

Full-length paper

3D-QSDAR Models of Polychlorinated Dibenzodioxins, Dibenzofurans, and Biphenyls Binding to the AhR

RICHARD D. BEGER*, DAN A. BUZATU, JON G. WILKES,

and JACKSON. O LAY JR.

*Division of Chemistry, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR
72079*

To whom correspondence should be addressed: RD Beger, Division of Chemistry, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR 72079-9502, USA, Phone: (870) 543-7080. FAX: (870) 543-7686. E-mail: rbeger@NCTR.FDA.GOV

Keywords: 3D-QSDAR; Aryl Hydrocarbon receptor (AhR); ^{13}C NMR; CoSA; CoSASA; CoSCSA; Dioxin; Furan; PCB

SUMMARY

We have developed a three-dimensional quantitative spectrometric data-activity relationship (3D-QSDAR) modeling technique which uses NMR spectral and structural information that is combined in a 3D-connectivity matrix. The 3D-connectivity matrix is built by displaying all possible assigned carbon NMR chemical shifts carbon-to-carbon connections and distances between the carbons. We selected 2D ^{13}C - ^{13}C COSY and a theoretical 2D ^{13}C - ^{13}C distance connectivity spectral slices from the 3D-connectivity matrix to produce a relationship among the 2D spectral patterns for polychlorinated dibenzofurans (PCDFs), dibenzodioxins (PCDDs), and biphenyls (PCBs) binding to the aryl hydrocarbon receptor (AhR). We refer to this technique as comparative structural connectivity spectra analysis (CoSCSA) modeling. All the QSDAR models were based on multiple linear regression analysis of the predicted ^{13}C NMR structure-connectivity spectral patterns. A 2.0 ppm resolution CoSCSA model for 26 PCDF compounds had an explained variance (r^2) of 0.97 and a leave-one-out (LOO) cross-validated variance (q^2) of 0.95. A 1.0 ppm resolution CoSCSA model for 14 PCDD compounds had an r^2 of 0.99 and a q^2 of 0.95. The 1.0 ppm resolution CoSCSA model for 12 PCB compounds had an r^2 of 0.97 and a q^2 of 0.97. A 1.0 ppm resolution CoSCSA model for all 52 compounds had an r^2 of 0.94 and a q^2 of 0.91. Conventional quantitative data-activity relationship (QSAR) modeling suffers from errors introduced by the assumptions for dielectrics, partial charges, and using only one structural conformation in calculated electrostatic potentials and the molecular alignment process. 3D-QSDAR modeling is not limited by such errors since electrostatic potential calculations and molecular alignment are not done.

Abbreviations: Three-dimensional quantitative spectrometric data-activity relationship (3D-QSDAR); Three-dimensional quantitative structure data-activity relationship (3D-QSAR); Aryl hydrocarbon receptor (AhR); Comparative spectral analysis (CoSA); Comparative structurally assigned spectral analysis (CoSASA); Comparative spectral connectivity spectral analysis (CoSCSA); heteronuclear multiple quantum correlation (HSMC); heteronuclear single quantum correlation (HSQC); Nuclear Magnetic Resonance (NMR); nuclear Overhauser effect spectroscopy (NOESY); Polychlorinated biphenyls (PCBs); Polychlorinated dibenzodioxins (PCDDs); Polychlorinated dibenzofurans (PCDFs), Principal components (PCs).

INTRODUCTION

Polychlorinated dibenzo-p-dioxins (PCDDs), dibenzofurans (PCDFs), and biphenyls (PCBs) are industrial compounds or byproducts that are widely distributed in the environment. They are known toxicants having a common receptor-mediated mechanism of action [1,2]. Many polychlorinated aromatic compounds cause toxic effects after binding to an intracellular cytosolic receptor called the aryl hydrocarbon receptor (AhR) [3,4]. Thymic atrophy, weight loss, immunotoxicity, acute lethality, and induction of cytochrome P4501A1 have all been correlated with the binding affinity of PCDDs, PCDFs, and PCBs to AhR [5,6]. This receptor controls not only the induction of the hepatic cytochrome P4501A1 but also the associated aryl hydrocarbon hydroxylase and 7-ethoxyresosufin O-deethylase activities [1]. Therefore, an important step in predicting the toxicity of PCDDs, PCDFs, and PCBs is being able to estimate each of their binding affinities to the AhR.

Quantitative structure-activity relationships (QSAR) are attempts to quantify the observed relationships between the structure of chemical compounds and the extent to which those compounds exhibit certain properties [7-14]. Quantitative spectrometric data-activity relationships (QSDAR) are based on the spectral-activity leg in the triangular structure-spectra-activity relationship. We have developed QSDAR models for binding to the corticosterone binding globulin [15], aromatase enzyme [16], and AhR [17]. One type of QSDAR, comparative spectral analysis (CoSA), produced models for corticosterone, aromatase enzyme and AhR that yielded better correlations and predictions than were seen with previous QSAR models.

These CoSA models were based on simulated ¹³C nuclear magnetic resonance (NMR) data. ACD Labs [18] now sells software that takes a chemical's structure and from that predicts its ¹³C NMR one-dimensional spectrum. Other ¹³C NMR prediction packages include artificial neural network [19] and NMRscape software from Spectrum Research Labs [20]. Therefore, the process of building the CoSA models was simple and rapid compared to corresponding QSAR models.

Our original QSDAR models were based on 1D CoSA of ¹³C NMR spectral data from a training set [15-17, 21]. Although the results of these QSDAR models were improvements compared to published QSAR results, we believed there was room for improvement. We thought that one way to improve 1D CoSA modeling was to add structural information to the model basis to produce a new technique modeling we called comparative structurally assigned spectral analysis (CoSASA) of NMR data on a 2D structural template [15,16]. Surprisingly, the CoSASA models were not as accurate as the structure-lacking CoSA models. We hypothesized that this inferior performance was caused by several factors including unanticipated nonlinear effects, a scale variation limitation that is also seen with some QSAR modeling and the information that is lost for very long side chains not represented in the template [15,16]. Another practical problem with CoSASA modeling was that each model was limited to structurally similar congeners that could be represented in relation to a single template.

Most QSAR and QSDAR attempts to produce a single, predictive model across multiple chemical classes have met with limited success. In the case of PCDDs, PCDFs, and PCBs, this challenge seems to be further aggravated by the great dependency of each molecule's AhR binding activity on its chlorination sites and on the way in which its molecular backbone conformation affects the spatial locations of the chlorine atoms. Estimation of molecular conformation for QSAR models typically uses energy-minimized structures rather than weighted average structural conformations. The latter arguably reflect more accurately the actual molecular characteristics. These factors explain why conventional QSAR models based on a mixture of PCDD, PCDF, and PCB congeners have not succeeded well [2,22]. Our previous 1D CoSA model for PCDD, PCDF, and PCB congeners was significantly robust with an r^2 of 0.85 and a q^2 of 0.71 [17], but we believed we could do better if we were able to add structural information to the model. We did not disclose the CoSASA modeling results of PCDD and PCDF binding to the AhR because they were not as good as the CoSA models [17].

The present research initiative avoids some of the foregoing problems by using the ^{13}C NMR spectral data for a test compound and adding the molecules structural connectivity information into a 3D-connectivity matrix. The 3D-connectivity matrix is built by displaying all possible carbon-to-carbon connections (through-bond and through-space) and their assigned carbon NMR chemical shifts. In this matrix representation the x-axis shows the chemical shifts of carbon i , the y-axis shows the chemical shift of carbon j , and the z-axis the distance between carbon i and carbon j (r_{ij}). Representation of a typical organic compound in this way dramatically extends the information content available as a basis for pattern recognition. In fact, the information in a 3D-connectivity matrix over-determines the compound's structure, so we can substantially reduce the information in the matrix that is used in a model without damaging the predictive power of the model. One way to reduce the information is to reduce the third (distance) dimension of the 3D matrix into a set of only four distance categories or 2D spectral planes. The first 2D plane represented the nearest neighbor through-bond connectivity plane. The three other 2D planes were constructed from compressing distance information on the z-axis, one for all short atom-to-atom through-space connections ($2.0 \text{ \AA} < r_{ij} < 3.0 \text{ \AA}$), one for all medium atom-to-atom through-space connections ($3.0 \text{ \AA} < r_{ij} < 5.0 \text{ \AA}$), and one for longer atom-to-atom through-space connections ($r_{ij} > 5.0 \text{ \AA}$). The exact distance range used in a model is not predetermined and can be adjusted to optimize a model's performance. Similarities between the pattern of 2D spectral data associated with the biological activity of the training set compounds and the corresponding spectral data for the test compound are detected to predict the extent to which the test compound should exhibit that biological activity.

Standard NMR instrumental techniques include 2D ^1H - ^1H COSY [23] experiments in which connectivity relationships through three bonds are found for nearest neighbor protons with an off diagonal cross peak. This experiment is similar to 2D ^{13}C - ^{13}C COSY experiments that contain analogous through bond connectivity spectral patterns. For the 3D-QSDAR methods developed in this work, the information contents of such COSY experiments are paralleled in the shortest distance layer of the connectivity matrix. Using the structurally assigned predicted spectra and adding the nearest neighbor information as cross peaks produces this layer.

Our 3D-QSDAR predictive methods also bears comparison to several other multi-dimensional NMR experimental techniques. ^{13}C - ^{13}C COSY experiments have similarities to 2D ^1H - ^{13}C heteronuclear single quantum correlation (HSQC) [24] and ^1H - ^{13}C heteronuclear multiple quantum correlation (HMQC) [25] NMR experiments that show the connectivity for carbons and their attached protons. In practice, 2D ^{13}C - ^{13}C COSY are seldom run because small molecules are rarely fully ^{13}C labeled. Even if the molecules were fully labeled, the ^{13}C through-bond connections usually are obtained directly from other NMR experiments like HCCH or indirectly by combining the information from ^1H - ^1H COSY with ^{13}C - ^1H HMQC and HMBC [28] NMR experiments.

In 3D-QSDAR, 2D ^{13}C - ^{13}C distance spectra that contain short, medium, and long through-space connectivity spectral patterns that are also produced by using the structurally assigned predicted spectra and selecting a distance range for nucleus to nucleus distance (r). This is analogous to ^1H - ^1H two-dimensional nuclear Overhauser effect spectroscopy (NOESY) NMR experiments where correlations through space are found for neighboring protons that are less than 5 Å away with an off-diagonal cross-peak. The size of the cross peak in the NOESY experiment is dependent on the distance between the protons, the mixing time of the experiment, and the number of different NOE spin diffusion pathways available for dipolar magnetization transfer. ^{13}C - ^{13}C NOESY experiments are for all practical purposes never executed, again because most small compounds are not fully ^{13}C labeled.

In 3D-QSDAR modeling the scientist can select which distance ranges to use in the model developing process. 2,3,7,8-tetrachlorodioxin is a strong binder in AhR and the presence of these four chlorine atoms is known to constitute an important factor in determining the compound's toxicity. The distance between positions 2 and 8 is 7.0 Å. 2,3,7,8-tetrachlorofuran is also a strong binder to AhR and the distance between its positions 2 and 8 is 6.76 Å. For these PCDDs, PCDFs, and PCBs we used a distance cutoff of 5.0 to 7.2 Å to capture this important aspect of their molecular geometry.

There are no NMR experiments that directly record structural distance information that is greater than 5 Å apart. This fact does not exclude the use of long distance in a 3D-QSDAR model, because the spectrum is simulated and does not represent or presume an energetically significant long distance interaction between the nuclei. The existence of a long range through-space in a particular i, j, r_{ij} bin connotes only that the molecule contains two atoms at this distance one has i 's and the other j 's chemical shift. Thus, the 3D conformational information used in 3D-QSDAR models is not registered with respect to an assumed structural backbone or assumed number of atoms in the molecule. It follows that, unlike QSAR methods, the 3D-QSDAR models, in either building or model use for prediction do not require any assumptions regarding the molecular alignment sequence or other docking events by which a molecule interacts with a biological receptor.

We noticed in our 1D CoSA results that many of the carbons that can not have chlorine attached to them and were attached to a oxygen in the middle of the ring system of PCDFs and PCDDs had high correlations to AhR binding [17]. This observation was not expected based on a known structure-AhR activity relationship. It was explicable as a consequence of certain NMR phenomena when applied to the simple 1D CoSA models. Because of the previous 1D CoSA observation, we constructed 3D-QSDAR models that used the $2.0 \text{ \AA} < r_{ij} < 3.0 \text{ \AA}$ distance spectra between positions 2 and 3 or positions 7 and 8 in PCDFs and PCDDs and these molecule's interior non-chlorinated atoms. For PCBs, this plane represented distances from positions 3, 4, and 5 of one phenyl ring or positions 3', 4', and 5' of the other phenyl ring.

This paper demonstrates that structural information combined with ^{13}C NMR spectra in the form of through bond and short and long range distances through space information can be used to produce a reliable, quantitative spectrometric data-activity relationship (QSDAR) model of PCDFs, PCDDs, and PCBs binding to the AhR. The work also demonstrates combination of all three compound types into a single, reliable model. We refer to this technique as comparative structural connectivity spectra analysis (CoSCSA).

PROCEDURES

Tables 1 column 3 contains previously reported [1,17] $\log EC_{50}$ binding data used for training these models. Each compound in Table 1 had their ^{13}C NMR spectra simulated using the ACD Labs CNMR predictor software, version

4.0 [18]. For QSDAR CoSCSA and CoSASA modeling we used the predicted ^{13}C NMR spectra. There were no chemical shift peaks outside of 107 to 159 ppm. The use of predicted rather than experimentally measured NMR chemical shifts is not necessary to build the QSDAR models, but it saves time, money and in this case prevents possible toxic exposures. Predicted ^{13}C NMR data points allow for the spectra to be independent of the solvent used. The CoSCSA and CoSASA modeling, LOO cross-validation, and prediction processes were completely computerized. The competitive *in vitro* binding affinities EC_{50} of PCDF, PCDD, and PCB compounds have been determined previously using [^3H]-2,3,7,8-tetrachlorodioxin as the radioligand and rodent hepatic cytosol as a source of the AhR [3,6,29-32].

Figure 1 shows the flow chart for the CoSCSA modeling procedure. The structures are used to predict 1D ^{13}C NMR spectra. We reduced the resolution of the 2D spectra to 1.0 or 2.0 ppm in both dimensions to populate many of the NMR bins for statistical analysis and to reduce the effects of uncertainties from the use of simulated spectra. These spectral widths were chosen because of convenience and because the 1.0 ppm spectral bin width was used successfully in prior QSDAR [15-17] and SDAR models based on experimental spectral data [33-35]. The 2D ^{13}C - ^{13}C NMR spectra were saved as two-dimensional bins under the peak within a certain spectral range and normalized to an integer. A single carbon to carbon connectivity was assigned an area of 100, two carbon to carbon connections in a bin had an area of 200, and so forth. This was done so that all the carbon to carbon connections would have a similar signal-to-noise ratio.

QSDAR models were produced by using the assigned ^{13}C NMR chemical shifts at the 12 positions in the PCDF and PCDD molecules, as shown in Figure 2. This requires 12 "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 name this modeling procedure comparative assigned spectra analysis (CoSASA).

The predicted NMR spectra were calculated by a substructure similarity technique called HOSE [36], which correlates similar structures with similar NMR chemical shifts. Therefore, the errors produced in the simulated NMR spectra were propagated through the similar structures found in the training set of the QSDAR models. This conveniently reduced the effective error when using the training set to predict unknown sample affinities for compound spectra predicted using the same HOSE routine.

We used the structurally assigned ^{13}C NMR spectra to produce a predicted 2D ^{13}C - ^{13}C COSY and theoretical 2D ^{13}C - ^{13}C distance spectra. The blue arrows in Figure 2A, 2D, and 2G show the through-bond neighboring carbon to carbon connections of PCDF, PCDD, and PCB molecules. These through bond carbon to carbon connections were used to simulate 2D ^{13}C - ^{13}C COSY spectra of the PCDF, PCDD, and PCB compounds. The green arrows in Figure 2B and 2E show the short range through-space carbon to carbon connections that are 2.0 to 3.0 Å apart from positions 2 and 3 or positions 7 and 8 in PCDF and PCDD molecules. Similar techniques were used to define the short distance spectra from positions 3, 4, 5 or 3', 4', and 5' in PCB molecules. Likewise the red arrows in Figure 2C, 2F, and 2I show the long range through space carbon to carbon connections that are 5.0 to 7.2 Å apart in PCDF, PCDD, or PCB molecules respectively. These carbon to carbon connections were used to produce a theoretical 2D ^{13}C - ^{13}C distance spectrum that had cross-peaks when two carbon were 5.0 to 7.2 Å apart for PCDF, PCDD, and PCB compounds. The 2D ^{13}C - ^{13}C COSY and 2D ^{13}C - ^{13}C distance spectra are symmetrical across the diagonal and for modeling purposes only half of each individual spectrum was used. The 2D ^{13}C - ^{13}C COSY and distance spectra for the compounds in the model were reduced to principal components of variation (PCs). The PCs from the 2D ^{13}C - ^{13}C COSY and distance spectra were combined. Forward multiple regression was performed on the combined set of PCs to produce a CoSCSA model.

All statistical analysis was performed by Statistica version 6.0 software [37]. The 3D-QSDAR models were produced such that the connectivity bins were evaluated with forward multiple linear regression analysis using only the most correlated PCs from both the 2D ¹³C-¹³C COSY and 2D ¹³C-¹³C distance connectivity spectra. This technique is known as principal component linear regression. We selected an optimal number of PCs in a CoSCSA model by a trail and error process, increasing the number until the r^2 either maximized or exceed 0.9 while specifying that the overall F-test and q^2 values were still increasing. For several of the models we stopped increasing the number of PCs at F-test maximum. For the other models the F-test continued to rise with the number of PCs until the number of components equaled the number of compounds in the training set. For those cases we limited the number of PCs used in the model to less than or equal to half the number of compounds in the training set.

Evaluations of the QSDAR models were done by the leave-one-out (LOO) cross-validation procedure in which each compound is systematically excluded from the training set and its inhibitor binding activity is predicted based on a model missing any contribution from that compound [38]. The cross-validated r^2 (termed q^2) that results from fitting predictions obtained by cross-validation experiments can be derived from $q^2=1-\text{PRESS}/\text{SD}$. Here 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. We believe that q^2 is a more valid measure than r^2 for assessing the reliability of a mathematical model intended for predictive applications.

RESULTS

Figure 3 plots the predicted binding versus experimental binding for four CoSCSA models of PCDFs that are based on only the ten most highly correlated PCs selected by the principal component linear regression procedure described above. Figures 3A and 3B are based on the COSY plus long range distance spectra using 1 or 2 ppm bins respectively. Figures 3C and 3D are based on the principal component regression of combined COSY, short- and long-range distance spectra using 1 or 2 ppm bins respectively. In figure 3A, the explained correlation (r^2) is 0.97 and a LOO cross-validated variance (q^2) is 0.90. The model in Figure 3B has an r^2 of 0.97 and a q^2 of 0.92. In figure 3C, the model has an r^2 of 0.95 and a q^2 of 0.94. The model in Figure 3D has an r^2 of 0.97 and a q^2 of 0.95. These are excellent results and are comparable to or better than our previous 1.0 ppm resolution CoSA model with an r^2 of 0.93 and a q^2 of 0.90 and the 2.0 ppm resolution CoSA model that had an r^2 of 0.82 and a q^2 of 0.72. It is noteworthy that 1 ppm bin widths produced superior performance for the 1D CoSA PCDF model but, with the added distance information available in these 3D-QSDAR models, it appears that the wider chemical shift range produces a more robust predictive model. The wider bin widths also produce a simpler pattern definition and reduce the number of computations.

Figure 4 plots the predicted binding versus experimental binding for four CoSCSA models of PCDDs that are based on seven to five PCs from principal component analysis. Figure 4A and 4B are the COSY plus long range distance spectra using 1 or 2 ppm bins respectively. Figure 4C and 4D are based on the linear regression of COSY, short-range distance and long-range distance PCs using 1 and 2 ppm bins respectively. In figure 4A, the r^2 is 0.99 and the q^2 is 0.95. The model in Figure 4B has an r^2 of 0.89 and a q^2 of 0.85. In figure 4C, the model has an r^2 is 0.94 and the q^2 is 0.83. The model in Figure 4D has an r^2 of 0.91 and a q^2 of 0.91. These modeling results range from to excellent. Most are better than the corresponding previously published 1.0 ppm resolution CoSA model with an r^2 of 0.87 and a q^2 of 0.52 and the 2.0 ppm resolution CoSA model that had an r^2 of 0.91 and a q^2 of 0.81.

Figure 5 plots the predicted binding versus experimental binding for four CoSCSA models of PCBs that are based on six or five PCs from regression analysis. Figure 5A and 5B are based on the combined COSY plus long range distance spectra using 1 or 2 ppm bins respectively. Figure 5C and 5D are based on the of COSY, short- and long-range distance spectra using 1 or 2 ppm bins respectively. In figure 5A, the explained correlation (r^2) is 0.98 and a LOO cross-validated variance (q^2) is 0.93. The model in Figure 5B has an r^2 of 0.96 and a q^2 of 0.79. In figure 5C, the r^2 is 0.97 and the q^2 is 0.97. The model in Figure 5D has an r^2 of 0.98 and a q^2 of 0.97. These are excellent results and are much better than either our previous 1.0 ppm resolution CoSA model with an r^2 of 0.87 and a q^2 of 0.45 or the 2.0 ppm resolution CoSA model that had an r^2 of 0.75 and a q^2 of 0.27.

Figure 6 plots the predicted binding versus experimental binding for four CoSCSA models of PCDFs, PCDDs, and PCBs that are based on 15 to 22 most correlated PCs from regression analysis. Figure 6A and 6B are the COSY plus long range distance spectra using 1 and 2 ppm bins respectively. Figure 6C and 6D are based on the COSY, short-range distance and long-range distance PCs using 1 and 2 ppm bins respectively. In figure 6A, the r^2 is 0.93 and the q^2 is 0.88. The model in Figure 6B has an r^2 of 0.83 and a q^2 of 0.65. In figure 6C, the r^2 is 0.83 and the q^2 is 0.84. The model in Figure 6D has an r^2 of 0.94 and a q^2 of 0.91. These are excellent results and are much better than our previous 1.0 ppm CoSA resolution model with an r^2 of 0.87 and q^2 of 0.67 or the 2.0 ppm resolution CoSA model that had an r^2 of 0.77 and q^2 of 0.61. The excellent results based on all three distance dimensions with 2 ppm bin widths (Figure 6D) may point to a best practice for modeling structurally divergent sets of compounds. If there are enough training set compounds (here 52) to represent the range of structural diversity (here PCDFs, PCDDs and PCBs) then wider bins including several distance ranges may give the most rugged, reliable, and valid models.

DISCUSSION

Table 2 summarizes the four CoSCSA and one CoSASA model for the PCDF compounds with respect to the n (number of PCs used), r^2 , q^2 , F and σ . Table 3 summarizes the four CoSCSA and one CoSASA model for PCDD compounds. Table 4 summarizes the four CoSCSA models for PCB compounds. Table 5 summarizes the four CoSCSA models of all 52 compounds. From these results, we conclude that the PCDF, PCDD, PCB, and the all 52 compounds CoSCSA models had contained enough information to generalize about the relevant substances' binding affinity to the AhR.

In Table 2 for PCDF compounds, all four CoSCSA models had a higher r^2 and q^2 than the corresponding 2D CoSASA model that associated spectral features with specific structural locations. In Table 3 for PCDD compounds three of four CoSCSA models had a higher r^2 and q^2 than the corresponding 2D CoSASA model. All four CoSCSA models for the 52 compounds represent a significant improvement over previously published modeling approaches (2, 22, 17, 39, and 40).

Our previous CoSA modeling paper on binding to the AhR included a complete comparison of CoSA models to other modeling techniques [2,17,22,37,38]. Almost all the CoSA models (for 26 PCDF, 14 PCDD, 12 PCB and combined 52 compounds) produced results at least equivalent to, and often far superior to other modeling methods. In this study almost all CoSCSA models showed some form of improvement over our earlier CoSA models. The CoSA models were based on selected "bins" from a 1D spectrum whereas the CoSCSA models were based on selected PCs from 2D ¹³C-¹³C COSY and 2D ¹³C-¹³C distance spectra. Thus there was more structure and spectral information available for the latter group of models. In many of the CoSCSA models the overall F score, r^2 , and q^2 were still increasing with " n " when, based on the PRINCIPAL COMPONENT REGRESSION selection rules, we stopped model building and calculated the results. We never used for the number of PCs a value more than half that of the number of compounds available for the training set. This choice avoided any possible criticism that the

method over fit the data because it contained too many degrees of freedom. However, the continued increase in overall F score and q^2 values with n, argues that over fitting had not yet occurred and that some continued improvement would still be possible and valid by using still more PCs.

From one point of view a 2D ^{13}C - ^{13}C COSY spectrum may be considered as a 2D ^{13}C - ^{13}C distance spectrum in which the distances between nuclei are less than 1.5 Å. The 2D ^{13}C - ^{13}C COSY spectrum and 2D ^{13}C - ^{13}C distance spectra may be viewed as a reduced form of a 3D matrix where the ^{13}C chemical shift comprise the x-axis and y-axis and atomic distance lies along the z-axis. For the 2D ^{13}C - ^{13}C COSY spectrum we are binning all distances less than 1.5 Å and projecting them onto a 2D plane. For the 2D ^{13}C - ^{13}C distance spectrum in these CoSCSA models we selected information from two short-range and a long-range spectra from positions 2, 3, 7, and 8 for PCDF and PCDD compounds or from positions 3, 4,5, 3', 4', and 5' in PCB compounds. We then compressed all the short-range and long-range information into separate 2D planes. For this particular modeling task we did not choose to use the structural information between 3.0 and 5.0 Å. The main reason for this curious choice is that presently we did not have NMR prediction software for ^{17}O in PCDF and PCDD compounds. The chemical shifts of the oxygen atom(s) are the major molecular structural effects between 3 to 5 Å distances from the 2 and 3 or 7 and 8 carbons. Had this information been available it might have helped the models, but even without it the results were excellent.

CONCLUSIONS

The CoSCSA modeling system can be applied to receptor binding systems for which the structure-activity relationship is unknown, a common situation faced by new drug discovery programs in the pharmaceutical industry. Producing QSAR models without detailed structural information is very unreliable and based on intuition. This work demonstrates superior performance by CoSCSA (based on multiple measures of predictability) in comparison to QSAR models, even those that include specific structural information such as CoMFA. Because CoSCSA modeling can be produced without subjective judgement and give very quick and accurate results it can be a valuable modeling system for any project that requires predictive structural models, such as new drug candidate discovery and qualification. CoSCSA modeling is ideally suited for dealing with high or medium throughput binding data. NMR chemical shifts are excellent descriptors of the nearby surrounding environment for each atom. By adding to the chemical shift information some nearest neighbor and distance-related structural information we have produced very accurate models of binding to AhR.

In this paper, we have not systematically optimized the size of the three-dimensional chemical shift/chemical shift/distance bins used in the ^{13}C - ^{13}C 2D distance spectra. In any event, without serious optimization efforts we have developed very accurate models of PCDD, PCDF, and PCB binding to AhR. Optimizing the bin dimensions and the number of distance categories spectra used in a CoSCSA model may be necessary in making predictive models for other biological endpoints.

TABLES

#	Compound	Experimental Log EC ₅₀
1	1-Cl-dibenzofuran	-5.53
2	2,8-diCl-dibenzofuran	-6.05
3	2,3,7-triCl-dibenzofuran	-8.10
4	2,3,8-triCl-dibenzofuran	-7.00
5	2,6,7-triCl-dibenzofuran	-7.35
6	1,2,3,6-tetraCl-dibenzofuran	-7.46
7	1,2,3,7-tetraCl-dibenzofuran	-7.96
8	1,2,4,8-tetraCl-dibenzofuran	-6.00
9	2,3,4,6-tetraCl-dibenzofuran	-7.46
10	2,3,6,8-tetraCl-dibenzofuran	-7.66
11	2,3,7,8-tetraCl-dibenzofuran	-8.60
12	1,2,3,7,8-pentaCl-dibenzofuran	-8.12
13	1,2,3,7,9-pentaCl-dibenzofuran	-7.40
14	1,2,4,7,9-pentaCl-dibenzofuran	-5.70
15	1,3,4,7,8-pentaCl-dibenzofuran	-7.70
16	2,3,4,7,8-pentaCl-dibenzofuran	-8.82
17	1,2,4,6,7,8-hexaCl-dibenzofuran	-6.08
18	2,3,4,6,7,8-hexaCl-dibenzofuran	-8.33
19	1,2,3,4,7,8-hexaCl-dibenzofuran	-7.64
20	1,2,3,6,7,8-hexaCl-dibenzofuran	-7.57
21	2,3,4,7,9-pentaCl-dibenzofuran	-7.70

22	2,3,4-triCl-dibenzofuran	-5.72
23	2,3-diCl-dibenzofuran	-6.33
24	2,6-diCl-dibenzofuran	-4.61
25	2-Cl-dibenzofuran	-4.55
26	4-Cl-dibenzofuran	-4.50
27	1-Cl-dibenzodioxin	-5.00
28	2,8-diCl-dibenzodioxin	-6.49
29	2,3,7-triCl-dibenzodioxin	-8.15
30	1,3,7,8-tetraCl-dibenzodioxin	-7.10
31	2,3,7,8-tetraCl-dibenzodioxin	-9.00
32	1,2,3,4,7-pentaCl-dibenzodioxin	-6.19
33	1,2,3,4,7,8-hexaCl-dibenzodioxin	-7.55
34	1,2,3,7,8-pentaCl-dibenzodioxin	-8.10
35	octaCl-dibenzodioxin	-6.00
36	1,2,3,4-tetraCl-dibenzodioxin	-6.88
37	1,2,4,7,8-pentaCl-dibenzodioxin	-6.96
38	1,2,4-triCl-dibenzodioxin	-5.88
39	2,3,6,7-tetraCl-dibenzodioxin	-7.79
40	2,3,6-triCl-dibenzodioxin	-7.66
41	2,2',4,4',5,5'-hexaCl-biphenyl	-5.10
42	2,2',4,4'-teraCl-biphenyl	-4.89
43	2,3,3',4,4',5-hexaCl-biphenyl	-6.30
44	2,3,3',4,4'-pentaCl-biphenyl	-6.15
45	2,3',4,4',5,5'-hexaCl-biphenyl	-5.80
46	2,3',4,4',5-pentaCl-biphenyl	-6.04
47	2,3,4,4',5-pentaCl-biphenyl	-6.38
48	2',3',4,4',5-pentaCl-biphenyl	-5.85

49	2,3,4,4'-tetraCl-biphenyl	-5.55
50	2,3,4,5-tetraCl-biphenyl	-4.85
51	3,3',4,4',5-pentaCl-biphenyl	-7.92
52	3,3',4,4'-tetraCl-biphenyl	-7.37

Table 1. In column two is the structures used in 3D-QSDAR models of binding to AhR and column three is the structure's experimental binding affinity.

Model	Size	N (PC)	r ²	q ²	F
1D CoSA (17)	1 ppm	5 Bins	0.93	0.90	54.7
2D CoSASA	--	6 Atoms	0.74	0.70	9.1
COSY + (5.0-7.2) Å	1 ppm	10	0.97	0.90	49.2
COSY + (5.0-7.2) Å	2 ppm	10	0.97	0.92	52.6
COSY + (2.0-3.0) Å + (5.0-7.2) Å	1 ppm	10	0.95	0.94	28.9
COSY + (2.0-3.0) Å + (5.0-7.2) Å	2 ppm	10	0.97	0.95	53.4

Table 2: 26 PCDF compound model performance parameters bin size, n (parameters used), r², q², and F.

Model	Size	N (PC)	r ²	q ²	F
1D CoSA (17)	2 ppm	5 Bins	0.91	0.81	15.9
2D CoSASA	--	5 Atoms	0.81	0.53	6.7
COSY + (5.0-7.2) Å	1 ppm	7	0.99	0.95	92.4
COSY + (5.0-7.2) Å	2 ppm	7	0.89	0.85	6.8
COSY + (2.0-3.0) Å + (5.0-7.2) Å	1 ppm	5	0.94	0.83	23.5
COSY + (2.0-3.0) Å + (5.0-7.2) Å	2 ppm	5	0.91	0.91	16.2

Table 3: 14 PCDD compound model performance parameters bin size, n (parameters used), r^2 , q^2 , and F.

Model	Size	N (PC)	r^2	q^2	F
1D CoSA (17)	2 ppm	5 Bins	0.87	0.45	8.1
COSY + (5.0-7.2) Å Distance	1 ppm	6	0.98	0.93	44.6
COSY + (5.0-7.2) Å Distance	2 ppm	5	0.96	0.79	6.9
COSY + (2.0-3.0) Å + (5.0-7.2) Å	1 ppm	5	0.97	0.97	44.3
COSY + (2.0-3.0) Å + (5.0-7.2) Å	2 ppm	5	0.98	0.97	47.0

Table 4: 12 PCB compound model performance parameters bin size, n (parameters used), r^2 , q^2 , and F.

Model	Size	N (PC)	r^2	q^2	F
1D CoSA (17)	1 ppm	15 Bins	0.87	0.67	16.6
COSY + (5.0-7.2) Å Distance	1 ppm	22	0.93	0.88	18.5
COSY + (5.0-7.2) Å Distance	2 ppm	15	0.83	0.65	11.5
COSY + (2.0-3.0) Å + (5.0-7.2) Å	1 ppm	18	0.83	0.84	11.9
COSY + (2.0-3.0) Å + (5.0-7.2) Å	2 ppm	15	0.94	0.91	28.8

Table 5: All 52 PCDF, PCDD, and PCB compound model performance parameters bin size, n (parameters used), r^2 , q^2 , and F.**REFERENCES**

- 1) Safe, S. *Crit. Rev. Toxicol.* 21 (1990) 50.
- 2) Mekemyan, O. G., Veith, G. D., Call, D. J. and Ankley, G. T., *Environ. Health Perspect.* 104 (1996) 1302.
- 3) Bhandiera, S., Sawyer, T., Romkes, M., Zmudzka, B., Safe, L., Mason, G., Keys, B. and Safe, S., *Toxicology*, 32 (1984) 131.
- 4) Mason, G., Farrell, K., Keys, B., Piskorska-Pliszczyńska, J., Safe, L. and Safe, S., *Toxicology*, 41 (1986) 21.

5) Mason, G., Sawyer, T., Keys, B., Bandiera, S., Romkes, M., Piskorska-Pliszczynska, J., Zmudzka, B. and Safe, S., *Toxicology*, 37 (1985) 1.

1. Bandiera, S., Safe, S. and Okey, A. B., *Chem. -Biol. Interact.*, 39 (1982) 259.
2. Cramer, R. D., Paterson, D. E. and Bunce, J. D., *J. Am. Chem. Soc.*, 110 (1988) 5959.
3. Tong, W., Perkins, R., Xing, L., Welsh, W. J. and Sheehan, D. M. ,*Endocrinology*, 138 (1997) 4022.
4. Hansch, C., and Leo, A. Exploring QSAR – Fundamentals and Applications in chemistry and biology. The American Chemical Society, Washington, D. C., 1995.
5. Oprea, T. I., Garcia, A. E., *J. Comput. Aid. Mol. Des.*, 10 (1996) 186.
6. Katritzky, A. R., Ignatchenko, E. S., Barcock, R. A. and Lobanov, V. S., *Anal. Chem.*, 66 (1994) 1799.
7. Katritzky, A. R., Mu, L., Labanov, V. S. and Karelson, M., *J. Phys. Chem.*, 100 (1996) 10400.
8. Fujita, T., Iwasa, J. and Hansch, C., *J. Am. Chem. Soc.*, 86 (1964) 5175.
9. Branbury, S. P. *Toxicology*, 25 (1995) 67.
10. Beger, R. D. and Wilkes, J. G., *J. Comput. Aid. Mol. Des.*, 15 (2001) 659.
11. Beger, R. D. and Wilkes, J. G., *J. Chem. Inf. Comput. Sci.*, 41 (2001) 1322.
12. Beger, R. D. and Wilkes, J. G., *J. Chem. Inf. Comput. Sci.*, 41 (2001) 1360.
13. *ACD/Labs CNMR* software version 4.0, Toronto, Canada.
14. Meiler, J., Meusinger, R. and Will, M., *J. Chem. Inf. Comp. Sci.*, 40 (2000) 1169.
15. Spectrum Research, NMRScope, Madison, WI
16. Bursi, R., Dao, T., van Wilk, T., de Gooyer, M., Kellenbach, E., and Verwer, P., *J. Chem. Inf. Comput. Sci.*, 39 (1999) 861.
17. Turner, D. B., Willett, P., Ferguson, A. M. and Heritage, T., *J. Comput. Aid. Mol. Des.*, 11 (1997) 409.
18. Aue, W. P., Bartholdi, E. and Ernst, R. R. *J. Chem. Phys.*, 24 (1976) 2229.
19. Bodenhausen, G. and Ruben, D. J., *Chem. Phys. Lett.*, 69 (1980) 185.

20. Bax, A. and Griffey, R. H., *J. Am. Chem. Soc.*, 105 (1983) 7188.
21. Randic', M. *J. Am. Chem. Soc.*, 97 (1975) 6609.
22. Burden, F. R. *Quant. Struct.-Act. Relat.*, 16 (1997) 309.
23. Bax, A. and Summers, M. F., *J. Am. Chem. Soc.*, 108 (1986) 2093.
24. Poland, A. and Knutson, J. C., *Annu. Rev. Pharmacol. Toxicol.*, 22 (1982) 517.
25. Poland, A., Glover, E. and Kende, A. S., *J. Biol. Chem.*, 251 (1976) 493.
26. Safe, S. *Crit. Rev. Toxicol.* 13, (1984) 319.
27. Safe, S. H. *Annu. Rev. Pharmacol. Toxicol.*, 26, (1986) 371.
28. Beger, R., Freeman, J., Lay Jr., J., Wilkes, J. and Miller, D., *Toxicol. Appl. Pharmacol.*, 169 (2000) 17.
29. Beger, R. D., Freeman, J. P., Lay, Jr., J. O., Wilkes, J. G. and Miller, D. W., *J. Chem. Inf. Comput. Sci.*, 41 (2001) 219.
30. Beger, R. D., Freeman, J. P., Lay, Jr., J. O., Wilkes, J. G. and Miller, D. W., *J. Chem. Inf. Comput. Sci.*, 40 (2000) 1449.
31. Bremser, W. HOSE - a Novel substructure Code. *Anal. Chim. Acta.*, 103 (1978) 355.
32. Statisica, StatSoft software, Tulsa, OK.
33. Cramer, R. D., Bunce, J. D. and Patterson, D. E., *Quant. Struct.-Act. Relat.*, 7 (1988) 18.
34. Rannug, U., Sjogren, M., Rannug, A., Gillner, M., Toftgard, R., Gustafsson, J.-A., Rosenkranz, H. and Klopman, G., *Carcinogenesis*, 12 (1991) 2007.
35. Kafafi, A. A., Afeefy, H. Y., Said, H. K. and Hakimi, J. M., *Chem. Res. Toxicol.*, 5 (1992) 856.

FIGURES

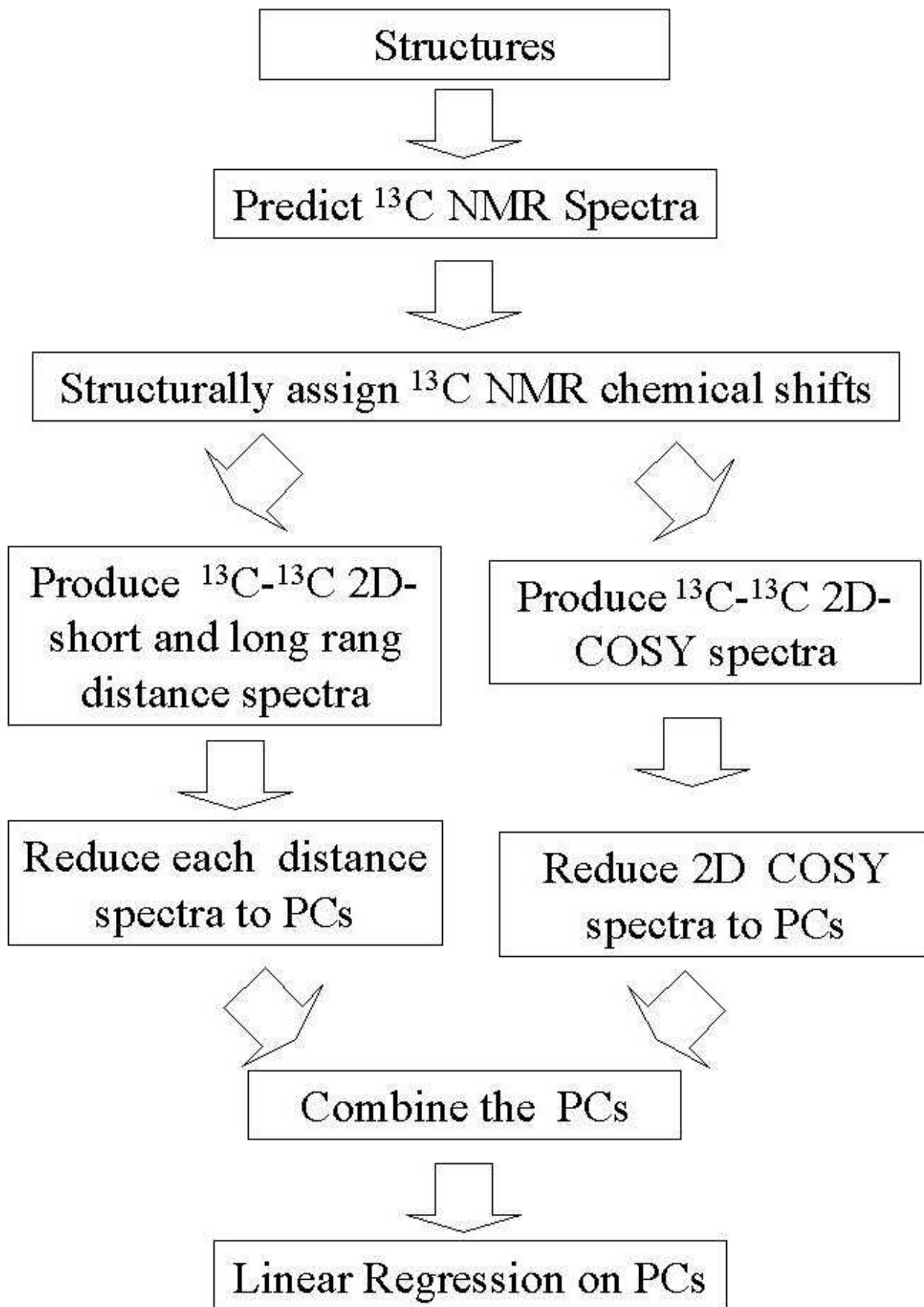


Figure1. The procedural flow chart for CoSCSA modeling.

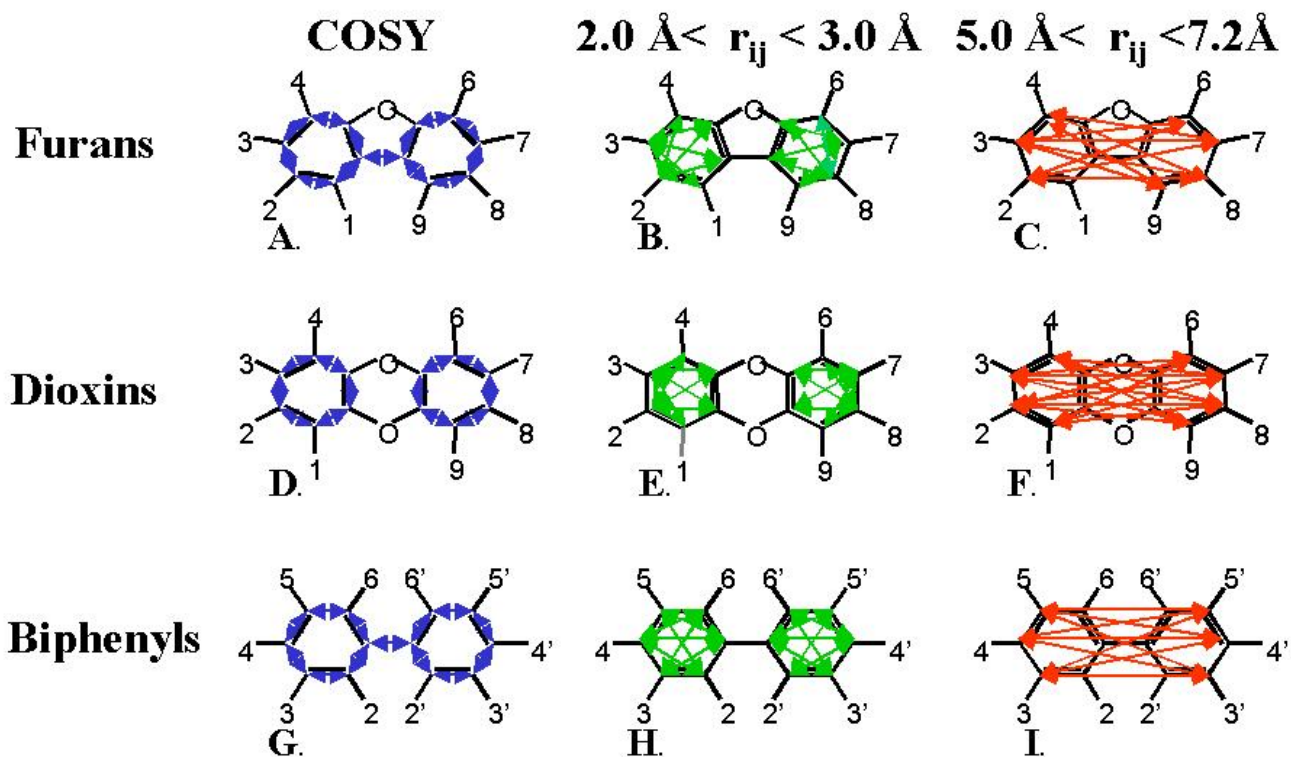


Figure 2. The arrows represent the 2D ^{13}C - ^{13}C COSY spectra for PCDFs A, PCDDs D, and PCBs G. The arrows represent short range $2.0 < r_{ij} < 3.0 \text{ \AA}$ 2D ^{13}C - ^{13}C distance spectra for PCDFs B, PCDDs E, and PCBs H. The arrows represent long range $5.0 < r_{ij} < 7.2 \text{ \AA}$ 2D ^{13}C - ^{13}C distance spectra for PCDFs C, PCDDs F, and PCBs I.

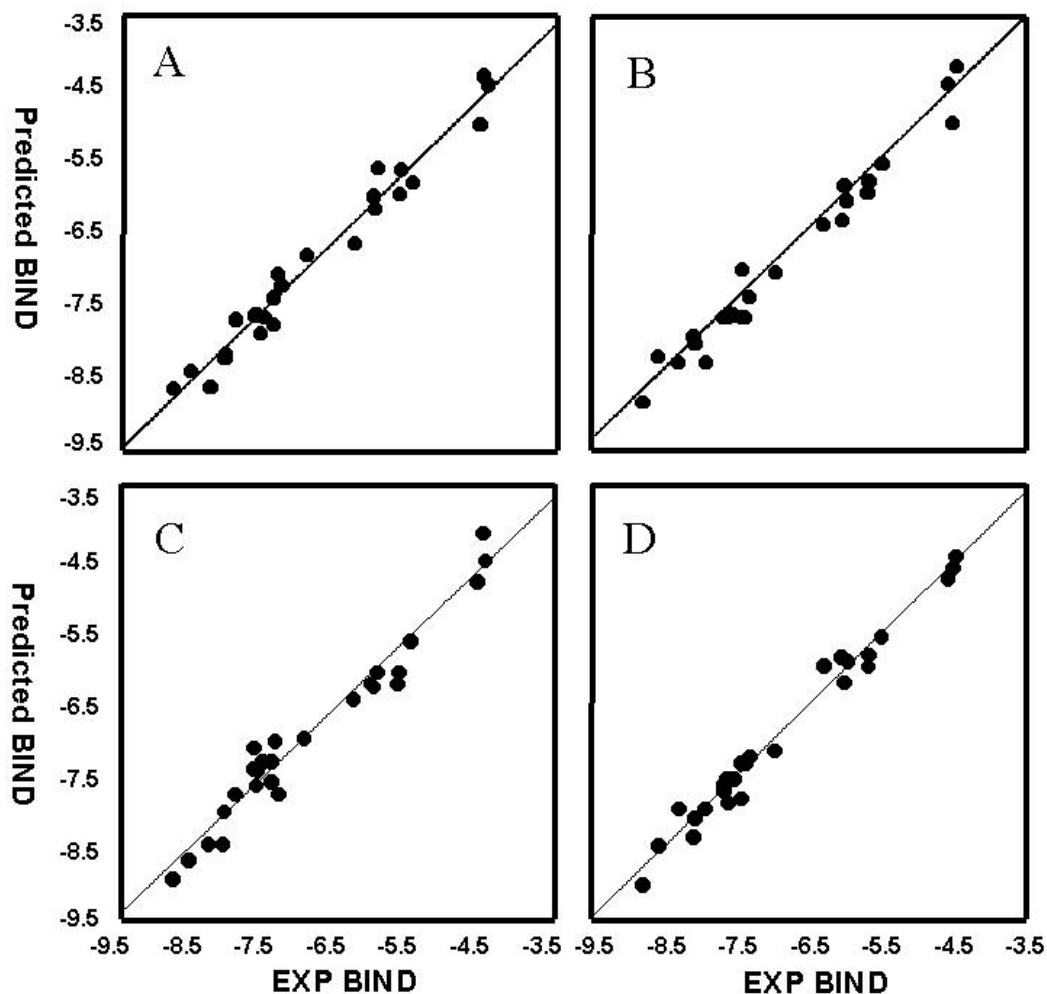


Figure 3. Plot of the predicted binding versus experimental binding for 26 PCDF compounds using 2D COSY plus long range distance spectra with 1.0 ppm (A) and 2.0 ppm bins (B). Plot of the predicted binding versus experimental binding for 26 PCDFS compounds using 2D COSY plus short and long range distance spectra with 1.0 ppm (C) and 2.0 ppm bins (D).

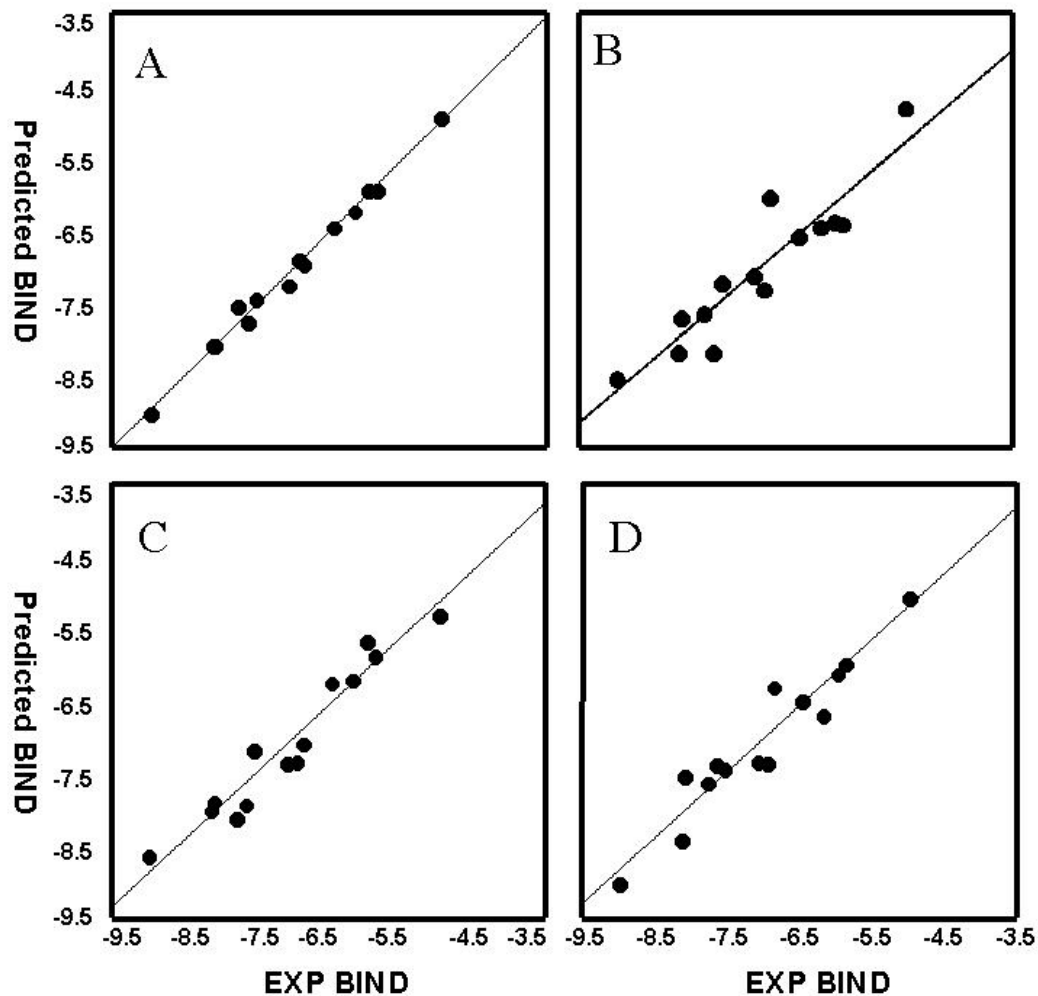


Figure 4. Plot of the predicted binding versus experimental binding for 14 PCDD compounds using 2D COSY plus long range distance spectra with 1.0 ppm (A) and 2.0 ppm bins (B). Plot of the predicted binding versus experimental binding for 26 PCDFS compounds using 2D COSY plus short and long range distance spectra with 1.0 ppm (C) and 2.0 ppm bins (D).

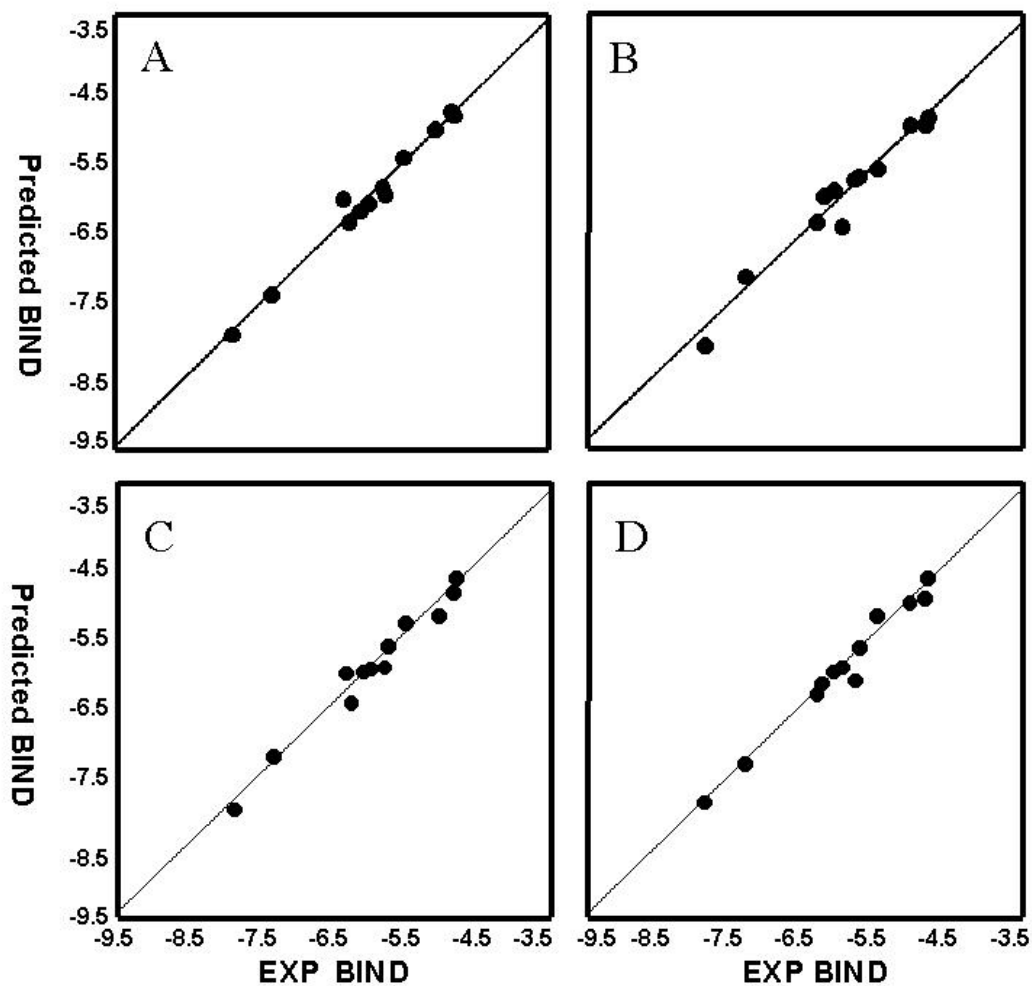


Figure 5. Plot of the predicted binding versus experimental binding for 12 PCB compounds using 2D COSY plus long range distance spectra with 1.0 ppm (A) and 2.0 ppm bins (B). Plot of the predicted binding versus experimental binding for 26 PCDFS compounds using 2D COSY plus short and long range distance spectra with 1.0 ppm (C) and 2.0 ppm bins (D).

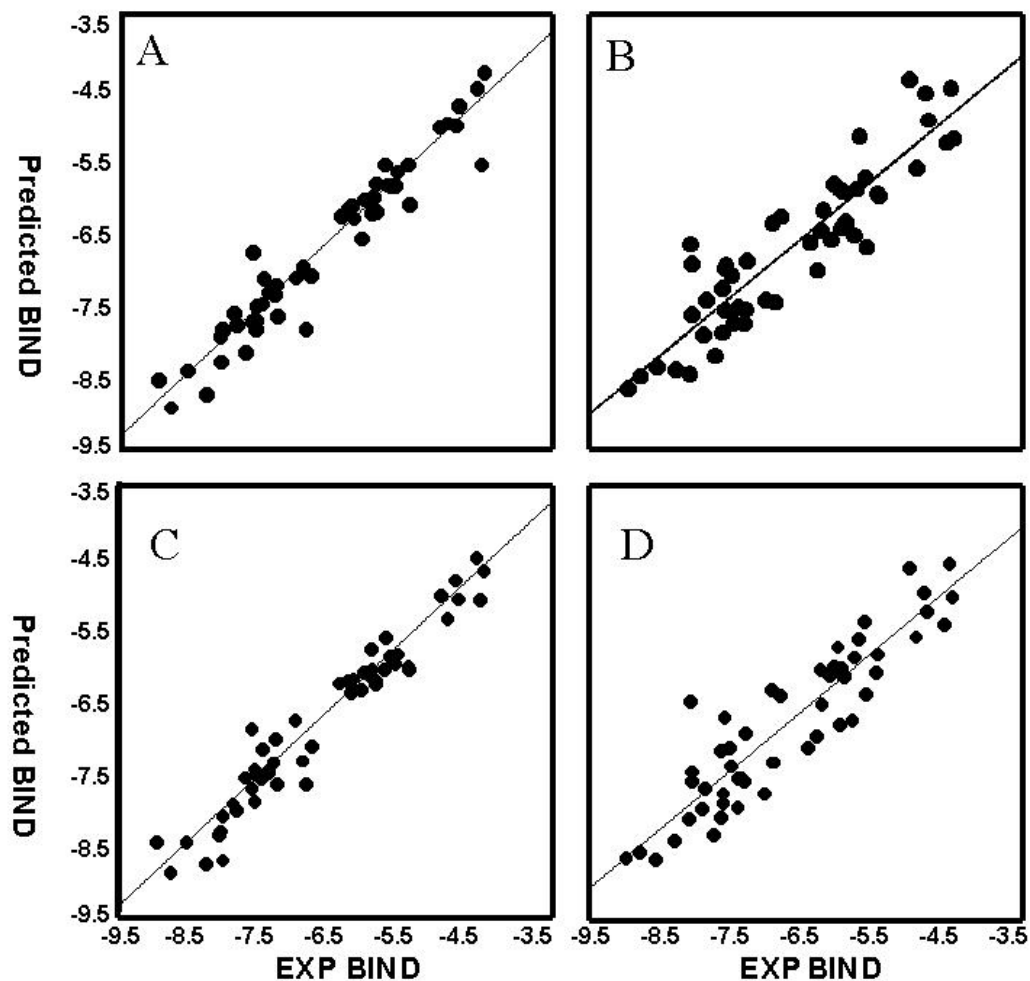


Figure 6. Plot of the predicted binding versus experimental binding for 26 PCDF, 14 PCDD, and 12PCB compounds using 2D COSY plus long range distance spectra with 1.0 ppm (A) and 2.0 ppm bins. (B). Plot of the predicted binding versus experimental binding for 26 PCDFS compounds using 2D COSY plus short and long range distance spectra with 1.0 ppm (C) and 2.0 ppm bins (D).