Three-Dimensional Quantitative Similarity−Activity Relationships (3D QSiAR) from SEAL Similarity Matrices

Hugo Kubinyi,*,† Fred A. Hamprecht,‡ and Thomas Mietzner†

Main Laboratory, BASF AG, D-67056 Ludwigshafen, Germany, and Department of Chemistry, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland

Received October 27, 1997

The program SEAL is suited to describe the electrostatic, steric, hydrophobic, and hydrogen bond donor and acceptor similarity of different molecules in a quantitative manner. Similarity scores A_F can be calculated for pairs of molecules, using either a certain molecular property or a sum of weighted properties. Alternatively, their mutual similarity can be derived from distances *d* or covariances *c* between SEAL-based property fields that are calculated in a regular grid. For a set of *N* chemically related molecules, such values form an $N \times N$ similarity matrix which can be correlated with biological activities, using either regression analysis and an appropriate variable selection procedure or partial least-squares (PLS) analysis. For the Cramer steroid data set, the test set predictivities ($r^2_{pred} = 0.53-0.84$) of different PLS models, based on a weighted sum of molecular properties, are superior to published results of CoMFA and CoMSIA studies ($r^2_{pred} = 0.31-0.40$), regardless of whether a common alignment or individual, pairwise alignments of all molecules are used in the calculation of the similarity matrices. Training and test set selections have a significant influence on the external predictivities of the models. Although the SEAL similarity score between two molecules is a single number, its value is based on the 3D properties of both molecules. The term 3D quantitative similarity−activity analyses (3D QSiAR) is proposed for approaches which correlate 3D structure-derived similarity matrices with biological activities.

Introduction

Rational approaches in medicinal chemistry are based on the consideration of relationships between chemical structures and biological activities. Lead structures, derived from natural products, biological concepts, or simply from screening hits, are structurally modified to optimize their biological activity, selectivity, and pharmacokinetic properties and to minimize their toxic and other side effects. To this end, structural elements are removed, added, or changed, e.g., by an isosteric replacement of atoms or groups, the variation of chain lengths, the exchange of linkers, the rigidization of a flexible molecule by introduction of bulky groups or additional ring systems, etc.[1−3] The underlying hypothesis of all structure−activity relationships is that similar molecules exert similar biological activities in a qualitative sense, i.e., in their mode of action, as well as quantitatively. Several dedicated investigations have provided evidence that the different structural elements within a molecule most often contribute in an additive manner to relative biological activities, expressed in a logarithmic scale. This additivity is based on the relationship between the free energies of ligand binding, ΔG, and the equilibrium constants of ligand binding to a protein, *K* (eq 1; ΔH = enthalpy and ΔS = entropy contributions). If the free energy ΔG of the ligand−protein interaction is an additive function of all individual interactions of the different parts of a ligand with its binding site, then also log *K* values are an additive molecular property.

$$\Delta G(\text{ligand}-\text{protein}) = \Delta H - T\Delta S = 2.303RT \log K \quad (1)$$

In quantitative structure−activity relationships (QSAR), biological activities are described either in terms of indicator variables that encode the presence or absence of certain chemical groups (Free−Wilson analysis) or in terms of physicochemical parameters (Hansch analysis).[4−8] Hansch analysis will predict similar activities of two molecules if they are similar in their physicochemical properties. It will predict different activities if they are dissimilar, with the only exception of nonlinear dependence on a certain physicochemical property; this dependence is frequently observed if transport and distribution in the biological system play an additional role, for example, in cell culture or animal studies.[4,5]

Three-dimensional quantitative structure−activity relationships (3D QSAR), especially comparative molecular field analysis (CoMFA),[9] correlate binding affinities and other biological activities with steric, electrostatic, and other 3D fields of the molecules. For this purpose, electronic properties of all compounds within a chemically related series are calculated and 3D structures are generated. A pharmacophore hypothesis is derived, and all molecules are superimposed in their hypothetical bioactive conformations to achieve a mutual alignment. A regular lattice is placed around the superimposed molecules in such a manner that the resulting box is in all directions several Ångstroms larger than the combined volume of all molecules. Then,

* To whom all correspondence should be addressed. E-mail: hugo.kubinyi@msm.basf-ag.de.
† BASF AG.
‡ ETH.

S0022-2623(97)00732-2 CCC: $15.00
Published on Web 06/16/1998

different molecular fields are calculated at each grid point of the lattice; the default distance between the grid points is 2 Å. Probe atoms or groups, e.g., a neutral carbon atom (probe for van der Waals interactions), a charged atom (probe for electrostatic interactions), a hydrogen bond donor or acceptor (probes for hydrogen bond interactions), are used to calculate interaction energy values at every grid point, for each molecule; the mathematical functions used for van der Waals and electrostatic interactions are the Lennard−Jones and Coulomb potentials, respectively.

Afterward, the data set is split into a training set for which a CoMFA model is derived and a test set to determine the external predictivity of the resulting model. In the next step, the molecular fields are correlated with the biological activities, using partial least-squares (PLS) analysis,[10,11] with or without variable selection. Theory, methods, and applications, as well as recent advances of 3D QSAR methods, have been reviewed (e.g., refs 12, 13).

Due to the relationships between molecular similarities and the corresponding variations in biological potencies, different quantitative expressions of chemical similarities have been investigated (e.g., refs 14−16). A common similarity measure is the Tanimoto coefficient T (eq 2) which compares certain features A of molecule 1 and B of molecule 2 with the features C that are common to both molecules. By its definition, T equals unity if both molecules are identical and its value is zero if they have nothing in common.

$$T = \frac{C}{A + B - C} \tag{2}$$

Indices which describe the electronic similarity of two molecules (A and B) are the Carbo index R_{AB}[17] and the Hodgkin index H_{AB},[18] based on the electron densities ρ_A and ρ_B of both molecules. In a more general form, any molecular property P can be used to calculate these similarity indices (eqs 3 and 4). As compared to the Hodgkin index, the Carbo index is more sensitive to the shape of the molecules than to the actual values of the corresponding property of both molecules. Modifications of these indices have been proposed to calculate electrostatic similarity values from grid-based fields (e.g., ref 19).

$$R_{AB} = \frac{\int P_A P_B \, dn}{(\int P_A^2 \, dn)^{1/2} (\int P_B^2 \, dn)^{1/2}} \tag{3}$$

$$H_{AB} = \frac{2\int P_A P_B \, dn}{\int P_A^2 \, dn + \int P_B^2 \, dn} \tag{4}$$

A different approach for the description of the similarity of molecules was chosen by Kearsley and Smith in their program SEAL.[20] SEAL defines a "similarity score" A_F between two molecules (A and B) in any relative orientation to each other (eq 5); r_{ij} is the distance between atoms i and j, α defines the distance dependence, w_E, w_S, etc., are user-attributed values to give different weights to electrostatic, steric, and other properties, q_i and q_j are partial charges at the atoms i (molecule A) and j (molecule B), and v_i and v_j are arbitrary powers (default = 3) of the van der Waals radii of atoms i (molecule A) and j (molecule B). Any other property, e.g., atom- or group-based hydrophobicity values or hydrogen bond donor and acceptor properties, may be added to the definition of w_{ij}.[21,22]

$$A_F = -\sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij} \, e^{-\alpha r_{ij}^2}; \; w_{ij} = w_E q_i q_j + w_S v_i v_j + \ldots \tag{5}$$

SEAL similarity scores that are based on weighted combinations of steric, electrostatic, and hydrophobic properties have been used to perform an objective and automated alignment of molecules, starting from several random orientations of two molecules relative to each other and reorienting one molecule with respect to the other one to achieve the largest mutual similarity score A_F (eq 5).[21] The advantage of this approach, as compared to several other methods, is its objective character. Despite the atom-based calculation procedure, the resulting alignment does not depend on the exact superposition of certain atoms. The method has been adapted to a flexible multiple alignment with simultaneous 3D structure optimization.[21,22] The calculations can be accelerated by the use of "united atoms", e.g., by combining all hydrogen atoms with their adjacent "heavy" atoms or by creating even larger "superatoms" (e.g., the center of an aromatic ring instead of all individual ring atoms).[22] The version used in this investigation considers, in addition to electrostatic, steric, and hydrophobic properties, also hydrogen bond donor and acceptor positions in the alignment of the molecules.[22]

The Lennard−Jones and Coulomb potentials that are typically used in CoMFA studies create relatively "hard" fields. They change their values from close to zero to very large numbers within a few tenths of an Ångstrom, i.e., within a small fraction of the commonly used grid distance of 2 Å. Cutoff values have to be defined to avoid values which would approach infinity at the atom centers. Significantly different CoMFA results are sometimes obtained after a shift or rotation of the box in which the fields are calculated, especially if the default grid distance of 2 Å is used and PLS analysis is performed with a variable selection procedure.[23] Several modifications of the CoMFA method have been proposed to avoid such problems.[13,23−26]

In a recently developed 3D QSAR method, comparative molecular similarity indices analysis (CoMSIA),[27,28] SEAL functions (eq 5) were used to calculate similarity fields of the molecules to different probe atoms and groups, in the same manner as CoMFA fields are calculated. Since the CoMSIA fields are based on Gaussian potentials, they are much "softer" than the CoMFA functions. In addition, they are good approximations to the cutoff-corrected Lennard−Jones and Coulomb potentials. Correspondingly, CoMSIA fields produce smoother contour maps than CoMFA fields.[27] Although not yet confirmed, it is to be expected that CoMSIA results are more or less invariant to variations of the box orientation.

The principle that QSARs are based on similarity hypotheses[29] seems to be trivial. However, so far only a few investigations have been performed on the rela-

tionships between molecular similarities and biological activities. In 1991, Rum and Herndon used $N \times N$ similarity matrices (N = number of compounds), derived from 2D topological descriptors, and stepwise regression analysis to correlate several columns of these matrices with biological activities.[30] Two years later, Good and Richards[31,32] performed systematic investigations to correlate 3D electronic similarities of molecules with their biological activities.[33] In this approach, first 3D structures of all molecules are generated and aligned in space, as in CoMFA studies. Then, similarity values between all pairs of molecules are calculated, using either the Carbo or the Hodgkin electrostatic potential similarity indices (eqs 3 and 4). Alternatively, Gaussian approximations of the Carbo and Hodgkin indices are used. In the last step, the resulting $N \times N$ similarity matrices are correlated with the biological activities of the molecules, using either neural networks[31] or PLS analysis.[32] In PLS analysis, better results are obtained after the elimination of irrelevant variables (i.e., variables not contributing to prediction) by application of the variable selection program GOLPE.[34]

At about the same time, Kubinyi used $N \times N$ lipophilicity distance matrices **D** ($d_{ij} = |\log P_i - \log P_j|$; P = *n*-octanol/water partition coefficient) to describe linear and nonlinear lipophilicity−activity relationships in a quantitative manner.[4] In fit and internal predictivity, the resulting models are as good as those using parabolic or bilinear functions.[4,29] Martin et al. confirmed the suitability of distance matrices **D** for the quantitative description of various nonlinear property-property relationships.[35] If **X** is the original $N \times M$ matrix of explanatory data (N rows, M columns), then all x_{ik} values are normalized (i.e., mean-value-centered and standardized) before, column by column. Afterward, every column has a mean value of 0 and a standard deviation of 1. An Euclidean distance matrix **D** is calculated using eq 6 (d_{ij} = distance between two molecules; M = number of physicochemical properties; x_{ik} and x_{jk} = properties k of two different molecules); alternatively, covariance matrices **C** (eq 7) can be used to derive quantitative structure−activity models.[29]

$$d_{ij} = \sqrt{\sum_{k=1}^{k=M} (x_{ik} - x_{jk})^2} \qquad (6)$$

$$c_{ij} = \sum_{k=1}^{k=M} x_{ik} \cdot x_{jk} \qquad (7)$$

A detailed investigation of several published data sets showed that Euclidean distance matrices **D** as well as covariance matrices **C** are well-suited to derive quantitative structure−activity relationships, starting from one or several physicochemical properties.[29] The correlation of one or very few columns of such matrices with the biological activity values produces quantitative models as good as or even better than classical Hansch regression models. In the case of nonlinear relationships between certain physicochemical properties and biological activities, distance matrices **D** are clearly superior to covariance matrices **C**.[29]

So far, similarity matrices were generated from 2D properties or from electrostatic properties, using Carbo and Hodgkin indices (e.g., refs 30−32, 36). The aim of this investigation is to describe the 3D similarity of molecules, using SEAL scores or SEAL similarity fields, and to derive 3D quantitative similarity−activity relationships (3D QSiAR) from $N \times N$ similarity matrices to prove whether such models have certain advantages as compared to other approaches. A new, important aspect of the current investigation is the concept of pairwise superpositions instead of a common alignment of all molecules.

The Cramer steroids (Table 1)[11] are chosen as a practical example. Although this data set is far from being perfectly suited for structure−activity studies, it has become a benchmark to investigate scope and limitations of new methods.[37] Special emphasis was given to a correct input of structures[38] and biological data, to a proper validation of all results, and to the prediction of the biological activities of different test sets.

Table 1. Structures and CBG Affinities of the Cramer Steroid Data Set[11,37,38]

no.	steroid	log *K* CBG
1	aldosterone	6.279
2	androstanediol	5.000
3	androstenediol	5.000
4	androstenedione	5.763
5	androsterone	5.613
6	corticosterone	7.881
7	cortisol	7.881
8	cortisone	6.892
9	dehydroepiandrosterone	5.000
10	deoxycorticosterone	7.653
11	deoxycortisol	7.881
12	dihydrotestosterone	5.919
13	estradiol	5.000
14	estriol	5.000
15	estrone	5.000
16	etiocholanolone	5.255
17	pregnenolone	5.255
18	17-hydroxypregnenolone	5.000
19	progesterone	7.380
20	17-hydroxyprogesterone	7.740
21	testosterone	6.724
22	prednisolone	7.512
23	cortisol 21-acetate	7.553
24	pregn-4-ene-3,11,20-trione	6.779
25	epicorticosterone	7.200
26	19-nortestosterone	6.144
27	16α,17-dihydroxypregn-4-ene-3,20-dione	6.247
28	16α-methylpregn-4-ene-3,20-dione	7.120
29	10-norprogesterone	6.817
30	11β,17,21-trihydroxy-2α-methylpregn-4-ene-3,20-dione	7.688
31	11β,17,21-trihydroxy-2α-methyl-9α-fluoropregn-4-ene-3,20-dione	5.797

Methods

Data Set. The structures of 31 steroids and their corticosteroid binding globulin (CBG) affinities were taken from ref 11; wrong structures were corrected according to refs 37, 38. Biological data are given in Table 1.

Geometries and Electronic Properties. First, 2D structures were sketched, using the program ISIS/Draw 2.1,[39] and controlled for correct stereochemistry. These were converted to 3D structures with CORINA, version 1.8.[40] Partial atomic charges were assigned using the AM1 method as implemented in MOPAC 7.[41] Electrostatic and hydrophobic properties were calculated from partial atomic charges and atomic hydrophobicities, respectively.[21,42] van der Waals radii were raised to

the third power to compute steric properties.[20,21] Hypothetical positions of hydrogen bond partners were generated in favorable geometries around the different donor and acceptor groups of the molecules, to determine their hydrogen bond donor and acceptor similarity.[22]

Alignments and Calculation of SEAL Similarity Scores. Generally, two kinds of alignment were performed: a biased one, based on a least-squares atom-by-atom match, and an objective one, using a SEAL-derived scoring function.

(a) In the first case, the rigid alignment, all 31 steroids were superimposed onto an unsubstituted steroid template (generated, consistent with the other compounds, via ISIS/Draw and CORINA), using an atom-by-atom least-squares fit as implemented in the SYBYL FIT option.[43] Following the ordinary steroid nomenclature, the carbons 5, 10, 13, and 14 (the atoms common to the A/B and C/D ring systems) were selected as fit centers. This relative orientation of all molecules was not changed in the subsequent analyses; different SEAL similarity scores (see below) were calculated in these orientations to generate the corresponding $N \times N$ similarity matrices (cf. Tables 4 and 5).

(b) In the second case, the SEAL alignments, only pairwise alignments were performed. One of each pair of structures was taken as a template for the superposition; for the second molecule of each couple, 50 start orientations (out of a much larger number of random start orientations), showing already a good overlap of hydrophobic and hydrophilic regions with corresponding regions of the reference molecule, were selected automatically. Then all-properties SEAL similarity scores A_F (eq 5) were calculated for each start orientation, using default parameters for the distance dependence α, the weights for the hydrophobic, electronic, and steric contributions, w_L, w_E, and w_S,[21] and the weight for hydrogen bond partner positions, w_H.[22] These default parameters resulted from an independent calibration, derived from 190 ligand pairs which mutually bind to the same protein. The experimental alignments of all ligand pairs could be reproduced in this investigation with an accuracy below 0.7 Å in about 33%, below 1.0 Å in 51%, and below 2.0 Å in nearly 90% of the cases.[22] These deviations have to be compared with the inherent accuracy limit of about 0.7 Å for the superposition of two experimentally determined ligand−protein complexes (for more details, see refs 21, 22).

Guided by the SEAL score A_F, we optimized the superpositions of the steroids until no further improvements in these values could be achieved.[21,22] Minor variations of the different weights did not significantly influence the SEAL similarity scores; on the other hand, single properties (e.g., hydrophobic or steric vs electronic or hydrogen bond properties) produced different alignments of certain pairs of steroids (see Results and Discussion).

All A_F values of the $N \times (N + 1)/2 = 496$ molecule pairs (for control purposes also the self-to-self alignments of identical molecules were performed) were normalized to $A_F(\max) = 1$,[21] in the same manner as the Carbo index (eq 3). Symmetrical $N \times N$ similarity matrices **S** were derived from different SEAL scores.

SEAL Similarity Fields and Calculation of Distance and Covariance Matrices. For both kinds of alignments, hydrophobic, steric, and electrostatic properties as well as their weighted combination (including also hydrogen bond donor and acceptor terms) were used to generate SEAL similarity fields of individual molecules as described in ref 27. Hydrogens were united with their heavy atoms. A box size of 24 × 24 × 24 Å was chosen to allow for a rotation of the steroids within the box, without the need to change the dimensions of the box. With a grid distance of 1 Å, 25 × 25 × 25 values were calculated for each field and all steroids; the resulting 15 625 values of the different molecular fields were concatenated into single vectors and subsequently normalized (i.e., their means were set to 0 and their standard deviations to 1). $N \times N$ distance matrices **D** and covariance matrices **C** of the different SEAL fields were computed from these field vectors, using eqs 6 and 7.

Box Rotation. As mentioned before, a marked dependence of CoMFA results on the orientation of the grid box has been observed in some studies.[23] To investigate the sensitivity of the field-based SEAL similarity approach, all pairs of steroids were rotated by 30°, 60°, ... 150° around all axes simultaneously. Since this procedure changed the orientation of each atom with respect to the surrounding grid points, no independent translations were performed.

Regression Analysis with Variable Selection. All regression models with up to three variables were investigated, and Fisher F values were used as the criterion for the "best" models. This statistical parameter especially favors small numbers of independent variables (eq 8; r = multiple correlation coefficient; n = number of compounds; k = number of independent variables); thus, it significantly reduces the risk of chance correlations, as compared to a standard deviation criterion. In addition, the significance of all variables and the mutual intercorrelations within the X variables were checked for each model.

$$F = \frac{r^2(n - k - 1)}{k(1 - r^2)} \tag{8}$$

Although there are highly efficient variable selection procedures, based on evolutionary[44] and genetic algorithms,[45] systematic search is extremely fast if only a few variables are involved (e.g, ref 46). In the current investigation (31 X variables), the F values of all 4991 models with up to three different X variables are calculated on a PC within a few seconds, using eq 9 ($r_{Y,(X1,...Xm)}$ = multiple correlation coefficient; $\mathbf{r}_{YX}$ = vector of r_{YXi} correlation coefficients; $\mathbf{R}_{XX}$ = matrix of $r_{Xi,Xj}$ correlation coefficients).[47]

$$r_{Y,(X1,...,Xm)} = (\mathbf{r}_{YX}{}^T \, \mathbf{R}_{XX}{}^{-1} \, \mathbf{r}_{YX})^{1/2} \tag{9}$$

Cross-Validation. Cross-validation procedures eliminate one or several data sets (i.e., compounds) from the training set, derive a quantitative model from the remaining objects, and predict the activity for the one or several objects which were not included in the derivation of the model.[48−50] The cross-validated squared correlation coefficient, Q^2, and the standard deviation of the predictions, s_{PRESS}, are calculated from the predictive residual sum of squares, PRESS = $\Sigma(y_{pred} - y_{obs})^2$, in the same manner as r^2 and s values are calculated from the unexplained variance $\Sigma(y_{calc} - y_{obs})^2$, to describe the quality of fit of the models. In the leave-one-out procedure, only one object is eliminated at a time and the process is repeated until all objects have been eliminated once and only once. For larger data sets, the elimination of several objects at a time, randomly or in a systematic manner, is recommended. Throughout this investigation, only leave-one-out cross-validation was performed to derive a measure of the internal predictivity of the models, within the training set.

y Randomization. As a much more reliable criterion for the risk of chance correlations, the affinity values of the steroids were re-ordered in a random manner (**y** scrambling), to determine the percentage of chance correlations that are as good as or even better than the best correlations found for the y values in their correct order; 100 different randomization runs were performed routinely, and the best models were individually derived for each randomization by systematic search, including up to three columns of the $N \times N$ matrices in the regression models. In some cases, even 1000 randomizations were performed to prove that 100 randomizations give representative results. No attempt was made to eliminate y-scrambled data sets with (fortuitous) high correlations between the real, original y values and the randomized y values.

Test Set Predictions. In the steroid data set, most often the steroids **1**−**21** are chosen as the training set and steroids **22**−**31** as the test set to determine the external predictivity of the models, expressed by r^2_{pred} and s_{pred}.[11] In a prior investigation, also another training set, steroids **1**−**12** and

Table 2. CoMFA[11] and CoMSIA Results[27] for the CBG Affinities (Table 1; training set, steroids **1**−**21**; test set, steroids **22**−**31**

analysis/statistical parameters	rigid alignment[a,b]		SEAL alignment[a]	
	CoMFA[b]	CoMSIA[a]	CoMFA[a]	CoMSIA[a]
fit (training set):				
r^2	0.897	0.941	0.947	0.937
s	0.397	0.320	0.303	0.330
cross-validation (training set):				
Q^2	0.662	0.662	0.598	0.665
s_{PRESS}	0.719	0.763	0.832	0.759
external predictivity (test set):				
r^2_{pred}	0.309[c]	nd	0.36	0.40

[a] Common alignment of all molecules.[27] [b] Reference 11. [c] The value of 0.65, given by Cramer et al.,[11] is wrong.[27,29]

23−**31**, was selected; compounds **13**−**22** served as the test set.[29] Both training and test set selections, called training and test sets I and II, respectively, are considered in this study.

PLS Analyses. Partial least-squares (PLS) analyses of the different data sets were performed, using the cross-validated s_{PRESS} value as the criterion for the number of significant vectors; up to five latent variables were calculated in each PLS analysis, but only up to three latent variables were allowed in the final model. Starting with one vector, the number of latent variables was derived from the first minimum of s_{PRESS}. In most cases, only one or two latent variables proved to be significant. For better comparison with the regression analyses, variables were scaled in the PLS fit and rescaled in the cross-validation runs.

Results and Discussion

The corticosteroid binding globulin affinities of the Cramer steroid data set (Table 1)[11] have been investigated by several groups, using different methods. Unfortunately, many of these studies suffer from incorrect structures and/or biological data in the investigations and the publications (for a recent review, see ref 37). Thus, no exact comparisons of published results with this study, obtained from correct structures and data, are possible. Nevertheless, a rough comparison shows that squared correlation coefficients r^2 in the range of 0.90−0.95 and standard deviations s around 0.3−0.4 are observed for the fit of steroids **1**−**21**; for internal predictivity, Q^2 values around 0.6−0.7 and s_{PRESS} values around 0.7−0.8 result. However, predictions of the test set (steroids **22**−**31**) yielded only r^2_{pred} values in the range of 0.3−0.4 (Table 2). This poor external predictivity can be attributed, at least in part, to the fact that several structural features within this test set, i.e., 2α-methyl, 9α-fluorine, 16α-methyl, and 21-acetoxy groups, are not included in the training set.

An earlier analysis of the steroid data showed that the different A/B ring junctions of the steroid skeletons are sufficient to explain the structure−activity relationships.[29] If a one-parameter Free−Wilson equation is derived for the training set (eq 10; the term 4,5-C═C- indicates the presence or absence of a cycloaliphatic 4,5-double bond in ring A of the steroids), the fit (r and s values) is worse than in the CoMFA and CoMSIA analyses (Table 2); however, the internal predictivity of this model, expressed by Q^2, is as good as the corresponding value of different CoMFA and other 3D QSAR studies.

$$\log K = 2.022(\pm 0.52)\ 4,5\text{-C═C-} + 5.186(\pm 0.36) \quad (10)$$

($n = 21$; $r = 0.882$; $s = 0.568$; $F = 66.41$; $Q^2 = 0.726$; $s_{PRESS} = 0.630$)

The test set predictivity of eq 10 is even better than the predictivity of published CoMFA and CoMSIA analyses (Table 2): $n = 10$; $r^2_{pred} = 0.477$; $s_{pred} = 0.733$. However, it is still relatively poor. In addition to the lack of certain structural features in the training set, weakly active compounds are somewhat under-represented in the test set. A better selection with respect to both problems seems to be, e.g., the choice of steroids **1**−**12** and **23**−**31** as the training set and steroids **13**−**22** as the test set (eq 11).[29] Despite a worse fit as compared to eq 10, the relevance of this model is proven by its excellent test set predictivity (compounds **13**−**22**): $n = 10$; $r^2_{pred} = 0.909$; $s_{pred} = 0.406$.

$$\log K = 1.667(\pm 0.75)\ 4,5\text{-C═C-} + 5.306(\pm 0.65) \quad (11)$$

($n = 21$; $r = 0.731$; $s = 0.697$; $F = 21.82$; $Q^2 = 0.454$; $s_{PRESS} = 0.754$)

Due to the fact that rigid alignments[11,27] and SEAL-derived alignments[27] were used in the CoMFA and CoMSIA analyses of this data set, both types of alignments were also performed in this investigation. The SEAL similarity scores of the self-to-self alignments can be taken as a criterion for the quality of the achieved superpositions. For all 31 steroids and both alignment procedures, the maximum values of 1.0000 were observed, as expected. For the pairwise alignments, the SEAL similarity scores depend on the differences in the structures, with a maximum for the pair **30/31** (differ only by a hydrogen/fluorine exchange in position 9α), $A_F = 0.9959$, and a minimum for the pair **15/23**, $A_F = 0.7995$ (compare Table 3). With only one exception, where the similarity score was slightly inferior, all SEAL-based objective alignments were better than or at least as good as the rigid alignments, indicating the efficiency of the automated SEAL-based alignments. Surprisingly, out of the $N \times (N + 1)/2 = 496$ different superpositions of the 31 steroids (weighted all properties), 79 pairs showed a head-to-tail alignment, i.e., ring D of one steroid was superimposed on ring A of the other steroid and ring A on ring D. This might be taken as an indication that some steroids, e.g. **9**, a 3-hydroxy-17-keto steroid, and **12**, a 3-keto-17-hydroxy steroid, are more closely related if identical functional groups are on top of each other than if the steroid skeletons are superimposed without a consideration of the differences in the hydrogen bond donor and acceptor functionalities (Figure 1). Fortunately, the similarity scores of these alignments do not differ too much from the best regular alignments. Thus, even if such head-to-tail alignments should be an artifact, this will not have a significant influence on the obtained results.

The observed head-to-tail superpositions demonstrate the importance of a weighted all-properties similarity score in the alignments; they are not obtained if only hydrophobic or steric properties are used in the align-

Table 3. Log *K* Values (CBG affinities) and Selected Similarity Vectors (*X* variables) of a Typical $N \times N$ Similarity Matrix **S** (SEAL alignment, all-properties similarity scores; eqs 13a−c)

no.	log *K*	*X-7*	*X-10*	*X-17*	*X-19*	*X-22*	*X-23*
1	6.279	0.9394	0.9601	0.8914	0.9181	0.9385	0.8723
2	5.000	0.8606	0.8905	0.9324	0.8947	0.8565	0.8205
3	5.000	0.8591	0.8862	0.9398	0.8888	0.8547	0.8195
4	5.763	0.8661	0.8977	0.9036	0.9223	0.8645	0.8456
5	5.613	0.8366	0.8693	0.8921	0.8929	0.8424	0.8382
6	7.881	0.9492	0.9784	0.8921	0.9404	0.9575	0.9062
7	7.881	1.0000	0.9344	0.8648	0.8965	0.9930	0.9308
8	6.892	0.9611	0.9460	0.8642	0.9080	0.9696	0.9149
9	5.000	0.8491	0.8815	0.9386	0.9029	0.8536	0.8309
10	7.653	0.9344	1.0000	0.9140	0.9624	0.9425	0.8863
11	7.881	0.9801	0.9554	0.8850	0.9166	0.9806	0.9180
12	5.919	0.8873	0.9282	0.9112	0.9360	0.8916	0.8597
13	5.000	0.8375	0.8721	0.8936	0.8704	0.8364	0.8000
14	5.000	0.8453	0.8515	0.8589	0.8444	0.8479	0.8075
15	5.000	0.8203	0.8698	0.8963	0.8957	0.8218	0.7995
16	5.255	0.8143	0.8292	0.8570	0.8505	0.8186	0.8139
17	5.255	0.8648	0.9140	1.0000	0.9459	0.8609	0.8431
18	5.000	0.9030	0.8789	0.9586	0.9081	0.9001	0.8851
19	7.380	0.8965	0.9624	0.9459	1.0000	0.9041	0.8914
20	7.740	0.9462	0.9183	0.9145	0.9513	0.9458	0.9259
21	6.724	0.8935	0.9156	0.8968	0.9278	0.8918	0.8590
22	7.512	0.9930	0.9425	0.8609	0.9041	1.0000	0.9393
23	7.553	0.9308	0.8863	0.8431	0.8914	0.9393	1.0000
24	6.779	0.8943	0.9333	0.9366	0.9703	0.8929	0.8743
25	7.200	0.9291	0.9728	0.8883	0.9360	0.9375	0.8817
26	6.144	0.8798	0.9215	0.9109	0.9316	0.8875	0.8616
27	6.247	0.9199	0.8591	0.8526	0.8883	0.9205	0.9014
28	7.120	0.9016	0.9483	0.9457	0.9848	0.9008	0.8803
29	6.817	0.8891	0.9563	0.9436	0.9934	0.8974	0.8854
30	7.688	0.9820	0.9428	0.8642	0.9050	0.9919	0.9383
31	5.797	0.9783	0.9372	0.8606	0.8997	0.9882	0.9351

ments. Whereas no experimental evidence for reverse binding modes of steroids is available (the 3D structure of the ligand-binding domain of the estrogen receptor has been published only recently[51]), bottom-up-top-down binding modes of some steroids to an antibody fragment have been observed.[52,53]

All alignments were based on the weighted combination of steric, electrostatic, hydrophobic, and hydrogen bond donor and acceptor properties. For the generation of the $N \times N$ similarity matrices **S**, pairwise SEAL scores of the steroids were also calculated from individual properties. Thus, for the rigid and the SEAL alignments, four different $N \times N$ matrices result in both series: hydrophobic, electrostatic, steric, and weighted hydrophobic/electrostatic/steric/donor/acceptor (all properties) similarity matrices.

Regression analyses were performed to correlate individual columns of the similarity matrices **S** with biological activities. To avoid chance correlations, the upper number of included variables (i.e., columns of the similarity matrices) in the regression models was restricted to three, the Fisher *F* value (eq 8) was used as the criterion to choose the best models, and the final models were checked for the significance of all regression variables and their mutual intercorrelations. A systematic search procedure (see Methods) was applied; in the training and test set selections, only columns of the similarity matrices **S**, for which the corresponding steroids were included in the training set, were also considered in the models. Typical models for the whole data set (steroids **1**−**31**) and the training sets I (**1**−**21**) and II (**1**−**12** and **23**−**31**) are presented in eqs 12a−c (rigid alignment; weighted all properties similarity matrix) and 13a−c (pairwise SEAL alignments; weighted all-properties similarity matrix; the values of the *X* variables of eqs 13a−c are listed in Table 3). The *X* variables correspond to columns of the similarity matrix **S**, i.e., *X-6* is the corticosterone similarity vector of all steroids, *X-7* is the cortisol similarity vector, and so on.

Rigid alignment, all compounds:

$$\log K = 15.56(\pm 3.92)\ X\text{-}6 - 7.494(\pm 3.51) \quad (12a)$$

($n = 31$; $r = 0.833$; $s = 0.608$; $F = 65.80$; $Q^2 = 0.658$; $s_{\mathrm{PRESS}} = 0.643$)

Training set I:

$$\log K = 14.70(\pm 3.86)\ X\text{-}7 - 6.684(\pm 3.38) \quad (12b)$$

($n = 21$; $r = 0.877$; $s = 0.578$; $F = 63.55$; $Q^2 = 0.731$; $s_{\mathrm{PRESS}} = 0.624$)

Test set I: $n = 10$; $r^2_{\mathrm{pred}} = 0.454$; $s_{\mathrm{pred}} = 0.748$

Training set II:

$$\log K = 16.15(\pm 5.59)\ X\text{-}6 - 8.077(\pm 5.08) \quad (12c)$$

($n = 21$; $r = 0.811$; $s = 0.597$; $F = 36.56$; $Q^2 = 0.592$; $s_{\mathrm{PRESS}} = 0.652$)

Test set II: $n = 10$; $r^2_{\mathrm{pred}} = 0.750$; $s_{\mathrm{pred}} = 0.674$

Pairwise SEAL alignments, all compounds:

$$\log K = 15.90(\pm 4.53)\ X\text{-}22 - 8.044(\pm 4.11) \quad (13a)$$

($n = 31$; $r = 0.800$; $s = 0.659$; $F = 51.66$; $Q^2 = 0.589$; $s_{\mathrm{PRESS}} = 0.704$)

Training set I:

$$\log K = 10.80(\pm 4.11)\ X\text{-}7 - 15.45(\pm 6.43)\ X\text{-}17 + 19.63(\pm 6.95)\ X\text{-}19 - 7.389(\pm 5.68) \quad (13b)$$

($n = 21$; $r = 0.958$; $s = 0.364$; $F = 63.55$; $Q^2 = 0.880$; $s_{\mathrm{PRESS}} = 0.440$)

Test set I: $n = 10$; $r^2_{\mathrm{pred}} = 0.598$; $s_{\mathrm{pred}} = 0.642$

Training set II:

$$\log K = 12.76(\pm 6.30)\ X\text{-}10 + 13.53(\pm 5.43)\ X\text{-}23 - 17.25(\pm 6.49) \quad (13c)$$

($n = 21$; $r = 0.881$; $s = 0.497$; $F = 31.05$; $Q^2 = 0.713$; $s_{\mathrm{PRESS}} = 0.562$)

Test set II: $n = 10$; $r^2_{\mathrm{pred}} = 0.784$; $s_{\mathrm{pred}} = 0.626$

What do eqs 12 and 13 tell us, in terms of structural similarity? Equation 12, starting from a rigid alignment and including all 31 steroids, shows that the similarity of the steroids to corticosterone, expressed by the similarity vector *X-6*, determines the binding affinities to CBG. If only the steroids of the training set I are considered, the similarity to cortisol, *X-7*, gives a slightly better model than *X-6*; using *X-6*, $r = 0.870$, $F = 59.18$, $Q^2 = 0.711$, and $r^2_{\mathrm{pred}} = 0.553$ result; thus, for external predictivity the *X-6* model is even slightly better than the *X-7* model. For training and test sets II, the fit and

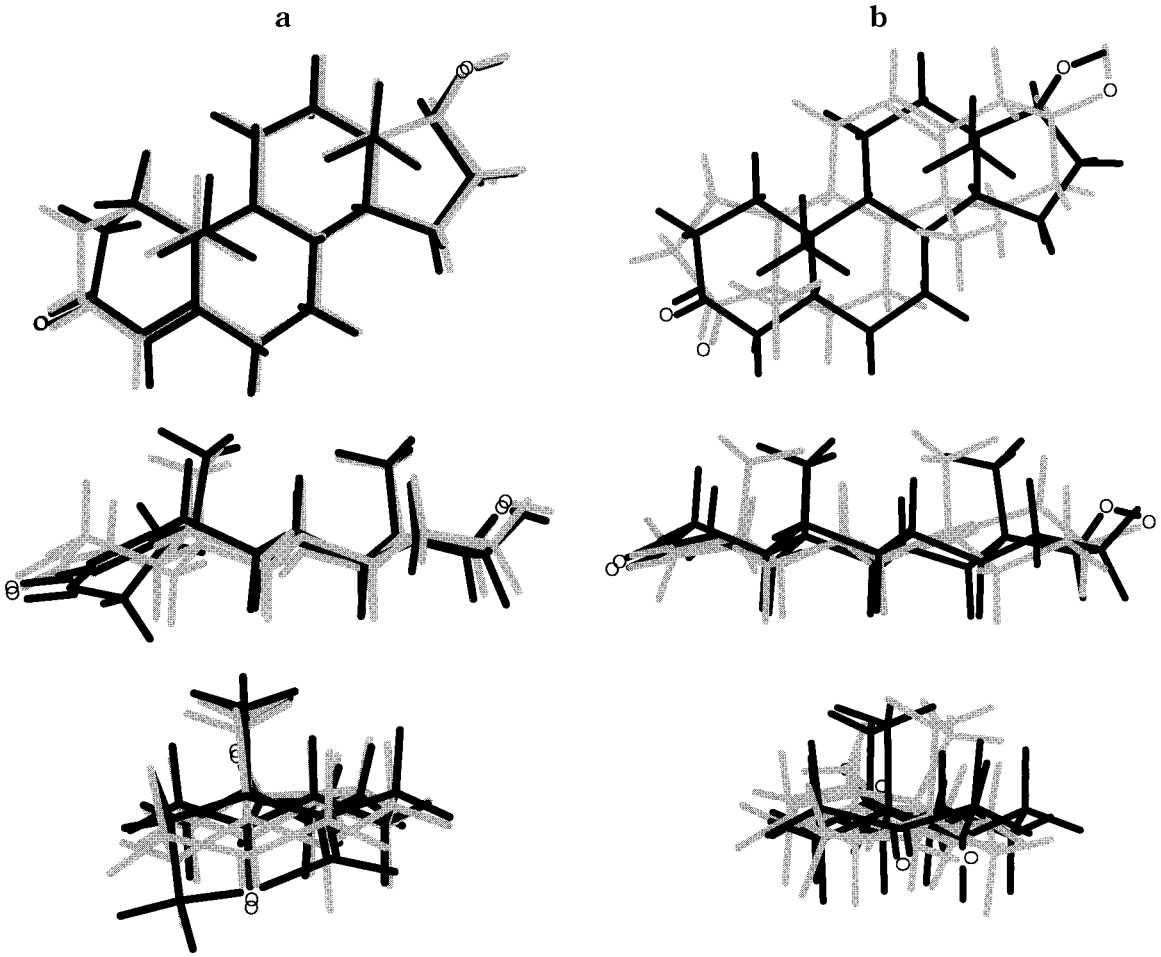

Figure 1. SEAL alignments of two pairs of steroids in three orthogonal views: upper diagrams, steroids viewed from above; middle diagrams, steroids viewed from the front; lower diagrams, steroids viewed along the axis C-3 to C-17; (a) steroids **12** and **21** (regular superposition; SEAL score 0.9872); (b) steroids **9** and **12** (head-to-tail superposition; SEAL score 0.9296).

internal predictivity of the best model is worse but the external predictivity is better (eq 12c). In general, the statistical parameters of eqs 13a−c correspond to those of eqs 12a−c; the variables of eqs 13b,c are not significantly interrelated (training set I: $r^2_{7,17} = 0.03$, $r^2_{7,19} = 0.20$, $r^2_{17,19} = 0.25$; training set II: $r^2_{10,23} = 0.07$). Figure 2 shows the deviations between predicted and observed log *K* values (pairwise SEAL alignments, all-properties similarity matrices; eqs 13a−c).

The internal and external predictivities (Q^2 and r^2_{pred} values) of all best regression models (up to three variables), starting from the two different alignment procedures and the different $N \times N$ similarity matrices, are summarized in Table 4 (upper part).

A closer inspection of these results indicates that in most cases hydrophobic, electrostatic, and steric as well as weighted all-properties similarity matrices yield comparable results. Only in one case a poor external predictivity is observed (SEAL alignment, training set II, hydrophobic similarity; $r^2_{\text{pred}} = 0.204$). For sufficient stability of the derived models the best selection seems to be a weighted combination of all properties.

For each analysis 100 runs with randomized *y* values were performed, again searching for the best models after re-ordering the *y* values in a random manner, to prove the significance of the observed models and to check the risk of chance correlations. For the eight different analyses that include all steroids, the fit and the internal predictivities of the real models were better than any of the 100 scrambled models, indicating a 100% confidence level. For the training and test sets I, the real models were better than 100% (internal predictivities) and 98−100% (external predictivities) of the *y*-scrambled models. For training and test sets II, the percentage rates for internal and external predictivities were 98−100% and 95−100%. Thus, all fit and predictivity parameters are statistically significant at a 95% level.

For the all-property analyses, all statistical parameters of the real analyses were better than any of the *y*-scrambled analyses. These numbers indicate a high stability of the obtained results with respect to the risk of chance correlations. An individual variable selection may produce, by chance, slightly better results than another variable selection (see discussion of eq 12b), but no models with fortuitous combinations of irrelevant variables are to be expected.

In addition to SEAL score-derived similarity matrices **S**, also distance matrices **D** and covariance matrices **C** were generated, starting from the different alignments and using the similarity field calculations implemented in CoMSIA.[27] In general, comparable results are observed for the fit of the whole data set and the training and test sets I and II. Table 4 shows that some internal predictivities are slightly worse than the corresponding results from the SEAL score-derived similarity matrices

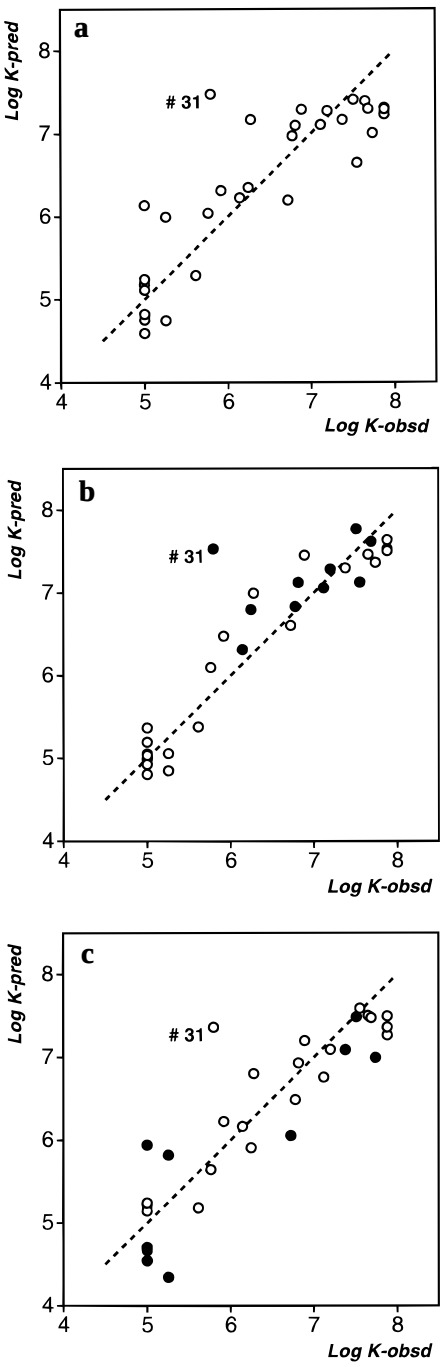

Figure 2. Comparison of predicted and observed log *K* values (pairwise SEAL alignments, all-properties similarity matrix, best regression models): (a) all compounds (eq 13a); (b) training and test sets I (eq 13b); (c) training and test sets II (eq 13c); open circles, leave-one-out cross-validation predictions; filled black circles, test set predictions. As in other published analyses of this data set (cf. ref 37), compound **31** is an outlier; the difference of 1.9 units in the log *K* values of compound **30** and its 9α-fluoro analogue **31** (Table 1) cannot be explained without special assumptions on the influence of the fluorine substituent on the binding affinities.

S. Some models (especially those derived from single-property-based similarity matrices) yield poor to unacceptable external predictivities; a negative r^2_{pred} value indicates that the predictions by the model are worse than taking the overall mean of the affinity values as the predictions (resulting in $r^2_{pred} = 0$). In one case even the all-properties-based external predictivity is much worse than for most other analyses (SEAL alignment, distance matrix; test set II, $r^2_{pred} = 0.146$). For those models also the *y*-randomization runs indicate a lack of statistical significance (i.e., percentage values smaller than 95%).

Table 4. Internal and External Predictivities of Regression Models, Derived from Different Similarity Matrices, Using All Compounds and the Training and Test Sets I and II (*F* criterion for the selection of models; only up to three variables were included in the models)

	all	test set I[a]		test set II[b]	
property	Q^2 (variables)	Q^2 (variables)	r^2_{pred}	Q^2 (variables)	r^2_{pred}
SEAL Similarity Matrices **S**, Rigid Alignment					
hydrophobic	0.785 (2)	0.785 (2)	0.684	0.545 (1)	0.655
electrostatic	0.687 (1)	0.846 (2)	0.534	0.677 (1)	0.736
steric	0.609 (1)	0.658 (1)	0.516	0.604 (1)	0.494
all properties[c]	0.658 (1)	0.731 (1)	0.454	0.592 (1)	0.750
SEAL Similarity Matrices **S**, Pairwise SEAL Alignments					
hydrophobic	0.552 (1)	0.635 (2)	0.716	0.704 (3)	0.204
electrostatic	0.628 (1)	0.842 (3)	0.411	0.583 (1)	0.557
steric	0.683 (2)	0.710 (2)	0.447	0.545 (1)	0.439
all properties[d]	0.589 (1)	0.880 (3)	0.598	0.713 (2)	0.784
Distance Matrices **D**, Rigid Alignment					
hydrophobic	0.701 (2)	0.664 (2)	0.752	0.508 (2)	0.823
electrostatic	0.601 (1)	0.860 (3)	0.530	0.198 (2)	0.795
steric	0.660 (1)	0.916 (3)	0.384	0.590 (1)	0.171
all properties	0.681 (2)	0.809 (3)	0.621	0.585 (2)	0.776
Distance Matrices **D**, Pairwise SEAL Alignments					
hydrophobic	0.590 (2)	0.542 (2)	0.746	0.565 (3)	0.656
electrostatic	0.529 (1)	0.597 (1)	0.350	0.746 (3)	0.755
steric	0.697 (3)	0.705 (1)	−0.114	0.228 (1)	0.559
all properties	0.634 (2)	0.721 (2)	0.401	0.329 (1)	0.146
Covariance Matrices **C**, Rigid Alignment					
hydrophobic	0.752 (3)	0.699 (3)	0.705	0.554 (2)	0.755
electrostatic	0.666 (1)	0.842 (2)	0.378	0.597 (1)	0.770
steric	0.626 (1)	0.853 (3)	0.599	0.629 (1)	0.545
all properties	0.535 (1)	0.848 (3)	0.571	0.475 (1)	0.632
Covariance Matrices **C**, Pairwise SEAL Alignments					
hydrophobic	0.422 (1)	0.593 (3)	0.680	0.433 (3)	0.146
electrostatic	0.601 (1)	0.651 (1)	0.481	0.561 (1)	0.662
steric	0.522 (1)	0.822 (3)	−0.191	0.572 (1)	0.483
all properties	0.767 (3)	0.836 (3)	0.611	0.654 (3)	0.855

[a] Training set, steroids **1**−**21**; test set, steroids **22**−**31**. [b] Training set, steroids **1**−**12** and **23**−**31**; test set, steroids **13**−**22**. [c] Eqs 12a−c. [d] Eqs 13a−c and Figure 2.

Models from SEAL field-derived distance and covariance matrices do not depend on the orientation of the molecules in the box, in which the fields are calculated. After a stepwise rotation of the steroids (see Methods), identical models with identical statistical parameters result after each individual rotation (numerical deviations occur only in the third decimal places), indicating that this method shows the desirable property of being box orientation-invariant.

PLS models are often more stable than regression models, especially if many highly interrelated variables are involved, as is the case for such $N \times N$ matrices. Thus, also PLS analyses were performed (Table 5). A comparison of Tables 4 and 5 provides evidence that in several cases inferior internal predictivities but better external predictivities result from the PLS models, as compared to the best regression models (Table 4). With one exception (distance matrix, SEAL alignment, steric field; test set I, $r^2_{pred} = 0.395$), all r^2_{pred} values are larger than 0.5, indicating a high degree of stability of the PLS models, as compared to the regression models. Variable

Table 5. Internal and External Predictivities of PLS Models, Derived from Different Similarity Matrices, Using All Compounds and the Training and Test Sets I and II (s_{PRESS} criterion for the selection of PLS components; only up to three PLS vectors included in the models)

	all	test set I[a]		test set II[b]	
property	Q^2 (vectors)	Q^2 (vectors)	r^2_{pred}	Q^2 (vectors)	r^2_{pred}
SEAL Similarity Matrices **S**, Rigid Alignment					
hydrophobic	0.667 (2)	0.631 (2)	0.709	0.579 (1)	0.759
electrostatic	0.679 (2)	0.750 (3)	0.515	0.634 (1)	0.675
steric	0.614 (1)	0.594 (1)	0.596	0.602 (1)	0.675
all properties	0.722 (1)	0.749 (1)	0.601	0.651 (1)	0.789
SEAL Similarity Matrices **S**, Pairwise SEAL Alignments					
hydrophobic	0.619 (1)	0.559 (1)	0.781	0.468 (1)	0.730
electrostatic	0.637 (1)	0.602 (1)	0.645	0.640 (1)	0.683
steric	0.605 (2)	0.624 (1)	0.648	0.582 (1)	0.535
all properties	0.722 (2)	0.768 (2)	0.660	0.622 (1)	0.754
Distance Matrices **D**, Rigid Alignment					
hydrophobic	0.722 (3)	0.590 (3)	0.777	0.373 (2)	0.644
electrostatic	0.737 (2)	0.806 (3)	0.502	0.641 (1)	0.781
steric	0.626 (1)	0.655 (1)	0.607	0.602 (1)	0.630
all properties	0.700 (1)	0.757 (1)	0.568	0.566 (1)	0.768
Distance Matrices **D**, Pairwise SEAL Alignments					
hydrophobic	0.605 (3)	0.490 (2)	0.715	0.354 (2)	0.625
electrostatic	0.696 (1)	0.691 (1)	0.645	0.668 (1)	0.664
steric	0.573 (1)	0.766 (3)	0.395	0.525 (1)	0.645
all properties	0.695 (2)	0.725 (1)	0.584	0.536 (1)	0.651
Covariance Matrices **C**, Rigid Alignment					
hydrophobic	0.650 (3)	0.526 (3)	0.701	0.458 (3)	0.685
electrostatic	0.694 (2)	0.701 (1)	0.576	0.607 (1)	0.754
steric	0.618 (1)	0.628 (1)	0.647	0.613 (1)	0.636
all properties	0.707 (2)	0.709 (1)	0.534	0.557 (2)	0.835
Covariance Matrices **C**, Pairwise SEAL Alignments					
hydrophobic	0.453 (2)	0.306 (1)	0.657	0.293 (2)	0.556
electrostatic	0.655 (1)	0.650 (1)	0.638	0.647 (1)	0.680
steric	0.586 (2)	0.491 (1)	0.507	0.494 (1)	0.539
all properties	0.748 (3)	0.723 (2)	0.659	0.612 (2)	0.762

[a] Training set, steroids **1**−**21**; test set, steroids **22**−**31**. [b] Training set, steroids **1**−**12** and **23**−**31**; test set, steroids **13**−**22**.

selection procedures did not significantly improve these results, especially with respect to the external predictivities.

If only the PLS analyses of the weighted all-properties similarity matrices are considered (Table 5), the external predictivities r^2_{pred} are between 0.53 and 0.66 for test set I and between 0.65 and 0.84 for test set II. These are excellent results, as compared to earlier CoMFA and CoMSIA studies, which produced test set I predictivities between 0.3 and 0.4 (Table 2).

In principle, one could expect that good internal predictivities (high Q^2 values) should be indicative for good external predictivities (high r^2_{pred} values). To analyze possible relationships between Q^2 and r^2_{pred} values, the results of all PLS analyses of the steroid data set were compared for the training and test sets I and II (Figure 3). Each point in the diagrams corresponds to one pair of Q^2 and r^2_{pred} values, resulting from different PLS analyses (including one to five latent variables), for all investigations listed in Table 5.

There is no relationship at all, a fact that was already observed in prior investigations.[25,54] High Q^2 values can be associated with very poor r^2_{pred} values and vice versa. The only possible conclusion is that external predictions are more stable (i.e., most r^2_{pred} values > 0.5) for the optimum number of PLS components (black dots in Figure 3) and that nearly all test set II predictivities are better than to be expected from the relatively poor internal predictivities of the models.

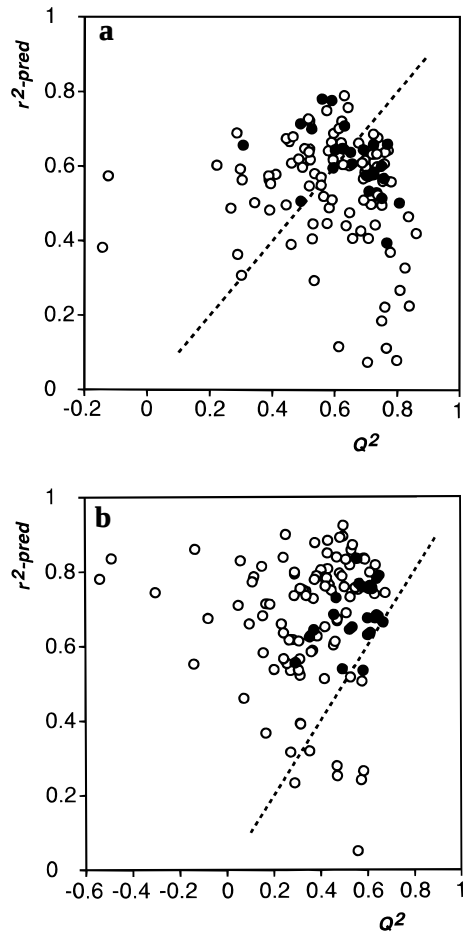

Figure 3. Relationships between Q^2 values of the training sets and r^2_{pred} values for the test sets. All 24 analyses of Table 5, with one to five latent variables in the PLS models (120 pairs of values), are included in the diagrams for the test and training sets I (a) and II (b); filled black circles, best models, selected by the s_{PRESS} criterion. The dashed lines indicate identical Q^2 and r^2_{pred} values, points in the upper left area indicate better r^2_{pred} values, and points in the lower right area indicate better Q^2 values. Whereas there is no relationship between the Q^2 and r^2_{pred} values, most r^2_{pred} values of the test set II are better than the cross-validated Q^2 values of the training set II.

Besides all advantages of 3D quantitative similarity−activity relationships, based on similarity matrices, there is one disadvantage, as compared to CoMFA and CoMSIA analyses: so far, no method could be developed which allows the calculation of 3D contour maps. If regression analyses are performed (e.g., eqs 12a−c and 13a−c), positive and negative coefficients of the similarity vectors to certain molecules within the series can be qualitatively interpreted, but the similarity coefficients cannot be decoded, i.e., projected onto the original property space of the molecules. In principle, it should be possible to approximate contour maps from the X variable regression coefficients (cf. eqs 12a−c and 13a−c), multiplied by the SEAL field coefficients of the corresponding molecules, but this has not yet been investigated in detail.

There is another problem, common to all quantitative structure−activity relationships. Sometimes similar

molecules have very different biological activities, caused by different ligand−protein interactions, different binding modes, or changes in the mode of action.[29,55] It will be a challenge for the future to extend 3D QSiAR also to such data sets.

Conclusions

It seems to be trivial that similar molecules show similar biological activities. Surprisingly enough, similarity matrices have not been used to derive QSARs until 1991.[30] Even then, only some investigations have been performed, and no pairwise alignments, instead of a common alignment of all molecules within a data set, have been performed. Pairwise superpositions are independent of all other alignments; thus, molecule 3 has no influence on the superposition of molecules 1 and 2, etc. This should offer a special advantage in the analysis of data sets where some compounds may have different conformational preferences or different binding modes. In a pairwise alignment procedure, such molecules do not cause any perturbation of the alignment of all other molecules and the similarity scores derived from these alignments. SEAL score-derived similarity matrices **S** do not depend on boxes and grids. Even if field-based $N \times N$ distance matrices **D** and covariance matrices **C** are calculated, the resulting analyses are invariant to box translation or rotation.

Considering different alignments, different matrices, and regression and PLS analyses, the following conclusions can be drawn from the results presented in Tables 4 and 5:

(1) There are no significant differences between a common alignment of all molecules and a pairwise SEAL alignment; this may be typical for rigid molecules as the steroids are. For flexible molecules, pairwise alignments might be more appropriate and could offer advantages.

(2) Weighted all-properties similarity matrices yield more stable regression models than models based on individual properties, especially if field-derived distance or covariance matrices are used; some individual property-derived distance and covariance matrices yield highly unstable results.

(3) In general, regression analyses produce models with good internal predictivities and fairly good external predictivities.

(4) PLS analyses tend to produce more stable models, with inferior internal predictivity but better external predictivity than regression analyses; there seem to be no significant differences between SEAL score-derived similarity matrices **S**, distance matrices **D**, and covariance matrices **C**.

(5) The external predictivities of all-properties-based PLS models, using either similarity matrices **S**, distance matrices **D**, or covariance matrices **C**, are much better ($r^2_{\text{pred}} = 0.53-0.84$; Table 5) than published CoMFA and CoMSIA results ($r^2_{\text{pred}} = 0.31-0.40$; Table 2).

(6) Variable selection did not produce better PLS models.

This investigation confirms also the crucial dependence of test set predictivities on the training set selection and the lack of a relationship between internal and external predictivities:

(1) Internal predictivities are most often better for training set I than for training set II, whereas the external predictivities for test set II are nearly always better than for test set I. The most important conclusion from this observation is that a broad variety of structural features should be covered by the training set molecules in order to allow reliable predictions.

(2) No conclusions on the test set predictivity can be derived from a mere consideration of the training set statistical parameters, at least if leave-one-out cross-validation is performed; more reliable predictions may result from larger data sets, by performing cross-validation in groups. Only a rigorous limitation of the number of variables or PLS vectors produces stable models and reliable external predictions, at the expense of the quality of fit and internal predictivity in the cross-validation runs.

Further data sets should be investigated to demonstrate scope and limitations of 3D quantitative similarity−activity relationships, based on pairwise alignments and correlation of the biological data with SEAL-derived similarity matrices **S**, **D**, or **C**.

Acknowledgment. Stimulating discussions with Jens Sadowski and Ute Abraham are gratefully acknowledged. Several other colleagues contributed by making manuscripts available before publication. Critical comments of an unknown reviewer helped to improve the quality and statistical validity of the results.

Note Added in Revision. After submission of this manuscript, a paper on three-dimensional quantitative structure−activity relationships from molecular similarity matrices and genetic neural networks (GNN) was published.[56] Using the steroid data set, better r_{FIT} and Q^2 values were obtained for the whole data set; the results were justified by cross-validation and *y* randomization runs. However, this study has two limitations: first, the Q^2 values are no independent measure of the validity of the models because they were used as the fitness criterion in the GNN variable selection runs; second, no training and test set selections were performed to check the external predictivities of the derived models.

References

(1) Hansch, C., Sammes, P. G., Taylor, J. B., Eds. *Comprehensive Medicinal Chemistry. The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds*; Pergamon Press: Oxford, 1990.
(2) Wolff, M. E., Ed. *Burger's Medicinal Chemistry*, 5th ed.; John Wiley & Sons: New York, 1995; Vol. 1.
(3) Wermuth, C. G., Ed. *The Practice of Medicinal Chemistry*; Academic Press: London, 1996.
(4) Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; VCH: Weinheim, 1993.
(5) Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
(6) van de Waterbeemd, H., Ed. *Chemometric Methods in Molecular Design*; VCH: Weinheim, 1995.
(7) van de Waterbeemd, H., Ed. *Advanced Computer-Assisted Techniques in Drug Discovery*; VCH: Weinheim, 1995.
(8) van de Waterbeemd, H., Ed. *Structure−Property Correlations in Drug Research*; Academic Press, R. G. Landes Co.: Austin, TX, 1996.
(9) Cramer, R. D., III; 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.

(10) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III. The Collinearity Problem in Linear Regression. The Partial Least Squares Approach to Generalized Inverses. *SIAM J. Sci. Stat. Comput.* **1984,** *5*, 735−743.
(11) Cramer, R. D., III. Partial Least Squares (PLS): Its Strengths and Limitations. *Perspect. Drug Discovery Des.* **1993**, *1*, 269−278.
(12) Kubinyi, H., Ed.; *3D QSAR in Drug Design. Theory, Methods and Applications*; ESCOM Science Publishers B.V.: Leiden, 1993.
(13) Kubinyi, H., Folkers, G., Martin, Y. C., Eds., *3D QSAR in Drug Design. Volume 2. Ligand-Protein Interactions and Molecular Similarity, Volume 3. Recent Advances*; Kluwer/ESCOM: Dordrecht, 1998.
(14) Johnson, M. A., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
(15) Sen, K., Ed. *Molecular Similarity I and II, Topics in Current Chemistry*; Springer-Verlag: Berlin, 1995; Vol. 173, 174.
(16) Dean, P. M., Ed. *Molecular Similarity in Drug Design*; Chapman & Hall: New York, 1995.
(17) Carbo, R.; Leyda, L.; Arnau, M. An Electron Density Measure of the Similarity Between Two Compounds. *Int. J. Quant. Chem.* **1980**, *17*, 1185−1189.
(18) Hodgkin, E. E.; Richards, W. G. Molecular Similarity Based on Electrostatic Potential and Electric Field. *Int. J. Quant. Chem., Quant. Biol. Symp.* **1987**, *14*, 105−110.
(19) Good, A. C. The Calculation of Molecular Similarity: Alternative Formulas, Data Manipulation and Graphical Display. *J. Mol. Graph.* **1992**, *10*, 144−151.
(20) Kearsley, S. K.; Smith, G. M. An Alternative Method for the Alignment of Molecular Structures: Maximizing Electrostatic and Steric Overlap. *Tetrahedron Comput. Methodol.* **1990**, *3*, 615−633.
(21) Klebe, G.; Mietzner, T.; Weber, F. Different Approaches Toward an Automatic Alignment of Drug Molecules: Application to Sterol Mimics, Thrombin and Thermolysin Inhibitors. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 751−778.
(22) Klebe, G.; Mietzner, T.; Weber, F. Methodological Developments and Strategies for a Fast Flexible Superposition of Drug-Size Molecules. *J. Comput.-Aided Mol. Des.* **1998**, *12*, in press.
(23) (a) Cho, S. J.; Tropsha, A. Cross-Validated R^2 Guided Region Selection for Comparative Molecular Field Analysis (CoMFA): A Simple Method to Achieve Consistent Results. *J. Med. Chem.* **1995**, *38*, 1060−1066. (b) Tropsha, A.; Cho, S. J. Cross-Validated R^2 Region Selection for CoMFA Studies. In *3D QSAR in Drug Design. Volume 3. Recent Advances*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Dordrecht, 1998; pp 57−69.
(24) (a) Kroemer, R. T.; Hecht, P. A New Procedure for Improving the Predictiveness of CoMFA Models and Its Application to a Set of Dihydrofolate Reductase Inhibitors. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 396−406. (b) Kroemer, R. T.; Hecht, P.; Guessregen, S.; Liedl, K. R. In *3D QSAR in Drug Design. Volume 3. Recent Advances*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Dordrecht, 1998; pp 41−56.
(25) Norinder, U. Single and Domain Mode Variable Selection in 3D QSAR Applications. *J. Chemomet.* **1996**, *10*, 95−105.
(26) (a) Pastor, M.; Cruciani, G.; Clementi, C. Smart Region Definition: A New Way To Improve the Predictive Ability and Interpretability of Three-Dimensional Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1997**, *40*, 1455−1464. (b) Cruciani, G.; Pastor, M.; Clementi, C. Region Selection in 3D-QSAR. In *Computer-Assisted Lead Finding and Optimization;* Proceedings of the 11th European Symposium on Quantitative Structure−Activity Relationships, Lausanne, 1996; van de Waterbeemd, H., Testa, B., Folkers, G., Eds.; Verlag Helvetica Chimica Acta and VCH: Basel and Weinheim, 1997; pp 381−395. (c) Cruciani, G.; Clementi, S.; Pastor, M. GOLPE-Guided Region Selection. In *3D QSAR in Drug Design. Volume 3. Recent Advances*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Dordrecht, 1998; pp 71−86.
(27) 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.
(28) Klebe, G. Comparative Molecular Similarity Indices Analysis − CoMSIA. In *3D QSAR in Drug Design. Volume 3. Recent Advances*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Dordrecht, 1998; pp 87−104.
(29) Kubinyi, H. A General View on Similarity and QSAR Studies. In *Computer-Assisted Lead Finding and Optimization*; Proceedings of the 11th European Symposium on Quantitative Structure−Activity Relationships, Lausanne, 1996; van de Waterbeemd, H., Testa, B., Folkers, G., Eds.; Verlag Helvetica Chimica Acta and VCH: Basel, Weinheim, 1997; pp 9−28.
(30) Rum, G.; Herndon, W. C. Molecular Similarity Concepts, 5. Analysis of Steroid-Protein Binding Constants. *J. Am. Chem. Soc.* **1991**, *113*, 9055−9060.
(31) Good, A. C.; So, S.-S.; Richards, W. G. Structure−Activity Relationships from Molecular Similarity Matrices. *J. Med. Chem.* **1993**, *36*, 433−438.
(32) Good, A. C.; Peterson, S. J.; Richards, W. G. QSAR's from Similarity Matrices. Technique Validation and Application in the Comparison of Different Similarity Evaluation Methods. *J. Med. Chem.* **1993**, *36*, 2929−2937.
(33) Good, A. C. 3D Molecular Similarity Indices and Their Application in QSAR Studies. In *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Chapman & Hall: New York, 1995; pp 24−56.
(34) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9−20.
(35) Martin, Y. C.; Lin, C. T.; Hetti, C.; DeLazzer, J. PLS Analysis to Detect Nonlinear Relationships between Biological Potency and Molecular Properties. *J. Med. Chem.* **1995**, *38*, 3009−3015.
(36) Horwell, D. C.; Howson, W.; Higginbottom, M.; Naylor, D.; Ratcliffe, G. S.; Williams, S. Quantitative Structure−Activity Relationships (QSARs) of N-Terminus Fragments of NK1 Tachykinin Antagonists: A Comparison of Classical QSARs and Three-Dimensional QSARs from Similarity Matrices. *J. Med. Chem.* **1995**, *38*, 4454−4462.
(37) Coats, E. A. The CoMFA Steroids as a Benchmark Data Set for Development of 3D QSAR Methods. In *3D QSAR in Drug Design. Volume 3. Recent Advances*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Dordrecht, 1998; pp 199−213.
(38) 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.
(39) ISIS/Draw 2.1; MDL Information Systems, 14600 Catalina St, San Leandro, CA 94577.
(40) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.
(41) Stewart, J. P. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−103.
(42) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three-Dimensional Structure Directed Quantitative Structure−Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(43) SYBYL Molecular Modeling Software; Tripos Inc., 1699 S. Hanley Rd, St. Louis, MO 63944.
(44) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(45) Kubinyi, H. Evolutionary Variable Selection in Regression and PLS Analyses. *J. Chemomet.* **1996**, *10*, 119−133.
(46) Kubinyi, H. Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution. *Quant. Struct.-Act. Relat.* **1994**, *13*, 393−401.
(47) Hartung, J.; Elpelt, B.; Klösener, K.-H. *Statistik. Lehr- und Handbuch der angewandten Statistik*, 3rd ed.; R. Oldenbourg Verlag: München, 1985; p 565.
(48) Cramer, R. D., III; Bunce, J. D.; Patterson, D. E.; Frank, I. E. Crossvalidation, Bootstrapping and Partial Least Squares with Multiple Regression in Conventional QSAR Studies. *Quant. Struct.-Act. Relat.* **1988**, *7*, 18−25; erratum **1988**, *7*, 91.
(49) Wold, S. Validation of QSAR's. *Quant. Struct.-Act. Relat.* **1991**, *10*, 191−193.
(50) Clark, M.; Cramer, R. D., III. The Probability of Chance Correlation Using Partial Least Squares (PLS). *Quant. Struct.-Act. Relat.* **1993**, *12*, 137−145.
(51) Brzozowski, A. M.; Pike, A. C. W.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engström, O.; Öhman, L.; Greene, G. L.; Gustafsson, J.-Å.; Carlquist, M. Molecular Basis of Agonism and Antagonism in the Oestrogen Receptor. *Nature* **1997**, *389*, 753−758.
(52) Arevalo, J. H.; Hassig, C. A.; Stura, E. A.; Sims, M. J.; Taussig, M. J.; Wilson, I. A. Structural Analysis of Antibody Specificity. Detailed Comparison of Five Fab′-Steroid Complexes. *J. Mol. Biol.* **1994**, *241*, 663−690.
(53) Wallimann, P.; Marti, T.; Fürer, A.; Diederich, F. Steroids in Molecular Recognition. *Chem. Rev.* **1997**, *97*, 1567−1608.

(54) Novellino, E.; Fattorusso, C.; Greco, G. Use of Comparative Molecular Field Analysis and Cluster Analysis in Series Design. *Pharm. Acta Helv.* **1995**, *70*, 149−154.
(55) Kubinyi, H. Similarity and Dissimilarity. A Medicinal Chemists View. In *3D QSAR in Drug Design. Volume 2. Ligand-Protein Interactions and Molecular Similarity*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Dordrecht, 1998; pp 225−252.
(56) So, S.-S.; Karplus, M. Three-Dimensional Quantitative Structure−Activity Relationships from Molecular Similarity Matrices and Genetic Neural Networks. 1. Method and Validations. *J. Med. Chem.* **1997**, *40*, 4347−4359.

JM970732A