# Receptor Surface Models. 2. Application to Quantitative Structure–Activity Relationships Studies

Mathew Hahn* and David Rogers*

*Molecular Simulations Incorporated, 16 New England Executive Park, Burlington, Massachusetts 01803-5297**

*Received December 6, 1994*®

A new technique for using receptor surface models in quantitative structure–activity relationship (QSAR) analysis is described. Receptor surface models provide compact, quantitative descriptors which capture three-dimensional information about putative receptor/ligand interactions. Receptor surface models can be constructed quickly, which allows the construction of multiple plausible models; a variable selection technique such as genetic function approxinmation (GFA) can then be used to suggest which receptor surface models provide the most valuable descriptors for QSAR. Advantages of this approach are shown by applying it against two previously-published and well-studied QSAR data sets. Our results indicate that the approach can model data as effectively as established 3D-QSAR techniques.

## 1. Background: QSAR Modeling

Quantitative structure–activity relationship (QSAR) modeling is an area of research pioneered by Hansch and Fujita;[1] QSAR attempts to model the "activity" of a series of compounds using measured or computed properties of the compounds. More recently, QSAR has been extended by including in the analysis three-dimensional information about the series, either through grid-based data such as the comparative molecular field analysis (CoMFA) approach,[2] or by three-dimensional shape descriptors, as illustrated by the molecular shape analysis (MSA) approach.[3] While the original Hansch work used multiple linear regression (MLR) to combine different descriptors in the data set, MLR has proven difficult or impossible to use for data sets that contained large numbers of descriptors. Subsequent work has demonstrated the ability of partial least-squares (PLS) regression to build models of data sets containing large numbers of descriptors.[4,5] For data sets containing large numbers of descriptors, in which the necessary information is localized in a relatively few descriptors, genetic function approximation (GFA) analysis[6] is a recent innovation that uses a genetic algorithm[7] to find an appropriate subset of descriptors, which are fitted in turn with MLR.

Many regression techniques develop a single model or a relatively small number of models. In contrast, the genetic function approximation algorithm develops a population of many models. The population of models is evolved by repeatedly performing the genetic crossover operation to recombine the terms of the better-performing models. Upon completion, one typically selects the model from the population with the best score. However, it is sometimes preferable to inspect many different models and select one or more models based on the appropriateness of the descriptors by applying chemical intuition, in addition to using the scores.

Of particular interest is the ability of GFA to discover nonlinear QSARs. Many techniques are dependent on the existence of linear relationships between descriptors in the data set and the activities. Even if a linear relationship is not apparent between some descriptor and the activity, some subrange of the descriptor may still have a significant linear relationship with the activity, even if the remainder of the range is uninformative. Linear modeling methods such as least-squares regression, stepwise regression, or partial least-squares regression will not discover these relationships or utilize such descriptors effectively.

GFA allows the discovery and use of nonlinear descriptors by using spline-based terms. If nonlinear relationships are suspected, the GFA process can be set to include splines. The splines used are *truncated power splines* and are denoted with angle brackets, where $<a - f(x)>$ is equal to zero if the value of $(a - f(x))$ is negative, otherwise it is equal to the value of $(a - f(x))$. For example, $<-9.852 - E_{\text{interact}}>$ is zero when $E_{\text{interact}} > -9.852$, and equal to $(-9.852 - E_{\text{interact}})$ otherwise. The constant $a$ is called the *knot* of the spline. When a spline term is created, the knot is set using the value of the descriptor in a random data sample. Because a spline term contains an extra constant, the scoring function (which takes into account the size of the model) counts splines the same as two linear terms. Therefore, splines are only included in the model if they reduce the training error more than two linear terms. This scoring helps eliminate bias toward splines when we mix spline and linear terms in the GFA process. (A more detailed description of the GFA algorithm can be found elsewhere.)[6]

## 2. Background: Receptor Surface Modeling

It is common in a QSAR analysis to have measured binding affinities for a set of compounds to a particular protein but not to have knowledge of the three-dimensional structure of the protein active site. A number of methods, called *receptor mapping techniques*, attempt to provide insight about the active site and to characterize receptor binding requirements. Often receptor mapping techniques are used to generate a hypothetical model of the actual receptor site. This is known as a *receptor site model*.[8-11] In this paper we describe a specific type of receptor site model, called a *receptor surface model (RSM)*. A compound is energy

* The authors can be reached via Internet as matth@msi.com. and drogers@msi.com.

® Abstract published in *Advance ACS Abstracts,* May 15, 1995.

0022-2623/95/1838-2091$09.00/0

minimized within the context of a receptor surface model to generate three-dimensional descriptors for use in QSAR analysis. (The method used for generating receptor surface models and descriptors has been described elsewhere.)[9]

Three-dimensional energetics descriptors can be calculated from either receptor surface model/ligand interactions or, alternatively, from actual protein/ligand interactions (if the protein is known), as advocated by Hopfinger.[12,13] These three-dimensional descriptors may be used alone or in combination with two-dimensional descriptors in QSAR analysis.

A receptor surface model is generated using some subset of the most active structures. The rationale underlying these models is that the most active structures tend to explore the best spatial and electronic interactions with the receptor, while the least active do not and tend to have unfavorable steric or electronic interactions. This is also the basis of the active analogue approach.[14]

As currently generated, receptor surface models are *exemplars* of the common features present in the most active compounds. This can be contrasted with methods which are *interpolative*, correlating differences in the parameters between the most and least active with activity.[15] By using only a subset of the most active models, the issues of conformational selection and alignment are reduced. Further, while it is possible to use interpolative methods in the generation of the receptor model, we found the exemplar approach to be simpler yet still effective when used with a series of closely related analogues.

Once chosen for the construction of the receptor surface model, the compounds must be aligned, preferably in conformations that reflect the active, "bound" conformations. The alignment may be achieved using any of a number of published methods.[16-22]

Receptor surface models are best constructed from a set of the most active analogues that are chosen to cover the variety of steric and electrostatic variations likely to appear in the test data. One approach is to visually inspect the training set and manually select a structurally diverse subset of the most active structures. Yet another approach is to automatically build a set of different receptor surface models from different combinations of the most active analogues, and then use a variable-selection technique such as GFA to discover the receptor surface model whose descriptors yield the best QSARs of the full training set. Our studies suggest that the selection of the actives is an important consideration, though models constructed with nonoptimal sets of compounds nearly always show some amount of predictiveness.

Once the desired receptor surface model has been constructed, all the structures in the training and test sets can be evaluated against the model. The evaluation consists of computing several energetic descriptors that are based upon the interactions between ligand and model. For this work, four descriptors are generated.

The first descriptors is the nonbonded energy of interaction between the ligand and the receptor surface model. This term is the sum of the nonbonded van der Waals and electrostatic energies, and is denoted $E_{\text{interaction}}$.

The second descriptor is the intramolecular strain energy (enthalpy) of the ligand inside the receptor surface model. Here, an energy minimization is performed in which the conformation of the ligand is optimized to adopt a minimum energy configuration with respect to the receptor model. This is analogous to minimizing a structure within an actual receptor by fixing all receptor atom positions, allowing for freedom in the ligand. This descriptor is denoted $E_{\text{inside}}$.

Another energy minimization is performed to calculate a third descriptor. The bound conformation structure (from descriptor $E_{\text{inside}}$) is minimized again in the absence of the receptor surface model influence and the internal strain energy that is *induced* by the receptor model is calculated. The descriptor for non-receptor-bound energy is called $E_{\text{relaxed}}$. Since this minimization will put the structure in the closest energy minima relative to the receptor-bound conformation, $E_{\text{relaxed}}$ is always less than or equal to $E_{\text{inside}}$.

The final descriptor, called $E_{\text{strain}}$, is the difference between $E_{\text{inside}}$ and $E_{\text{relaxed}}$. This descriptor corresponds to a $\delta$ strain energy between a bound conformation and its closest relaxed unbound conformation. It does not indicate anything about the $\Delta$ strain between the bound conformation energy and the global energy minimum. If a conformational search has produced a global energy–minimum energy ($E_{\text{global}}$) a descriptor corresponding to $E_{\text{inside}} - E_{\text{global}}$ could be used in addition to, or in place of, $E_{\text{strain}}$.

These descriptors are representing components of the binding energies of ligands in the putative receptor site; they may be useful for QSAR modeling in cases where the activity is correlated with the ligand binding energies. This correlation is frequently (though not always) found.

This paper will illustrate the utility of these descriptors in QSAR and demonstrate them to be a compact and effective representation of three-dimensional ligand interaction information.

## 3. Analysis of the Corticosteroid–Globulin Binding Data Set

We applied the receptor surface modeling–QSAR combination to a standard data set consisting of steroid binding data. This data set has been previously studied using several different techniques, including CoMFA[2] and Compass.[23] The data set consists of 31 steroids assayed for binding affinity to the transport protein, corticosteroid binding globulin (CBG). These steroids are shown in Figure 1. The training set consisted of the first 21 molecules in the series and was used to generate a QSAR model. This model was then used to predict the affinity of the remaining 10 molecules.

This data set has relatively little conformational flexibility and is relatively simple to align, although Jain et al. suggest that the importance of conformational flexibility may be underestimated.[23] The reduced role of conformation and alignment in this data set allows better examination of the effects of energetics separate from the complications induced by issues of multiple conformations and alignments.

To build a QSAR model from the 21 compound training set (compounds **1–21**), an initial receptor surface model was constructed from the six most active steroids (compounds **6**, **7**, **10**, **11**, **19**, and **20**). We chose

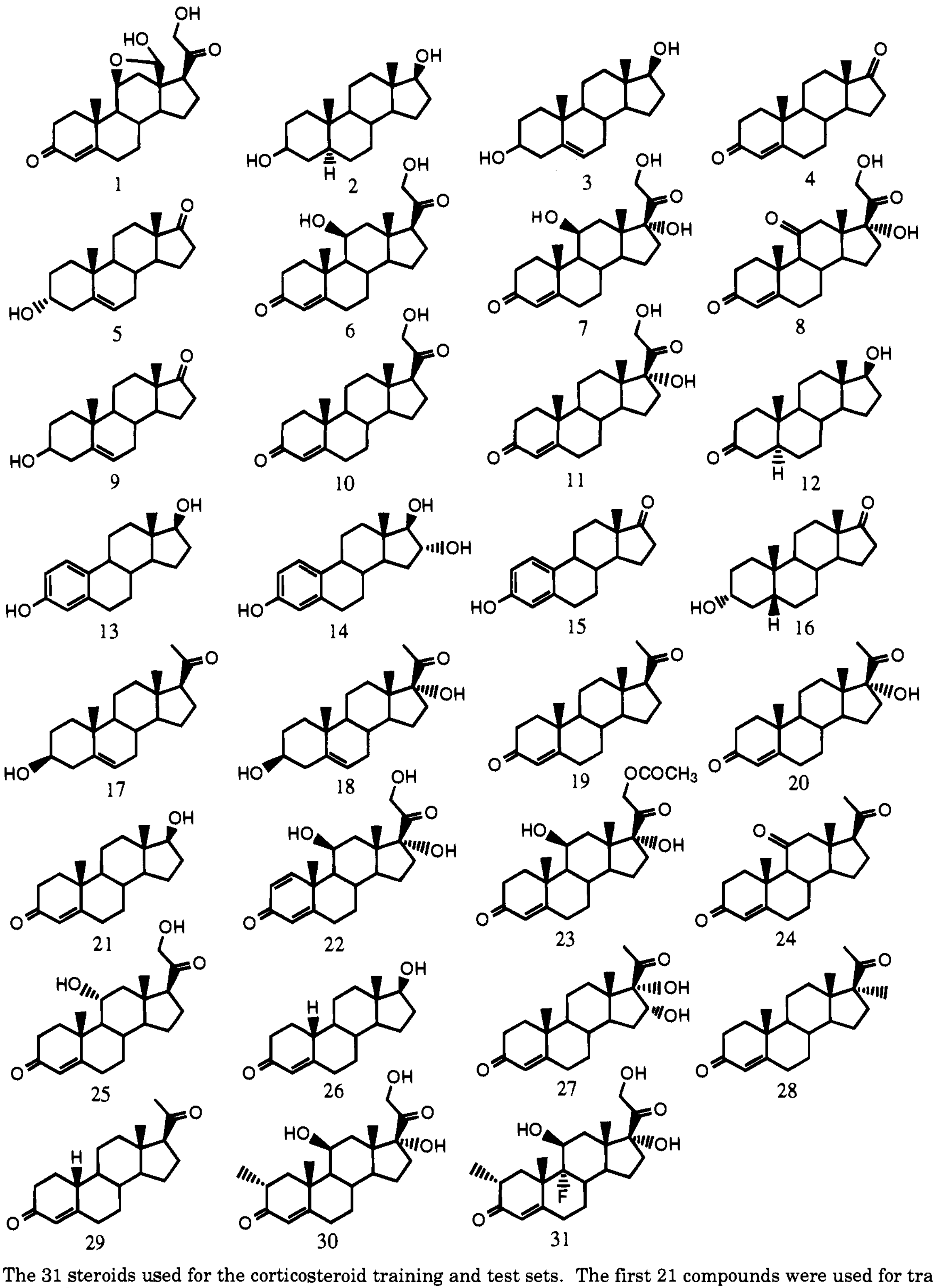


**Figure 1.** The 31 steroids used for the corticosteroid training and test sets. The first 21 compounds were used for training and the remaining 10 for testing.

the top six because they covered the range of structural variation seen in the active compounds. A low-energy conformation for each molecule was generated by minimizing from standard steroid conformations. The six most active molecules were then aligned to minimize backbone ring system and side chain RMS differences. A receptor surface model was then constructed around the six aligned compounds. This model is shown at the top of Figure 2.

This initial receptor surface model completely surrounds the aligned molecules; it is a *closed* receptor surface model.[9] The model is sterically overconstrained; previously unseen steric variation in a test compound is assumed to be detrimental. While this may or may not be the case with respect to the actual binding of the molecule to a receptor site, we argue it is reasonable for the model to initially assume such variations are detrimental, since there is no way to validate their utility given the information in the training set. Receptor surface models are conservative in this sense and tend to underpredict the activity of novel steric variants.

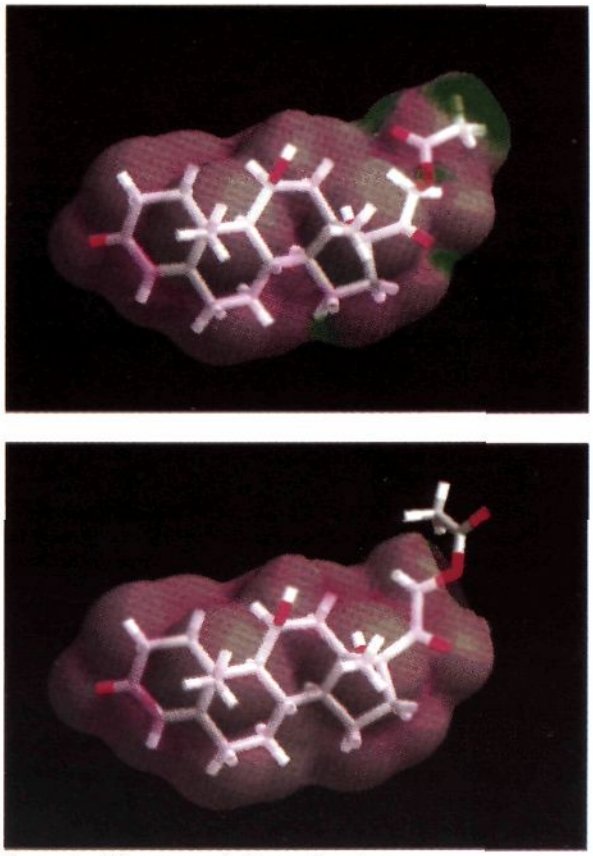


**Figure 2.** The top model is a closed receptor surface model generated from the six most active compounds in the corticosteroid data set. The bottom model is an open receptor surface model. A test molecule (compound **23**) has been minimized within the model, and the VDW interaction energy has been mapped onto the surface. Purple denotes regions of favorable VDW interaction, and green denotes regions of unfavorable VDW interactions. The visualization of interaction energy focuses attention onto the acetoxy group on C-21, which assumes a highly-strained geometry to fit into the model. The open model shows no such strain; the opening in the region around C-21 allows the acetoxy group to extend outside of the model and eliminates the severe strain seen with this molecule in the closed model. This shows how the receptor surface model can first assist the user in pinpointing regions of poor interaction and then allow the editing of the model to incorporate the user's estimation of critical and noncritical regions.

The chemist, however, may wish to allow variation in specific regions; for example, to represent solvent openings in a receptor site, to represent the lack of knowledge about specific regions of the receptor site, or to allow scientifically justified steric variation suggested by a novel test molecule. For any of these reasons, the receptor surface model can be made *open* by removing regions of the surface. Portions of the test molecules can then extend through these openings, and upon evaluation will not be considered detrimental. (Further discussion of open and closed receptor surface models can be found in ref 9.)

We illustrate opening the receptor surface model with the corticosteroid data set in Figure 2. Compound **7** and **23** are identical except that compound **23** has an acetoxy group in place of an hydroxy group on C-21. Both compounds have similar activity. Two possible explanations for this similarity in activity follow. First, the acetoxy group may be hydrolyzed at physiological pH before binding, yielding compound **17**. A closed receptor mode predicts this compound correctly. Second, the acetoxy group may not be hydrolyzed and is accommodated in the receptor active site (or is exposed to solvent). We can handle the second case and allow the acetoxy group to reside outside of the receptor surface by operating the surface at C-21. Such editing is not required for CoMFA and Compass since they do not make the same sterically conservative assumption that is used during the construction of the receptor surface model; however, this can be dangerous, since there is no *a priori* penalty for novel steric variants. For the receptor surface model, such variants are penalized unless the user makes the direct decision to allow the variant by opening the model. (Of course, the user must be careful to apply the opened model only against compounds where the steric variation in the opening is deemed reasonable.)

$\text{CBG} = 3.476 - 0.223 * \mathbf{E_{interact}}$

| | |
|---|---|
| N: 21 | |
| $r^2$: | 0.702 |
| Regression-only CV-$r^2$: | 0.646 |
| Test set $r^2$ (including 23): | 0.006 |
| Test set $r^2$ (excluding 23): | 0.696 |

**Figure 3.** The QSAR model generated with the energetic descriptors from the closed receptor surface model for the corticosteroid binding data set. The model rated best contained only a single linear term of $E_{\text{interact}}$. The model predicts poorly against the test set, primarily due to compound **23**, whose activity is significantly underestimated. This is reflected in the test set $r^2$, which is poor when compound **23** is included but good when it is excluded.

$\text{CBG} = 3.498 - 0.236 * \mathbf{E_{interact}}$

| | |
|---|---|
| N: 21 | |
| $r^2$: | 0.664 |
| Regression-only CV-$r^2$: | 0.628 |
| Test set $r^2$: | 0.652 |

**Figure 4.** The QSAR model generated with the energetic descriptors from the open receptor surface model for the corticosteroid binding data set. The model rated best contained only a single linear term of $E_{\text{interact}}$. While the $r^2$ may appear only moderate against the training set, the nearly equivalent value for $r^2$ when applied to the test set is strongly suggestive of a predictive model.

After building the closed and open receptor surface models, energetic descriptors for all 31 compounds in the training and test sets were generated for each of the two receptor surface models. Scatterplots of the descriptors versus activity showed no sign of nonlinearity, so we decided to use only linear terms in the models.

The best model generated using the descriptors from the closed receptor surface model is given in Figure 3, along with the number of compounds $N$, the correlation coefficient squared ($r^2$) over the training set, the regression-only cross-validated $r^2$, and $r^2$ over the test set, both with and without compound **23**. The receptor surface model-based descriptors, $E_{\text{interact}}$ and $E_{\text{strain}}$, the CBG activity values, the predictions, and residuals of the QSAR are shown in Table 1. The model generated using the descriptors from the open receptor surface model is given in Figure 4. The value of $E_{\text{interact}}$, the CBG activity values, and the predictions and residuals of this new QSAR model, CoMFA, and Compass are given in Table 2. Scatterplots of the activity versus prediction for the

**Table 1.** Corticosteroid Data Descriptor for the Closed Receptor Surface Model[a]

| | | | | $RSM_{closed}$ | |
|---|---|---|---|---|---|
| no. | $E_{interact}$ | $E_{strain}$ | CBG | predicted | residual |
| 1 | −16.6 | 10.4 | 6.279 | 6.995 | 0.716 |
| 2 | −10.4 | 10.8 | 5.000 | 5.862 | 0.862 |
| 3 | −9.8 | 1.8 | 5.000 | 5.711 | 0.711 |
| 4 | −8.4 | 0.0 | 5.763 | 5.342 | −0.420 |
| 5 | −6.1 | 6.5 | 5.613 | 4.764 | −0.848 |
| 6* | −17.7 | 0.0 | 7.881 | 7.324 | −0.556 |
| 7* | −18.8 | 0.0 | 7.881 | 7.619 | −0.261 |
| 8 | −18.0 | 2.9 | 6.892 | 7.413 | 0.521 |
| 9 | −6.2 | 1.5 | 5.000 | 4.795 | −0.204 |
| 10* | −16.7 | 0.0 | 7.653 | 7.096 | −0.556 |
| 11* | −17.7 | 0.0 | 7.881 | 7.363 | −0.517 |
| 12 | −12.2 | 2.4 | 5.919 | 6.296 | 0.377 |
| 13 | −8.9 | 7.3 | 5.000 | 5.511 | 0.511 |
| 14 | −8.4 | 10.3 | 5.000 | 5.399 | 0.399 |
| 15 | −6.3 | 6.8 | 5.000 | 4.825 | −0.174 |
| 16 | −5.3 | 11.4 | 5.225 | 4.570 | −0.654 |
| 17 | −11.2 | 1.9 | 5.225 | 6.025 | 0.800 |
| 18 | −12.1 | 1.4 | 5.000 | 6.274 | 1.2745 |
| 19* | −13.8 | 0.0 | 7.380 | 6.650 | −0.729 |
| 20* | −14.9 | 0.0 | 7.740 | 6.913 | −0.826 |
| 21 | −12.2 | 0.0 | 6.724 | 6.297 | −0.426 |
| 22 | −18.1 | 0.0 | 7.512 | 7.505 | −0.007 |
| 23 | −2.7 | 132.4 | 7.553 | 4.083 | −3.469 |
| 24 | −13.9 | 2.0 | 6.779 | 6.575 | −0.203 |
| 25 | −15.7 | 7.0 | 7.200 | 6.975 | −0.224 |
| 26 | −11.6 | 0.0 | 6.114 | 6.060 | −0.053 |
| 27 | −14.6 | 8.0 | 6.247 | 6.720 | 0.473 |
| 28 | −13.7 | 0.8 | 7.120 | 6.520 | −0.599 |
| 29 | −13.401 | 0.0 | 6.817 | 6.461 | −0.355 |
| 30 | −16.1 | 11.4 | 7.688 | 7.070 | −0.617 |
| 31 | −11.6 | 20.6 | 5.797 | 6.049 | 0.252 |

[a] This table contains the indexes $E_{interact}$ and $E_{strain}$ for the closed receptor surface model, the corticosteroid binding affinity CBG, and the predictions and residuals. The training set was comprised of compounds **1**−**21**; the test set was comprised of compounds **22**−**31**. The compounds with the asterisk (*) after their index were used in the construction of the receptor surface model.

test compounds derived using the open receptor surface QSAR model, CoMFA, and Compass are shown in Figure 5.

The experiment described indicates the superior performance of the open receptor model as a source of descriptors for QSAR model generation. This improvement came about because we were able to use qualitative knowledge (from other studies and visualization) in designing the opening, followed by use of these quantitative descriptors to build QSAR models. This illustrates one of the important strengths of receptor surface model-based QSAR modeling: the ability to move between *qualitative* and *quantitative* descriptions, combining insights gained from each in building the final QSAR model. In this particular case, the visualization of the strain around C-21, combined with chemical knowledge about the offending acetoxy group, guided the construction of the C-21 opening in the receptor surface model. This opening then led to a quantitative improvement in the predictivity of the resulting QSAR model.

Receptor surface models can give results that are quantitatively different from other 3D-QSAR methods. For example, previous studies using CoMFA[2] or Compass[23] substantially overestimate the activity of compound **31** (see Table 1). Compound **31** is a fluorine derivative of compound **30** but is about 100 times less active. CoMFA predicts the compound to be 10 times more active than it is. Compass predicts this compound to be the most active of the 10 test molecules, when in fact it is the least active. The receptor surface model-derived QSAR correctly predicts that compound **31** is the least active. This is because the evaluation procedure has an empirical solvation correction term that penalizes polar groups placed in hydrophobic regions (see ref 9). The polar fluorine atom is positioned in a region that is predicted to be hydrophobic based upon the examination of the six most active compounds. Purely statistical techniques such as CoMFA do not directly use higher-level chemical knowledge when building their models. The receptor surface model allows such knowledge to be incorporated as part of the evaluation of energetics.

Further, the ability to minimize the test molecule within the receptor surface model allows other subtle chemical relationships to be handled. For example, starting with compound **31**, one can compare the effect of the fluorine on atom C-9 against the addition of a methyl group on atom C-2. The small fluorine atom on C-9 introduces about the same strain as does the addition of a larger methyl group to atom C-2. Since the evaluation procedure minimizes a molecule inside the surface, the methyl group can be reasonably accommodated in the surface by sliding the A ring laterally away from the surface and the concomitant small adjustment of angles and bonds in the remaining rings.

Finally, the chemist, using high-level chemical knowledge, can refine the receptor surface model to accommodate unanticipated variation in the test molecules. For example, in the closed receptor surface model, there is no possibility of accommodating the acetate group in compound **23** without strain. This causes a large underestimation of activity for this compound with the closed model. The open receptor surface model is able

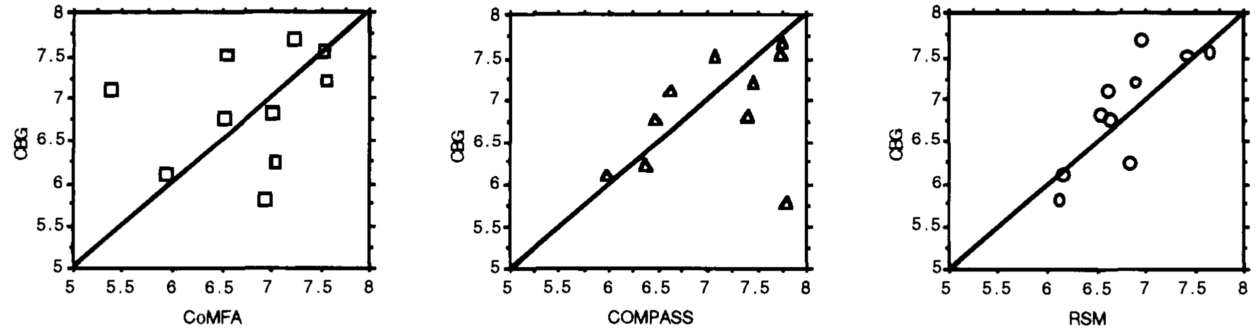


**Figure 5.** Scatterplots of CBG versus predicted activity for the 10 test compounds using CoMFA, Compass, and the open receptor surface QSAR model. The chemical knowledge represented in the open receptor surface model reduces the number of outliers compared with either CoMFA or Compass.

**Table 2.** Corticosteroid Data Descriptor for the Open Receptor Surface Model[a]

| no. | $E_{\text{interact}}$ | CBG | $\text{RSM}_{\text{open}}$ predicted | CoMFA predicted | Compass predictted | $\text{RSM}_{\text{open}}$ residual | CoMFA residual | Compass residual |
|---|---|---|---|---|---|---|---|---|
| 1 | −14.72 | 6.279 | 6.975 | | 6.012 | 0.696 | | −0.267 |
| **2** | −10.015 | 5.000 | 5.864 | | 5.156 | 0.864 | | 0.156 |
| **3** | −9.567 | 5.000 | 5.758 | | 5.021 | 0.758 | | 0.021 |
| 4 | −7.903 | 5.763 | 5.365 | | 6.836 | −0.397 | | 1.073 |
| **5** | −4.511 | 5.613 | 4.563 | | 5.118 | −1.049 | | −0.495 |
| **6*** | −16.097 | 7.881 | 7.301 | | 7.84 | −0.579 | | −0.041 |
| **7*** | −17.265 | 7.881 | 7.577 | | 7.691 | −0.303 | | −0.190 |
| **8** | −16.439 | 6.892 | 7.382 | | 7.771 | 0.490 | | 0.879 |
| **9** | −5.763 | 5.000 | 4.859 | | 4.995 | −0.140 | | −0.005 |
| 10* | −15.182 | 7.653 | 7.058 | | 7.682 | −0.567 | | 0.029 |
| 11* | −16.247 | 7.881 | 7.336 | | 7.614 | −0.544 | | −0.267 |
| **12** | −11.767 | 5.919 | 6.278 | | 6.107 | 0.359 | | 0.188 |
| **13** | −9.037 | 5.000 | 5.633 | | 4.989 | 0.633 | | −0.011 |
| **14** | −8.503 | 5.000 | 5.507 | | 4.851 | 0.507 | | −0.149 |
| **15** | −6.248 | 5.000 | 4.974 | | 4.912 | −0.025 | | −0.088 |
| **16** | −4.396 | 5.225 | 4.536 | | 5.377 | −0.688 | | 0.152 |
| **17** | −10.632 | 5.225 | 6.010 | | 5.525 | 0.785 | | 0.300 |
| **18** | −11.610 | 5.000 | 6.241 | | 5.215 | 1.241 | | −0.215 |
| 19* | −13.204 | 7.380 | 6.617 | | 7.473 | −0.762 | | 0.093 |
| 20* | −14.280 | 7.740 | 6.871 | | 7.248 | −0.868 | | −0.492 |
| 21 | −11.924 | 6.724 | 6.315 | | 6.955 | −0.408 | | 0.231 |
| **22** | −16.799 | 7.512 | 7.417 | 6.544 | 7.062 | −0.094 | −0.968 | −0.450 |
| **23** | −16.670 | 7.553 | 7.646 | 7.540 | 7.729 | 0.093 | −0.013 | 0.176 |
| **24** | −13.232 | 6.779 | 6.647 | 6.526 | 6.462 | −0.131 | −0.253 | −0.317 |
| **25** | −14.334 | 7.200 | 6.905 | 7.546 | 7.466 | −0.294 | 0.346 | 0.266 |
| **26** | −11.392 | 6.114 | 6.164 | 5.955 | 5.994 | 0.050 | −0.159 | −0.120 |
| **27** | −13.932 | 6.247 | 6.841 | 7.057 | 6.383 | 0.594 | 0.810 | 0.136 |
| **28** | −13.074 | 7.120 | 6.623 | 5.384 | 6.625 | −0.496 | −1.736 | −0.495 |
| **29** | −12.808 | 6.817 | 6.550 | 7.009 | 7.403 | −0.266 | 0.192 | 0.586 |
| **30** | −14.660 | 7.688 | 6.9475 | 7.227 | 7.741 | −0.740 | −0.461 | 0.053 |
| **31** | −11.387 | 5.797 | 6.116 | 6.937 | 7.779 | 0.319 | 1.14 | 1.982 |

[a] This table contains the compound index, the energy of interaction of the analogue in the receptor surface model, the predictions for the models from receptor surface modeling, CoMFA, and Compass, and the residuals from receptor surface modeling, CoMFA, and Compass. (CoMFA predicted and residuals for the training set not available.)

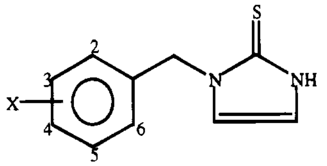


**Figure 6.** The shared structure of the dopamine $\beta$-hydroxylase inhibitors.

to accommodate the acetate group and gives a good estimation of activity.

## 4. Analysis of the $\beta$-Hydroxylase Inhibitor Data Set

The dopamine $\beta$-hydroxylase inhibitor data set is a set of 47 1-(substituted-benzyl)imidazole-2(3*H*)-thiones with associated inhibitory activities described in the work of Kruse et al.[24] These inhibitors effectively reduce blood pressure and are used for treatment of cardiovascular disorders related to hypertension. The series of analogs are of the general form shown in Figure 6. We were interested in studying this data set because the compounds had been studied using molecule shape analysis (MSA) by Burke and Hopfinger.[25] The original Kruse et al. study contained 52 compounds although only 25 were used for QSAR generation. The less active compounds had their activity reported as percent inhibition at fixed concentration, and the remaining compounds had their activity reported as $-\log(\text{IC}_{50})$ values. Burke and Hopfinger chose 47 of the 52 compounds for their study, estimating $-\log(\text{IC}_{50})$ for the samples reported with percent inhibition at fixed concentration. Five compounds were considered problematical due to differences in structure, flexibility, and charge from the remaining 47 compounds and were not used in their study, nor were their $-\log(\text{IC}_{50})$ estimates reported. While we realized such removal of samples can bias the resulting QSAR, we wished to compare our work with the results of Burke and Hopfinger, and so we used this same 47 compound subset in our QSAR study. The original compound numbers are retained from the Kruse et al. study.

- $V_0$: Common overlap steric volume against the most-active compound
- $\pi_0$: Molecular lipophilicity
- $\pi_4$: Water/octanol fragment constant of the 4-substituent
- $Q_6$: The partial atomic charge on atom 6
- $Q_{3,4,5}$: Sum of partial atomic charges on atoms 3, 4, and 5

**Figure 7.** The QSAR descriptors generated by Burke and Hopfinger.

Linear free energy descriptors were used by Kruse et al. to construct their QSARs. Burke and Hopfinger[25] constructed QSARs for this data set; they generated five descriptors for each of the compounds. The QSAR descriptors generated by Burke and Hopfinger are shown in Figure 7.

Burke and Hopfinger proposed two models. The first model contained six terms and a constant and used the complete set of 47 compounds. The second model contained three terms and a constant and used 45 of the compounds (compounds **23** and **39** were identified as outliers and removed). The models generated with the Burke and Hopfinger descriptors are shown in Figure 8. Most critical to the modeling was the descriptor $V_0$, a shape-based descriptor which reflected the common steric volume between the most active compound and a given test compound.

| | |
|---|---|
| $-\log(IC_{50}) = 52.27$ | |
| $- 116.9 * V_o$ | Common overlap steric volume against most-active compound |
| $+ 69.1 * V_o^2$ | Square of Vo |
| $+ 2.06 * Q_{3,4,5}$ | Sum of partial atomic charges on atoms 3, 4, and 5 |
| $- 4.68 * Q_6$ | The partial atomic charge on atom 6 |
| $+ 0.0465 * \pi_0^2$ | Molecular lipophilicity |
| $- 0.578 * \pi_4$ | Water/octanol fragment constant of the 4-substituent |

N: 47;
$r^2$: 0.828

| | |
|---|---|
| $-\log(IC_{50}) = 52.15$ | |
| $- 117.5 * V_o$ | Common overlap steric volume against most-active compound |
| $+ 70.4 * V_o^2$ | Square of Vo |
| $+ 2.32 * Q_{3,4,5}$ | Sum of partial atomic charges on atoms 3, 4, and 5 |

N: 45 (samples 23 and 39 removed)
$r^2$: 0.810

**Figure 8.** The two models constructed in the study of Burke and Hopfinger. The first model was constructed using all 47 compounds in the data set. The second model was constructed using the first model to identify two outliers in the data set, removing those outliers, and constructing a new model over the reduced data set.

**Table 3.** Dopamine $\beta$-Hydroxylase Inhibitor Data Set with Burke/Hopfinger Descriptors

| no. | $Q_{345}$ | $Q_6$ | $\pi_4$ | $\pi_0^2$ | $V_o$ | substituents | $-\log(IC_{50})$ | predicted | residual |
|---|---|---|---|---|---|---|---|---|---|
| **2** | −0.03 | 0.04 | 0.00 | 19.86 | 0.816 | 2,6-$Me_2$ | 3.00 | 3.51 | 0.51 |
| 4 | 0.06 | 0.07 | 0.00 | 21.01 | 0.842 | 2,6-$Cl_2$ | 3.15 | 3.55 | 0.40 |
| **6** | −0.03 | 0.19 | 0.00 | 10.02 | 0.748 | 2,6-$(OME)_2$ | 3.30 | 2.96 | −0.34 |
| **7** | 0.04 | 0.00 | 0.00 | 14.98 | 0.908 | 2-Cl | 3.45 | 3.81 | 0.36 |
| **8** | 0.05 | 0.02 | 0.00 | 14.49 | 0.896 | 2-Me | 3.47 | 3.63 | 0.16 |
| **9** | 0.28 | 0.01 | −0.02 | 6.79 | 0.824 | 3,4-$(OMe)_2$ | 3.47 | 3.67 | 0.20 |
| 10 | 0.07 | 0.01 | 0.88 | 16.33 | 0.894 | 4-$CF_3$ | 3.70 | 3.28 | −0.42 |
| 11 | 0.10 | 0.04 | −0.02 | 17.68 | 0.855 | 3-$CF_3$,4-OMe | 3.76 | 3.63 | −0.13 |
| **12** | 0.17 | 0.10 | −0.02 | 22.21 | 0.763 | 2,6-$Cl_2$,4-OMe | 3.81 | 4.18 | 0.37 |
| **13** | 0.07 | 0.02 | 0.00 | 14.49 | 0.944 | 4-Me | 3.83 | 4.15 | 0.32 |
| 14 | 0.16 | 0.01 | 0.86 | 16.17 | 0.917 | 4-Br | 3.94 | 3.65 | −0.29 |
| **15** | 0.30 | 0.00 | −0.02 | 14.44 | 0.876 | 3-Br,4-OMe | 4.08 | 4.13 | 0.05 |
| **16** | 0.31 | −0.01 | −0.02 | 11.29 | 0.897 | 3-F,4-OMe | 4.13 | 4.17 | 0.04 |
| **17** | 0.01 | 0.03 | 0.00 | 9.47 | 0.883 | 2-OMe | 4.13 | 3.19 | −0.94 |
| **18** | 0.16 | 0.02 | −0.02 | 13.88 | 0.885 | 3-Me,4-OMe | 4.16 | 3.77 | −0.39 |
| **19** | 0.01 | 0.03 | 0.00 | 6.21 | 0.947 | 2-OH | 3.24 | 3.64 | 0.40 |
| **20** | 0.19 | 0.01 | −0.02 | 5.21 | 0.883 | 3-$NO_2$,4-OMe | 3.45 | 3.46 | 0.01 |
| **21** | 0.10 | 0.01 | −0.02 | 7.45 | 0.898 | 4-OMe | 3.69 | 3.47 | −0.22 |
| **22** | 0.20 | −0.01 | 0.00 | 9.47 | 0.900 | 3-OMe | 3.80 | 3.87 | 0.07 |
| **23** | 0.20 | 0.00 | 0.00 | 6.21 | 0.986 | 3-OH | 3.83 | 4.82 | 0.99 |
| **24** | 0.11 | 0.04 | −0.67 | 15.16 | 0.920 | 3-$CF_3$,4-OH | 3.92 | 4.28 | 0.36 |
| **25** | 0.17 | 0.08 | 0.71 | 28.06 | 0.798 | 2,4,6-$Cl_3$ | 3.99 | 3.81 | −0.18 |
| **26** | 0.12 | 0.00 | 0.00 | 21.01 | 0.908 | 2,5-$Cl_2$ | 4.01 | 4.26 | 0.25 |
| **27** | 0.16 | 0.01 | 0.71 | 14.98 | 0.948 | 4-Cl | 4.02 | 4.05 | 0.03 |
| **28** | 0.17 | 0.10 | −0.67 | 19.03 | 0.827 | 2,6-$Cl_2$,4-OH | 4.12 | 3.96 | −0.16 |
| **29** | 0.42 | 0.19 | −0.67 | 9.36 | 0.951 | 2,3,5,6-$F_4$,4-OH | 4.21 | 4.32 | 0.11 |
| **30** | 0.13 | 0.02 | −0.28 | 8.59 | 0.942 | 4-$NO_2$ | 4.28 | 4.14 | −0.14 |
| 31 | 0.12 | 0.00 | 0.00 | 21.01 | 0.904 | 2,3-$Cl_2$ | 4.28 | 4.23 | −0.05 |
| **32** | 0.16 | 0.02 | −0.67 | 9.86 | 0.964 | 3-Me,4-OH | 4.31 | 4.81 | 0.50 |
| **33** | 0.17 | 0.03 | 0.14 | 10.90 | 0.989 | 4-F | 4.33 | 4.81 | 0.48 |
| **34** | 0.30 | 0.04 | −0.02 | 14.46 | 0.902 | 3,5-$Cl_2$,4-OMe | 4.33 | 4.10 | −0.23 |
| **35** | 0.50 | −0.06 | −0.02 | 12.77 | 0.900 | 3,5-$F_2$,4-OMe | 4.44 | 4.89 | 0.45 |
| **36** | 0.02 | 0.00 | 0.00 | 17.98 | 0.948 | H | 4.48 | 4.36 | −0.12 |
| **37** | 0.22 | 0.03 | −0.67 | 5.13 | 0.952 | 3-$NO_2$,4-OH | 4.51 | 4.48 | −0.03 |
| **38** | 0.19 | 0.00 | 0.71 | 21.01 | 0.951 | 3,4-$Cl_2$ | 4.55 | 4.48 | −0.07 |
| **39** | 0.14 | 0.00 | 0.71 | 36.77 | 0.874 | 2,4-$Cl_2$ | 4.77 | 4.41 | −0.36 |
| 40 | 0.02 | 0.00 | −0.67 | 11.36 | 0.959 | 3-Br,4-OH | 4.92 | 4.60 | −0.32 |
| 41 | 0.15 | 0.02 | 0.00 | 14.98 | 0.986 | 3-Cl | 4.92 | 5.03 | 0.11 |
| **42** | 0.25 | −0.01 | 0.00 | 10.9 | 0.991 | 3-F | 5.25 | 5.28 | 0.03 |
| **44** | 0.13 | 0.03 | −0.67 | 6.21 | 0.989 | 4-OH | 5.59 | 4.98 | −0.61 |
| **45** | 0.19 | 0.01 | 0.00 | 21.01 | 0.993 | 3,5-$Cl_2$ | 5.62 | 5.57 | −0.05 |
| **46** | 0.28 | 0.01 | −0.67 | 8.33 | 0.983 | 3,4-$(OH)_2$ | 5.66 | 5.36 | −0.30 |
| **48** | 0.19 | 0.02 | −0.67 | 10.37 | 0.999 | 3-Cl,4-OH | 5.70 | 5.54 | −0.16 |
| **49** | 0.30 | −0.01 | −0.67 | 7.03 | 1.000 | 3F,4-OH | 5.82 | 5.78 | −0.04 |
| **50** | 0.43 | −0.07 | 0.00 | 11.86 | 0.991 | 3,5-$F_2$ | 5.92 | 5.98 | 0.06 |
| **51** | 0.27 | −0.03 | −0.67 | 14.46 | 1.000 | 3,5-$Cl_2$,4-OH | 6.17 | 6.15 | −0.02 |
| **52** | 0.50 | −0.06 | −0.67 | 7.09 | 1.000 | 3,5-$F_2$,4-OH | 7.13 | 6.42 | −0.71 |

The descriptors generated by Burke and Hopfinger, the activities of the compounds, and the predictions using their first model are shown in Table 3.

The energetic descriptors calculated by a receptor surface model were tested for their ability to represent three-dimensional information critical to the estimation of activity, and hence their ability to serve as components of QSAR models in situations where three-dimensional effects must be considered.

A receptor surface model was constructed from a

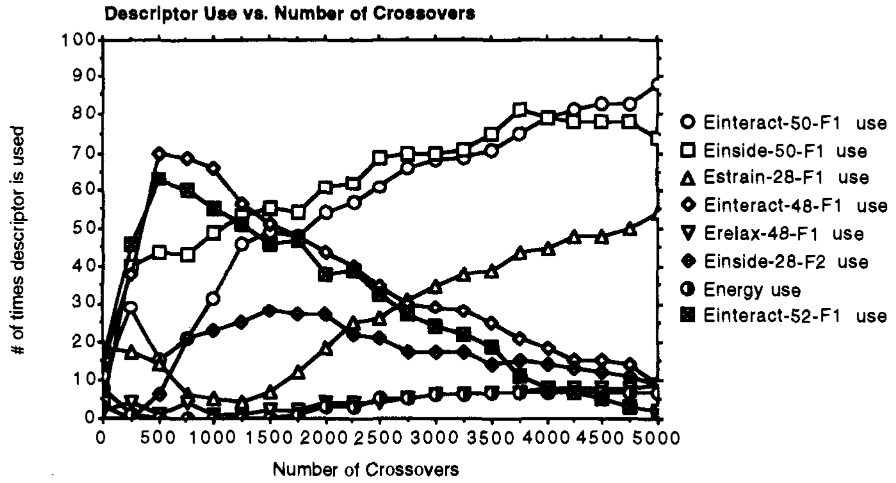


**Figure 9.** A graph of descriptor use versus the number of crossover operations for the seven most used descriptors in the multiple receptor surface data set. The descriptors are of the form Name-XX-FY, where XX is the starting compound for the series (XX-52), and *Y* is the tolerance (1 = 0.1 Å, 2 = 0.2 Å). It is clear from this graph that the descriptors generated from the receptor surface model of the three most active compounds at a tolerance of 0.1 Å is the preferred model.

subset of the most active compounds, their associated activities, and a steric "tolerance". (This tolerance is the distance from the van der Waals surface of the overlapped molecules that the receptor surface is constructed.) The question arises as to the appropriate number of compounds and the appropriate tolerance to use in constructing a model. In the previous experiment, we used intuition to select the set of actives to use in model construction. In this study, we used GFA to select among models constructed with various parameters.

Twelve receptor surface models were created that combined six activity ranges and two tolerances. The activity ranges contained either the most active compound (**52**), the three most active (**50**–**52**), the five most active (**48**–**52**), the eight most active (**44**–**52**), the 25 most active (**28**–**52**), or all the compounds (**2**–**52**). In each case, the compounds used to generate the models were first aligned to the most active compound (**52**). The conformation chosen for most active compound was the same as the conformation of the shape reference compound in the Burke and Hopfinger study. The tolerances were either 0.1 Å out from the combined van der Waals surface or 0.2 Å out.

After creation of the models, each compound in the data set was evaluated against each of the models to calculate the four receptor surface-based descriptors. The descriptors from the multiple receptor models were placed into one table for a total of 48 receptor surface descriptors.

The GFA algorithm was applied to this data set, allowing both linear and nonlinear terms. In this case, we were not interested in the specific QSAR models discovered but in the relative number of times each of the receptor surface based descriptors were used in the population. This gives an indication of which receptor surface model provides the highest quality descriptors for QSAR modeling. In effect, the receptor surface models were taking part in a competition to see which could provide the most useful descriptors for QSAR model building.

The frequency of use of the eight most used descriptors versus the number of crossovers is displayed in Figure 9.

$$
\begin{aligned}
-\log(\mathrm{IC}_{50})_{\text{P-QSAR}} = {} & 3.762 \\
& + 0.296 * \langle -10.203 - \mathbf{E_{interact}} \rangle \\
& + 0.089 * \langle 26.855 - \mathbf{E_{inside}} \rangle
\end{aligned}
$$

N: 47

| | |
|---|---|
| $r^2$: | 0.808 |
| Regression-only CV-$r^2$: | 0.788 |
| Fully CV-$r^2$: | 0.669 |

**Figure 10.** The number of data samples *N*, the correlation coefficient $r^2$, the regression-only cross-validated $r^2$, and the fully cross-validated $r^2$ for the top model derived using GFA with the Hopfinger descriptors augmented with the receptor surface model descriptors. Only the descriptors derived from the receptor surface model were chosen for use in the top-rated model. The $r^2$ scores show this model to be as good over the full 47-compound data set as the Burke and Hopfinger model was over the 45-compound reduced data set.

The most useful descriptors are derived from the receptor surface model constructed from the three most active compounds (i.e., **50**–**52**) at a tolerance of 0.1 Å. The graph shows them being rapidly discovered and used in nearly every model by the end of the evolution. From this evidence, we selected this receptor surface model constructed from the three most active compounds for a more formal and thorough analysis of the data set.

(An intriguing possibility suggested by the graph in Figure 9 is that different descriptors may be best derived from different receptor surface models. In this case, it appears that receptor surface models generated from relatively few active molecules give the best interaction energies, but models generated from many active molecules give the best strain energies. More study would be needed to determine whether this is a true effect or merely a statistical artifact.)

We analyzed the dopamine β-hydroxylase inhibitors with GFA using linear polynomials and linear splines. The data set contained four-receptor surface descriptors generated from the receptor surface model of the three most active compounds at a steric tolerance of 0.1 Å. These receptor surface descriptors were combined with the Burke and Hopfinger QSAR descriptors shown in Figure 7. The population of QSAR models was evolved for 5000 generations. The best QSAR model, as rated by the GFA's lack-of-fit score, is shown in Figure 10. The receptor surface descriptors and the predictions of this model are shown in Table 4. This model explains

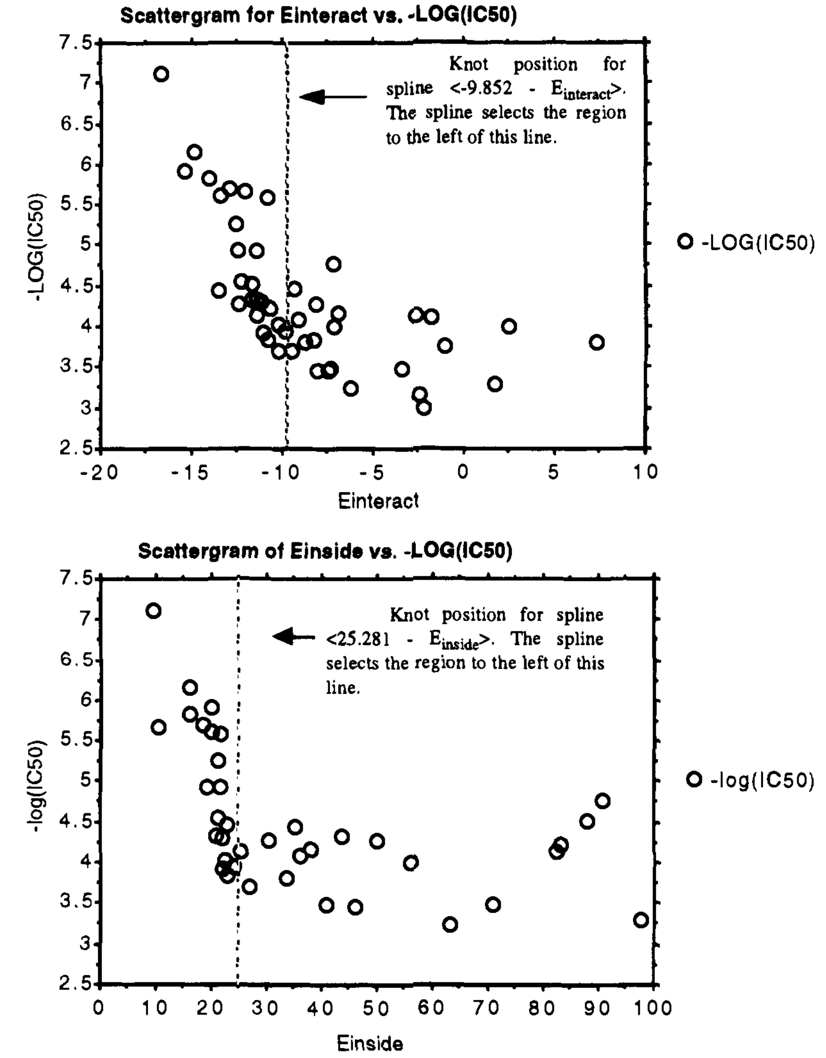
**Table 4.** Receptor Surface Descriptors and Prediction using Top Model

| no. | substituents | $E_{\text{inside}}$ | $E_{\text{interact}}$ | $E_{\text{relax}}$ | $E_{\text{strain}}$ | $-\log(IC_{50})$ | predicted | error |
|---|---|---|---|---|---|---|---|---|
| 2 | 2,6-$Me_2$ | 177.234 | −2.181 | 28.271 | 148.963 | 3.00 | 3.77 | 0.77 |
| 4 | 2,6-$Cl_2$ | 181.649 | −2.451 | 37.168 | 144.48 | 3.15 | 3.77 | 0.62 |
| 6 | 2,6-$(OME)_2$ | 97.672 | 1.744 | 49.283 | 48.388 | 3.30 | 3.77 | 0.47 |
| 7 | 2-Cl | 45.891 | −8.077 | 32.557 | 13.333 | 3.45 | 3.77 | 0.32 |
| 8 | 2-Me | 40.658 | −7.350 | 31.403 | 9.25 | 3.47 | 3.77 | 0.30 |
| 9 | 3,4-$(OMe)_2$ | 70.863 | −3.450 | 30.901 | 39.962 | 3.47 | 3.77 | 0.30 |
| 10 | 4-$CF_3$ | 112.597 | −10.272 | 22.610 | 89.986 | 3.70 | 3.88 | 0.18 |
| 11 | 3-$CF_3$,4-OMe | 197.764 | −1.150 | 40.604 | 157.160 | 3.76 | 3.77 | 0.01 |
| 12 | 2,6-$Cl_2$,4-OMe | 318.513 | 7.321 | 38.191 | 280.321 | 3.81 | 3.77 | −0.04 |
| 13 | 4-Me | 22.842 | −8.290 | 22.446 | 0.395 | 3.83 | 4.02 | 0.19 |
| 14 | 4-Br | 24.111 | −9.851 | 21.916 | 2.194 | 3.94 | 3.89 | −0.05 |
| 15 | 3-Br,4-OMe | 35.949 | −9.144 | 20.404 | 15.544 | 4.08 | 3.77 | −0.31 |
| 16 | 3-F,4-OMe | 25.280 | −11.496 | 18.706 | 6.573 | 4.13 | 4.22 | 0.09 |
| 17 | 2-OMe | 82.315 | −2.707 | 31.395 | 50.920 | 4.13 | 3.77 | −0.36 |
| 18 | 3-Me,4-OMe | 38.115 | −6.980 | 22.072 | 16.043 | 4.16 | 3.77 | −0.39 |
| 19 | 2-OH | 63.337 | −6.185 | 36.716 | 26.62 | 3.24 | 3.77 | 0.53 |
| 20 | 3-$NO_2$,4-OMe | 193.627 | −7.456 | 55.189 | 138.437 | 3.45 | 3.77 | 0.32 |
| 21 | 4-OMe | 26.854 | −9.535 | 22.226 | 4.628 | 3.69 | 3.77 | 0.08 |
| 22 | 3-OMe | 33.713 | −8.829 | 25.553 | 8.160 | 3.80 | 3.77 | −0.03 |
| 23 | 3-OH | 22.610 | −10.866 | 22.610 | −0.000227 | 3.83 | 4.32 | 0.49 |
| 24 | 3-$CF_3$,4-OH | 21.838 | −11.141 | 21.843 | −0.00498 | 3.92 | 4.47 | 0.55 |
| 25 | 2,4,6-$Cl_3$ | 284.909 | 2.430 | 34.786 | 250.122 | 3.99 | 3.77 | −0.22 |
| 26 | 2,5-$Cl_2$ | 56.056 | −7.186 | 32.140 | 23.906 | 4.01 | 3.77 | −0.24 |
| 27 | 4-Cl | 22.502 | −10.203 | 21.765 | 0.736 | 4.02 | 4.15 | 0.13 |
| 28 | 2,6-$Cl_2$,4-OH | 258.102 | −1.846 | 39.274 | 218.827 | 4.12 | 3.77 | −0.35 |
| 29 | 2,3,5,6-$F_4$,4-OH | 83.035 | −10.763 | 26.832 | 56.203 | 4.21 | 4.02 | −0.19 |
| 30 | 4-$NO_2$ | 30.354 | −12.382 | 24.717 | 5.636 | 4.28 | 4.46 | 0.18 |
| 31 | 2,3-$Cl_2$ | 49.832 | −8.134 | 32.649 | 17.183 | 4.28 | 3.77 | −0.51 |
| 32 | 3-Me,4-OH | 21.928 | −11.159 | 21.931 | −0.002 | 4.31 | 4.47 | 0.16 |
| 33 | 4-F | 20.979 | −11.402 | 20.984 | −0.00531 | 4.33 | 4.63 | 0.30 |
| 34 | 3,5-$Cl_2$,4-OMe | 43.727 | −11.667 | 30.936 | 12.790 | 4.33 | 4.27 | −0.06 |
| 35 | 3,5-$F_2$,4-OMe | 35.077 | −13.486 | 24.495 | 10.581 | 4.44 | 4.77 | 0.33 |
| 36 | H | 22.737 | −9.333 | 22.516 | 0.220 | 4.48 | 4.03 | −0.45 |
| 37 | 3-$NO_2$,4-OH | 87.982 | −11.663 | 36.421 | 51.56 | 4.51 | 4.27 | −0.24 |
| 38 | 3,4-$Cl_2$ | 21.288 | −12.367 | 20.710 | 0.577 | 4.55 | 4.87 | 0.32 |
| 39 | 2,4-$Cl_2$ | 90.716 | −7.190 | 32.866 | 57.849 | 4.77 | 3.77 | −1.00 |
| 40 | 3-Br,4-OH | 19.326 | −12.484 | 19.316 | 0.00976 | 4.92 | 5.10 | 0.18 |
| 41 | 3-Cl | 21.492 | −11.490 | 21.498 | −0.00567 | 4.92 | 4.60 | −0.32 |
| 42 | 3-F | 21.165 | −12.607 | 21.165 | 0.000252 | 5.25 | 4.95 | −0.30 |
| 44 | 4-OH | 21.572 | −10.850 | 21.449 | 0.122 | 5.59 | 4.42 | −1.17 |
| 45 | 3,5-$Cl_2$ | 20.051 | −13.435 | 20.055 | −0.00419 | 5.62 | 5.29 | −0.33 |
| 46 | 3,4-$(OH)_2$ | 10.527 | −12.037 | 10.529 | −0.00256 | 5.66 | 5.87 | 0.21 |
| 48 | 3-Cl,4-OH | 18.586 | −12.933 | 18.589 | −0.00262 | 5.70 | 5.30 | −0.48 |
| 49 | 3F,4-OH | 16.042 | −13.989 | 16.041 | 0.000872 | 5.82 | 5.85 | 0.03 |
| 50 | 3,5-$F_2$ | 20.013 | −15.402 | 20.014 | −0.00133 | 5.92 | 5.83 | −0.09 |
| 51 | 3,5-$Cl_2$,4-OH | 16.051 | −14.886 | 16.056 | −0.00523 | 6.17 | 6.10 | −0.07 |
| 52 | 3,5-$F_2$,4-OH | 9.671 | −16.736 | 9.674 | −0.00282 | 7.13 | 7.26 | 0.13 |

the entire training set (no outliers removed) nearly as well as either of the models of Burke and Hopfinger and contains fewer descriptors. (If we count each spline-based term as the equivalent of two linear terms, this four-term model performs as well as the six-term Burke and Hopfinger model.) None of the Burke and Hopfinger descriptors are chosen in the best-rated models, including their molecular shape descriptor $V_o$. This suggests the information provided by $V_o$ is replacable by receptor surface model descriptors. This suggestion is confirmed by calculating the correlation coefficient $r^2$ between $V_o$ and $E_{\text{interact}}$. They are highly correlated, with an $r^2$ of 0.78. That $E_{\text{interact}}$ rather than $V_o$ is chosen by the GFA procedure implies that $E_{\text{interact}}$ contains additional useful information in its variance noncorrelated with $V_o$. This demonstrates the ability of receptor surface model descriptors to provide a compact representation of three-dimensional information for QSAR modeling of binding data.

Perhaps surprisingly, even though GFA is parameterized to favor linear terms over spline terms, the best model contained only spline terms. Splines are preferred if the data contains nonlinear relationships. This is understandable if we look at a scatterplot of $E_{\text{interact}}$ and $E_{\text{inside}}$ versus $-\log(IC_{50})$. These scatterplots are shown in Figure 11. The dotted lines show the location of the spline know, which is where the spline separates the linear region from the nonlinear region. The spline term in each case has discovered a linear relationship between the points to the left of the knot and the activity. Samples to the right of the knot cause the spline term to return 0.0. The QSAR model uses spline terms to expose, for the most active compounds, the linear relationship between the descriptors and activity. Without splines, these relationships could have been obscured or missed entirely.

There are other techniques available for nonlinear studies; see, for example, the review of Sekulic et al.[26] One possibility is the direct introduction of nonlinearity into the PLS process ("nonlinear-PLS").[27] Minimally, the user can visually inspect the scatterplots of the descriptors versus activity and create new descriptors (such as quadratic terms) from the nonlinear descriptors. Notwithstanding, there are advantages to our approach: the nonlinearities are discovered automatically, which may be difficult to do visually as the number of descriptors increases; the spline terms are easy to interpret, unlike quadratic terms or PLS latent


Scattergram for Einteract vs. -LOG(IC50)

Knot position for spline <-9.852 - $E_{interact}$>. The spline selects the region to the left of this line.

Scattergram of Einside vs. -LOG(IC50)

Knot position for spline <25.281 - $E_{inside}$>. The spline selects the region to the left of this line.

**Figure 11.** Scatterplots of the descriptors $E_{\text{interact}}$ and $E_{\text{inside}}$ versus $-\log(\text{IC}_{50})$. The dotted lines show the location of the knot; the spline term in each case has discovered a linear relationship between the points to the left of the knot and the activity. Samples to the right of the knot cause the spline term to return 0.0. In this case, the model for activity uses spline terms that reflect relationships present for only the most active compounds.

variables; and the technique allows but does not require the construction of nonlinear models, since the spline terms are eliminated if they do not significantly increase the performance of the model.

The receptor surface descriptors and the predictions of the best model are shown in Table 4.

The above results confirm that the receptor surface models allow us to model the dopamine $\beta$-hydroxylase inhibitor training compounds better, but in most cases what we are interested in is predictiveness: how well can we estimate the activity of compounds outside the training set? This is commonly estimated using cross-validation and, preferably, randomization testing.

Cross-validation is done by dividing the training set into some number of groups, called cross-validation groups. Each group is left out in turn, and the remaining groups are used to build a model of activity. The samples in the left-out group are then predicted using this model. At the end of the process, all samples have been predicted. Commonly, each cross-validation group may contain only one sample; this is leave-one-out cross-validation. In any case, the final estimate is usually expressed as a *cross-validated* $r^2$.

For the dopamine $\beta$-hydroxylase data set, we divided the samples into 16 cross-validation groups. Leaving out each group in turn, we used PLS on the data set containing the 10 two-dimensional QSAR descriptors and the four three-dimensional receptor surface model based descriptors. As shown in Figure 10, the nonlinear GFA analysis of the data gave a cross-validated $r^2$ of 0.669; the PLS analysis performed much worse, with a cross-validated $r^2$ of 0.471.

The relatively poor result from PLS is due to the nonlinearities in the data variables. A linear technique such as PLS cannot discover these values automatically; GFA was able to discover them and so create superior models and a superior cross-validation score.

Another technique for validating a model-generation process is called a randomization test.[28] A *randomization test* is a validation test that answers the question: what is the probability that the model construction process could have found a model that scored this well by random chance?

The test is conducted as follows. After using some model-building process on a data set, the $Y$ variable (usually activity) is scrambled and the model-construction process repeated. This new model, generated against the data set with randomized activities, is scored against the same randomized data set. Repeating this process multiple times gives statistical confidence limits

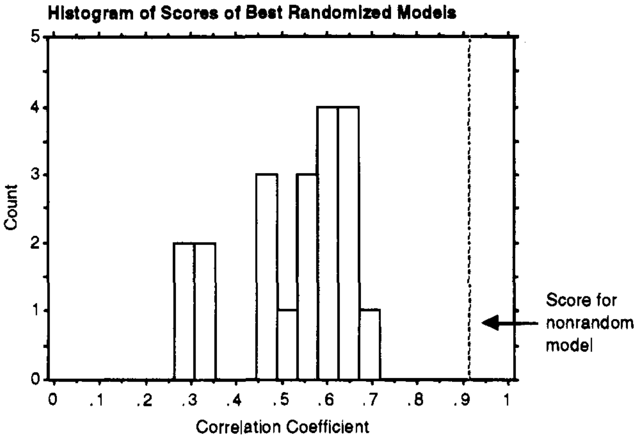


**Figure 12.** Histogram of the results from 20 randomization tests. The score for the nonrandom model is shown as a dotted line. The results show that the QSAR model for the nonrandomized data is predictive at >95% confidence level. The mean for the randomized experiments is 0.525; the standard deviation is 0.132.

about the quality of the model generated by the process against the actual (i.e., unscrambled) data. For example, if the original model has a better score than nine randomized models, you can state that the original model is predictive at the 90% confidence level. If it scores better after 19 tests, it is predictive at the 95% confidence level. If is scores better after 99 tests, it is predictive at the 99% confidence level.

A randomization test was performed on the dopamine $\beta$-hydroxylase inhibitor data set by creating 20 separate tables with randomized activity data. For each table, a receptor model was created from the three "most active" compounds in that table. GFA was run against each table, and the correlation coefficient of the top model in the population was recorded. A histogram of the results is shown in Figure 12.

The randomization test confirms that the receptor surface model is indeed predictive and is not the fortuitous result of random chance. This test was especially important for this experiment, as the receptor surface model was constructed using the three most active compounds; a leave-one-out cross-validation procedure will always leave two of the three most active molecules available for receptor surface building and so leaves some doubt as to the true lack of bias in the procedure. However, the randomization testing, combined with the cross-validation tests, strengthen the case that the receptor surface modeling process using GFA analysis is indeed discovering predictive QSAR models.

## 5. Experimental Section

All experiments were conducted on a Silicon Graphics Indigo/R4000, running under the IRIX 4.0.5 operating system. The receptor surface models described in this paper can each be generated in less than a minute. The evaluation of each compound with a surface model to generate energy descriptors requires only a few seconds per structure.

## 6. Conclusions

A new technique is proposed for using receptor surface models in QSAR analysis. This approach is effective for the analysis of data sets where activity information is available but the structure of the receptor site is unknown. An important aspect of receptor surface models is their ability to visualize putative receptor/ligand interactions in a qualitative and intuitive manner; this can help guide the chemist in the construction and refinement of better receptor surface models, leading to better quality descriptors and more predictive QSAR models.

Receptor surface models provide compact, quantitative descriptors which capture three-dimensional information about a putative receptor site. These descriptors may be used alone or in combination with more traditional 2D descriptors. Such combined QSAR models may better reflect the combination of mechanisms (transport, binding, absorption, etc.) responsible for drug activity.

A receptor surface model allows higher level chemical knowledge to be utilized during both model generation and model evaluation. An example of this is the ability of the receptor surface model utilize knowledge of hydrophobic interactions to better predict the activity of a fluorinated steroid which has been difficult to predict correctly with other methods.

Receptor surface models and their descriptors are generated quickly. Numerous alternate receptor surface models can be constructed with varying combinations of active structures, surface fit tolerances, and alignments. A variable selection technique like GFA can be used to suggest which receptor surface model or models are likely most informative. GFA also facilitates the discovery of nonlinear relationships by allowing spline models; this makes explicit the location of the discontinuity in the relationship between energy-derived terms and activity. Such relationships are not easily discovered using linear modeling tools such as PLS.

Receptor surface models could be applied against more flexible data sets, although in this case the selection of an appropriate conformation and alignment for the training compounds is likely to be critical for quality results.

Our application of receptor surface models against previously described data sets indicates that the approach can model the data as effectively as established techniques. This functionality is available as part of Molecular Simulations Incorporated's Cerius$^2$ modeling environment.[29]

**Acknowledgment.** The authors would like to thank Dr. A. J. Hopfinger for his advice and assistance relative to the dopamine $\beta$-hydroxylase inhibitor data set and Scott Kahn and Peter Sprague for their careful (and long-anticipated) reviews of the manuscript drafts.

## References

(1) Hansch, C.; Fujita, T. $\varrho$-$\sigma$-$\pi$ Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616.

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

(3) Burke, B.; Hopfinger, A. J. Molecular Shape Analysis: A Formalism to Quantitatively Establish Spatial Molecular Similarity. In *Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley and Sons: New York, 1990; pp 11–73.

(4) Wold, S.; Wold, H.; Ruhe, A.; Dunn, W. J., III. The Colinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM J. Sci. Stat. Comp.* **1984**, *5*, 735–743.

(5) Glen, W. G.; Dunn, W. J., III; Scott, D. R. Principal Components Analysis and Partial Least Squares Regression. *Tetrahedron Comput. Methodol.* **1989**, *2*, 279–356.

(6) Rogers, D.; 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) Holland, J. *Adaptation in Artificial and Natural Systems*; University of Michigan Press: Ann Arbor, MI, 1975.

(8) Snyder, J. P.; Rao, S. N.; Koehler, K. F.; Vedani, A. Minireceptors and Pseudoreceptors. In *3D QSAR in Drug Design: Theory Methods and Applications*; Kubinyi, H., Ed.; Escom: Leiden, 1993; pp 336–354.

(9) Hahn, M. Receptor Surface Models. 1. Definition and Construction. *J. Med. Chem.* **1995**, *38*, 2080–2090.

(10) Kato, Y.; Itai, A.; Iitaka, Y. A novel method for superimposing molecules and receptor mapping. *Tetrahedron* **1987**, *43*, 5229–5236.

(11) Doweyko, A. M. The Hypothetical active site lattice. An approach to modelling active sites from data on inhibitor molecules. *J. Med. Chem.* **1988**, *31*, 1396–1406.

(12) Hopfinger, A. J.; Nakata, Y.; Max, N. Quantitative structure-activity relationship of anthracycline antitumor activity and cardiac toxicity based upon intercalation calculations. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, 1981; p 431.

(13) Hopfinger, A. J.; Kawakami, Y. QSAR analysis of a set of benzothiopyranoindazole anti-cancer analogs based on their DNA intercalation properties as determined by molecular dynamics simulation. *Anti-Cancer Drug Design* **1992**, 7, 203–217.

(14) Marshall, G. R. Binding Site Modeling of Unknown Receptors. In *3D QSAR in Drug Design: Theory Methods and Applications*; Kubinyi, H., Ed.; Escom: Leiden, 1993; pp 80–116.

(15) Walters, D. E.; Hinds, R. M. Genetically Evolved Receptor Models: A Computational Approach to Construction of Receptor Models. *J. Med. Chem.* **1994**, *37*, 2527–2535.

(16) Kearsely, 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.

(17) Dammkoehler, R. A.; Karasak, S. F.; Berkely Shands, E. F.; Marshall, G. R. Constrained search of conformational hyperspace. *J. Comput.-Aided Mol. Design* **1989**, *3*, 3–21.

(18) Perkins, T. D.; Dean, P. M. An exploration of a novel strategy for superimposing several flexible molecules. *J. Comput.-Aided Mol. Design* **1993**, 7, 155–172.

(19) Blaney, J. M.; Dixon, J. S. A good ligand is hard to find: Automatic docking methods. *Perspect. Drug Disc. Design* **1993**, *1*, 301–319.

(20) Martin, Y. C.; Bures, M. G.; Danahar, E. A.; DeLazzar, J.; Lico, I.; Pavlik, P. A. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput.-Aided Mol. Design* **1993**, 7, 83.

(21) Hoffmann, R.; Langer, T. Use of the CATALYST program as a new alignment tool for 3D QSAR. In *Proceedings of the 10th European Symposium on Structure-Activity Relationships: QSAR and Molecular Modeling*; Prous Science Publishers: Barcelona, Spain, 1994.

(22) Barnum, D.; Greene, J.; Smellie, A. Identification of Common Functional Configurations. *J. Chem. Inf. Comput. Sci.* In press.

(23) Jain, A.; Koile, K.; Chapman, D. Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark. *J. Med. Chem.* **1994**, *37*, 2315–2327.

(24) Kruse, L. I.; Kaiser, C.; DeWolf, W. E., Jr.; Frazee, J. S.; Ross, S. T.; Wawro, J.; Wise, M.; Flaim, K. E.; Sawyer, J. L.; Erickson, R. W.; Ezekiel, M.; Ohlstein, E. H.; Berkowitz, B. A. *J. Med. Chem.* **1987**, *30*, 486.

(25) Burke, B. J.; Hopfinger, A. J. 1-(Substituted-benzyl)imidazole-2(3H)-thione Inhibitors of Dopamine beta-Hydroxylase. *J. Med. Chem.* **1990**, *33*, 274–281.

(26) Sekulic, S.; Seasholtz, M. B.; Wang, Z.; Kowalski, B. R.; Lee, S. E.; Holt, B. R. *Anal. Chem.* **1993**, *65*, 835.

(27) Berglund, A.; Ränner, S.; Wold, S. Nonlinear QSAR Problems. Possible Preprocessings and Nonlinear PLS Algorithms. In *Proceedings of the 10th European Symposium on Structure-Activity Relationships: QSAR and Molecular Modeling*; Prous Science Publishers: Barcelona, Spain, 1994.

(28) Fisher, R. The Principles of Experimentation, Illustrated by a Psycho-Physical Experiment. In *The Design of Experiments*, 8th ed.; Hafner Publishing: New York, 1966.

(29) Available from Molecular Simulations Incorporated, 16 New England Executive Park, Burlington, MA 94107.

JM940815G