J-CAMD 366

Objective models for steroid binding sites of human globulins

Jurgen Schnitker, Ramesh Gopalaswamy and Gordon M. Crippen*

College of Pharmacy, University of Michigan, 428 Church Street, Ann Arbor, MI 48109-1065, U.S.A.

Received 19 March 1996
Accepted 31 May 1996

Keywords: QSAR; Voronoi; Receptor mapping; Steroids

Summary

We report the application of a recently developed alignment-free 3D QSAR method [Crippen, G.M., J. Comput. Chem., 16 (1995) 486] to a benchmark-type problem. The test system involves the binding of 31 steroid compounds to two kinds of human carrier protein. The method used not only allows for arbitrary binding modes, but also avoids the problems of traditional least-squares techniques with regard to the implicit neglect of informative outlying data points. It is seen that models of considerable predictive power can be obtained even with a very vague binding site description. Underlining a systematic, but usually ignored, problem of the QSAR approach, there is not one unique type of model but, rather, an entire manifold of distinctly different models that are all compatible with the experimental information. For a given model, there is also a considerable variation in the found binding modes, illustrating the problems that are inherent in the need for 'correct' molecular alignment in conventional 3D QSAR methods.

Introduction

The binding of steroid molecules to corticosteroid- and testosterone-binding human globulin has long been recognized as a challenging test case for the evaluation of new methods related to the identification of quantitative structure–activity relationships (QSARs) [1]. Thus the system has also repeatedly been studied with three-dimensional (3D) QSAR techniques [2], i.e., methods that include an explicit representation of the 3D ligand structure. There have been reports using not only the very popular technique of Comparative Molecular Field Analysis (CoMFA) [3–6], but also the recent and rather elaborate method 'Compass' [7], a similarity matrix method [6], and another CoMFA-type approach [8,9]. Apart from the benchmarking aspect, the steroid–protein system also continues to be of considerable fundamental interest, as there is still no direct structural information on the ligand-binding domains of the corresponding receptors [10].

In this paper, we report on a study of the human steroid–protein system with the recently proposed EGSITE (Energy and Geometry of SITE models) technique. EGSITE differs from other 3D QSAR methods in a number of crucial points. Although it is the eventual outcome of an evolution of techniques that has been labeled the 'Voronoi' approach to binding site modeling [11–13], EGSITE is probably most easily understood independent from its precursors. Its conceptual basis is in the mathematical field of interval analysis, and the relevant formalism has been presented in some detail in the original publication of the method [14]. In the following, we will summarize the essential characteristics of the new method; an in-depth description of the computational algorithm can be found in the original paper.

First, EGSITE leads to binding site models that are objective in the sense that they require no input from the user as to which conformations of the training set molecules are important, nor how the active molecules are to be superimposed at the binding site. This goes beyond even the automatic superposition algorithms employed by the most advanced 3D QSAR methods [7,15,16], which are so critical to their results. In most methods, one begins by determining some kind of molecular superposition, explicitly or implicitly, either by subjective input from the user or by some fixed algorithm. Then the activities of the compounds are correlated with molecular

*To whom correspondence should be addressed.

properties, some of which depend on this fixed alignment. In EGSITE, all sorts of different conformations and superpositions are considered throughout the calculation, resulting in a final 'optimal alignment' for each molecule, which has the property that no other choice of conformation and positioning in the site model produces a more favorable interaction with the site. The initially fixed alignments in other methods do not necessarily share this property. We believe EGSITE derives considerable power from including the positive information that the optimal alignments explain the observed activities, plus the negative information that other alignments do not. The more recent versions of Compass [17] agree with our approach in that the initial alignment (choice of poses, in their terminology) is subject to (a starting point dependent) revision so as to achieve a better fit to the given binding affinities, but they retain the traditional emphasis of superposition of molecules on each other, rather than their positioning relative to the site model. This is still qualitatively different from EGSITE's requirement that the final calculated positioning of each molecule in the site model is energetically optimal and also fits the data.

A second and most important characteristic of EGSITE is that it does not involve some kind of least-squares fitting. The method rather associates an error interval with each and every data point, and EGSITE then attempts to find a set of simultaneous binding modes for all molecules within the given error bars. A single disallowed binding mode is sufficient to invalidate a tentative site model, in stark contrast to conventional fitting procedures. This approach is taken in recognition of the fact that it is typically the *outlying* binding constants that add the most useful information to a given training set. In methods that use least-squares fitting, the outliers are easily swept aside by the abundance of more regular data points, naturally an effect that only gets exacerbated by any redundancy in the data set. By design, EGSITE is not prone to this perpetual problem of traditional fitting methods.

Finally, EGSITE leads to binding site models of deliberate minimum complexity. The description chosen is in terms of a number of convex 'regions', such that every atom of a given molecule in a given binding mode falls into one of the regions. Requiring each region to be geometrically convex simplifies the description of the structure of the site and reduces the combinatorial complexity without significantly constraining the flexibility of the method, especially for such simple site models as in this study. For example, the structural formula of naphthalene consists of two (convex) hexagons sharing a common side, yet the combined figure is not convex. The total number of regions is held as small as possible; for example, only two regions – one genuine binding site region and one solvent region – are sometimes sufficient in the steroid–protein system studied in this paper. The only properties of a region are its convexity, a lower and an upper bound for its diameter, and lower and upper bounds for the distances relative to the other regions. (Chirality relationships also enter if there are at least four binding site regions, see Ref. 11.) The construction of a site model consists of making the diameter and distance bounds just as precise as necessary to explain the experimental binding data. More detail (in the form of additional regions and/or more precise diameter and distance bounds) only emerges as new molecules are added to the training set, or if the binding data are provided with smaller error bars. This minimalist approach avoids the all too common pitfall of QSAR models with numerous adjustable parameters: the models serve to reproduce the data from the training set exceedingly well, but at the cost of diminishing predictive power.

While the previous features of EGSITE clearly depart from those of other 3D QSAR methods, the physicochemical description of the molecules is rather conventional, with atom-specific parameters that may include hydrophobicity, molar refractivity, partial charge, or any other suitable atomic descriptor. A deduced binding site model associates a corresponding set of physicochemical parameters with each of the regions, thus completing the binding site description. By multiplying the physicochemical property of a given atom with the corresponding value for the region that the latter lies in, and by then taking the proper sums over all atoms and over all physicochemical properties, the total binding energy of a given molecule is obtained. An important detail is that in the abstract space of physicochemical parameters a given binding site model is generally not just a point but a polyhedron of finite extent, with its bounds given by a number of nonredundant linear inequalities (see the original publication [14]; also note that this polyhedron in the space of physicochemical parameters should not be confused with the distance intervals that describe a binding site model in *geometrical* space). The result is that in the prediction mode of EGSITE, it is energy *intervals*, rather than unique numbers, that are associated with the binding modes of a given molecule to a given site model.

Most of the actual computational effort in EGSITE is spent on checking for the energetic feasibility of binding modes by solving linear inequalities in the space of physicochemical parameters, carried out again and again while exploring a geometric search tree of binding site models. Having the scaling properties of a combinatorial problem, the total effort grows rapidly with the number of spatial locations per molecule that has to be handled. The number of atoms in typical compounds of medicinal interest may be as large as a hundred, making an exhaustive testing of all possible atom-onto-region mappings infeasible. One way to address this computational bottleneck is the use of a united-atom representation of the molecules. With current computational resources, up to about 10 'superatoms' per molecule can be handled quite easily, and this is also the approach taken here.

A preliminary application of EGSITE to the binding of cocaine analogues to its nerve membrane receptor was presented in the first publication [14]. In the present paper, the emphasis is on an application and assessment of the new method to a benchmark test case. It will be seen that the predictive power of EGSITE is comparable to that of other well-recognized QSAR methods. Contrary to the other methods, however, this performance level is achieved with a unique, unbiased approach that does not rely on a preconceived or predetermined single binding mode for each molecule in the training set. Perhaps the most important aspect of the 'objective' approach taken here is then that it leads to a vivid illustration of some fundamental shortcomings of QSAR methods. We will see that there can be far more than one model that fits the given data, a simple fact that is hardly ever appreciated. Furthermore, we will see that the so-called alignment problem is in fact as serious as it has sometimes been suggested.

In the next section, we first give some technical details of the calculations, including a few developments and improvements of EGSITE since the original publication. We then present in the following two sections the results of applying EGSITE to two systems, involving the binding of 31 steroid test molecules to corticosteroid-binding globulin and testosterone-binding globulin. In both cases, we compare with findings from other QSAR studies of the same systems. In the final section, we present the conclusions.

Methods

We studied the same diverse set of 31 steroids as in the classic paper of Cramer et al. [3] and the more recent papers of Kellogg et al. [4], Good et al. [6], and Jain et al. [7]. The molecular structures shown in Fig. 1 were taken from Ref. 7 (note that there are a number of errors in the structures shown in Refs. 3 and 6). For simplicity, we use exactly the same compound labels as in the papers of Good et al. [6] and Jain et al. [7]. The experimental binding affinities, G_{obs}, that we seek to fit and predict are $-\log K_{diss}$ for the dissociation constants from both human corticosteroid-binding globulin (CBG) [18] and testosterone-binding globulin (TBG) [19], which are exactly the same data used in Ref. 7. We will refer to these values as G_{obs}, as shown in Table 1. The assumed experimental errors and deviations between observed and calculated values always have units on this logarithmic scale.

The energies of molecular conformations were calculated with the modeling package Cerius2 [20] using the MM2 force field option [21] and with missing parameters obtained from an MM2(91) force field file [22] and from the literature [23]. For each steroid, an intermediate set of conformations representative for the total conformational space was chosen with the following iterative procedure. After building and energetically minimizing a given molecule, first the distance geometry program DGEOM [24] was used to generate a set of 100 structures. These 100 structures were then all energetically minimized with Cerius2, and duplicate structures were removed with the program 'Padre' [25]. Then the next iteration was initiated by using the original molecule and DGEOM to generate an additional set of 100 structures, etc.; the entire cycle was repeated until two consecutive iterations did not produce any additional nonduplicate structures.

For each steroid, a final small set of at most three conformations was chosen by only retaining all those conformations from each intermediate set whose energy lay within 2.5 kcal/mol of that of the most favorable one. In 18 out of the 31 cases, this set already comprised no more than either one, two, or three conformations. In the 13 remaining cases (the extreme being steroid **23** with a total of 21 conformations in the intermediate set), a final triple of conformations was selected by calculating the root-mean-square difference between all possible pairs of distance matrices and then finding the particular triple with the largest sum of these differences. The conformational coverage produced by such a set is surprisingly good [26], but it would be interesting to explore the effect of including more conformations, particularly for **23**. In the final count, 10 steroids are described by just one conformation (steroids **2**, **3**, **5**, **9**, **12**, **15**, **16**, **19**, **25**, and **29**), six steroids by two conformations (**4**, **6**, **13**, **14**, **21**, and **26**), and the remaining 15 steroids by three conformations.

In the program EGSITE, a few changes and improvements to the algorithm have been made since the original publication [14]. The first site model region is now explicitly declared to be the solvent region, with physicochemical parameters that are zero and hence an interaction energy with the ligand that always vanishes, since binding affinity of a molecule to the receptor is measured relative to the unbound, solvated ligand. Second, the program can now explicitly address the issue of conformational flexibility by being able to handle ligands in terms of entire sets of rigid conformations, such as described in the previous paragraph. Each conformation in the final set is treated equally, rather than adding the relative conformational energy to the calculated binding affinity, or choosing them with Boltzmann weighted probability.

For the sake of completeness, we will also mention another modification which is a conceptually rather subtle one. The training set determines a set of clustered intersuperatom distances which are used to describe the site geometry [14], so that the sizes and relative positions of the site's regions depend to some extent on the set of molecules under consideration. When the same site was subsequently used for prediction purposes, the test set of

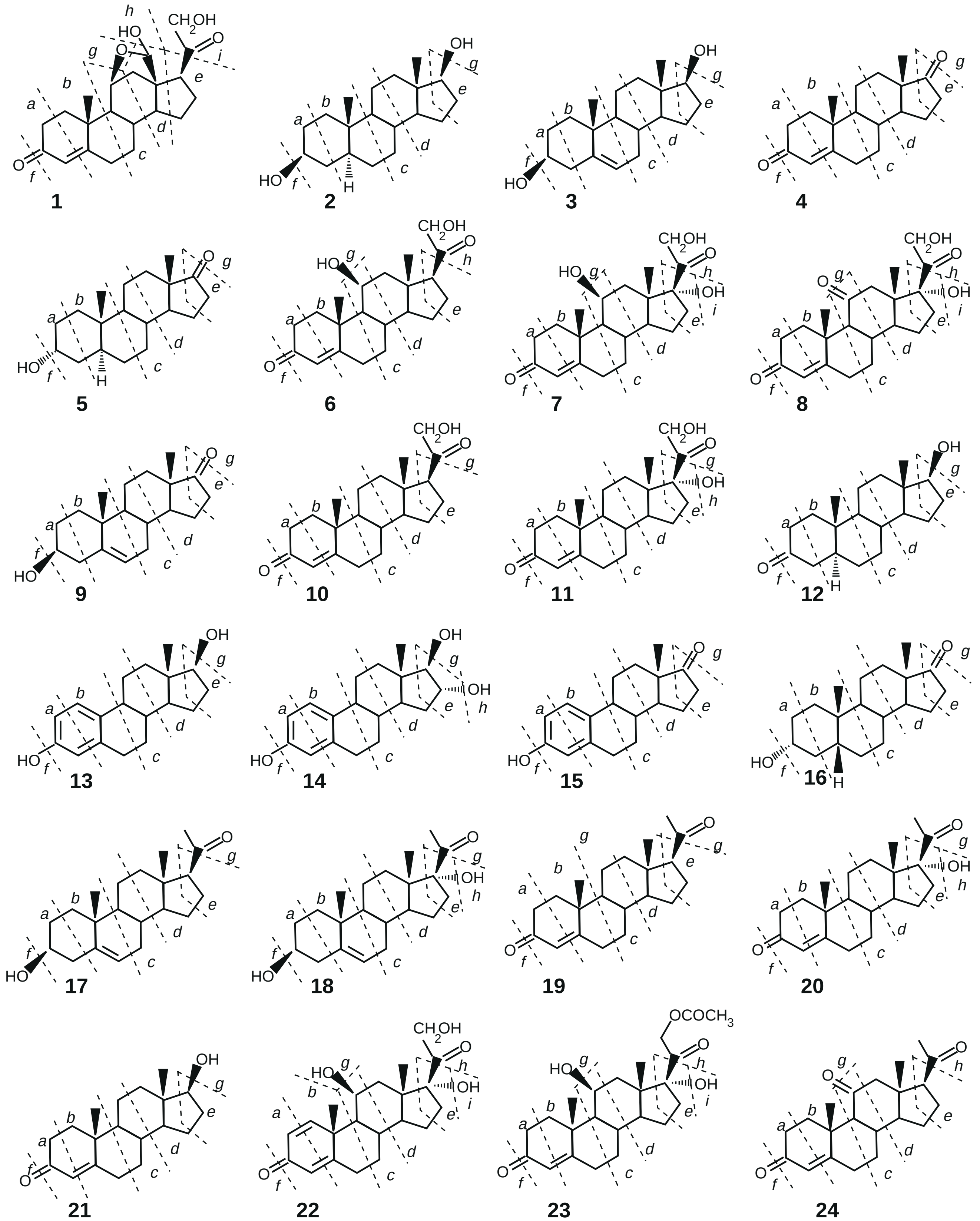


Fig. 1. Superatom grouping of the steroid atoms in the majority of runs. The steroid backbone is partitioned as indicated by the italicized letters *a*–*j*, always giving rise to at least five superatoms. Additional superatoms arise if there are oxygen- or halogen-containing substituents in one or several of the eight indicated positions (methyl substituents are included in the corresponding fused ring superatoms, rather than being counted separately).

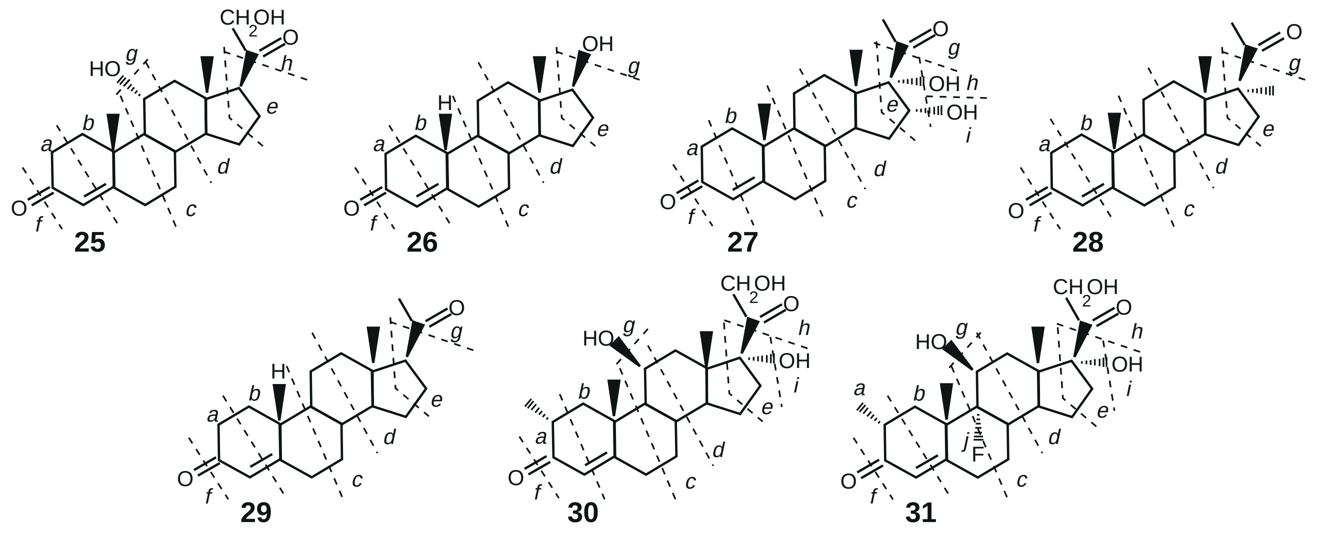


Fig. 1. (continued).

molecules altered the site geometry. Now the training set distances are always carried with the site, regardless of the test set.

The program EGSITE was run with hydrophobicity data, molar refractivity data, and Gasteiger partial charges [27] as provided by the molecular modeling program 'Galaxy' [28]. By using a special Fortran code linked to the Galaxy package, the declaration of superatoms – and all other input file preparation for EGSITE – could be carried out with the molecular display of a graphical user interface. For most runs, the atoms of the steroids were grouped into 7 to 10 superatoms. Five of the superatoms, *a–e*, corresponded to a uniform partitioning of the steroid backbone as shown in Fig. 1. The remaining two to five superatoms, *f–j*, represented the substituents added on to the backbone, as also shown in the figure. In each case, the position of the superatom was defined as the centroid (average of the Cartesian coordinates) of the contributing atomic positions. The physicochemical parameter of a superatom is simply the sum of the corresponding atomic values.

An important aspect of setting up an EGSITE run is the specification of the error bars for the binding affinities of the ligands. If the error bars are large, EGSITE quickly finds a satisfactory site model involving very few regions. As the error bars are decreased, more and more regions are required, involving ever-increasing computer time. Eventually for very small error bars, there may be no site model that explains the input. This can even occur when the given binding data are entirely correct, due to approximations in the model such as imperfect and nonadditive physicochemical parameters. Conceptually, it would seem most straightforward to work with actual experimental error values, but the original experimental papers for this study do not give realistic estimates. Comparing the results for the few compounds treated by two different laboratories leads to an estimate of ±0.5 (see the section on predictions for affinities of other compounds below).

Hence, it is in practice more expedient to initially run EGSITE with a rather wide error interval for the binding affinities, such as $G_{obs} \pm 1.5$. The error interval can subsequently be reduced if the calculation of a more detailed site model is desired and computationally feasible. The specified error intervals for the steroid–protein system in this study are always given in the following sections in connection with the results of a given run. Much of this study is concerned with the effect of error bars on the number of required regions, rather than the single model that results from error bars deduced from an error analysis of the experiments.

In the majority of runs, the training set was the first 21 molecules **1**–**21**, and the test set the remaining molecules **22**–**31**. Some site models were constructed with all 31 molecules **1**–**31** included in the training set. The construction of a two- or three-region site model, for a training set of 21 molecules and with 7 to 10 superatoms and up to three conformations per molecule, will typically take some 5–10 h on a 150 MHz Silicon Graphics workstation. A corresponding prediction run for the remaining 10 molecules adds to this a few more hours.

Binding site models for corticosteroid-binding globulin

The steroid–protein system most extensively benchmarked before [3,4,6,7] is for a set of 21 training molecules and an additional set of 10 test molecules binding to corticosteroid-binding globulin (CBG). Subsequently, we will describe the CBG binding site models that are found with EGSITE, compare their predictive power with those of the previous methods, and draw some prelimi-

TABLE 1

EXPERIMENTAL AFFINITIES FOR THE BINDING OF CBG WITH MOLECULES **1**–**21** AND PREDICTED BINDING INTERVALS (VIA 21-FOLD HOLD-ONE-OUT CROSS-VALIDATION) FOR MODELS C2.1 AND C3.1

Steroid	CBG			TBG
	G_{obs}	G_{pred}		G_{obs}
		C2.1	C3.1[a]	
1	6.279	[6.93,7.10]	[7.67,7.73]	5.322
2	5.000	[5.31,5.49]	[6.08,6.19]	9.114
3	5.000	[5.23,5.41]	[5.65,5.91]	9.176
4	5.763	[6.55,6.68]	[5.82,5.85]	7.462
5	5.613	[5.90,5.97]	[5.43,5.97]	7.146
6	7.881	[5.26,5.46]	[7.12,7.34]	6.342
7	7.881	[7.13,7.30]	[7.12,7.25]	6.204
8	6.892	[8.31,8.33]	[7.10,7.17]	6.431
9	5.000	[6.04,6.26]	[4.64,5.04]	7.819
10	7.653	[6.58,6.67]	[6.73,6.74]	7.380
11	7.881	[6.75,6.91]	[6.87,6.91]	7.204
12	5.919	[5.97,6.02]	[5.16,5.51]	9.740
13	5.000	[4.93,5.13]	[5.72,5.77]	8.833
14	5.000	[5.21,5.29]	[8.57,8.65]	6.633
15	5.000	[5.52,5.60]	[4.46,4.49]	8.176
16	5.225	[5.90,5.97]	[5.46,5.90]	6.146
17	5.225	[5.64,5.85]	[5.49,5.86]	7.146
18	5.000	[6.50,6.52]	[6.35,6.38]	6.362
19	7.380	[6.35,6.40]	[6.58,6.67]	6.944
20	7.740	[6.58,6.68]	[6.97,7.01]	6.996
21	6.724	[5.99,6.05]	[5.92,5.98]	9.204
r^2		[+0.23,+0.35]	[+0.20,+0.28]	
(r^2)*		[+0.57,+0.71]	[+0.63,+0.79]	
(r^2)**		[+0.33,+0.49]	[+0.50,+0.61]	

The r^2 values describe the correlation between all 21 data points from experiment and binding site models. As described in the text, (r^2)* and (r^2)** refer to certain subsets of 18 and 17 data points, respectively.

[a] The three-region cross-validation of molecules **14**, **17**, and **19** was carried out with error intervals for G_{exp} of ±0.95, ±1.05, and ±1.05, respectively, and for all other molecules with an error interval of ±0.9.

nary conclusions concerning the general significance of the results. In addition, we will study the sensitivity of EGSITE results towards three important specifications, namely the choice of physicochemical parameters, the choice of united atoms or 'superatoms', and the representation (or neglect) of conformational flexibility.

Two- and three-region binding site models

For a given training set of molecules, EGSITE models become successively more detailed if the binding affinities are specified with higher accuracy. Conversely, models will generally also become more detailed if, at a given binding constant accuracy, the size of the training set is increased. We will first illustrate the former by discussing the site models that are found if the training set consists of molecules **1**–**21**, with binding constants as given in Table 1.

If the binding constant accuracy is specified as ±1.1, the site models of minimum complexity consist of two regions. For two-region models, an exhaustive search of the solution space of EGSITE is easily carried out, and we find that in this case there is a total of two such models. They are listed in Table 2 as C2.1 and C2.2. The site description is of course extremely vague, but it is still useful as will be seen below. The first region is always the solvent region, which by construction is of infinite extent, i.e., it has a diameter from the interval [∞,∞]. The second of the two regions is the genuine binding site region. The diameter of this region is between 13.4 Å and ∞ for the first one of the two models, and between 12.4 and 13.1 Å for the second model. The site and solvent regions may be at any distance from each other, as the intervals for the corresponding interregion distance in Table 2 are [0,∞] for both models.

The physicochemical parameters w of the regions m, together with the corresponding parameters v of the atoms a in the molecule, lead to the binding energy G according to

$$G = \sum_{a \in m} (v_{hp,a} \cdot w_{hp,a \to m} + v_{mr,a} \cdot w_{mr,a \to m} - v_{ch,a} \cdot w_{ch,a \to m}) \quad (1)$$

where hp refers to hydrophobicity, mr to molar refractivity, and ch to partial charge. Larger positive values of G correspond to stronger binding, and the notation a → m indicates the atom-onto-region mapping of the best binding mode (EGSITE does an exhaustive scan of all binding modes and then counts the best one found). The negative sign in the last term simply accounts for the fact that it is unlike charges that attract each other. Note that the significance of the *absolute* values of the region parameters w is not obvious since the binding energy G is given in some arbitrary logarithmic units and since we do not try to specify prefactors of the sums in Eq. 1. For example, while we take in Eq. 1 the atomic partial charges v_{ch} in multiples of the unit charge, we do not know the unit of the corresponding region parameters w_{ch}. The same is true for the region parameters w_{hp} and w_{mr}. Also note that all site interactions are taken *relative* to interactions with the solvent. For example, an either positive or negative molar refractivity parameter w_{mr} indicates that the site–molecule interaction is either favorable or unfavorable, respectively, relative to polarizability-induced site–solvent interactions.

It can be seen in Table 2 that the physicochemical parameters of both two-region models are indistinguishable within the accuracy stated. In either case, the genuine binding site region has a negative hydrophobicity parameter of −0.2, i.e., it is hydrophilic, and interacts favorably with polarizable parts of the ligands as the molar refractivity parameter is positive with a value of +0.07. The charge parameter is +2.4, suggesting favorable interactions with negative partial charges of the ligands.

The optimal binding modes of the 21 training molecules (not shown) are also the same for both two-region models. The fused ring system in the middle of the molecule always binds to the site, and only superatoms attached to either one or both ‘ends’ of the molecule are exposed to the solvent. Thus, in 14 of the 21 cases hydroxy- or oxo-groups from both ends of the molecule remain unbound, while in the remaining seven cases this is only true for the group labelled *f*. Note that although the energetic parameters and the binding modes of the training molecules are the same for both two-region models, the geometric parameters are nevertheless somewhat different, and hence the predictions of the two models for the binding affinities of other compounds may still differ (see below).

If the error interval of the binding affinities on input drops below ±1.1, the number of regions even in the least complex site models increases from two to three. In Table 2, we list the first three of the three-region models that are found with an error interval of ±0.9, namely models C3.1, C3.2, and C3.3. We actually identified five such models. We do not list the fourth and fifth models in Table 2, however, because they seem to be rather similar to C3.2 and C3.3. It is well possible that there are even more three-region models, but an exhaustive search of the solution space for models of this complexity can be computationally demanding, and was not attempted in this case.

Models C3.1 and C3.2 in Table 2 are very different from each other, and this now extends not only to the geometric but also to the energetic parameters. Thus, for the first model both site regions are hydrophilic, while they are both hydrophobic for the second model. The sign also differs for the molar refractivity parameter of one of the regions. For both models, there is one positively and one negatively charged region, but the magnitudes of the charge parameters again differ significantly. Although C3.2 in Table 2 is very different from C3.1, C3.3 qualitatively resembles C3.2. As mentioned above, the fourth and fifth models (not listed) also seem to be of the same type.

In reflection of the vastly different properties of C3.1 and C3.2, the optimum binding modes also differ considerably. These are shown in Figs. 2 and 3. Sometimes the entire molecule binds to the site, and sometimes only a very few superatoms. Most striking is the fact, however, that not a single molecule shows the same binding pattern for the two models. Furthermore, even within a given model there is a large variation in the apparent orientation of the molecule relative to the binding site. There is clearly no trivial alignment rule for either model. As can be anticipated on the basis of the similarities in the site properties mentioned above, the optimum binding modes of the third to fifth three-region model (not shown) resemble those of the second model. For example, 18 out of 21 binding modes for the third model are the same as for the second model, only molecules **1** and **12** bind with an opposite orientation, and molecule **3** binds in its entirety to one of the genuine site regions.

Increasing the size of the training set by 10 molecules so as to encompass molecules **1**–**31** also leads to increased model complexity. With an error interval for the binding

TABLE 2

GEOMETRIC AND ENERGETIC PROPERTIES OF TWO- AND THREE-REGION[a] SITE MODELS FOR THE BINDING OF CBG WITH MOLECULES **1**–**21**

Site model	Region geometry (Å)			Region energetics		
				Hydrophobicity	Molar refractivity	Charge
C2.1	[∞,∞]	[0,∞]		0	0	0
		[13.4,∞]		–0.2	0.07	2.4
C2.2	[∞,∞]	[0,∞]		0	0	0
		[12.4,13.1]		–0.2	0.07	2.4
C3.1	[∞,∞]	[0,∞]	[0,∞]	0	0	0
		[4.0,∞]	[0,∞]	0.2	0.19	5.3
			[11.0,∞]	–0.1	0.07	–5.4
C3.2	[∞,∞]	[0,∞]	[1.6,∞]	0	0	0
		[11.2,∞]	[0,∞]	3.3	–0.41	21.2
			[13.1,∞]	0.1	0.06	–9.6
C3.3	[∞,∞]	[0,∞]	[1.6,∞]	0	0	0
		[11.2,∞]	[0,∞]	4.2	–0.50	23.7
			[12.6,12.9]	0.1	0.06	–10.4
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Interregion distance intervals are shown as upper triangular matrices, and the physicochemical parameters associated with each region’s row are given at the right.

[a] Two other models were also found that are of the same qualitative type as C3.2 and C3.3; it is possible that there are more three-region models, even of a completely new type.

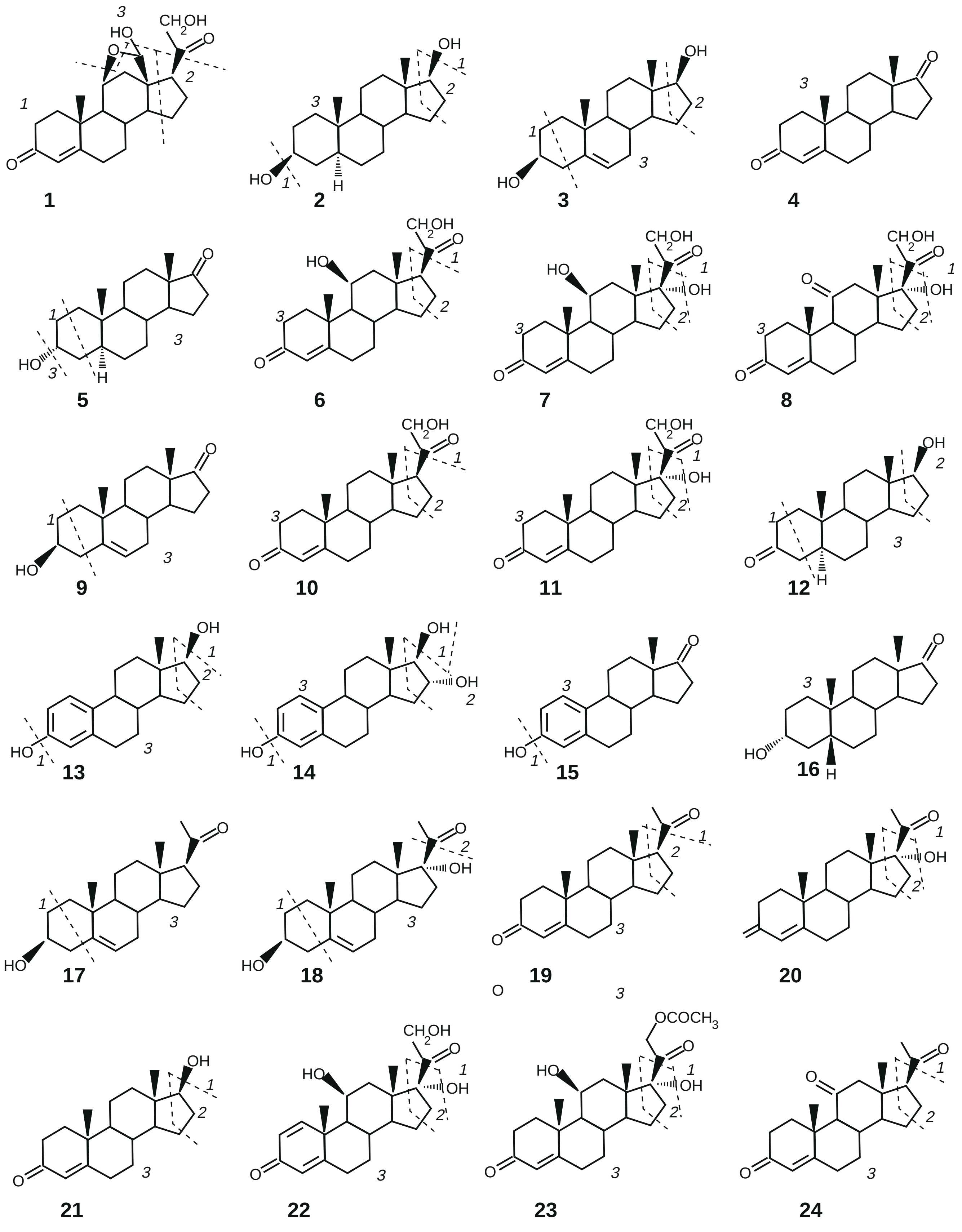


Fig. 2. Optimum binding modes of the 21 steroids in the training set and the 10 steroids in the test set to site model C3.1. Atoms binding in regions 1 = solvent, 2, and 3 are indicated, corresponding to region ordering in Table 2.

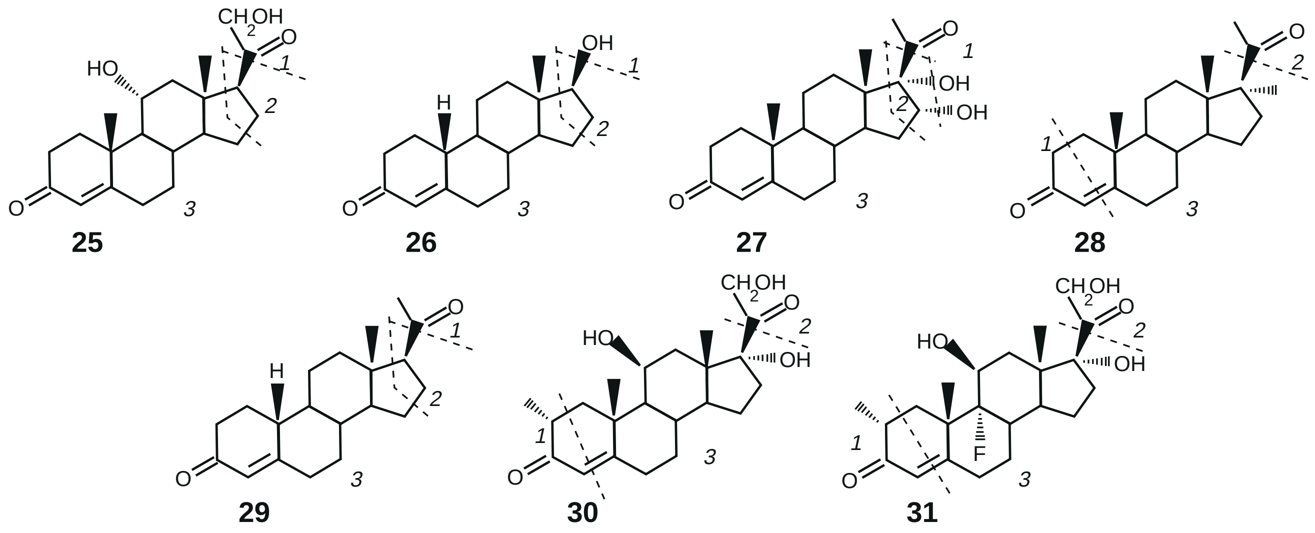


Fig. 2. (continued).

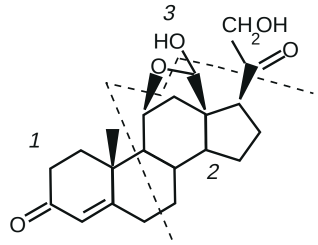
affinities of ±1.1, the site model must become a three-region model, rather than a two-region model as seen with the previous training set of 21 molecules. The first model found is somewhat similar to C3.1, with site hydrophobicity parameters of –0.4 and –0.5, molar refractivity parameters of +0.19 and +0.09, and charge parameters of +4.6 and –2.0. The binding modes are similar too, except that there are now more molecules (10 out of 31, as opposed to previously 2 out of 21) where all superatoms bind to the two site regions. Naturally, one expects that a more complete search with this new training set would again lead to the identification of additional site models.

Internal validation of binding site models

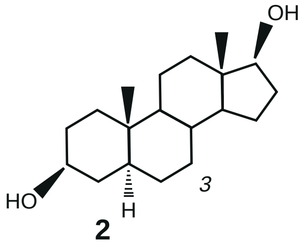
In the previous studies of the same system [3,6,7], 'hold-one-out' cross-validation was used to evaluate the found structure–activity relationships. However, this kind of analysis cannot uncritically be applied to the kind of binding site models found here. As EGSITE does not do any 'fitting' of the data points, its conceptual basis is fundamentally different from that of the more traditional methods, regardless of the fitting technique used by the latter, be it partial least squares [3] or a nonlinear, neural-network-based algorithm [6,7,15]. In the traditional methods, outlying points that are of crucial importance for the construction of the site model just contribute with some small weight to the total result, with the weight basically depending on the number of data points involved. Thus, the omission of any one point will have slight, but not necessarily dramatic, consequences. In EGSITE, however, the significance of crucial outlying points is fully retained in the construction of the site model; hence, even if *any* one of those points is left out, the binding site model is expected to change greatly. The net effect is that cross-validation should lead to seemingly bad results for a few outlying data points. Conversely, the failure of cross-validation for a particular molecule actually provides a very interesting piece of information, namely it indicates the crucial importance of this molecule for the construction of the site model.

Thus, while there are no formal problems with the calculation of cross-validated r^2 values [29] also in our case, the results have to be interpreted very carefully. Specifically, we expect a strong increase in the calculated r^2 value if some molecules are removed from the analysis, namely those that are effectively outliers. Note that this has nothing to do with the removal of 'bad' points, i.e., points that are obviously beyond the reach of the model, that is often carried out while calculating correlation coefficients [3,6,7]. In our case, outlying points in the cross-validation analysis arise because of the crucial importance of the corresponding molecules for the model construction, and *not* necessarily because of inadequacies of the method. If there are any inadequacies of the method as such, the corresponding consequences will show up superimposed on the effect just described.

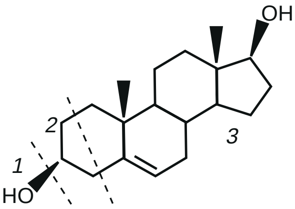
These expectations are borne out by the data. For models C2.1 and C3.1, the predicted binding intervals via cross-validation are given in Table 1, and in the case of C3.1 also in Fig. 4. Correlation of the predicted and experimental affinities, under inclusion of all 21 data points, leads to r^2 values of +0.23 to +0.35 for C2.1 and +0.20 to +0.28 for C3.1, depending on which points within the predicted intervals are chosen. As expected, this performance is worse than that reported for the similarity matrix (0.53), CoMFA (0.69), and Compass (0.89) methods [3,6,7]. However, on removing the three seemingly worst data points (molecules **6**, **8**, and **18** for the two-region model, and **1**, **14**, and **18** for the three-region model), the correlation coefficient rises quickly, with new cross-validated $(r^2)^*$ values of +0.57 to +0.71 and +0.63 to



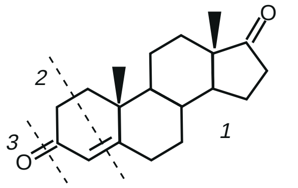
1



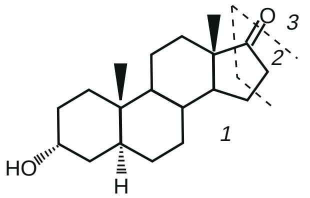
2



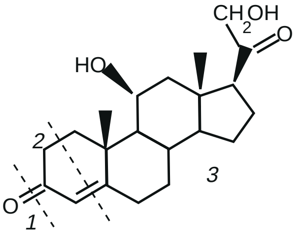
3



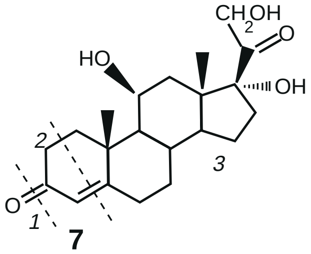
4



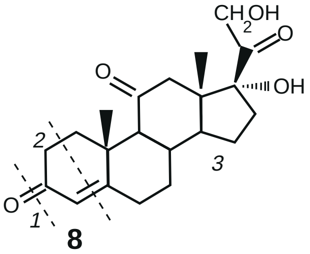
5



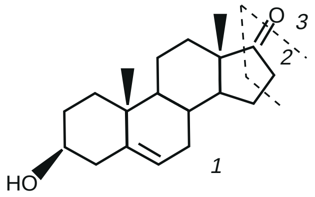
6



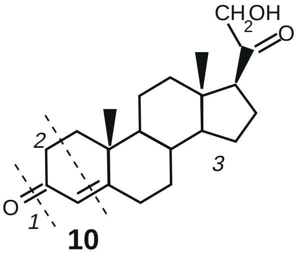
7



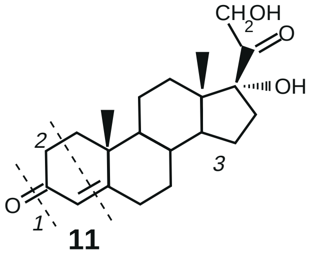
8



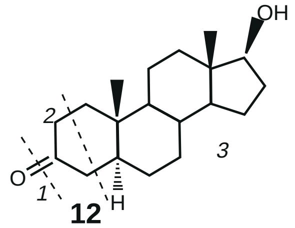
9



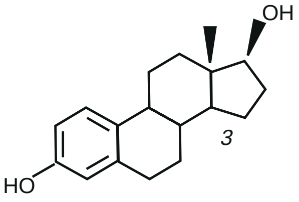
10



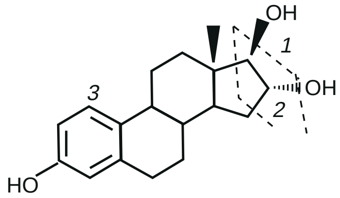
11



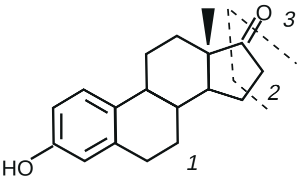
12



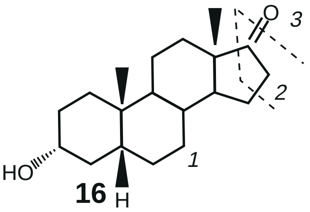
13



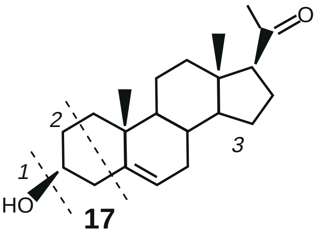
14



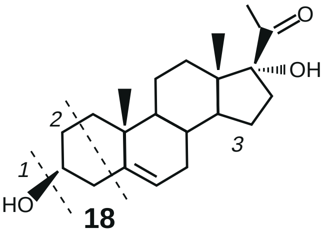
15



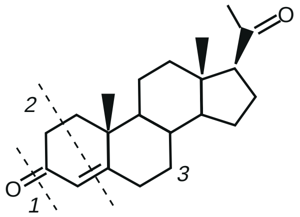
16



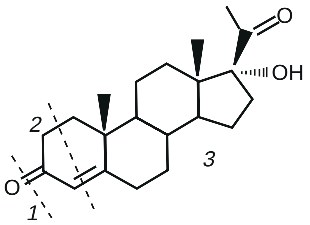
17



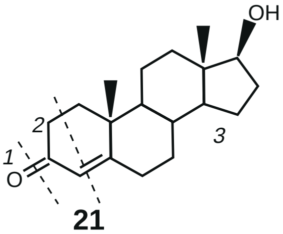
18



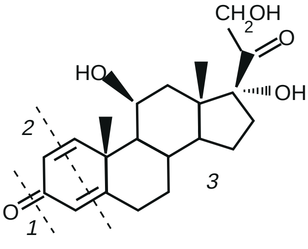
19



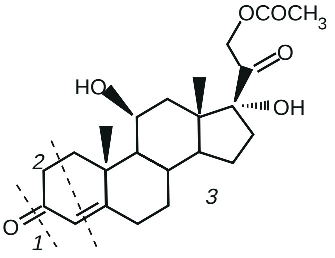
20



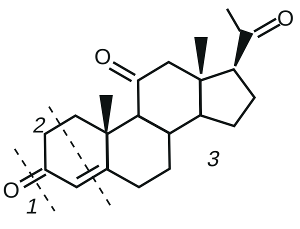
21



22



23



24

Fig. 3. The same as in Fig. 2, except for site model C3.2.

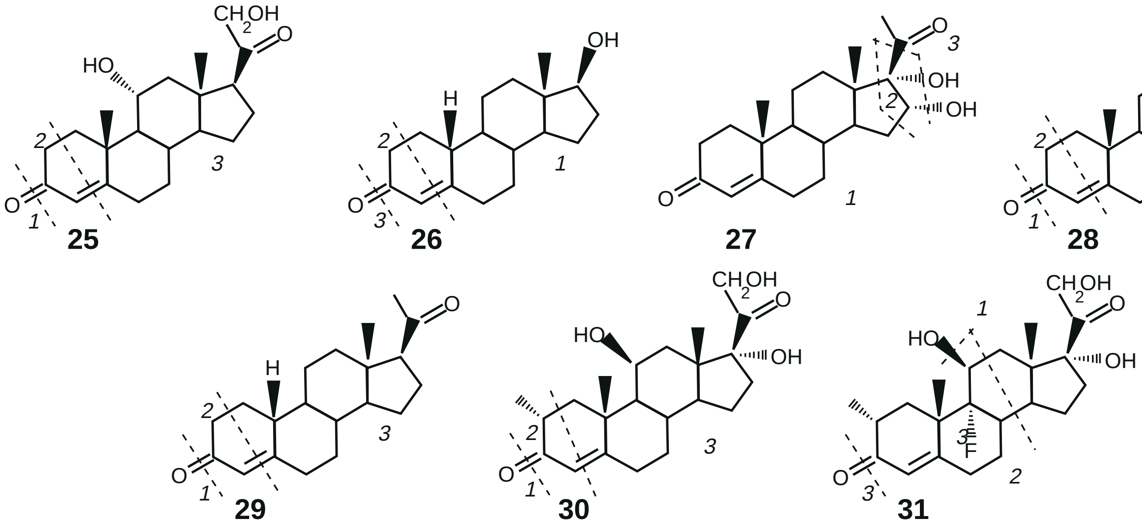


Fig. 3. (continued).

+0.79, respectively. Since we are all so used to dealing with data fitting methods analogous to least squares, this looks like some kind of trick to make EGSITE look good. In fact, as explained above, this is really telling us that these extremely simple site models depend critically on a small subset of the training set, but only slightly on the rest. Given this critical subset upon which the unusual fitting methodology of EGSITE is so dependent, it produces models having 5 to 10 adjustable parameters that give cross-validation scores as good as those for methods involving thousands of adjustable parameters.

Alternatively, on examining the properties of the models generated by the cross-validation procedure we can eliminate those which do not seem to map onto the original site model for the complete 21-molecule training set. Such a lack of mapping may be due to two reasons. First, the site model may genuinely change because of the omission of a very crucial data point, as explained above. In addition, there is the problem that the predicted affinities for the 'hold-one-out' models in Table 1 are always those for the *first* site model found. The algorithm with which EGSITE explores the geometric search tree of site models [14] is, although arbitrary, well defined, making it likely that a given first model will 'map' onto another first model that is found with a very similar training set of mostly the same molecules in the same order. However, there is no rigorous requirement for such a mapping, and any violations introduce additional noise into the data presented in Table 2, noise that is related to the true multiplicity of site models. In any case, upon examination of the physicochemical region parameters, one finds the largest deviations for molecules **8**, **9**, **18**, and **20** for the two-region models, and for molecules **2**, **14**, **17**, and **19** for the three-region models. After removal of the corresponding data points, the cross-validated $(r^2)^{**}$ values are +0.33 to +0.49 and +0.50 to +0.61, respectively.

Thus, we find that the cross-validation analysis of the CBG–steroid system indicates a respectable performance of EGSITE, comparable to that of any of the other methods. As a by-product of the analysis, we have also learned something regarding the identity of the molecules that are of crucial relevance for the construction of the binding site model. Traditional fitting methods do not provide equivalent information in an equally direct way.

Predictions for affinities of other compounds

To assess the predictive power of our very simple 21-molecule binding site models, we also calculated binding affinities for molecules **22**–**31**, as shown in Table 3 and Fig. 5. It should be kept in mind that there is a considerable uncertainty associated with the experimental reference values. Because of the scarcity of independent studies from different laboratories, we will not try to quantify the uncertainty. Based on the little evidence that is available [18,19], we estimate that for some compounds the experimental error bar may be as large as ±0.5.

For the two-region models, most of the prediction ranges are within 0.7 log units of the experimental value. The only clear failure of model C2.1 (Fig. 5a) is the prediction of the fluoride-containing molecule **31** that has also consistently been mispredicted by the other methods [3,6,7]. The failures of C2.2 are the predictions of molecules **23** and **25**. Particularly noteworthy is the fact that while the difference of almost 2 orders of magnitude in the affinities of the very similar molecules **30** and **31** is not reproduced by C2.1, it *is* reproduced by C2.2, in contrast to all the methods in the literature [3,6,7]. Since the physicochemical parameters of both two-region models are virtually the same (see Table 2), this must have to do with steric requirements related to the finite region diameter of the second model. Of course, the effect may be fortuitous and we caution against overinterpreting the result. Its main significance is in illustrating what can be accomplished even with a primitive two-region model.

The performance of the three-region models is also good, with the qualification that the pronounced difference between the affinities of compounds **30** and **31** is not correctly reproduced by any of the models. Also molecules **27** and **28** are significantly mispredicted by model C3.1 (Fig. 5b), and molecule **27** is somewhat mispredicted by C3.2 (Fig. 5c) and C3.3. Interestingly enough, C3.2 and C3.3 lead to very similar predictions even though their properties differ more than those of the two two-region models. Apparently, the similarity or dissimilarity of site models does not always indicate in a straightforward way whether predictions are going to resemble each other or not.

As already emphasized by Jain et al. [7], one would also like to evaluate the overall quality of the predictions with a measure that directly relates to the usefulness of a QSAR method in drug design where the central problem is in deciding on which compound to study next. A nonparametric correlation coefficient that quantifies the *rank* correlation of experimental and calculated values [30] is an adequate tool for this purpose, and we follow the example of Jain et al. [7] and calculate Kendall's τ measure. The interpretation of Kendall's τ is particularly straightforward since it has a value of +1 for perfect correlation between the rankings of experimental and calculated values, and a value of −1 for complete anticorrelation [30]. Usually τ is calculated for two lists of numbers, G_{obs} versus G_{pred}, but in our case the latter are intervals. If we think of the G_{pred} for each molecule as three possible numbers (the lower bound of the interval, the upper bound, and the midpoint), then for the 10 molecules in the test set, the 10 G_{obs} are compared with all $3^{10} = 59\,049$ lists of different possible combinations of G_{pred} values, resulting in a range of τ's. Therefore we report τ intervals corresponding to the minimum and maximum of the 59 049 τ values.

Using the predicted affinity intervals from Table 3, we find $\tau = [+0.24,+0.42]$ for C2.1 and $\tau = [+0.07,+0.38]$ for C2.2. For the three-region models, the ranges are [+0.02, +0.24] for C3.1, [−0.11,+0.38] for C3.2, and [−0.16,+0.69] for C3.3. All these figures are well within the range of performance of the CoMFA (+0.28) and Compass (+0.46) methods [3,7], illustrating the practical value of EGSITE for drug design purposes.

General significance of prediction results

The evident value of EGSITE as a predictive tool is naturally most welcome. However, there are two other conclusions concerning the nature of QSAR methods in general, whose significance cannot be overemphasized. First, we recall that the number of adjustable parameters in our models is almost negligible compared to that used by any of the other methods, be it CoMFA, Compass, or the similarity matrix method. The extreme case is provided by the two-region models which may be somewhat inferior to the three-region models, but are still valuable, and if nothing else they vividly illustrate what could literally be called the austerity of the method. The two-region models contain no more than *six* parameters (the three physicochemical parameters of the site region, the lower and upper bounds of the site diameter, and the lower bound of the site–solvent distance) whose values are determined by the algorithm (the few remaining parameters are fixed at values of either 0 or ∞). This has to be

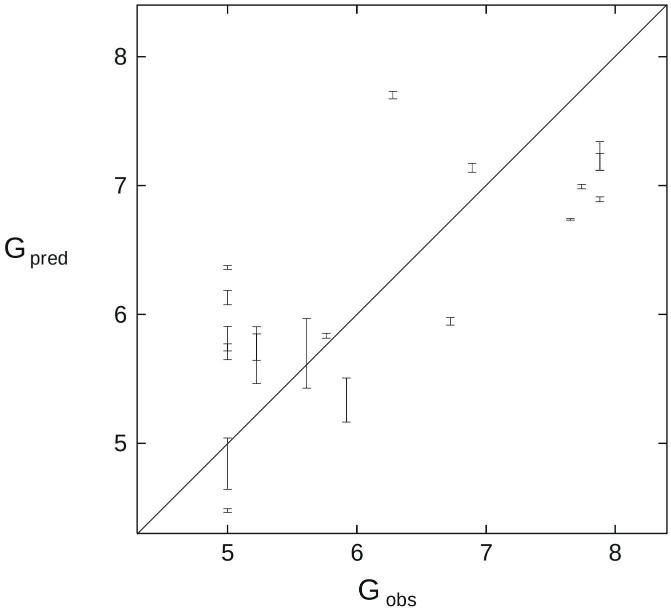


Fig. 4. Measured CBG binding affinities for molecules **1**–**21** versus binding intervals (vertical bars) as predicted by model C3.1 via 21-fold cross-validation.

TABLE 3

EXPERIMENTAL AFFINITIES FOR THE BINDING OF CBG WITH MOLECULES **22**–**31** AND PREDICTED BINDING INTERVALS FROM VARIOUS TWO- AND THREE-REGION BINDING SITE MODELS (SEE TABLE 2) THAT WERE OBTAINED WITH MOLECULES **1**–**21** AS THE TRAINING SET

Steroid	G_{obs}	G_{pred}				
		C2.1	C2.2	C3.1	C3.2	C3.3
22	7.512	[7.21,7.37]	[7.20,7.39]	[7.34,7.41]	[7.09,7.44]	[7.05,7.63]
23	7.553	[7.71,7.87]	[5.79,6.16]	[7.23,7.31]	[7.57,7.87]	[7.59,7.89]
24	6.779	[7.26,7.45]	[7.26,7.45]	[6.95,7.03]	[6.83,7.02]	[6.80,7.00]
25	7.200	[7.00,7.09]	[5.26,5.46]	[7.09,7.16]	[7.00,7.21]	[6.99,7.20]
26	6.114	[5.75,5.81]	[5.75,5.80]	[6.65,6.71]	[6.46,6.56]	[6.46,6.59]
27	6.247	[6.93,7.00]	[6.93,7.01]	[7.37,7.45]	[7.05,7.48]	[7.08,7.95]
28	7.120	[6.53,6.61]	[6.54,6.61]	[5.23,5.63]	[7.17,7.47]	[7.14,7.43]
29	6.817	[6.09,6.14]	[6.09,6.14]	[6.52,6.58]	[6.57,6.86]	[6.55,6.88]
30	7.688	[7.41,7.51]	[7.37,7.52]	[6.25,6.68]	[6.65,7.06]	[6.66,7.21]
31	5.797	[7.39,7.48]	[5.71,5.87]	[6.23,6.62]	[7.57,8.13]	[7.07,8.11]
τ		[0.29,0.42]	[0.07,0.38]	[0.02,0.24]	[–0.11,0.38]	[–0.16,0.69]

compared with the several parameter values that are associated with each of the hundreds (Compass, similarity method) or even thousands (CoMFA) of grid points or matrix elements that are present in the traditional fitting algorithms. We are led to conclude that the detailed site models generated by the other methods may be vastly overdetermined, in ways that their statistical tools, such as partial least squares, are unable to counteract. For example, partial least squares in a CoMFA analysis can select a very low dimensional, statistically significant subspace out of the whole parameter space, but even this outlines a rather detailed spatial picture of regions and substituents that appears to determine binding affinity.

Second, we have found that the first and second of the three-region models for the CBG system are very different and still have comparable predictive power. We have no reason to doubt that this is a typical result: for a training set of just one or two dozen compounds, there are a *multitude* of site models that are all ‘true’ in the sense that none of them can be eliminated without access to additional information. This is an aspect that all too easily gets lost in the presentation and application of the sophisticated traditional methods where, for whatever reasons, there is only *one* outcome of the algorithm, and accordingly only one binding site model. The conceptual or algorithmic limitations of a given method should not be confused with the suggestion that there is only one model that is compatible with the experimental information.

Dependence on the choice of physicochemical parameters

The identification of the actually relevant physicochemical parameters is a crucial step in the formulation of any QSAR model. For example, the importance of the steric, electrostatic, and other fields in CoMFA has been studied in several steroid–protein systems [4–6], and systematic studies along these lines have also been carried out for the Compass and similarity matrix methods [6,7].

In our approach, each binding site model is associated with a polyhedron in the space of physicochemical parameters, as mentioned above and described elsewhere [14]. If there are any redundant parameters, this polyhedron will have a large diameter in some directions. Even if there are no redundant parameters, the issue of possible weighting factors in Eq. 1 still arises. The latter unfortunately requires deciding on the factors that convert from physicochemical parameters to absolute energies, and this intricate problem will not be addressed here. The more basic redundancy problem, however, can easily be studied by calculating binding site models under systematic omission of one or several of the physicochemical parameters.

While we generally find that all three parameters – molar refractivity, Gasteiger partial charge [27], and hydrophobicity – are relevant, omission of the molar refractivity parameter clearly has the most pronounced effect. If we search for a CBG site model with only the partial charge and hydrophobicity parameters (as before with molecules **1**–**21** as the training set; the error interval for the binding affinities is ±1.5), a rather strange two-region site model with a hydrophobicity parameter of −5.6 and a charge parameter of −18.4 is found. Only a single superatom of each training molecule binds to the site region. The model is practically useless for the prediction of the affinities of molecules **22**–**31**.

In CoMFA studies of the CBG system, hydrophobic interactions were found to be unimportant and leaving out the electrostatics was actually seen to improve the performance of the method [4,6]. Our corresponding observations with EGSITE are not quite that pronounced, but they also do not literally contradict the older findings. Thus even under removal of either the partial charge or the hydrophobicity from the original list of parameters, useful site models can still be identified, and their predictive power is not noticeably diminished. In the first case, without the Gasteiger partial charge and

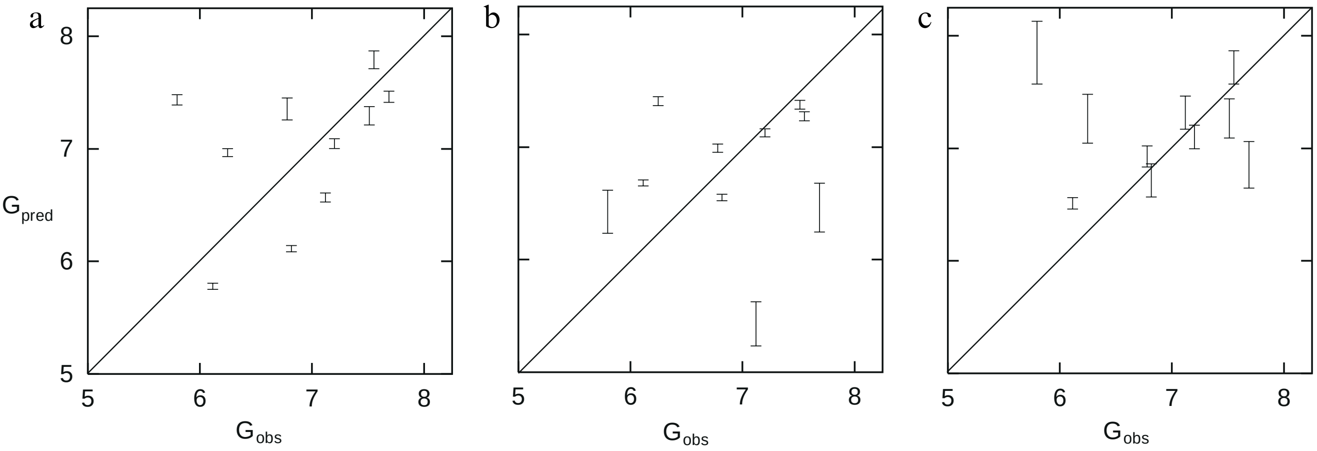


Fig. 5. Measured CBG binding affinities for molecules **22**–**31** versus binding intervals (vertical bars) predicted from (a) site model C2.1; (b) site model C3.1; and (c) site model C3.2.

with a binding affinity uncertainty of ±1.25, we find a plausible three-region site model with hydrophobicity parameters of +1.6 and –0.8 and molar refractivity parameters of +0.06 and +0.10. The binding modes of the training molecules are rather diverse, and the affinities of compounds **22**–**31** are predicted with a Kendall's τ = [–0.16,+0.42]. Similarly, omission of the hydrophobicity parameter (with an affinity error interval of ±1.25) leads to a two-region site model with a molar refractivity parameter of +0.06 and a charge parameter of +1.9. In this case, the affinities of molecules **22**–**31** are at least satisfactorily predicted with τ = [+0.11,+0.29].

These findings are underlined if we attempt to construct site models employing a single physicochemical parameter. The molar refractivity is by far the best choice. With it as the only descriptor and with an affinity uncertainty of ±1.25, a three-region site model with parameters of +0.05 and +0.08 and a surprisingly high predictive power can be identified (the affinities of molecules **22**–**31** are predicted with τ = [+0.29,+0.38]). No comparable success is seen when the only parameter is either the partial charge or the hydrophobicity. With affinity error intervals of ±1.35 and ±1.5, respectively, three-region site models are found with charge parameters of +8.1 and +26 in the first case and hydrophobicity parameters of +4.1 and –1.3 in the second case. The affinities of molecules **22**–**31** are predicted with τ values of only [–0.56,+0.07] and [–0.02,+0.07], respectively.

It should be pointed out, however, that the affinities of the most and the least active of the training molecules **1**–**21** are separated by no more than about 3 log units. Hence with fairly large affinity error intervals, such as ±1.25 or ±1.5, there is a certain lack of separation into active and inactive compounds, and site models may exist that have a more or less constant energy term for all training molecules. This is more easily accomplished with simple one-parameter models if the given *atomic* physicochemical parameter always has the same sign (and does not vary too widely). This is precisely the case for the molar refractivity parameter, and may partly explain the observations described above.

In general it should be kept in mind that previous studies suggest a great variation of the parameter dependence from one steroid–protein system to another. In systematic CoMFA studies, for example, it was seen that use of a steric field only gives the best results with the CBG and progesterone receptors [5,6], while use of an electrostatic field only works best with TBG [6] and the androgen receptor [5]. Kellogg et al. [4] find that for the CBG system basically any combination of fields works well as long as it includes a steric and/or electrostatic field. Regardless of the validity of these old results, it seems prudent to be wary of generalizations regarding the relevance of selected physicochemical parameters in steroid–protein systems.

Dependence on the choice of superatoms

Most of the site models in this paper were calculated with a grouping of the atoms in molecules **1**–**31** into either seven (16 cases), eight (seven cases), nine (seven cases), or 10 (molecule **31** only) superatoms. Since the computational requirements of EGSITE grow rapidly with the number of molecular centers, it is clearly of interest to determine if a smaller number of superatoms could possibly suffice. We have therefore also undertaken runs where all 31 molecules were grouped into either five, three, two, or just one superatom(s). In the last case, the molecules, of course, no longer have any internal structure, and finding a site model effectively corresponds to a classic Free-Wilson analysis of the CBG system.

A minor complication is the fact that the inclusion of the atomic partial charge among the physicochemical parameters becomes increasingly meaningless if the number of superatoms becomes smaller and smaller, simply because the net charges of the larger and larger molecular fragments start to vanish. In practice, it is even worse as the charges of the extended superatoms of the overall electroneutral molecules essentially become random numbers, because of the arbitrariness of the algorithm by which they are defined. We therefore left the partial charge out in all runs with fewer than the original 7 to 10 superatoms.

It is seen that with five superatoms the performance of EGSITE in the CBG system is somewhat worse than before, but the results are not yet useless. With an accuracy of ±1.1 for the affinities of molecules **1**–**21**, already a three-region site model is obtained, rather than a two-region model as found with 7 to 10 superatoms. The hydrophobicity parameters are –0.2 and 0.0, and the molar refractivity parameters are +0.01 and +0.12. The affinities of molecules **22**–**31** are predicted with $\tau = [-0.02,+0.20]$, the main failure being the underpredicted affinity of compound **30**.

If only three superatoms are used to describe each of the molecules, the performance is much worse. With an affinity uncertainty of ±1, again a three-region model is found, but this model now mispredicts 4 of the 10 affinities in the test set of molecules **22**–**31**, and τ for the predictions is [–0.24,–0.20].

With two superatoms and a binding affinity uncertainty of ±1, no binding site model with up to four regions exists. Similarly, there is no site model with up to five regions if the now structureless molecules consist of just one superatom and the affinity uncertainty is again ±1.

Dependence on accounting for conformational flexibility

Since steroids with their four fused rings are generally rather rigid molecules, it is not immediately clear that conformational flexibility is necessarily a very sensitive issue for the CBG system studied here. It still has to be recognized that some molecules have rotatable side chains, in particular steroid **23**. In the sophisticated Compass method, the so-called pose selection is the most crucial step of the calculation, and it has thus been claimed that the conformational selection of the molecules is extremely important [7]. Unfortunately, Compass intertwines the two separate issues of binding site alignment and conformational description. This is not the case in EGSITE, where the alignment problem does not come up at all, and where the conformational issue is addressed in a very direct way by including as many conformations of a given molecule as desired.

Up to this point, the construction of all site models was undertaken with up to three conformations per molecule, and one of the multiple conformations was always the lowest energy conformation. To examine the actual importance of the issue of conformational flexibility, we have also carried out a variety of runs with a sparser description of the conformational space, namely by either including only the lowest energy conformation for each molecule or by essentially picking conformations at random.

We find, perhaps not surprisingly, that it becomes increasingly important to represent some molecules by multiple conformations as the complexity of the site model increases. Thus two-region models are less affected than three-region models. If we only include the lowest energy conformation for each molecule and work with a binding affinity uncertainty of ±1.1, we can calculate a two-region model whose geometrical and physicochemical parameters are almost the same as those for the first model in Table 1. Kendall's τ for the predictions of molecules **22**–**31** even seems improved to [+0.33,+0.42], but this may be a fortuitous result.

Even for two-region models, the situation gets somewhat worse if we replace the lowest energy conformation

TABLE 4
GEOMETRIC AND ENERGETIC PROPERTIES OF SOME SITE MODELS THAT WERE FOUND FOR THE BINDING OF TBG WITH MOLECULES **1**–**21**

Site model	Region geometry (Å)			Region energetics		
				Hydrophobicity	Molar refractivity	Charge
T2.1	[∞,∞]	[0,∞]		0	0	0
		[7.0,9.4]		0.3	0.16	–2.9
T3.1	[∞,∞]	[0,∞]	[0,∞]	0	0	0
		[4.0,∞]	[0,∞]	3.8	0.90	2.5
			[9.6,∞]	–1.4	–0.01	27.0
T3.2	[∞,∞]	[2.7,∞]	[0,∞]	0	0	0
		[7.0,∞]	[0,∞]	0.2	–0.01	–20.9
			[5.6,∞]	0.9	0.23	–17.4
T3.3	[∞,∞]	[4.6,∞]	[0,∞]	0	0	0
		[7.2,∞]	[0,∞]	1.5	–0.04	–26.7
			[5.6,∞]	1.0	0.22	–18.7
T3.4	[∞,∞]	[4.8,∞]	[0,∞]	0	0	0
		[6.2,∞]	[0,∞]	–5.2	–0.47	65.4
			[6.6,∞]	2.9	0.12	34.0

TABLE 5
PREDICTED INTERVALS FOR THE BINDING OF TBG WITH MOLECULES **22**–**31**, USING SITE MODELS FROM TABLE 4

Steroid	G_{pred}				
	T2.1	T3.1	T3.2	T3.3	T3.4
22	[7.06,7.22]	[6.55,6.94]	[6.45,7.53]	[5.67,7.40]	[7.06,7.11]
23	[5.83,5.86]	[5.19,6.41]	[6.30,7.54]	[5.35,7.55]	[6.76,6.81]
24	[8.71,8.79]	[7.58,7.84]	[4.99,6.93]	[4.96,6.96]	[10.87,10.96]
25	[5.83,5.86]	[6.46,6.80]	[6.65,7.82]	[5.22,7.83]	[6.76,6.81]
26	[8.04,8.09]	[8.27,8.74]	[8.25,8.43]	[8.26,9.05]	[9.04,9.16]
27	[8.66,8.71]	[7.66,7.86]	[8.06,9.96]	[8.01,10.01]	[6.67,6.72]
28	[8.66,8.71]	[7.59,7.83]	[5.36,5.72]	[5.34,5.73]	[10.14,10.22]
29	[6.43,6.46]	[7.58,7.84]	[7.82,8.33]	[7.80,8.43]	[8.23,8.28]
30	[7.06,7.22]	[6.55,6.94]	[6.45,7.54]	[5.67,7.40]	[4.71,5.86]
31	[6.16,6.53]	[6.55,6.94]	[6.37,7.42]	[5.67,7.30]	[5.51,6.43]

The compounds have not been assayed experimentally.

by one conformation picked at random from the intermediate pool of conformations described in the Methods section. The two-region site model now found (again with an affinity error of ±1.1) still has very similar geometric and energetic parameters, but there are more mispredictions for the affinities of molecules **22**–**31**. The mispredictions include molecule **23**, and τ drops to [–0.07,+0.16]. Once we represent each molecule by a set of several randomly picked conformations, the predictive power of the then found two-region model is again improved.

For three-region models, the predictions are already dramatically worse if each molecule is given by only its lowest energy conformation. Under the latter conditions and with a binding affinity uncertainty of ±1, we get a three-region model that exhibits a huge hydrophobicity parameter of +95 for one of the regions. The affinity of compound **31** is mispredicted by 27 orders of magnitude, and most of the predictions of the other compounds in the test set are also very unsatisfactory.

Binding site models for testosterone-binding globulin

We have also determined some binding site models for TBG, using the same training set of 21 steroid molecules as in the previous section and experimental affinities as listed in Table 1. Working with affinity error intervals of ±1.7 and ±1.5, we found one two-region model and five three-region models, respectively, listed in Table 4. Just as in the CBG case, the search of the solution space for the three-region models was not exhaustive, and thus there may be additional unidentified site models. (The second three-region model is not included in Table 4 because it happens to be essentially the same as the first model, with only the labels of the two genuine binding sites swapped. What is shown are the remaining four three-region models.)

An examination of the energetic region properties in Table 4 shows that there seems to be even more variety and fragmentation in the solution space of this system than in the CBG case. In particular, there is no pair of three-region models that would obviously be in the same class. There is, on the other hand, less variation in the binding modes (not shown) of the individual molecules in the various site models. For all site models and for almost all compounds, the middle part of the molecule (superatoms *a*, *b*, *c*, and *d*) is exposed to the solvent. It is almost always the ends of the molecule (superatom *f* and superatom *e*, along with *e*'s substituents) that bind to the genuine site regions.

The TBG binding affinities of molecules **22**–**31** have not yet been assayed experimentally. The corresponding EGSITE predictions in Table 5, calculated with the two- and three-region models from Table 4, can thus not be evaluated at this time. Since the aim of this work was primarily to compare the performance of our method with other approaches, we have focussed on the better-studied CBG data. Therefore, cross-validation calculations for our TBG sites have not yet been performed. Since the fitting of the TBG data caused no special difficulties, we expect prediction results comparable to what we found for CBG. For all three-region models, molecule **26** is predicted to bind more strongly than most of the other test molecules, but there are also cases such as that of molecule **24** which is at the same time the least active compound for the third three-region model in the table and the most active one for the fourth model. This is a prime example of how drastic the ramifications of the multiplicity of site models can be.

Conclusions

We have shown for a benchmark test case that binding site models can be found without any form of subjective molecular alignment. Superficially speaking, the same can be said for Compass [7,15] with its automated pose selection, for example. The distinction is that EGSITE not only calculates a predicted binding mode for each molecule, but ensures that this mode has a better calculated binding strength than any other mode. This optimality condition puts a great constraint on possible solutions while retaining this aspect of physical reality in the model. In contrast, an automated alignment procedure must somehow make an equivalent choice without knowing whether the subsequently derived model really favors its alignment.

Our approach to data fitting is more reminiscent of linear programming than of linear regression. Our results may depend critically on a few of the compounds, rather than the more uniform weighting seen in other methods. This has the disadvantage that a single erroneous compound in the training set will lead to incorrect site models. If, however, all the input data are correct, the addition of a single new compound can lead to a dramatic change in the site models, just as one new observation sometimes leads to a dramatic shift in one's thinking. Stated this way, EGSITE's behavior seems appropriate, but it does make it difficult to assess our results with traditional statistical measures, such as correlation coefficients and cross-validation.

The choice of physicochemical descriptors, the specification of the superatoms, and the representation of the conformational space remain as issues that need to be approached with some care. Except for the need to use superatoms, these problems of course also arise in any other 3D QSAR method. With the specifications used here, we obtain a very reasonable performance of the method, but further optimization should certainly be possible.

One shortcoming of the current implementation of EGSITE is the fact that site models have to be identified one at a time by a systematic exploration of the tree of solutions. Although the ad hoc search algorithm is rather sophisticated and puts strong emphasis on the exploration of 'promising' branches of the search tree [14], the generation of multiple solutions can still be computationally demanding. Ideally one would want to obtain in a direct way a representative sample of all the existing binding site models, rather than having to tediously explore the solution tree in a systematic manner. This could theoretically be accomplished if the systematic exploration of the search tree was replaced by a random walk in solution space, preferably in a parallel sense along multiple strands and using a genetic algorithm. An implementation of this approach is currently being developed (Crippen, G.M., work in progress).

The most important lesson of this study is not so much the quantitative performance of EGSITE versus other methods on these standard test sets, but rather the fact that relaxing some implicit assumptions reveals how drastic these assumptions are:

(1) 3D QSAR methods generally focus on finding one optimal superposition of the molecules. When we instead consider all different binding modes, we find many different choices for a single site geometry leading to equally good fits to the training set but sometimes quite different predictions for test molecules. Having seen this in so many different cases, we conclude this is a general phenomenon that is of course never observed by methods that assume there is only one best superposition.

(2) If an algorithm seeks the optimal explanation for the given binding data, it will never explore alternative explanations of equal quality. We, instead, find several distinct site geometries that each explain the observed binding, up to the given accuracy. It is disturbing that we cannot choose among these different possibilities without resorting to more experimental input, but that is better than the false security of finding exactly one 'best' solution.

(3) We are aware of no other 3D QSAR method that systematically explores different levels of geometric detail. For example, a CoMFA analysis at a given grid spacing carefully reduces the energetic degrees of freedom (and hence detail) via PLS, but the fixed superposition determines once and for all the composite molecular envelope at a rather high resolution. This is equivalent to a regression fitting of data that reduces one large set of (energetic) variables to a single linear combination while leaving hundreds of (geometric) variables in 10th-order polynomials. Leaving one compound out of the training set still gives a very complicated geometric picture having energetic features sufficiently stable that the deleted compound is reasonably predicted. However, even good cross-validation results in a framework that always has great geometric detail cannot prove that the detail is essential. EGSITE demonstrates that all this careful alignment and resulting protrusions and hollows are largely irrelevant, since it can fit the same data to comparable accuracy with the most primitive site shapes.

Acknowledgements

Partial support of this work by an Upjohn research award is gratefully acknowledged.

References

1 Stouch, T.R. and Jurs, P.C., J. Med. Chem., 29 (1986) 2125.

2 Kubinyi, H. (Ed.) 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, The Netherlands, 1993.

3 Cramer, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.

4 Kellogg, G.E., Semus, S.F. and Abraham, D.J., J. Comput.-Aided Mol. Design, 5 (1991) 545.

5 Loughney, D.A. and Schwender, C.F., J. Comput.-Aided Mol. Design, 6 (1992) 569.

6 Good, A.C., So, S.-S. and Richards, W.G., J. Med. Chem., 36 (1993) 433.

7 Jain, A.N., Koile, K. and Chapman, D., J. Med. Chem., 37 (1994) 2315.

8 Norinder, U., J. Comput.-Aided Mol. Design, 4 (1990) 381.

9 Norinder, U., J. Comput.-Aided Mol. Design, 5 (1991) 419.

10 Carlstedt-Duke, J., Nilsson, L. and Norinder, U., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, The Netherlands, 1993, pp. 373–385.

11 Crippen, G.M., J. Comput. Chem., 8 (1987) 943.

12 Bradley, M.P. and Crippen, G.M., J. Med. Chem., 36 (1993) 3171.

13 Srivastava, S. and Crippen, G.M., J. Med. Chem., 36 (1993) 3572.

14 Crippen, G.M., J. Comput. Chem., 16 (1995) 486.

15 Jain, A.N., Dietterich, T.G., Lathrop, R.H., Chapman, D., Critchlow Jr., R.E., Bauer, B.E., Webster, T.A. and Lozano-Perez, T., J. Comput.-Aided Mol. Design, 8 (1994) 635.

16 Hahn, M. and Rogers, D., J. Med. Chem., 38 (1995) 2091.

17 Jain, A.N., Harris, N.L. and Park, J.Y., J. Med. Chem., 38 (1995) 1295.

18 Mickelson, K.E., Forsthoefel, J. and Westphal, U., Biochemistry, 20 (1981) 6211.

19 Dunn, J.F., Nisula, B.C. and Rodbard, D., J. Clin. Endocrin. Metab., 53 (1981) 58.

20 Cerius2, Molecular Simulations Inc., Burlington, MA, U.S.A., 1994.

21 Allinger, N.L., J. Am. Chem. Soc., 99 (1977) 8127.

22 MM2(91), Allinger, N.L., University of Georgia, Athens, GA, U.S.A., 1991.

23 Lipkowitz, K.B., QCPE Bull., 12 (1992) 6.

24 DGEOM, Blaney, J.M., Crippen, G.M., Dearing, A. and Dixon, J.S., Copyright DuPont Corporation, 1990; QCPE Program 590, Indiana University, Bloomington, IN, U.S.A.

25 'Padre', Stahl, M. and Walters, P., University of Arizona, Tucson, AZ, U.S.A., 1995.

26 Smellie, A., Kahn, S.D. and Teig, S.L., J. Chem. Inf. Comput. Sci., 35 (1995) 285.

27 Gasteiger, J. and Marsili, M., Tetrahedron, 36 (1980) 3219.

28 Galaxy, Copyright Ghose, A.K., 1995; AM Technologies, San Antonio, TX, U.S.A.

29 Wold, S., Technometrics, 20 (1978) 397.

30 Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P., Numerical Recipes in C, 2nd ed., Cambridge University Press, Cambridge, U.K., 1992.