Validation of EGSITE2, a Mixed Integer Program for Deducing Objective Site Models from Experimental Binding Data

Gordon M. Crippen

College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109-1065

Received March 31, 1997[⊗]

EGSITE2 represents a substantial advance in a long series of methods for calculating receptor site models given only specific binding data. Compared to our most recently reported technique, EGSITE [Schnitker et al. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 93−110] the user no longer has to simplify the structures of the molecules in the training set by clustering the atoms into a few superatoms. The only remaining source of subjectivity is the user's choice of compounds for the training set, which can be surprisingly few in number. Then EGSITE2 automatically produces typically several different models that explain the observed binding without outliers. The models are remarkably simple but have substantial predictive power for any sort of test compound, with an estimation of the uncertainty of the prediction. Validation of the method is reported for four standard test cases: triazines and pyrimidines binding to dihydrofolate reductase, steroids binding to corticosteroid-binding globulin and to testosterone-binding globulin, and peptides binding to angiotensin-converting enzyme.

Introduction

The problem being addressed in this work is a commonly occurring one in drug discovery: given only the experimentally determined binding affinities of a few compounds for one receptor site, construct a model of the site that explains the data and can be used for predictions. Of course, every QSAR method addresses this problem in one way or another, but the general approach we have been pursuing[1,2] differs in fundamental ways from most others.

Particularly when one wants to account for the binding of chemically diverse compounds, it is no longer permissible to construct different structure−activity relations for each set of close homologs involved. Excluding outliers and attributing the remaining unexplained variance to random experimental errors becomes untenable. This line of reasoning naturally leads away from the traditional reliance on linear regression and toward instead fitting the activity of each compound in the training set to within some predetermined limits, corresponding to the error bars on the experimental values. Consequently, the results can depend strongly on the presence of certain key compounds in the training set, instead of giving nearly the same result when any one compound is excluded, as in standard cross-validation.

The other main philosophical cornerstone of our approach is the treatment of molecular superposition and similarity. Let the binding *mode* denote a particular positioning of a ligand in the receptor site, including overall translation, rotation, and choice of internal conformation. The real ligands adopt the mode of lowest free energy, given the particular environment created by the receptor. In contrast, many QSAR methods assume homologous compounds will bind so as to place their common chemical groups in the same mode, or perhaps it is left to the user to guess the binding mode of each. We, on the other hand, view the mode of each compound in the training and test sets to be an outcome of the calculation, corresponding to the most favorable way the ligand can be fitted into the model site. The resulting superposition of active ligands depends on the site model derived from the given binding affinities. The same compounds interacting with a different receptor would produce a different superposition. Hence we view molecular similarity in an absolute sense to be an ill-defined concept; it only makes sense to talk about molecular superposition with respect to a real or hypothetical receptor site.[3]

In an effort to avoid overinterpreting the data, we have tried to produce the simplest site model(s) required to explain the experimental results. Even then, there is seldom a unique model at that level of simplicity, so we search for whole sets of models. In general, each model gives somewhat different predicted binding modes and affinities, so each test compound gets a *range* of calculated activities. In other words, not only are there error bars on the experimental data coming into this calculation, but there are error bars on the resulting predictions.

The last general principle is that there should be as little subjective input from the user as possible. Not only is he relieved of having to suggest alignments of the training molecules, but also no pharmacophore hypothesis is required. In our most recently described method,[1] EGSITE, the user still had to simplify the molecular structures by grouping atoms together into a few superatoms in order to make the search for binding modes computationally feasible. Now in EGSITE2, this requirement has been removed. Every molecule is represented by all atoms, hydrogens included. Of course there will always be subjective inputs from the user, such as which ligands to choose for the training set, and chosen values of the various adjustable parameters that are inherent in any such computer program.

While EGSITE2 is the latest in a series of increasingly powerful and increasingly objective methods built around these general design principles, the algorithm is substantially different from anything that has been described previously. In the Methods section, this is

[⊗] Abstract published in *Advance ACS Abstracts,* September 1, 1997.

S0022-2623(97)00211-2 CCC: $14.00

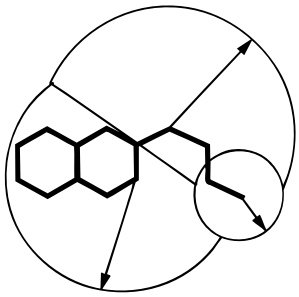

Figure 1. A schematic illustration of partitioning a molecule (heavy lines) into subsets corresponding to three regions having radii indicated by the arrows. The atoms at the centers of these subsets and regions are located at the tails of the respective arrows.

presented in generally readable language, leaving some of the technicalities to the Appendix. Key definitions of terms highlighted in italic type are found in the Methods section. Validation of the method is reported in the Results section for four standard test cases: triazines and pyrimidines binding to dihydrofolate reductase, steroids binding to corticosteroid-binding globulin and to testosterone-binding globulin, and peptides binding to angiotensin-converting enzyme.

Methods

Training Set Preparation. For each compound chosen to be in the *active training set*, there must be given the experimentally determined *binding affinity* expressed on some quasi-free-energy scale, such as $-\log K_i$, where zero corresponds to being free in solution, and greater, positive values correspond to stronger, more favorable binding. Each ligand's binding is given as a range, taken either from the given error bars, or arbitrarily set around a single value as a way of expressing the required accuracy of fit for the model. Each molecule is subjected to a *conformational search* under some standard molecular mechanics force field, such that several fairly low-energy conformers are located that differ substantially from one another. Since these calculations are done *in vacuo* as a way of presenting the subsequent algorithm with a variety of shapes these molecules might assume when bound to the receptor, the details of how the search is carried out are probably unimportant, as long as the shapes of the final receptor models are vague. For the set of steroids, MM2 within MSI's Cerius[2] software[4] was used, whereas the Dreiding force field[5] was used for the DHFR and ACE inhibitors. Independent of conformation, each atom is assigned three *physicochemical parameters*: an atomic contribution to log *P* and to the molar refractivity,[6] and a Gasteiger partial charge.[7]

Site Definition. A site model consists of a small number of *regions*, the first one always representing the solvent and the others corresponding to different pockets in the receptor where parts of the ligands will experience some sort of environment that is taken to be constant throughout the region. Each region has associated with it three adjustable *interaction parameters* corresponding to the fixed hydrophobicity, polarity, and charge properties of the atoms. This vector of parameters is always zero for the solvent region. When an atom lies inside one of the other regions, it makes an additive contribution to the total calculated binding affinity equal to the dot product of the atomic property vector with the interaction parameter vector. The geometry of the site is described broadly in terms of the diameter of each region and the greatest and least distances between a point in one region and a point in another. The underlying picture is that each region has a center in space, and it encompasses all points that are within the region's radius from its center, unless they are within a second region's sphere and closer to the second one's center (Figure 1). Note that if the region diameters were all infinite, this would correspond to the Voronoi polyhedra we had used in our previous work. The solvent region's diameter is always taken to be so large that every ligand in any conformation could fit completely inside it.

Binding Modes. Let the binding *mode* of a ligand molecule to a site model denote which of the several conformations is chosen and which atoms are assigned to lie in which regions. Every atom must lie in one and only one region, since the site regions are intended to describe all accessible space and they are supposed to be nonoverlapping. Furthermore, the assignments are far from arbitrary but rather conform to the definition of the site geometry. The subset of atoms assigned to a particular region has one of its atoms designated as the center, which should coincide in space with the corresponding region's center. Then the other atoms of the subset are all those lying within the region's radius but not closer to another subset's center and within that one's radius (Figure 1). Any molecule always has available the trivial binding mode where all atoms are in the solvent region, regardless of conformation or center atom. It is possible that the same choices of center atoms could lead to different partitionings of the remaining atoms among the subsets, depending on the chosen conformation. It is also possible that some atom falls into no subset, in which case the mode is said to have an *error*. A mode may not be in error yet still not be geometrically compatible with another site model, for example, because the diameter of a subset of atoms exceeds the diameter of the region it is assigned to. Then we say the mode is *disallowed* with respect to that site. Given the current site interaction parameters, a mode may be *suboptimal*, if its calculated binding affinity is less than the weakest experimental value, or *superoptimal*, if stronger than the strongest limit value, or otherwise *in-range*.

Basis Set of Modes. Clearly it is not practical to consider all possible modes for a molecule having 50 atoms and a site having as few as four regions. We go on the working assumption that the number of modes for each molecule in the training set that are actually essential toward defining a satisfactory site is vastly smaller than the total. As a heuristic, the algorithms starts by building typically 10 modes for each molecule, discovered in a random search that selects those 10 that are maximally different from one another in terms of the aggregate physicochemical properties of the subsets and the subset diameters and relative positions. At this point, the only thing known about the site is the number of regions chosen by the user, where one initially tries as few as two regions and gradually increases the number until one or more site models can be found. Hence the partitioning of the atoms in these initial modes is done assuming the least geometrically restrictive site geometry, namely, all regions are very large in diameter and all touch each other. Later in the calculation, this initial basis of modes is gradually augmented by including superoptimal modes that are discovered, on the assumption that these are important constraints on the adjustable features of the site: the region diameters, the upper and lower bounds on the interregion distances, and the three interaction parameters associated with each region. In what follows, the basis is generally kept to fewer than 150 modes in total for all training molecules, because the computer time tends to rise rapidly with basis size.

Initial Search for Site Geometries. Suppose for a moment that the initial mode basis covers all possible modes for the training compounds. Then the site geometry and energetic parameters must be adjusted in such a way that the following statement is true: "Every molecule must have at least one in-range mode and no superoptimal ones. Every mode is either geometrically allowed or disallowed. If it is disallowed, its calculated binding affinity is of no consequence, but there must be one or more geometric reasons for being incompatible with the site. In contrast, an allowed mode must not be superoptimal and must be compatible with the site geometry in every way." As is explained in the Appendix, this statement is translated into a set of linear inequalities involving the continuously variable geometric and energetic parameters of the site and a number of Boolean variables having values of either 0 for false or 1 for true. If any solutions can be found for the inequalities, one seeks the one involving the least restrictive site geometry. Optimizing such a linear function subject to linear constraints involving both continuous

and discrete variables is called *mixed integer programming*, which we solve using the standard commercial software Cplex.[8]

If no solution can be found in a tolerable amount of time, the user is forced to try more regions or more basis modes or a different training set. If on the other hand, a solution is found, it corresponds to a particular site geometry and energetics, and also to a particular combination of in-range modes selected from the basis. Alternative site models can be found by introducing an additional linear constraint that excludes this combination of in-range modes. Sometimes the next site that is found in this way has exactly the same geometry as one seen before, so the algorithm simply adds yet another exclusion constraint and continues on until either a maximal number of different site geometries have been found (typically 10) or there are no further solutions to the mixed integer program. In other words, this phase of the calculation produces a set of candidate site geometries that satisfy the very restricted initial mode basis set.

Adjustment of Interaction Energies. For each candidate site in turn, start with the initial mode basis and seek the strongest binding mode of each molecule. All the modes in the basis are either disallowed or not superoptimal, but a dedicated search given the current site generally locates a new mode for most molecules that is superoptimal. These are added to the basis, and now a restricted mixed integer program is solved for the adjustable interaction parameters, keeping the geometry of the site fixed. The linear inequalities to be solved correspond to the following statement: "Considering only the allowed modes in the basis, there must be at least one in-range mode for each molecule and no superoptimal ones." If the program is unable to find a solution in a tolerable amount of time (about 1 h of CPU time on an SGI R5000 workstation), the candidate site geometry is discarded, and the next one is tried. More often, a solution is found, but now the revised interaction parameters permit new superoptimal modes (at most one per molecule in the training set), these are added to the basis, and the new restricted integer program is tried. Eventually, the candidate site is either discarded or no new superoptimal modes can be found, in which case the resulting site is kept as one of the final solutions to the training set.

The reason it was necessary to split up the determination of candidate site geometries and subsequent refinement of their interaction parameters is a matter of practicality. Adjusting geometry and energetics simultaneously is a big problem made rapidly worse by additional modes in the basis. Adjusting the interaction parameters is a much smaller problem, but repeated several times as superoptimal modes are successively added. Once again, as the basis expands, the mixed integer program has many more combinations of in-range modes to sort through in search of a solution. By the time a solution site is found, the basis may have increased from 40 modes in the initial basis to over 100. Since many of the additional modes may have little bearing on any other site, they are removed before going on to the next candidate site.

Site Assessment and Predictions. Because the training sets must consist of only a few compounds, lest even the initial basis become unwieldy, the resulting solution sites may completely satisfy the training set but will not necessarily have strong predictive power unless the test compounds present no novel challenges to the site models. One approach is to solve a series of problems, starting with a very small training set and gradually adding those test compounds that are particularly badly predicted. Another approach built in to the current program is to have a so-called *passive training set*, as opposed to the active set referred to in all the steps above. The role of the passive set is to simply calculate the error of every solution site, defined as the sum of the amount of under- or overprediction for all the incorrectly predicted passive compounds. In this way, one can choose those solutions that have exceptional predictive promise and discard those that have explained the binding of the active training compounds on the basis of modes that apparently are not generally valid.

Whether searching for superoptimal modes in the previous step, assessing the site error by the passive training set, or directly making predictions for test compounds, the same procedure is used to find the modes having the highest calculated binding affinity. Except for very small molecules and very few regions, there are so many possible modes that an exhaustive search is out of the question. Instead, the program searches systematically over the number of atom subsets and their assignment to regions, but the choice of center atom for each subset is made at random several times, and then if any improvement can be made by shifting to an adjacent center atom, the change is accepted until no improvement is possible. This is analogous to minimizing the energy of a molecule starting from several random initial conformations. Only three random tries are required for two- and three-region sites in our experience, but four-region sites can require as many as 10, and even then the best mode is not necessarily found. Thus a site may be declared to be a solution prematurely, and underpredictions may not always be correct. A lot of effort has gone in to this part of the algorithm, and it was the motivation for departing from strict Voronoi polyhedra, because there a small shift in the center atom tends to often produce a drastic change in the partitioning of the molecule.

Analysis of Predictions. If more than one solution is found and accepted as having an adequately low error with respect to any passive training set, then each site in general will yield a different predicted optimal binding mode and binding affinity for any test compound. The variety of final sites reflects the inability of the training set to exclude all but the right answer, and the range of resulting predicted binding affinities for each test compound reflects the uncertainty in the prediction. Frequently the predicted binding interval overlaps the experimentally determined interval but lies in part outside it. This we refer to as an *excess range* prediction, being neither clearly wrong nor reliably right. If the predicted interval lies entirely within the experimental one, the prediction is *correct*; otherwise it is a clear *overprediction* or *underpredition*.

Given that we are dealing with intervals instead of single values, we have to modify the customary scatter plot to graphically display the observed vs predicted binding affinities. For example, Figure 6 represents each compound in the test set by a line segment running from (x, y) = (observed lower limit, predicted upper limit) to (observed upper limit, predicted lower limit). Any segment that crosses the observed = predicted diagonal line is either a correct or excess prediction; lying entirely above is a clear overprediction, and entirely below is a clear underprediction. An exact match between the observed and predicted intervals produces the diagonal of a square whose center is on the observed = predicted line, for example, (0.0, 3.0) to (3.0, 0.0) in Figure 9. Predicted intervals that are narrower than the observed ones appear as relatively horizontal line segments; conversely, predicted intervals that are broader than the observed are more vertical lines.

In order to compare our results with those of others, it is convenient to have quantitative measures of prediction accuracy. One way is to compare the centers (i.e. the mean of the lower and upper limits) of the observed intervals with the centers of the predicted intervals by means of the customary standard deviation σ (root mean square deviation between observed and predicted) and the correlation coefficient, ρ. One simple way to treat the scattering of predictions obtained for each compound when tested against the full set of site models is to calculate c_{pred}, the fraction of all predicted values over all site models and test compounds that fell in the observed binding interval for the respective compound.

Another measure is Kendall's τ, which is $+1$ if the ordering of predicted values agrees completely with the observed values, -1 if the ordering is completely backwards, and 0 if random. It consists of comparing every pair of compounds and noting whether the two observed intervals are nonoverlapping and whether the two predicted intervals are nonoverlapping. If so, the predicted intervals are "concordant" with the observed if they have the same ordering; otherwise they are "discordant". Let c = the number of concordant compound pairs, d = the number of discordant compound pairs, e = the number of overlapping experimental pairs, and p = the number of overlapping predicted pairs. Then

$$\tau = \frac{c - d}{\sqrt{c + d + e}\sqrt{c + d + p}}$$

which gives values near zero if there are many overlapping predicted or observed intervals. We will refer to this comparison of intervals as τ_{int}. Alternatively, one can compare the centers of the observed and predicted intervals, so as to be more comparable with conventional studies that produce a single predicted value. In that case, $e = p = 0$, usually, and the result will be denoted by τ_{cntr}.

Results

DHFR Inhibitors. Here we reexamined the same set of triazine and pyrimidine inhibitors of *L. casei* dihydrofolate reductase (DHFR) used to validate an earlier approach.[10] The compounds' structures (Figures 2 and 3) and observed binding affinities (up to an arbitrary estimate of their accuracy) are summarized in Tables 1 and 2. Suppose we wanted to construct a training set as a subset of triazines **1a**−**4a** and pyrimidines **1b**−**4b**. As seen in the first three rows of Table 3, choosing a single compound for the training set produces several site models of the very simplest kind, namely only two regions, but their predictive performance on the remaining 46 compounds is not impressive by any measure. When **1a** is the training compound, the other triazines all have excess predictions by more than 1.2 log units, while the pyrimidines all have excess predictions of more than 5.1 units. Using **1b** instead improves the excess predictions of pyrimidines slightly to >4.8, but now the triazines all have excess predictions >7.9, except for **22a**, which is nearly correctly predicted. Switching to **3b** has little effect on the triazine predictions (all being in excess), but it does reduce the excess pyrimidine predictions to >1.5 units. For this third row of the table, the greatest excess prediction among **1a**−**4a** and **1b**−**4b** is 14.8 for **3a**. A reasonable heuristic is to keep adding the worst predicted compound to the test set, resulting in the fourth row training set, {**3a**, **3b**}. Once again essentially all the predictions are in excess, by >1.3 for triazines and >1.1 for pyrimidines. For once, $\rho > 0$ and $\tau_{cntr} > 0$, but τ_{int} remains essentially zero (until the last row of the table, where a single solution produces prediction intervals of zero width.) Now the worst predicted compound among **1a**−**4a** and **1b**−**4b** is **4b** at 6.2 excess.

So far, the sets of site models have been extremely simple, giving rise to very broad prediction intervals, saying in effect that they have not learned enough from these small training sets to be generally useful, but at least they do not give clearly erroneous predictions. The tendency is for the width of the excess predictions to gradually narrow as the training sets are improved. Following our heuristic, training set {**3a**, **3b**, **4b**} still finds two-region solutions, but these are now so strongly constrained that 50% of the predictions are correct. In fact, compounds **9b**, **10b**, **11b**, **12b**, **14b**, **15b**, **16b**, **18b**, **19b**, and **20b** are correctly predicted, five compounds are overpredicted, and the rest are excess predictions. The worst prediction among the first four triazines and pyrimidines is **4a** in excess by 1.6, so the next training set becomes {**3a**, **4a**, **3b**, **4b**}. Now for the first time, no two-region solutions can be found, and the program located 14 different three-region solutions before the run was terminated. This is typical behavior for these algorithms: increasing the training set (or decreasing the error bars on the observed binding) tends to produce more geometric detail in the site models.[13] Of course, if there is a solution in n regions, there is always a solution in $n + 1$ regions, only we always stop enumerating them at the smallest possible n. If both two- and three-region solutions had been used in the predictions for the previous training sets, the results would have been much more vague. Now we have an apparent drop in predictive power for the four-compound training set, mostly because there are suddenly more adjustable geometric and energetic parameters.

When the training set consists of six compounds, namely the three triazines **1a**, **3a**, and **4a**, and the three pyrimidines **1b**, **2b**, and **3b** (last row of Table 3), EGSITE2 finds no solution for three regions, but one solution in four regions, which is shown in Table 4. Since only a single solution was found, the passive training set plays no role, and the predictions consist of only single values, not intervals. The detailed predictions for each compound are included in Tables 1 and 2. Altogether, 19 of 41 test compounds were correctly predicted, 10 were underpredicted by less than 0.4, and the rest were overpredicted, particularly **5a**, **7a**, and **8b**. The binding of methotrexate was underpredicted by 1.9 log units, and the predicted optimal binding mode does not agree with the crystal structure (Protein Data Bank entry 4DFR). For example, the glutamate moiety does not reach into the solvent, but instead prefers the very hydrophilic region 3.

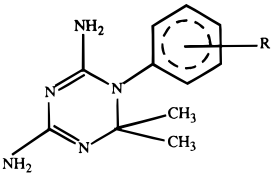

Figure 2. 4,6-Diamino-1,2-dihydro-2,2-dimethyl-1-(substituted phenyl)-*S*-triazine.

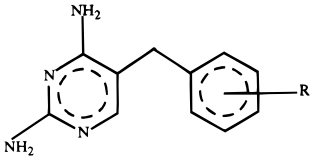

Figure 3. 2,4-Diamino-5-(substituted phenyl)pyrimidine.

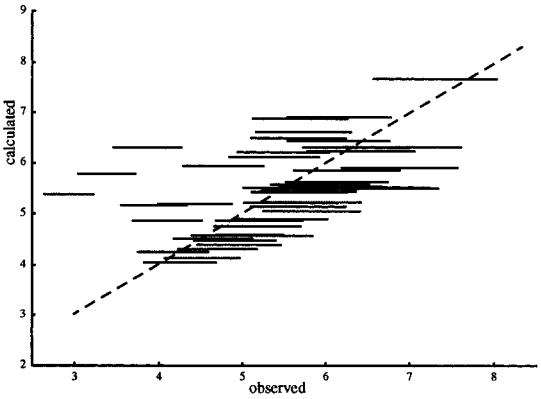

Figure 4. Observed vs calculated binding of DHFR inhibitors (Tables 1, 2, and 4). Observed binding affinities are intervals, represented by horizontal line segments. Any that cross the dashed line representing observed = calculated are correct predictions.

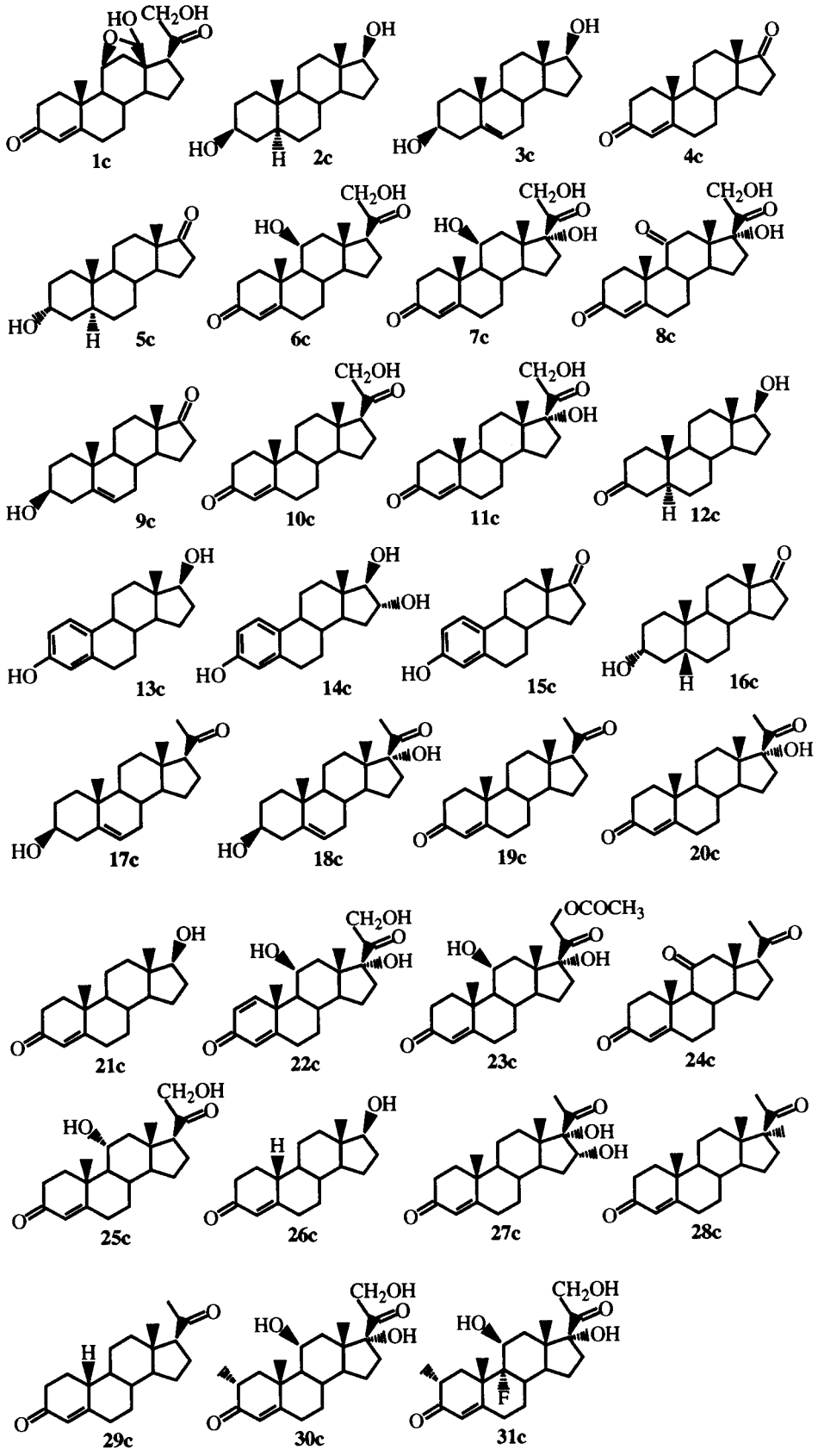

Figure 5. Steroid structures.

It is interesting to compare these results with our earlier work,[10] where the same molecules were grouped into three to five superatoms each, eight compounds were used in the training set ({**1a**−**4a**, **1b**−**4b**}), and no one region was constrained to represent the solvent. A solution was found for three nonsolvent regions, although one of these had weak interactions with ligands and thus approximated the solvent. There were 23 of 39 correct predictions, and **5a**, **7a**, and **8b** were also outliers. The combinatorial search was greatly simplified by considering small numbers of superatoms, and when methotrexate was simplified into three su-

Table 1. Observed and Calculated Binding of Triazine Inhibitors of DHFR

compd	R[a]	obsd[b]	calcd
1a[c]	H	[4.23, 5.17]	4.32 ok
2a	3-I	[4.66, 5.70]	4.76 ok
3a[c]	3-OBzCl$_2$	[5.01, 6.13]	5.51 ok
4a[c]	4-OCH$_3$	[3.69, 4.51]	4.87 hi
5a	3-SO$_2$NH$_2$	[2.64, 3.22]	5.38 hi
6a	3-COCH$_3$	[3.82, 4.68]	4.05 ok
7a	3-OH	[3.46, 4.27]	6.30 hi
8a	3-CF$_3$	[4.29, 5.25]	5.94 hi
9a	3-F	[4.39, 5.37]	4.58 ok
10a	3-CN	[4.78, 5.84]	4.58 lo
11a	3-CH$_3$	[4.46, 5.46]	4.40 lo
12a	3-CH$_2$CH$_3$	[4.68, 5.49]	4.60 lo
13a	3-OCH$_3$	[4.07, 4.97]	4.14 ok
14a	3-OCH$_2$CH$_3$	[4.67, 5.71]	4.58 lo
15a	3-OPr	[5.02, 6.14]	5.22 ok
16a	3-OHx	[5.12, 6.26]	6.88 hi
17a	3-OBz	[5.11, 6.25]	5.43 ok
18a	3-OCH$_2$Ph	[5.91, 7.23]	5.52 lo
19a	4-OH	[4.42, 5.40]	4.48 ok
20a	4-NH$_2$	[3.55, 4.33]	5.17 hi
21a	4-I	[3.99, 4.87]	5.19 hi
22a	4-CH$_3$	[3.75, 4.59]	4.26 ok
23a	4-F	[4.18, 5.12]	4.52 ok

[a] See Figure 2. [b] − log K ± 10%; ref 11. [c] Training set.

Table 2. Observed and Calculated Binding of Pyrimidine Inhibitors of DHFR

compd	R[a]	obsd[b]	calcd
1b[c]	H	[4.68, 5.72]	4.87 ok
2b[c]	3-OBu	[5.52, 6.74]	5.63 ok
3b[c]	4-I	[6.00, 7.34]	5.50 lo
4b	3,4,5-(OCH$_3$)$_3$	[6.19, 7.57]	5.91 lo
5b	3-F	[4.84, 5.92]	6.12 hi
6b	3-CH$_2$OH	[5.10, 6.24]	6.49 hi
7b	4-NH$_2$	[4.92, 6.02]	4.90 lo
8b	3,5-(CH$_2$OH)$_2$	[5.16, 6.30]	6.61 hi
9b	4-F	[5.10, 6.24]	5.14 ok
10b	3,4-(OH)$_2$	[5.26, 6.42]	5.23 lo
11b	3-OH	[5.24, 6.40]	5.49 ok
12b	4-CH$_3$	[5.25, 6.41]	5.05 lo
13b	3-CH$_2$OBu	[4.94, 6.04]	6.21 hi
14b	3-CH$_3$	[5.20, 6.36]	5.44 ok
15b	4-OCH$_3$	[5.62, 6.88]	5.86 ok
16b	4-OBu	[5.73, 7.00]	6.31 ok
17b	4-NHCOCH$_3$	[5.44, 6.66]	5.55 ok
18b	3-OCH$_3$	[5.34, 6.52]	5.58 ok
19b	3-OBz	[5.54, 6.76]	6.44 ok
20b	3-CF$_3$	[5.54, 6.78]	6.91 hi
21b	3-CF$_3$, 4-OCH$_3$	[6.57, 8.03]	7.68 ok
22b	3,4-(OCH$_3$)$_2$	[6.22, 7.61]	6.32 ok
23b	3,5-(OCH$_3$)$_2$	[5.78, 7.06]	6.24 ok
24b	3,5-(OH)$_2$	[3.04, 3.72]	5.78 hi

[a] See Figure 2. [b] − log K ± 10%; ref 12. [c] Training set.

Table 3. DHFR Models from Various Training Sets

training set	no. of regions	no. of solutions	c_{pred} (%)	σ[a]	ρ	τ_{cntr}
1a	2	5	11.3	2.13	−0.01	0.07
1b	2	8	11.7	3.61	−0.47	−0.28
3b	2	17	17.0	2.33	0.03	0.05
3a, **3b**	2	12	19.2	1.85	0.15	0.21
3a, **3b**, **4b**	2	6	50.0	1.05	0.22	0.38
3a, **4a**, **3b**, **4b**	3	14	43.3	2.08	0.21	0.31
1a, **3a**, **4a**, **1b**, **2b**, **3b**	4	1	48.8	0.89	0.51	0.45

[a] Units of log K.

peratoms (pteridine, p-aminobenzoyl, and glutamate), it was predicted to bind in a mode that superimposed well on the crystal structure of the DHFR/methotrexate complex. We conclude that EGSITE2 would have to go to five regions and extremely lengthy calculations to improve on our current result for this dataset, but at least it has produced a result of quality comparable to the earlier work and suggests that it can extract more information from a given training set by using all atoms.

Table 4. DHFR Four-Region Site Model

region geometry (Å)				region energetics		
				HP	MR	charge
∞	∞	∞	∞	0.00	0.00	0.00
0	8.13	∞	∞	1.17	−0.07	4.31
0	0	∞	∞	−1.09	0.00	−3.20
0	0	1.21	∞	0.42	0.01	−1.81

Table 5. Observed and Calculated Binding of CBG and TBG with Steroids **1c**−**31c**

	CBG		TBG	
compd[a]	obsd[b]	calcd	obsd[c]	calcd
1c	[5.28, 7.28]	[7.28, 7.80][d] xs	[4.32, 6.32]	4.46[e] ok
2c	[4.00, 6.00]	[5.85, 6.11][f] xs	[8.11, 10.11]	8.11[e] ok
3c	[4.00, 6.00]	[6.11, 6.26][f] hi	[8.18, 10.18]	8.04 lo
4c	[4.76, 6.76]	[5.44, 6.87][f] xs	[6.46, 8.46]	5.82 lo
5c	[4.61, 6.61]	[5.56, 5.94][f] ok	[6.15, 8.15]	5.89 lo
6c	[6.88, 8.88]	[7.15, 7.23][f] ok	[5.34, 7.34]	5.42 ok
7c	[6.88, 8.88]	[7.54, 7.63][f] ok	[5.20, 7.20]	5.21 ok
8c	[5.89, 7.89]	[7.34, 7.37][f] ok	[5.43, 7.43]	7.53 hi
9c	[4.00, 6.00]	[6.25, 6.51] hi	[6.82, 8.82]	7.07 ok
10c	[6.65, 8.65]	[6.78, 6.85] ok	[6.38, 8.38]	6.88 ok
11c	[6.88, 8.88]	[6.99, 7.15] ok	[6.20, 8.20]	6.30 ok
12c	[4.92, 6.92]	[5.56, 5.94] ok	[8.74, 10.74]	6.61 lo
13c	[4.00, 6.00]	[5.26, 5.64] ok	[7.83, 9.83]	7.38 lo
14c	[4.00, 6.00]	[5.94, 6.02] xs	[5.63, 7.63]	6.01 ok
15c	[4.00, 6.00]	[4.97, 5.47] ok	[7.18, 9.18]	6.62 lo
16c	[4.22, 6.22]	[5.56, 6.41] xs	[5.15, 7.15]	5.80 ok
17c	[4.22, 6.22]	[6.62, 6.75] hi	[6.15, 8.15]	6.36 ok
18c	[4.00, 6.00]	[6.82, 7.01] hi	[5.36, 7.36]	5.40 ok
19c	[6.38, 8.38]	[6.38, 6.55] ok	[5.94, 7.94]	6.46 ok
20c	[6.74, 8.74]	[6.45, 6.81] xs	[6.00, 8.00]	6.06 ok
21c	[5.72, 7.72]	[5.73, 6.06] ok	[8.20, 10.20]	6.46 lo
22c	[6.51, 8.51]	[7.70, 7.84] ok		
23c	[6.55, 8.55]	[8.24, 9.27] xs		
24c	[5.78, 7.78]	[6.31, 6.71] ok		
25c	[6.20, 8.20]	[7.15, 7.23] ok		
26c	[5.11, 7.11]	[5.47, 5.76] ok		
27c	[5.25, 7.25]	[6.99, 7.44] xs		
28c	[6.12, 8.12]	[6.20, 6.78] ok		
29c	[5.82, 7.82]	[5.84, 6.20] ok		
30c	[6.69, 8.69]	[7.82, 7.85] ok		
31c	[4.80, 6.80]	[7.79, 8.07] hi		

[a] See Figure 5. [b] − log K_{diss} ± 1.0; ref 14. [c] − log K_{diss} ± 1.0; ref 15. [d,e] Active training set. [f] Passive training set.

Binding to CBG. Here we tested EGSITE2 on the same set of 31 steroids binding to human corticosteroid-binding globulin (CBG)[14] that has served as a benchmark for several computational methods.[1,16−19] The structures and numbering of the compounds is as in Jain et al.,[19] and the set of conformers for each was derived as before,[1] except between six and nine of the lowest energy conformers were kept for each compound to represent conformational flexibility. The observed binding affinity was taken to be −log K_{diss} ± 1.0, where the assumed error limits are comparable to our previous work, but real experimental errors are not known. Using only **1c** as the active training set, along with **2c**−**8c** as passive training compounds, the program easily found its given limit of 10 site models, each having only two regions. These represented a total of five different geometries, but those where the nonsolvent region was not very large gave prediction errors on the passive set in the range of 0.89−2.95. The four best sites (Table 6) had errors of only 0.29−0.44. These site models agree qualitatively with our previous results on CBG using the program EGSITE.[1] These models were tested

Table 6. The Four CBG Two-Region Site Models

model	region geometry (Å)[a]		region energetics		
			HP	MR	charge
solvent	∞	∞	0.00	0.00	0.00
1	0	∞	−0.42	0.09	0.50
2	0	∞	−0.22	0.08	1.27
3	0	∞	−0.26	0.08	0.67
4	0	13.7	−0.27	0.08	1.68

[a] Each of the four models consists of a solvent region described in the first row of the table and one binding pocket described by one of the other four subsequent rows. Thus the geometry of the first model is given in terms of the 2 × 2 matrix of interregion distances given in the first two rows.

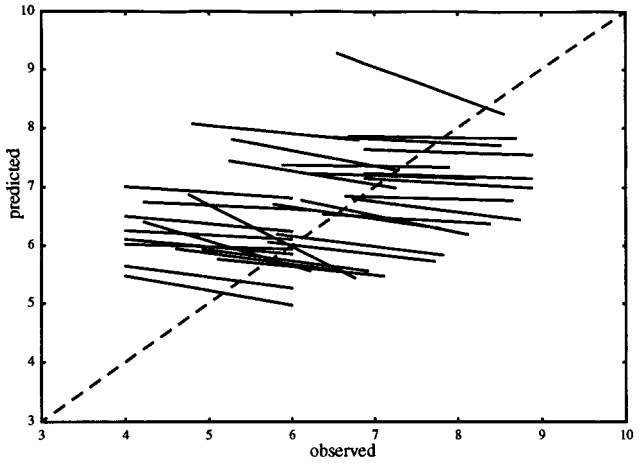

Figure 6. Observed vs calculated binding of the 31 CBG ligands to four site models (Tables 5 and 6). Since both observed and calculated affinities are intervals, each compound is represented by a diagonal line segment running from the minimal observed and maximal calculated to the maximal observed and minimal calculated. Any that cross the dashed line representing observed = calculated are correct predictions.

against the remaining 23 test compounds (Table 5), and we also checked the active and passive compounds for predicted binding, realizing that a renewed search for optimal binding modes may fail to find ones as good as those located in the training procedure, or it may discover superoptimal modes.

The one training compound, **1c**, was overpredicted by 0.52 by one of the four models, but correctly predicted by the rest. Of the seven passive training compounds, four were correctly predicted, two had excess ranges, and one was high by 0.1. Of the 23 test compounds, 14 were correct, one was in excess by 0.02, four were substantially in excess, and four were high. These results are summarized graphically in Figure 6. This dataset is particularly favorable to EGSITE2 because although steroids do bind strongly and specifically to CBG, their binding can be rationalized in terms of an exceptionally vague or nonspecific site geometry.

At first glance it may seem outrageous that on the basis of a single training compound, only four out of 30 were clearly mispredicted, and for six the predictions were neither clearly right nor wrong. Remember, however, that the training procedure ensures that there is at least one in-range mode and no superoptimal modes out of the enormous number possible for that one training compound. In other words, one compound may constitute many more than one constraint on the site model. Contrast this with a more conventional method where the purported binding mode is somehow chosen in advance, and then the model's parameters are adjusted so that the calculated binding affinity agrees with the observed. This really is a matter of fitting one data point with several adjustable parameters, and in general there will be many alternative binding modes of that same molecule having superoptimal calculated affinities. The predicted binding for other compounds can differ a great deal, depending on which single mode of the training compound was selected and which set of parameter values was used to fit it, out of the many possible. In terms of drawing a graph, this is equivalent to plotting a single data point and then drawing a broad fan of many different lines through it. As is well-recognized, the fit becomes well-behaved and the predictions become useful only when there are substantially more data points than adjustable parameters. What EGSITE2 has done in this example of a single training compound is to search over the many candidate optimal binding modes and the space of adjustable parameters for one such that that mode is in-range and no other mode is superoptimal. In the graph drawing analogy, one compound has generated many data points so that only a narrow bundle of lines fit them.

Table 7. The One TBG Two-Region Site Model

region geometry (Å)		region energetics		
		HP	MR	charge
∞	∞	0.00	0.00	0.00
0	11.56	0.002	0.07	−3.07

In a comparison of the 23 predicted binding affinity intervals with the assumed observed intervals, $\tau_{int} = 0.27$. In terms of interval centers, our standard deviation in prediction was 0.98 log units, the correlation coefficient was 0.52, and $\tau_{cntr} = 0.40$. In comparison, Jain et al.[19] used **1c**−**21c** as their training set and then tested the 10 compounds **22c**−**31c**. Their standard deviation was 0.70, the correlation coefficient was 0.40, and $\tau = 0.46$. For them, **31c** was a noticeable outlier, whereas it is not in this study. Their results on cross-validation would be another point of comparison, except that it is totally inappropriate in our approach. Our training sets are of minimal size, and each compound contributes essential information toward determining the site. Because of this lack of redundancy, the leave-one-out cross-validation protocol would always give very bad predictions, as demonstrated with the DHFR inhibitors.

Binding to TBG. Here we examined the standard test case of 21 steroids (Table 5) binding to testosterone-binding globulin (TBG)[15] treated in the same way as the larger CBG data set. If only **1c** is used as the training set, there are always substantial underpredictions of particularly **2c** and **3c** in the site models initially sampled. This is in agreement with the general experience that the TBG data set is more challenging. Going to an active training set consisting of **1c** and **2c** resulted in only one site model (Table 7), and thus there is in effect no passive training set involved. In predictions, the active training set was all correct, as were 11 out of the 19 test compounds. **8c** was marginally overpredicted by 0.1, and **3c** was marginally low; six other test compounds were underpredicted. See Figure 7.

Quantitatively for the 19 test compounds, $\tau_{int} = 0.37$. The interval centers tests give a standard deviation of 1.36 log units, a correlation coefficient of 0.60, and τ_{cnt}

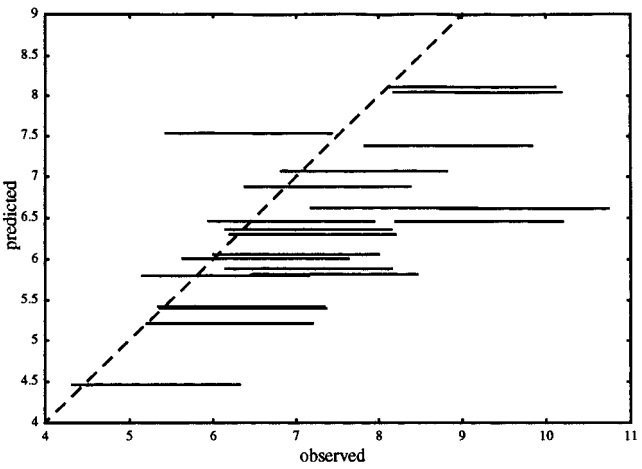

Figure 7. Observed vs calculated binding of the 21 TBG ligands in the single site model (Tables 5 and 7), drawn as in Figure 4.

= 0.49. This is very comparable to our CBG results, but it did require two compounds in the training set instead of one. Since Jain et al. used all 21 compounds for training, they have only cross-validation results for TBG.

ACE Inhibitors. Inhibitors of angiotensin converting enzyme (ACE) have been studied by detailed conformational search methods to produce a narrowly defined required pharmacophore.[20,21] This dataset would therefore likely be particularly challenging for EGSITE2, which tends to search for very vague site models, as in the CBG example. Worse yet, the 30 compounds are especially conformationally flexible (Figure 8) and structurally diverse, which makes it more difficult to determine the optimal binding mode for a molecule in a given trial site model, especially as the number of regions increases and their diameters decrease. Here we number the structures as before,[20] but we convert the IC_{50}s to a logarithmic scale of binding affinity and add an arbitrary "error" of ±0.5 log units. Compounds **1d**−**28d** are potent ACE inhibitors, but **29d** and **30d** are only qualitatively described as inactive, so here they are assigned very weak binding intervals of [0.0, 3.0], well below any of the other compounds. Conformational flexibility is treated only approximately in terms of a broad sampling of nine low-energy structures per compound.

Not surprisingly, the inactive compounds are crucial for determining site models. When the active training set consists of **1d**, **24d**, and **29d**, EGSITE2 finds no two-region solutions, but easily enumerates the given limit of 20 different three-region solutions. Of these, 18 have prediction errors between 30 and 37 on a passive training set consisting of the 12 compounds **2d**, **4d**, **7d**−**9d**, **11d**, **13d**, **17d**, **26d**−**28d**, and **30d**. The other two solutions had some inconsequential differences in geometry but identical interaction energy parameters and thus gave identical predictions, such as total errors against the passive training set of only 14.6 log units. Against the true test set of 15 compounds, only four were correctly predicted (**18d**, **20d**, **22d**, and **25d**), seven were underpredicted, and four were overpredicted. In particular, the other inactive compound, **30d**, is substantially overpredicted. Clearly this is an example of a small training set being satisfied in many different ways, only a tiny fraction of which have any predictive value on test compounds.

Expanding the training set to **1d**, **24d**, **29d**, and **30d**, EGSITE2 quickly eliminates all three-region solutions and embarks on a lengthy search for four-region sites. After about 100 h on an SGI Indy R5000 workstation, 18 different geometries were considered, of which two developed into solutions after several cycles of searching for superoptimal modes and adjusting the energies (Table 9). When used for predictions, all four compounds in the training set were correct, as they should be. No test compounds were overpredicted, 13 were underpredicted (although **13d** was only 0.05 low), and the other 13 had predicted ranges of binding that overlapped the observed range. Since many of the line segments in Figure 9 are steeply inclined, it is clear that the predicted range was often much greater than the observed. Comparing the mean predicted values to the mean observed over the 26 test compounds, yields $\sigma =$ 2.6 and $\rho = 0.31$. Comparing the observed intervals to the predicted ones gives $\tau_{int} = 0.03$, due to the many overlaps, but comparing the interval centers yields τ_{cntr} = 0.26.

In the earlier work of Marshall and coworkers,[20,21] the aim was to determine a common pharmacophore among the active compounds **1d**−**28d**, rather than to quantitatively fit the observed IC_{50}s. They determined five distances between an amide carbonyl oxygen, its bonded carbon, a carboxyl oxygen, and the position where the ACE zinc atom would be when complexed by, for example, the sulfhydryl group in **1d**−**5d**, **11d**, **13d**, **16d**,**18d**, **19d**, and **22d**. Other zinc liganding groups were selected for the other compounds. Initially,[20] they determined one set of such distances to a precision of 0.15 Å, but their later work[21] found a second substantially different set as well and relaxed the estimated precision of both to 0.5 Å. In our terms, this would correspond to two different site models, each consisting of five regions: solvent and four pockets having diameters of roughly 1 Å. Our results are in vague agreement insofar as EGSITE2 was forced into the most detailed site models seen to date, in spite of its built-in bias toward low-resolution geometries, but a five-region model such as theirs is currently not a very feasible calculation. On the basis of only two active compounds in the training set, **1d** and **24d**, and a much coarser search over the available conformations, we find no consistent pharmacophore whatever. For example, the calculated binding mode of **1d** in our first model (Figure 10) puts the sulfhydryl sulfur atom in region 2, one methylene hydrogen in region 3, and the rest in the solvent. The second model seizes the oxygen atom of one amide carbonyl in region 2 and its bonded carbon in region 3, while leaving the rest of the molecule out in the solvent. At this stage in the development, one should not take these results too seriously, other than to note that those who diligently search for a preconceived pharmacophore can find one, while those without such preconceptions do not necessarily reach the same conclusion.

In more recent work, CoMFA was used with a training set of 68 compounds, which included **1d**−**28d**, and then tested on 20 other ACE inhibitors.[22,23] These studies used a much larger training set and a different test set of somewhat smaller size than we did, and of

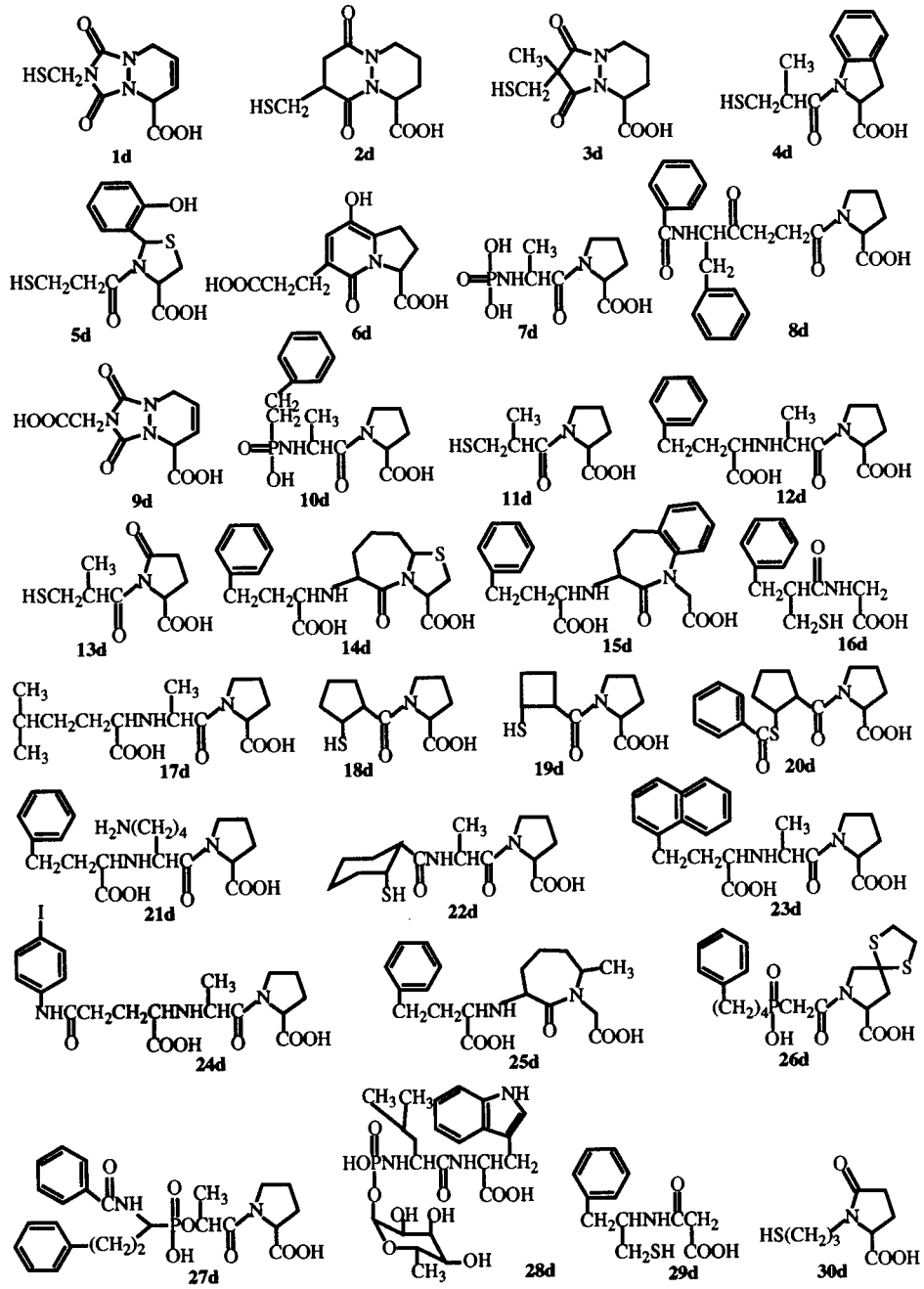

Figure 8. Structures of ACE inhibitors.

course the emphasis in CoMFA is on producing a least squares fit between the single (best estimated) observed binding affinity of each compound and the single calculated value. Nevertheless, we can attempt to compare our results with theirs by using the midpoints of our predicted binding intervals as our predicted values and comparing their "predictive R^2" values with ours. In terms of the notation of this paper,

$$R^2 = 1 - \frac{\sigma^2}{\langle(\text{obsd} - \langle\text{obsd}\rangle_{\text{train}})^2\rangle_{\text{test}}}$$

which amounts to comparing the standard deviation of the predictions vs observed values for the test set to the standard deviation between the observed binding affinities of the test compounds and the mean observed binding of the training compounds. If the prediction does no better than to assume every test compound binds at the mean level of the training compounds, then $R^2 = 0$; better predictions give $R^2 > 0$. Our results in Table 8 for four training compounds and 26 test compounds corresponds to $R^2 = 0.47$. The first CoMFA study[22] (68 training compounds and 20 test compounds) gave $R^2 = 0.53$, while the alignment rule used in the second study produced $R^2 = 0.46$. In other words, we can predict a few more compounds on the basis of a much smaller training set to comparable accuracy, according to this measure of quality.

Conclusions

It is possible to derive very simple receptor site models from remarkably small training sets and yet have a predictive quality comparable to other methods. Results depend on exactly which compounds are used for training, and the correctness and decisiveness of the predictions tend to improve by adding compounds to the

Table 8. Observed and Calculated Binding of ACE Inhibitors

compd[a]	obsd[b]	calcd
1d	[5.6, 6.6]	[6.41, 6.42][c] ok
2d	[6.9, 7.9]	[3.01, 3.03] lo
3d	[5.5, 6.5]	[3.03, 7.41] xs
4d	[7.9, 8.9]	[3.29, 7.45] lo
5d	[7.7, 8.7]	[6.02, 6.84] lo
6d	[7.5, 8.5]	[3.20, 7.10] lo
7d	[8.3, 9.3]	[3.24, 3.32] lo
8d	[8.0, 9.0]	[4.48, 5.31] lo
9d	[8.5, 9.5]	[2.85, 6.41] lo
10d	[7.6, 8.6]	[3.57, 9.58] xs
11d	[7.1, 8.1]	[3.01, 7.45] xs
12d	[8.4, 9.4]	[3.73, 7.70] lo
13d	[7.5, 8.5]	[3.08, 7.45] lo
14d	[8.7, 9.7]	[6.75, 7.70] lo
15d	[8.0, 9.0]	[4.40, 9.07] xs
16d	[6.5, 7.5]	[3.59, 6.59] xs
17d	[8.1, 9.1]	[4.15, 7.26] lo
18d	[6.9, 7.9]	[3.34, 7.31] xs
19d	[6.8, 7.8]	[3.10, 7.27] xs
20d	[7.2, 8.2]	[5.77, 8.39] xs
21d	[8.4, 9.4]	[3.50, 10.92] xs
22d	[8.0, 9.0]	[3.38, 7.53] lo
23d	[8.5, 9.5]	[4.45, 8.81] xs
24d	[9.1, 10.1]	[9.10, 9.10][c] ok
25d	[8.0, 9.0]	[4.04, 7.76] lo
26d	[7.9, 8.9]	[6.43, 11.06] xs
27d	[7.4, 8.4]	[4.11, 9.20] xs
28d	[8.3, 9.3]	[5.39, 11.66] xs
29d	[0.0, 3.0]	[0.00, 3.00][c] ok
30d	[0.0, 3.0]	[0.00, 3.00][c] ok

[a] See chemical structures in Figure 8. [b] From ref 20. − log IC_{50} ± 0.5. [c] Active training set.

Table 9. Two ACE Four-Region Site Models

region geometry (Å)				region energetics		
				HP	MR	charge
∞	∞	∞	∞	0.00	0.00	0.00
0	9.21	6.81	6.01	−0.71	−0.16	−0.99
0	0	2.40	8.01	−1.76	0.52	−0.88
0	0	0	1.20	61.38	−17.98	−31.00
∞	∞	∞	∞	0.00	0.00	0.00
1.20	9.21	1.60	0.80	0.40	0.03	2.45
1.60	1.20	0.80	6.01	−8.62	1.03	−6.10
0	0	1.20	0.80	7.38	−0.17	14.71

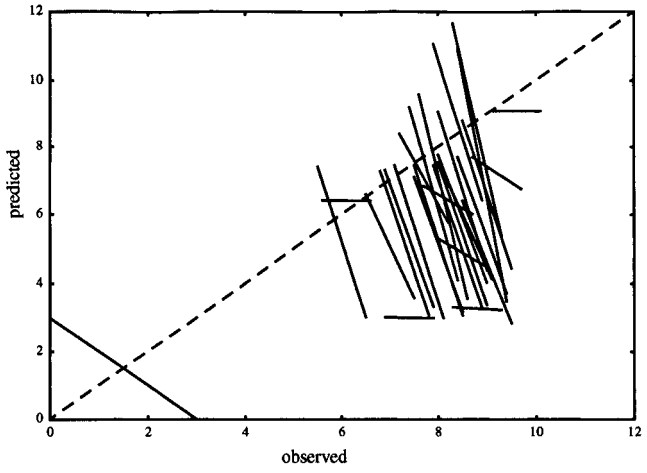

Figure 9. Observed vs calculated binding of the 30 ACE inhibitors to two site models (Tables 8 and 9), drawn as in Figure 4.

training set. The key idea behind our approach is that some compounds can be especially informative about the sort of receptor they must interact with, rather than relying on statistical methods to filter out the significant

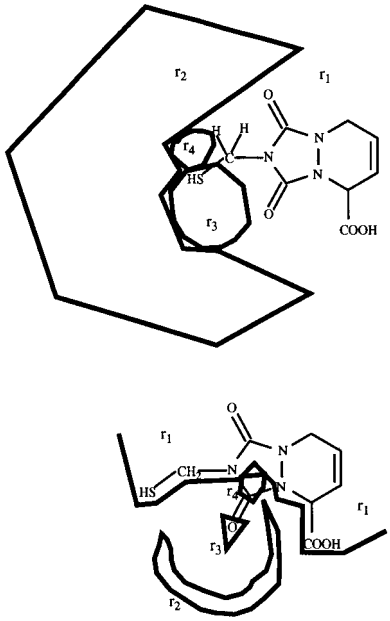

Figure 10. Sketch of the two ACE site models drawn roughly to scale with **1d** in its calculated optimal binding mode. The heavy lines indicate a possible way to draw the boundaries of the regions so as to agree with the interregion distance bounds in Table 9.

features of a large training set. The way we wring so much information out of so few compounds is to require that the best binding mode out of all possible ones for each molecule must interact with the site model to give a calculated binding strength in agreement with the observed value. From tests on standard data sets, we conclude that this method is likely to work best in situations where relatively nonspecific interactions account for the binding, so that a satisfactory model can be quite simple.

Appendix

In the Methods section, the EGSITE2 algorithm was explained in general outline with emphasis on motivation, design principles, and performance considerations. The precise description of the entire method is contained in 6500 lines of C++ code, which is excessive detail. In the interests of explaining EGSITE2 well enough that it could in principle be reproduced, the following explanation of the mixed integer program formulations is required.

For the ith mode in the basis, let Y_i = the Boolean variable for mode i being in-range (1 = true, 0 = false), g_i = Boolean allowed geometrically, h_{ij} = Boolean disallowed for the jth reason out of n_{ih} possible reasons. The geometry of the site is expressed as G_{Ukl} = the upper bound on the distance between a point in region k and a point in region l, and G_{Lkl} = the corresponding lower bound, taken over all pairs of the n_r regions. The corresponding atomic subsets implied by the mode have d_{Ukl} = the upper bound on the distance between an atom that is supposed to lie in region k and an atom that lies in region l; and d_{Lkl} is the corresponding lower bound on interatomic distances. The region interaction parameters are concatenated into a list of variables S_k, so that if for a particular binding mode the sum of the corresponding fixed atomic physicochemical property values for those atoms in that region is p_k, then the calculated binding affinity is $\sum_k S_k p_k$. The molecule to

Table 10. Inequalities Used in Mixed Integer Programs

mathematical statement	verbal equivalent
$(d_{Ukl} + 1)g_i - G_{Ukl} \leq 0$	If the mode is allowed, the maximal interatomic distance must be less than the maximal interregion distance.
$(-d_{Ukl} + 1 + M)h_{ij} + G_{Ukl} \leq M$	Otherwise the mode is disallowed by reason of the maximal interatomic distance exceeding the maximal interregion distance.
$d_{Lkl} + (M - d_{Lkl} + 1)g_i \leq M$	If the mode is allowed, the minimal interatomic distance must be greater than the minimal interregion distance.
$(d_{Lkl} + 1)h_{ij} - G_{Lkl} \leq 0$	Otherwise the mode is disallowed by reason of the minimal interatomic distance being less than the minimal interregional distance.
$-\sum_k p_k S_k + (L_i + M)Y_i \leq M$	If the mode is in-range, its calculated binding affinity must be at least the lower observed limit, but if not, it is unconstrained from below.
$\sum_k p_k S_k + (M - L_i)g_i + (L_i - U_i)Y_i \leq M$	If the mode is allowed, its binding affinity must never be superoptimal, but if it is not in-range, the calculated affinity must be below even the lower observed limit.
$\sum h_{ij} + n_{hi}g_i \leq n_{hi}$	If the mode is allowed, it must not be disallowed for any reason.
$-\sum h_{ij} - g_i \leq -1$	Either the mode is allowed or it is disallowed for at least one reason.
$Y_i \leq g_i$	If the mode is in-range, it must be geometrically allowed.
$-\sum_{\text{modes of molec}} Y_i \leq -1$	At least one of the modes for a given molecule must be in-range.
$\sum^{n}_{\text{all modes}} (2Y_{i,\text{ old}} - 1)\, Y_i \leq n - 1$	Exclude the old combination of in-range modes for all molecules expressed as the old values $Y_{i,\text{ old}}$.
$\sum G_{Ukl} - \sum G_{Lkl} < \sum G_{Ukl,\text{ old}} - \sum G_{Lkl,\text{ old}} - 1$	Exclude the old site geometry.

which mode i corresponds has experimental binding in the range $[L_i, U_i]$. Let M be a large positive number that is bigger than any experimental binding affinity or the diameter of any molecule in any conformation. Now in terms of this notation, Table 10 gives the linear constraints associated with a mode and the matching English translation. The first four equations apply to whatever distance comparisons are nontrivial, each one counting as a different possible reason for disallowing a mode. If $k = l$, the minimal interregion distance is trivially zero; if one of the regions is the solvent, interregion upper bounds are always effectively infinite. As described previously, the distances between atomic subsets are rescaled to odd integers so that these inequalities are satisfied with a margin of 1, making the scaled interregion distances even integers.

In order to find a candidate site geometry from the initial mode basis, every basis mode contributes several inequalities having the types given in the first nine rows of Table 10. Exactly how many depends on n_{ih}, which depends on n_r and whether a subset of atoms is assigned to the solvent in mode i. In addition, there is one inequality for each molecule to ensure that at least one of its modes in the basis is in-range (Table 10, row 10). Of course all the Boolean variables are bounded between 0 and 1, and it helps the numerical stability of the linear program to require $-M \leq S_k \leq M$ for all the $3(n_r - 1)$ interaction parameters, and $0 \leq G_{ULkl} \leq M$ for all the adjustable geometric parameters. A linear program (LP) is the minimization of a linear objective function subject to a set of linear inequalities and equalities. An LP is called feasible if there is some solution to the set of inequalities; otherwise it is said to be infeasible. Here the LP tries to minimize $\sum_{k,l}(G_{Lkl} - G_{Ukl})$, which biases the solutions toward unrestrictive geometry, i.e. large diameter regions that touch one another. That way test molecules tend not to be excluded from regions unless there was a good reason in the training set.

Solving the LP by standard methods embodied in Cplex leads generally to a solution where most of the Boolean variables have intermediate values, such as 0.5. Inspection of the set of inequalities reveals that some of the Boolean variables are much more influential than others at forcing other variables toward 0 or 1 if they themselves are set to 0 or 1. Solving the mixed integer program (MIP) therefore consists of exploring a huge tree of possible 0/1 choices for the Boolean variables, beginning by first choosing alternative values for the most influential variables, solving the remaining LP, and trying to constrain remaining Boolean variables that are not already at their limits if the LP was feasible. This is the standard branch-and-bound approach to solving a MIP, except that the initial ordering of variables is crucial to speed, and no bounding condition is used beyond the possible infeasibility of the associated LP. The process terminates upon finding the first combination of assigned values of the Boolean variables such that all the rest have become 0 or 1 and the LP is feasible, exploring the entire tree far enough to determine that all branches eventually become infeasible, or giving up after 10^4 tree nodes have been explored.

If there are n Boolean variables in the MIP, all possible solutions can be viewed as the 2^n corners of an n-dimensional hypercube. The linear constraints on the continuous variables tend to exclude many of these corners, but generally many others remain. When EGSITE2 is searching for different candidate site geometries, it locates one feasible corner, notes down the corresponding site geometry and interaction parameters, and then adds a cutting plane (Table 10, row 11) that excludes only that one corner from further consideration. This continues until either all feasible corners have been cut off or the maximum number of candidate sites has been produced. This strategy is very conservative in that each such cutting plane excludes exactly the one combination of in-range modes, but this may not lead immediately to a different site geometry because another combination of in-range modes may be compatible with the same geometry. There is also the drawback that the exclusion is done with respect to a certain set of binding modes, rather than directly excluding a combination of geometric parameters. Thus if the set of modes is changed, the cutting planes may become invalid. A less conservative strategy is to employ the cutting plane given in line 12 of Table 10, which excludes an old set of geometric parameters and not much else since the LP always seeks to minimize $\sum(G_{Lkl} - G_{Ukl})$.

When EGSITE2 is adjusting only the interaction parameters, it considers only those modes in the basis that are geometrically allowed ($g_i = 1$) so that each of

these contributes exactly two inequalities from Table 1, namely those involving the S_k. Then for each molecule there is the constraint given in row 10 of the table, where the sum runs of course over only the allowed modes. This still is a mixed integer program since the Y_i are Boolean, but it involves many fewer variables and constraints than the full search for both site geometry and interaction parameters. The objective function attempts to maximize the sum of the calculated binding affinities over all molecules in the training set, but this is probably not important.

References

(1) Schnitker, J.; Gopalaswamy, R.; Crippen, G. M. Objective models for steroid binding sites of human globulins. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 93−110.

(2) Crippen, G. M. Intervals and the deduction of drug binding site models. *J. Comput. Chem.* **1995**, *16*, 486−500.

(3) Bradley, M.; Richardson, W.; Crippen, G. M. Deducing molecular similarity using Voronoi binding sites. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 750−755.

(4) Molecular Simulations, Inc., 9685 Scranton Rd., San Diego, CA; http://www.msi.com.

(5) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: A generic force field for molecular simulations. *J. Phys. Chem.* **1990**, *94*, 8897−8909.

(6) Ghose, A. K.; Pritchett, A.; Crippen, G. M. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships III: Modeling hydrophobic interactions. *J. Comput. Chem.* **1988**, *9*, 80−90.

(7) Mortier, W. J.; Van Genechten, K.; Gasteiger, J. Electronegativity equalization: Application and parametrization. *J. Am. Chem. Soc.* **1985**, *107*, 829−35.

(8) CPLEX Optimization, Inc., Suite 279, 930 Tahoe Blvd., Incline Village, NV; http://www.cplex.com.

(9) Press, W; Teukolsky, S.; Vetterling, W.; Flannery, B. *Numerical Recipes in C*, 2nd ed.; Cambridge University Press, Cambridge, 1992.

(10) Bradley, M. P.; Crippen, G. M. Voronoi modeling: The binding of triazines and pyrimidines to *L. casei* dihydrofolate reductase. *J. Med. Chem.* **1993**, *36*, 3171−3177.

(11) Hansch, C.; Hathaway, B. A.; Guo, Z. R.; Dias Selassie, C.; Dietrich, S. W.; Blaney, J. M.; Langridge, R.; Volz, K. W.; Kaufman, B. T. Crystallography, quantitative structure−activity relationships, and molecular graphics in a comparative analysis of inhibition of dihydrofolate reductase from chicken liver and *Lactobacillus casei* by 4,6-diamino-1,2-dihydro-2,2-dimethyl-1-(substituted-phenyl)-*S*-triazines. *J. Med. Chem.* **1984**, *27*, 129−143.

(12) Hansch, C.; Li, R.-L.; Blaney, J. M.; Langridge, R. Comparison of inhibition of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase by 2,4-diamino-5-(substituted-benzyl)-pyrimidines: Quantitative structure−activity relationships, X-ray crystallography, and computer graphics in structure−activity analysis. *J. Med. Chem.* **1982**, *25*, 777−784.

(13) Boulu, L. G.; Crippen, G. M. Voronoi Binding Site Models: Calculation of Binding Modes and Influence of Drug Binding Data Accuracy. *J. Comput. Chem.* **1989**, *10*, 673−682.

(14) Mickelson, K. E.; Forsthoefel, J.; Westphal, U. Steroid−protein interactions. Human corticosteroid binding globulin: Some physicochemical properties and binding specificity. *Biochemistry* **1981**, *20*, 6211−6218.

(15) Dunn. J. F.; Nisula, B. C.; and Rodbard, D. Transport of steroid hormones: Binding of 21 endogenous steroids to both testosterone-binding globulin and corticosteroid-binding globulin in human plasma. *J. Clin. Endocrin. Metab.* **1981**, *53*, 58−68.

(16) 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*, 5959−67.

(17) Kellogg, G. E.; Semus, S. F.; Abraham, D. J. HINT: A new method of empirical hydrophobic field calculation for CoMFA. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 545−52.

(18) Good, A. C.; So, S.-S.; Richards, W. G. Structure−activity relationships from molecular similarity matrices. *J. Med. Chem.* **1993**, *36*, 433−8.

(19) Jain, A. N.; Koile, K.; Chapman, D. Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.* **1994**, *37*, 2315−27.

(20) Mayer, D.; Naylor, C. B.; Motoc, I.; Marshall, G. R. A unique geometry of the active site of angiotensin-converting enzyme consistent with structure-activity studies. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 3−16.

(21) Dammkoehler, R. A.; Karasek, S. F.; Shands, E. F. B.; Marshall, G. R. Constrained search of conformational hyperspace. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 3−21.

(22) DePriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 3D-QSAR of angiotensin-converting enzyme and thermolysin inhibitors: A comparison of CoMFA models based on deduced and experimentally determined active site geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372−5384.

(23) Waller, C. L.; Marshall, G. R. Three-dimensional quantitative structure−activity relationship of angiotensin-converting enzyme and thermolysin inhibitors. II. A comparison of CoMFA models incorporating molecular orbital fields and desolvation free energies based on active-analog and complementary-receptor-field alignment rules. *J. Med. Chem.* **1993**, *36*, 2390−2403.

JM970211N