On the Stability of CoMFA Models

James L. Melville and Jonathan D. Hirst*

School of Chemistry, University of Nottingham, University Park, Nottingham NG7 2RD, U.K.

Received February 5, 2004

Abrupt, smooth, and box methods for the calculation of electrostatic and steric field values in the comparative molecular field analysis (CoMFA) 3D QSAR technique are assessed on three diverse data sets of medicinal chemistry interest. While the standard CoMFA settings are robust to small changes in the position of the lattice, superior results may sometimes be obtained by use of only one field. However, if only the electrostatic field is used, then sometimes large differences between models are apparent. This appears to be due to a lack of column dropping, and these difficulties can be remedied by use of the box method.

INTRODUCTION

Since its inception in 1988,[1] the comparative molecular field analysis (CoMFA) approach to three-dimensional quantitative structure−activity relationships (3D QSAR) has proved enormously popular—a Web of Science search reveals over 1000 citations of the original paper, which does not take into account the additional unpublished studies in the pharmaceutical industry. Standard CoMFA involves sampling the steric and electrostatic interactions (by means of the Lennard−Jones and Coulomb potentials, respectively) between a series of aligned molecules and an array of probe atoms arranged in a rectilinear grid. The relation between the interactions and the response variable (e.g., the activity) is established by use of the partial least-squares (PLS) regression method.[2]

Despite its undoubted popularity, CoMFA analyses can be highly sensitive to the parameters used to construct the matrix of 3D descriptors, e.g., the type of partial charges used for the electrostatic field,[3] the method of calculation of steric fields,[4,5] and the grid spacing[6] and dimensions.[7] In particular, the sensitivity of results to translations of the grid relative to the aligned molecules has been noted.[6,8,9] Some techniques harness this variation to improve the statistical quality of the regression.[10,11] A variable selection approach may also be employed to reduce the instability.[8]

Another possibility, rarely mentioned in the literature, is the use of alternative methods within CoMFA to generate the steric and electrostatic fields.[12] In the box option, the probe atom is replaced with eight probe atoms, located at the corners of a box, the center of which is the location of the original probe. The interaction between each of the corner probes and the molecule is calculated, and the mean interaction is assigned to the probe at the center of the cube. This method was briefly alluded to by Cramer and co-workers,[12] who reported an improvement of over 0.2 q^2 units with unpublished data, but there is no indication in the literature of any widespread adoption of this technique in the 10 years since this was first brought to the attention of medicinal chemists—one exception is a study by Kroemer and co-workers on HIV proteinase inhibitors,[13] where the box method gave slightly superior results to the standard fields with a smoothed transition to the cutoff. Additionally, models based on the inclusion of different fields[14] or varying parameter settings[15] may display different sensitivities to small changes in the lattice position.

Given the widespread use of CoMFA and the potential significant improvement suggested by earlier results,[12] in this paper we investigate the effect of the different methods of generating steric and electrostatic descriptors within CoMFA. The methods are validated using three different data sets, and we investigate the effect of the form of dielectric function, grid spacing, and partial charge calculation. While there is some evidence that CoMFA is relatively insensitive to the form of geometry optimization[16] and that the exact bioactive conformation of the molecules may not be necessary for good results,[17] incorporating conformational flexibility presents challenges. Several methods have been suggested to account for multiple conformation[10,18,19] and alignment[20] possibilities within the CoMFA methodology, but in order to simplify the analysis, we shall not consider these. The results of this investigation are assessed not just on the statistical results, but also on the effect of sensitivity to grid placement.

DATA SETS AND COMPUTATIONAL METHODS

Data Sets. Three data sets were chosen from the literature on the basis of a previous successful application of CoMFA and optimized aligned coordinates being readily available or that the geometry optimization was straightforward and unambiguous alignment rules were presented. Data set 1 consisted of the 31 CoMFA steroids.[1] Despite some well-known difficulties with the set, it has been widely used to illustrate additions and modifications to CoMFA and other 3D QSAR techniques and is considered a benchmark for validating such methods.[21] As the original data set contained some structural errors, the corrected set provided by Gasteiger was used.[22] Data set 2 consisted of 38 D_2 antagonists, the subject of a recent study by Boström and co-workers.[23] Data set 3 consisted of 34 polychlorinated dibenzofurans known to bind to the Ah receptor[20,24] (it was pointed out by a referee that some studies using this data set, e.g., ref 24, contain

* Corresponding author phone: +44-115-951-3478; fax: +44-115-951-3562; e-mail: jonathan.hirst@nottingham.ac.uk.

10.1021/ci049944o CCC: $27.50
Published on Web 05/08/2004

errors and duplications in the structures). Data sets 1 and 2 were both originally split into training and test sets, but we did not apply this splitting in our study. While the use of an external test set for model validation is popular,[25] for the size of data set that we consider here (which is quite typical in 3D QSAR studies), a more accurate assessment can be obtained by use of cross-validation.[26,27]

Geometry Optimization and Alignment. The structures and activities of the molecules in the three data sets are provided in the Supporting Information. The atomic coordinates provided by Gasteiger were used for data set 1 and also optimized at the BLYP/6-31G* level in the quantum chemistry program QChem 2.0.[28] For data set 2, the coordinates of the molecules were taken from the original study[23] and used as-is. The structures in data set 3 were built in PC Spartan Pro, version 1.0.8 (Wavefunction, Inc., Irvine, CA), and optimized at the BLYP/6-31G* level with QChem 2.0. The alignments used were those specified in previous studies.[1,23,24]

Variable Calculation. The two standard CoMFA fields are considered: electrostatic and steric fields. In most work, a distance-dependent $1/r$ dielectric function (the default setting) is applied to the electrostatic field, but a constant dielectric function is sometimes employed. In this work, we consider both functions. The default cutoff, $E_{cut} = 30$ kcal mol^{-1}, is applied to both fields. There are two methods of applying the cutoff. The abrupt method sets all probe values that experience an interaction energy greater than E_{cut} to E_{cut}. However, the default setting in CoMFA applies a smoothing function. For all field values greater than $1.2E_{cut}$, a value of E_{cut} is applied, and for field values between $0.8E_{cut}$ and $1.2E_{cut}$ the following interpolation is applied[12]

$$E_{smooth} = \frac{E^2 - 2E_{up}E + {E_{lo}}^2}{2(E_{lo} - E_{up})} \quad (1)$$

where E is the original field value, $E_{lo} = 0.8E_{cut}$, and $E_{up} = 1.2E_{cut}$.

An alternative to the smooth option is the box option, in which the single probe atom is replaced by eight probe atoms, making the corners of a cube, centered on the coordinates of the original probe. The length of the side of the cube is two-thirds of the normal lattice spacing. The interactions between each probe atom and the target molecule are calculated, and then the average of these interactions is assigned to the original probe coordinates.

In this study, the Lennard−Jones parameters and van der Waals radii were taken from the Tripos force field[29] and the partial charges were calculated by the semiempirical AM1 method[30] using Spartan. Default CoMFA parameters were employed in this study: a dielectric constant of 80; the lattice used for each data set was extended at least 4 Å in all directions from the aligned molecules; an sp^3 carbon atom with a +1 charge was used as a probe; a lattice spacing of 2 Å was employed. Where more than one field was considered at once, block scaling (COMFA_STD) was employed along with column dropping, where electrostatic descriptors at lattice points that registered a steric field value larger than the cutoff value were set to the average of all electrostatic values that were not within the steric envelope. Column filtering was employed to reduce the number of variables for both electrostatic and steric fields by 80%. Removing low-variance columns in this way had a negligible effect on the quality of the model while affording an appreciable speed increase.

Regression models were built with these descriptors using PLS,[2] employing the SIMPLS algorithm of de Jong.[31] The program was written in-house in C++. The quality of the regression model is tested via cross-validation. Leave-one-out (LOO) was employed for this purpose. While LOO possesses some less than desirable theoretical properties,[32] in practice it is overwhelmingly the method of choice in QSAR studies and gives an accurate evaluation of the predictive quality of a model.[26,27] Repeating some of the results given below using 3-fold cross-validation with up to 200 separate splits gave extremely similar results but with, as expected, slightly lowered q^2 values. As this had no effect on discriminating among the different models for a particular data set, we used LOO, which has the dual advantage of being faster and providing more easily replicable results than n-fold cross-validation.

As recommended by Kubinyi and Abraham,[33] the optimum number of components (ONC) was chosen as that which corresponded to the first minimum in the standard error of cross-validation SE_{CV}

$$SE_{CV} = \sqrt{\frac{PRESS}{N - ONC - 1}} \quad (2)$$

where N is the number of molecules and the predicted residual sum of squares, PRESS, is given by

$$PRESS = \sum_{n=1}^{N} (y_{n,obs} - y_{n,pred})^2 \quad (3)$$

where $y_{n,pred}$ and $y_{n,obs}$ are the predicted and observed activities of molecule n, respectively. To avoid chance correlations, no more than $N/5$ components were extracted.

While SE_{CV} is considered more informative than q^2,[12] the LOO q^2 is perhaps the more widespread and intuitive parameter, so we shall quote q^2 values as an indicator of the *predictive quality* of a model and the ONC as a measure of the *parsimony* of the model—while a high q^2 is desirable, the more components that are extracted, the greater the risk of overfitting. To assess the effect of small grid translations on the quality of the regression, for each set of parameters, cross-validation was repeated with the grid translated by a small amount along the principal axes of the lattice. All 27 possible ±0.1 Å displacements along one, two, or all three of the x, y, and z directions, relative to the default lattice position, were considered. The mean and standard deviation of the q^2 and ONC values for the 27 runs is reported.

RESULTS AND DISCUSSION

Smooth versus Box Fields. We first compare the field values that arise in box-generated fields and the smoothed versions. Figure 1 illustrates the effect of the different field value calculations for aldosterone, one of the molecules in the CoMFA steroid set. With the centroid of aldosterone located at the point (0,0,0) in the lattice and with its long axis approximately oriented along the x-axis, the field values measured by sp^3 C atom with a +1 charge placed at 1 Å

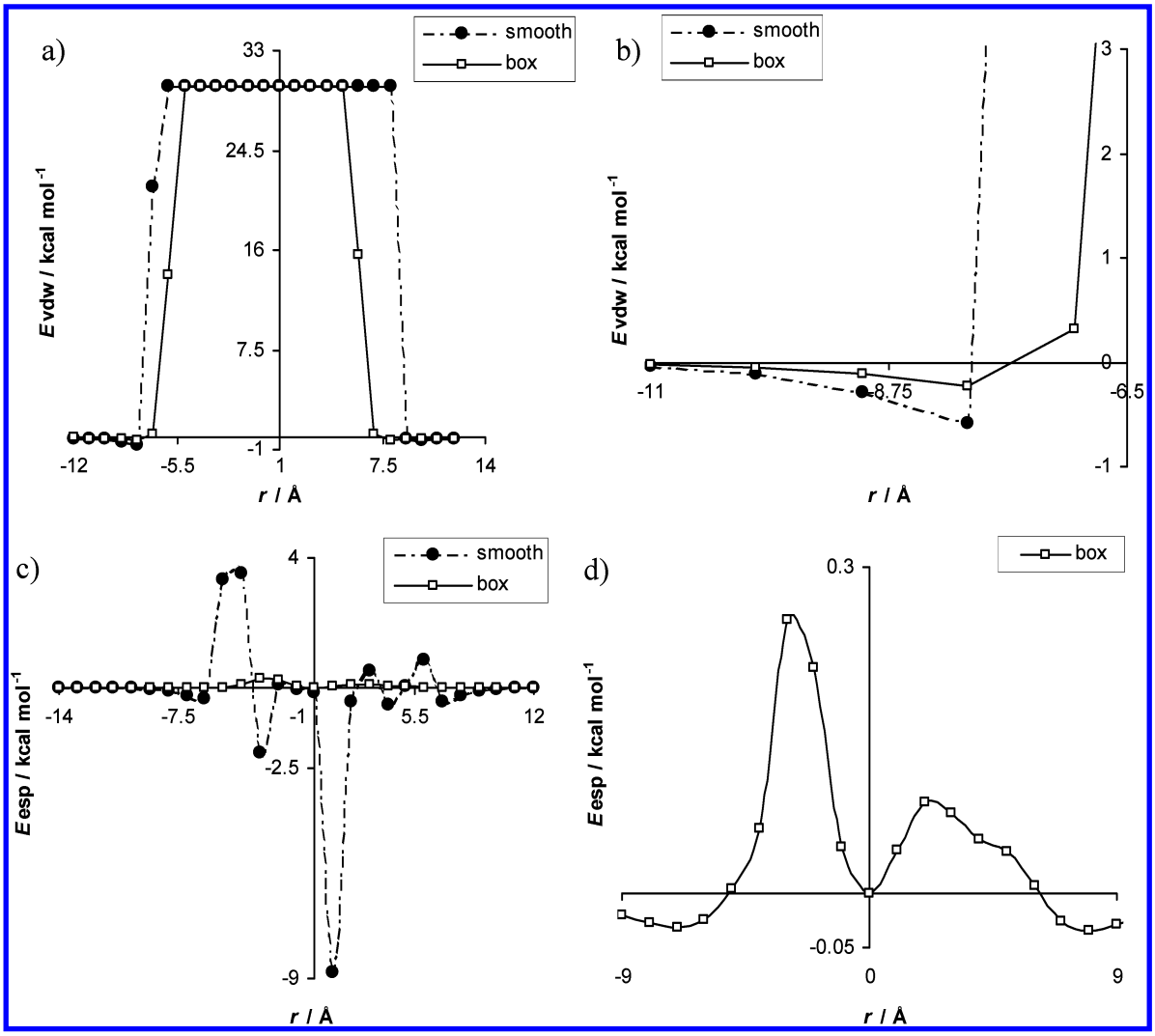

Figure 1. Molecular interaction field plot along the long axis of aldosterone. The distance, *r*, is measured relative to its centroid. The dashed line shows the change in field value using the smoothed field, and the solid line shows the effect of using the box option. The interactions are shown for (a) the steric field, (b) a closer look at the region where the steric field goes from negative to positive, indicating the interior of the molecule is being probed, (c) the electrostatic field, and (d) a plot over a smaller distance to show the variation of the box-generated field, with the smoothed field removed for clarity.

intervals along the line (*x*,0,0) were plotted using both the smooth- and box-generated fields. Figure 1a shows the change in the steric field. For both the smooth- and box-generated fields, the steric interaction rises swiftly at the molecular surface to take on a value of the cutoff, 30 kcal mol^{-1}. It can be seen that the box field rises slightly less steeply and measures a value of cutoff over a slightly smaller distance. Figure 1b shows in more detail one of the regions where the steric interaction reaches its minimum. There is little difference between the behavior of the two fields. Thus, we should perhaps expect CoMFA to be reasonably insensitive to the choice of field generation for the steric field. On the other hand, the differences between the electrostatic fields (generated using AM1 partial charges, a $1/r$ dielectric function, and a dielectric constant of 80), shown in Figure 1c,d, are more pronounced. Toward the interior of aldosterone, the smoothed field varies greatly over quite short distances. In contrast, the variation in the boxed field is imperceptible on the scale of the changes in the smoothed field. Figure 1d focuses on the change in value for the box-generated field with the smoothed field values removed to aid clarity. Some variation in field value is certainly visible but on a much smaller scale compared to smoothed field values. Hence, without column dropping, we may be likely to observe large differences between the models from a box-generated electrostatic field and from the smooth option.

Cross-Validation. For each of the three data sets, the three different methods of descriptor calculation were carried out: abrupt, smooth, and box for both electrostatic and steric fields. In all cases, as expected, the smoothed fields gave results indistinguishable from those with the abrupt cutoff; therefore, only results for the smooth and box methods are reported. The statistics reflecting the quality of the resultant models are shown in Tables 1−3 using the q^2 values with the ONC in parentheses.

Some general trends can be observed across all three data sets. The standard CoMFA setting, with both steric and electrostatic fields included, utilizing column dropping, a $1/r$ dielectric function, and the smooth cutoff results in good models in all cases but not the best model, although the difference is very small. While the box and smooth methods are comparable when considering the steric field, if using

Table 1. LOO q^2 and ONC Values for Different Field Types with the 31 CoMFA Steroids at 2 Å

field type[a]	smooth[b] q^2 (ONC)	box[b] q^2 (ONC)
S	0.71 (1)	0.71 (1)
E	0.65 (3)	0.77 (1)
Ed	0.72 (2)	0.73 (2)
SE	0.69 (3)	0.77 (1)
SEd	0.74 (1)	0.73 (1)
E-r	0.56 (2)	0.71 (2)
Ed-r	0.69 (1)	0.71 (1)
SE-r	0.65 (2)	0.74 (2)
SEd-r	0.72 (1)	0.73 (1)

[a] Key: S, steric field; E, electrostatic field with a dielectric constant of 80; d, column dropping applied to the electrostatic field; r, used a $1/r$ distance-dependent dielectric function. [b] The first number is the q^2 value. The number in parentheses represents the optimal number of components.

Table 2. LOO q^2 and ONC Values for Different Field Types with 38 D_2 Antagonists at 2 Å

field type[a]	smooth[b] q^2 (ONC)	box[b] q^2 (ONC)
S	0.73 (5)	0.76 (4)
E	0.30 (2)	0.71 (4)
Ed	0.81 (5)	0.79 (5)
SE	0.69 (6)	0.78 (4)
SEd	0.77 (4)	0.78 (4)
E-r	0.12 (2)	0.57 (6)
Ed-r	0.82 (5)	0.79 (4)
SE-r	0.54 (2)	0.79 (6)
SEd-r	0.78 (5)	0.79 (4)

[a] Key: S, steric field; E, electrostatic field with a dielectric constant of 80; d, column dropping applied to the electrostatic field; r, used a $1/r$ distance-dependent dielectric function. [b] The first number is the q^2 value. The number in parentheses represents the optimal number of components.

Table 3. LOO q^2 and ONC Values for Different Field Types with the 34 Polychlorinated Dibenzofurans at 2 Å

field type[a]	smooth[b] q^2 (ONC)	box[b] q^2 (ONC)
S	0.70 (4)	0.69 (5)
E	0.74 (4)	0.72 (2)
Ed	0.74 (2)	0.71 (3)
SE	0.72 (4)	0.74 (3)
SEd	0.74 (2)	0.75 (3)
E-r	0.58 (4)	0.65 (2)
Ed-r	0.75 (2)	0.72 (2)
SE-r	0.72 (3)	0.74 (3)
SEd-r	0.74 (2)	0.74 (3)

[a] Key: S, steric field; E, electrostatic field with a dielectric constant of 80; d, column dropping applied to the electrostatic field; r, used a $1/r$ distance-dependent dielectric function. [b] The first number is the q^2 value. The number in parentheses represents the optimal number of components.

the electrostatic field, the box method is noticeably superior, in particular for the electrostatic field on its own. The difference is most marked when using a $1/r$ dielectric function. For the D_2 antagonists (Table 2), an increase in q^2 of 0.45 units is found; there is an associated increase in ONC from 2 to 6, which appears to be a consequence of going from a nonpredictive model to a predictive one. For the polychlorinated dibenzofurans, while there is no increase in q^2, the ONC is reduced from 6 to 3, a decrease in model complexity that is to be welcomed. In all cases, the very best models are found when using the electrostatic field with column dropping and the very worst when not using column dropping. Quite remarkable improvements are observed for the D_2 antagonists. Therefore, while it would appear that for these three data sets, the steric and electrostatic fields are encoding essentially the same information about ligand−receptor interactions, it is of benefit to calculate the steric fields and use them to modify the electrostatic fields but not to include them in the regression if one is concerned about obtaining the most predictive models possible. However, as the addition of the steric field is not greatly deleterious to the model, it may be of value to include it for the purposes of interpretation.

Model Interpretation. Apart from monitoring the effect on q^2 and the ONC, another way to assess the effect of the smoothing and box method is to examine the change in the coefficients that result from the PLS regression. Due to the large number of coefficients, it is usual to examine these visually. As an example, we shall consider the CoMFA steroids and plot the 5% largest positive and 5% largest negative standardized coefficients (i.e., β*STDEV) relating to the electrostatic field using the $1/r$ dielectric function. Aldosterone was overlaid for reference. For visualization purposes, we used the program VMD.[34]

Figure 2 shows the coefficient plot without using column dropping, for (a) the smooth-generated field and (b) the box-generated field. Figure 3 shows the equivalent plots but this time using column dropping. It can be seen that there is a small difference between smooth field plots and the box field plots. Both, however, give qualitatively the same results with the positive and negative coefficients distributed in the same regions of space. By comparing Figures 2 and 3 it can also be seen how column dropping reduces the importance of lattice points close to and within the van der Waals surface of the molecules, which makes the plots slightly clearer and easier to interpret. Once again, however, the positive and negative coefficients are located in similar regions as when column dropping is not employed. In line with the reduced effect on q^2 and ONC values, the effect of either box or smooth field generation on the steric coefficient plots is small, so we do not show these.

The difference in performance between the smooth and box method appears to be due to the steep increase in electrostatic field near atomic centers. This can be concluded from the following observations: (1) the poorest models obtained for all three data sets occur when using the electrostatic field without column dropping and employing a $1/r$ dielectric, where the increase in the electrostatic field value is effectively $1/r^2$—in all cases, using the box method improves the predictive ability of the model; (2) the box method tends to produce much shallower increases in the field value (as illustrated in Figure 1); (3) column dropping, which essentially removes lattice points within the van der Waals radius of a molecule (and which would therefore likely record a large field value), substantially ameliorates the problem with the smooth fields, leading to increases in q^2 in all cases and for the CoMFA steroids and the polychlorinated dibenzofurans a reduction in the ONC. With the box method, where the variance in field value is substantially smaller, column dropping appears to have little effect and in some cases even reduces the q^2 value. From these results, it would seem that a modest increase in model quality and parsimony may be achievable by trying a constant dielectric function rather than the distance-dependent version and by

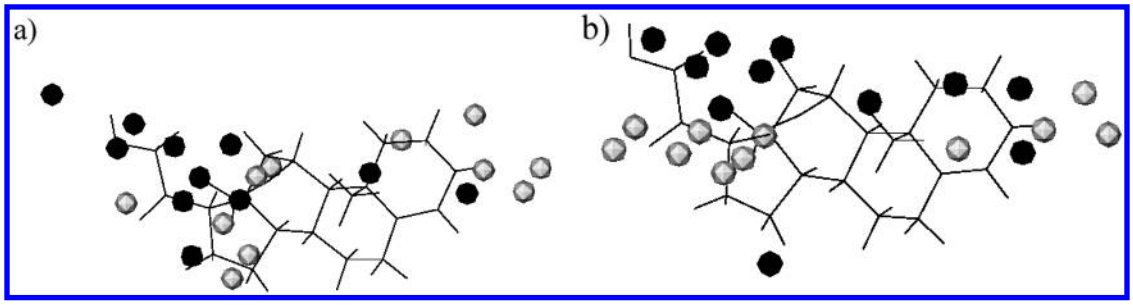

Figure 2. The 5% largest positive and 5% largest negative electrostatic standardized coefficient (β*STDEV) plots for the CoMFA steroids without column dropping and using a 1/*r* dielectric and either (a) the smooth cutoff option or (b) the box field generation option. Aldosterone is overlaid as a reference. Black spheres represent lattice points associated with positive coefficients, and white spheres represent those associated with negative coefficients.

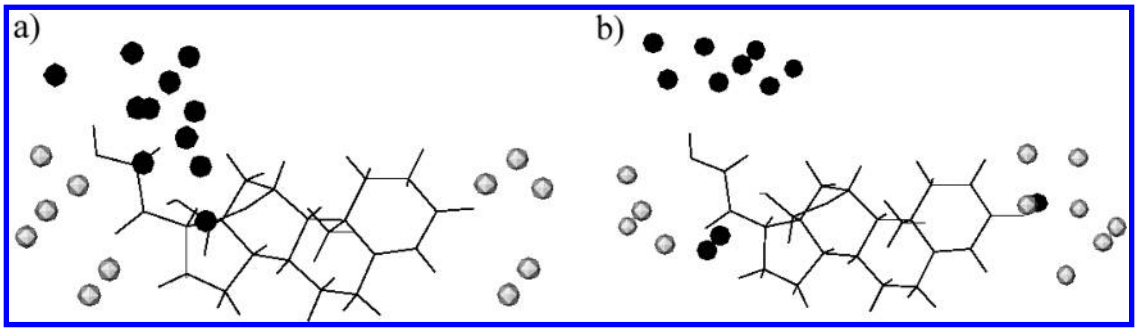

Figure 3. The 5% largest positive and 5% largest negative electrostatic standardized coefficient (β*STDEV) plots for the CoMFA steroids with column dropping and using a 1/*r* dielectric and either (a) the smooth cutoff option or (b) the box field generation option. Aldosterone is overlaid as a reference. Black spheres represent lattice points associated with positive coefficients, and white spheres represent those associated with negative coefficients.

Table 4. Mean and Standard Deviation of LOO q^2 and ONC Values for 27 Lattice Displacements for the 31 CoMFA Steroids at 2 Å

field type[a]	smooth[b] q^2 (ONC)	box[b] q^2 (ONC)
S	0.71 ± 0.01 (1.0 ± 0.0)	0.68 ± 0.01 (1.3 ± 0.5)
E	0.63 ± 0.06 (2.9 ± 0.8)	0.74 ± 0.02 (1.2 ± 0.4)
Ed	0.71 ± 0.00 (2.0 ± 0.0)	0.73 ± 0.01 (2.0 ± 0.0)
SE	0.69 ± 0.03 (2.9 ± 0.5)	0.75 ± 0.02 (1.1 ± 0.3)
SEd	0.73 ± 0.00 (1.0 ± 0.0)	0.73 ± 0.01 (1.0 ± 0.0)
E-r	0.53 ± 0.11 (2.3 ± 1.2)	0.64 ± 0.06 (2.4 ± 1.2)
Ed-r	0.69 ± 0.01 (1.5 ± 0.5)	0.71 ± 0.01 (1.4 ± 0.5)
SE-r	0.66 ± 0.05 (2.7 ± 0.9)	0.71 ± 0.03 (1.9 ± 0.7)
SEd-r	0.72 ± 0.01 (1.0 ± 0.0)	0.72 ± 0.01 (1.0 ± 0.2)

[a] Key: S, steric field; E, electrostatic field with a dielectric constant of 80; d, column dropping applied to the electrostatic field; r, used a 1/*r* distance-dependent dielectric function. [b] The first two numbers are the mean and standard deviation of the q^2 values, respectively. Numbers in parentheses are the mean and standard deviation of the optimal number of components.

Table 5. Mean and Standard Deviation of LOO q^2 and ONC Values for 27 Lattice Displacements for the 38 D_2 Antagonists at 2 Å

field type[a]	smooth[b] q^2 (ONC)	box[b] q^2 (ONC)
S	0.74 ± 0.02 (5.1 ± 0.4)	0.76 ± 0.01 (4.5 ± 0.5)
E	0.32 ± 0.28 (2.4 ± 1.4)	0.73 ± 0.05 (3.5 ± 0.7)
Ed	0.80 ± 0.01 (5.0 ± 0.0)	0.80 ± 0.01 (5.1 ± 0.3)
SE	0.67 ± 0.13 (4.6 ± 1.9)	0.78 ± 0.01 (3.9 ± 0.4)
SEd	0.77 ± 0.01 (4.0 ± 0.2)	0.78 ± 0.01 (4.1 ± 0.3)
E-r	0.20 ± 0.13 (2.2 ± 0.7)	0.47 ± 0.16 (3.8 ± 1.7)
Ed-r	0.80 ± 0.02 (4.8 ± 0.6)	0.81 ± 0.01 (4.3 ± 0.5)
SE-r	0.58 ± 0.10 (3.4 ± 1.4)	0.73 ± 0.06 (4.3 ± 1.3)
SEd-r	0.79 ± 0.01 (5.3 ± 0.7)	0.79 ± 0.00 (4.0 ± 0.0)

[a] Key: S, steric field; E, electrostatic field with a dielectric constant of 80; d, column dropping applied to the electrostatic field; r, used a 1/*r* distance-dependent dielectric function. [b] The first two numbers are the mean and standard deviation of the q^2 values, respectively. Numbers in parentheses are the mean and standard deviation of the optimal number of components.

using the box method to generate the fields (possibly without column dropping). The box method certainly gives more consistent results for the parameters considered here.

Stability to Grid Translation. Despite the possibility of improving the statistics with the box method presented above, the proposed advantage of the box method is to reduce the dependence of the results with small movements in the lattice.[12] To investigate this, the above regressions were repeated for all 27 possible combinations of either −0.1, 0, or +0.1 Å in the *x*, *y*, and *z* directions. The results are shown in Tables 4−6. Averaging the results over the 27 different lattice positions does not change the general pattern of behavior of the different models across all data sets. It can be seen that using a 1/*r* dielectric function with electrostatic and steric columns with column dropping gives close to the best results in all three cases but not always the very best. From an examination of the standard deviations, the least stable results are found when using electrostatic fields without column dropping. It can be observed that applying column dropping to these models, in addition to the superior predictive power noted in the previous section, also results in an increase in stability. For combined steric and electrostatic fields, some instability is apparent on translating the lattice, particularly using the smoothed fields for the D_2 antagonists. In these cases, once again, column dropping substantially reduces this instability. These results are in contrast to the findings of Folkers, Merz, and Rognan,[15] who reported the opposite effect in a study of HSV1 TK inhibitors. Once again, the problem appears to be traceable to the increase in electrostatic field values near atomic centers, as

Table 6. Mean and Standard Deviation of LOO q^2 and ONC Values for 27 Lattice Displacements for the 34 Polychlorinated Dibenzofurans at 2 Å

field type[a]	smooth[b] q^2 (ONC)	box[b] q^2 (ONC)
S	0.71 ± 0.01 (3.4 ± 0.6)	0.73 ± 0.02 (4.8 ± 0.5)
E	0.71 ± 0.06 (4.4 ± 1.1)	0.72 ± 0.00 (2.0 ± 0.0)
Ed	0.74 ± 0.01 (2.6 ± 0.5)	0.72 ± 0.01 (2.0 ± 0.0)
SE	0.72 ± 0.01 (3.3 ± 0.5)	0.74 ± 0.00 (3.0 ± 0.0)
SEd	0.74 ± 0.00 (2.8 ± 0.4)	0.74 ± 0.00 (3.1 ± 0.3)
E-r	0.30 ± 0.48 (4.0 ± 1.7)	0.65 ± 0.01 (2.1 ± 0.3)
Ed-r	0.75 ± 0.01 (2.1 ± 0.6)	0.73 ± 0.00 (2.0 ± 0.0)
SE-r	0.71 ± 0.02 (3.4 ± 0.8)	0.75 ± 0.00 (3.3 ± 0.5)
SEd-r	0.74 ± 0.01 (2.1 ± 0.8)	0.74 ± 0.01 (3.0 ± 0.0)

[a] Key: S, steric field; E, electrostatic field with a dielectric constant of 80; d, column dropping applied to the electrostatic field; r, used a $1/r$ distance-dependent dielectric function. [b] The first two numbers are the mean and standard deviation of the q^2 values, respectively. Numbers in parentheses are the mean and standard deviation of the optimal number of components.

column dropping or using the box method stabilizes the results. The smoother fields that are generated using CoMSIA[35] have been anticipated[36] to be more stable to lattice displacements than the standard CoMFA field, and recent work appears to confirm this.[37] This may also be indicative that the less steep increases of the box-generated fields are responsible for the greater stability. When using the default settings of smoothed fields with column dropping and a $1/r$ dielectric, there is no clear advantage to either the smoothed cutoff or the box method—for the CoMFA steroids, both methods give the same ONC; for the D_2 antagonists, the box method requires the extraction of fewer components, but this situation is reversed for the polychlorinated dibenzofurans. However, in the latter case, the box method gives consistently more stable results.

Effect of Partial Charge Calculation. Repetition of the lattice displacement study using Gasteiger−Marsili charges[38] (available in the Supporting Information) reveals the same pattern of behavior, suggesting that the observed behavior is not an artifact of the method of charge calculation. The only difference was that the quality of the regressions (as measured by q^2) was, in general, slightly lower. This finding is in line with previous observations that semiempirical charges are more accurate representations of the charge distribution in the molecules.[16]

Effect of Lattice Spacing. At a lattice spacing of 1 Å (using the AM1 charges), lattice displacements had a very small effect. Models that had performed poorly at 2 Å (e.g., smoothed electrostatic fields on their own) were as good as the steric and combined fields. It may be asked why, if as we hypothesize the steep increase in the electrostatic field near the atomic centers causes poor predictions and a lack of model stability, this effect is not seen using a 1 Å lattice? Our reasoning is as follows: when column dropping is not applied, electrostatic interactions can take large values, leading to large variances at these lattice points. It has been established that PLS, while being robust to chance correlation,[39] will not uncover all relationships between variables and is sensitive to scaling.[40] As a result, the electrostatic variables within the van der Waals surface of the molecules, despite being irrelevant to ligand−receptor interactions, will effectively "drown out" the influence of the variables with smaller variance but greater relevance. Hence, column dropping will uncover more relevant correlations in the data by reducing the influence of the electrostatic variables within the van der Waals surface. Lattice points that are located close to the "surface" of several aligned molecules will have a relatively high variance, but in contrast to electrostatic variables within the van der Waals surface of a molecule, they will be *relevant* to ligand−receptor interactions. The higher variances of these points make them less likely to be drowned out by the interior electrostatic variables.

On average, there will be more of these variables with a 1 Å lattice, and therefore, column dropping is less vital at a higher resolution; we note, however, that column dropping improves the predictive quality and stability of the models more often than not at 1 Å as well as at coarser resolutions. Furthermore, no 1-Å model was superior to the best 2-Å model, which suggests that in the majority of cases any extra "signal" to the model was accompanied by an equal proportion of "noise". Thus, for these data sets, little seems to be gained from the extra computational burden of using more lattice points, compared to a careful choice of parameters at 2 Å. These results are also given in the Supporting Information.

CONCLUSION

In this study we considered the effect of different methods of field generation available to users of CoMFA over three different data sets. These have implications for both the quality of the models produced and their robustness with respect to small lateral displacements of the lattice. The default CoMFA setup (2 Å spacing, including steric and electrostatic settings, dropping electrostatic points at large steric values, use of a $1/r$ dielectric function) gave good quality models that were relatively insensitive to the lattice movement. However, better models in some cases were obtained with the electrostatic fields by use of the constant dielectric function, and if a nondefault setting is chosen (in particular, if electrostatic fields are used on their own), then rather different behavior may be observed—the results are normally a lot more sensitive to the movement of the lattice. As there can be a large difference in results obtained with the box method, on one hand, and the abrupt or smooth methods, on the other, we also urge practitioners to state which method they use. The box method attenuates the variability of the models, resulting in, on average, larger q^2 values than the equivalent field generated with the smoothed cutoff field. While fewer components were not always extracted, this was observed in some cases, so it may well be worth repeating the standard CoMFA analysis with the box method as a matter of course when using this 3D QSAR method. The box method therefore represents a useful, simple, and immediately applicable method that can potentially increase the quality of the PLS model. To some extent, it may also obviate the need for a tedious iterative procedure often employed, wherein the grid is moved with respect to the aligned molecules until an "optimum" positioning is found. In the absence of validation with external predictions, we would question the justification for choosing one particular placement out of all the possible placements simply on the basis of the largest q^2 value. For these reasons, we suggest that users of CoMFA consider the box method for generating the field descriptors.

ACKNOWLEDGMENT

We thank the Gatsby Foundation for funding and EPSRC for an equipment grant (GR/R62052/01) for computers.

Supporting Information Available: Structures and observed activity values for data sets 1−3; tabulated results of the grid displacements for all three data sets repeated at a lattice spacing of 1 Å with AM1 semiempirical charges and at 2 Å with Gasteiger−Marsili charges. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

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

(2) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(3) Allen, M. S.; LaLoggia, A. J.; Dorn, L. J.; Martin, M. J.; Constantino, G.; Hagen, T. J.; Koehler, K. F.; Skolnick, P.; Cook, J. M. Predictive Binding of β-Carboline Inverse Agonists and Antagonists via the CoMFA/GOLPE Approach. *J. Med. Chem.* **1992**, *35*, 4001−4010.

(4) Kroemer, R. T.; Hecht, P. Replacement of steric 6−12 potential-derived interaction energies by atom-based indicator variables in CoMFA leads to models of higher consistency. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 205−212.

(5) Sulea, T.; Oprea, T. I.; Muresan, S.; Chan, S. L. A Different Method for Steric Field Evaluation in CoMFA Improves Model Robustness. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1162−1170.

(6) Kellogg, G. E. Finding Optimum Field Models for 3D QSAR. *Med. Chem. Res.* **1997**, *7*, 417−427.

(7) Bucholtz, E. C.; Tropsha, A. The Effect of Region Size on CoMFA Analyses. *Med. Chem. Res.* **1999**, *9*, 675−685.

(8) Cho, S. J.; Tropsha, A. Cross-Validated R^2-Guided Region Selection for Comparative Molecular Field Analysis: A Simple Method To Achieve Consistent Results. *J. Med. Chem.* **1995**, *38*, 1060−1066.

(9) Brusniak, M.-Y. K.; Pearlman, R. S.; Neve, K. A.; Wilcox, R. E. Comparative Molecular Field Analysis-Based Prediction of Drug Affinities at Recombinant D1A Dopamine Receptors. *J. Med. Chem.* **1996**, *39*, 850−859.

(10) Kroemer, R. T.; Hecht, P. A new procedure for improving the predictiveness of CoMFA models and its application to a series of dihydrofolate reductase inhibitors. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 396−406.

(11) Wang, R. X.; Gao, Y.; Liu, L.; Lai, L. H. All-Orientation Search and All-Placement Search in Comparative Molecular Field Analysis. *J. Mol. Model.* **1998**, *4*, 276−283.

(12) Cramer, R. D., III; DePriest, S. A.; Patterson, D. E.; Hecht, P. E. The Developing Practice of Comparative Molecular Field Analysis. In *3D QSAR in drug design−theory, methods and applications*; Kubinyi, H., Ed.; ESCOM Science Pub.: Leiden, 1993; pp 443−485.

(13) Kroemer, R. T.; Ettmayer, P.; Hecht, P. 3D-Quantitative Structure−Activity Relationships of Human Immunodeficiency Virus Type-1 Proteinase Inhibitors: Comparative Molecular Field Analysis of 2-Heterosubstituted Statine Derivatives−Implications for the Design of Novel Inhibitors. *J. Med. Chem.* **1995**, *38*, 4917−4928.

(14) Wilcox, R. E.; Huang, W.-H.; Brusniak, M.-Y. K.; Wilcox, D. M.; Pearlman, R. S.; Teeter, M. M.; DuRand, C. J.; Wiens, B. L.; Neve, K. A. CoMFA-Based Prediction of Agonist Affinities at Recombinant Wild-Type versus Serine to Alanine Point Mutated D2 Dopamine Receptors. *J. Med. Chem.* **2000**, *43*, 3005−3019.

(15) Folkers, G.; Merz, A.; Rognan, D. CoMFA: Scope and Limitations. In *3D QSAR in drug design−theory, methods and applications*; Kubinyi, H., Ed.; ESCOM Science Pub.: Leiden, 1993; pp 583−618.

(16) Kim, K. H.; Greco, G.; Novellino, E. A critical review of recent CoMFA applications. *Perspect. Drug Discovery Des.* **1998**, *12/13/14*, 257−315 and references therein.

(17) Cramer, R. D. Topomer CoMFA: A design methodology for rapid lead optimization *J. Med. Chem.* **2003**, *46*, 374−388.

(18) Nicklaus, M. C.; Milne, G. W. A.; Burke, T. R., Jr. QSAR of conformationally flexible molecules: Comparative molecular field analysis of protein-tyrosine kinase inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 487−504.

(19) Broughton, H. B.; Gordaliza, M.; Castro, M.-A.; Miguel del Corral, J. M.; San Feliciano, A. Modified CoMFA methods for the analysis of antineoplastic effects of lignan analogues. *J. Mol. Struct.* (*THEOCHEM*) **2000**, *504*, 287−294.

(20) Lukacova, V.; Balaz, S. Multimode Ligand Binding in Receptor Site Modeling: Implementation in CoMFA. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2093-2105.

(21) Coats, E. A. The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. *Perspect. Drug Discuss. Des.* **1998**, *12/13/14*, 199−213.

(22) http://www2.chemie.uni-erlangen.de/services/steroids/index.html.

(23) Boström, J.; Böhm, M.; Gundertofte, K.; Klebe, G. A 3D QSAR Study on a Set of Dopamine D4 Receptor Antagonists. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1020−1027.

(24) Bradley, M.; Waller, C. L. Polarizability Fields for Use in Three-Dimensional Quantitative Structure−Activity Relationship (3D-QSAR). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1301−1307.

(25) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graph. Model.* **2002**, *20*, 269−276.

(26) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing Model Fit by Cross-Validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579−586.

(27) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1−12.

(28) Kong J.; White, C. A.; Krylov, A. I.; Sherrill, D.; Adamson, R. D.; Furlani, T. R.; Lee, M. S.; Gwaltney, S. R.; Adams, T. R.; Ochsenfeld, C.; Gilbert, A. T. B.; Kedziora, G. S.; Rassolov, V. A.; Maurice, D. R.; Nair, N.; Shao, Y.; Besley, N. A.; Maslen, P. E.; Dombroski, J. P.; Daschel, H.; Zhang, W.; Korambath, P. P.; Baker, J.; Byrd, E. F. C.; Van Voorhis, T.; Oumi, M.; Hirata, S.; Hsu, C.-P.; Ishikawa, N.; Florian, J.; Warshel, A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M.; Pople, J. A. Q-Chem 2.0: A high-performance ab initio electronic structure program package. *J. Comput. Chem.* **2000**, *21*, 1532−1548.

(29) Clark, M.; Cramer, R. D., III; Van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10*, 982−1012.

(30) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(31) de Jong, S. SIMPLS: an alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251−263.

(32) Shao, J. Linear Model Selection by Cross-Validation. *J. Am. Stat. Ass.* **1993**, *88*, 486−494.

(33) Kubinyi, H.; Abraham, U. Practical Problems in PLS Analyses. In *3D QSAR in drug design−theory, methods and applications*; Kubinyi, H., Ed.; ESCOM Science Pub.: Leiden, 1993; pp 717−728.

(34) Humphrey, W.; Dalke, A.; Schulten, K. VMD−Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33−38.

(35) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130−4146.

(36) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. Three-Dimensional Quantitative Similarity−Activity Relationships (3D QSiAR) from SEAL Similarity Matrices. *J. Med. Chem.* **1998**, *41*, 2553−2564.

(37) Hou, T. J.; Xu, X. J. Three-dimensional quantitative structure−activity relationship analyses of a series of cinnamamides. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 123−132.

(38) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity−A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(39) Clark, M.; Cramer, R. D. The probability of chance correlation using Partial Least Squares (PLS). *Quant. Struct.−Act. Relat.* **1993**, *12*, 137−145.

(40) Ortiz, A. R.; Pastor, M.; Palomer, A.; Cruciani, G.; Gago, F.; Wade, R. C.; Reliability of comparative molecular field analysis models: Effects of data scaling and variable selection using a set of human synovial fluid phospholipase A(2) inhibitors. *J. Med. Chem.* **1997**, *40*, 1136−1148.

CI049944O