Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins

Richard D. Cramer, III,* David E. Patterson, and Jeffrey D. Bunce

*Contribution from Tripos Associates, 1699 South Hanley Road,
St. Louis, Missouri 63144.* *Received January 5, 1988*

Abstract: Comparative molecular field analysis (CoMFA) is a promising new approach to structure/activity correlation. Its characteristic features are (1) representation of ligand molecules by their steric and electrostatic fields, sampled at the intersections of a three-dimensional lattice, (2) a new "field fit" technique, allowing optimal mutual alignment within a series, by minimizing the RMS field differences between molecules, (3) data analysis by partial least squares (PLS), using cross-validation to maximize the likelihood that the results have predictive validity, and (4) graphic representation of results, as contoured three-dimensional coefficient plots. CoMFA is exemplified by analyses of the affinities of 21 varied steroids to corticosteroid- and testosterone-binding globulins. Also described are the sensitivities of results to the nature of the field and the definition of the lattice and, for comparison, analyses of the same data using various combinations of other parameters. From these results, a set of ten steroid-binding affinity values unknown to us during the CoMFA analysis were well predicted.

A major goal in chemical research is to predict the behavior of new molecules, using relationships derived from analysis of the properties of previously tested molecules. Relationships derived primarily by empirical analysis of a data table, whose columns are numerical property values and whose rows are compounds, usually taking the form of a linear equation, are called quantitative structure/activity relationships (QSAR).[1]

Especially in biological applications, it has long been agreed that the most relevant numerical property values would be *shape-dependent*. Work on comparative molecular field analysis (CoMFA) began 12 years ago with two additional observations: (1) at the molecular level, the interactions which produce an observed biological effect are usually *non-covalent*; and (2) molecular mechanics force fields, most of which treat noncovalent (non-bonded) interactions only as *steric* and *electrostatic* forces, can account precisely for a great variety of observed molecular properties.[2] Thus it seems reasonable that a suitable sampling of the steric and electrostatic fields surrounding a set of ligand (drug) molecules might provide all the information necessary for understanding their observed biological properties. However, the emergence of a practical CoMFA methodology had to await a new method of data analysis, partial least squares (PLS),[3] which can derive robust linear equations from tables having many more columns than rows, and a number of advances in the methodology of molecular graphics.

Other "3D-QSAR" methodologies have been described. The molecular shape (MS) approaches, developed independently by Simon et al.[4] and by Hopfinger,[5] compare *net*, rather than location-dependent, differences in molecular connectivities, volumes, and/or fields. A second approach, the "distance geometry" method of Crippen,[6] provides validation of a "site-point" hypothesis, a list of binding set coordinates and properties that must be proposed by the investigator. A prototype version of the CoMFA method is called "DYLOMMS".[7] In related work, for exploring binding modes of ligands to receptors, Goodford[8] advocates the display of probe-interaction "grids", similar to those used in CoMFA, while Hansch, Blaney, Langridge, et al.[9] have shown the complementarity of QSAR and molecular graphics in understanding enzyme inhibitor data.

Below we describe the main features of the CoMFA approach, exemplifying its use by analyzing the binding affinities of 21 varied steroid structures to human corticosteroid-binding globulins (CBG) and testosterone-binding globulins[10] (TBG). In this series, the comparative rigidity of the steroid nucleus allows the conformational variable to be neglected, and the in vitro, particularly simple, character of the test system minimizes the importance of non-receptor-related, hence non-shape-related, compound differences on the experimental observations.[11] We then investigated the sensitivity of the excellent results obtained to critical model assumptions. For the purpose of comparison, we have also analysed these steroid binding data using both classical and other "molecular shape" parameters, in various combinations. Finally, toward the end of this work, we were informed of additional corticosteroid binding data,[12] and thus were able to test the ability of our model to predict the binding constants of ten more, structurally diverse, steroids.

Computational Methods

CoMFA Methodology. The overall data flow of a CoMFA analysis appears in Figure 1. Its top two panels show how the data table is constructed from the field values at the lattice intersections. These automatically calculated parameters are the energies of steric (van der Waals 6-12) and electrostatic (Coulombic, with a $1/r$ dielectric) interaction between the compound of interest, and a "probe atom" placed at the various intersections of a regular three-dimensional lattice, large enough to surround all of the compounds in the series, and with a 2.0 Å separation between lattice point unless otherwise stated. The van der Waals A/B values were taken from the standard Tripos force field[13] and the atomic charges were calculated by the method of Gasteiger and Marsili.[14] Unless stated otherwise, the probe atom had the van der Waals properties of sp^3 carbon and a charge of +1.0. Wherever the prove atom experiences a steric repulsion greater than "cutoff" (30 kcal/mol

* Author to whom all correspondence should be addressed.

[1] Martin, Y. C. *Quantitative Drug Design*; Marcel Dekker: New York, 1978.

[2] Burkert, U.; Allinger, N. L. *Molecular Mechanics*; American Chemical Society: Washington, DC, 1982.

[3] Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735.

[4] Simon, Z.; Badileuscu, I.; Racovitan, T. *J. Theor. Biol.* **1977**, *66*, 485. Simon, Z.; Dragomir, N.; Plauchithiu, M. G.; Holban, S.; Glatt, H.; Kerek, F. *Eur. J. Med. Chem.* **1980**, *15*, 521.

[5] Hopfinger, A. J. *J. Am. Chem. Soc.* **1980**, *102*, 7196.

[6] Ghose, A. K.; Crippen, G. M. *J. Med. Chem.* **1985**, *28*, 333 and references therein.

[7] Cramer, R. D., III; Milne, M. Abstracts of the ACS Meeting, April 1979, COMP 44. Wise, M.; Cramer, R. D.; Smith, D. M.; Exman, I. In *Quantitative Approaches to Drug Design*; Dearden, J. C., Ed.; Elsevier: Amsterdam, 1983; p 145. Wise, M. in *Molecular Graphics and Drug Design*; Burgen, A. S. V., Roberts, G. C. K., Tute, M. S., Elsevier: New York, 1986; pp 183-194. Cramer, R. D., III; Bunce, J. D. In *QSAR in Drug Design and Toxicology*; Hadzi, D., Jerman-Blazic, B., Eds.; Elsevier: New York, 1987; p 3.

[8] Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849.

[9] Hansch, C.; Hathaway, B. A.; Guo, Z. R.; Selassie, C. D.; Dietrich, S. W.; Blaney, J. M.; Langridge, R.; Volz, K. W.; Kaufman, B. T. *J. Med. Chem.* **1984**, *27*, 129.

[10] Dunn, J. F.; Nisula, B. C.; Rodbard, D. *J. Clin. Endocrin. Metab.* **1981**, 63.

[11] Cramer, R. D., III *Quant. Struct. Act. Pharmacol., Chem. Biol.* **1983**, *2*, 7, 13. Yunger, L. M.; Cramer, R. D., III *Quant. Struc. Act. Relat. Pharmacol., Chem. Biol.* **1983**, *2*, 149.

[12] Westphal, U. *Steroid-Protein Interactions II*; Springer-Verlag: Berlin, 1986.

[13] Vinter, J. G.; Davis, A.; Saunder, M. R. *J. Comp.-Aided Mol. Design* **1987**, *1*, 31.

[14] Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219.

0002-7863/88/1510-5959$01.50/0

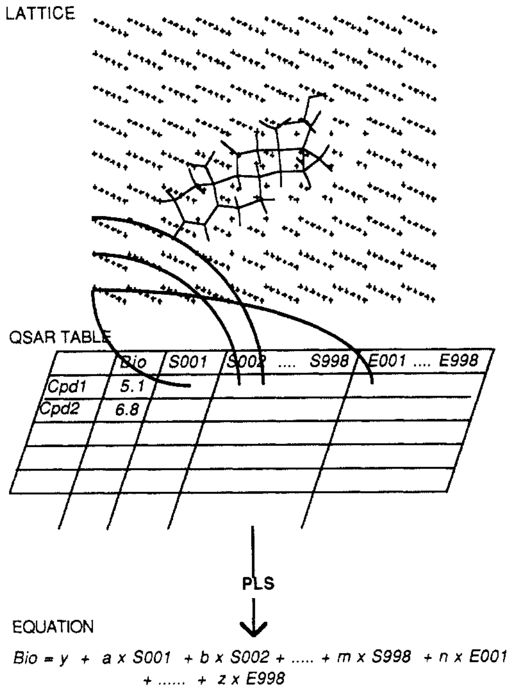

Figure 1. The process of comparative molecular field analysis (CoMFA).

in these studies), the steric interaction is set to the value "cutoff", and the electrostatic interaction is set to the mean of the other molecules' electrostatic interactions at the same location.

To extract a stable QSAR from this unusually proportioned data table, with its highly underdetermined character resulting from many more columns than rows, the partial least-squares (PLS) method developed by Wold et al.[15] is used. Because the units of all independent variables are the same (kcal/mol), the columns are *not* autoscaled. PLS components are extracted as long as the cross-validated r^2 ("predictive r^2") increases. The number of cross-validation groups was four (either five or six compounds "predicted" in each run).

Our implementation of PLS[16] also rotates the PLS solution back into the original data space, thus generating a "conventional" QSAR equation, r^2, and s values. This QSAR equation (bottom panel of Figure 1) contains a potentially non-zero coefficient for each column in the data table, two for each lattice point, and can therefore be contoured in three-dimensional space, just like any other expression associating a numerical value with known locations in space.

Molecular Models and the Alignment Rule. The names and structures used to generate QSARs are shown in Figure 2. Because of the relative rigidity of the steroid nucleus, each compound was represented by a single molecular model. Models were constructed, starting with coordinates taken from the Cambridge Crystallographic Data Base,[17] by minimizing using the standard Tripos force field, searching the side chain torsional space on a 10-deg grid, and minimizing the final structure again. Coordinates in MACCS format of all molecular models used are given in the Supplementary Material, and also are available on PC diskettes from the authors.

The "alignment rule", the positioning of a molecular model within the fixed lattice, is by far the most important input variable in CoMFA, since the relative interaction energies depend strongly on relative molecular positions. At present, conformation selection and trial alignment rules are the exclusive responsibility of the chemist. However, a new "Field Fit" procedure can be used to increase field similarity within a series of molecules. In the "Field Fit" operation, the RMS difference in the sum of steric and electrostatic interaction energies, averaged across all (possibly weighted) lattice points, between that molecule and some template molecule or set of molecules, is minimized with respect to the six rigid-body degrees of freedom and/or any user-specified torsion angles. Expressed differently, with reference to the middle panel of Figure 1, "Field Fitting" Cpd 2 to Cpd 1 would correspond to minimizing the sum of squared differences between the values in all but the first column of the first and second rows of the table, by altering the position and/or torsion angles of Cpd 2. Also required for satisfactory results are a steric repulsion beyond the lattice boundary and, when torsion angles are varied, the conventional molecular mechanics internal energy calculated using the same force field. Minimization is performed by the Simplex method,[18] with step sizes such that individual atoms initially move no more than 0.2 Å. Convergence occurs when successive function evaluations vary less than 1%. As with any minimization, field fit is likely to be successful only if the final geometry is expected to closely resemble (be "downhill" from) the starting geometry.

However, the field-fit procedure was not necessary for these steroid data. The "alignment rule" here was simply the rigid-body, least-squares fitting of the 3, 5, 6, 13, 14, and 17 carbon atoms of each molecule to the corresponding atoms of deoxycortisol, a steroid showing good binding to both globulins.

Comparison of Fields. As a measure of the lattice-point-by-lattice-point similarity of two molecular fields, or of two other conformable vectors such as CoMFA QSAR equations, the correlation coefficient r is familiar and appropriate. Two similar fields, i.e., fields in which high values and low values are observed at corresponding spatial locations, will have an r approaching 1.0; two complementary fields will have an r approaching −1.0; and dissimilar fields will have an r close to 0.0.

Only conformable fields can be compared directly, so when non-conformable fields A and B are to be compared (e.g., two coefficient contour maps based on different lattices, hence different point locations), the field B must be remapped onto A. Only those points in B lying completely within the lattice of A are remapped. The value given to each point in remapped B is the inverse-distance-weighted average of the non-"missing" values at the eight lattice points in the surrounding parallelepipid of A.

In either case, to avoid misinterpretations caused by "spiking", the value at a location is first replaced by the mean of the (up to) 27 values defined by the points making up a cube centered on the location.

Other QSAR Parameters. The Hopfinger molecular shape analysis parameters V_0 and P_u were calculated as originally defined.[19] Algorithmic improvements involved basing the three-dimensional integration of steric/electrostatic field (P_u) on three applications of the trapezoidal rule,[20] rather than a fixed grid sample, assuring convergence and improving computational time, and a new method for calculation of van der Waals volume.[21] The reference molecule was the most active molecule in each set, and the same alignment rule was used as in CoMFA, except that in P_u evaluation, following Hopfinger, each molecule was then translated to position its center of mass on the origin. Because of problems with convergence, whenever P_u evaluation extended to "within" a steroid, P_u was evaluated only for a shell defined by spheres of radii 10.0 and 18.0 Å.

Molar refractivities were summed from tabulated group refractivities.[1] Melting points and six log P values were also taken from the literature.[22] Log P for the remaining steroids was calculated from the experimentally known value for testosterone with use of Rekker fragment constants,[1] and, for unsaturation, the cyclohexene/cyclohexane log P difference. Except for CoMFA itself, all the parameters used in QSAR, also including the coding of five substructural or "dummy" variables, appear in Table I.

Predictions. Of three additional human corticosteroid binding globulin data sets listed by Westphal,[12] the free energies of binding in Table V-23[23] showed the best correlation[24] with the logarithms of the binding constants reported by Dunn, Nisula, and Rodbard,[10] among compounds common to both data sets. The only criterion in selecting the ten compounds for prediction, shown in Figure 3, was the availability of related structures in the Cambridge Crystal Database.[17] Models were constructed and aligned as described above. The actual log K values were calculated by using the correlation equation.[24]

Results

Comparison of QSAR Parameters. The most important results of this work, a comparison of the abilities of different sets of parameters to fit and predict steroid binding potencies, appear

[15] Lindberg, W.; Persson, J.-A.; Wold, S. *Anal. Chem.* **1983**, *55*, 643.

[16] Cramer, R. D., III; Bunce, J. D.; Patterson, D. E.; Frank, I. E. *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1988**, *7*, 18.

[17] Cambridge Crystallographic Database, University Chemical Laboratory, Lensfield Rd., Cambridge CB2 1EW, England.

[18] Nelder, J. A.; Mead, R. *Comp. J.* **1965**, *7*, 308.

[19] Hopfinger, A. J. *J. Med. Chem.* **1983**, *26*, 990.

[20] Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes*; Cambridge Universtiy Press: Cambridge, UK, 1986; pp 102–130.

[21] A traditional point-sampling algorithm, made more rapid by pre-indexing the atomic coordinates with bitsets.

[22] Valvani, S. C.; Yalkowsky, S. H. In *Physical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A.; Valvani, S. C., Eds.; Marcel Dekker: New York, 1980.

[23] Mickelson, K. E.; Forsthoefel, J.; Westphal, U. *Biochemistry* **1981**, *20*, 6211.

[24] The cross-validated r^2 value between free energy (ref 23) and log K (ref 10) for the eleven compounds common to both studies was 0.963 for the equation log K = 1.228 + 0.578 (DG).

Table I. Steroids Used for 3-D QSAR Study[a,c]

compound	TeBG	CBG	MR	MP	log P	calcd log P	A	B	C	D	E
aldosterone	−5.322	−6.279	88.76	164		0.78	0	0	0	0	3
androstanediol	−9.114	−5.000	81.98			4.15	0	0	1	0	0
androstenediol	−9.176	−5.000	79.92	184		3.75	0	0	1	0	1
androstenedione	−7.462	−5.763	77.18	173		2.60	0	0	0	0	3
androsterone	−7.146	−5.613	80.61	185		3.59	0	0	0	0	0
corticosterone	−6.342	−7.881	89.18	181	1.94	1.36	0	0	0	0	3
cortisol	−6.204	−7.881	90.70	213	1.55	0.72	0	0	1	0	3
cortisone	−6.431	−6.892	89.33	222	1.42	0.42	1	0	1	0	3
dehydroepiandrosterone	−7.819	−5.000	78.55	152		3.29	0	0	0	0	1
deoxycorticosterone	−7.380	−7.653	85.45	141	2.90	2.90	0	0	0	0	3
deoxycortisol	−7.204	−7.881	89.18	213		2.72	0	0	0	0	3
dihydrotestosterone	−9.740	−5.919	80.61	181		3.69	0	0	1	0	1
estradiol	−8.833	−5.000	71.32	173		1.59	0	1	1	0	6
estriol	−6.633	−5.000	72.82	282		0.80	0	1	1	1	6
estrone	−8.176	−5.000	69.97	255	1.88	1.90	0	1	0	0	6
etiocholanolone	−6.146	−5.255	80.63			3.69	0	0	0	0	0
pregnenolone	−7.146	−5.255	87.53	149		3.40	0	0	0	0	2
hydroxy pregnenolone	−6.362	−5.000	89.07			2.41	0	0	1	0	2
progesterone	−6.944	−7.380	86.14	131	3.87	3.92	0	0	0	0	1
hydroxy progesterone	−6.996	−7.740	87.68	22		2.91	0	0	1	0	1
testosterone	−9.204	−6.724	78.57	155	3.29[b]	3.29	0	0	1	0	3

[a] Calculated from equation derived from ref 22 (log $P = 1.454 - 0.809 \log S_w - 0.011$ MP; $r^2 = 0.856$, $s = 0.305$). [b] Used as base value to calculate log P values. [c] A = 11-ketone, B = ring A aromatic (estrogens), C = 17-hydroxy substituted, D = 16-hydroxy substituted, E = number of sp^2 carbons in ring A.

Table II. Selected Results of QSAR Studies

			cross-validated		conventional	
independent variables	method	degrees of freedom	r^2	"press"	r^2	s
		A. Corticosteroid-Binding Globulin				
calcd MR	Mult	1	0.311	0.950	0.427	0.911
calcd *l*og P, MP	Mult	2	0.180	1.256	0.028	1.199
calcd MR, A, B, C, D	PLS	1	0.200	1.024	0.337	0.980
V_0, $P_u(16, 18)$, E	PLS	2	0.562	0.764	0.696	0.691
CoMFA	PLS	2	0.662	0.719	0.897	0.397
		B. Testosterone-Binding Globulin				
calcd MR	Mult	1	0.223	1.040	0.299	1.039
calcd log P	Mult	1	0.226	1.015	0.368	0.964
calcd MR, A, B, C, D	PLS	5	0.416	0.902	0.689	0.778
V_0, $P_u(10,12)$, E	PLS	1	−0.217	1.301	0.251	1.073
CoMFA	PLS	2	0.555	0.849	0.873	0.453

in Table II. The two sections of Table II summarize the QSAR for TBG and CBG, respectively. Each line of Table II corresponds to a different QSAR analysis. The first three columns in Table II contain the following inputs: the dependent variable set; the method of deriving the QSAR; and the number of terms or PLS components. The last four columns report the resulting goodness-of-fit. The most important of these is the first column, containing the "predictive" or cross-validated r^2. The latter three columns list the conventional r^2 and the standard errors of the cross-validated predictions ("press") and of the model residuals.

Cross-validation evaluates a model not by how well it *fits* data but by how well it *predicts* data. While useful in many situations,[25] cross-validation is critical for validating the badly underdetermined CoMFA models. In cross-validation of PLS, the analysis is repeated with a randomly chosen subset of the compound-rows excluded, and the resulting model is used to "predict" the biological property value of interest for the excluded compounds, as schematized in Figure 4. This procedure is repeated until every such property value has been "predicted" by a model from whose derivation it was excluded. A "cross-validated r^2" or "predictive r^2" may then be defined, completely analogously to the definition of the conventional r^2, as

$$\text{cross-validated } r^2 = (\text{SD} - \text{press})/\text{SD}$$

where SD is the sum of squared deviations of each biological property value from their mean and press, or predictive sum of squares, is the sum, over all compounds, of the squared differences between the actual and "predicted" biological property values. Note that negative r^2 values will arise whenever press is larger than SD, that is, whenever the biological property values are better estimated by "the mean of all values" than by the model under consideration.

From the r^2 and s values in the last columns of Table II, it is evident that, for these data, the QSAR obtained with CoMFA parameters have greater predictive, and also correlative, power than do QSAR based on any other combination of parameters.[26]

CoMFA Coefficient Contour Maps. The QSAR produced by a CoMFA, with its hundreds or thousands of terms, is usefully represented as a three-dimensional "coefficient contour" map. Figures 5 and 6 show stereo color views of such maps, for the steric aspect only, of both the CBG- and TBG-binding CoMFA QSAR's. (The electrostatic maps are for this set of data almost featureless.) To help in visualization, a blue model of a very strongly bound and a red model of a very weakly bound steroid (cortisol and estradiol for CBG and testosterone and aldosterone for TBG, respectively) are superimposed within each map. In general, the colored polyhedra in each map surround all lattice points where the QSAR strongly associates changes in steroid field values with changes in binding affinity. More specifically, the polyhedra surround lattice points where the scalar products of the associated QSAR coefficient and the standard deviation of all values in the corresponding column of the data table are higher or lower than a user-specified value.

[25] Wold, S. *Technometrics* **1978**, *20*, 397. For a philosophical exposition of this and the related "bootstrapping" technique, see Diaconis, P.; Efron, B. *Sci. Am.* **1984**, 116.

[26] The results shown in the bottom line of Table I have recently been verified by Y. C. Martin and T. Lin of Abbott Laboratories, using our model coordinates, probe interaction energies calculated by Goodford's GRID program (ref 8), and their own implementation of PLS.

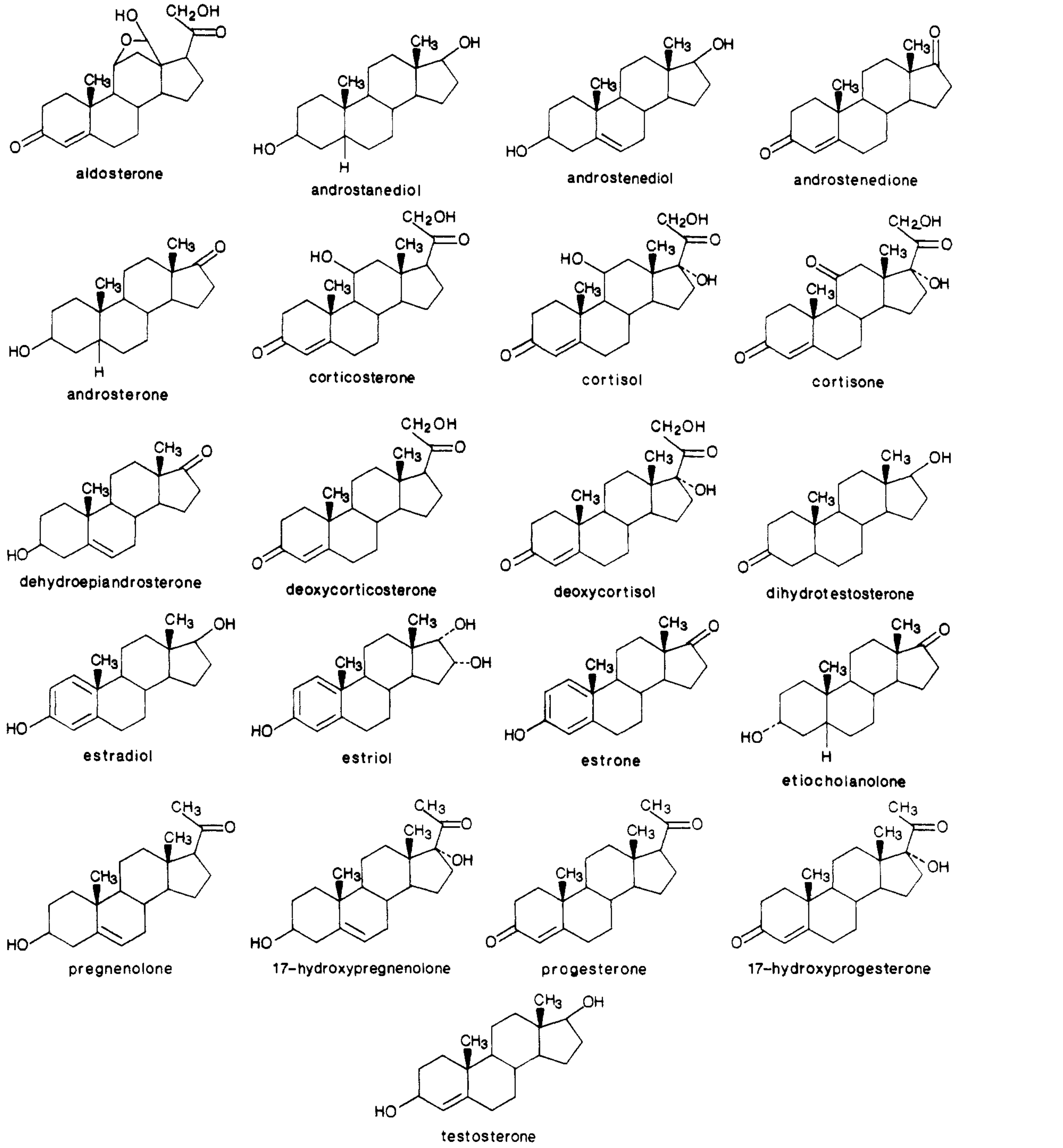

Figure 2. The 21 steroid structures used to derive the CoMFA QSARs.

Color is also used to code the direction and magnitude of these differential interactions. In these steric maps (Figures 5 and 6), blue and cyan polyhedra surround regions where more bulk is "good" (the steric column-variance-weighted QSAR coefficients are less than −0.1 in value within blue and less than −0.01 within cyan, so binding is expected to increase with increases in steric bulk) while red and yellow polyhedra surround regions where less bulk is "good" (steric QSAR coefficients greater than +0.1 within red and greater than +0.01 within yellow). The numerical data used to construct these "coefficient contour" maps, the QSAR coefficients and the data table, are available on request from the authors.

Robustness of CoMFA. A lengthy series of studies, summarized in Tables III–V, explore model "robustness", the dependency of the final CoMFA results on various model parameters, specifically the steric and electrostatic properties of the probe atom and the locations of the lattice intersections. Robustness is a critical concern, for in the absence of structural information about the binding site, the choices of probe atom and lattice location and spacing are blind guesses.[27] Clearly the intermediate numerical results schematized in Figure 1, the field values in the table, and the resulting QSAR must depend strongly on these parameters. However, it is not obvious how greatly the three final results important for design will be affected, specifically the associated cross-validated r^2, the predictions using the QSAR, and the locations of the polyhedra (assessed by the correlation coefficient between two QSAR scaled by column standard deviation, as discussed in the Experimental Section).

Table III shows the effect of probe atom steric properties on final CoMFA results, for both CBG and TBG binding models. The top panel of Table III lists the cross-validated r^2 values obtained for different sized probe atoms, from hydrogen with a

[27] When the receptor is known, CoMFA allows replacement of the probe interaction energies by the receptor-atom-by-atom receptor-ligand interaction energy. However, we have little experience yet with this procedure.

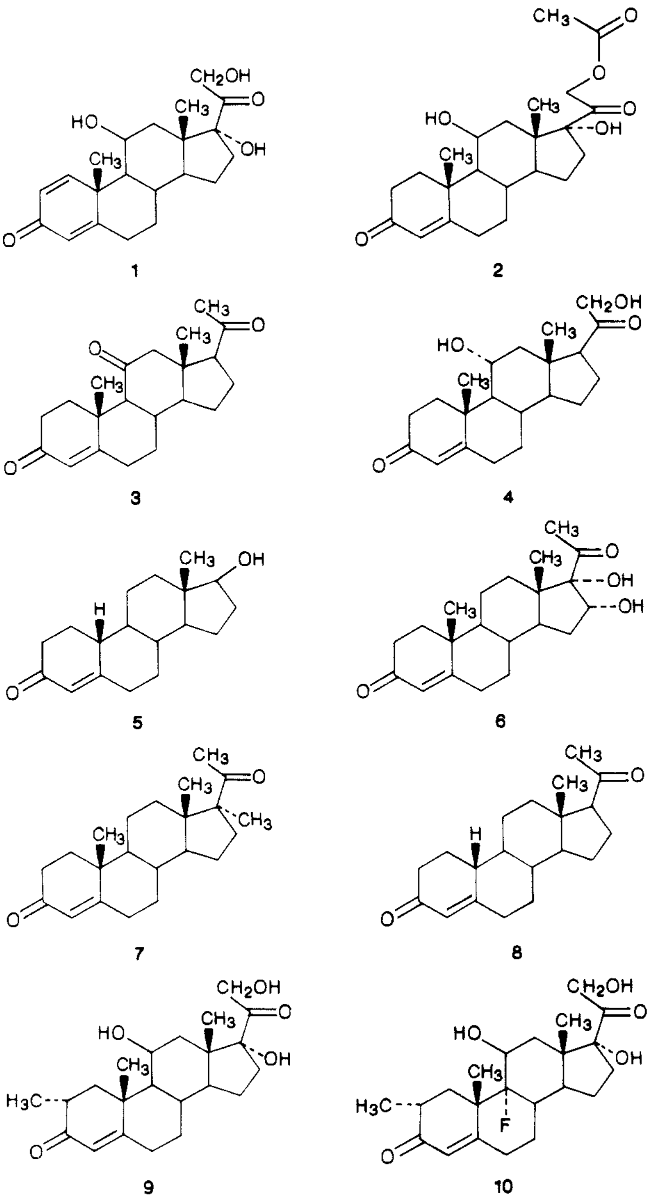

Figure 3. The 10 steroid structures for which corticosteroid-binding globulin affinities were predicted.

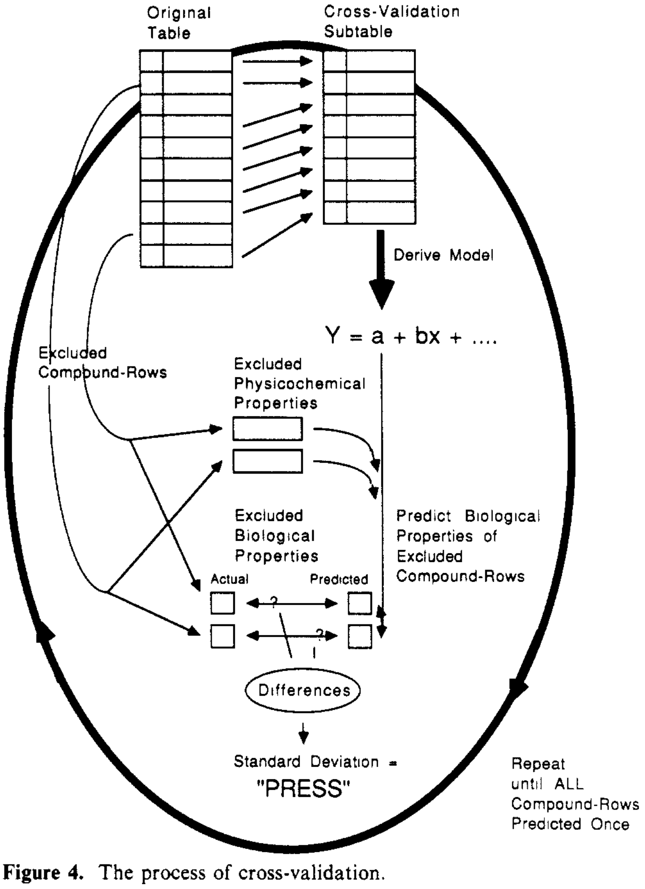
Original Table
Cross-Validation Subtable
Derive Model
Y = a + bx + ...
Excluded Compound-Rows
Excluded Physicochemical Properties
Excluded Biological Properties
Actual
Predicted
?
Predict Biological Properties of Excluded Compound-Rows
Differences
Standard Deviation = "PRESS"
Repeat until ALL Compound-Rows Predicted Once

Figure 4. The process of cross-validation.

VDW radius of 1.08 Å to calcium and 2.75 Å. The middle panel shows the individual binding affinities predicted by using the different QSAR for the molecules in Figure 3, along with their averages and standard deviations. The bottom panel compares "polyhedra locations", as a matrix whose upper right triangle and lower left triangle contain the *r* value comparisons for the TBG- and CBG-binding models, respectively. For example, the value of 0.539 in the lowest left corner of Table III measures the correlation between the sets of column-standard-deviation-weighted QSAR coefficients obtained from the CBG data, when the probe atom had the VDW radii of 1.08 and 2.75 Å. In summary, Table III shows that, for this data set and alignment rule, any sensible value for the steric properties of the probe atom will yield similar final results with CoMFA.

Table IV similarly shows the effect of various translations of the lattice with respect to the set of steroid molecules. While the results in general resemble those of Table III, there is an important difference. The QSAR for TBG only (upper right hand triangle of panel C) varies substantially as the origin of the grid shifts. We believe this to be a consequence of the accidental alignment between the principal axes of the planar steroids and the grid axes. To test this hypothesis, the CoMFA procedure was provisionally modified so that a lattice point represents the interaction averaged over that point and the eight points at the corners of a surrounding cube, rather than just the energy at that point. The results (not shown) showed a much greater similarity in the resulting QSAR's, only the −1, −1, −1 field shift producing a very dissimilar equation.

Table V shows the effect of large changes in lattice spacings on final CoMfA results. Because the fields encompass such widely different sampling intervals, meaningful comparison of field values was difficult and panel C was omitted. The cross-validated r^2 results, in the top panel, suggest that, at least for these molecules, the 2.0-Å spacing between lattice points was a good choice.

Prediction Results. The prediction studies are summarized in Figures 3 and 7. Figure 7 is a plot of actual vs predicted binding affinities (the actual data can be found in any of Tables III through V). Although this plot has the form of the residual plot in, for example, a conventional QSAR study, note that there is an important difference. Most such plots represented how well a model *reproduces* data used in its derivation. Figure 7 shows how well the CoMFA-derived model *predicts* results unknown (to us) at the time of model derivation. The "predictive r^2" value associated with all points in Figure 7 is 0.65, while the value associated with all points except 1, 9, 10, is 0.81. (See discussion for the structural rationale for excluding these points.) Note that the SD for these predictive r^2 computations refers to the mean of the original 21 compounds, not to the 10 predicted; the latter span a smaller range of affinities.

Discussion

In this, its first, application, only the CoMFA procedure provided useful levels of correlation for both of these sets of steroid binding data. This finding is consistent with previous studies. When conventional QSAR parametrizations are used, steroids have always been difficult molecules to correlate successfully.[28]

Despite the large number of descriptors in CoMFA, it is important to understand that this result is most definitely not a simple matter of "given enough parameters, any data can be fit". While assuredly anything can be *fit*, an adventitious correlation will fail

[28] Stouch, T. R.; Jurs, P. C. *J. Med. Chem.* **1986**, *29*, 2125 and references therein.

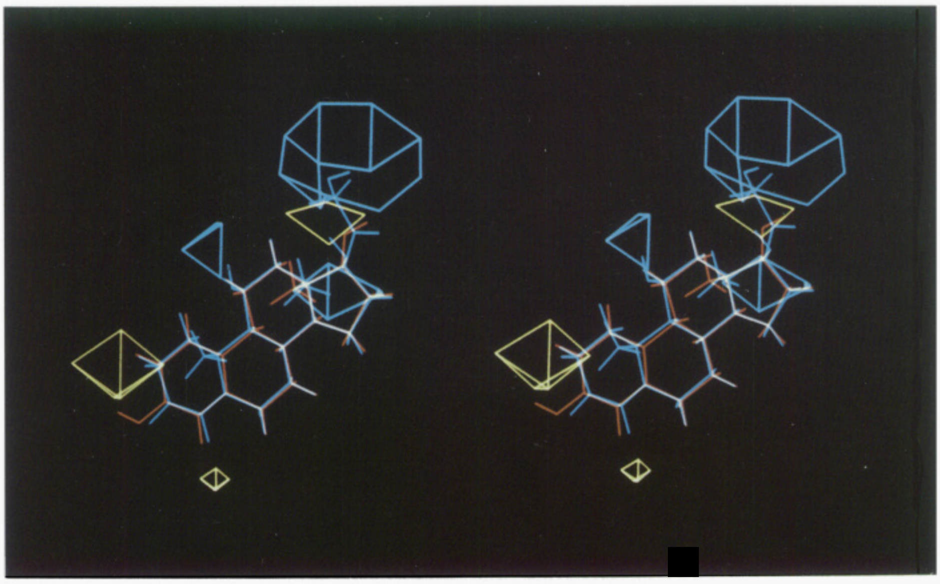

Figure 5. Stereoscopic views of the major steric features of the QSAR for steroid binding to corticosteroid-binding globulin (CBG). Yellow and red contours surround regions where a lower steric interaction would increase binding (the QSAR coefficient times the standard deviation of the corresponding column greater than +0.01 and +0.1, respectively). Blue and cyan contours surround regions where a higher steric interaction would increase binding (less than −0.01 and −0.1, respectively). The red molecule (estradiol) is poorly bound to CBG and the blue molecule (cortisol) strongly bound.

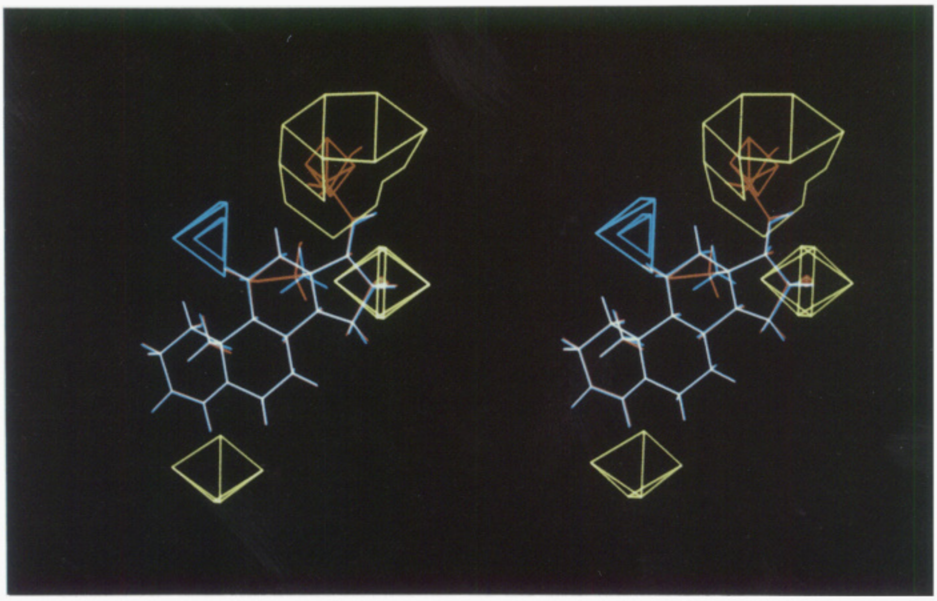

Figure 6. Stereoscopic views of the major steric features of the QSAR for steroid binding to testosterone-binding globulin (TBG). See legend of Figure 5 for the color coding of the contoured regions. The red molecule (aldosterone) is weakly bound to TBG and the blue molecule (testosterone) strongly bound.

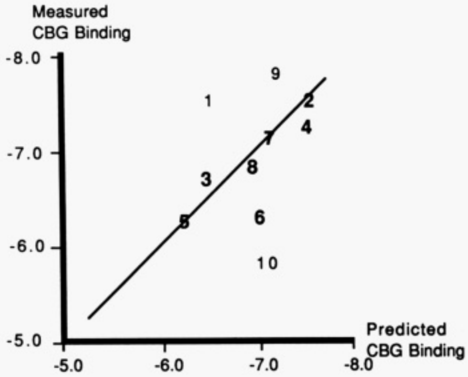

Figure 7. Predicted vs measured affinities to CBG of ten steroids. The numbers, keyed to the structures in Figure 3, are centered on the data locations. Bold face numerals indicate structures with ring geometries and substitution patterns represented among the structures of Figure 2.

the cross-validation test of *predictive* ability, integral to PLS. To provide ourselves with empirical support for this assertion, we used PLS to correlate hundreds of columns of random numbers with one of the columns of 21 biological data. Although *conventional* r^2 values higher than 0.99 usually resulted, in none of a dozen trials did the associated cross-validated r^2 even take on a positive value.

Instead, we believe that the CoMFA technique succeeds because of the high relevance of its descriptors to the structure–activity correlation problem. First, consider their physicochemical nature. Steric and electrostatic interactions are the primary, often the only, non-covalent interactions in molecular mechanics force fields. Recent successes in quantifying ligand–enzyme binding by using such force fields in a perturbational treatment[29] are consistent

[29] Lybrand, T.; McCammon, J. A.; Wipff, G. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 833. Bash, P. A.; Singh, U. C.; Langridge, R.; Kollman, P. A. *Science* **1987**, *236*, 564.

Table III. Comparison of CoMFA Results with Different Atom Probes

A. Cross-Validated r^2

VDW radius	CBG (n = 2)	CBG (n)	TGB (n = 2)	TBG (n)
1.08 (H)	0.728	0.740 (3)	0.647	0.647 (2)
1.36 (O)	0.678	0.713 (4)	0.647	0.647 (2)
1.45 (N)	0.649	0.686 (4)	0.640	0.650 (5)
1.52 (C)[a]	0.662	0.677 (4)	0.555	0.577 (3)
1.72 (S)	0.673	0.759 (4)	0.520	0.520 (2)
1.75 (P)	0.693	0.748 (4)	0.562	0.562 (2)
2.05 (I)	0.702	0.751 (3)	0.660	0.660 (2)
2.75 (Ca)	0.728	0.754 (3)	0.541	0.541 (3)

B. Predictions of Corticosteroid Activity, by Compound

VDW radius	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1.08	−6.992	−7.705	−7.141	−7.446	−6.261	−7.076	−5.295	−7.067	−7.253	−7.093
1.36	−7.087	−7.638	−6.984	−7.527	−6.390	−7.085	−5.336	−6.974	−7.129	−6.968
1.45	−7.026	−7.564	−6.766	−7.527	−6.239	−7.192	−5.419	−6.846	−7.099	−6.845
1.52[a]	−6.544	−7.540	−6.526	−7.546	−5.955	−7.057	−5.384	−7.009	−7.227	−6.937
1.72	−7.047	−7.638	−6.176	−7.487	−6.261	−7.344	−5.707	−6.833	−7.846	−7.789
1.75	−7.015	−7.715	−6.346	−7.345	−6.103	−7.231	−5.693	−7.005	−7.785	−7.744
2.05	−7.152	−7.778	−6.862	−7.439	−6.097	−7.014	−5.508	−7.020	−7.756	−7.698
2.75	−7.008	−7.815	−6.983	−7.360	−5.945	−7.157	−5.282	−6.910	−7.593	−7.575
mean	−6.984	−7.764	−6.723	−7.460	−6.156	−7.145	−5.453	−6.958	−7.461	−7.331
std dev	0.173	0.091	0.317	0.072	0.148	0.102	0.158	0.080	0.331	0.380
actual	−7.512	−7.553	−6.779	−7.200	−6.144	−6.247	−7.120	−6.817	−7.688	−5.797

C. Correlation Coefficient between QSAR (r)

VDW radius	1.08	1.36	1.45	1.52	1.72	1.75	2.05	2.75	
1.08		0.771	0.736	0.451	0.423	0.414	0.351	0.495	TBG
1.36	0.782		0.753	0.659	0.678	0.649	0.565	0.591	
1.45	0.839	0.940		0.718	0.628	0.595	0.498	0.529	
1.52[a]	0.571	0.677	0.704		0.917	0.881	0.768	0.675	
1.72	0.701	0.735	0.785	0.799		0.990	0.927	0.790	
1.75	0.697	0.738	0.782	0.804	0.996		0.962	0.822	
2.05	0.747	0.735	0.772	0.749	0.924	0.925		0.833	
2.75	0.539	0.540	0.568	0.540	0.646	0.651	0.683		
	CBG								

[a] The standard parameter setting.

with the hypothesis underlying CoMFA, that a suitable sampling of the steric and electrostatic interactions of a ligand would suffice to answer most questions about its possible receptor interactions. Second, consider the truly three-dimensional, shape-dependent nature of a row of CoMFA descriptors. The "shape analysis" parameters used in some other "3D-QSAR" methods are actually aggregate indices, which describe "shape" only to the same extent, for example, that the "shape" of a sculpture is described by measuring its differential weight or volume. Since drug/receptor interaction is universally believed to depend mainly on shape complementarity, descriptors that actually differentiate the details of drug topography should be advantageous.[30] Finally, consider that each CoMFA parameter represents the interaction energy of an entire ligand, not just the interaction of a more or less arbitrarily selected substructure of the ligand.

The most significant evidence for the utility of these CoMFA-derived models is the good prediction results of Figures 3 and 7. In general, QSAR publications seldom attempt predictions. In any case, with most other QSAR parametrizations, which are closely tied to topological structure, predictions would not be possible for such a diversity of structures. Note that these predictions are even better than a casual glance at the point distribution in Figure 7 might suggest, because the point cloud is centered in the upper right of the graph, showing that the model correctly predicted high affinity for these molecules as a group. It is also encouraging that the poorest predictions are for molecules that are dissimilar from any molecule in Figure 2. Specifically, compound **1** in Figure 7 is the only dienone A-ring among the 31 compounds of Figures 3 and 7; compounds **9** and **10** have the only 2 substituents, and compound **10** also has the only fluorine and the only 9 substituent. In summary, we believe these predictions, certainly among the more convincing ever made for biological properties, suggest a most practical value for the CoMFA technique in the context of molecular discovery research programs.

Are all of these promising results somehow an artifact of the steroid dataset? Two types of artifact might be considered. The first, an apparent pattern within numbers which actually are uncorrelated, might seem likely in this underdetermined model. However, a random pattern among such a structurally varied group of 21 compounds seems most unlikely to predict the properties of a second group of 10 similarly varied molecules. The second, more subtle, type of artifact might be that the observed pattern is actually a trivial consequence of some more simple relationship among these data. But, as shown in Table I, we were not able to find other good correlates for these data.

At this writing, an even better level of predictive performance has been obtained in a study of structurally diverse and conformationally labile angiotensin-converting enzyme (ACE) inhibitors,[31] and results much more consistent with the crystal structure of the actual binding geometry than would be expected from the structural variations were obtained in a study of dihydrofolate reductase inhibitors related to trimethoprim.[32] These results strengthen our belief in the power of the CoMFA methodology. However, the steroid dataset, because of its structural variety combined with conformational simplicity, remains the most compelling and complete application of the technique.

In addition to its modelling power, physicochemical realism, and predictive performance, the current version of CoMFA has

[30] Marshall, G. R.; Motoc, I. In *Molecular Graphics and Drug Design*; Burgen, A. S. V., Roberts, G. C. K., Tute, M. S., Eds.; Elsevier: New York, 1986; pp 117–156.

[31] Mayer, D.; Naylor, C. B.; Cramer, R. D., III; Marshall, G. R., manuscript in preparation.

[32] Naylor, C. B.; Mayer, D.; Motoc, I.; Dammkoehler, R. A.; Cramer, R. D., III; Marshall, G. R., manuscript in preparation.

Table IV. Comparison of CoMFA Results with Different Offsets of Lattice

A. Cross-Validated r^2

	Å offset						
field no.	X	Y	Z	CBG (n = 2)	CBG (n)	TGB (n = 2)	TBG (n)
1[a]	0.0	0.0	0.0	0.662	0.677 (4)	0.555	0.577 (3)
2	−0.5	0.0	0.0	0.693	0.734 (4)	0.707	0.723 (8)
3	0.0	0.0	−0.5	0.697	0.754 (3)	0.660	0.668 (6)
4	−0.5	−0.5	0.0	0.652	0.669 (3)	0.694	0.694 (2)
5	−0.5	−0.5	−0.5	0.753	0.755 (3)	0.636	0.636 (2)
6	−1.0	−1.0	0.0	0.646	0.646 (2)	0.631	0.631 (2)
7	−1.0	−1.0	−0.5	0.559	0.584 (4)	0.681	0.714 (2)
8	−1.0	−1.0	−1.0	0.779	0.779 (2)	0.500	0.535 (2)

B. Predictions of Corticosteroid Activity, by Compound

	1	2	3	4	5	6	7	8	9	10
1[a]	−6.544	−7.540	−6.526	−7.546	−5.955	−7.057	−5.384	−7.009	−7.227	−6.937
2	−6.959	−7.868	−6.873	−7.581	−6.239	−7.236	−5.672	−7.000	−7.406	−7.297
3	−6.872	−7.856	−7.144	−7.496	−6.250	−7.107	−5.511	−7.022	−7.753	−7.708
4	−7.089	−7.822	−6.725	−7.329	−6.111	−7.015	−5.488	−6.806	−7.207	−7.150
5	−7.224	−7.810	−6.196	−7.453	−6.243	−7.005	−5.244	−7.051	−7.804	−7.805
6	−7.780	−7.930	−6.592	−7.408	−5.876	−7.113	−5.857	−6.625	−7.804	−7.668
7	−7.292	−7.792	−6.841	−7.313	−6.111	−6.853	−5.868	−6.828	−7.722	−7.563
8	−6.857	−7.797	−7.033	−7.610	−6.182	−7.167	−5.409	−7.030	−7.616	−7.467
mean	−7.077	−7.802	−6.831	−7.467	−6.121	−7.169	−5.554	−6.921	−7.567	−7.449
std dev	0.345	0.108	0.197	0.104	0.131	0.109	0.109	0.211	0.143	0.281
actual	−7.512	−7.553	−6.779	−7.200	−6.144	−6.247	−7.120	−6.817	−7.688	−5.797

C. Correlation Coefficient between QSAR (r)

	amount shifted											
field	X	Y	Z	1	2	3	4	5	6	7	8	
1[a]	0.0	0.0	0.0		0.814	0.106	0.736	0.174	0.583	−0.171	0.408	TBG
2	−0.5	0.0	0.0	0.773		0.136	0.919	0.240	0.716	−0.109	0.585	
3	0.0	0.0	−0.5	0.272	0.630		0.509	0.630	0.517	0.645	0.697	
4	−0.5	−0.5	0.0	0.599	0.877	0.703		0.187	0.700	−0.048	0.597	
5	−0.5	−0.5	−0.5	0.531	0.726	0.860	0.543		0.186	0.145	0.135	
6	−1.0	−1.0	0.0	0.664	0.903	0.736	0.961	0.701		−0.127	0.506	
7	−1.0	−1.0	−0.5	0.435	0.652	0.887	0.507	0.759	0.405		0.048	
8	−1.0	−1.0	−1.0	0.587	0.833	0.745	0.849	0.677	0.791	0.736		
				CBG								

[a] The standard parameter setting.

Table V. Comparison of CoMFA Results with Different Lattice Spacing

A. Cross-Validated

	spacing, Å						
field no.	X	Y	Z	CBG (n = 2)	CBG (n)	TBG (n = 2)	TBG (n)
1	1.0	1.0	1.0	0.750	0.769 (3)	0.625	0.625 (2)
2	2.0	1.0	1.0	0.756	0.777 (3)	0.588	0.588 (2)
3	1.0	1.0	2.0	0.702	0.741 (4)	0.679	0.679 (2)
4[a]	2.0	2.0	2.0	0.662	0.677 (4)	0.555	0.577 (3)
5	3.0	3.2	2.0	0.473[b]	0.473 (1)	0.554	0.554 (2)
6	4.5	4.0	2.0	0.400[b]	0.449 (4)	0.604	0.604 (2)

B. Predictions of Corticosteroid Activity, by Compound

field no.	1	2	3	4	5	6	7	8	9	10
1	−6.629	−7.744	−6.594	−7.518	−6.650	−7.409	−5.247	−7.373	−7.908	−7.800
2	−6.746	−7.681	−6.938	−7.478	−6.038	−7.128	−5.305	−6.966	−7.541	−7.371
3	−6.949	−7.742	−6.663	−7.499	−5.942	−7.047	−5.486	−6.826	−7.633	−7.499
4[a]	−6.544	−7.540	−6.526	−7.546	−5.955	−7.057	−5.384	−7.009	−7.227	−6.937
5	−6.780	−7.150	−6.906	−7.066	−5.301	−6.809	−7.052	−6.962	−7.162	−7.114
6	−6.599	−6.887	−7.013	−7.041	−5.243	−6.894	−6.941	−7.028	−7.023	−6.956
mean	−6.708	−7.457	−6.740	−7.358	−5.855	−7.057	−5.903	−7.027	−7.416	−7.280
std dev	0.135	0.327	0.171	0.216	0.477	0.190	0.778	0.168	0.305	0.310
actual	−7.512	−7.553	−6.779	−7.200	−6.144	−6.247	−7.120	−6.817	−7.688	−5.797

[a] The standard parameter setting. [b] One component was used instead of 2 since 1 component had a higher cross-validated r^2 than 2 components in fields 5 and 6 of this table.

other promising attributes. First, the only inputs needed are models of all the molecular forms, the lattice description, and, usually, an explicit "alignment rule". The most important outputs are the cross-validated r^2, the "coefficient contour" map displays, and model predictions. Because handling these inputs and outputs does not seem to require specialized experience with parametrizations or statistics, and because the design of new molecules predicted by the current model to have improved properties seems to be quite as straightforward a process as Figures 5 and 6 suggest, CoMFA may be a technique that any interested chemist can effectively use. Second, its computational demands are modest by current standards. In general running times are proportional to the various problem dimensions. For example, in this work, which typically involved the dimensions of 21 compounds, 500–1000 lattice points, and five cross-validation groups, individual runs took about a CPU hour on a MicroVAX II. Finally, the technique appears extremely general, being directly applicable to any series of molecules for which alignable models can be constructed and whose desired property is believed to result from an alignment-dependent non-covalent molecular interaction. Indeed, the approach seems relevant to the discovery of "receptors" of known or inferrable structure as well as of ligands, for example, improved zeolite or enzyme catalysts.

A basic CoMFA concept, representation of a molecule by a vector sampling of its electrostatic and steric fields, can be useful in applications other than molecular design via QSAR. Goodford uses such a "grid" as a visual guide in docking a ligand to a receptor.[8] More formally, the "field fit" method of minimizing differences between molecular fields might be used to investigate possible ways that a guest molecule might fit into a cavity of known structure (maximixing rather than minimizing field differences), or to predict which of several conformations of a molecule may be responsible for an observed property.

There are some potential difficulties to be noted with CoMFA. First are the inter-related operational issues of specifying an initial "alignment rule" and the "active conformation(s)" for each individual compound, within a series of interest. The well-established "active analog" approach of Marshall, Barry, Dammkoehler, et al.,[33] which identifies active conformers on the basis of a conformational search constrained by distances between user-specified atoms, often fulfills this need. Second is a reminder that, even though PLS provides a robust self-consistent QSAR, with cross-validation ensuring a high probability of predictive utility, the system remains inherently underdetermined, with many times more coefficients to be evaluated than compounds. Almost certainly there are other QSAR equally consistent with any given set of compounds and data, as for example was suggested in the TBG binding data triangle of panel C of Table IV. This consideration underlines a caution against the temptation to overinterpret the "contour coefficient" maps, for example, as "receptor maps". All possibly relevant aspects of a ligand–receptor interaction surely cannot be explored with test results for a few dozen compounds. In practice, however, maps of column variance, similar to "contour coefficient maps", can help in delineating the less explored volumes of a lattice, and "difference maps" can highlight differences between QSAR for exploration by further synthesis and testing. A fourth caution is that CoMFA will often fail when a few of the molecules are very dissimilar from *all* others, in both shape and property-of-interest value, because of the impossibility of predicting the behavior of the dissimilar molecules from the others and the consequent failure of the cross-validation test. In this situation, although the PLS method can derive models without the cross-validation constraint, there is a much higher risk of chance correlation[34] and CoMFA results should be viewed only as tentative hypotheses to guide synthesis, with no expectation of success in predictions. Finally, the molecular mechanics snapshot of steric and electrostatic nonbonded enthalpies, which produces the CoMFA parameters, does not include entropically based factors such as hydrophobicity,[35] which also contribute to ligand binding, and which may have shape-related aspects inadequately described in the CoMFA parametrization. Of course, other molecular descriptors, appropriately weighted, such as octanol/water log *P*, can be included with the probe interaction energies before CoMFA QSAR derivation.

Conclusion. In the analysis of two sets of steroid-binding globulin data, the new structure/property correlation technique of comparative molecular field analysis (CoMFA) shows a number of uniquely promising attributes, including inherent generality, ease of use, superior analytic power, and predictive utility.

Acknowledgment. Support from NIH (Grant R44 GM33662), via the SBIR program, is most gratefully acknowledged. For calling our attention to the original steroid binding problem, and to the prediction data set, we thank David Rodbard, via Ioan Motoc, and John Katzenellenbogen, respectively. R.D.C. thanks Graham Richards and Svante Wold for critical technical suggestions and Garland Marshall, Margaret Wise, and Richard Dammkoehler for helpful discussions and encouragement.

Registry No. 1, 50-24-8; **2**, 50-03-3; **3**, 516-15-4; **4**, 600-67-9; **5**, 434-22-0; **6**, 595-77-7; **7**, 1239-79-8; **8**, 57-83-0; **9**, 3836-17-7; **10**, 432-34-8; aldosterone, 52-39-1; androstanediol, 571-20-0; androstenediol, 521-17-5; androstenedione, 63-05-8; androsterone, 53-41-8; corticosterone, 50-22-6; cortisol, 50-23-7; cortisone, 53-06-5; dehydroepiandrosterone, 53-43-0; deoxycorticosferone, 64-85-7; deoxycortisol, 152-58-9; dehydrotestosterone, 521-18-6; estradiol, 50-28-2; estriol, 50-27-1; estrone, 53-16-7; etiocholanolone, 53-42-9; pregnenolone, 145-13-1; hydroxy pregnenolone, 12041-98-4; progesterone, 57-83-0; hydroxy progesterone, 68-96-2; testosterone, 58-22-0.

Supplementary Material Available: Coordinates and connection lists for all structures in Figures 2 and 3 (33 pages). Ordering information is given on any current masthead page.

[33] Marshall, G. R.; Barry, C. D.; Bosshard, A. E.; Dammkoehler, R. A.; Dunn, D. A. In *Computer-Aided Drug Design*; Olson, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979; ACS Symp. Series No. 112, p 205.

[34] Topliss, J. G.; Edwards, R. P. *J. Med. Chem.* **1979**, *22*, 1238. Wold, S.; Dunn, W. J., III *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 6.

[35] While the experimental evidence for the classical dogma of hydrophobicity has been widely debated (ref 31 of Cramer, R. D. III, *J. Am. Chem. Soc.* **1980**, *102*, 1837, references some dissenting views), recent MonteCarlo simulations, for example, the solvent effect on the cis/gauche equilibrium of butane (Jorgensen, W. L. *J. Chem. Phys.* **1982**, *29*, 5757), are more compelling.