	2	1, 4	0.01607 × OVE + 0.98393 × COU	0.743	0.790
	3	1, 4, 8	0.01108 × OVE + 0.98892 × COU	0.774	0.820
	4	1, 4, 8, 5	0.00696 × OVE + 0.99304 × COU	0.779	0.828
	5	1, 4, 5, 8, 7	0.00101 × OVE + 0.99899 × COU	0.767	0.834

[a] MLR using one up to five principal axes. Overlaplike (OVE) and Coulomb-like (COU) MQSM have been used. The dimensions have been selected by the most predictive variables method with a threshold of 1% of variance.

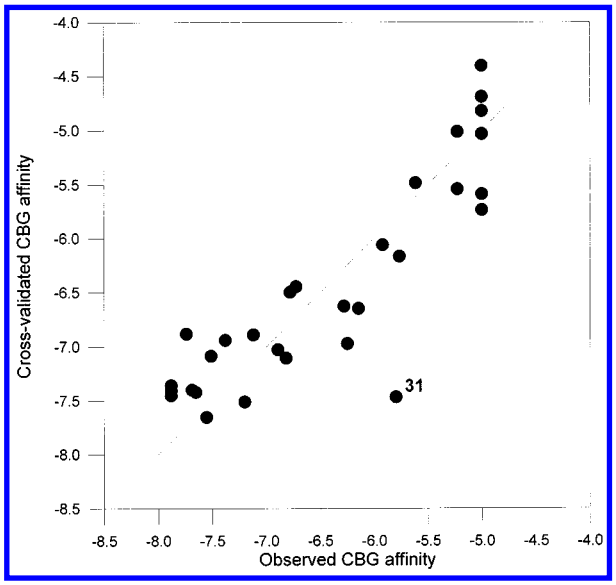

Figure 2. Cross-validated versus experimental CBG affinities. TQSAR model: two quantum similarity matrixes (Coulomb, overlap), four PCs. Molecule **31** has been explicitly marked.

power of the models derived from the individual steric and electrostatic descriptors is comparable to those derived from the combined matrixes.[14] Finally, the consistency of the PCs selected by the most predictive variables method is maintained.

The best model is found when four principal axes (five parameters) are used. It yields $r^2 = 0.828$ and $q^2 = 0.779$ values. A representation of the experimental activities versus the predicted ones for the optimal model is shown in Figure 2.

Again compound **31** is a pronounced outlier (29% relative error), and steroids **2**, **9**, **18**, **20**, and **27** are slightly distant from the diagonal line (about 11−14% error). For the other TQSAR models this behavior is repeated: molecule **31** is the worst described, with relative errors about 25−30%, and a small series of molecules also present small deviations. In all the models built the predicted affinity of molecule **31** is higher than the observed value.

To illustrate the influence of point **31** in the regression statistics, we have recalculated the TQSAR model excluding molecule **31** of the set. The optimal model is found when Coulomb (96%) and overlap (4%) matrixes are combined in a linear regression with five PCs (1, 4, 8, 3, and 5): $r^2 = 0.909$ and $q^2 = 0.872$. These values are much better than those attained when molecule **31** is included, and in this case no relative error larger than 15% is found. This difference

in the predictive coefficient is due to the fact that q^2 is very sensitive to any very large error of prediction, and just one very bad prediction can produce a low q^2.

Triple-Density Molecular Quantum Similarity Measures as a Tool to Remove Outliers. TD MQSM, defined in a previous section, can be considered as matrix representations of the molecule whose density function acts as an operator, in the basis set where the overall densities belong. Further, they can be also viewed as subjective perceptions of the similarity relationships between the objects compared, seen from the viewpoint of the compound that plays the role of operator.[39] This distortion can be seen by plotting a bi- or tridimensional representation of the set with the principal coordinates as axes. It can then be seen that the molecule acting as an operator usually tends to be located in an isolated position far from the rest of the compounds, even though it possesses a structure similar to the other ones.[77] This fact can be used precisely to eliminate, or at least to reduce, the statistical influence of the desired molecules. To do this, the TD MQSM corresponding to the molecules poorly described by the model, in our case, compound **31**, will be calculated and introduced as possible descriptors of a new TQSAR model. These parameters, weighted with an adequate coefficient, will relocate the remodeled molecules and will possibly remove their statistical influence.

This procedure is not only oriented to improve the training set adjustment. Thus, if molecules with similar features to the outlier (with, for instance, fluorine atoms at the same position but with different substituents at other positions) were analyzed in order to predict their CBG affinity, pressumably the model including similarity matrix TD31 would improve its predictive ability. Due to the rather generalized fact that molecular similarity tends to classify correctly objects of homogeneous systems but usually is not able to give any information about a dissimilar object introduced in the set, TD matrixes overcome this problem by a specific parametrization oriented to the outliers. Thus, if a dissimilar molecule is introduced in the training set (such as compound **31**), the corresponding TD matrix will improve its description, and it will also improve the description of those molecules from the test set similar to the outlier.

TQSAR Model Using Two-Body Operators and TD Similarity Matrixes. As indicated in the last section, TD MQSM can be introduced in the TQSAR algorithm as possible descriptors of different statistical or molecular features, such as the presence of outliers. Thus, three quantum similarity matrixes, corresponding to the two two-body operators, OVE and COU; and the TD MQSM for the outlier **31**, TD31; are introduced to construct the TQSAR model, giving the model a chance to select any of them. The results obtained are shown in Table 4.

This more sofisticated TQSAR model yields the most valuable results. The best TQSAR model is found when two matrixes and six PCs are used (eight parameters), yielding $q^2 = 0.842$ and $r^2 = 0.903$ values. The TD similarity matrix associated with the molecule-operator **31** possesses more than half the TQSAR model weight (64%), and its presence removes the influence of this point in the predictive model, as expected. In particular, the relative error of compound **31** decreases to 8.5%.

With one less PC, the results are of the same quality: $q^2 = 0.833$ and $r^2 = 0.890$, and the error in the molecule **31** prediction reduces to 4.6%. In this scheme the consistency in the weights, the matrixes involved, and the PCs chosen as MLR parameters is no longer maintained. Table 5 gathers the obtained cross-validated CBG affinities, and a graphical representation of the actual activities versus the predicted ones for the eight-parameter model is shown in Figure 3, where no outliers are observed.

Another possibility could consist of introducing all the possible TD MQSM, together with the two-body operators (overlap and Coulomb), in the TQSAR model and then choose the ν optimal descriptors from the existing 33 (2 + 31) matrixes. The main problem in this method is that a large number of descriptors can produce undesirable chance correlations. In addition, this procedure is really cumbersome due to the large computational time required, both for computing the TD MQSM by means of the optimal molecular superpositions and for calculating all the possible

Table 4. TQSAR Model for Linear Combinations of Two Operators[a]

ν	k	selected PCs	TQSAR model	q^2	r^2
2	3	1, 4, 8	0.98279 × COU + 0.01721 × TD31	0.789	0.831
	4	1, 4, 5, 10	0.39526 × TD31 + 0.60474 × OVE	0.796	0.852
	5	1, 4, 6, 9, 11	0.36384 × OVE + 0.63616 × TD31	0.833	0.890
	6	1, 4, 6, 9, 11, 8	0.36290 × OVE + 0.63710 × TD31	0.842	0.903
	7	1, 4, 6, 9, 11, 8, 5	0.36120 × OVE + 0.63880 × TD31	0.837	0.905

[a] MLR using three up to seven principal axes. Three quantum similarity matrixes ($\mu = 3$) have been used (see text for details). Trident MQSM for molecule **31** have been denoted by TD31. The dimensions have been selected by the most predictive variables method with a threshold of 1% of variance.

Table 5. Observed, Cross-Validated, and Relative Errors of the 31 CBG Affinities for the TQSAR Model[a]

compound	observed CBG affinity (pK_a)	cross-validated CBG affinity (pK_a)	% relative error (absolute value)
1	−6.279	−6.605	5.18
2	−5.000	−5.636	12.72
3	−5.000	−4.701	5.98
4	−5.763	−6.710	16.44
5	−5.613	−5.404	3.72
6	−7.881	−7.994	1.43
7	−7.881	−7.799	1.05
8	−6.892	−6.934	0.61
9	−5.000	−4.632	7.36
10	−7.653	−7.498	2.03
11	−7.881	−7.268	7.78
12	−5.919	−5.322	10.08
13	−5.000	−4.918	1.64
14	−5.000	−4.901	1.98
15	−5.000	−4.689	6.22
16	−5.225	−5.456	4.42
17	−5.225	−5.664	8.41
18	−5.000	−5.619	12.38
19	−7.380	−6.943	5.92
20	−7.740	−6.771	12.51
21	−6.724	−6.551	2.58
22	−7.512	−7.609	1.29
23	−7.553	−7.669	1.53
24	−6.779	−6.452	4.83
25	−7.200	−7.503	4.21
26	−6.144	−6.549	6.59
27	−6.247	−6.832	9.36
28	−7.120	−6.895	3.15
29	−6.817	−7.202	5.65
30	−7.688	−7.550	1.80
31	−5.797	−6.288	8.46

[a] Two quantum similarity matrixes (overlap, TD31), six PCs.

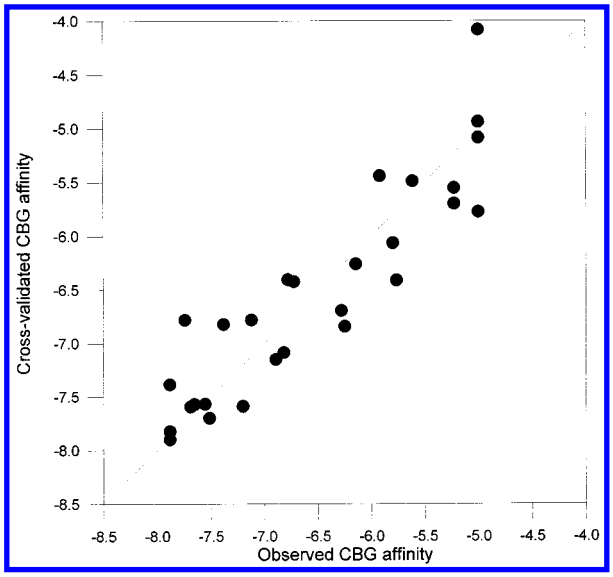

Figure 3. Cross-validated versus experimental CBG affinities. TQSAR model: three quantum similarity matrixes (overlap, TD31, TD2), five PCs.

combinations of ν matrixes. In reference to this last observation, it must be noted that it is reasonable to think that $n \gg n'$, n' being the number of outliers of the system. This inequality yields to

$$\binom{2+n}{\nu} \gg \binom{2+n'}{\nu}$$

where the second term corresponds to the overlap and Coulomb operators. The systematic procedure presented here does not ensure the best description of the studied property, since some TD matrixes have been rejected a priori, but provides a faster and more easily interpretable model than the one that will result of employing all the similarity matrixes.

Predictive Ability of TQSAR. The handicap of any QSAR approach consists of the property prediction for a set of compounds that have not been used to build the model, the so-called test set. It is assumed that the structures of the test set will be similar to the training set ones. In the steroid set this is usually performed by dividing the compounds into two classes: a training set of 21 compounds and a test set made up of the remaining 10. Thus, only the 21-steroid set is used to elaborate the predictive model, and it is then applied on the test set. In this case, the use of the TD31 matrix will not be so interesting, since the outlier is not included in the training set. It must be remembered that the criterion of matrix selection is based on the optimization of the q^2 coefficient, and this is a statistical parameter related to the training set. Thus, even when TD31 operator is given a chance to be selected, the algorithm will reject it.

In this case, only the electrostatic descriptor, namely, Coulomb MQSM, will be employed to construct the predictive model. The optimal QSAR model is found when the SDEP coefficient for the test set is the highest, and this does not necessarily coincide with the highest q^2 value for the training set. Combinations of overlap and Coulomb operators improve the q^2 cross-validated coefficient but not the SDEP. The best results are found when six PCs are used, yielding $q^2 = 0.832$ and $r^2 = 0.939$ for the 21-steroid set. The standard deviation of the errors of prediction for the test set is SDEP $= 0.526$. The predicted values for each compound are shown in Table 6, where they are also compared to the results found by other approaches. In this model molecule **31** is a pronounced outlier, as in the rest of the cases. If we use the optimal result found in the previous section, that is, matrixes and weights given in Table 4 (case $\nu = 2$, $k = 6$), the SDEP coefficient decreases to a very low value: SDEP $= 0.234$ for the test set. The influence of the fluorine atom is removed. Nevertheless, it is important to remark that these selected matrixes and the corresponding weights were obtained *with the knowledge of the actual CBG affinity of the steroid test set*, and therefore the derived predictive results are not significant.

Table 6. Comparison of TQSAR Predictions for the Test Set with Other 3D QSAR Approaches[a]

steroid	actual activity	CoMFA[b]	compass	RSM	MS-WHIM	PARM	TQSAR
22	−7.512	−7.883	−7.062	−7.417	−7.300	−7.449	−7.237
23	−7.553	−7.430	−7.729	−7.646	−8.332	−8.037	−7.879
24	−6.779	−6.642	−6.462	−6.647	−6.821	−6.601	−6.648
25	−7.200	−7.705	−7.466	−6.905	−7.445	−6.015	−7.809
26	−6.144	−6.495	−5.994	−6.164	−6.121	−6.246	−6.832
27	−6.247	−6.962	−6.383	−6.841	−6.901	−5.742	−7.318
28	−7.120	−6.848	−6.625	−6.623	−6.532	−6.925	−7.363
29	−6.817	−6.816	−7.403	−6.550	−6.838	−6.100	−7.540
30	−7.688	−7.767	−7.741	−6.948	−7.860	−6.108	−7.628
31	−5.797	−7.793	−7.779	−6.116	−7.491	−5.991	−7.537
	SDEP	0.716	0.705	0.380	0.662	0.504	0.526

[a] SDEP coefficient is also given. [b] Corrected results from ref 26.

Validation of the TQSAR Model. Although linear regression models are less problematic than neural networks in relation to possible fortuitous chance correlations, since the scores are few and orthogonal, some kind of statistical validation of the model must be equally performed. In any predictive model, and TQSAR is not an exception, there exists the possibility that the selection of variables, and in our case, also the weights of the matrix descriptors, is performed in such a way that an excellent fit and even valuable predictions are obtained, even though no meaningful correlation between molecular structure and biological properties exist. To illustrate this fact, the common randomization test[76,78] has been adopted, in the same way as other authors.[19,29] In this technique the elements of the n-dimensional activity vector **y** are shuffled by an arbitrary number of random exchanges in their positions, and the QSAR model is constructed with the modified activities. In the current study, there have been generated 80 randomized affinity vectors. If a real structure−activity relationship exists, the best correlations will be obtained for the correctly ordered activities. Otherwise, if apparently valuable predictions are found when randomized responses are used, the model is not reliable. Different degrees of freedom of the model have been varied to construct the TQSAR models: the number of MLR parameters (k), the number of MQSM involved (μ), and the number of matrixes combined (ν). Figure 4 shows the representation of q^2 against r^2 for the real and randomized activities. In it, a clear separation between the two situations is observed. On one hand, statistical parameters for correctly ordered affinities account for the highest values in all the cases, being located at the top right corner of the plot. On the other hand, correlation and prediction coefficients for the random ordered activities present low values, as expected.

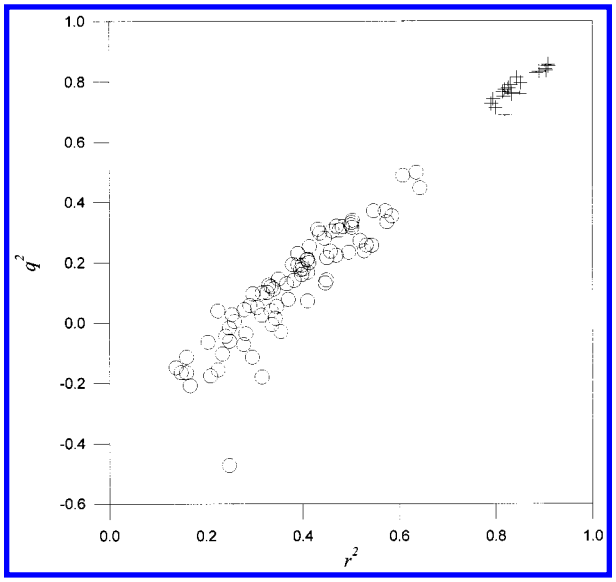

Figure 4. Plot of q^2 against r^2 for the real (+) and randomized (○) steroid activities.

Note that q^2 can achieve negative values when random shuffled activities are used. This happens when the PRESS factor is larger than SD (see eq 16); that is, when the property is better estimated by the mean of all the activities than by the model considered.

These results confirm the validity of tuned MQSM to generate satisfactory molecular descriptors and the robustness of principal axes regressions as statistical methods to extract valuable structure−property relationships.

Overview of Previous Results and Comparison with Other QSAR Studies. As has been indicated in the Introduction, the studied steroid family has become a frequently used training set for novel QSAR methodologies. Comparison with previous studies has to be made with caution due to the errors found by the Gasteiger group[20] in some steroid structures. In the same way, several studies do not take all 31 steroids used in the present paper as training set but only a reduced set of 21, considering the remaining 10 compounds as a test set. A brief survey of the works reported will be shown in the following. Studies dealing with larger steroid sets[79−83] will not be commented on here.

Cramer et al.[14] published the first 3D QSAR study on this data set. They proposed the comparative molecular field analysis (CoMFA) model on the reduced set of 21 steroids, obtaining a cross-validated statistic $q^2 = 0.662$ when two partial-least-squares (PLS) components were used.[84] Good et al.,[15] using the same set of 21 steroids, constructed a linear model on different molecular similarity matrixes using PLS statistics and a neural network pattern recognition technique. They reported an optimal result of $q^2 = 0.761$, obtained when the CoMFA-shape method and two PLS parameters were used. Oprea et al.[16] employed the minimal steric difference (MTD) method to correlate the steroid structure to their activity. The results for the 21 steroids were somewhat better than those attained by the CoMFA model, achieving $q^2 = 0.70$ and $r^2 = 0.93$ values. Nevertheless, this predictive coefficient value has to be considered with caution due to the fact that the cross-validation has not been performed by means of the leave-one-out method; instead the data were divided into two subseries of equal size. The authors also used the remaining 10-steroid set as a test set, obtaining valuable predictions. Jain et al.[17] presented the Compass program, which calculated the ligand properties only around the molecular surface and selected automatically the optimal conformation and alignment of the compounds in relation to the biological property studied. A neural network correlated the structural descriptors to the activities of the 21 steroids, and a $q^2 = 0.89$ value was found when steric and polar features were combined. Klebe et al.[18] described a variant of the CoMFA method, called CoMSIA (comparative molecular similarity indices analysis). In it, a mixture of steric, electrostatic, and hydrophobic similarity indices was used to predict the steroid binding affinities. This method overcomes grid dependencies of the results and quantifies the contribution of each descriptor. A $q^2 = 0.665$ value was obtained when four PLS components were used. Hahn and Rogers[19] proposed a receptor surface model (RSM) as a tool to capture information about receptor/ligand interactions, constructing a simple linear QSAR model based uniquely on the nonbonded energy of interaction between the ligand and the receptor surface model, the sum of the nonbonded van der Waals and electrostatic energies. A nonlinear variable selection technique, the genetic function approximation, was also used. They obtained a $q^2 = 0.628$ value. Wagener et al.[20] noticed the errors in the topology and stereochemistry of some steroids used in the previous studies and recompiled the data set from the original literature. They also presented a QSAR model based on the autocorrelation of molecular surface properties using a combination of Kohonen and feedforward neural networks. Their optimal result was $q^2 = 0.63$ for the full set of 31 steroids. Silverman and Platt[21] utilized the moments of molecular mass and charge distributions as source of QSAR descriptors in an alignment-free method called CoMMA (comparative molecular moment analysis). They studied both the reduced set of 21 steroids ($q^2 = 0.828$ with three PLS components) and the entire set of 31 ($q^2 = 0.689$ with six PLS components). Kellogg et al.[22] derived a 3D QSAR field based on the electrotopological state formalism, the E state, together with an index describing the polarity of hydrogens, the HE state. The classic 21-steroid set was examined with this methodology, and the best model was found when both E and HE states were combined, with $q^2 = 0.803$ and $r^2 = 0.979$ when three PLS components were used. Anzali et al.[23] improved the Wagener method, which used a self-organizating Kohonen neural network to analyze surface properties. In this study, the CBG affinities are not predicted quantitatively but in a qualitative way. Thus, the data activity is divided into three classes of comparable size, and a Kohonen neural network is used for mapping the molecular surface properties into two dimensions. The 2D Kohonen maps provide a useful tool to elucidate the structural factors responsible for biological activity. Norinder[24] created an approach called TDQ (three-dimensional QSAR) that combined two descriptions: the first one, related to the shape and hydrogen-bonding properties, similar to the Compass characterization, and the second one, the usual nonbonding and electrostatic properties used in the field analyses. From all the results reported, the two PLS component model yielding $q^2 = 0.68$ and $r^2 = 0.90$ has been selected. Although these values are worse than those attained by Compass, better predictions for the test set are found ($q^2 = 0.82$ when molecule **31** is excluded). Schnitker et al.[25]

studied the benchmark problem by means of the EGSITE (energy and geometry of SITE models) technique. This method is an evolution of the Voronoi approach to binding site modeling. EGSITE avoids dependence on the molecular alignment and does not include any kind of fitting, neither PLS nor nonlinear neural network algorithms. In contrast, this model fixes an error interval for each point and then attempts to find a set of simultaneous binding modes for all molecules within the error bars. Hydrophobicity data, molar refractivities, and Gasteiger partial charges were used as input data. Bravi et al.[26] proposed the MS−WHIM approach, where molecular surface (MS) properties were taken as a source of the weighted holistic invariant molecular (WHIM) indices. These indices are dealt with by PCA/PLS analysis and optional factorial and fractional factorial design strategies. A value of $q^2 = 0.631$ is obtained when two parameters are used. Predictions for the test set are satisfactory, except for molecules **23** and **31**. This method overcomes possible problems due to molecular alignment, since it is rotation- and translation-invariant. Turner et al.[27] presented a novel molecular descriptor (EVA) based upon calculated infrared range vibrational frequencies and used it in a QSAR study. This descriptor is also invariant to both rotation and translation of the structures concerned. Among the data sets studied, the authors tried to correlate the steroid CBG affinity. They found $q^2 = 0.83$ when two PLS components were employed. Parretti et al.[28] described an algorithm for molecular alignment based upon Monte Carlo optimization of molecular similarity indices (Carbó and Hodgkin−Richards indices). The similarity matrixes are treated with PLS technique and used to correlate the 21-steroid set activity and to perform a prediction of it for the 10-steroid test set. The optimal results are obtained when the electrostatic descriptor is used, yielding $q^2 = 0.780$ and $r^2 = 0.819$ values when only one PLS component is used. When the entire set (31 steroids) is used, $q^2 = 0.734$ and $r^2 = 0.764$ values are obtained. It is important to notice that none of the above studies describes correctly molecule **31**, considered explicitly by almost all of them as an outlier.

So and Karplus[29] reported recently a study where the predictive results were substantially improved. They used a combination of shape and electrostatic similarity matrixes dealt with by a nonlinear genetic neural network, and they obtained a cross-validated statistic of $q^2 = 0.94$ with six descriptors. This result is far from the ones obtained by any of the previous works, and it is the first study where molecule **31** is not found to be an outlier. Lobato et al.[12] studied the entire steroid set with two different methodologies: MQSM, a first stage of the present study, and topological quantum similarity indices (TQSI). In reference to the first method, a MQSM matrix was transformed by means of principal component analysis, and a MLR was built to correlate the activities. No selection of axes was performed. An optimal $q^2 = 0.705$ value was found when three PCs were used. On the other hand, the second model was constructed by selecting the optimal subset of TQSI such that the binding affinities are best adjusted. This occurred when four variables were used, yielding a $q^2 = 0.775$ value. This last method also removed the influence of the fluorine atom. Tominaga and Fujiwara[30] reported a study where the 21-steroid set was examined. They presented a method for improving the variable selection in PLS, consisting of an exclusion of those grid points in the CoMFA fields that contributed less to the predictive abilities of the model. The optimal results, once the possible overfitted models were excluded, were found when three components were used, yielding $q^2 = 0.807$ and $r^2 = 0.892$. Finally, Chen et al.[31] improved Hahn's QSAR approach based uniquely on the ligand−receptor energy of interaction, applying a genetic evolved algorithm to produce an atomic-level pseudoreceptor model (PARM). They studied the original 21-steroid set, achieving a $q^2 = 0.806$ value. This model was used to predict the remaining 10 steroid activities, obtaining good predictions.

Table 7. Cross-Validated Coefficients for the Different QSAR Models[a]

QSAR model	q^2	steroid set	QSAR model	q^2	steroid set
CoMFA	0.662 (2)	21	MS-WHIM	0.631 (2)	21
similarity matrixes	0.761 (2)	21	EVA	0.83 (2)	21
MTD	0.70	21	similarity indices	0.780 (1)	21
compass	0.89	21	similarity indices	0.734 (1)	31
CoMSIA	0.665 (4)	21	genetic N.N.	0.94 (6)	31
RSM	0.628	21	MQSM	0.705 (3)	31
Wagener's	0.63	31	TQSI	0.775 (3)	31
CoMMA	0.828 (3)	21	Tominaga's	0.807 (3)	21
CoMMA	0.689 (6)	31	PARM	0.806	21
E-state	0.803 (3)	21	TQSAR	0.832 (6)	21
TDQ	0.68 (2)	21	TQSAR	0.842 (6)	31

[a] In parentheses appear the number of components of the regressions, if any. For a brief description of the methods, see text.

All these results are summarized in Table 7, where the comparison can be more easily performed.

TQSAR Approach as an Alternative to Grid-Based QSAR Models. The present QSAR methodology, TQSAR based on MQSM, obtains comparable results to other highly predictive QSAR models, even when they use more sophisticated nonlinear neural network techniques. However, *the advantages of the model proposed are mainly related to obtaining the molecular descriptors*. Thus, quantum similarity matrixes constructed as indicated in the first section are better than those derived from the construction of stereoelectronic maps, done by the evaluation of relevant molecular energy fields in a 3D rectilinear grid surrounding the compound, since they overcome problematic dependencies on grid parameters (size, spacing, location), on the kind of charge selected, on the probe atom chosen, and on the relative orientation of the molecule into the grid. TQSAR also overcomes the molecular alignment problem, since the alignment of the molecules is generated by optimizing their similarity. This sort of free alignment is similar to Parretti's description.[28] Moreover, overlap and Coulomb MQSM are more consistent shape and electrostatic descriptors, respectively, than the usual measurement of the Lennard-Jones 6−12 and Coulomb potentials at the grid points, because these exhibit singularities at atomic positions, which are gotten around by setting the potentials at these points to zero. They also avoid the truncation at arbitrarily fixed cutoff values near the atomic positions due to the extremely large contributions of the potentials there. Result consistency in grid-based QSAR approaches is a problem discussed by various authors,[85−87] and different improvements to overcome these drawbacks have been proposed.[18,88−92] In this sense, the TQSAR method based on MQSM is close to the

CoMSIA spirit,[18] where these problems were exposed and tried to be solved. The higher authority of the structural descriptors used can also be extended to those QSAR methods dealing with other molecular effects, such as lipophilicity, inductive effects, resonance effects, etc. As indicated by Dunn et al.,[84] chemists still do not agree on how to define these effects, and the only "objective" molecular descriptors are the collective electron distributions, which are precisely the source of the MQSM, and subsequently of the TQSAR parameters. This can also be connected with the modern discussion on the significance of the reality of the wave functions and the role of density functions as quantum object descriptors.[93−95]

In addition, TQSAR models elaborated following the previous indications allow a quantitative determination of the contribution of both shape and electrostatic descriptors in the form of the coefficients weighting their associated quantum similarity matrixes. This method is more easily interpretable than other mixtures lacking any systematic process that appear in other QSAR methods.

FINAL CONCLUSIONS

The TQSAR approach, a new QSAR model based on tuned MQSM, has been described. The method is based on the premise that "similar" molecules will possess "similar" properties. In most cases, the relevant features for activity are given just by molecular fragments (the pharmacophore), and the rest of the molecule does not contribute significatively to the interaction. This fact is especially true when one deals with congeneric compounds, as in the steroid case. Molecular similarity is able to quantify the resemblance between the elements of a molecular set, and when dealing with chemicals of a same family, it focuses the information on the fragmental differences.

A multilinear regression has been used to correlate the structural descriptors, the principal axes of the quantum similarity matrixes, with the biological activities of a standard molecular set: 31 steroids that binds to the CBG receptor. Satisfactory correlations and predictions have been obtained, justifying the usefulness of the method. These results are comparable to those attained with other QSAR methods, including nonlinear ones. The randomization test has confirmed the real predictive character of the TQSAR approach and, definitively, of the MQSM, stating that the results obtained are not due to chance correlations.

A new application for TD MQSM has been also proposed: they can be used in QSAR models as corrective molecular descriptors, to reduce the statistical influence of some molecules poorly described. This work also states that simple multilinear regressions are enough to achieve good adjustments and predictions, indicating the validity of MQSM as robust sources of QSAR parameters. The n (31) columns of the tuned similarity matrix turn out to be as effective as the thousand of columns needed for CoMFA or similar grid-based approaches. As has been widely discussed, the TQSAR method depicts a useful alternative to grid-based QSAR models, that avoids the problems derived from the dependence of the solution on the parameters of the grid.

Further, it must be stressed that the whole set of procedures reported here are based on the application of quantum mechanics principles and current statistical tools. Thus, no more arbitrary elements than those present in these basic descriptors are introduced.

Finally, we admit that only one example of application is not enough to judge realistically a novel QSAR model, and for this reason more studies on different situations (other molecular systems, biological properties, or chemical mechanisms) are currently in progress in our laboratory.

ACKNOWLEDGMENT

This research has been partially supported by CICYT Grant SAF 96-0158 and the Fundació Maria Francesca de Roviralta. We thank Dr. J. Gasteiger and co-workers for providing the molecular geometries of the family studied in this paper. One of us (D.R.) also wants to gratefully acknowledge Dr. C. M. Cuadras for his enlightening advice and comments on the statistical aspects of the work. Thanks also to the Centre de Supercomputació de Catalunya (CESCA) and the Centre Europeu de Paral·lelisme de Barcelona (CEPBA) for a generous amount of computation time.

REFERENCES AND NOTES

(1) Hansch, C.; Fujita, T. ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.
(2) Martin, Y. C. *Quantitative Drug Design. A Critical Introduction*; Marcel Dekker: New York, 1978.
(3) Carbó, R.; Arnau, J.; Leyda, L. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189.
(4) Wold, S.; Johansson, E.; Cocchi, M. PLS−Partial Least-Squares Projections to Latent Structures. In *3D QSAR in Drug Design*; Kubinyi, H., Ed.; ESCOM: Leiden, The Netherlands, 1993; pp 523−550.
(5) Tetko, I. V.; Luik, A. I.; Poda, G. I. Application of Neural Networks in Structure−Activity Relationships of a Small Number of Molecules. *J. Med. Chem.* **1993**, *36*, 811−814.
(6) Ajay, A. A Unified Framework for Using Neural Networks To Build QSARs. *J. Med. Chem.* **1993**, *36*, 3565−3571.
(7) Stuper, A. J.; Jurs, C. Classification of Psychotropic Drugs as Sedatives or Tranquilizers Using Pattern Recognition Techniques. *J. Am. Chem. Soc.* **1975**, *97*, 182−187.
(8) Kowalski, B. R.; Bender, C. F. Pattern Recognition. II. Linear and Nonlinear Methods for Displaying Chemical Data. *J. Am. Chem. Soc.* **1973**, *95*, 686−693.
(9) McFarland, J. W.; Gans, D. J. Cluster significance analysis contrasted with three other quantitative structure−activity relationship models. *J. Med. Chem.* **1987**, *30*, 46−49.
(10) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I. Fuzzy Adaptive Least Squares and Its Application to Structure−Activity Studies. *Quant. Struct.-Act. Relat.* **1992**, *11*, 325−331.
(11) Fradera, X.; Amat, L.; Besalú. E.; Carbó-Dorca, R. Application of Molecular Quantum Similarity to QSAR. *Quant. Struct.-Act. Relat.* **1997**, *16*, 25−32.
(12) Lobato, M.; Amat, L.; Besalú, E.; Carbó-Dorca, R. Structure−Activity Relationships of a Steroid Family using Quantum Similarity Measures and Topological Quantum Similarity Indices. *Quant. Struct.-Act. Relat.* **1997**, *16*, 465−472.
(13) Amat, L.; Robert, D.; Besalú, E.; Carbó-Dorca, R. Molecular Quantum Similarity Measures Tuned 3D QSAR: An antitumoral family validation study. *J. Chem. Inf. Comput. Sci.* **1998**, *39*, 624−631.
(14) Cramer, R. D.; Patterson, 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.
(15) Good, A. C.; So, S.-S.; Richards, W. G. Structure−Activity Relationships from Molecular Similarity Matrixes. *J. Med. Chem.* **1993**, *36*, 433−438.
(16) Oprea, T. I.; Ciubotariu, D.; Sulea, T. I.; Simon, Z. Comparison of the Minimal Steric Difference (MTD) and Comparative Molecular Field Analysis (CoMFA) Methods for Analysis of Binding of Steroids to Carrier Proteins. *Quant. Struct.-Act. Relat.* **1993**, *12*, 21−26.
(17) Jain, A. N.; Koile, K.; Chapman, D. Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark. *J. Med. Chem.* **1994**, *37*, 2315−2327.

(18) 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.
(19) Hahn, M.; Rogers, D. Receptor Surface Models. 2. Application to Quantitative Structure−Activity Relationships Studies. *J. Med. Chem.* **1995**, *38*, 2091−2102.
(20) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic *Ah* Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.
(21) Silverman, B. D.; Platt, D. E. Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR without Molecular Superposition. *J. Med. Chem.* **1996**, *39*, 2129−2140.
(22) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. E-state fields: Applications to 3D QSAR. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 513−520.
(23) Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polanski, J. The comparison of geometric and electronic properties of molecular surfaces by neural networks: Application to the analysis of corticosteroid-binding globulin activity of steroids. *J. Comput.-Aided Mol. Design* **1996**, *10*, 521−534.
(24) Norinder, U. 3D-QSAR Investigation of the Tripos Benchmark Steroids and some Protein-Tyrosine Kinase Inhibitors of Styrene Type using the TDQ Approach. *J. Chemom.* **1996**, *10*, 533−545.
(25) Schnitker, J.; Gopalaswamy, R.; Crippen, G. M. Objective models for steroid binding sites of human globulins. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 93−110.
(26) Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. MS−WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 79−92.
(27) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. Evaluation of a novel infrared range vibration-based descriptor (EVA) for QSAR studies. 1. General application. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 409−422.
(28) Parretti, M. F.; Kroemer, R. T.; Rothman, J. H.; Richards, W. G. Alignment of Molecules by the Monte Carlo Optimization of Molecular Similarity Indices. *J. Comput. Chem.* **1997**, *18*, 1344−1353.
(29) So, S.-S.; Karplus, M. Three-Dimensional Quantitative Structure−Activity Relationships from Molecular Similarity Matrixes and Genetic Neural Networks. 1. Method and Validations. *J. Med. Chem.* **1997**, *40*, 4347−4359.
(30) Tominaga, Y.; Fujiwara, I. Prediction-Weighted Partial Least-Squares Regression Method (PWPLS) 2: Application to CoMFA. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1152−1157.
(31) Chen, H.; Zhou, J.; Xie, G. PARM: A Genetic Evolved Algorithm To Predict Bioactivity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 243−250.
(32) Carbó, R.; Domingo, L. LCAO-MO Similarity Measures and Taxonomy. *Int. J. Quantum Chem.* **1987**, *23*, 517−545.
(33) Besalú, E.; Carbó, R.; Mestres, J.; Solà, M. Foundations and Recent Developments on Molecular Quantum Similarity. *Top. Curr. Chem.* **1995**, *173*, 31−62.
(34) Carbó, R.; Besalú, E. Theoretical Foundations of Quantum Molecular Similarity. In *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*; Carbó, R., Ed.; Kluwer: Dordrecht, The Netherlands, Amsterdam, 1995; pp 3−30.
(35) Carbó-Dorca, R.; Besalú, E.; Amat, L.; Fradera, X. Quantum Molecular Similarity Measures: concepts, definitions, and applications to quantitative structure−property relationships. In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1996; Vol. 1, pp 1−42.
(36) Löwdin, P. O. Quantum Theory of Many-Particle Systems. I. Physical Interpretations by Means of Density Matrixes, Natural Spin−Orbitals, and Convergence Problems in the Method of Configurational Interaction. *Phys. Rev.* **1955**, *97*, 1474.
(37) McWeeny, R. The density matrix in many-electron quantum mechanics. I. Generalized product functions. Factorization and physical interpretation of the density matrixes. *Proc. R. Soc. London* **1959**, *A253*, 242.
(38) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *B136*, 864.
(39) Carbó, R.; Calabuig, B.; Besalú, E.; Martínez, A. Triple Density Molecular Quantum Similarity Measures: A General Connection Between Theoretical Calculations and Experimental Results. *Mol. Eng.* **1992**, *2*, 43−64.
(40) Constans, P.; Carbó, R. Atomic Shell Approximation: Electron Density Fitting Algorithm Restricting Coefficients to Positive Values *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1046−1053.
(41) Constans, P.; Amat, L.; Fradera, X.; Carbó-Dorca, R. Quantum Molecular Similarity Measures (QMSM) and the Atomic Shell Approximation (ASA). In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1996; Vol. 1, pp 187−211.
(42) Amat, L.; Carbó-Dorca, R. Quantum Similarity Measures under Atomic Shell Approximation: First-Order Density Fitting using elementary Jacobi Rotations. *J. Comput. Chem.* **1997**, *18*, 2023−2039.
(43) ASA coefficients and exponents can be seen and downloaded from the WWW site http://iqc.udg.es/cat/similarity/ASA/funcset.html.
(44) Constans, P.; Amat, L.; Carbó-Dorca, R. Towards a Global Maximization of the Molecular Similarity Function: Superposition of Two Molecules. *J. Comput. Chem.* **1997**, *18*, 826−846.
(45) Carbó-Dorca, R. Tagged Sets, Convex Sets and Quantum Similarity Measures. *J. Math. Chem.* **1998**, *23*, 353−364.
(46) Carbó-Dorca, R. Fuzzy Sets and Boolean Tagged Sets. *J. Math. Chem.* **1997**, *22*, 143−147.
(47) Mezey, P. G. *Shape in Chemistry: an introduction to molecular shape and topology*; VCH: New York, 1993.
(48) Hodgkin, E. E.; Richards, W. G. Molecular similarity based on electrostatic potential and electric field. *Int. J. Quantum Chem.* **1987**, *14*, 105−110.
(49) Kubinyi, H. A general view on similarity and QSAR studies. In *Computer-Assisted Lead Finding and Optimization: Current Tools for Medicinal Chemistry*; van de Waterbeemd, H., Testa, B., Folkers, G., Eds.; VHCA: Basel, Switzerland, and Wiley-VCH: Weinheim, Germany, 1997; pp 7−28.
(50) Rekker, R. F. The Hydrophobic Fragmental Constants. Its Derivation and Application. A Means of Characterization Membrane Systems. In *Pharmacochemistry Library*; Nauta, W. T.; Rekker, R. F., Eds.; Elsevier: New York, 1977; Vol. 1.
(51) Exner, O. In *Advances in Free Energy Relationships*; Chapman, N. B.; Shorter, J., Eds.; Plenum: New York, 1972; p 1.
(52) Verloop, A.; Hoogenstraaten, W.; Tipker, J. In *Drug Design*; Ariens, E. J., Ed.; Academic Press: New York, 1976; Vol. VII, p 165.
(53) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure−Property Modeling. In *Reviews in Computational Chemistry II*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1991; pp 367−422.
(54) Solà, M.; Mestres, J.; Oliva, J. M.; Duran, M.; Carbó, R. The Use of Ab Initio Quantum Molecular Self-Similarity Measures to Analyze Electronic Charge Density Distributions. *Int. J. Quantum Chem.* **1996**, *58*, 361−372.
(55) Amat, L.; Carbó-Dorca, R.; Ponec, R. Molecular Quantum Similarity Measures as an Alternative to Log P Values in QSAR Studies. *J. Comput. Chem.* **1998**, *19*, 1575−1583.
(56) Carbó, R.; Calabuig, B.; Vera, L.; Besalú, E. Molecular Quantum Similarity: Theoretical framework, ordering principles and visualization techniques. *Adv. Quantum Chem.* **1994**, *25*, 253−313.
(57) Mardia, K. V.; Kent, J. T.; Bibby, J. M. *Multivariate Analysis*; Academic Press: London, 1979; p 397.
(58) Van de Waterbeemd, H. Chemometric Methods Used in Drug Discovery. In *Structure−Property Correlations in Drug Research*; Van de Waterbeemd, H., Ed.; Academic Press: San Diego, CA, 1996; pp 61 and 65.
(59) Mardia, K. V. Some properties of classical multidimensional scaling. *Commun. Statist. Theor. Methods* **1978**, *A7*, 1233−1241.
(60) Eckart, C.; Young, G. Approximation of one matrix by another of lower rank. *Psychometrika* **1936**, *1*, 211−218.
(61) Cuadras, C. M.; Arenas, C. A distance based regression model for prediction with mixed data. *Commun. Statist. Theor. Methods* **1990**, *19*, 2261−2279.
(62) Cuadras, C. M.; Arenas, C.; Fortiana, J. Some computational aspects of a distance-based model for prediction. *Commun. Statist. Simul.* **1996**, *25*, 593−609.
(63) Kshirsagar, A. M. *Multivariate Analysis*; Marcel Dekker: New York, 1972.
(64) Carbó, R.; Besalú, E.; Amat, L.; Fradera, X. Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-properties relationships (QSPR). *J. Math. Chem.* **1995**, *18*, 237−246.
(65) Montgomery, D. C.; Peck, E. A. *Introduction to Linear Regression Analysis*; John Wiley: New York, 1992.
(66) Allen, D. M. The prediction sum of squares as a criterion for selecting variables. Technical Report 23, Department of Statistics, University of Kentucky, 1971.
(67) Allen, D. M. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **1974**, *16*, 125−127.
(68) Good, A. C.; Peterson, S. J.; Richards, G. H. QSAR's from Similarity Matrixes. Technique Validation and Application in the Comparison of Different Similarity Evaluation Methods. *J. Med. Chem.* **1993**, *36*, 2929−2937.

(69) Carbó, R.; Besalú, E. Definition, mathematical examples and quantum chemical applications of nested summation symbols and logical Kronecker deltas. *Comput. Chem.* **1994**, *18*, 117−126.
(70) Carbó, R.; Besalú, E. Definition and quantum chemical applications of nested summation symbols and logical functions: Pedagogical artificial intelligence devices for formulae writing, sequential programming and automatic parallel implementation. *J. Math. Chem.* **1995**, *18*, 37−72.
(71) Amat, L.; Robert, D.; Besalú, E. TQSAR−SIM. Institute of Computational Chemistry, University of Girona: Girona, Spain, 1997.
(72) Dunn, J. F.; Nisula, B. C.; Rodbard, D. Transport of Steroid Hormones: Binding of 21 Endogenous Steroids to Both Testosterone-Binding Globulin and Corticosteroid-Binding Globulin in Human Plasma. *J. Clin. Endocrin. Metab.* **1981**, *53*, 58−68.
(73) The steroid structures can be seen and downloaded in the Gasteiger's group Website: http://schiele.organik.uni-erlangen.de/services/steroids/.
(74) These same structures can also be found in the IQC Website: http://iqc.udg.es/cat/similarity/QSAR/steroids/index.htm.
(75) Stouch, T. R.; Jurs, P. C. Computer-Aided Studies of the Structure−Activity Relationships between the Structure of Some Steroids and Their Antiinflamatory Activity. *J. Med. Chem.* **1986**, *29*, 2125−2136.
(76) Wold, S.; Eriksson, L. Statistical validation of QSAR results. In *Chemometric Methods in Molecular Design*; van der Waterbeemd, H., Ed.; VCH Publishers: New York, 1995; Vol. 2, pp 309−318.
(77) Robert, D.; Carbó-Dorca, R. Analyzing the Triple Density Molecular Quantum Similarity Measures with the INDSCAL model. *J. Chem. Inf. Comput. Sci.* **1998**, *39*, 620−623.
(78) Fisher, R. The Principles of Experimentation, Illustrated by a Psycho-Physical Experiment. In *The Design of Experiments*, 8th ed.; Hafner Publishing: New York, 1966.
(79) Norinder, U.: Experimental design based 3D-QSAR analysis of steroid-protein interactions: Application to human CBG complexes. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 381−389.
(80) Norinder, U. 3D-QSAR analysis of steroid/protein interactions: The use of difference maps. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 419−426.
(81) Rum, G.; Herndon, W. C. Molecular Similarity Concepts. 5. Analysis of Steroid-Protein Binding Constants. *J. Am. Chem. Soc.* **1991**, *113*, 9055−9060.
(82) Simon, Z.; Bohl, M. Structure−Activity Relations in Gestagenic Steroids by the MTD Method. The Case of Hard Molecules and Soft Receptors. *Quant. Struct-Act. Relat.* **1992**, *11*, 23−28.
(83) Waszkowycz, B.; Clark, D. E.; Frenkel, D.; Li, J.; Murray, C. W.; Robson, B.; Westhead, D. R. PRO_LIGAND: An approach to de novo molecular design. 2. Design of novel molecules from molecular field analysis (MFA) models and pharmacophores. *J. Med. Chem.* **1994**, *37*, 3994−4002.
(84) Dunn, W. J.; Wold, S.; Edlund, U.; Hellberg, S. Multivariate structure−activity relationships between data from a battery of biological tests and an ensemble of structure descriptors: the PLS method. *Quant. Struct.-Act. Relat.* **1984**, *3*, 131−137.
(85) Folkers, G.; Merz, A.; Rognan, D. CoMFA: Scope and Limitations. In *3D QSAR in Drug Design*; Kubinyi, H., Ed.; ESCOM: Leiden, The Netherlands, 1993; pp 583−618.
(86) Cramer III, R. D.; DePriest, S. A.; Patterson A.; Hecht, P. The Developing of Comparative Molecular Field Analysis. In *3D QSAR in Drug Design*; Kubinyi, H., Ed.; ESCOM: Leiden, The Netherlands 1993; pp 443−485.
(87) Simon, Z. Comparative molecular field analysis. Critical comments. *Rev. Roum. Chem.* **1992**, *37*, 323−325.
(88) Cho, S. J.; Tropsha, A. Cross-Validated R^2-Guided Region Selection for Comparative Molecular Field Analysis: A Simple Method To Achieve Consistent Results. *J. Med. Chem.* **1995**, *38*, 1060−1066.
(89) Kroemer, T. R.; Hecht, P. Replacement of Steric 6−12 Potential-Derived Interaction Energies by Atom-Based Indicator Variables in CoMFA Leads to Models of Higher Consistency. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 205−212.
(90) Greco, G.; Novellino, E.; Silipo, C.; Vittoria, A. Comparative Molecular Field Analysis on a Set of Muscarinic Agonists. *Quant. Struct.-Act. Relat.* **1991**, *10*, 289−299.
(91) Klebe, G.; Abraham, U. On the Prediction of Binding Properties of Drug Molecules by Comparative Molecular Field Analysis. *J. Med. Chem.* **1993**, *36*, 70−80.
(92) Sulea, T.; Oprea, T. I.; Muresan, S.; Chan, S. L. A Different Method for Steric Field Evaluation in CoMFA Improves Model Robustness. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1162−1170.
(93) Bell, J. S. *Speakable and unspeakable in Quantum Mechanics*. Cambridge University Press: Cambridge, England, 1993.
(94) Goldstein, S. Quantum Theory Without Observers—Part One. *Phys. Today*, **1998** (March), 42−46.
(95) Goldstein, S. Quantum Theory Without Observers—Part Two. *Phys. Today*, **1998** (April), 38−42.

CI980410V