Models of Polychlorinated Dibenzodioxins, Dibenzofurans, and Biphenyls Binding Affinity to the Aryl Hydrocarbon Receptor Developed Using ^{13}C NMR Data

Richard D. Beger* and Jon G. Wilkes

Division of Chemistry, National Center for Toxicological Research, Food and Drug Administration,
Jefferson, Arkansas 72079-9502

Received December 31, 2000

Quantitative spectroscopic data-activity relationship (QSDAR) models for polychlorinated dibenzofurans (PCDFs), dibenzodioxins (PCDDs), and biphenyls (PCBs) binding to the aryl hydrocarbon receptor (AhR) have been developed based on simulated ^{13}C nuclear magnetic resonance (NMR) data. All the models were based on multiple linear regression of comparative spectral analysis (CoSA) between compounds. A 1.0 ppm resolution CoSA model for 26 PCDF compounds based on chemical shifts in five bins had an explained variance (r^2) of 0.93 and a leave-one-out (LOO) cross-validated variance (q^2) of 0.90. A 2.0 ppm resolution CoSA model for 14 PCDD compounds based on chemical shifts in five bins had an r^2 of 0.91 and a q^2 of 0.81. The 1.0 ppm resolution CoSA model for 12 PCB compounds based on chemical shifts in five bins had an r^2 of 0.87 and a q^2 of 0.45. The models with more compounds had a better q^2 because there are more multiple chemical shift populated bins available on which to base the linear regression. A 1.0 ppm resolution CoSA model for all 52 compounds that was based on chemical shifts in 12 bins had an r^2 of 0.85 and q^2 of 0.71. A canonical variance analysis of the 1.0 ppm CoSA model for all 52 compounds when they were separated into 27 strong binding and 25 weak binding compounds was 98% correct. Conventional quantitative structure-activity relationship (QSAR) modeling suffer from errors introduced by the assumptions and approximations involved in calculated electrostatic potentials and the molecular alignment process. QSDAR modeling is not limited by such errors since electrostatic potential calculations and molecular alignment are not done. The QSDAR models provide a rapid, simple and valid way to model the PCDF, PCDD, and PCB binding activity in relation to the aryl hydrocarbon receptor (AhR).

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] Some 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 the induction of the hepatic cytochrome P4501A1 and 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.

Multi-class prediction is a formidable challenge in QSAR modeling due to its methodological requirement that structures to be predicted are aligned to a reference compound known to interact strongly with the receptor. Most QSAR 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 an energy minimized structure rather than weighted average structural conformation. 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.[2,7]

QSAR is based on the assumption that there is a relationship between the structure and activity of a compound. QSAR modeling results have been able to show that receptor binding of a compound can be predicted from a combination of electrostatic potentials and geometrical structural analysis.[8−10] However, using a specific molecular conformation derived from computer modeling of each compound dramatically extends the number of calculations required to define the model. Moreover, the selection of the most appropriate 3D conformation 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 contrast, a significant advantage of QSDAR is that it is not necessary either to perform laborious quantum mechanical calculations or to use the 3D conformations of the molecules for electrostatic calculations, both of which are done in many QSAR techniques.[11−15]

* Corresponding author phone: (870)543-7080; fax: (870)543-7686; e-mail: rbeger@nctr.fda.gov.

10.1021/ci000312l Published 2001 by the American Chemical Society
Published on Web 05/12/2001

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 3D geometry.[16] Ab initio quantum mechanical calculations of ^{13}C chemical shift tensors in proteins reveal that they are dependent on the structural environment.[17] This implies that the chemical shifts respond to and represent the effects of the molecule's structural environment. Given a large enough training set from which to generalize, these chemical shift patterns can be associated with the molecular characteristics corresponding to biological interactions with proteins. They can be associated with biological activity. 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. One way to describe the benefit of using NMR data in this way is that NMR uses the laws of nature to "calculate" the average quantum mechanical magnetic moment energy of every carbon nucleus in the molecule. It is analogous to defining a structure activity relationship without requiring use of any conformational assumptions or approximations. The effects of substituents on ^{13}C NMR chemical shifts can be felt from as far as five bonds away or directly through space. Relative, not absolute energies in NMR spectra are used in QSDAR. This fact is reflected in the typical units for NMR chemical shifts, parts per million (ppm), dimensionless numbers defined with respect to a reference compound's NMR signals.

^{13}C nuclear magnetic resonance (NMR) chemical shifts have been used to predict and refine chemical structures.[18,19] This works because there is a well-defined relationship between the structures and 3D conformations of a molecule and its ^{13}C NMR shifts. ACD Labs[20] now sells software that performs the converse association: it takes a chemical's structure and from that predicts its^{13}C NMR one-dimensional spectrum. Other ^{13}C NMR prediction packages include artificial neural networks[21] and NMRscape software from Spectrum Research Labs.[22]

Using QSAR modeling methods, receptor binding of a compound can be predicted, based in part upon electrostatics and geometrical structure. 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).[11−15] By combining the ^{13}C NMR data into a composite set of descriptors for CoSA and putting them into statistical software programs for comparative spectral analysis, it would be possible to produce a QSDAR model of the inhibitor compound binding to the enzyme.

We are now developing conceptually novel modeling methods that use patterns in spectral data of molecules to predict biological characteristics such as receptor binding affinities. These methods avoid many of the problems associated with conventional QSAR. They can be implemented as general classifiers, categorizing congeners as strong, medium, or weak binders. For example, experimental ^{13}C NMR and electron ionization mass spectrometric (EI MS) data have been used to produce a reliable classification for spectrometric data-activity relationship (SDAR) models of the estrogen receptor system.[23,24]

An extension of the SDAR concept uses somewhat different pattern recognition techniques to produce a quantitative prediction of binding affinity, directly analogous to QSAR results. We use the acronym QSDAR for such spectral data models that give a quantitative rather than only a classification prediction. For example, we have produced comparative spectral analysis (CoSA) models giving quantitative affinity predictions for 30 steroidal inhibitors binding to corticosteroid binding globulin.[25] These models were based on simulated ^{13}C nuclear magnetic resonance (NMR) data. One of these CoSA models yielded better correlations and predictions than were seen with comparative molecular field analysis (CoMFA). Similarly, the binding activity of 45 progestagen steroids to a steroid receptor have been quantitatively modeled with simulated ^{13}C NMR spectra by CoSA.[26] This CoSA model also yielded better explained correlations than were seen with CoMFA methods.

Because NMR chemical shifts at an atom respond to substituents up to five bonds removed, in QSDAR each chemical shift functions as a quantum mechanical descriptor for a 4−8 atom structural moiety.[23−25] These quantum mechanically based descriptors are used in a manner similar to those in current QSAR models that break the molecule into secondary structural motifs. This paper demonstrates that simulated ^{13}C NMR spectral data can be used to produce a reliable, quantitative spectrometric data-activity relationship (QSDAR) model of PCDFs, PCDDs, and PCBs binding to the AhR.

PROCEDURES

The compounds in Table 1 had their ^{13}C NMR spectra simulated using the ACD Labs CNMR predictor software, version 4.0.[20] Simulated rather than actual ^{13}C NMR spectra were used throughout this analysis for self-consistency: that is, similar errors in the predicted chemical shifts will be produced by similar substructures. The use of predicted chemical shifts is not necessary to build the QSDAR models, but it saves time and money and in this case prevents possible toxic exposures. The QSDAR modeling, 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[3,6,27−30]

For CoSA modeling we used the unassigned simulated ^{13}C NMR data points. Unassigned one-dimensional (1D) ^{13}C NMR chemical shifts were segregated into bins over a 106−160 ppm range. There were no chemical shift peaks outside the 106−160 ppm range. Because of uncertainties in the simulated ^{13}C NMR spectra, we reduced the resolution of the chemical shift peak to 1.0 and 2.0 ppm. Another reason to reduce the resolution of the NMR spectra is a need to populate many of the NMR bins for statistical analysis. The pattern recognition software used was Statistica version 5.5.[31] The spectral widths were chosen because of convenience and because the 1.0 ppm spectral bin width was used successfully in prior QSDAR[24] and SDAR models based on experimental spectral data.[23,24] The spectral width of the bins was used to

Table 1. Structures of 26 PCDFs, 14 PCDDs, and 12 PCBs Used in CoSA Models of Binding to the AhR

no.	compound	exptl log EC_{50}	CoSA(12) LOO predicted log EC_{50}	
			1.0	2.0
1	1-Cl-dibenzofuran	−5.53	−6.69	−6.84
2	2,8-diCl-dibenzofuran	−6.05	−5.77	−5.40
3	2,3,7-triCl-dibenzofuran	−8.10	−8.08	−8.49
4	2,3,8-triCl-dibenzofuran	−7.00	−7.20	−7.26
5	2,6,7-triCl-dibenzofuran	−7.35	−7.04	−7.19
6	1,2,3,6-tetraCl-dibenzofuran	−7.46	−8.33	−7.88
7	1,2,3,7-tetraCl-dibenzofuran	−7.96	−7.62	−6.91
8	1,2,4,8-tetraCl-dibenzofuran	−6.00	−6.20	−6.91
9	2,3,4,6-tetraCl-dibenzofuran	−7.46	−7.47	−8.12
10	2,3,6,8-tetraCldibenzofuran	−7.66	−8.63	−8.57
11	2,3,7,8-tetraCl-dibenzofuran	−8.60	−8.58	−8.34
12	1,2,3,7,8-pentaCl-dibenzofuran	−8.12	−8.22	−8.22
13	1,2,3,7,9-pentaCl-dibenzofuran	−7.40	−7.75	−7.28
14	1,2,4,7,9-pentaCl-dibenzofuran	−5.70	−6.66	−6.82
15	1,3,4,7,8-pentaCl-dibenzofuran	−7.70	−7.80	−7.04
16	2,3,4,7,8-pentaCl-dibenzofuran	−8.82	−7.23	−7.71
17	1,2,4,6,7,8-hexaCl-dibenzofuran	−6.08	−6.52	−6.59
18	2,3,4,6,7,8-hexaCl-dibenzofuran	−8.33	−7.65	−7.19
19	1,2,3,4,7,8-hexaCl-dibenzofuran	−7.64	−7.32	−6.89
20	1,2,3,6,7,8-hexaCl-dibenzofuran	−7.57	−7.02	−7.50
21	2,3,4,7,9-pentaCl-dibenzofuran	−7.70	−7.23	−6.07
22	2,3,4-triCl-dibenzofuran	−5.72	−6.39	−5.28
23	2,3-diCl-dibenzofuran	−6.33	−5.74	−6.10
24	2,6-diCl-dibenzofuran	−4.61	−4.23	−4.76
25	2-Cl-dibenzofuran	−4.55	−4.93	−4.40
26	4-Cl-dibenzofuran	−4.50	−5.43	−5.52
27	1-Cl-dibenzodioxin	−5.00	−3.32	−5.55
28	2,8-diCl-dibenzodioxin	−6.49	−6.61	−7.16
29	2,3,7-triCl-dibenzodioxin	−8.15	−7.77	−7.85
30	1,3,7,8-tetraCl-dibenzodioxin	−7.10	−7.50	−7.96
31	2,3,7,8-tetraCl-dibenzodioxin	−9.00	−9.29	−8.54
32	1,2,3,4,7-pentaCl-dibenzodioxin	−6.19	−7.26	−7.01
33	1,2,3,4,7,8-hexaCl-dibenzodioxin	−7.55	−7.30	−7.50
34	1,2,3,7,8-pentaCl-dibenzodioxin	−8.10	−7.32	−7.46
35	octaCl-dibenzodioxin	−6.00	−5.77	−5.53
36	1,2,3,4-tetraCldibenzodioxin	−6.88	−6.56	−7.12
37	1,2,4,7,8-pentaCl-dibenzodioxin	−6.96	−6.55	−6.71
38	1,2,4-triCl-dibenzodioxin	−5.88	−6.22	−6.05
39	2,3,6,7-tetraCl-dibenzodioxin	−7.79	−8.31	−6.65
40	2,3,6-triCl-dibenzodioxin	−7.66	−7.89	−7.41
41	2,2′,4,4′,5,5′-hexaCl-biphenyl	−5.10	−3.99	−6.30
42	2,2′,4,4′-teraCl-biphenyl	−4.89	−6.00	−6.00
43	2,3,3′,4,4′,5-hexaCl-biphenyl	−6.30	−5.75	−6.66
44	2,3,3′,4,4′-pentaCl-biphenyl	−6.15	−6.34	−5.84
45	2,3′,4,4′,5,5′-hexaCl-biphenyl	−5.80	−6.39	−6.23
46	2,3′,4,4′,5-pentaCl-biphenyl	−6.04	−5.77	−4.91
47	2,3,4,4′,5-pentaCl-biphenyl	−6.38	−6.18	−6.66
48	2′,3′4,4′,5-pentaCl-biphenyl	−5.85	−6.38	−5.85
49	2,3,4,4′-tetraCl-biphenyl	−5.55	−5.82	−6.26
50	2,3,4,5-tetraCl-biphenyl	−4.85	−5.88	−5.41
51	3,3′,4,4′,5-pentaCl-biphenyl	−7.92	−6.90	−6.23
52	3,3′,4,4′-tetraCl-biphenyl	−7.37	−6.38	−6.33

reduce sensitivity to any errors introduced through use of simulated NMR data. 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 eliminate line width variations due to shimming, temperature, and predicted line shapes. Thus, in the 1.0 ppm resolution CoSA model, there were 55 bins, each of which was populated, or not, depending on the pattern of simulated chemical shifts. The number of bins analyzed by the CoSA models were reduced by using only the number of remaining bins that had more than one chemical shift "hit" in the bin. This reduced the number of available bins in each 1.0 ppm resolution CoSA model to 26 bins for the 26 PCDF compounds, 18 bins for the 14 PCDD compounds, 15 bins for the 12 PCB compounds, and 40 bins for the combined model of all 52 compounds. This reduced the number of available bins in each 2.0 ppm resolution CoSA model to 17 bins for the 26 PCDF compounds, 13 bins for the 14 PCDD compounds, 9 bins for the 12 PCB compounds, and 22 bins for the combined model of all 52 compounds.

The predicted NMR spectra were calculated by a substructure similarity technique called HOSE,[32] which correlates similar structures with similar NMR chemical shifts. CoSA QSDAR models were produced in which the spectral width bins were evaluated with partial least squares (PLS) multiple regression analysis using only the most correlated individual spectral bins. Therefore, the errors produced in the simulated NMR spectra were propagated through to 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.

Before regression analysis was performed on the simulated ^{13}C NMR data, the spectral bin columns with all zeros and spectral bin columns with only one nonzero number were removed from the data set input to the Statistica software program. This accelerated the pattern recognition calculations without affecting the quality of the result.

Table 1 contains the LOO predicted log EC_{50} values for 1.0 and 2.0 ppm CoSA models of 26 PCDF, 14 PCDD, and 12 PCB compounds combined using 12 NMR spectral bins. Previously reported[1] log EC_{50} binding data used for training these models is shown is column 2 of Table 1.

Use of "unassigned" chemical shifts means that this CoSA approach does not require the identification of each shift with the carbon atom that produced it. The CoSA QSDAR models were based on the individual unassigned spectral bins. This technical detail is particularly important in the case of models involving disparate backbone structures for which it makes no sense to associate a particular shift with a particular atom. For PCBs used in a set that also contains PCDDs or PCDFs, it is not possible consistently and meaningfully to number carbon atoms and associated chemical shifts even though every molecule in the set contains exactly 12 carbon atoms in two six member aryl hydrocarbon rings. The problem would be worse if the set included compounds with more or less than 12 carbon atoms. The possibility of basing patterns on unassigned chemical shifts greatly expands the utility of the QSDAR CoSA methods for predicting the binding characteristics of structurally dissimilar compounds.

In addition to the QSDAR models, a nonquantitative canonical analysis was performed on all 52 compounds from Table 1. Using one canonical variate we separated the 52 compounds into 27 strong binding compounds if the log of EC_{50} was lower than −7.0 and 25 weak binding compounds if the log of the EC_{50} was greater than −7.0. Canonical analysis was performed using the same 15 NMR chemical shift bins used in the 52 compound CoSA model.

Evaluations of the QSDAR models were done by the LOO cross-validation procedure in which each compound is systematically excluded from the training set, and its inhibitor

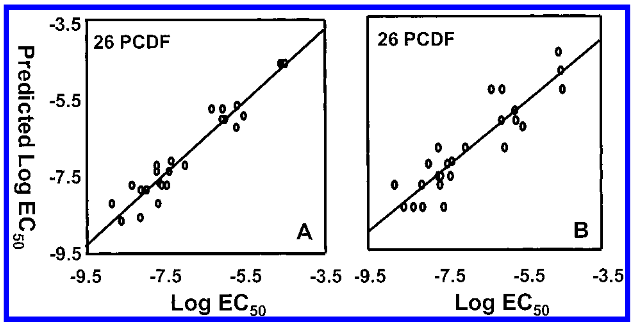

Figure 1. Plot of the predicted binding versus experimental binding for 1.0 ppm (1.A) and 2.0 ppm (1.B) CoSA models of 26 PCDF compounds using five spectral bins.

binding activity is predicted by a model missing any contribution from that compound.[33] The cross-validated r^2 (termed q^2) can be derived from $q^2 = 1 - \text{PRESS/SD}$. 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 1 plots the predicted binding versus experimental binding for the 26 PCDF compound 1.0 ppm (1.A) and 2.0 ppm (1.B) CoSA model that used only the five most correlated spectral bins. The explained correlation (r^2) of this 1.0 ppm resolution model is 0.93, and LOO cross-validated variance (q^2) is 0.90. These five spectral bins corresponded in order of correlation strength to the chemical shift frequencies in the 156, 130, 124, 157, and 117 ppm. Although, the chemical shifts were unassigned in building the model, we noticed that those in the 156 and 157 ppm bins monitored carbon atoms bound to the oxygen in the dibenzofuran. The chemical shifts in the 130 ppm bin tended to monitor the carbon 3 and 8 positions when a chlorine atom was attached to it, whereas chemical shifts in the 117 ppm bin tended to monitor any carbon positions for a lack of an attached chlorine. The explained correlation (r^2) of this 2.0 ppm resolution model is 0.82 and LOO cross-validated variance (q^2) is 0.72. These five spectral bins corresponded in order of correlation strength to the chemical shift frequencies in the 156, 130, 116, 154, and 132 ppm. We also noticed that the chemical shifts in the 156, 130, and 116 bins were involved in the 1.0 ppm CoSA model of PCDFs.

These results exemplify an interesting characteristic of QSDAR modeling. It is known that chlorine substitutions at positions 2, 3, 7, and 8 are most strongly associated with AhR binding for PCDDs and PCDFs.[1] But the most strongly correlated single signal is obtained in this QSDAR model for the two carbons at which chlorine substitution is not possible. A qualitative explanation for this observation is that signals obtained from the carbon attached to the furan oxygen near the center of the molecule contain information (from up to a five bond distance) that represent a greater proportion of the molecule's total chlorination characteristics. Also, there is a nonlinearity consideration. AhR binding decreases for fewer as well as for more than four chlorines. Therefore, the magnitude of signals obtained at the number 3 or number 8 carbon positions (spectrally degenerate) will not be linearly related to the binding phenomenon. In this case, a better single representation of binding affinity is obtained from the population of carbons adjacent to the furan oxygen positions. We have observed a similar phenomenon with the other endpoints studied earlier.[25]

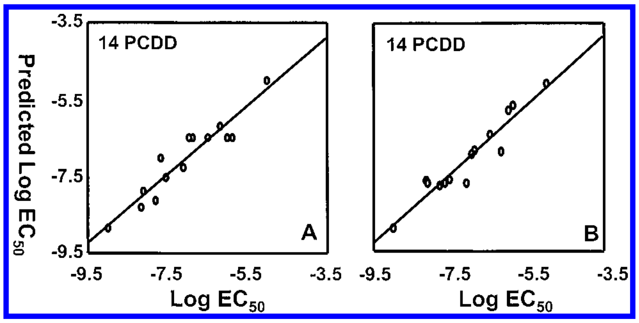

Figure 2. Plot of the predicted binding versus experimental binding for 1.0 ppm (2.A) and 2.0 ppm (2.B) CoSA models 14 PCDD compounds using five spectral bins.

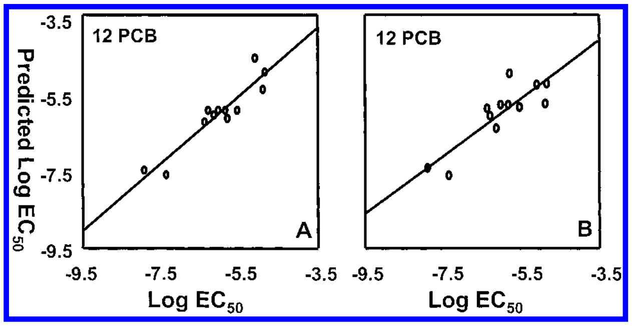

Figure 3. Plot of the predicted binding versus experimental binding for 1.0 ppm (3.A) and 2.0 ppm (3.B) CoSA models 12 PCB compounds using five spectral bins.

Figure 2 shows the predicted versus experimental binding for the 14 PCDD compound 1.0 ppm (2.A) and 2.0 ppm (2.B) CoSA model when only the five most correlated spectral bins were included. The explained correlation (r^2) of this model is 0.87 and LOO cross-validated variance (q^2) is 0.52. The five spectral bins corresponded in order of correlation strength to the chemical shifts of 128, 119, 142, 129, and 121 ppm. Analogous to the PCDF case discussed above, we noticed that chemical shifts in the 142 ppm bins monitored carbons bound to one of the oxygens in the dioxin. Chemical shifts in the 128 and 129 ppm bin tended to monitor carbon atoms that had a chorine atom attached to them. The explained correlation (r^2) of the 2.0 ppm resolution CoSA model is 0.91, and the LOO cross-validated variance (q^2) is 0.81. The five spectral bins corresponded in order of correlation strength to the chemical shifts of 128, 116, 130, 142, and 124 ppm. Again analogous to the PCDF case discussed above, we noticed that chemical shifts in bins 128, 129, and 142 are included in the 2.0 ppm resolution CoSA model of PCDDs, while 119 and 121 are not.

Figure 3 shows the predicted versus experimental binding for a 12 PCB compound 1.0 ppm (3.A) and 2.0 ppm (3.B) CoSA model when using only the five most correlated spectral bins. The explained correlation (r^2) of the 1.0 ppm resolution CoSA model is 0.87, and the LOO cross-validated variance (q^2) is 0.45. These five spectral bins corresponded to the chemical shift frequencies in the 125, 131, 136, 135,

Table 2. 1.0 ppm Resolution CoSA Model Performance Parameters *n* (Bins Used), r^2, q^2, and the Order of Specific NMR Bins Used

compounds used in CoSA model	*n*	r^2	q^2	NMR bins used
26 PCDF	5	0.93	0.90	156, 130, 124, 157, 117
26 PCDF	3	0.75	0.65	156, 130, 124
14 PCDD	5	0.87	0.52	127, 129, 142, 128, 133
14 PCDD	3	0.81	0.56	128, 133, 142
12 PCB	5	0.87	0.45	125, 131, 136, 135,128
12 PCB	3	0.72	0.39	125, 131, 136
26 PCDF + 14 PCDD + 12 PCB	15	0.87	0.67	156, 143, 142, 119, 153, 117, 157, 126, 135, 137, 124, 108, 111, 113, 147
26 PCDF +14 PCDD + 12 PCB	12	0.85	0.71	156, 143, 142, 119, 153, 117, 157, 126, 135, 137, 124, 108,

Table 3. 2.0 ppm Resolution CoSA Model Performance Parameters *n* (Bins Used), r^2, q^2, and the Order of Specific NMR Bins Used

compounds used in CoSA model	n	r^2	q^2	NMR bins used
26 PCDF	5	0.82	0.72	156, 130, 116, 154, 132
26 PCDF	3	0.75	0.67	156, 130, 116
14 PCDD	5	0.91	0.81	128, 130, 116, 142, 124
14 PCDD	3	0.83	0.74	128, 130, 116
12 PCB	5	0.75	0.27	124, 132, 134, 138,140
12 PCB	3	0.66	0.30	124, 132, 140
26 PCDF + 14 PCDD + 12 PCB	15	0.77	0.51	156, 142, 116, 152, 118, 144, 128, 136, 130, 154, 126, 114, 124, 134, 146
26 PCDF + 14 PCDD + 12 PCB	12	0.75	0.61	156, 142, 116, 152, 118, 144, 128, 136, 130, 154, 126, 114

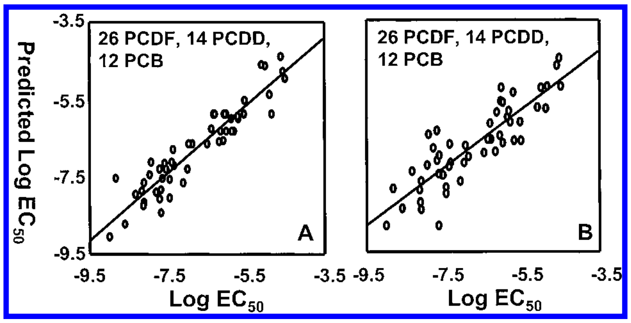

Figure 4. Plot of the predicted binding versus experimental binding for 1.0 ppm (4.A) and 2.0 ppm (4.B) CoSA models 52 compounds using 15 spectral bins.

128 ppm. We determined after the CoSA analysis that the chemical shifts in the 136 and 135 bins monitored carbon 1 and 1′ in the phenyl rings, and chemical shifts in the 131 bin tended to monitorcarbons 3 and 3′ for a chorine atom attached to it. The explained correlation (r^2) of the 2.0 ppm resolution CoSA model is 0.75 and LOO cross-validated variance (q^2) is 0.25. These five spectral bins corresponded to the chemical shift frequencies in the 124, 132, 138, 136, and 126 ppm. Analogous to the PCDF and PCDD case discussed above, we noticed that chemical shifts in bins 125 and 136 are in included in the 2.0 ppm resolution CoSA model of PCBs, while 131, 135, and 128 are not. The large drop off between explained and cross-validated correlations happens because of the reduced number of patterns in the training set and the consequent vulnerability of the model to significant changes as each spectrum is removed in the LOO cross-validation process. This factor is a general characteristic of SDAR modeling. The quality of the model increases dramatically as a function of the number of spectra available for training.

Figure 4 represents predicted versus experimental binding for a combined 26 PCDF, 14 PCDD, and 12 PCB compound 1.0 ppm (4.A) and 2.0 ppm (4.B) CoSA model based on the 12 most correlated spectral bins. The explained correlation (r^2) of the 1.0 ppm CoSA model is 0.85 and LOO cross-validated variance (q^2) is 0.71. Within this model the explained correlation (r^2) of the PCDFs in this model is 0.92, the LOO cross-validated variance (q^2) is 0.6, r^2 of the PCDDs in this model is 0.89 and q^2 is 0.75, and r^2 of the PCBs in this model is 0.75 and q^2 is 0.45. We see that PCDFs and PCDDs seem to be learning broad patterns inherent in both classes in this model. The PCBs do not seem to be learning broad patterns inherent in other classes in this model, a fact that may be due to their ability to freely rotate their phenyl rings or the variability and accuracy of the training set. Of course, the PCDFs and PCDDs do not share this PCB capability. The r^2 of the 2.0 ppm CoSA model for PCBs is 0.75 and q^2 is 0.27.

The PCDFs, PCDDs, and the 52 compound combination CoSA models had a standard error (SE) between 0.1 and 0.6 and $p < 0.001$. The PCB models had a standard error (SE) between 0.5 and 0.6 and $p < 0.02$. Table 3 is a summary of CoSA model performance parameters: *n* (number of bins used), r^2, q^2, and the specific NMR bins used in the model. The results for PCDFs, PCDDs, and the 52 compound combination are excellent by any measure but particularly in comparison to published SAR results based on quantum mechanics for these compounds[2] and QSAR results based on infrared descriptors and CoMFA techniques.[7]

Figure 5 shows a discriminant function score plot for the canonical analysis of the nonquantitative SDAR based on 15 NMR spectral bins. Compounds with a gray background are strong binders and compounds with a white background, weak binders to the AhR receptor. The plot shows a separation between the 27 strong and the 25 weak binders, except for 1,2,4,6,7,8-hexachlorodibenzofuran, which was marginally predicted as a strong binder. The discriminant function score seen in Figure 5 was obtained by multiplying the 15 spectral factor canonical weights by the spectra in the 15 bins for the compound. Figure 6 shows the factor weights for the canonical variate analysis split into bins that bias for strong (bars pointing up) and weak binding (bars pointing down).

DISCUSSION AND CONCLUSIONS

Table 2 is a summary of these four 1.0 ppm CoSA models' characteristics and performance with respect to four parameters: *n* (number of bins used), r^2, q^2, and a list of specific

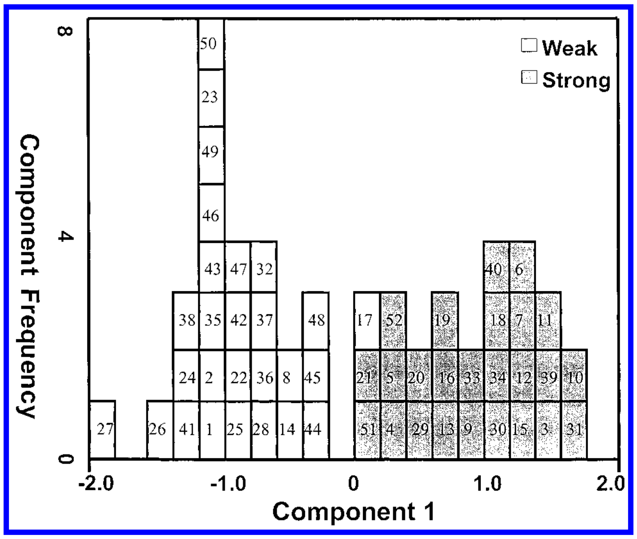

Figure 5. Plot of the discriminant function for canonical variate analysis of 26 PCDF, 14 PCDD, and 12 PCB compounds using 15 chemical shift bins.

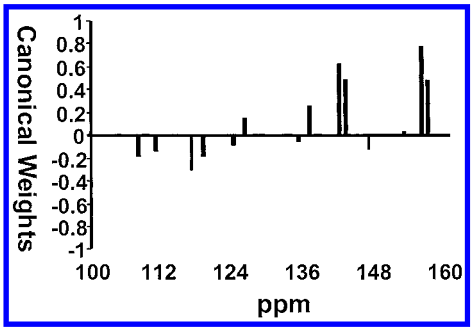

Figure 6. Plot of the factor loadings used in canonical variate analysis of 26 PCDF, 14 PCDD, and 12 PCB compounds using 15 chemical shift bins.

NMR bins used in each model. The 26 PCDF CoSA model had a higher q^2 than the models of 14 PCDD and 12 PCB CoSA models because there were more spectra in that model, as discussed above. However, there were also more chemical shifts in different bins from which to draw constructive correlations. The greater number of spectra available and the greater number of bins both derive from the lesser symmetry of PCDF molecules compared to PCDDs or PCBs. From these results, we conclude that the 26 compound PCDF and 52 compound CoSA models were provided enough information to generalize about the relevant substances' binding affinity to the enzyme. By relevant substances, we mean that one would use the 26 PCDF model only to predict the affinities for other PCDFs or the 52 combined PCDD, PCDF, and PCB model to predict affinities for any of these three types of polychlorinated hydrocarbon. While the 14 PCDD and 12 PCB compound CoSA models represent a significant improvement over previously published QSAR-based predictive approaches, they still need more compounds in the model to be satisfactorily robust. Table 3 is a summary of similar results for the 2.0 ppm CoSA model's characteristics and performance with respect to the same four parameters.

Our QSDAR models did not use compounds that contain bromine, so a complete comparison would not be fair to any other models that did include brominated aryl hydrocarbons.[2,7,34,35] With that caveat we present these comparative statistics. Our 1 ppm resolution CoSA model based on five chemical shift bins for the 26 PCDF compounds had an r^2 of 0.93 and q^2 of 0.90. We used the structural parameters Lmax, HOMOs, E(HOMO−LUMO), log P, and GIW (the geometric analogue of Weiner topological indices) produced by Mekenyan and co-workers[2] to produce a five component model for 25 PCDF compounds (all 26 PCDF compounds except for 237-trichlorodibenzofuran) that had an r^2 of 0.85 and q^2 of 0.71. The best model for 39 dibenzofurans proposed by Turner, one that used three infrared EVA molecular descriptors, had an r^2 of 0.96 and a q^2 of 0.73 and a six component QSAR CoMFA model had an r^2 of 0.85 and a q^2 of 0.72. This paper's 2 ppm resolution CoSA model based on five chemical shifts bins for the 14 PCDD compounds had an r^2 of 0.91 and a q^2 of 0.81. The structural parameters Lmax, HOMOs, E(HOMO−LUMO), log P, and GIW presented by Mekenyan and co-workers[2] were used to produce a five component model for 14 PCDD compounds that had an r^2 of 0.95 and q^2 of 0.82. The model for 25 dibenzodioxins proposed by Turner that used two infrared EVA molecular descriptors had an r^2 of 0.88 and a q^2 of 0.65, and a two component QSAR CoMFA model had an r^2 of 0.88 and a q^2 of 0.73.[7] Our CoSA model based on five bins for the 12 PCB compounds had an r^2 of 0.87 and a q^2 of 0.45. The model for 33 biphenyls proposed by Turner that used one infrared EVA molecular descriptor had an r^2 of 0.72 and a q^2 of 0.16, and a three component QSAR CoMFA model had an r^2 of 0.87 and a q^2 of 0.49.[7] The structural parameters Lmax, HOMOs, E(HOMO−LUMO), Log P, and GIW calculated by Mekenyan and co-workers[2] were used to produce a model for 12 PCB compounds that had an r^2 of 0.95 and q^2 of 0.79. Clearly the 12 PCB compounds does not contain structural diversity and enough examples in the training set to support one-dimensional NMR or IR spectra-activity relationship models.

The CoSA model based on 12 chemical shifts bins for the 52 polychlorinated aromatic compounds had an r^2 of 0.85 and a q^2 of 0.71. We used the structural parameters Lmax, HOMOs, E(HOMO−LUMO), log P, and GIW produced by Mekenyan and co-workers[2] to produce a model for 25 PCDF (all 26 PCDF compounds except for 237-trichlorodibenzofuran), 14 PCDD, and 12 PCB compounds that had an r^2 of 0.72 and q^2 of 0.60. The model for 77 dibenzodioxins, dibenzofurans, and PCBs proposed by Turner that used three infrared EVA molecular descriptors had an r^2 of 0.87 and a q^2 of 0.68, and the QSAR CoMFA model with six components had an r^2 of 0.88 and a q^2 of 0.71.[7]

The CoSA models for 26 PCDF, 14 PCDD, and combined 52 compounds provided results at least equivalent to and often superior to other modeling methods. QSDAR has further significant advantages with respect to objectivity in development of appropriate molecular descriptors and the rapidity with which the results can be obtained using inexpensive spectral prediction and pattern recognition software operated on ordinary PC platforms.

A canonical analysis using nonquantitative SDAR techniques of the 52 polychlorinated aromatic compounds showed, as displayed in Figure 5, a separation of the 27 strong binders and 25 weak. The SDAR approach does not attempt to quantify the relative binding affinity of the model compounds, only to group them into one of a small number of categories, in this case only two. This type of approach may find use applied to the preliminary assessment of complex mixtures of the type in which PCDD, PCDF, and PCB contamination's typically occur. We envision the possibility that predicted spectral patterns of pure compounds could be used to build an SDAR classification into which the real spectra of mixtures may be input as unknowns. By the clustering of the mixture spectrum into either the strong or weak binder groups one might infer the mixture's relative toxicity with respect to AhR-mediated effects. This capability has not been demonstrated by the work presented here. However, it seems that any predictive scheme based on structure−activity relationships models could not possibly address mixture phenomena in which there is no known single structure whose binding affinity would be estimated. Approaches based on spectral data at least seem to offer a possible solution if one is willing to settle for a nonquantitative or semiquantitative estimate, that is, a classification.

Finally, use of simulated ^{13}C NMR data enables modeling of both structurally similar and dissimilar compounds with respect to their receptor interactions. The accuracy of the first three 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. The accuracy of the QSDAR and SDAR models demonstrates the same capability for dissimilar compound sets. Like electrotopological-states calculations, ^{13}C NMR spectral data contain information about the electronic structure and topological environment of each atom in a molecule.[36,37] Data from other calculations may be added to ^{13}C NMR spectral data to produce a QSDAR model that has a better LOO cross-validated variance than that seen in QSAR models either based only on separate calculations for electrostatics and steric interactions or on spectral data alone. The cross-validated variance of QSDAR models based on simulated ^{13}C NMR data may improve as the errors introduced by simulation of the ^{13}C NMR data are further reduced.

Our CoSA modeling which 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. 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 and 2 ppm bins seem to work best.

One possible reason that CoSA modeling 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 possible reason for the power of CoSA modeling is the data in every bin is completely "orthogonal" to every other spectral bin. This allows for the information in any bin to constructively add information to the model.

The use of predicted chemical shifts is not necessary to build the SDAR models. We have, for example, used experimental chemical shifts to model estrogen receptor binding and biodegradation.[23,24,38] The use of predicted chemical is not necessary to build the SDAR models. The use of predicted chemical shifts allows the modeler to save the time and money required for buying or making the compounds and running the NMR spectra. It also limits toxic exposures to compounds with high cancer risks and produces a completely computer driven model that is faster and quicker than QSAR modeling.

REFERENCES AND NOTES

(1) Safe, S. Polychlorinated biphenyls (PCBs), dibenzo-p-dioxins (PCDDs), dibenzofurans (PCDFs), and related compounds: Environmental and mechanistic considerations which support the development of toxic equivalency factors (TEFs). *Crit. Rev. Toxicol.* **1990**, *21*(1), 50−88.

(2) Mekemyan, O. G.; Veith, G. D.; Call, D. J.; Ankley, G. T. A QSAR evaluation of Ah receptor binding of halogenated aromatic xenobiotics. *Environ. Health Perspect.* **1996,** *104*, 1302−1310.

(3) Bhandiera, S.; Sawyer, T.; Romkes, M.; Zmudzka, B.; Safe, L.; Mason, G.; Keys, B.; Safe, S. Polychlorinated bibenzofurans (PCDFs): Effects of structure on binding to the 2,3,7,8-TDDD cytosolic receptor protein, AHH induction and toxicity. *Toxicology* **1984**, *32*, 131−144.

(4) Mason. G.; Farrell, K.; Keys, B.; Piskorska-Pliszczynska, J.; Safe, L.; Safe, S. Polychlorinated dibenzo-p-dioxins: Quantitative in vitro and in vivo structure−activity relationships. *Toxicology* **1986**, *41*, 21−31.

(5) Mason, G.; Sawyer, T.; Keys, B., Bandiera, S.; Romkes, M.; Piskorska-Pliszczynska, J.; Zmudzka, B.; Safe, S. Polychlorinated dibenzo-p-dioxins: Correlation between in vitro and in vivo structure−activity relationships. *Toxicology* **1985**, *37*, 1−12.

(6) Bandiera, S.; Safe, S.; Okey, A. B. Binding of polychlorinated biphenyls classified as either phenobarbitone-, 3-methylcholanthrene- or mixed type inducers to cytosolic Ah receptor. *Chem. −Biol. Interact.* **1982**, *39*, 259−277.

(7) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. Evaluation of novel infrared range vibration-based descriptor (EVA) for QSAR studies. 1. General application. *J. Comput.-Aided Design* **1997**, *11*, 409−422.

(8) Cramer, R. D.; Paterson, 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.

(9) Tong, W.; Perkins, R.; Xing, L.; Welsh, W. J.; Sheehan, D. M. QSAR models for binding of estrogenic compounds to estrogen receptor α and β subtypes. *Endocrinology* **1997,** *138*, 4022−4025.

(10) Hansch, C.; Leo, A. *Exploring QSAR − Fundamentals and Applications in chemistry and biology;* The American Chemical Society: Washington, DC, 1995.

(11) Oprea, T. I.; Garcia, A. E. Three-dimensional quantitative structure−activity relationships of steroid aromatase inhibitors. *J. Comput.-Aided Mol. Design* **1996**, *10*, 186−200.

(12) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S. Prediction of gas chromatographic retention times and response factors using a general quantitative structure−property relationship. *Anal. Chem*. **1994**, *66*, 1799−1807.

(13) Katritzky, A. R.; Mu, L.; Labanov, V. S.; Karelson, M. Correlation of boiling points with molecular structure. 1. A training set of 298 diverse organics and a test set of 9 simple inorganics. *J. Phys. Chem.* **1996**, *100*, 10400−10407.

(14) Fujita, T.; Iwasa, J.; Hansch, C. A new substituent constant, π, derived from partition coefficient. *J. Am. Chem. Soc.* **1964**, *86*, 5175−5180.

(15) Branbury, S. P. Quantitative structure−activity relationship and ecological risk assessment: An overview of predictive aquatic toxicology research. *Toxicology* **1995**, *25*, 67−89.

(16) Emsley, J. W.; Feeney, J.; Sutcliffe, L. H. *High-Resolution Nuclear Magnetic Resonance;* Pergamon Press Ltd.: Oxford, 1965; Vol. I, pp 1−287.

(17) De Dios, A. C.; Pearson, J. G.; Oldfield, E. Secondary and tertiary structural effects on protein NMR chemical shifts: an *ab initio* approach. *Science* **1993**, *260*, 1491−1496.

(18) Beger, R. D.; Bolton, P. H. Protein ϕ and ι dihedrals restraints determined from multidimensional hypersurface correlations of backbone chemical shifts and their use in the determination of protein tertiary structures. *J. Biomol. NMR* **1997**, *10*, 129−142.

(19) Wishart, D. S.; Sykes, B. D. Chemical shifts as a tool for structure determination. *Methods Enzymol*. **1994**, *239*, 363−92.

(20) *ACD/Labs CNMR software version 4.0*; Toronto, Canada.

(21) Meiler, J.; Meusinger, R.; Will, M. Fast determination of 13 C NMR chemical shifts using artificial neural networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1169−1176.

(22) Spectrum Research, NMRScape, Madison, WI.

(23) Beger, R.; Freeman, J.; Lay, J., Jr.; Wilkes, J.; Miller, D. ^{13}C NMR and EI mass spectrometric data-activity relationship (SDAR) model of estrogen receptor binding. *Toxicol. Appl. Pharmacol.* **2000**, *169*, 17−25.

(24) Beger, R. D.; Freeman, J. P.; Lay, J. O., Jr.; Wilkes, J. G.; Miller, D. W. The Use of ^{13}C NMR Spectrometric Data to produce a Predictive Model of Estrogen Receptor Binding Activity. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 219−224.

(25) Beger, R. D.; Wilkes, J. G. Developing ^{13}C NMR quantitative Spectrometric data-activity relationship (QSDAR) Models to the Corticosteroid Binding Globulin. *J. Comput.-Aided Mol. Design.* **2001,** in press.

(26) Bursi, R.; Dao, T.; van Wilk, T.; de Gooyer, M.; Kellenbach, E.; Verwer, P. Comparative spectra analysis (CoSA): Spectra as three-dimensional molecular descriptors for the prediction of biological activities. *J. Chem. Inf. Comput. Sci***.** **1999,** *39*, 861−867.

(27) Poland, A.; Knutson, J. C. 2,3,7,8-tetrachlorodibenzo-p-dioxin and related halogenated aromatic hydrocarbons: Examination of the mechanism of toxicity. *Annu. Rev. Pharmacol. Toxicol.* **1982**, *22*, 517−554.

(28) Poland, A.; Glover, E.; Kende, A. S. Sterospecific, high affinity binding of 2,3,7,8-tetrachlorodibenzo-p-dioxin by hepatic cytosol: evidence that the binding species is the receptor for induction of aryl hydrocarbon hydroxylase. *J. Biol. Chem.* **1976,** *251*, 493−494.

(29) Safe, S. Polychlorinated biphenyls (PCBs) and polybrominated biphenyls (PBBs): Biochemistry, toxicology and mechanism of action. *Crit. Rev. Toxicol.* **1984**, *13*, 319−95.

(30) Safe, S. H. Comparative toxicology and mechanism of action of polychlorinated dibenzo-p-dioxins and dibenzofurans. *Annu. Rev. Pharmacol. Toxicol.* **1986**, *26*, 371−399.

(31) StatSoft software, Tulsa, OK.

(32) Bremser, W. HOSE − a Novel substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355−365.

(33) Cramer, R. D.; Bunce, J. D.; Patterson, D. E Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. *Quant. Struct.-Act. Relat.* **1988**, *7*, 18−25.

(34) Rannug, U.; Sjogren, M.; Rannug, A.; Gillner, M.; Toftgard, R.; Gustafsson, J.-A.; Rosenkranz, H.; Klopman, G. Use of artificial intelligence in structure-affinity correlations of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) receptor ligands. *Carcinogenesis* **1991**, *12*, 2007−2015.

(35) Kafafi, A. A.; Afeefy, H. Y.; Said, H. K.; Hakimi, J. M. A new structure for Ah receptor binding. Polychlorinated dibenzo-p-dioxins and dibenzofurans. *Chem. Res. Toxicol.* **1992**, *5*, 856−862.

(36) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. E-state fields: Applications to 3D QSAR. *J. Comput.-Aided Mol. Design* **1996**, *10*, 513−520.

(37) De Gregorio, C.; Kier, L. B.; Hall, L. H. QSAR modeling with electrotopological state indices: Corticosteroids. *J. Comput.-Aided Mol. Design* **1998**, *12*, 557−561.

(38) Beger, R. D.; Freeman, J. P.; Lay, J. O., Jr.; Wilkes, J. G.; Miller, D. W. Producing ^{13}C NMR, Infrared Absorption and EI Mass spectrometric data monodechlorination models of Chlorobenzenes, Chlorophenols, and Chloroanilines. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1449−1455.

CI000312L