Quantitative Structure−Activity Relationship Studies of Progesterone Receptor Binding Steroids

Sung-Sau So,*[,†,‡] Steven P. van Helden,[§] Vincent J. van Geerestein,[§] and Martin Karplus*[,†,⊥]

Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, Department of Computational Medicinal Chemistry, N. V. Organon, P.O. Box 20, 5340 BH Oss, The Netherlands, and Institut le Bel, Université Louis Pasteur, 4 Rue Blaise Pascal, 67000 Strasbourg, France

Received October 18, 1999

The selection of appropriate descriptors is an important step in the successful formulation of quantitative structure−activity relationships (QSARs). This paper compares a number of feature selection routines and mapping methods that are in current use. They include forward stepping regression (FSR), genetic function approximation (GFA), generalized simulated annealing (GSA), and genetic neural network (GNN). On the basis of a data set of steroids of known in vitro binding affinity to the progsterone receptor, a number of QSAR models are constructed. A comparison of the predictive qualities for both training and test compounds demonstrates that the GNN protocol achieves the best results among the 2D QSAR that are considered. Analysis of the choice of descriptors by the GNN method shows that the results are consistent with established SARs on this series of compounds.

I. INTRODUCTION

The generation of a wide range of molecular and substituent descriptors based on molecular modeling techniques has become a routine part of quantitative structure−activity relationship (QSAR) studies. This has created a challenging problem in the selection of the best variables to use for the QSARs. In many situations typical of the drug design process, QSAR applications are trying to solve an underdetermined problem in which there are more variables (descriptors) than objects (compounds). Moreover, the underlying physical properties of the drugs that are correlated with their biological responses are often unknown so that a priori feature selection is not possible in most cases. This difficulty has been summarized by Kubinyi in suggesting that "selection of variables is time-consuming, difficult and, despite many different statistical criteria for the evaluation of the resulting models, a highly subjective and ambiguous procedure."[1] In recent years a number of automated objective feature selection algorithms have been proposed to remove the subjectivity and ambiguity from the selection procedure.[1−28] In this study we investigate several methods which represent some of the most successful combinations of feature selection and feature mapping tools. These include the forward stepping regression (FSR),[29] the genetic function approximation (GFA),[6,12,30] generalized simulated annealing (GSA)[5,10] and genetic neural network (GNN).[16,17,22] All of the methods are constructed to deal with high-dimensional data sets (i.e. a large number of descriptors and often a limited number of compounds).

To test the various methods, we use a range of progestagens, which were first introduced as oral contraceptives in the late 1950s. A large number of progestagenic steroids have been prepared. Compounds of this class all share the same molecular skeleton and display a wide variety of substituents with a generally well-defined orientation. For this reason, this class of compounds is particularly suitable for both 2D and 3D QSAR analyses. A number of studies using various molecular parameters and QSAR techniques have been published.[31−42] The most closely related work is that by van Helden et al.[42] in which an attempt was made to combine a genetic algorithm (GA) and a neural network to develop a QSAR for a set of 56 progestagens with known in vitro binding affinity to the progesterone receptor. They selected molecular descriptors using a combined genetic algorithm and multiple regression method as implemented in the GFA module of Cerius[2].[43] The selected descriptors were examined further by a neural network to derive a QSAR by using cross-validation to obtain optimal parameters. A major drawback of their approach is that the GFA descriptors, which give the best regression model, are not necessarily optimal for a nonlinear neural network. Use of an integrated system that directly couples a GA for feature selection and a neural network for correlation, such as GNN, is more appropriate. In this paper, we report a comparative study of some recently developed QSAR methods, which include GNN, on the same steroid data set.

II. EXPERIMENTAL SECTION

Data Set. The structural features of the compounds used in this study are shown in Table 1; see also Figure 1. The relative binding affinities (RBAs) of these compounds, also listed in Table 1, have been determined using standard procedures.[44] Progestagenic activity was measured relative to compound **7** which was set to 100% (log RBA = 2). These

[†] Harvard University.
[‡] Present address: Hoffmann-La Roche Inc., 340 Kingsland St., Nutley, NJ 07110.
[§] N. V. Organon.
[⊥] Université Louis Pasteur.

10.1021/ci990130v CCC: $19.00
Published on Web 05/02/2000

Table 1. Steroid Data Set

compd	log RBA	Δ	6α	7α	11β	13β	17β	17α	other
					(a) Training Set				
1	0.48	5(10)	H	CH_3	H	CH_3	OH	C≡CH	
2	1.51	4	H	H	H	CH_3	OH	CH_2CH_3	
3	0.78	4	CH_3	H	H	CH_3	OH	C≡CCH_3	10β-CH_3
4	1.74	4	CH_3	H	H	CH_3	C(O)CH_3	OC(O)CH_3	10β-CH_3
5	1.20	4	CH_3	H	H	CH_3	OH	C≡CH	
6	1.59	4 6	Cl	H	H	CH_3	C(O)CH_3	OC(O)CH_3	10β-CH_3
7	2.00	4	H	H	H	CH_3	C(O)CH_2OH	H	16α-CH_2CH_3
8	1.15	4	H	H	H	CH_3	C(O)CH_3	H	10β-CH_3
9	−0.30	5(10)	CH_3	H	H	CH_3	OH	C≡CH	
10	2.08	4	H	H	CH_3	CH_3	OH	C≡CH	
11	1.90	4	H	H	H	CH_2CH_3	OH	C≡CH	
12	1.23	4	H	H	OCH_3	CH_3	OH	C≡CH	
13	2.28	4	H	H	$=CH_2$	CH_2CH_3	OH	C≡CH	
14	1.34	4	H	H	CH_2Cl	CH_3	OH	C≡CH	
15	1.85	4	H	H	CH_2CH_3	CH_3	OH	C≡CH	
16	1.57	4	H	H	CH_2CH_2Cl	CH_3	OH	C≡CH	
17	1.60	4	H	H	CH=CH_2	CH_3	OH	C≡CH	
18	1.90	4	H	H	C≡CH	CH_3	OH	C≡CH	
19	0.00	4	H	H	CH_2OCH_3	CH_3	OH	C≡CH	
20	1.00	4 9 11	H	H	H	CH_2CH_3	OH	C≡CH	
21	2.11	4 9	H	H	H	CH_3	C(O)CH_2OH	H	16α-CH_2CH_3
22	1.48	4	H	CH_3	$=CH_2$	CH_3	OH	C≡CH	
23	1.20	4	H	CH_3	OCH_3	CH_3	OH	C≡CH	
24	1.15	4 9 11	H	H	H	CH_3	OH	C≡CH	
25	1.08	4	H	H	H	CH_3	OH	C-iC_3H_7	
26	2.26	4 15	H	H	H	CH_2CH_3	OH	C≡CH	
27	2.15	4 15	H	H	$=CH_2$	CH_3	OH	C≡CH	
28	1.69	4 15	H	H	H	CH_3	OH	C≡CH	
29	1.51	4	H	H	H	CH_3	OH	C≡CH	15α-CH_3
30	0.30	4	H	H	H	CH_2CH_3	OH	C≡CH	12 $=CH_2$
31	0.48	4	H	H	H	CH_2CH_3	OH	C≡CH	12β-CH_3
32	2.26	4	H	H	CH_3	CH_3	C(O)CH_3	OC(O)CH_3	H
33	2.05	4	H	H	OH	CH_3	C(O)CH_3	OC(O)CH_3	10β-CH_3
34	1.32	4	H	H	C≡N	CH_3	OH	C≡CH	
35	0.90	4	H	CH_3	CH=CH_2	CH_3	OH	C≡CH	
36	−0.52	4	H	CH_3	CH_2OCH_3	CH_3	OH	C≡CH	
37	0.70	4	H	CH_3	C≡N	CH_3	OH	C≡CH	
38	1.20	4	H	H	H	CH_3	OH	C=CH	
39	1.11	4	H	H	CHF_2	CH_3	OH	C≡CH	
40	1.61	4	H	CH_3	CH_3	CH_3	OH	C≡CH	
41	0.79	4	H	CH_3	=CH−C_4H_9 (*E*)	CH_3	OH	C≡CH	
42	1.40	4	H	CH_3	C≡CH	CH_3	OH	C≡CH	
43	1.99	4	H	H	=$CHCH_3$ (*E*)	CH_3	OH	C≡CH	
					(b) Test Set				
44	1.18	4	H	CH_3	H	CH_3	OH	C≡CH	
45	1.40	4	H	H	H	CH_3	OH	C≡CH	
46[a]	0.65	4 6	H	H	H	CH_3	C(O)CH_3	H	10α-CH_3
47	0.49	5(10)	H	H	H	CH_3	OH	C≡CH	
48	2.03	4	H	H	$=CH_2$	CH_3	OH	C≡CH	
49	1.56	4 9	H	H	H	CH_3	C(O)CH_3	CH_3	
50	1.43	4	H	CH_3	H	CH_3	OH	$CH_2CH=CH_2$	
51	2.30	4 15	H	H	$=CH_2$	CH_2CH_3	OH	C≡CH	
52[a]	1.61	4	H	H	H	CH_3	C(O)CH_3	bridge	17α-OCH(C_2H_5)O-
53	1.53	4	H	H	=CHF (*E*)	CH_3	OH	C≡CH	14α,10ß-CH_3
54	2.28	4	H	H	=CHF (*Z*)	CH_3	OH	C≡CH	
55	1.29	4	H	CH_3	CH_2CH_3	CH_3	OH	C≡CH	
56[a]	1.70	4	H	H	bridge	bridge	OH	C≡CH	11β-$(CH_2)_3$-13β

[a] Structural outlier.

compounds were selected from literature[42] and the Organon database of steroids on the basis of two important criteria. First, the RBAs of these compounds with the progesterone receptor should cover a wide range, preferably several orders of magnitude. In the present data set, the extremes of affinity differ by a factor of 440. Second, the structures should offer a wide variation in substituents and positions of unsaturation in the steroid skeleton. Particular care was taken to include pairs of compounds that differ only in a single feature. For example, **1**/**44, 45**/**47**, and **5**/**9** differ solely in that they possess a Δ^4 or a $\Delta^{5(10)}$ double bond, respectively; **53**/**54** have *E*- and *Z*-fluoromethylene substituents; **17**/**35** and **15**/**55** have a hydrogen or a methyl group in position 7α; **37**/**42** and **18**/**34** are isoelectronic, an ethynyl group having been replaced by a nitrile; **13**/**51** and **27**/**48** differ by a Δ^{15} double bond, whereas **26**/**51** and **45**/**48** differ by the presence of a 11-methylene substituent. Finally, **2**, **38**, and **45** have a single, double, and triple $C_{20}-C_{21}$ bond, respectively.

The distribution between training and test sets was made somewhat arbitrarily in such a way that each of the sets

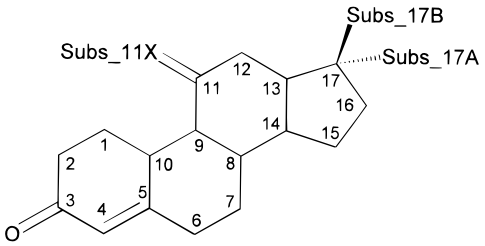

Figure 1. Steroid structure and substituent naming scheme (see text).

would contain both high- and low-RBA compounds, and at least one example of each basic molecular skeleton (Δ^4, $\Delta^{5(10)}$, $\Delta^{4,9}$, Δ^{15}, 17β-pregnan-20-one, etc.). This was achieved by sorting all compounds by RBA and then placing every fourth compound in the test set. A few compounds with similar RBA were switched between test set and training set so as to include the desired structural variations in the test set. Compounds **53** and **54** represent a special case. They differ only in the configuration of the fluoromethylene group but have a markedly different RBA. Both compounds were placed in the test set to see whether the QSAR models would be capable of distinguishing them, without any similar compounds in the training set. Also three compounds were included to assess the performance of the various methods on extrapolation beyond the scope of the training set. They have molecular skeletons that are unique and are regarded as structural outliers in the data set. They are dydrogesterone (**46**), which has the unusual retro (9β,10α) configuration; proligestone (**52**), with a unique 14α,17α bridge; and **56**. In **56** the 11β,13β bridge would seem at first sight to emulate a combination of 13β- and 11β-ethyl groups (such as in **15** and **55**). However, in the latter compounds, the ethyl groups are known to occupy an anti conformation that is not possible in **56**. Moreover, steric repulsion causes the steroid to adopt a convex shape, as demonstrated by X-ray structures.[45,46]

Three-dimensional starting structures were obtained by converting 2D diagrams into 3D coordinates using CORINA.[47] The 3D structures were loaded into the Chem-X program to perform all modeling calculations.[48] Atomic charges were calculated using MOPAC[49] AM1 version 6.0, and all structures were minimized using the Chem-X MME force field.[48] Unless stated otherwise, Chem-X default values were used for all modeling tasks.

Chem-X provides tools to calculate a wide range of properties of (or parts of) a molecule. To automate these calculations, the atoms of all structures were numbered in a consistent way. The carbon atoms in the steroid ring system were numbered C_1 through C_{17} by convention,[50] and properties of substituents connected to each of these carbon atoms were given a corresponding number in their name (see Figure 1). Stereochemistry in a steroid is normally indicated by α (a substituent below the plane of the steroid) or β (a substituent above the plane of the steroid). In the calculated properties these substituents were distinguished by adding a residue name ("A" or "B") to the names of the properties; e.g. VOL_11B encodes the volume of the substituent attached at the 11β position of the steroid. Substituents connected to a sp^2 carbon atom in the ring were treated separately since they are approximately parallel to the plane of the steroid; such substituents were labeled with residue name "X", as shown in Figure 1. Finally, for compounds that have a bridge between two substituent positions (e.g. **46** and **52**), the bridging atoms were evenly divided to create a pair of pseudosubstituents at each of the attachment positions.

Table 2. Properties Calculated with Chem-X

(a) Whole-Molecule Properties

MME	molecular mechanics energy (kcal)
MOLWGT	molecular weight (atomic units)
DIPOLE	dipole moment (De)
VOLTOT	total molecular volume (Å^3)
SURTOT	total surface area of molecule) (Å^2)
HOMO	energy of HOMO (AM1) (eV)
LUMO	energy of LUMO (AM1)) (eV)
ELEC	total electronic energy (AM1) (eV)
HEAT	heat of formation (AM1) (kcal)

(b) Atom and Substituent Properties

CHC	atomic charge of steroid ring atoms C1 to C17 (au)
CH_SKEL	sum of atomic charges of all steroid ring atoms (au)
CH_	sum of atomic charges of atoms in a given substituent (au)
VOL_	total volume of atoms in a given substituent (Å^3)
SUR_	total surface area of atoms in a given substituent (Å^2)
DIP	dipole moment of atoms in a given substituent (De)

The QSAR module of Chem-X was used to calculate a total of 52 properties (Table 2) for each of the compounds. These properties, which were chosen to reflect the structural diversity in the data set, included nine whole-molecule properties and 43 properties for individual atoms or substituents. The resulting data set[51] with 52 properties and the log RBA value for each molecule were used for analysis.

METHOD

The various methods for obtaining a QSAR and their extensions that are used in this paper have been described thoroughly in previous publications. Consequently, their full implementation details are not discussed here. We list them and describe them briefly below. The methods all attempt to correlate biological activity with a selected set of descriptors rather than making use of the entire descriptor set as in methods such as principal component regression or partial least squares. All QSAR models contain eight descriptors based on the empirical rule of Topliss that there should be at least five or six data points in the training set per descriptor used.[52] Either regressions or 8−2−1 scaled conjugate gradient (SCG) neural networks[53] are utilized for mapping.

Different Approaches Used for Comparison. (a) Subjective Selection. Since the present study of progestagenic steroids represents a case where some SARs are already known, a subjective selection of descriptors has been made using such knowledge. Previous SARs suggested that the key substituents that influence progesterone receptor binding are at positions 11 and 17. Thus, a major effort has been made to synthesize compounds that provide a diverse set of substituents at these positions. Such experimental bias is evident in this steroid data set, since most structural variations occur at these two positions. For position 11 (14 different non-hydrogen substituents in 23 compounds), two considerations were used in the development of the subjective QSAR model. First, both steric and electrostatic effects should be characterized, and therefore VOL_11B and DIP11 were selected. Second, CH_11X was selected to distinguish between a sp^2 and sp^3 carbon atom. For position 17 the two descriptors VOL_17A and CH_17B seemed appropriate as judged from the nature of the substituents at the α and β

positions; e.g. the 9 different α substituents vary significantly in bulk, and the 3 different β substituents contain polar hydroxyl and/or carbonyl functional groups. Two other structural features are known to play important roles. They are the substituents at position 13 and the double bond between C_9 and C_{10}. The descriptors VOL_13B and CHC10 were introduced to describe these effects. To complete the set of eight descriptors, inclusion of a molecular property was regarded as appropriate. The descriptor LUMO was chosen because it was the most linearly correlated property ($r = -0.29$ for the 43 compounds in the training set) with log RBA among the 9 whole-molecule descriptors.

Since some of the selected descriptors may be of nonlinear nature, a neural network was utilized to perform the model-free mapping of these descriptors with respect to binding affinities.[54]

(b) Forward Stepping Regression. One of the simplest, chemometric methods considered here is the combination of a forward stepping selection procedure and a multiple linear regression. In this method descriptors are added until no more significant variables ($F > 2.5$, where F is the result from an F-test) can be found. The QSAR+ module in the commercial Cerius^2 package was used for this calculation.[29]

(c) Genetic Function Approximation. Genetic function approximation is a hybrid method formulated by Rogers and Hopfinger that makes use of a genetic algorithm and regression to construct QSAR equations.[6] A suitable set of descriptors is chosen by a genetic algorithm, and the selected descriptors are utilized to build a nonlinear QSAR regression equation. Nonlinear correlations in the data are explicitly dealt with by use of the descriptors in spline, quadratic, offset quadratic, and quadratic spline functions.[6] A spline function is denoted $\langle x - a \rangle$, where a is the knot parameter. This function returns the value of the argument, if it is positive, and zero otherwise. The method has been implemented in the Cerius^2 package, and it was used here without modification.[29] The smoothness parameter was kept at the default value of 1.0, and the length of an individual was fixed at 8 descriptors. A total of 200 individuals were evolved over 5000 new generations.

(d) Generalized Simulated Annealing. Simulated annealing (SA) is an optimization method based on a physical annealing process.[55] The method is similar to a Monte Carlo simulation[56] but contains a slowly decreasing temperature factor. At the beginning of the optimization, solutions of lesser quality are accepted more readily according to a Boltzmann-type probability. The slow cooling process makes possible an exploration of the configurational space and facilitates escapes from local minima.[57] As the temperature drops, the system theoretically converges to the optimal solution. However, this is guaranteed only over an infinitely long annealing period. The earlier applications of SA in QSAR, termed generalized SA (GSA),[5] had been used in conjugation with multiple linear regression, but the most recent report indicated that an improvement resulted from its use with a neural network.[10] We have implemented the GSA algorithm based on the literature in combination with a neural network that used the scaled conjugated gradient (SCG) training algorithm.[53]

(e) Genetic Neural Networks. The genetic neural network is a hybrid method that uses a genetic algorithm for the variable selection and a neural network for the mapping.[16,17,22] An evolutionary programming technique and a SCG-type neural network were used here. The simulation methodology has been described by other reports[16,17] and is not repeated here.

Statistical Validation. In QSAR studies, especially in the field of CoMFA,[58] cross-validation is used extensively to assess the predictivity of a model. Although it has been suggested that leave-one-out (LOO) cross-validation tends to overestimate the predictivity,[59] the method is still widely used because it is easy to apply and gives a reproducible indication of the predictive power. If a sufficient number of compounds is available, an alternative (and arguably better) way of QSAR validation tests the predictivity for a set of compounds that have not been used in the derivation of the model. In this study, we used 43 compounds for training and cross-validation and 13 additional compounds as a test set. This protocol simulates the situation where a set of known compounds (the training set) is available for deriving and validating a model and where the activities of the new compounds (the test set) are predicted at a later stage.

Two statistical quantities are used to assess the performance of the different QSAR models. The definitions of r^2 and q^2 are given in eqs 1 and 2. The Pearson correlation coefficient r_{trn}^2 gives the quality of fit obtained for the compounds in the training set. The cross-validated correlation coefficient q^2 estimates the predictivity of the QSAR model using a leave-one-out (LOO) procedure. The overall accuracy of the test set predictions is indicated by r_{tst}^2.

$$r^2 = \Big(\sum_{i=1}^{N} (\text{activity}_{\text{calc},i} - \text{activity}_{\text{calc,mean}})(\text{activity}_{\text{obs},i} - \text{activity}_{\text{obs,mean}})\Big)^2 / \Big(\sum_{i=1}^{N} \text{activity}_{\text{calc},i}^{\,2} - N\text{activity}_{\text{calc,mean}}^{\,2}\Big) \Big(\sum_{i=1}^{N} \text{activity}_{\text{obs,i}}^{\,2} - N\text{activity}_{\text{obs,mean}}^{\,2}\Big) \quad (1)$$

$$q^2 = 1 - \frac{\sum_{i=1}^{N} (\text{activity}_{\text{calc},i} - \text{activity}_{\text{obs},i})^2}{\sum_{i=1}^{N} (\text{activity}_{\text{obs},i} - \text{activity}_{\text{obs,mean}})^2} \quad (2)$$

III. RESULTS

Table 3 presents a summary of the results obtained with the different methods. The selection of properties is displayed graphically in Figure 2 to give a better insight into the distribution of the properties over the steroid skeleton.[13] Predictions of individual compounds in both the training (i.e. from the cross-validation results) and test sets are shown in the scatter plots (Figure 3).

Subjective Selection. The eight-descriptors neural network QSAR model is not very good in fitting the training data ($r_{trn}^2 = 0.590$). The weakness of the subjective selection becomes even clearer when the predictive performance of the model is assessed (Figure 3a). The model yields a q^2 value of -0.188, which means that the predictions are no better than chance. This result is in line with the prediction accuracy for the test compounds, also shown in Figure 3a. The value of r_{tst}^2, including the predictions for the three

Table 3. Results of the Five Eight-Descriptor QSAR Models (SS = Subjective Selection; FSR = Forward Stepping Regression; GFA = Genetic Function Approximation; GSA = Generalized Simulated Annealing; GNN = Genetic Neural Network)[a,b]

	SS	FSR	GFA	GSA	GNN
r_{trn}^2	0.590	0.721	0.772	0.860	0.880
q^2	−0.188	0.535	0.626	0.635	0.717
r_{tst}^2 with outliers	0.108	0.089	0.145	0.144	0.526
r_{tst}^2 without outliers	0.127	0.408	0.506	0.488	0.610
MME		●		●	
CH_17B	●	●			●
CHC13		●	●	●	
VOL_11TOT		●			
VOL_11B	●	●	$\langle$●−$a\rangle^2$	●	
SUR_11B		●			
SUR_11X		●			●
ELEC		●	●	●	●
SUR_17B			●	●	
LUMO	●				
DIP11	●				
CH_11X	●				
VOL_17A	●				
CHC17			●		
CHC_11TOT			●		
CHC11B			●^{2}		
CHC14			●^{2}	●	
CH_SKEL				●	
VOL_13B	●			●	●
DIP17					●
MOLWGT					●
VOL_6A					●
CHC10	●				●

[a] A ● mark indicates that the descriptor on the left is used by the corresponding model. [b] The numbers refer to the position of the substituents, and stereochemistry is indicated by A and B for α and β, respectively. Substituents connected to a sp^2 carbon in the steroid skeleton are indicated by X. "TOT" refers to all atoms in all substituents connected to the corresponding carbon atom in the steroid skeleton.

structural outliers (**46**, **52**, and **56**), is 0.108, and it increases to only 0.127 when they are removed from the calculation. The model is seen to perform very poorly on the most potent compounds in the test set; they are all predicted to have a low binding affinity. This result underlines the difficulty in selecting appropriate descriptors based simply on existing SAR information. Even though the descriptors refer to positions that are known to be important for binding, the selection of descriptors at these positions is inadequate for describing potent compounds. The descriptors chosen for position 11, which is known to be very important, are quite different from those used in all the other methods. The above findings suggest that inclusion of descriptors from other positions and refinement in variables selection in the key positions are necessary for improving the predictivity.

Forward Stepping Regression. Forward stepping regression selected eight significant descriptors based on the criterion that each of the chosen descriptors have a partial F value greater than 2.5. The resulting QSAR equation is given as follows:

$$\log \mathrm{RBA} = 1.98 + 9.83 \times 10^{-3}\mathrm{MME} + 1.93 \times 10\mathrm{CH\_17B} - 2.33 \times 10\mathrm{CHC13} - 5.66 \times 10^{-2}\mathrm{VOL\_11TOT} - 8.38 \times 10^{-2}\mathrm{VOL\_11B} + 1.52 \times 10^{-1}\mathrm{SUR\_11B} + 7.23 \times 10^{-2}\mathrm{SUR\_11X} + 5.46 \times 10^{-5}\mathrm{ELEC}$$

The most striking feature of this QSAR model is that half

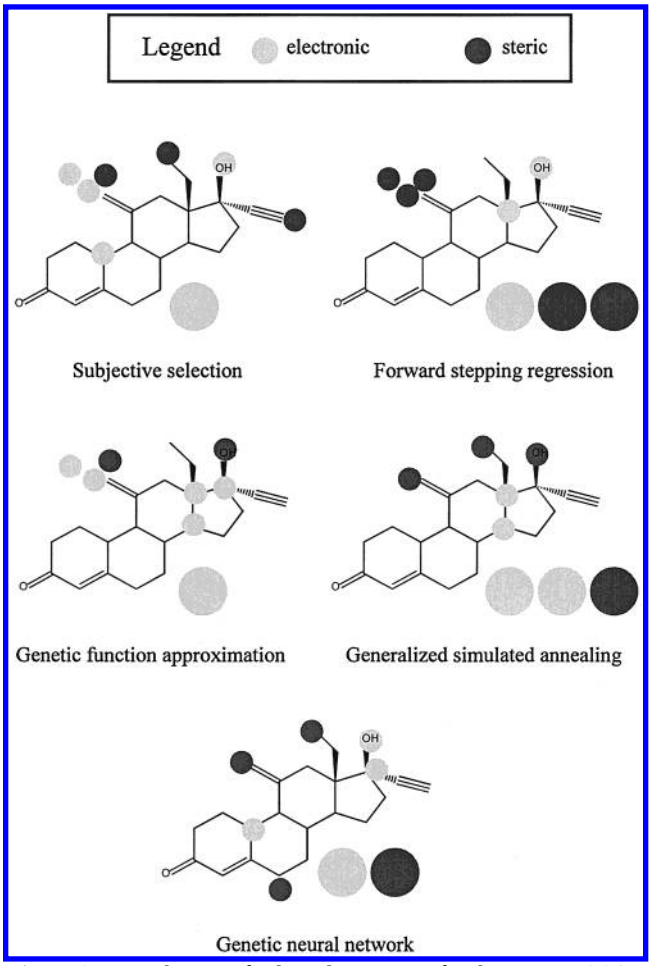

Figure 2. Distribution of selected properties for the various QSAR models: Subjective selection; forward stepping regression; genetic function approximation; generalized simulated annealing; genetic neural network. For each model the small circles on the steroid indicate what type of properties are selected for that particular position. A light gray circle indicates an electrostatic property, and a dark circle indicates a steric property. Whole molecule properties are presented as large circles near the molecules with the same color coding. More than one light or dark circle indicate the number of selected properties.

of the selected descriptors encode steric properties at position 11. This is in good agreement with the established SAR which suggests that the C_{11} substituents play the most important role in the activity. However, from a statistical standpoint the repeated use of this descriptor class is a source of concern. In particular the combination of VOL_11B and SUR_11B, which are highly correlated, results in unreliable estimates for the corresponding coefficients in the QSAR equation. The selection of CH_17B and CHC13 makes chemical sense because it is known that these two positions are functionally important. However, there is no explicit variable that characterizes the Δ^9 pattern. This model uses two whole molecule descriptors (MME and ELEC) to complement the above positional descriptors.

The ability to fit the training data is fairly good (r_{trn}^2 = 0.721), and the cross-validated q^2 has a reasonable value of 0.535, suggesting that the predictivity of the QSAR is unlikely to be due to chance. The results are shown in Figure 3b. The test set predictions (r_{tst}^2 = 0.089) are also shown in Figure 3b. The prediction of **54** is much too low: it is predicted to have a binding almost equal to **53** because SUR_11X does not distinguish between these two similar

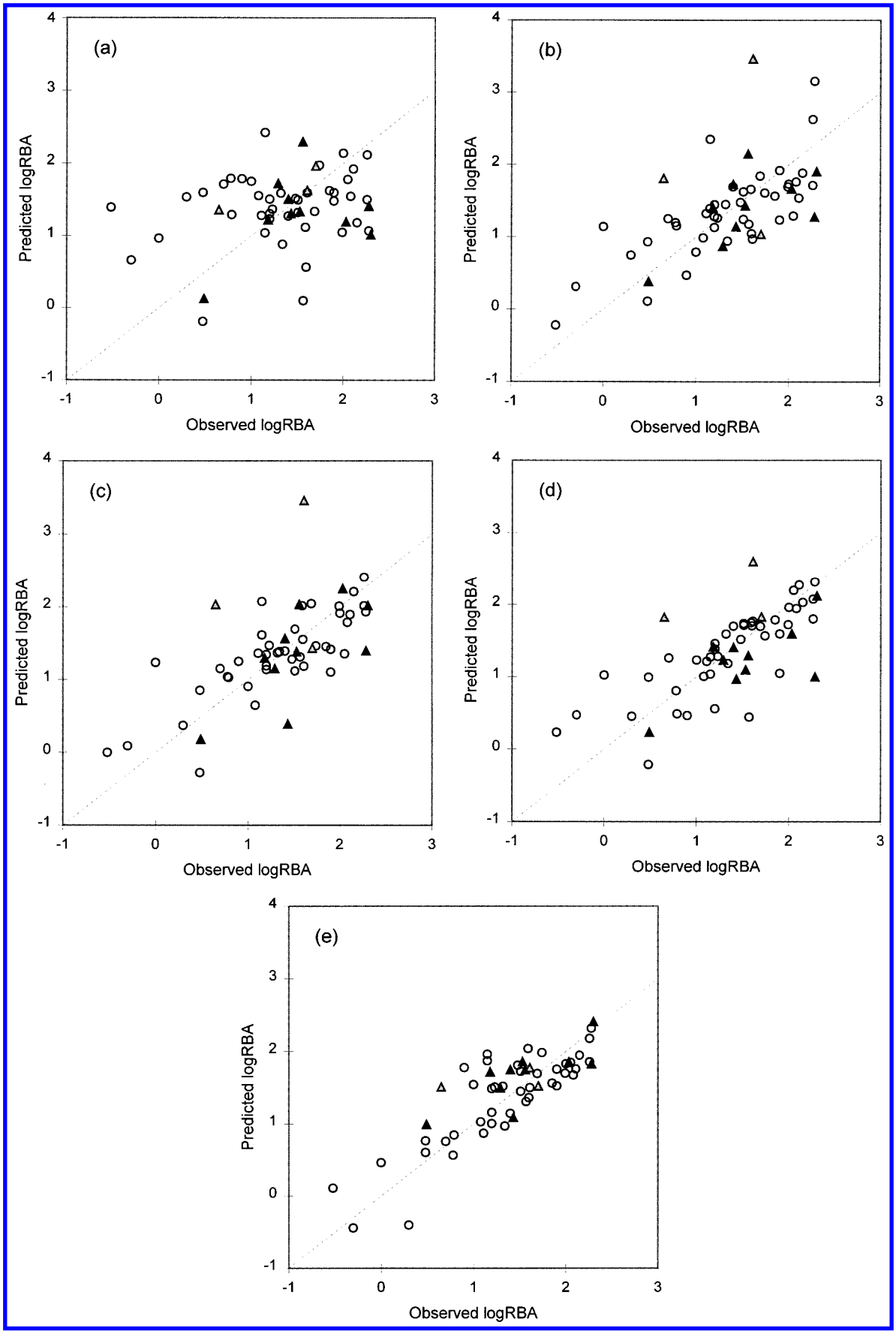

Figure 3. Plots of predicted log RBA versus observed log RBA for the various QSAR models: (a) subjective selection; (b) forward stepping regression; (c) genetic function approximation; (d) generalized simulated annealing; (e) genetic neural network. The open circles in the plot denote the predictions of training compounds using LOO cross-validations; the solid triangles denote the predictions of the test compounds; and the gray triangles are the structural outliers. The dotted line represents a perfect correlation between observed and predicted binding affinity.

compounds. The rest of the predictions seem reasonable (r_{tst}^2 without outliers = 0.408); for the three structural outliers, the predictions are poor.

One known problem with this method is its sensitivity to multiple local minima so that the final solution is often not optimal. Despite this limitation, the current regression QSAR is satisfactory given its simplicity and the ease of construction. It is definitely superior to the model based upon a subjective selection of descriptors. The calculation is faster than the subsequent methods; it requires approximately a minute on an 175 MHz R4400 Silicon Graphics workstation. Thus, the procedure seems suitable as a quick preliminary check on the quality of the data set, e.g., to determine whether it is necessary to generate more descriptors (or include a larger set of compounds) or nonlinear terms before a reliable QSAR can be built.

Genetic Function Approximation. The GFA method implemented in Cerius[2] was used to select the best nonlinear combination of properties for an eight-descriptor model. The following regression equation was found after 5000 generations for 200 individuals:

$$\log \mathrm{RBA} = 6.68 \times 10^{-1} + 5.06\mathrm{CHC17} + 9.28 \times 10^{-5}\mathrm{ELEC} - 3.35\mathrm{CH\_11TOT} - 2.35 \times 10(\mathrm{CH\_11B})^2 + 1.48 \times 10^{-1}\mathrm{SUR\_17B} - 2.10 \times 10\mathrm{CHC13} - 4.41 \times 10^{-1}\langle \mathrm{VOL\_11B} - 5.16 \times 10\rangle^2 - 1.36 \times 10^2(\mathrm{CHC14})^2$$

As in the previous case the whole molecule property ELEC was selected. Consistent with the SAR analysis, this selection contained both steric and electrostatic aspects of positions 11 and 17 and also included the variation of position 13. The role of CHC14 is less obvious. It may be that the charge of atom C_{14} reflects the presence of a double bond between the neighboring atoms C_{15} and C_{16}, which has been shown to have a positive effect on binding. The predictions shown in Figure 3c are comparable to the results of the FSR calculation. Two of the structurally different compounds (**46** and **52**) are again predicted to have too high a binding affinity. The predicted binding affinities of compounds **50** and **54** are underestimated by the GFA equation. For **50**, this is related to a unusually low partial charge on atom C_{17} (CHC17) due to a unique 17α substituent that is present only in this test compound.

The performance of this QSAR model, as determined by the statistical measures r_{trn}^2 (0.772), q^2 (0.626), and r_{tst}^2 (0.145 and 0.506 without outliers), shows an improvement over the previous models (Table 3). The improvement probably results from a more thorough evaluation of different descriptor combinations through the use of the genetic algorithm. The nonlinear terms, though limited in functional forms, appear to be important contributions for obtaining correlation and prediction; i.e., CH_11B and CHC14, which were not previously chosen, are present as squared functions. The descriptor VOL_11B is assigned to a quadratic spline function by GFA. The use of this relatively complex function has prompted us to reexamine the correlation between VOL_11B and log RBA, which is shown in the scatter plot (Figure 4). The plot suggests that a spline function is appropriate for modeling such a distribution of data. The knot (a) is expected to be located somewhere between 5 and 12 (dotted vertical line) because of an apparent linearity between log RBA and VOL_11B, once the latter reaches the threshold value defined by the knot. However, there is a problem with the position of the knot, which is at $a = 51.59$ (the solid vertical line on the plot). The fact that only one (the least active) compound has a nonzero $\langle \mathrm{VOL\_11B} - 51.59\rangle^2$ value makes the exponent following the spline function redundant. In essence, this term appears to have been introduced only to obtain a better fit to the experimental data for a single compound, which is a questionable choice in terms of a meaningful and predictive model.

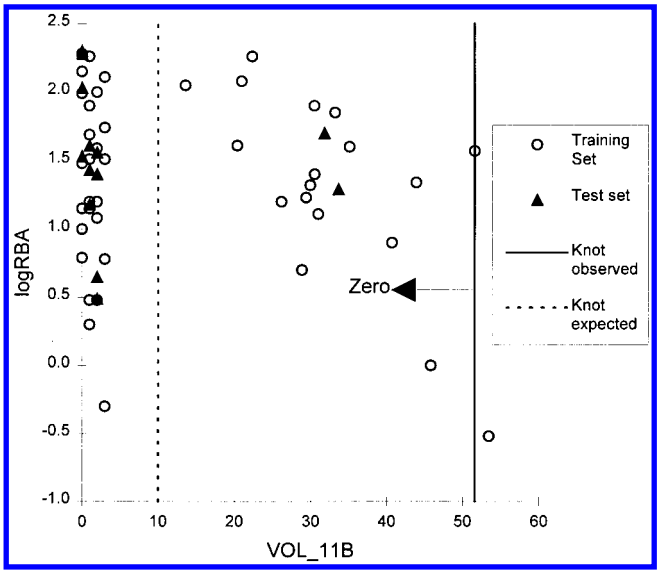

Figure 4. Plot of observed log RBA versus the VOL_11B descriptor values for the compounds in the data set. The solid line shows the actual position of the knot by GFA; the spline term returns zero for all of the compounds on the left side of the line. The dotted line seems a more logical placement of the knot of the spline function because it would clearly separate the different trends.

The result described here demonstrates the limitations in the generation of nonlinearity by QSAR regression equations. Automatic inclusion of nonlinear terms still appears to be problematic and limited in scope. It is conceivable that cross-product terms can be important, though an exhaustive evaluation of all possibilities is not feasible with the current implementation. For this reason, it is advantageous to make use of intrinsically nonlinear techniques, such as a neural network, to introduce nonlinear QSARs.

Generalized Simulated Annealing. Ten independent GSA runs were made using different initialization seeds. The model with the highest predictivity in terms of cross-validation was considered as the optimal GSA model. The eight descriptors in this QSAR are as follows: MME, CH_SKEL, VOL_13B, CHC13, CHC14, VOL_11B, SUR_17B, and ELEC. This set shares five descriptors with the GFA model. There is also a correspondence with the SAR models, as half of the descriptors come from the key positions (positions 11, 13, and 17). Three of the descriptors involve whole molecule properties.

The use of a neural network probably plays a role in the improvement in r_{trn}^2 (0.860). Nevertheless, the increase in predictivity ($q^2 = 0.635$) estimated by cross-validation is not significant relative to the GFA method (Table 3 and Figure 3d). The accuracy of the test set prediction ($r_{tst}^2 = 0.144$ and 0.488 without outliers) is also found to be similar to GFA.

Although simulated annealing can deal with the problem of local minima, in our experience this algorithm appeared very sensitive to the temperature control: i.e., its success depends on the choice of annealing schedule.[57] This dependency can be circumvented in part by performing multiple runs with different initial conditions (as in this case). However, if multiple models are involved, it is more appropriate to make use of a genetic algorithm. Because in a GA there are interactions among the various evolving QSAR models, a better solution is likely to be found more rapidly than by the use of many independent SA trials.[60]

Genetic Neural Networks. The eight descriptors chosen in the top GNN model are as follows: DIP17, MOLWGT, VOL_13B, VOL_6A, CHC_10, CHC_17B, SUR_11X, and ELEC. Interestingly, the set has relatively little overlap with other models. The most notable omissions are the two descriptors CHC13 and VOL_11B that are present in the other objective QSAR models. ELEC is the only descriptor present in all the models, although its high occurrence frequency is difficult to explain. Possibly it is due to the fact that this global electronic measure can provide additional information on the molecular size (the correlation coefficient between ELEC and VOL_TOT is −0.57). As expected, the descriptors include positions 11, 13, and 17. Interestingly, a descriptor from the 6-position is selected for the first time. Also, it is the only case where position 10 is represented, other than in the subjective selection. Overall, the selection by GNN appears to cover a wider range of structural features than the other approaches. Thus, the predictions provided by this QSAR should be more reliable for novel analogs, particularly when the structural modifications occur at the positions where a substituent descriptor has been represented in the model.

The calculated (r_{trn}^2) and cross-validated (q^2) correlation coefficients for the training set are 0.880 and 0.717, respectively. This improvement over the nonlinear GFA and the GSA models indicates that a neural network combined with a genetic algorithm is more appropriate than a regression and simulated annealing scheme for this problem. The test set predictions are good; $r_{tst}^2 = 0.526$ with and 0.610 without the three structural outliers (Table 3 and Figure 3e).

A standard randomization test is employed to ensure that the good predictivity of the GNN model is not a result of chance correlation.[52,61−63] In this method, the output values (the biological activities) of the compounds are shuffled randomly. They are trained against real input descriptors values, and the correlation and predictivity of the QSAR model is determined. The whole procedure is repeated on many different scrambled data sets. The rationale behind this test is that the significance of the real QSAR model would be suspect if there is a strong correlation between selected descriptors with these randomized response variables. Figure 5 shows a plot of q^2 against r_{trn}^2 for 50 such runs (crosses), together with the point (solid circle) corresponding to the real QSAR. The separation of the true QSAR point from those corresponding to random response variables provides compelling evidence that the QSAR is the result of genuine correlation between the chosen descriptors and activity.

Functional Analysis of the GNN Descriptors. Since the GNN QSAR has the best predictivity, the choice of descriptors has been evaluated with an established neural network monitoring scheme.[16,64] First the GNN model was fully trained in the normal way. After training, the variation of predicted binding affinity was monitored by changing the descriptor value of one (or more) specified network inputs, while the remaining inputs were kept at the values corresponding to those of a chosen template compound. This procedure was repeated for all network inputs. In this way the functional dependence of the various descriptors can be determined. We thank a reviewer for bringing up the point that the shapes of the curves obtained can be somewhat dependent on the exact values chosen for the constant inputs. To keep the extent of extrapolation to a minimum when drawing the functional dependence maps, it is appropriate to use the descriptor values of an existing analog as the template. Furthermore, it is interesting to predict the types of structural modifications of an existing compound that should lead to higher predicted potencies. For this reason, compound **45**, a precursor of many more potent progestagens, was used as the template molecule. Figure 6a shows the functional dependence plot of four relatively uncorrelated GNN descriptors (CHC10, VOL_6A, SUR_11X, and VOL_13B), and parts b and c of Figure 6 show the variation of binding affinity as functions of the CH_17B/DIP17 and MOLWGT/ELEC pairs in the form of a contour plot.

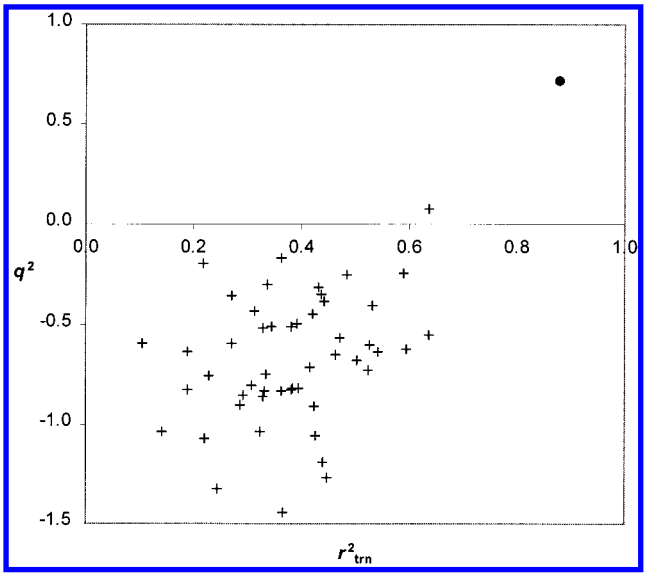

Figure 5. Scatter plot for q^2 against r_{trn}^2 for the real QSAR (●) and those with randomized activity values (+).

CHC10 is chosen by the GA probably because of its ability to encode the three types of structural variations at position 10 that are observed in the data set. The majority of the compounds have a hydrogen substituted C_{10} atom which carries a partial charge of −0.07. A few compounds have 10β-methyl substituents that increase the CHC10 value to approximately −0.03. The remaining compounds contain either Δ^9 or $\Delta^{5(10)}$ double bond; the unsaturation increases the electron density at C_{10}, which lowers the CHC10 value to −0.10. Figure 6a shows that the binding affinity is at a maximum when C_{10} is unsubstituted and saturated. Oxidation to unsaturated analogs triggers a rapid decrease in binding affinity, a manifestation of the fact that all but one of the compounds with unsaturated C_{10} have low log RBA (the average log RBA of the seven compounds with unsaturated C_{10} is 0.92 as compared to 1.37 for the data set). A methyl substitution also weakens binding, though to a lesser extent, as shown in the plot. For example, the presence of the

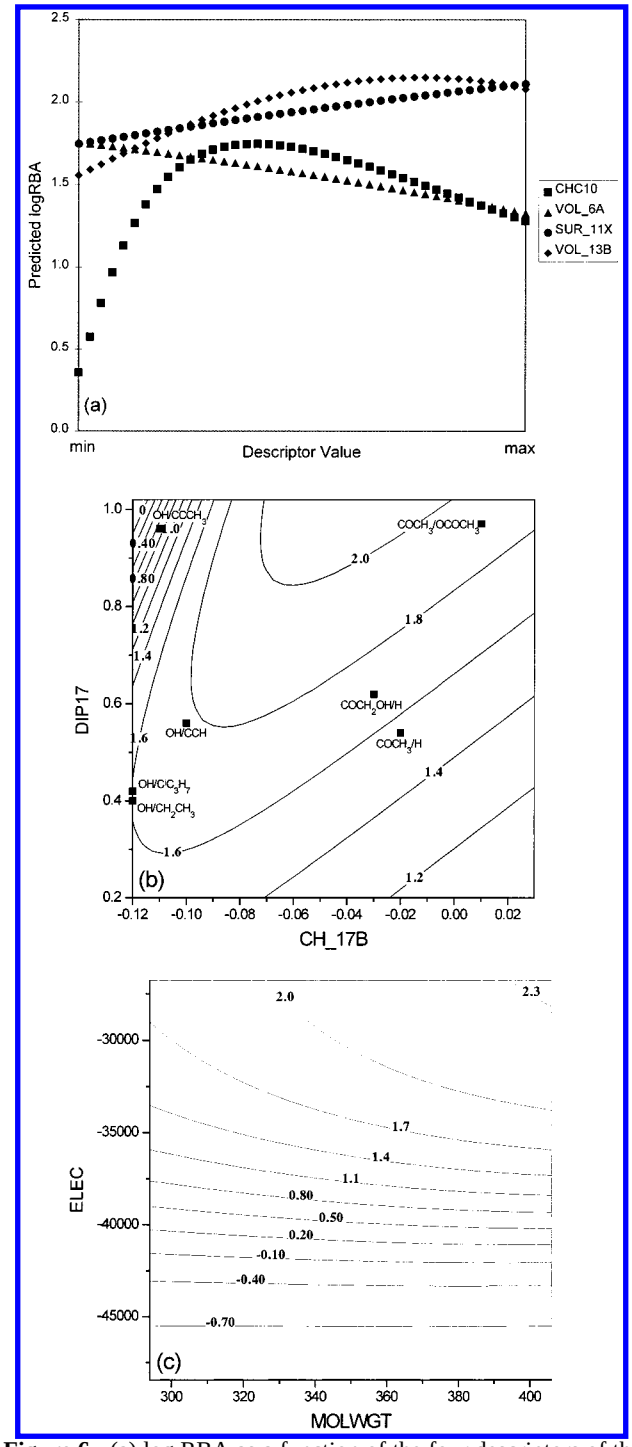

Figure 6. (a) log RBA as a function of the four descriptors of the GNN model, using a common precursor (**45**) as a template (see text). The minimum and maximum values for the four descriptors are as followed: CHC10, from −0.11 to −0.01; VOL_6A, from 0.0 to 29.4; SUR_11X, from 0.0 to 68.9; VOL_13B, from 18.8 to 43.2. (b) Contour map showing the variation of logRBA as functions of both CH_17B and DIP17. The different 17β/-α substituent pairs in the training set are also shown. (c) A contour map showing the variation of log RBA as functions of both MOLWGT and ELEC. The dotted line corresponds to a perfect anticorrelation between the two descriptors. The 43 training compounds are also displayed in the plot and are labeled by their log RBA values.

10β-methyl group in compounds **3** and **33** appears to reduce the binding affinity, as compared to **5** and **32**. On a related note, we notice that only this and the subjective QSAR do not substantially overestimate the affinity of compound **46** in the test set. We suggest that, due to the lack of a specific 10A descriptor, the presence of the unusual 10α-methyl substitution in **46** can only be conveyed by the CHC10 descriptor. Since the electronic effect of a methyl group is nonstereospecific (i.e., CHC10 for 10α- or 10β-methyl $\sim$ −0.03), the two QSAR models that include this descriptor give a lower predicted affinity that is consistent with other 10β-methyl analogs. The apparent role of CHC10 presented in the analysis above also brings up an important issue. While the dependence of activity as a function of CHC10 is consistent with the structural features in the data set, there is still a question of whether the change of binding affinity can be uniquely attributed to electrostatic interaction at C_{10}. Since the 10-methyl derivatives are in general less active, it is quite reasonable to assume that steric conflicts with the receptor may also be a key factor. However, given that the number of substituents at position 10 is rather limited (either H or CH_3), it would not be appropriate to draw a conclusion at the present time.

The functional dependence of VOL_6A suggests that an increase in bulk at this substituent position weakens binding somewhat. This trend agrees with the SAR that compounds **9** and **5** bind to the progesterone receptor less tightly than **47** and **45**, respectively, where the two sets of compounds only differ by a methyl substitution at the 6α-position.

The GNN model indicates that binding can be enhanced by increasing the SUR_11X value. This descriptor is nonzero when the C_{11} substituent is connected to the steroid backbone via a double bond (e.g. C_{11}═CH_2). Without exception, compounds with this feature are far more potent than those without (e.g. **13** > **11**, **22** > **23**, **27** > **28** and **48** > **12**, **14**−**19**, **45**). As already mentioned, because SUR_11X does not distinguish between *E*- or *Z*-fluoromethylene as in **53** and **54**, the two structural isomers are predicted with an almost identical binding affinity (about 1.85); their observed log RBA values are actually quite different (1.53 and 2.28).

The positive slope of the VOL_13B dependence plot indicates that an increase of volume at this position will enhance binding. Indeed, several examples in the data set show that the 13β-ethyl compounds have stronger binding than the corresponding methyl analogs (**11** > **45**, **26** > **28**, **13** > **48**, and **51** > **27**). The only exception to this trend is the **20**/**24** pair. One possible explanation is that, based on conformation analysis, the 13β-ethyl group **20** has a greater conformational freedom relative to the other ethyl analogs in the data set. The gain in the free energy in binding, presumably enthalpic and arising from favorable van der Waals interactions, is offset by a greater entropic loss when the conformation of the ethyl group is restricted by the receptor environment.

Because CH_17B and DIP17 are related (i.e., DIP17 has contributions from both 17β and 17α substituents), their influence on binding affinity is monitored by a systematic variation of both parameters, while keeping the other six descriptors at the template values. The results are shown in Figure 6b, where the contours display the different levels of binding affinity. It suggests that optimal binding is attained by a high dipole moment and a CH_17B value close to −0.05. On the same figure, the various 17β/-α substituent pairs in the training set are plotted at their corresponding descriptor values. It appears that binding affinity increases in the order $COCH_3/OCOCH_3$ > OH/CCH > $COCH_2OH/H$ > OH/C-^{i}C_3H_7, OH/CH_2CH_3 > $COCH_3/H$ > $OH/CCCH_3$.

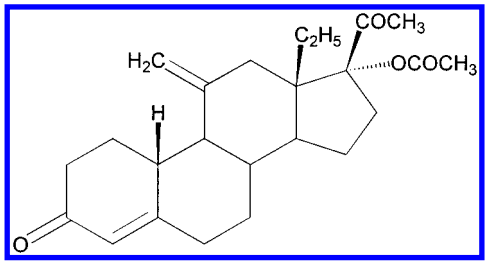

Figure 7. Chemical structure of a potent analog predicted by the GNN model.

Since the total electronic energy of a molecule becomes increasingly negative as its molecular weight increases, it is not unexpected that the two whole molecule descriptors, MOLWGT and ELEC, are strongly anticorrelated ($r = -0.97$) for the current data set. Nevertheless, the absence of either descriptor resulted in a significant decrease in predictivity (standard GNN model $q^2 = 0.72$; without ELEC $q^2 = 0.48$; without MOLWGT $q^2 = 0.61$). This suggests that there is useful information in the variance that is uncorrelated between the two descriptors which accounts for the binding. Figure 6c is a contour plot showing the variation of binding affinity as functions of the two descriptors. It clearly shows that ELEC exerts a greater variation on the binding affinity, as compared to MOLWGT. Specifically, the figure suggests that for large negative ELEC, MOLWGT is not so important; only at higher ELEC values, MOLWGT becomes more significant. This may also explain why ELEC is preferably selected in most of the other QSAR models as well.

In summary, a dependency plot provides useful visualization and can be applied to suggest new compounds.[17] The above analysis points to several structural modifications to **45** which could lead to a more potent progestagen. One such compound, shown in Figure 7, is suggested by combining the various substituent fragments that exhibit positive influence on binding.

IV. CONCLUDING DISCUSSIONS

A number of descriptor selection QSAR methods have been applied to the same group of descriptors for a set of progestagenic steroids. It is clear that selection of descriptors by an experienced practitioner does not always lead to a satisfactory QSAR. With a data set of this complexity, an objective selection derived from a chemometric study can lead to better models. The use of stepwise regression and the genetic functional approximation regression seem appropriate for an initial screening of the data set because these two methods are computationally inexpensive.[21] However, the resulting QSARs are not very good, particularly because of the way in which nonlinearity is treated. To obtain improved QSARs, an intrinsically nonlinear mapping method, such as a neural network, is required. When used in conjunction with a descriptor selection method, such as simulated annealing or a genetic algorithm, excellent results are obtained. In this application, as in earlier studies,[16,17,22] a combination of genetic algorithm and neural network provides the best results for training and test sets. The GNN model is further validated by a randomization test which confirms that the QSAR is not a result of chance correlation. A detailed functional analysis of the descriptors indicates that the *quantitative* SAR is consistent with established *qualitative* SAR.

In contrast to more conventional statistical techniques, a genetic algorithm does not result in a single model but rather a number of models, often of similar quality. From an analysis point of view one might consider this a disadvantage because the interpretation is less straightforward. On the other hand, in many situations there is no single best solution[6] and the set of QSARs provided by the GA properly reflects this situation. One thing that has not been addressed in this paper is the way to utilize multiple models derived from GNN. A number of studies have suggested that optimal use of a composite model will lead to more robust predictions.[6,16,65,66]

ACKNOWLEDGMENT

We wish to thank Dr. H. Hamersma for stimulating and helpful discussions on the QSARs of progestagens. We would like to acknowledge the reviewers for their insightful comments that helped to improve this manuscript. This work is supported in part by a grant supplement from the GOALI program of the National Science Foundation.

REFERENCES AND NOTES

(1) Kubinyi, H. Variable selection in QSAR studies. I. An evolutionary algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285−294.
(2) Leardi, R.; Boggia, R.; Terrile, M. Genetic algorithms for a strategy for feature selection. *J. Chemom.* **1992**, *6*, 267−281.
(3) Leardi, R.; Boggia, R.; Terrile, M. Genetic algorithms as a strategy for feature selection. *J. Chemom.* **1992**, *6*, 267−281.
(4) Wikel, J. H.; Dow, E. R. The use of neural networks for variable selection in QSAR. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 645−651.
(5) Sutter, J. M.; Kalivas, J. H. Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. *Microchem. J.* **1993**, *47*, 60−66.
(6) Rogers, D. R.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure−activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(7) 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.
(8) Luke, B. T. Evolutionary programming applied to the development of quantitative structure−activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−1287.
(9) Leardi, R. Application of genetic algorithms feature selection under full validation conditions and to outlier detection. *J. Chemom.* **1994**, *8*, 65−79.
(10) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated descriptor selection for quantitative structure−activity relationships using generalized simulated annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.
(11) Wessel, M. D.; Jurs, P. C. Prediction of normal boiling points for a diverse set of industrially important organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 841−850.
(12) Hahn, M.; Rogers, D. Receptor surface models. 2. Application to quantitative structure−activity relationships studies. *J. Med. Chem.* **1995**, *38*, 2091−2102.
(13) Maddalena, D. J.; Johnston, G. A. R. Prediction of receptor properties and binding affinity of ligands to benzodiazepine/GABAA receptors using artificial neural networks. *J. Med. Chem.* **1995**, *38*, 715−724.
(14) Sutter, J. M.; Jurs, P. C. Prediction of aqueous solubility for a diverse set of heteroatom-containing organic compounds using a quantitative structure−activity relationship. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 100−107.
(15) Kubinyi, H. Evolutionary variable selection in regression and PLS analyses. *J. Chemom.* **1996**, *10*, 119−133.
(16) So, S.-S.; Karplus, M. Evolutionary optimization in quantitative structure−activity relationship: an application of genetic neural network. *J. Med. Chem.* **1996**, *39*, 1521−1530.
(17) So, S.-S.; Karplus, M. Genetic neural networks for quantitative structure−activity relationships: Improvements and application of benzodiazepine affinity for benzodiazepine/$GABA_A$ receptors. *J. Med. Chem.* **1996**, *39*, 5246−5256.
(18) Mitchell, B. E.; Jurs, P. C. Prediction of autoignition temperature of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 538−547.

(19) Engelhardt, H. L.; Jurs, P. C. Prediction of supercritical carbon dioxide solubility of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 478−484.
(20) Hasegawa, K.; Miyashita, Y.; Funatsu, K. GA strategy for variable selection in QSAR studies: GA based PLS analysis of calcium channel antagonists. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 306−310.
(21) Mitchell, B. E.; Jurs, P. C. Prediction of aqueous solubility of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489−496.
(22) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.
(23) Leardi, R.; González, A. L. Genetic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemom. Intell. Lab. Syst.* **1998**, *41*, 195−207.
(24) Shi, L. M.; Fan, Y.; Myers, T. G.; O'Conner, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the NCI anticancer drug discover databases: Genetic function approximation for the QSAR study of anticancer ellipticine analogues. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 189−199.
(25) Hasegawa, K.; Funatsu, K. GA strategy for variable selection in QSAR studies: GAPLS and D-optimal designs for predictive QSAR model. *J. Mol. Struct. (THEOCHEM)* **1998**, *425*, 255−262.
(26) Hasegawa, K.; Kimura, T.; Funatsu, K. GA strategy for variable selection in QSAR studies: Enhancement of comparative molecular binding energy analysis by GA-based PLS method. *Quant. Struct.-Act. Relat.* **1999**, *18*, 262−272.
(27) Hasegawa, K.; Kimura, T.; Funatsu, K. GA strategy for variable selection in QSAR studies: Application of GA-based region selection to a 3D-QSAR study of acetylcholinesterase inhibitors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 112−120.
(28) Kimura, T.; Hasegawa, K.; Funatsu, K. GA strategy for variable selection in QSAR studies: GA-based region selection for CoMFA modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 276−282.
(29) *Cerius*^{2}, Version 2.0; Molecular Simulations Inc.: San Diego, CA, 1996.
(30) Kowar, T. R. Genetic function approximation experimental design (GFAXD): A new method for experimental design. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 858−866.
(31) Teutsch, G.; Weber, L.; Page, G.; Shapiro, E. L.; Herzog, H. L.; Neri, R.; Collins, J. A. Influence of 6-Azido and 6-Thiocyanato Substitution on Progestational and Corticoid Activities and a Structure−Activity Correlation in the D^6-6-Substituted Progesterone Series. *J. Med. Chem.* **1973**, *16*, 1370−1376.
(32) Coburn, R. A.; Solo, A. J. Quantitative Structure−Activity Relationships among Steroids. Investigations of the Use of Steric Parameters. *J. Med. Chem.* **1976**, *19*, 784−754.
(33) Lee, D. L.; Kollman, P. A.; Marsh, F. J.; Wolff, M. E. Quantitative relationships between steroid structure and binding to putatve progesterone receptors. *J. Med. Chem.* **1977**, *20*, 1139−1146.
(34) van den Broek, A. J.; Broess, A. I. A.; van den Heuvel, M. J.; de Jongh, H. P.; Leemhuis, J.; Schönemann, K. H.; Smits, J.; de Visser, J.; van Vliet, N. P.; Zeelen, F. J. Strategy in drug research. Synthesis and study of the progestational and ovulation inhibitory activity of a series of 11b-substituted-17a-ethynyl-4-estren-17b-ols. *Steroids* **1977**, *30*, 481−510.
(35) Moriguchi, I.; Komatsu, K.; Matsushita, Y. Pattern recognition for the study of structure−activity relationships. *Anal. Chim. Acta* **1981**, *133*, 625−636.
(36) Daux, W. L.; Griffin, J. F.; Rohrer, D. C. Steroid conformation, receptor binding, and hormone action. In Horn, A. J., de Ranter, C. J., Eds.; X-ray crystallography and drug action, Course of the International School of Crystallography 9th; Oxford University Press: Oxford, U.K., 1984; pp 406−426.
(37) Belaisch, J. Chemical classification of synthetic progestogens. *Rev. Fr. Gynecol. Obstet.* **1985**, *80*, 473−477.
(38) Doré, J. C.; Gilbert, J.; Ojasoo, T.; Raynaud, J. P. Correspondence analysis applied to steroid receptor binding. *J. Med. Chem.* **1986**, *29*, 54−60.
(39) Hopper, H. O.; Hammann, P. The influence of structure modification on progestrone and androgen receptor binding of norethisterone. Correlation with nucluear magnetic resonance signals. *Acta Endocrinol. (Copenhagen)* **1987**, *115*, 406−412.
(40) Ojasoo, T.; Doré, J. C.; Gilbert, J.; Raynaud, J. P. Binding of Steroids to the Progestin and Glucocorticoid Receptors Analyzed by Correspondence Analysis. *J. Med. Chem.* **1988**, *31*, 1160−1169.
(41) Loughney, D. A.; Schwender, C. F. A comparison of progestin and androgen receptor binding using the CoMFA technique. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 569−581.
(42) van Helden, S. P.; Hamersma, H.; van Geerestein, V. J. Prediction of progesterone receptor binding of steroids using a combination of genetic algorithms and neural networks. In *Genetic Algorithm in Molecular Modeling*; Devillers, J., Ed.; Academic Press: London, U.K., 1996.
(43) *Cerius*^{2}, Version 1.5; Molecular Simulations Inc.: San Diego, CA, 1995.
(44) Bergink, E. W.; van Meel, F.; Turpijn, E. W.; van der Vies, J. Binding of progestagens to receptor proteins in MCF-7 cells. *J. Steroid Biochem.* **1983**, *19*, 1563−1570.
(45) Rohrer, D. C.; Hazel, J. P.; Duax, W. L.; Zeelen, F. J. 11b-Methyl-19-nor-17a-pregn-4-en-20-yn-17b-ol ($C_{21}H_{30}O$). *Cryst. Struct. Commun.* **1976**, *5*, 543−546.
(46) van Geerestein, V. J. Structure of the enthanol solvate of 13-ethyl-11b-methyl-18-norlynestrenol. *Acta Crystallogr., Sect. C* **1988**, *44*, 376−378.
(47) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1990**, *3* (6C), 537−547.
(48) *Chem-X*, Version 1.5 Chemical Design Ltd.: Chipping Norton, U.K., 1995.
(49) Stewart, J. J. P. MOPAC: A semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−105.
(50) *The CRC Handbook*; CRC Press: Boca Raton, FL, 1995.
(51) Available upon request from S. P. van Helden; e-mail: s.helden@organon.oss.akzonobel.nl.
(52) Topliss, J. G.; Edwards, R. P. Chance factors in studies of quantitative structure−activity relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(53) Møller, M. F. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks* **1993**, *6*, 525−533.
(54) Maggiora, G. M.; Elrod, D. W. Computational neural networks as model-free mapping devices. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 732−741.
(55) Kirkpatrick, S.; Gelatt, C. D. J.; Vecchi, M. P. Optimization by simulated annealing. *Science* **1983**, *220*, 671−680.
(56) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.
(57) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical recipes in C*; Cambridge University Press: Cambridge, U.K., 1992.
(58) 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.
(59) Shao, J. Linear-model selection by cross-validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486−494.
(60) Cartwright, H. M. *Applications of artificial intelligence in chemistry*; Oxford University Press: Oxford, U.K., 1993.
(61) Livingstone, D. J.; Manallack, D. T. Statistics using neural networks: Chance effects. *J. Med. Chem.* **1993**, *36*, 1295−1297.
(62) van de Waterbeemd, H. Chemometric methods in Molecular Design; VCH: Weinheim, Germany, 1995.
(63) Jouan-Rimbaud, D.; Massart, D. L.; de Noord, O. E. Random correlation in variable selection for multivariate calibration with a genetic algorithm. *Chemom. Intell. Lab. Syst.* **1996**, *35*, 213−220.
(64) So, S.-S.; Richards, W. G. Application of neural networks: Quantitative structure−activity relationships of the derivatives of 2,4-diamino-5-(substituted-benzyl)pyrimidines as DHFR inhibitors. *J. Med. Chem.* **1992**, *35*, 3201−3207.
(65) Rogers, D. Evolutionary statistics: Using a genetic algorithm and model reduction to isolate alternate statistical hypotheses of experimental data. *The Seventh International Conference on Genetic Algorithms*; Morgan-Kaufmann: San Francisco, 1997.
(66) So, S.-S.; Karplus, M. A comparative study of ligand-receptor complex binding affinity prediction methods based on glycogen phosphorylase inhibitors. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 243−258.

CI990130V