*Journal of Computer-Aided Molecular Design*, **15:** 659–669, 2001.
*KLUWER/ESCOM*



# Developing $^{13}$C NMR quantitative spectrometric data-activity relationship (QSDAR) models of steroid binding to the corticosteroid binding globulin

Richard D. Beger* & Jon G. Wilkes
*Division of Chemistry, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR 72079, USA*

Received 8 September 2000; accepted 25 May 2001

*Key words:* computer modeling, corticosteroid binding globulin, QSDAR, CoSA, CoSASA, $^{13}$C NMR

**Summary**

We have developed four quantitative spectrometric data-activity relationship (QSDAR) models for 30 steroids binding to corticosteroid binding globulin, based on comparative spectral analysis (CoSA) of simulated $^{13}$C nuclear magnetic resonance (NMR) data. A QSDAR model based on 3 spectral bins had an explained variance ($r^2$) of 0.80 and a cross-validated variance ($q^2$) of 0.78. Another QSDAR model using the 3 atoms from the comparative structurally assigned spectral analysis (CoSASA) of simulated $^{13}$C NMR on a steroid backbone template gave an explained variance ($r^2$) of 0.80 and a cross-validated variance ($q^2$) of 0.73. Positions 3 and 14 from the steroid backbone template have correlations with the relative binding activity to corticosteroid binding globulin that are greater than 0.52. The explained correlation and cross-validated correlation of these QSDAR models are as good as previously published quantitative structure-activity relationship (QSAR), self-organizing map (SOM) and electrotopological state (E-state) models. One reason that the cross-validated variance of QSDAR models were as good as the other models is that simulated $^{13}$C NMR spectral data are more accurate than the errors introduced by the assumptions and approximations used in calculated electrostatic potentials, E-states, HE-states, and the molecular alignment process of QSAR modeling. The QSDAR models developed provide a rapid, simple way to predict the binding activity of a steroid to corticosteroid binding globulin.

## Introduction

Many different types of models have been developed to predict the binding activity for the compound-receptor system of the corticosterone binding globulin [1]. These corticosteroid binding globulin models include the standard quantitative structure-activity relationship (QSAR) [2], the hybrid electrotopological state (E-state) [3], the self-organizing map (SOM) [4], and the combination QSAR E-state models [5]. Previously, we have demonstrated that $^{13}$C NMR and electron ionization mass spectrometric (EI MS) spectrometric data can be used to produce a reliable classification for spectrometric data-activity relationship (SDAR) models of the estrogen receptor system [6]. The combination of $^{13}$C NMR, infrared absorption (IR), and (EI MS) data has been used to produce a reliable classification SDAR model of monodechlorination rates [7]. The binding activity of 45 progestagens has been quantitatively modeled with simulated $^{13}$C NMR spectra by comparative spectral analysis (CoSA) methods [8]. The binding activity to the aryl hydrocarbon receptor of 26 polychlorinated dibenzofurans, 14 polychlorinated dibenzo-p-dioxins, and 12 polychlorinated biphenyls has been quantitatively modeled with simulated $^{13}$C NMR spectra by CoSA methods [9]. This CoSA model using simulated $^{13}$C NMR data yielded higher cross-validated correlations than were seen with comparative molecular field analysis (CoMFA) methods. However, the simulated $^{13}$C NMR spectrometric data were treated as if they were real experimental data without taking into account the errors produced in the simulated $^{13}$C NMR data and

*To whom correspondence should be addressed: E-mail: rbeger@NCTR.FDA.GOV

had very low leave-one-out cross-validated correlations. In this paper, we demonstrate by accounting for errors produced that simulated $^{13}$C NMR spectral data can be used to produce quantitative spectrometric data-activity relationship (QSDAR) models of steroids binding to the corticosterone binding globulin by CoSA methods.

Nishikawa and Tori showed that the reactivity of the β-lactam ring of 3-substituted and 3-methylene substituted cephalosporins toward alkaline hydrolysis had a correlation with changes in the $^{13}$C NMR chemical shift of selected carbon atoms [10]. A similar attempt to utilize $^{13}$C NMR chemical shift derived parameters as structure descriptors was made by IJzerman et al. who derived structure descriptors from the assigned $^{13}$C NMR shifts of the aromatic carbons in a family of N-*tert*-butylphenylethanolamines [11]. IJzerman et al. subtracted the value of the chemical shift for the carbon atoms in benzene from the assigned chemical shift values for each of the carbons in the benzene ring of each of the substituted N-*tert*-butylphenylethanolamines to yield sets of structure descriptors for the compounds. A correlation was found between certain linear combinations of these structure descriptors and the $\beta_2$-andrenoreceptor intrinsic activity of the substituted N-*tert*-butylphenylethanolamines. Both papers showed that the models produced from atomically assigned $^{13}$C NMR spectra were better than models produced by quantum mechanical structural parameters [10, 11].

The power of QSDAR is that it is unnecessary to solve any quantum mechanic calculations or use the structures of molecules for electrostatic calculations as is done in QSAR techniques [2, 12–16]. QSAR is based on the assumption that there is a relationship between structure and activity of a compound. QSAR modeling results show that receptor binding of a compound can be predicted from a combination of electrostatics potentials and geometrical structural analysis [2, 8, 15, 16].

Electrotopological states (E-states) were developed to describe in one calculation the combination of electronic composition and steric environment for every non-hydrogen atom in the molecule [3]. E-states take into account the atom type, its valence state, and degree of adjacency, which are then calculated in a three-dimensional grid map around the molecule for every atom at every grid point. The distance of the atom to the grid point is multiplied by a function of the distance that is supposed to take into account the behavior of electrostatics. E-state QSAR models use many different distance functions to produce their final models. The hydrogen electrotopological state (HE-state) index [3] increases the accuracy of QSAR models based on E-states but drastically increases the number of calculations needed to produce the model. A combination of QSAR with E-states modeling reduces the number of calculations needed to produce a QSAR-E-state model while keeping the strong leave-one-out (LOO) cross-validated $r^2$ [5]. The QSAR/ E-state combination model reduced the number of calculations by removing the grid and applying their E-state calculations to only specific atoms on the backbone of steroids.

Self-organizing maps (SOM) were produced to reduce the dimensionality of the input into neural networks QSAR models [4]. SOM are two-dimensional maps representing a transformation of three-dimensional molecules. SOM may contain molecular van der Waals surfaces and electrostatic potentials of a reference molecule that is used for comparison to other molecules.

$^{13}$C nuclear magnetic resonance (NMR) chemical shifts have been used to predict and refine chemical structures [17, 18]. Conversely, the chemical structure of a compound has been used to predict its $^{13}$C NMR chemical shifts [19]. The $^{13}$C NMR spectrum of a compound contains frequencies that correspond directly to the quantum mechanical properties of the molecule. The quantum mechanical description of a molecule depends largely on its electrostatic features and geometry [20]. *Ab initio* quantum mechanical calculations of $^{13}$C chemical shift tensors in proteins reveal that they are dependent on the structural environment [21]. ADC Labs now sells software that will take a $^{13}$C NMR one-dimensional spectrum and predict a structure [22].

The frequencies obtained from $^{13}$C NMR spectroscopic data correspond directly to the energies obtained when solving the quantum mechanical Schrodinger equation for a nuclear magnetic moment transition [20]. The NMR chemical shifts (quantum energies) are strongly dependent on the electrostatic potential energy of the carbon nucleus and the type of orbital (wave function) surrounding the carbon nucleus. The magnetic and kinetic energy terms in the quantum mechanical calculation of the nuclear magnetic dipole transition are small compared to the electrostatic component. The $^{13}$C NMR spectrum is similar to the E-state index calculations because they are both largely dependent on electrostatics and valence states. E-states are calculated at every point in

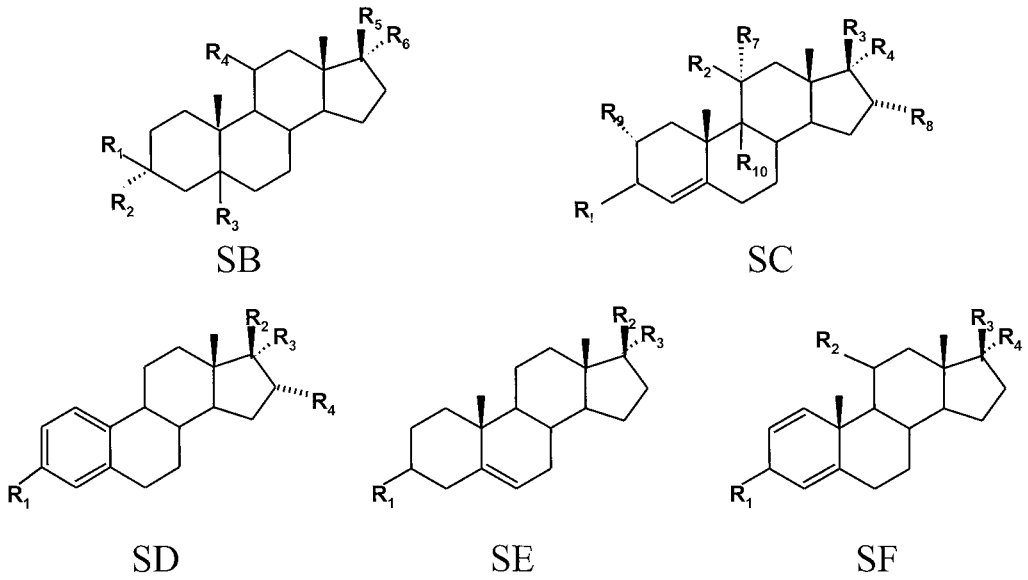


*Figure 1.* Structures SA - SE used with Table 1 for the corticosteroid binding globulin steroid series.

a three-dimensional grid while using many different distance functions, but NMR uses the laws of nature to instantly 'calculate' the quantum mechanical magnetic moment energy of every carbon nucleus in the molecule without using approximations for partial charges, dielectric constants, or wave functions.

Typically, $^{13}$C NMR chemical shifts in the 0 to 100 ppm range are associated with carbon atoms that have $sp^3$ orbitals, with the more upfield shifts having a positive electrostatic potential (like methyl groups) and the downfield shifts having a more negative electrostatic potential (like ester bonds). Likewise, $^{13}$C NMR chemical shifts in the 100 to 220 ppm range are associated with carbon atoms that have $sp^2$ and sp orbitals, with the more upfield shifts having a positive electrostatic potential (like benzyl groups) and the downfield shifts having a more negative electrostatic potential (like carbonyl groups). The effect of substituents on $^{13}$C NMR chemical shifts can be felt from as far as five bonds away or directly through space. The absolute energies in NMR spectra are not used in QSDAR because NMR spectra are given as parts per million (ppm) chemical shifts that are dimensionless numbers defined with respect to a reference compound.

Using QSAR modeling results, receptor binding of a compound can be predicted, based in part upon electrostatics and geometrical structure [15, 16, 23]. Therefore, we postulated that we could use $^{13}$C NMR data in much the same way that QSAR uses constitutional, topological, geometrical, electrostatic, and quantum descriptors to model receptor binding of a compound with comparative molecular field analysis (CoMFA) [12–16, 23]. By combining the simulated $^{13}$C NMR data into a composite set of descriptors and putting them into statistical software programs for comparative spectral analysis, it is possible to produce a QSDAR model of the corticosteroid globulin binding compounds.

Current quantitative structure-activity relationship (QSAR) and structure-activity relationship (SAR) models use computer modeling of the molecule or break the molecule into secondary structural pieces [12–16, 24, 25]. Many calculations are used in QSAR, SOM or E-states models. Using a specific molecular structure for computer modeling of each compound dramatically extends the number of calculations required to define the model. Moreover, the selection of the most appropriate 3D structure for each molecule requires a number of assumptions. The necessary simplifying assumptions in some cases give results that are hard to replicate or are inaccurate.

In QSDAR, each NMR chemical shift or IR absorption peak functions as a quantum mechanical identifier of a 4 to 8 atoms secondary structural moiety. These quantum mechanical identifiers are used in a manner similar to that in current QSAR and SAR models that break the molecule into secondary structural motifs.

*Table 1.* Structures of corticosteroids used in QSDAR models of corticosteroid binding globulin data.

| # | Structure[a] | $R_1$ | $R_2$ | $R_3$ | $R_4$ | $R_5$ | $R_6$ | $R_7$ | $R_8$ | $R_9$ | $R_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | SB | OH | H | H | H | OH | H | | | | |
| 2 | SE | OH | OH | H | | | | | | | |
| 3 | SC | =O | H | =O | | | | H | H | H | H |
| 4 | SB | H | OH | H | H | =O | | | | | |
| 5 | SC | =O | OH | $COCH_2OH$ | H | | | H | H | H | H |
| 6 | SC | =O | OH | $COCH_2OH$ | OH | | | H | H | H | H |
| 7 | SC | =O | =O | $COCH_2OH$ | OH | | | | H | H | H |
| 8 | SE | OH | =O | | | | | | | | |
| 9 | SC | =O | H | $COCH_2OH$ | H | | | H | H | H | H |
| 10 | SC | =O | H | $COCH_2OH$ | OH | | | H | H | H | H |
| 11 | SB | =O | | H | H | OH | H | | | | |
| 12 | SD | OH | OH | H | H | | | | | | |
| 13 | SD | OH | OH | H | OH | | | | | | |
| 14 | SD | OH | =O | | H | | | | | | |
| 15 | SB | H | OH | H | H | =O | | | | | |
| 16 | SE | OH | COMe | H | | | | | | | |
| 17 | SE | OH | COMe | OH | | | | | | | |
| 18 | SC | =O | H | COMe | H | | | H | H | H | H |
| 19 | SC | =O | H | COMe | OH | | | H | H | H | H |
| 20 | SC | =O | H | OH | H | | | H | H | H | H |
| 21 | SF | =O | OH | $COCH_2OH$ | OH | | | | | | |
| 22 | SC | =O | OH | $COCH_2OCOMe$ | | | | H | H | H | H |
| 23 | SC | =O | =O | COMe | H | | | | H | H | H |
| 24 | SC | =O | H | $COCH_2OH$ | H | | | OH | H | H | H |
| 25[b] | SC | =O | H | OH | H | | | H | H | H | H |
| 26 | SC | =O | H | COMe | OH | | | H | OH | H | H |
| 27 | SC | =O | H | COMe | H | | | H | Me | H | H |
| 28[b] | SC | =0 | H | COMe | H | | | H | H | H | H |
| 29 | SC | =O | OH | $COCH_2OH$ | OH | | | H | H | Me | H |
| 30 | SC | =O | OH | $COCH_2OH$ | OH | | | H | H | Me | F |

[a]Structures according to references [1, 22, 23].
[b]H (hydrogen) instead of Me at $C_{10}$ steroid skeleton.

## Procedures

The 30 compounds specified in Table 1 and Figure 1 have known steroid inhibitor binding affinities to the aromatase enzyme [1–5, 26, 27] shown in Table 2. All the simulated $^{13}C$ NMR spectra were simulated using the ACD Labs CNMR predictor software, version 4.0 [22]. We used the $^{13}C$ NMR data points in the same way that QSAR uses comprehensive descriptors for structural and statistical analyses (CODESSA) [28].

One QSDAR model was produced by using the assigned $^{13}C$ NMR chemical shifts at the 20 positions in the steroid backbone template, as shown in Figure 2. This requires 20 'bins' in which the corresponding intensity is each carbon's simulated $^{13}C$ NMR chemical shift. This model combines structural information with the assigned simulated $^{13}C$ NMR chemical shifts. We like to call this procedure comparative structurally assigned spectra analysis (CoSASA). Two CoSASA QSDAR models were developed, one

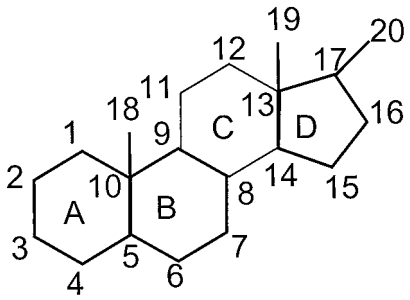


*Figure 2.* Steroid template carbon atom numbering.

*Table 2.* Real and predicted binding activities (BA) to the corticosterone binding globulin.

| # | Binding activity | CoSASA (3 atoms) | CoSASA (3 PCs) | CoSA (3 shifts) | CoSA (3 PCs) |
|---|---|---|---|---|---|
| 1 | 5.00 | 4.89 | 5.16 | 5.92 | 5.30 |
| 2 | 5.00 | 4.91 | 5.74 | 5.12 | 5.40 |
| 3 | 5.76 | 6.54 | 6.12 | 6.14 | 6.67 |
| 4 | 5.61 | 4.81 | 4.76 | 5.08 | 5.63 |
| 5 | 7.88 | 7.55 | 7.47 | 7.76 | 7.68 |
| 6 | 7.88 | 7.08 | 7.03 | 6.92 | 7.50 |
| 7 | 6.89 | 6.99 | 7.49 | 7.16 | 7.63 |
| 8 | 5.00 | 5.02 | 5.10 | 5.12 | 5.70 |
| 9 | 7.65 | 7.40 | 7.18 | 6.97 | 7.78 |
| 10 | 7.88 | 6.92 | 6.78 | 7.04 | 6.95 |
| 11 | 5.92 | 6.74 | 6.17 | 6.10 | 6.06 |
| 12 | 5.00 | 5.06 | 5.33 | 5.12 | 4.90 |
| 13 | 5.00 | 5.09 | 5.43 | 5.12 | 5.28 |
| 14 | 5.00 | 5.16 | 4.34 | 5.80 | 5.32 |
| 15 | 5.23 | 4.88 | 4.85 | 5.10 | 5.30 |
| 16 | 5.23 | 6.37 | 6.20 | 5.10 | 6.21 |
| 17 | 5.00 | 5.67 | 5.78 | 5.12 | 5.57 |
| 18 | 7.38 | 7.50 | 7.17 | 7.11 | 6.64 |
| 19 | 7.74 | 6.93 | 6.78 | 8.28 | 6.71 |
| 20 | 6.72 | 6.32 | 6.80 | 6.09 | 6.71 |
| 21 | 7.51 | 6.93 | 6.64 | 7.86 | 6.85 |
| 22 | 7.55 | 7.08 | 7.11 | 6.99 | 7.93 |
| 23 | 6.78 | 7.44 | 7.98 | 7.17 | 7.40 |
| 24 | 7.20 | 7.56 | 6.92 | 7.92 | 6.89 |
| 25 | 6.14 | 6.25 | 6.88 | 6.08 | 5.26 |
| 26 | 6.25 | 6.96 | 6.99 | 7.24 | 7.10 |
| 27 | 7.12 | 7.43 | 6.37 | 7.09 | 5.27 |
| 28 | 6.82 | 7.10 | 7.12 | 7.08 | 6.75 |
| 29 | 7.69 | 7.13 | 7.07 | 7.06 | 7.39 |
| 30 | 5.80 | 6.76 | 7.20 | 6.14 | 6.05 |

model was based on the three individual assigned carbon atom chemical shifts and the other model was based on three principal components built from all the assigned carbon atom chemical shifts in the steroid backbone.

In another QSDAR model that did not use the steroid backbone template, we used the unassigned simulated $^{13}$C NMR data points for CoSA. Unassigned 1D $^{13}$C NMR chemical shifts were segregated into bins over a 0 to 256 ppm range. The CoSA QSDAR models were produced with a spectral width of the bins that was 1.0 ppm. The $^{13}$C NMR spectra were saved as the area under the peak within a certain spectral range and normalized to an integer. A single chemical shift frequency in the 1.0 ppm spectral bin was assigned an area of 100, two chemical shifts in the 1.0 ppm spectral bin had an area of 200, and so forth. This was done so that all the spectra would have a similar signal-to-noise ratio and to reduce the error produced in the chemical shift prediction of the $^{13}$C NMR spectra for a given compound. The spectral width of 1.0 ppm was used because of convenience and 1.0 ppm is approximately twice the average uncertainty of 0.53 ppm from the predicted $^{13}$C NMR data of all the carbon nuclei in the 30 compounds. The most important reason to use a 1 ppm or larger spectral bin width is to collect enough chemical shift hits in the bin to produce strong and reliable statistical correlations to build a PLS model on. In the 1.0 ppm resolution CoSA model, there are 256 bins, each of which is populated or not depending on the pattern of simulated chemical shifts. This approach does not require an identification of the shift with the carbon that produced it. Therefore, its predictions are not limited to compounds having a steroid backbone.

The analysis of each $^{13}$C NMR QSDAR model was done by the leave-one-out (LOO) cross-validation procedure where each compound is systematically excluded from the training set and its binding activity is predicted by the model [29]. The cross-validated $r^2$ (termed $q^2$) can be derived from $q^2 = 1-(\text{PRESS})/\text{SD}$. Where PRESS is the sum of the differences between the actual and predicted activity data for each molecule during LOO cross-validation, and SD is the sum of the squared deviations between the measured and mean activities of each molecule in the training set.

The pattern recognition software used was Statistica version 5.5 [30]. The simulated $^{13}$C NMR spectroscopic data for all 30 corticosteroid binding compounds in Table 1 were input into the software. Figure 2 shows the steroid backbone template that was used to produce a QSDAR model for the corticosteroid binding globulin. The steroid backbone template QSDAR model used only the simulated $^{13}$C NMR chemical shifts for the 20 positions found in the steroid backbone template. When there was no carbon atom in positions 18, 19, and 20 a zero was entered for the $^{13}$C NMR chemical shift. Two QSDAR models that are produced when using the steroid backbone template, one using partial least squares (PLS) multiple regression analysis of 3 atoms (bins) and the other with 3 principal components based on all 20 chemical shifts. The second type of QSDAR model is produced when the spectral width of the $^{13}$C NMR

data is 1.0 ppm. Before PLS regression analysis was done on the simulated $^{13}C$ NMR data, the spectral bin columns with all zeroes and spectral bin columns with only one non-zero number were removed from the data set input to the Statistica software program. We removed the bins with only one hit in the bin because when these bins are used in LOO cross-validation a correlation matrix can not be obtained. The 1.0 ppm resolution QSDAR models were evaluated with PLS multiple regression analysis using only the most correlated 3 spectral bins and the other with 3 not correlated principal components based on the all the spectral bins.

## Results

Table 2 contains the actual binding activity of the 30 steroidal compounds and the predicted binding activity from the two CoSASA and two CoSA QSDAR modles. Table 3 contains a comparison of the model performance parameters $n$, $r^2$, $q^2$, and number of components for the QSAR [2], HE-state/E-state [3], E-state [3], SOM [4], combination QSAR/E-state [5] and four QSDAR models. All four QSDAR models have a strong correlation ($r^2$) and cross-validated variance ($q^2$) and are favorable when compared to the published models of previously the corticosteroid binding globulin. The statistical results were further tested and validated by randomizing the binding activity data and the best statistical correlation occurred using actual binding data. The explained correlation ($r^2$) between three completely randomized binding data sets and simulated $^{13}C$ NMR data was always less than 0.12 for all four QSDAR models.

Figure 3A is a plot of the predicted binding versus experimental binding for the 20 atom models when using only the 3 most correlated carbon atom's chemical shifts. The NMR chemical shifts used in this model were from the steroid template positions 3, 14, and 20. The explained correlation ($r^2$) of this model is 0.80 and the cross-validated variance ($q^2$) is 0.73, which indicates self-consistency and good predictive capability. Figure 3B is a plot of the predicted binding versus experimental binding for 20 atom models using the top 3 principal components, each based on all 20 atoms. The explained correlation ($r^2$) is 0.68 and the cross-validated variance ($q^2$) is 0.60, which indicates self-consistency and average predictive capability.

Figure 4A is a plot of the predicted binding versus experimental binding for the 1.0 ppm resolution models when using only the 3 most correlated spectral bins. The explained correlation ($r^2$) of this model is 0.80 and the cross-validated variance ($q^2$) is 0.78, which indicates strong self-consistency and very good predictive capability. The 3 spectral bins corresponded to the chemical shift frequencies of 32.00–32.99, 43.00–43.99, and 55.00–55.99 ppm. The 32 ppm spectral bin was usually associated with positions 6 and 7 on the steroid template but this fact was not used in the correlation analysis. Similarly, the 43 ppm spectral bin was associated primarily with position 13 on the steroid template and the 55 ppm spectral bin had a large correlation with position 9 on the steroid template. Figure 4B is a plot of the predicted binding versus experimental binding for the 1.0 ppm resolution models when using 3 principal components. The explained correlation ($r^2$) of this model is 0.71 and the cross-validated variance ($q^2$) is 0.65, which indicates self-consistency and good predictive capability.

The steroid backbone QSDAR model was based on the chemical shifts frequency change of atoms from carbon atoms at the 3, 14, and 20 positions on the template. The 1.0 ppm resolution QSDAR model was based on the chemical shifts frequencies that for the most part watched for the chemical shifts of atoms in the 6, 7, 9, and 13 positions to go into a certain frequency range. The reason the steroid backbone and 1.0 ppm resolution QSDAR models are not based on the same atoms are that the chemical shift frequencies from the atoms in the steroid backbone model may fall into chemical shift frequency ranges that do not have a high correlation to the binding activity. All the QSDAR models of corticosteroid binding globulin had a standard error (SE) between 0.5 and 0.7 and $p < 0.00001$. The best performing QSDAR models are the ones based on 1.0 ppm resolution using either the 3 most significant spectral bins or the first 3 principal components of variation.

Figure 5 shows the explained variance ($r^2$) and cross-validated variance ($q^2$) results for the CoSASA QSDAR models as a function of the (Figure 5A) number of atoms or (Figure 5B) number of principal components. Figure 6 shows the explained variance ($r^2$) and cross-validated variance ($q^2$) results for the 1.0 ppm resolution CoSA QSDAR models versus the (Figure 6A) number of spectral bins or (Figure 6B) number of principal components. Figures 5 and 6 show that the explained variance and cross-validated variance continues to rise together with the addition of

*Table 3.* Model performance parameters $n$, $r^2$, $q^2$, and number of components

| Model | $n$ | $r^2$ | $q_1^2$ | Components |
|---|---|---|---|---|
| QSAR (2) | 31 | 0.72 | 0.68[a] | 3 (PCs) |
| HE state/ E-state (3) | 31 | 0.98[a]/0.96[b] | 0.80[a]/0.76[b] | 3[a](PCs)/5[b] (PCs) |
| E-state (3) | 31 | 0.96[a]/0.96[b] | 0.79[a]/0.67[b] | 3[a](PCs)/4[b] (PCs) |
| SOM (4) | 31 | 0.85 | – | 3 (PCs) |
| QSAR/E-state (5) | 30 | 0.82 | 0.78 | 3 (atoms) |
| CoSASA | 30 | 0.80 | 0.73 | 3 (atoms) |
| CoSASA | 30 | 0.68 | 0.60 | 3 (PCs) |
| CoSA (1 ppm) | 30 | 0.80 | 0.78 | 3 (bins) |
| CoSA (1 ppm) | 30 | 0.71 | 0.65 | 3 (PCs) |

[a]1.0 Ångstrom models.
[b]2.0 Ångstrom models

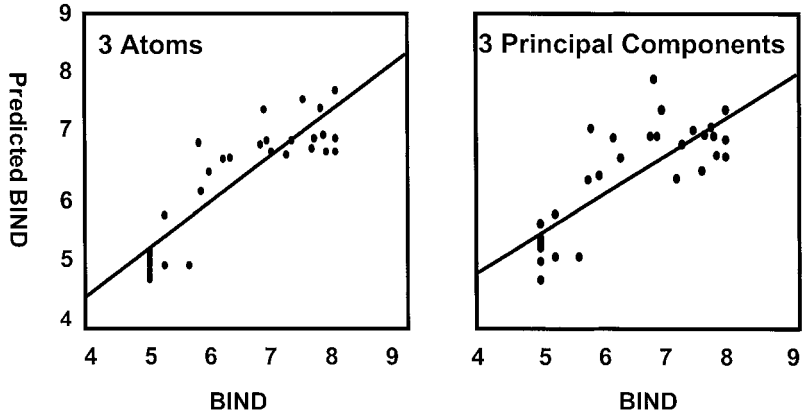


*Figure 3.* Plot of the predicted binding versus experimental binding for the 17 atom QSDAR models A. 3 atoms B. 3 Principle components.

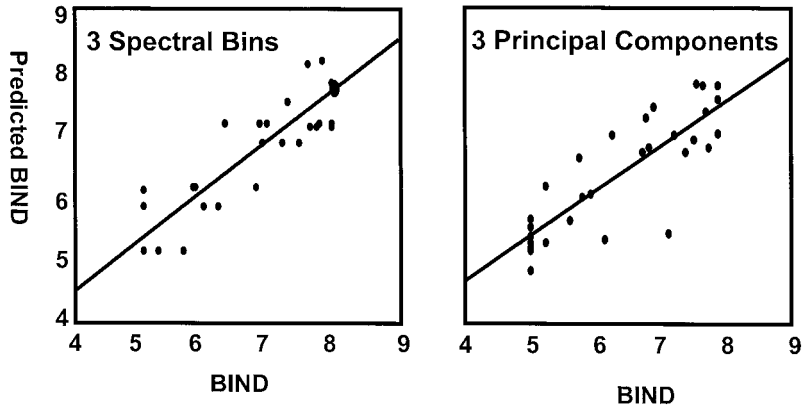


*Figure 4.* Plot of the predicted binding versus experimental binding for 1.0 ppm resolution QSDAR models A. 3 spectral bins B. 3 Principle components.

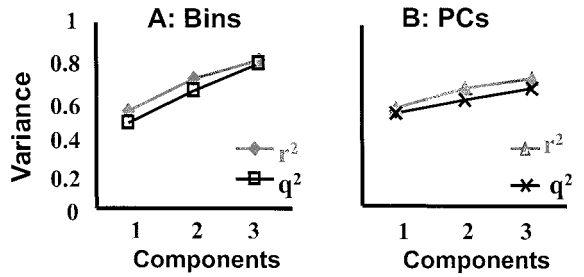


*Figure 5.* Plot of steroid backbone template QSDAR model parameters $r^2$and $q^2$ versus number of atoms or number of principal components used to produce the model. A. 3 atoms B. 3 Principle components.

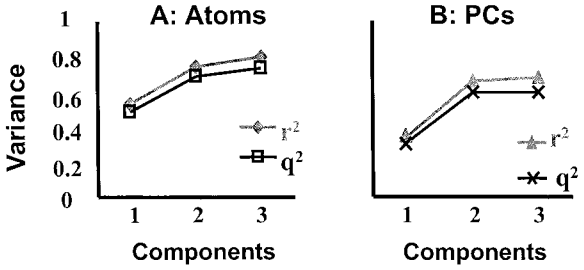


*Figure 6.* Plot of 1.0 ppm resolution QSDAR model parameters $r^2$and $q^2$ versus number of spectral bins or number of principal components used to produce the model. A. 3 spectral bins B. 3 Principle components.

the number atoms, spectral bins, or principal components used in any of the QSDAR models. Since the explained variance and cross-validated variance do not fluctuate wildly from each other in a model, this demonstrates that the simulated $^{13}$C NMR spectral data contains information relevant to a steroid inhibitor binding to the corticosteroid binding globulin. Figures 5 and 6 show that the QSDAR models based on principal components have a lower explained variance and cross-validated variance than models based on individual atoms or spectral bins. One reason for this is that the principal components are including $^{13}$C NMR data that does not have a high correlation with steroid binding to the corticosteroid binding globulin. The ability to predict binding is weakened by the inclusion of spectral 'noise' into the model. The CoSA model based on bins (Figure 6A) is still rising where all the other models have flatten or fallen off.

## Discussion

Two of the four QSDAR models have a $q^2$ greater than the 0.71 seen for the QSAR model. The HE-state and E-state models have a greater $r^2$ than all the QSDAR models but these models are very computational-intensive with many distance formulas used for every point in the grid. Still, the 1.0 ppm resolution QSDAR model based on 3 spectral bins has an explained variance ($r^2$) of 0.80 and a cross-validated variance ($q^2$) of 0.78, which are close to the $r^2$ of 0.97 and $q^2$ of 0.80 for the combination HE-state/E-state model. The 1.0 ppm resolution CoSA QSDAR model based on 3 spectral bins has a $q^2$ (0.78) better than the $q^2$ of 0.76 from the combination HE-state/E-state models with a 2 Angstrom grid density based on four principal components. The 1.0 ppm resolution QSDAR models predictability are better or comparable to the predictability for QSAR, HE-state/E-state, and E-state models. The E-state has an advantage over $^{13}$C NMR in that it uses other non-hydrogen atoms like oxygen and nitrogen besides carbon atoms. The combination HE-state/E-state uses all the atoms in the molecule to produce a QSAR model and should perhaps be compared to a $^{13}$C NMR and $^{1}$H NMR QSDAR model.

The CoSASA QSDAR model based on 3 atoms has an $r^2$ and $q^2$ that are almost equal to but slightly smaller than the $r^2$ and $q^2$ determined from the combination QSAR/E-state model based on 3 atoms. This confirms that the E-state model calculations that take into account atom type, valence electron states, and the degree of adjacency calculate energies that are similar to a $^{13}$C NMR chemical shift energy so that E-states are approximations to $^{13}$C NMR spectral data. An explanation for the fact that the explained variance of the QSDAR models was generally lower than the explained variance of the other models is that the use of large bins can combine non-correlated information and information is lost by the removal of bins with only one 'hit' in them. An explanation for the fact that the cross-validated variance of the QSDAR model was as good as the other models is that even simulated NMR spectral data are more accurate than the errors introduced by solvent effects, partial charges, dielectrics, and structural conformations used during the calculation of electrostatic potentials. All of the assumptions and approximations are prone to produce significant error. $^{13}$C NMR spectral data takes into account all structural conformations and complete solvent effects to produce a 'quantum mechanical energy' that represents the average structural environment for every carbon atom in the molecule.

The difference between the uncertainty of the predicted spectra and the true error of the predicted spectra is unknown. Figure 7 displays the standard

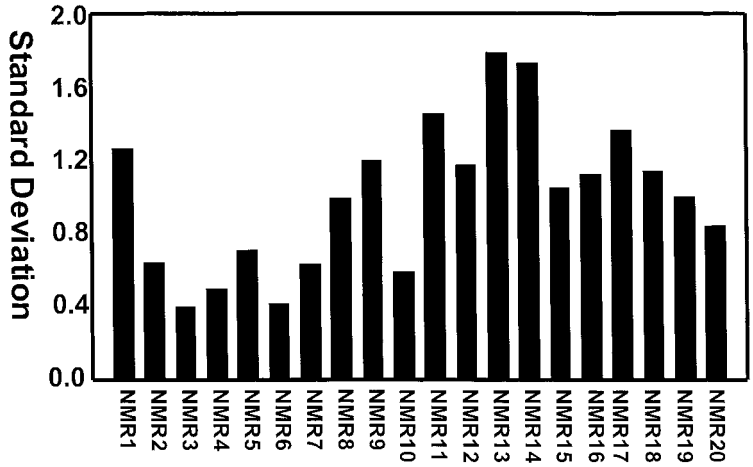


*Figure 7.* Plot of the standard deviation of uncertainty versus the carbon atom number system displayed in Figure 2.

deviation of uncertainty for the $^{13}$C NMR chemical shift prediction for the carbon atoms used in CoSASA modeling. The standard deviation of uncertainty varies from 0.39 to 1.77 ppm. The median uncertainty of carbon atoms 1 to 20 was 0.58 ppm. Carbon atoms 3, 14, and 20 were used in the CoSASA '3-atom' model and had standard deviation of uncertainty of 0.39, 1.72, and 0.83 ppm, respectively. The total variation in $^{13}$C NMR chemical shift was 143.7 ppm for atom 3, 18.4 ppm for 14, and 212.8 ppm for atom 20. The standard deviation of carbon 14 was almost 10% of its total variation, whereas the standard deviation of carbon atom 3 and 20 are less than 0.4% of their total variation. The standard deviation of carbon 14 is large enough that it could effect the variance of the CoSASA models. We see that the uncertainty of the predicted spectra does not compromise the ability of the models with high total variances. The fact the CoSASA model obtained results that correspond to the places where changes in the structure occur and these positions correspond to places that are known to effect binding is very encouraging. The CoSASA model is not strongly affected by the uncertainty of errors in the predicted NMR spectra.

Bursi used 8192 spectral bins and chemical shift peak shaping for their CoSA model of progestagens binding activity. The 1 ppm resolution CoSA QSDAR model presented here started with only 256 spectral bins, a number then reduced to 94 spectral bins when all the columns with only zeroes or with only one non-zero entry were removed. Removing the bins with one hit in the bin reduces the explained variance $r^2$ of the CoSA model but does not lower the LOO cross-validated variance $q^2$. The 1.0 ppm QSDAR model takes into account the average uncertainty in the predicted $^{13}$C NMR data. It therefore reduces the information content of the spectrum by reducing the number of spectral bins and losing the shape of the chemical shift peak. Still, the 1.0 ppm resolution CoSA QSDAR models retained enough information by increasing the number of chemical shifts in many spectral bins to produce reliable models of binding to the corticosteroid binding globulin. The NMR chemical shift peak has information about atom adjacency, solvent effects, and average structural conformation but the shape of the peak is greatly affected by shimming and temperature dependent dynamics. We see that the inclusion of average uncertainty into the simulated $^{13}$C NMR data does not affect the ability of simulated $^{13}$C NMR data to be used to model the binding affinity of structurally similar compounds to a receptor.

## Conclusions

The accuracy of the $^{13}$C NMR QSDAR model predictions shows that simulated $^{13}$C NMR spectra can be used in PLS regression analysis to model binding of structurally similar compounds to a receptor. Like, E-states calculations, simulated $^{13}$C NMR spectral data have information about the electronic structure and topological environment of an atom [3, 5]. The combination of electronic structure, topological information, and solvent effects allows simulated $^{13}$C NMR spectral data to produce a model that is as accurate and reliable as QSAR models based on separate calcula-

tions for electrostatics and steric interactions. We took precautions to reduce errors from the simulation of $^{13}$C NMR data, nevertheless, the cross-validated variance of QSDAR models based on simulated $^{13}$C NMR data should improve as the errors introduced by the simulation of the $^{13}$C NMR data are further reduced by improved spectral simulation programs.

The CoSA modeling that uses bins and does not use all the chemical shift bins is similar to QSAR modeling which removes the data from points in space where the energy calculated is always too small or large. The choice and number and size of bins necessarily avoids the extremes. Too large a bin size inappropriately lumps distinct spectral information into the same category and too small a bin size suffers from false distinctions based on reduced average bin occupancy values that adversely affect the statistics needed to identify and confirm the pattern. If one uses a huge number of bins, the results will be a model with excellent $r^2$ and pitiful $q^2$, just as Bursi reported, experimentally without an exhaustive search we have found that 1 ppm and 2 ppm bins seem to work best.

One possible reason that CoSA modeing is better than CoMFA modeling is the information in the model is being presented in a more "digital" like fashion, whereas the information in a CoMFA model is given in a more 'analog' fashion. In electronics, it has been proven that information presented in 'digital' has a higher signal to noise ratio than information presented in 'analog'. The large signal to noise is found in CoSA modeling because a chemical shift is a 'hit' inside the bin or does not 'hit' inside the bin. CoSA modeling is an attempt to digitize the modeling process. Another reason for the power of CoSA modeling is the data does use approximations in partial charges, dielectric, solvent interactions, and the average conformation in calculating electrostatic effects as done in CoMFA and other modeling techniques.

A major benefit of the QSDAR model approach is that simulated spectra can be saved and used for other compound-receptor systems by simply exchanging the relative binding affinities of the corticosteroid-receptor system for those appropriate to the alternative system to be modeled. The $^{13}$C NMR spectral data do not change and they can be used as comprehensive descriptors in new QSDAR models of many different biological endpoints. The only requirement is the availability of a suitable training set for which the strength in relation to that endpoint has already been determined. QSDAR modeling is not meant to replace QSAR, but can be used as an alternative when QSAR modeling is unreliable.

## Acknowledgements

We thank Christie McKenzie a teacher from the STRIVE (Science Teachers Research Involvement for Vital Education) program at NCTR who performed data management and some data analysis in this work.

## References

1. Mickelson, K.E., Forsthoefel, J. and Westphal, U., Biochemistry, 20 (1981) 6211.
2. Good, A.C., So, S.-S. and Richards, W.G., J. Med. Chem., 36 (1993) 433.
3. Kellogg, G.E., Kier, L.B., Gaillard, P. and Hall, L.H., J. Comput. Aid Mol. Des., 10 (1996) 513.
4. Polanski, J., J. Chem. Inf. Comput. Sci., 37 (1997) 553.
5. De Gregorio, C., Kier, L.B. and Hall, L.H., J. Comput. Aid. Mol. Des., 12 (1988) 557.
6. Beger, R., Freeman, J. Lay Jr., J., Wilkes, J. and Miller, D., Toxicol. Appl. Pharmacol., 169 (2000) 17.
7. Beger, R., Freeman, J., Lay Jr., J., Wilkes, J. and Miller, D., J. Chem. Inf. Comput. Sci., 40 (2000) 1449.
8. Bursi, R., Dao, T., van Wijk, T., de Gooyer, M., Kellenbach, E. and Verwer, P. J. Chem. Inf. Comput. Sci., 39 (1999) 861.
9. Beger, R.D. and Wilkes, J.G., J. Chem. Inf. Comput. Sci., (2001) in press.
10. Nishikawa, J. and Tori, K., J. Med. Chem., 27 (1984) 1657.
11. IJzerman, A.P., Bultsma, T. and Timmerman, H., J. Med. Chem., 29 (1986) 549.
12. Katritzky, A.R., Ignatchenko, E.S., Barcock, R.A. and Lobanov, V.S., Anal. Chem., 66 (1994) 1799.
13. Katritzky, A.R., Mu, L., Labanov, V.S. and Karelson, M., J. Phys. Chem., 100 (1996) 10400.
14. Fujita, T., Iwasa, J. and Hansch, C.A, J. Am. Chem. Soc., 86 (1964) 5175.
15. Branbury, S.P., Toxicology, 25 (1995) 67.
16. Cramer, R.D., Paterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
17. Beger, R.D. and Bolton, P.H., J. Biomol. NMR, 10 (1997) 129.
18. Wishart, D.S. and Sykes B.D., Methods Enzymol., 239 (1994) 363 .
19. Kvasnicka, V., J. Math. Chem., 6 (1991) 63.
20. Emsley, J.W., Feeney, J. and Sutcliffe, L.H., High Resolution Nuclear Magnetic Resonance, Vol. I. Pergamon Press Ltd, Oxford, 1965; Chapter 8, p. 287.
21. De Dios, A.C., Pearson, J.G. and Oldfield, E., Science, 260 (1993) 1491.
22. ACD/Labs website,http://www.acdlabs.com/.
23. Hansch, C. and Leo, A., Exploring QSAR – Fundamentals and Applications in Chemistry and Biology. The American Chemical Society, Washington, D.C., 1995.
24. Klopman, G., J. Am. Chem. Soc., 106 (1984) 7315.
25. Klopman, G., Quant. Struct. Act. Rel., 11 (1992) 176.
26. Good, A.C., Peterson, S.J. and Richards, W., J. Chem. Inf. Comput. Sci., 36 (1993) 2929.

27. Wagener, M., Sadowski, J. and Gasteiger, J., J. Am. Chem. Soc., 117 (1995) 7769.
28. Collantes, E., Tong, W. and Welsh, W., J. Anal. Chem., 68 (1996) 2038.
29. Cramer, R.D., Bunce, J.D. and Patterson, D.E. Quant. Struct.-Act. Relat., 7 (1988) 18.
30. StatSoft website, http://www.statsoft.com/.