Multimode Ligand Binding in Receptor Site Modeling: Implementation in CoMFA

Viera Lukacova and Stefan Balaz*

College of Pharmacy and Center for Protease Research, North Dakota State University, Sudro Hall 8,
Fargo, North Dakota 58105

Received May 21, 2003

Receptor site modeling methods usually use one binding mode (conformation and/or orientation) for each ligand in a 1:1 complex with receptor. Multiple modes should be considered instead because (1) they have frequently been observed experimentally; (2) in a series, ligands can bind in single yet different modes; and (3) a series may only exhibit one but unknown mode and a few plausible modes must be examined. For multimode binding, the observed ligand/receptor association constant is the sum of the association constants that characterize individual binding modes. This relation, when applied to Comparative Molecular Field Analysis (CoMFA), results in a dependence of the observed binding energy on the probe energies that is nonlinear in optimized parameters. The dependence was linearized to allow parameter optimization by the partial least-squares method that was used iteratively until self-consistency. In addition to the standard CoMFA output, the procedure objectively selects one or a few optimal binding modes out of a dozen or more modes that are considered for each ligand. The approach was applied to published data for binding of 34 polychlorinated dibenzofurans to the aryl hydrocarbon receptor. Descriptive and predictive abilities of the 16-mode model were significantly better than for the one-, two-, and four-mode models. Predominantly, edge-aligned modes were selected that are seldom used in CoMFA. Since inclusion of multimode binding only changes the form of the correlation equation and does not affect the number of optimized parameters, the improvement is believed to be due to a more realistic description.

INTRODUCTION

Interactions of ligands with macromolecules are multifarious due to a variety of loosely defined binding sites, at which ligands can bind with various stoichiometries and in differing orientations. To define our scope in this diverseness, the following terms will be used in the present context. The *binding site* is the pocket or cavity of a macromolecule that is not much larger than interacting ligands so that only one ligand molecule can bind per site. For the studied system under given conditions, only one class of binding sites is assumed to exist. The (*binding*) *mode* denotes a specific bound conformation and/or orientation of a ligand molecule in the binding site. *Multimode binding* is encountered when a ligand binds to one binding site of a macromolecule with an overall 1:1 stoichiometry, whereby individual bound ligand molecules representing one molecular species exhibit different binding modes. If all bound ligand molecules were depicted simultaneously, the structures would, at least partially, overlap due to the limited size of the binding site. The *mode prevalence* is the ratio of the number of ligand molecules bound in the given mode to the total number of bound ligand molecules. For multimode binding, free energies of binding do not differ sufficiently for individual binding modes to warrant dominance of a single mode; rather, they determine the mode prevalences via the Boltzmann probabilities. With continuing sophistication of methods for structure determination, multiple binding modes have been observed experimentally with increased frequency.[1]

Conceptual receptor site models (also called three-dimensional quantitative structure−activity relationships−3D-QSAR) use, in the absence of macromolecular structural information, structures and binding affinities of ligands to deduce the hypothetical shape and properties of the binding site and to predict affinities of nontested ligands. If the ligands have a common scaffold, the scaffold conformations or orientations may vary among the bound ligands even if each ligand only has one binding mode.[2]

A 3D-QSAR analysis starts with an alignment of ligands or their placement in a putative binding site. Ambiguousness of this step increases with ligand flexibility. Among many subjective options, the most common starting points are ligand alignments, which utilize either atom-based or property-based superpositions according to a pharmacophore hypothesis.[3] The alignments of the ligand set are evaluated one at a time on the basis of calibration and prediction statistical indices. Occasionally, alternate binding modes are examined for the worst-fitting ligands.[4−7] This step reduces the subjectivity in the model construction even if finally one mode is selected for each ligand.[5,8] A brute-force examination of alternate binding modes of all ligands is practically impossible for any real-world QSAR problem due to a combinatorial explosion in the number of required 3D-QSAR analyses: $M_1 \times M_2 \times ... \times M_L$ analyses are needed for L ligands, if the *i*th ligand binds in M_i binding modes.[9]

Historically, the first 3D-QSAR method considering multiple binding modes was Comparative Molecular Field Analysis (CoMFA).[10] The procedure characterizes the hypothetical binding site by regression coefficients assigned to the probe/ligand interaction energies in individual points

* Corresponding author phone: (701)231-7749; fax: (701)231-8333; e-mail: stefan.balaz@ndsu.nodak.edu.

10.1021/ci034100a CCC: $25.00
Published on Web 10/28/2003

of a grid encompassing the aligned ligands. Multimode ligand binding is represented by a field of interaction energies that are obtained as weighed averages of the energies for individual modes, with either identical[6,11] or different[12] weights corresponding to anticipated mode prevalences. However, the prevalences depend, in a nonlinear way, on the products of the optimized regression coefficients with the energies (see eqs 3 and 4, and the accompanying text below). Therefore, an a priori representation of the multimode binding using just the probe/ligand interaction energies is not feasible, even if the prevalences would be known.

4D-QSAR analysis[13] relies on ensemble averaging to incorporate conformational flexibility and alignment freedom. The averaging has a similar effect as the field weighing in CoMFA. Many ligand conformations are generated by molecular dynamics and placed into the grid in different alignments. Occupancies of individual grid cells are used as descriptors in the analysis that includes partial least squares (PLS) and genetic function approximation.[14] The result is a manifold of 3D-QSAR models.

Functions similar to Boltzmann probabilities were used in pseudoreceptor models considering different conformations[15] and protonation states of ligands.[16] Interestingly, no proton affinities of individual ligands were needed in the latter analysis. The approach optimizes the atom composition in a multiatom envelope representing the pseudoreceptor by a genetic algorithm with crossovers and transcription errors. The result is a fuzzy family of several hundred models, which are visualized using most frequently occurring atom types (the frequencies were not given but probably were rather low).

This study describes a conceptual incorporation of the multimode binding with objective optimization of the mode prevalences into the most widely used 3D-QSAR method, CoMFA. The approach is applied to binding of polychlorinated dibenzofurans (PCDFs) to the aryl hydrocarbon receptor.[17,18] Several QSAR analyses of these data were published,[17,19] including two CoMFA studies.[20] Surprisingly, the results were rather modest, even though no common complicating factors (neither chemical attributes as conformational flexibility or ionization nor biological attributes such as subcellular distribution, metabolism, or different mechanisms leading to biological response) were encountered. We examined another complicating factor: multimode PCDF binding to the receptor due to the symmetry of the dibenzofuran skeleton.

THEORETICAL BACKGROUND

The Observed Association Constant of a Ligand Binding in Multiple Modes. For practical reasons, usually the total drug-receptor association constant K_i is only determined in experimental studies, and no attempt is made to analyze the population of bound molecules. Schematically, the fast and reversible 1:1 interaction of the receptor site R with the ith ligand L_i that binds in multiple modes is depicted as

$$\begin{tikzcd}
 & \mathrm{LR}_{i1} \\
\mathrm{L}_i + \mathrm{R} \arrow[ur, leftrightarrow, "K_{i1}"] \arrow[r, leftrightarrow, "K_{i2}"] \arrow[dr, leftrightarrow, "K_{im}"'] & \mathrm{LR}_{i2} \\
 & \mathrm{LR}_{im}
\end{tikzcd} \qquad (1)$$

The binding of the ligand in individual modes in the binding site is characterized by partial association constants K_{ij} (j = 1, 2, ... m; the number of binding modes m can differ for individual ligands, but the subscript i is omitted to avoid double subscripts). The total concentration of receptor site/ligand complexes is equal to the sum of concentrations of individual complexes in which the ligand binds in different binding modes. The observed overall association constant can then be expressed as[21]

$$K_i = \frac{[\mathrm{LR}_i]}{[\mathrm{L}_i] \times [\mathrm{R}]} = \frac{[\mathrm{LR}_{i1}] + [\mathrm{LR}_{i2}] + ... + [\mathrm{LR}_{im}]}{[\mathrm{L}_i] \times [\mathrm{R}]} = \sum_{j=1}^{m} K_{ij} \qquad (2)$$

The simple eq 2 is in accord with published analyses of formally analogous situations: the statistical thermodynamic[22] and equilibrium[21,23] treatment of multimode binding in ligand/protein interactions and kinetic analyses of reversible unimolecular reactions leading to different isomers[24] or other products.[25] Using eq 2, multimode binding can be incorporated into any conceptual 3D-QSAR method.

It should be noted that the ligand molecules in eq 1 represent one molecular species, most frequently the nonionized molecules. Equation 2 is not valid for binding of multispecies populations of ligand molecules created by ionization, isomerism, or tautomerism. Therefore, simple additivity of the partial association constants cannot be used in multispecies 3D-QSAR correlations, contrary to a recent suggestion.[16]

Multimode Binding in CoMFA. The association constant for a ligand binding in a particular binding mode is correlated to the ligand/probe interaction energies in CoMFA as[10,21]

$$K_{ij} = \exp(C_0 - C_\mathrm{e} \times E_{ij} + \sum_{k=1}^{f \times g} C_k \times X_{ijk}) \qquad (3)$$

The summation goes through $f \times g$ independent variables, where f is the number of used fields (steric, electrostatic, sometimes hydrophobic, polarizability, or hydrogen bonding fields) and g is the number of the used grid points. Occasionally, the number of variables can be lower since not all fields need to be used in each grid point. The independent variables, X_{ijk}, are the energies of interaction between a probe placed in the kth grid point and the ith ligand molecule in the jth binding mode. The regression coefficients C_k characterize significance of field contributions in each grid point for overall binding. Conformational energy E_{ij}, although seldom used in the one-mode CoMFA analysis, can be of importance when several binding modes are considered.[26] Standard one-mode CoMFA uses logarithmized eq 3 (with j omitted because j = 1) that is linear in the optimized regression coefficients C. For multimode ligand binding, the correlation equation of the observed association constant K_i with the probe/ligand interaction energies X_{ijk} results from combination of eqs 2 and 3:

$$K_i = \sum_{j=1}^{m} \exp(C_0 - C_\mathrm{e} \times E_{ij} + \sum_{k=1}^{f \times g} C_k \times X_{ijk}) \qquad (4)$$

The j-summation goes through m binding modes. Equation 4 is nonlinear in regression coefficients C.

Linearization of the Correlation Equation for CoMFA. The number ($f \times g$) of optimized coefficients C is usually much higher than the number of tested compounds. In original formulation,[10] the PLS technique was adopted to cope with this situation. To use the same approach, eq 4 needs to be linearized. One of the possibilities is to use the first two terms in the Taylor expansion of the exponentials in eq 4 as $\exp(x) \approx \exp(M) \times (1 - M + x)$ for x approaching M. Each exponential in eq 4 represents a K_{ij} and can be expanded around the number $M = \ln K_{ij}$ without introducing much error (less than 10% if the exponent x is from the interval $\langle \ln K_{ij} - 0.5; \ln K_{ij} + 0.5 \rangle$):

$$K_i = \sum_{j=1}^{m} K_{ij} \times \left(1 - \ln K_{ij} + C_0 - C_e \times E_{ij} + \sum_{k=1}^{f\times g} C_k \times X_{ijk}\right) \tag{5}$$

Collecting the terms with the regression coefficients C:

$$K_i - \sum_{j=1}^{m} K_{ij} \times (1 - \ln K_{ij}) = C_0 \times \sum_{j=1}^{m} K_{ij} - C_e \times \sum_{j=1}^{m} K_{ij} \times E_{ij} + \sum_{k=1}^{f\times g} C_k \times \sum_{j=1}^{m} K_{ij} \times X_{ijk} \tag{6}$$

The partial association constants K_{ij}, too, depend on the regression coefficients C and the intramolecular and interaction energies E_{ij} and X_{ijk}, respectively, according to eq 3. For now, let us ignore this fact. All the variables on the right side of eq 6 are summed through all m binding modes and contain K_{ij}. To avoid a large variation in the variables, eq 6 was normalized[27] using $1/K_i$:

$$1 - \sum_{j=1}^{m} \frac{K_{ij}}{K_i} + \sum_{j=1}^{m} \frac{K_{ij}}{K_i} \times \ln K_{ij} = C_0 \times \sum_{j=1}^{m} \frac{K_{ij}}{K_i} - C_e \times \sum_{j=1}^{m} \frac{K_{ij}}{K_i} \times E_{ij} + \sum_{k=1}^{f\times g} C_k \times \sum_{j=1}^{m} \frac{K_{ij}}{K_i} \times X_{ijk} \tag{7}$$

Eq 7 is the final form of the linearized eq 5 that was used in iterative optimization. Each ratio K_{ij}/K_i represents, after optimization, the prevalence of the jth mode in the binding of the ith ligand (cf. eq 2). This fact can be utilized in a physical interpretation of eq 7. Once the optimization procedure gets into its final stages, the sum of prevalences of all modes approaches unity. In this situation, the first two terms on the left side mutually cancel, and the remaining term becomes practically equal to the weighed average of $\ln K_{ij}$ (see eq 8 below). The independent variables on the right side of eq 7 are the intramolecular energies E_{ij} and the interaction energies for individual grid points X_{ijk}, both multiplied by prevalences of individual modes. In summary, eq 7 eventually converges to a form that is reminiscent of the weighed-fields approach.[12] This fact, however, should not be understood as a justification for the weighed-fields approach because eq 7 is a linearized form of the complete eq 4 and is meant for iterative optimization only (see below). The association constants of individual modes K_{ij} cannot be fixed before optimization as suggested by the weighed-fields approach because they depend on the regression coefficients C and energies E_{ij} and X_{ijk} according to eq 3. Initial estimates of K_{ij} are optimized along with the regression coefficients C.

PLS analysis is used for optimization, since the number of C is usually higher than the number of compounds.[28] Once C are obtained by PLS, new K_{ij} values are calculated from eq 3 and used to update the variables of eq 7. Repeating the procedure until self-consistency provides a stable set of optimized regression coefficients C. More details on the procedure can be found below.

METHODS

Studied Data Set. The multimode CoMFA procedure was applied to published binding data of a set of 34 polychlorinated dibenzofurans (PCDFs) to the aryl hydrocarbon (Ah) receptor[17,18] (Table 1). The data set was split into two subsets, the training set and the test set, consisting of 22 and 12 derivatives, respectively. Only the training data set was used for model calibration. The compounds for the test set were selected so that (i) their binding affinities were evenly distributed within the range of pEC_{50} values; (ii) all degrees of substitution were included (number of compounds − number of chlorine substituents: 1−0, 1−1, 1−2, 2−3, 3−4, 3−5, and 1−6); and (iii) each chlorine position was represented at least once. The test set also contained the congeners **1**, **4**, and **15** for which only semiquantitative estimates of binding affinity were available.

Structure Optimization. All molecules were built de novo using the sketch option of Sybyl.[30] Since chlorine substituents in adjacent ring positions preclude the use of standard force fields and simplified charge schemes, full geometry optimization and calculation of Mulliken partial atomic charges in vacuo were done with the GAMESS suite of programs[31] using ab initio approach with restricted Hartree−Fock wave function and basis set 6-31G(d). The charges were in qualitative agreement with data for chlorinated dibenzo-*p*-dioxins[32,33] as the closest published systems that were studied by similar or better methods.

Alignment. The prototypical Ah receptor ligand, 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD), in the planar conformation, was chosen as a template for alignment of PCDF molecules. PCDF molecules are comparatively rigid, planar, and similar in size. All superpositions were constructed in the way that the dibenzo-*p*-dioxin and dibenzofuran skeletons substantially overlapped. In absence of contrary experimental evidence, we assumed that PCDF molecules bind to identical receptor parts as displaced TCDD. This assumption was used as the first choice, and, since the results were satisfactory, no additional subspaces, neither in directions perpendicular to the TCDD skeleton plane nor reaching far beyond in the plane occupied by TCDD, were explored. The multimode CoMFA analyses were used to systematically examine 2, 4, and 16 binding modes for each ligand.

The two and four binding modes were constructed by superposition of 1, 4, 6, and 9 carbon atoms of TCDD and PCDF (Figure 1). The dibenzofuran skeleton is symmetric around the y-axis so the PCDF molecules can bind in the forward mode (A) as well as in the reversed (B) mode. The modes A and B were used in the two-mode CoMFA analysis.

The TCDD molecule contains two oxygen atoms, while PCDF molecules only have one oxygen. Hypothetically, a PCDF molecule could be oriented in the binding site with

Table 1. Polychlorinated Dibenzofurans with Binding Affinities to the Ah Receptor[17,18]

		pEC$_{50}$[b]				
			calculated/predicted[d]			
PCDF no.	Cl position(s)[a]	observed[c]	1 mode	2 modes	4 modes	16 modes
01[e]		<3.000	2.735	−0.349	1.783	1.980
02	2	3.553	3.333	3.607	3.758	3.554
03	3	4.377 ± 0.058	4.487	4.375	4.383	4.379
04[e]	4	<3.000	3.220	1.409	1.620	2.201
05	2 3	5.326	4.993	4.888	5.186	5.334
06	2 6	3.609	3.841	3.640	3.869	3.642
07[e]	2 8	3.590	5.218	4.837	4.563	3.901
08[e]	1 3 6	5.357	5.975	5.860	4.121	4.516
09[e]	1 3 8	4.071	7.005	6.891	5.688	5.923
10	2 3 4	4.721	4.964	4.827	4.798	4.713
11	2 3 8	6.000 ± 0.041	6.611	6.166	6.040	6.007
12	2 6 7	6.347	6.421	6.306	6.386	6.344
13	1 2 3 6	6.456	6.394	6.418	6.508	6.460
14[e]	1 2 3 7	6.959	8.598	8.584	9.570	8.458
15[e]	1 2 4 8	<5.000	5.861	5.470	5.241	5.129
16[e]	2 3 4 6	6.456	5.336	5.496	4.533	5.539
17	2 3 4 7	7.602	7.773	7.717	7.661	7.609
18	2 3 4 8	6.699	6.420	6.125	6.662	6.700
19	2 3 6 8	6.658	6.412	6.638	6.987	6.655
20	2 3 7 8	7.387 ± 0.059	7.027	7.060	6.166	7.392
21	1 2 3 4 8	6.921	6.895	6.860	6.926	6.920
22	1 2 3 7 8	7.128 ± 0.105	7.228	7.185	7.069	7.127
23	1 2 3 7 9	6.398	6.468	6.400	6.445	6.407
24	1 2 4 6 7	7.169	6.949	7.086	7.107	7.167
25	1 2 4 6 8	5.509	5.487	5.565	5.479	5.510
26[e]	1 2 4 7 8	5.886	5.818	5.630	5.283	5.172
27	1 2 4 7 9	4.699	4.633	4.942	4.653	4.693
28	1 3 4 7 8	6.699	6.589	6.731	6.706	6.694
29[e]	2 3 4 7 8	7.824 ± 0.028	6.706	6.671	6.356	7.314
30[e]	2 3 4 7 9	6.699	6.155	6.682	6.653	6.770
31	1 2 3 4 7 8	6.638	6.726	6.668	6.732	6.643
32	1 2 3 6 7 8	6.569 ± 0.137	6.608	6.604	6.696	6.562
33	1 2 4 6 7 8	5.081	5.288	5.141	5.211	5.098
34[e]	2 3 4 6 7 8	7.328 ± 0.036	6.195	6.035	6.480	6.356

[a] For numbering of the PCDF skeleton, see Figure 1. [b] Experimental EC$_{50}$ values (in mol/L) are given in the form[29] of pEC$_{50}$ = −log(EC$_{50}$) ∼ logK_i. [c] When not reported, standard deviation data not available. [d] Calculated and predicted pEC$_{50}$ values for the training set and the test set, respectively, come from eq 4 with regression coefficients optimized using the training set. [e] The test set member, not used in calibration of the model.

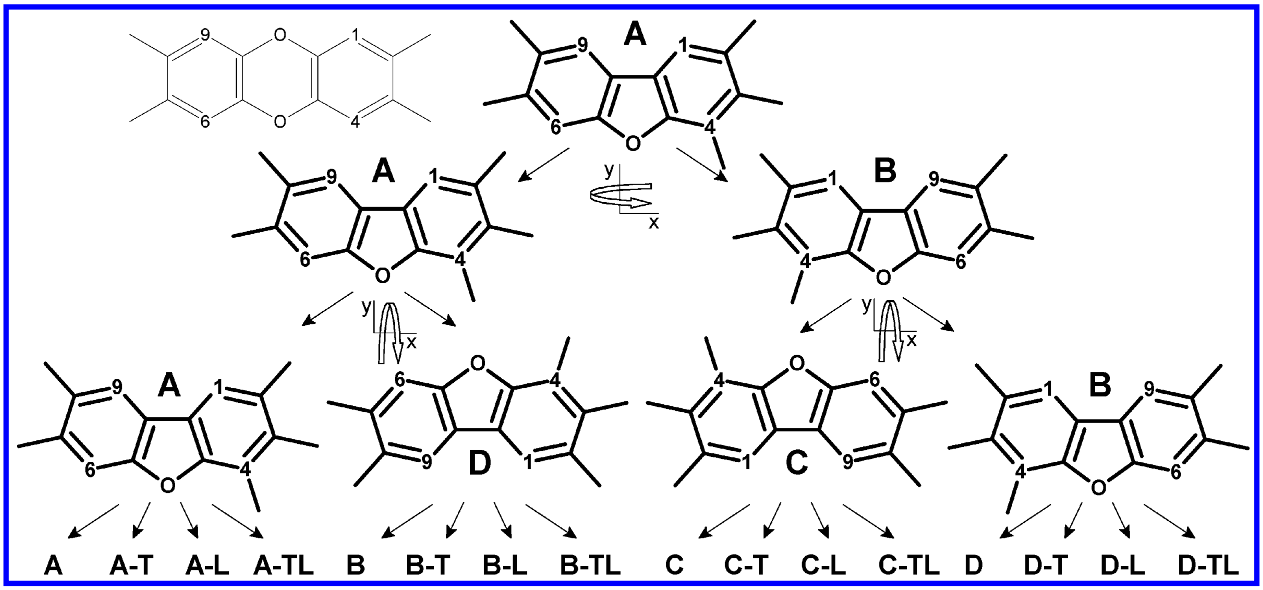

Figure 1. Alignments of PCDF congeners (compound **29**, Table 1 shown) in 2, 4, and 16 modes. Carbons 1, 4, 6, and 9 of TCDD (thin lines) and of the PCDF congeners were superimposed in different combinations. Mode A corresponds to the IUPAC nomenclature; other modes are obtained from the standard orientation (A) by flipping around the *y*-axis (mode B), *x*-axis (mode D), and both *x*- and *y*-axes (mode C). To obtain 16 modes, each of the modes A−D had the alignment modified by shifting the PCDF molecule, in the plane of the skeletons, to the top (T), the left side (L) or the top left corner (TL) of the box enclosing the TCDD molecule (for details see Figure 2).

oxygen atom in the position opposite to that in modes A and B, in the forward or reverse mode (modes D and C).

The four hypothetical binding modes differ in the 180°−rotation of the molecule around the *x*- and/or *y*-axes (Figure

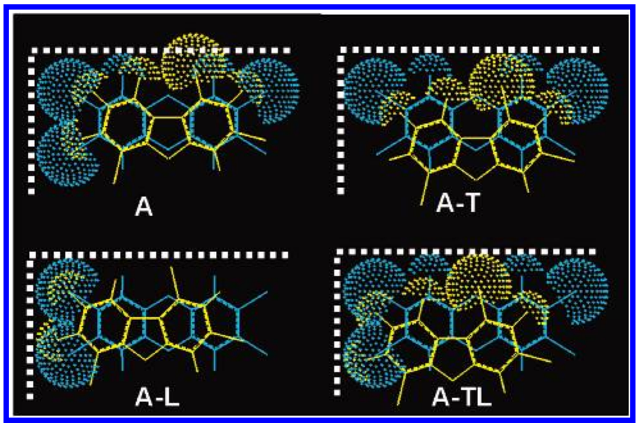

Figure 2. Sixteen binding modes were generated by translation of existing four binding modes A−D (mode A for compound **8** in Table 1 shown in yellow) in the plane of the skeletons to obtain the overlap of the van der Waals surfaces at the top (A−T), left (A−L), and both left and top (A−TL) sides of TCDD molecule (cyan).

1). The modes A−D were used in the four-mode CoMFA analysis.

Implementation of binding modes as suggested above assumes strong interactions of aromatic rings or oxygens with the binding site, which would hold skeletons of all PCDF molecules in the same position within the receptor site, regardless of their actual substitution (shape). This scenario does not seem plausible because the PCDF molecules exhibit a five-order span of binding affinities (Table 1), and the substitution pattern plays an important role.

If the skeleton-centering interactions are not dominant, the PCDF molecules could shift from the skeleton-superimposed positions inside the binding site to engage in the same attractive steric interactions as the TCDD molecule. In the absence of structural information, we decided to roughly describe the shape of the binding cavity as a rectangular box surrounding the TCDD molecule. The focus on the "two-dimensional" PCDF binding in the plane of the TCDD molecule, as described above, eliminates the two walls of the box that are parallel to the skeleton plane. Consideration of four symmetric binding modes allows for further reduction because the placement of boundary walls on opposite sides of the box should provide comparable results. Based on these assumptions, two putative boundary walls of the binding site were set: one on the left side and the other on the upper side of the TCDD molecule. The exact positions of these walls were given by the van der Waals surfaces of chlorines in positions 7 and 8 and hydrogens in positions 1 and 9 of TCDD. From each of the four modes A−D created by the 180°−rotation around the *x*- and *y*-axes and superposition of the skeletons, three more modes were formed by translation of the PCDF molecule in the skeleton plane to the left wall, the top wall, and both left and top walls so that the van der Waal surfaces of the ligand atoms touch the planes (Figure 2). An illustration of the resulting alignments for 1, 2, 4, and 16 modes is provided in Figure 3.

CoMFA Interaction Energy Calculations. For each ligand in each binding mode, steric and electrostatic interaction energies X_{ijk} were calculated at each lattice intersection of a regularly spaced rectangular grid (2 Å in each coordinate direction). An sp^3 carbon atom with the charge +1 was used as a probe. The grid extended approximately 4.0 Å in every direction away from the aligned molecules. The grid's coordinates were as follows: −8 to 8 Å along the *x*-axis, −6 to 6 Å along the *y*-axis, and −4 to 4 Å along the *z*-axis, with the center of the TCDD molecule being placed in the origin. The maximum allowable steric and electrostatic energy values were set to 30 kcal/mol. The electrostatic energy term was not calculated in the grid points where the steric energy term reached the limit value (30 kcal/mol). The distance-dependent dielectric constant was used.

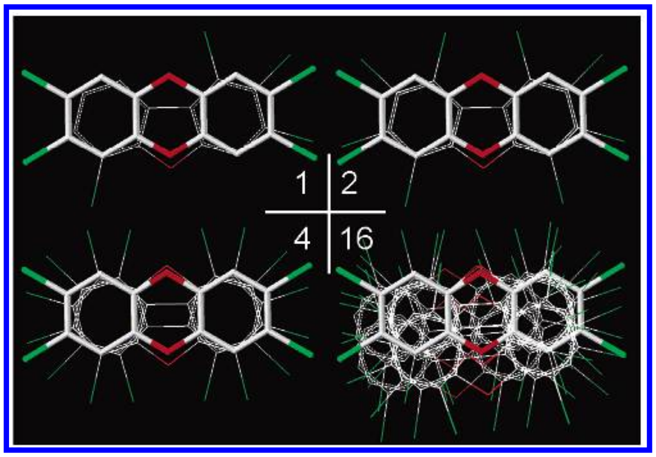

Figure 3. Alignments of a PCDF congener (**13**, Table 1, thin lines) to the TCDD molecule (cylinders) in one or several modes (the numbers shown). The one-, two-, and four-mode alignments are based on superposition of skeleton carbons 1, 4, 6, and 9 in different orientations (Figure 1). In the 16-mode alignment, the PCDF molecules are matched with the top and left parts of the TCDD molecule.

The energies X_{ijk} in each grid point were entered into the *energy table*. Its dimensions are as follows: the number of rows given by the product of the ligand number (34) and the number of modes (1, 2, 4, or 16) and the number of columns equal to the number of used grid points (315) multiplied by the number of used fields (2). Normally,[26] the intramolecular energies E_{ij} would also form a column in the table. The energy table was populated once at the beginning of the optimization.

Variable preselection was used for multimode analyses. Most of the 630 X_{ijk} values were insignificant due to low variation throughout the set of PCDF molecules. Only the variables with the standard deviation SD > 3 and sustained variability (as described in part Results and Discussion: Optimization of Regression Coefficients) were selected for optimization.

Using Smart Region Definition (SRD) within GOLPE,[34] the chosen columns were sorted into groups carrying similar information. First, the PCA model with dimensionality five was created. Then SRD was performed in the chemometrical space of PCA loadings. The most informative variables (seeds, the total number 31) were selected, and the other variables were assigned to the seeds if their distance from the seed was less than 1.0 Å. Resulting Voronoi polyhedra were merged, if the distance between their seeds was 2.0 Å or less and if they contained the same information as assessed by the correlations of the average values of all, positive, and negative point energies in the regions (Pearson's $R > 0.8$ for averages of all point energies and $R > 0.5$ for averages of positive and negative point energies). Some groups contained variables corresponding to grid points symmetrically placed in space. Because the PCDF molecules have a symmetric skeleton and the majority of binding modes are symmetric as well, the points aligned symmetrically are expected to carry similar information. Therefore, the symmetric groups were merged.

Optimization. Iterative optimization based on eq 7 requires good initial estimates. Since the number of regression coefficients C precludes the exhaustive search, a forward-selection procedure (described in detail in Results and Discussion) was adopted. From the groups containing similar information, the variables were gradually added to the current correlation equation according to following criteria: (i) grid points corresponding to the chosen variables were evenly distributed around the molecule, (ii) all groups were represented more or less equally, and (iii) both fields were equally represented, but the variables with steric and electrostatic interaction energies might not correspond to the same grid points.

The left-side and right-side variables of eq 7 were organized into the *data table*. Since the summations in eq 7 go through all binding modes, the data table only contains one row per ligand. Initial estimates of the regression coefficients C were used to calculate the K_{ij}/K_i ratios using eq 3. Alternatively, starting estimates for the prevalences of individual binding modes K_{ij}/K_i can be chosen either en bloc for the whole set or individually for each ligand. The composite variables for each of the regression coefficients C and the left side in eq 7 were calculated for each ligand and entered into the data table after each iteration.

Iterative PLS procedure optimizes the nonlinear regression coefficients C in eq 4 for the given set of variables using the linearized eq 7. Each iteration in the Iterative PLS Procedure includes four steps: (i) setting the values of the regression coefficients C, which can either be initial estimates or the set from the previous run plus the new regression coefficients; (ii) calculation of the partial association constants K_{ij} using eq 3; (iii) calculation of the variables of the linearized eq 7 using the resulting K_{ij} values; and (iv) calculation of the regression coefficients C in eq 7 by PLS. The dependence of the fit quality on the number of components in each PLS run exhibits several extremes with unpredictable positions (data not shown). Therefore, in step (iv) of each iteration, the PLS correlation equations for the numbers of components ranging from one to the maximum[35] had to be enumerated, and no extreme-seeking methods (e.g. golden section[36]) could be applied to find the optimal number of components faster and to shorten the computations.

The Iterative PLS Procedure is finished when the change in the sum of squares of errors (SSE) falls below certain limit in consecutive iterations or when a limit number of iterations is reached, whatever occurs earlier. Among correlation equations formed in individual iterations, the best equations for a particular set of independent variables are then chosen based on the lowest SSE. The best equations are statistically characterized by calculation of the correlation coefficient (R), the predictive sum of squares of deviations (PRESS), and the predictive correlation coefficient (q) based on the leave-one-out cross-validation with the optimal number of components.[37]

Run Time. All calculations except the quantum mechanical characterization were done in the QSAR module of Sybyl,[30] with nonstandard procedures coded in the Sybyl Programming Language. The presented analysis required about 20 days on a SGI Octane with two R10000 processors with 250 MHz clock speed and 512 MB RAM. The time still compares very favorably with the period that would be needed to examine even a small portion of possible combinations of alternate binding modes using the one-mode approach. The multimode procedure can easily be optimized for faster execution by algorithm adjustment, recoding the procedure in a faster language, as well as by parallelization. Speed was not the main concern in the presented proof-of-the-concept study.

RESULTS AND DISCUSSION

PCDF Binding to Ah Receptor. The application of the multimode CoMFA procedure is demonstrated using published data on binding of PCDFs (Table 1) to the Ah receptor. Binding affinities of 34 congeners (Table 1) were carefully determined (cf. very low standard errors, where available) in a single laboratory[17,18] as the displacement of radiolabeled TCDD (see Figure 1 for structure). PCDFs (Table 1, Figure 1) do not contain rotatable bonds and differ only in the number and positions of the chlorine substituents. These characteristics make the data set a superb object for 3D-QSAR studies.[17,19,20] We examined multiple binding modes of PCDFs due to symmetry of their skeleton as a factor that could improve the correlation.

Multimode Alignments. Multimode binding was systematically analyzed for 2, 4, and 16 hypothetical modes of each ligand (Figure 1). Two and four binding modes were generated by atom-based superposition of the PCDF and TCDD skeletons in different orientations. Sixteen modes resulted from translation in the skeleton plane of the PCDF ligands in each of the superposition-based four modes to ensure the contact with the left, top, or both left and top walls of putative binding site represented by a box enclosing the TCDD molecule (Figure 2). These edge-based alignments are seldom used in the 3D-QSAR studies, despite their obvious rationale. An illustration of all alignments for one PCDF congener is provided in Figure 3. For 16 modes, the aligned molecules represent a really complex cluster.

Variable Preselection. Selection of variables to be added to the correlation equation was done with the aim to maintain the maximum amount of information per added variable. This intent led to three basic criteria for variable selection: (1) high and sustained variability, (2) a minimal number of collinear variables carrying similar information, and (3) more or less even distribution of selected grid points around the aligned molecules.

Variability of independent variables X_{ijk} is usually characterized by the standard deviation (SD). This approach, however, does not uncover the situations when variability is high due to one or two extreme X_{ijk} values that substantially differ from the rest. Technically, this situation arises when only one or two ligands in the series occupy certain parts of the 3D-space[34] or exhibit strong electrostatic interactions in the space where other aligned ligands have nonpolar atoms. These singularities lead to correlations that lack physical meaning. The leave-one-out cross-validation effectively eliminates the correlations with singularities due to only one extreme X_{ijk} value per variable; two or three extreme X_{ijk} values per variable go unnoticed. The singularity problem might be less severe in one-mode CoMFA because of a high number of variables that are included in the final correlation. In multimode CoMFA, the forward-selection procedure could pick the variables with singularities and use them as indicator variables to account for the binding differences of respective ligands. Sustained variability, i.e., variability due to at least four different interaction energy values, was checked for variables exhibiting a large difference between the median and the mean values.

Singularities are frequently treated improperly in the CoMFA analyses: the columns with singularities are eliminated, but the ligands that cause the singularities are retained in the data set, although the interactions of some parts of their molecules are not included in the model due to eliminated columns. Since the description of interactions of these ligands is incomplete, they must be excluded from analysis. If, after model calibration, their predicted affinities agree with experimental data, a conclusion can be made that the regression coefficients for the singular grid points are close to zero and, consequently, the singularity space is probably a water-filled cavity.

Optimization of the Regression Coefficients. The PLS technique works with equations that are linear in regression coefficients. To enable PLS optimization, eq 4 was linearized and normalized (eq 7). Iterative application of eq 7 requires good initial estimates of the regression coefficients. Since the high number of variables precludes an exhaustive evaluation of possible initial estimates, we used a forward-selection method (Figure 4) with gradual addition of variables, which were selected on the basis of high and sustained variability,[38] low collinearity, and even spatial distribution. The SRD procedure within GOLPE[34] classified the variables into several groups, from which the variables were gradually added to the working set so that the grid points corresponding to the chosen variables were evenly distributed around the molecules, and all groups and both fields were represented more or less equally. The optimization process consists of three phases: Outlining, Shaping, and Detailing (Figure 4). The regression coefficients can be optimized using other approaches, e.g. genetic algorithms.

In the first phase of optimization (Outlining, Figure 4), only six variables (three steric and three electrostatic interaction energies X_{ijk}) were used to enable a complete brute-force evaluation of a large number of initial estimate sets within a reasonable time. For the sets of regression coefficients C resulting from all plus/minus combinations of numbers 0.1, 0.2, ... 1.0, the correlations were first examined using the total binding constants K_i calculated from eq 4 without optimization. The 15% of the C sets with the lowest SSE were optimized by the Iterative PLS Procedure. The Outlining phase provided the correlations with R^2 ranging from 0.3 to 0.7 for various considered numbers of modes. The results for 16 binding modes are shown in Figure 5.

The best 5% of the C sets advanced to the second phase (Shaping, Figure 4). Here, the group of six new variables was initially added to the best C sets with zeros as the initial estimates and all regression coefficients C_k were optimized by the Iterative PLS Procedure. The correlation improved to R^2 in the range 0.6−0.8 (Figure 5). Further additions of groups of 12 variables led to further increase in the values of R^2 between 0.9 and close to 1.0 (Figure 5). Twenty-four variables were sufficient to obtain a satisfactory correlation for 16 modes (Figure 5) as well as for two and four modes (data not shown). However, at least 36 columns (Figure 5) for 16 binding modes (48 columns for two and four binding modes − data not shown) were needed to obtain a correlation that would also have good leave-one-out predictions. After

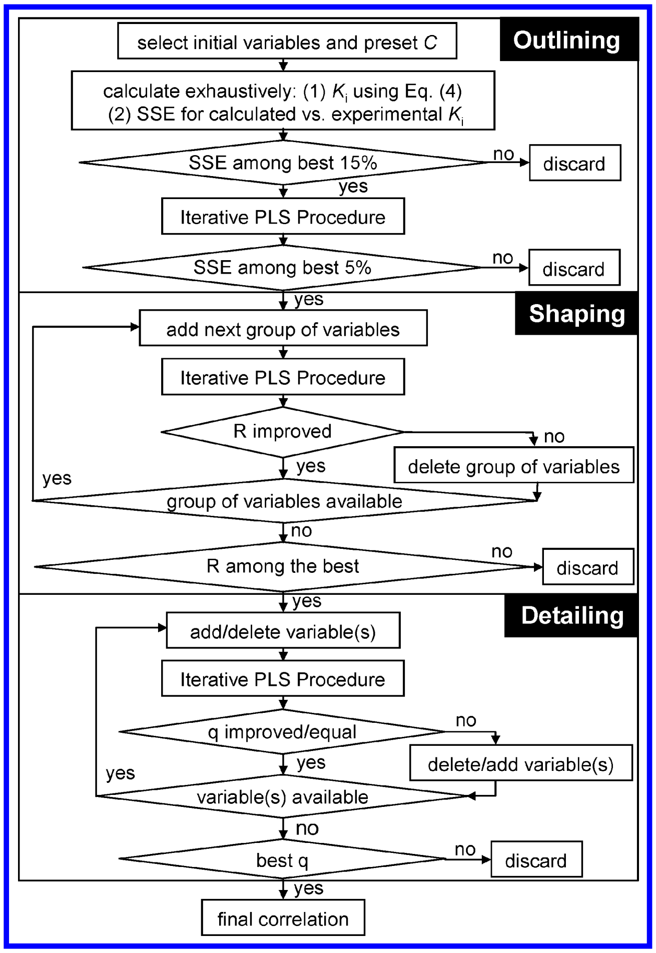

Figure 4. Forward-selection procedure for optimization of regression coefficients in eq 4 by iterative use of linearized eq 7. Outlining, Shaping, and Detailing are three phases of the optimization. In detailing, additions preceded deletions.

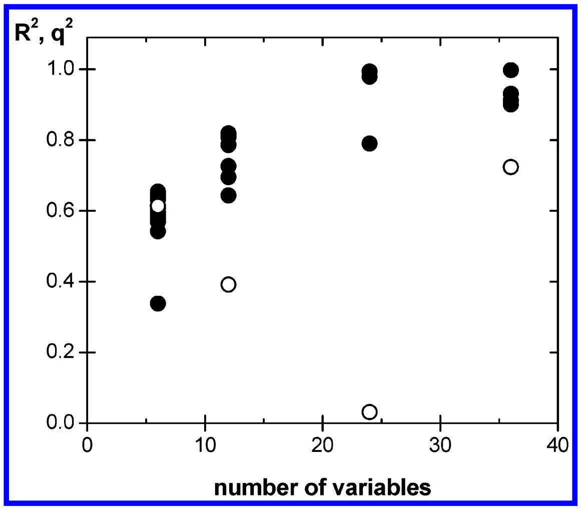

Figure 5. Improvement of the fit (R^2 − ●) and leave-one-out predictions (q^2 - ○) of the correlation with 16 binding modes with addition of variables in the second phase (Shaping − Figure 4). The shown q^2 values are valid for the sequence that led to the best correlation in the second phase.[37] The data for the six variables represent some results of the first phase (Outlining − Figure 4).

Table 2. Quality of the CoMFA Models Considering Various Numbers of Binding Modes Treated by the Present Approach and by the Weighed-Fields Approach[12]

	CoMFA				
	1 mode	2 modes	4 modes	16 modes	weighed fields[a]
no. of variables	570	26	22	22	570
		Training Set[b]			
SSE	1.097	0.771	1.799	0.002	2.613
R^2	0.963	0.974	0.940	0.999	0.944
PRESS	6.391	2.120[c]	2.476[c]	1.167[c]	16.782
q^2	0.786	0.929[c]	0.917[c]	0.961[c]	0.639
		Test Set[b]			
RMS pred. error	1.430	1.349	1.447	1.002	*a*

	predictions for semiquantitative data: pEC$_{50}$				
compound[d] (pEC$_{50}$)	1 mode	2 modes	4 modes	16 modes	weighed fields[a]
1 (<3.000)	2.735	−0.349	1.783	1.980	2.609
4 (<3.000)	3.220	1.409	1.620	2.201	3.445
15 (<5.000)	5.861	5.470	5.241	5.129	5.230

[a] All compounds with the precise binding affinities were used as the training set; the test set only comprised compounds **1**, **4**, and **15** with semiquantitative affinities (Table 1). [b] The ligands forming the training and test sets are specified in Table 1. [c] Not used to assess predictivity.[37] [d] Structures of the ligands are in Table 1.

this point, addition of further 12-variable groups did not significantly improve any statistical parameter. In fact, the leave-one-out correlation coefficient q^2 decreased,[37] implying that the new variables (or most of them) brought just more noise to the correlation.

The best correlation equations from the second phase (see Figure 5 for its statistical indices for the 16-mode analysis) were fine-tuned in the third phase of model development (Detailing, Figure 4). The groups of 6 or 12 variables, which were rejected in the second phase (Shaping), were broken into smaller sets of 1−6 variables and tested. For the 2- and 16-mode analyses, no further improvement was reached. For four modes, the correlation that could not be further improved by addition of variables contained 52 variables. The Detailing phase (Figure 4) was finished with the one-by-one elimination of variables. The variables, showing the lowest variability when multiplied by the corresponding regression coefficient *C*, were omitted if the impact on the q^2 value was negligible.[37]

Statistical Evaluation of Mode Addition. Calibration statistics (SSE and R^2 for the training set) and prediction statistics (RMS predicted error for the test set[39]) for the best correlations for each analyzed number of binding modes are summarized in Table 2. The q^2 and PRESS values for the leave-one-out cross-validation are also shown for completeness. However, these values only characterize predictivity of the one-mode model and the weighed-field model; they cannot be used for the multimode models because the omission of the ligands for cross-validation occurred in the last stages of optimization.

The one-mode CoMFA analysis for the training set provided better calibration indices (Table 2) than the published studies with the whole data set. The leave-one-out predictive ability slowly increased with the number of the latent variables and peaked at 10 latent variables. For illustration, the q^2 values were 0.097 (2), 0.344 (4), 0.508 (6), and 0.773 (8) for the numbers of latent variables given in parentheses. However, the test set-based predictive ability was not impressive: the statistics for the test set were rather insignificant and two (**4** and **15**, Table 1) out of three congeners with semiquantitative data had incorrectly predicted affinities (Table 2).

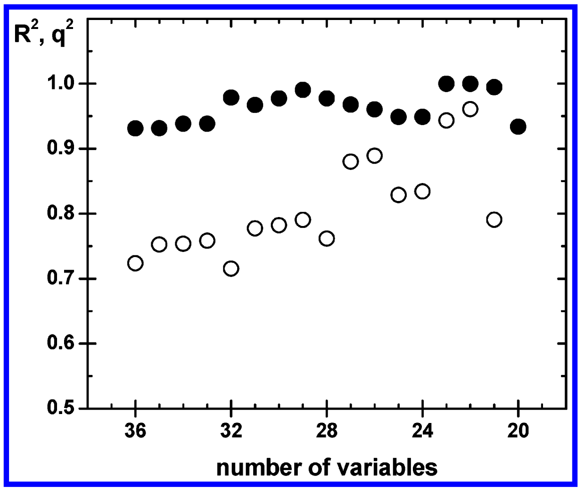

Figure 6. Development of R^2 (●) and q^2 (○) of the correlation with 16 binding modes during the third optimization phase (Detailing − Figure 4). Addition of more columns did not improve the correlation (data not shown). The stepwise removal of columns increased q^2 for the training set.[37] The indices for the best correlations for the given numbers of variables are shown. The correlation with 22 variables was selected as final. For 20 variables, q^2 was out of scale.

The two- and four-mode correlations show no improvement in the predictive ability (SSE and R^2 for the test set) in comparison with the one-mode correlation. This fact can be explained by inspection of the individual mode prevalences for the 16-mode model (Table 3). The modes formed by the skeleton alignments (A, B, C, and D) do not contribute to overall binding for any of the compounds in neither training nor testing data sets. Incorporation of the edge-based modes related to a putative receptor site that are only available in the 16-mode setup was necessary for a significant improvement in both the calibration and prediction statistics. These results also suggest that the enhancement of the model quality is not due to an increased number of possibilities resulting from addition of binding modes. For this assumption to be true, the two- and four-mode setups would have to provide better statistical indices than the one-mode approach. As can be seen in Table 2, this is not the case.

The overall best correlation was produced for 16 starting binding modes for each compound (see Figure 6 for the Detailing phase of the model development). The 16-mode CoMFA model compares favorably with the one-mode model as illustrated in Table 2 and Figure 7: calibration indices (SSE, R^2 for the training set) and, more importantly, the prediction indices (RMS prediction error for the test set) are significantly better for the multimode approach. Noticeably, the worst one-mode predictions (**7** and **9**, Table 1) were substantially improved by the multimode approach. The 16-mode predictions for the semiquantitative data (**1**, **4**, and **15**, Table 1) are also the best among the used approaches, with only a close miss for congener **15**.

Statistical indices for CoMFA with weighed contributions of individual modes,[12] although included in Table 2, will be analyzed later.

Table 3. Prevalences of Individual Binding Modes (Figures 1−3) for the 16-Mode Correlation

	prevalence for the mode^{b}															
PCDF no.^{a}	A	A−T	A−L	A−TL	B	B−T	B−L	B−TL	C	C−T	C−L	C−TL	D	D−T	D−L	D−TL
01^{c}	–	–	–	–	–	–	–	–	–	**0.50**^{d}	–	–	–	**0.50**^{d}	–	–
02	–	–	–	–	–	–	–	–	–	–	–	–	–	**0.82**	0.09	0.09
03	–	–	–	–	–	–	–	–	–	**0.38**	–	**0.56**	–	0.05	–	0.01
04^{c}	–	–	–	–	–	–	–	–	–	0.06	0.03	–	–	**0.89**	0.02	-
05	–	0.01	–	–	–	0.02	–	0.03	–	–	–	–	–	**0.14**	**0.19**	**0.61**
06	–	–	–	–	–	–	–	–	–	–	0.08	–	–	–	**0.92**	–
07^{c}	–	0.03	–	–	–	0.03	–	–	–	0.03	**0.15**^{d}	**0.29**^{e}	–	0.03	**0.15**^{d}	**0.29**^{e}
08^{c}	–	**0.64**	–	–	–	**0.18**	–	**0.18**	–	–	–	–	–	–	–	–
09^{c}	–	–	–	0.01	–	**0.43**	–	**0.48**	–	–	–	–	–	–	–	0.08
10	–	0.03	–	–	–	0.01	–	0.01	–	–	–	–	–	0.01	**0.94**	–
11	–	–	–	–	–	0.02	–	0.03	–	–	–	–	–	–	**0.15**	**0.80**
12	–	–	–	–	–	–	–	–	–	-	0.01	–	–	**0.44**	–	**0.55**
13	–	**1.00**	–	–	–	–	–	–	–	–	–	–	–	–	–	–
14^{c}	–	**0.30**	–	**0.33**	–	–	–	–	–	–	–	–	–	**0.14**	–	**0.23**
15^{c}	–	0.08	–	–	–	–	–	–	–	–	–	–	–	–	**0.92**	–
16^{c}	–	–	–	–	–	–	–	–	–	–	–	–	–	0.01	**0.99**	–
17	–	0.01	–	0.01	–	–	–	–	–	–	–	–	–	**0.51**	–	**0.47**
18	–	0.02	–	–	–	0.01	–	0.01	–	–	–	–	–	–	**0.96**	–
19	–	–	–	–	–	**0.49**	–	**0.46**	–	–	–	–	–	–	0.05	–
20	–	**0.23**^{d}	–	**0.27**^{e}	–	**0.23**^{d}	–	**0.27**^{e}	–	–	–	–	–	–	–	–
21	–	0.01	–	–	–	–	–	–	–	–	–	–	–	–	**0.99**	–
22	–	**0.41**	–	**0.45**	–	0.03	–	0.03	–	–	–	–	–	0.03	–	0.05
23	–	**0.46**	–	**0.52**	–	0.01	–	0.01	–	–	–	–	–	–	–	–
24	–	**0.57**	–	**0.43**	–	–	–	–	–	–	–	–	–	–	–	–
25	–	0.02	–	–	–	–	–	–	–	–	–	–	–	–	**0.98**	–
26^{c}	–	**0.45**	–	**0.48**	–	0.04	–	0.03	–	–	–	–	–	–	–	–
27	–	**0.41**	–	**0.40**	–	**0.18**	–	0.01	–	–	–	–	–	–	–	–
28	–	–	–	0.01	–	**0.53**	–	**0.46**	–	–	–	–	–	–	–	–
29^{c}	–	**0.45**	–	**0.51**	–	0.02	–	0.02	–	–	–	–	–	–	–	–
30^{c}	–	**0.49**	–	**0.49**	–	0.01	–	0.01	–	–	–	–	–	–	–	–
31	–	**0.14**	–	**0.67**	–	0.10	–	0.09	–	–	–	–	–	–	–	–
32	–	**0.46**	–	**0.42**	–	0.06	–	0.06	–	–	–	–	–	–	–	–
33	0.02	**0.50**	0.02	**0.42**	–	0.02	–	–	–	–	0.02	–	–	–	–	–
34^{c}	0.01	**0.26**^{d}	0.01	**0.22**^{e}	0.01	**0.26**^{d}	0.01	**0.22**^{e}	–	–	–	–	–	–	–	–
RO^{f}	–	**0.20**	–	**0.17**	–	0.08	–	0.07	–	0.03	0.01	0.02	–	0.10	**0.22**	0.09

[a] Structures of ligands are given in Table 1. [b] Calculated as K_{ij}/K_i, with the partial association constants K_{ij} obtained from eq 3. The hyphen indicates prevalence or relative occupancy equal to 0.00. Significant modes that contributed more than 10% to overall binding are shown in boldface. [c] The test set member. [d,e] The modes with identical superscripts are equivalent due to symmetrical substitutions. [f] Relative occupancy of individual modes.

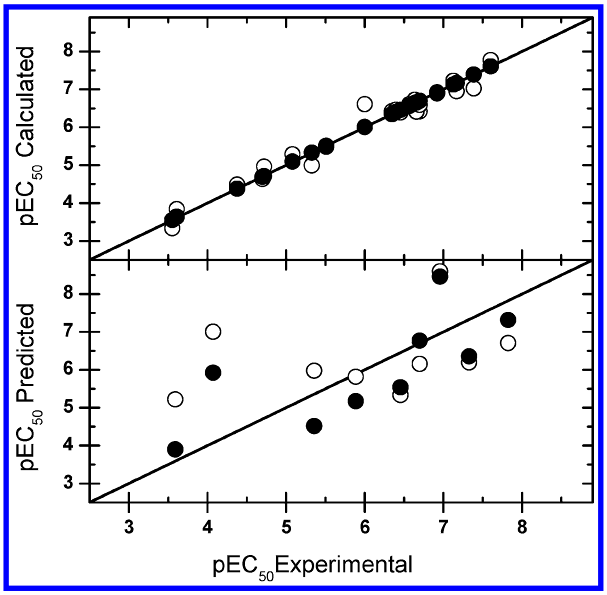

Figure 7. Calculated and predicted affinities for the test (top) and the training (bottom) sets, respectively, versus experimental affinities for the correlations considering one binding mode (○) and 16 binding modes (●). Predicted binding affinities are for compounds in the test data set using the model developed for the training data set.

It is worth mentioning that the multimode approaches do not optimize more regression coefficients C than the one-mode approach. The coefficients C are associated with the grid points and used fields that do not depend on the number of analyzed ligand binding modes. Essentially, the maximum numbers of coefficients in multimode CoMFA and in the one-mode approach, either in standard setting or with the weighed fields, are identical. If the same optimization procedure would be used for both one-mode and multimode approaches, the resulting number of the regression coefficients C for the one-mode approaches could possibly be lower because more grid points would be eliminated due to insignificant variation in the field values than in the multimode approach. However, the optimization procedures differ. The forward-selection procedure for the multimode approach naturally minimizes the number of used variables, so the multimode CoMFA actually ends up with much fewer optimized coefficients than traditional approaches. For instance, the presented multimode correlations only contain ~5% of variables included in the one-mode models.

Optimized mode prevalences are calculated as $K_{ij}/K_i = [LR_{ij}]/[LR_i]$ (cf. eq 2), whereby K_{ij} are obtained from eq 3. The mode prevalence distribution for the 16-mode correlation is summarized in Table 3. Even though 16 modes (Figures 1−3) were considered initially, the procedure selected 1−4 modes that were significantly represented (with the prevalences higher than 10%) for each compound.

Most ligands exhibit one binding mode (11 compounds: **1**, **2**, **4**, **6**, **10**, **13**, **15**, **16**, **18**, **21**, and **25,** Table 3) or two binding modes (18 compounds: **3**, **7**, **9**, **11**, **12**, **17**, **19**, **20**, **22**−**24**, **26**, **28**−**30**, and **32-34**). Four congeners (**5**, **8**, **27**, and **31**) bind in three binding modes and one congener (**14**) binds in four binding modes. In some cases (Table 3), two or four equivalent binding modes were observed and counted as one or two modes, respectively. This mode equivalency was caused by symmetrical chlorine positions on the dibenzofuran skeleton. One pair of equivalent modes (for compound **1**, Table 3) or two pairs of equivalent modes (for compounds **7**, **20**, and **34**) were identified.

Relative occupancy of individual modes for the studied data set (the last line in Table 3) can be estimated by summing up prevalences of individual modes for all ligands (Table 3) and dividing them by the maximum occupancy (34, the total number of compounds). The 16 modes can be classified according to increasing relative occupancy into three groups: A, B, C, D, A−L, and B−L (occupancy 0.00); B−T, B−TL, C−T, C−L, C−TL, D−T, and D−TL (0.01−0.10); and A−T, A−TL, and D−L (>0.10). Mode C and its derivatives have the lowest representation in the mode distribution. Only three compounds (**1**, **3,** and **7**, Table 3) bind in the four C-related modes in excess of 10% (combined for equivalent modes).

Individual groups of modes exhibit increasing relative occupancies (RO) in the following order: the skeleton-aligned modes A, B, C, D (the sum of RO = 0.00), the left-aligned modes A−L, B−L, C−L, D−L (0.23), the top-left aligned modes A−TL, B−TL, C−TL, D−TL (0.35), and the top-aligned modes A−T, B−T, C−T, D−T (0.41). The skeleton-aligned modes A−D are practically not represented in the optimized mode distribution. This is an interesting outcome considering that the skeleton-based alignments are most frequently used in 3D-QSAR work. Both previous CoMFA studies[20] of the PCDF binding to the Ah receptor also used the skeleton-based alignments.

The left-aligned modes are present mostly for the D mode. The only compound that shows significant contribution from other left-aligned mode is compound **7**, which is symmetrical and hence the C−L and D−L modes are identical. Interestingly, when a single mode contributes to overall binding of a ligand in excess of 90% it is almost exclusively this left-aligned D mode, except the compounds **1** and **13** (Table 1): the former only shows binding in the top-aligned mode with oxygen in upward position (symmetrical molecule with binding in C−T and D−T mode) and the latter binds only in the A−T mode.

Binding Site Maps. The significant regression coefficients C in respective grid points provide a spatial description of the binding site that corresponds to the differences in interaction fields of ligands between the receptor site and bulk water. The description represents actual shape and properties of the binding site, if the aligned ligands have no common axis of flexibility. Such axes are usually encountered when the analyzed ligands share a common flexible skeleton. In these cases, the shape and properties of the subregions around the flexibility axes are characterized but not their relative positions.

Several regions of the lattice can be distinguished: (1) the *Core* where all the aligned ligand molecules overlap; (2) the *Interface* that contains the quantitative description of binding; and (3) the *Bulk* that is not important for binding. The Core and Bulk are undefined because the energies in their grid points were eliminated from the analysis due to low variance in the interaction energies (the energies are equal to the maximum and approaching zero, respectively). The Interface contains the grid points with optimized regression coefficients that represent the shape and properties of the binding site. The Interface is not defined in parts where all aligned ligands are identical, and consequently, the energies in surrounding grid points are dropped from the analysis due to low variance. The contributions to binding from the identical ligand parts in the Interface are summed up, with other contributions, in the regression coefficient C_0 (eqs 3 and 4). The Interface can also contain grid points with regression coefficients that have optimized values close to zero. These grid points indicate open, *water-filled regions*.

For a good prediction, it is essential that the molecule of a new ligand, as aligned according to specified rules, (1) completely fills the Core, (2) is identical with the training set ligands in the undefined parts of the Interface; and (3) does not extend into the Bulk. Only the last requirement is commonly recognized.[40] Ideally, the software should automatically check whether these conditions are met when making predictions.

The commonly published contour maps only show the grid points with the major electrostatic and steric potentials. This information indicates where and what kind of modifications to make on the lead molecule to achieve better affinity. However, the conventional maps do not show several important features of the binding site model: the Core as well as the undefined parts and water-filled regions of the Interface. Therefore, a different way of pictorial representation of the binding site models was attempted.

The models of the Ah binding site[41] as developed by approaches considering varying number of binding modes are illustrated in Figure 8. The color-coding follows the traditional CoMFA scheme, except the black, gray, cyan, and white colors that indicate the Core, the Interface with undefined and zero coefficients, and the Bulk, respectively. No undefined Interface points were found in the present models. Grid points with electrostatic interactions are shown as concentric cycles, with the outer circle indicating the steric properties. If such point has the zero steric coefficient, the outer circle is cyan to indicate the water-filled area with electrostatic field. The Core and Bulk regions for multimode models were defined using only the significant modes (Table 3).

The main difference between the one-mode and multimode maps is the absence of attractive electrostatic interactions in the lateral areas of the skeleton in the latter. Electrostatic interactions of chlorines are weak due to their low charges. In multimode models, electrostatic interactions are seen around the middle part of the Core, to discriminate the modes based on binding of PCDF oxygens. The Core is significantly smaller for the 16-mode model, where the ligands are allowed to shift in the skeleton plane. In all models, favorable steric regions (green) are mostly localized near the lateral positions on both sides of the skeleton.

Weighed-Field Approach. The one-mode CoMFA approach treats the problem of multiple binding modes differently from the presented multimode procedure. It uses the standard one-mode procedure, whereby the field representing

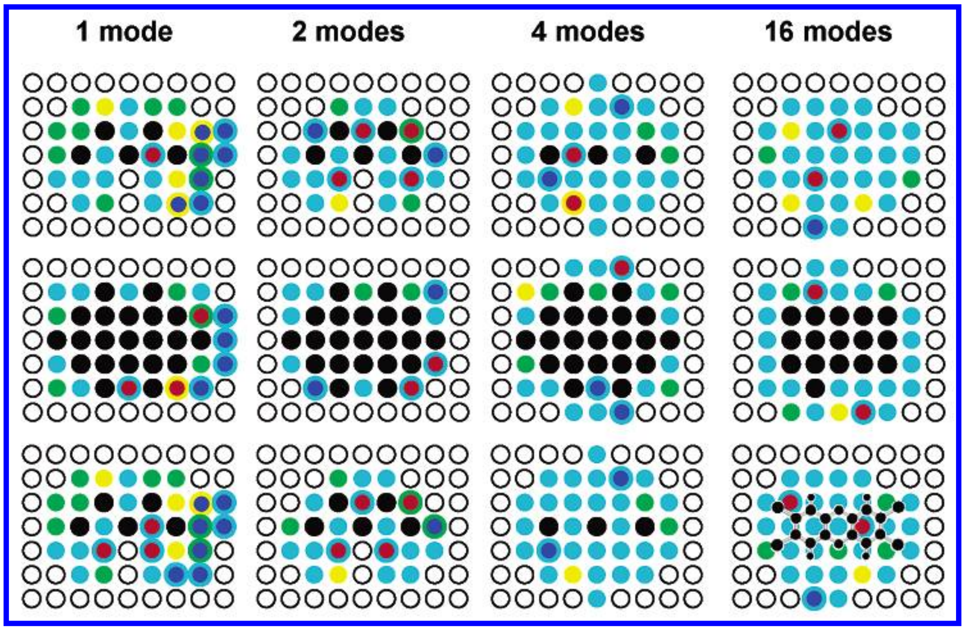

Figure 8. Characterization of the binding site models developed by CoMFA analysis for the CH_3^+ probe. Three middle *xy*-layers of the 9 × 7 × 5 grid are shown with *z* (Å) = −2 (top), 0 (middle), and 2 (bottom); two borderline *xy*-layers (*z* = −4 and 4 Å) are omitted because they were exclusively composed of the bulk points. Color-coding: core − black; interface: steric attractive − green, steric repulsive − yellow, electrostatic attractive − blue, electrostatic repulsive − red, water-filled regions − cyan; bulk − white with black borders. For one-mode analysis, all grid points except the core were included in the model. Here, only 10% of the most significant coefficients are shown. The template TCDD molecule is shown for reference in a plane that did not contain any core (black) points. The planar TCDD molecule is actually positioned in the middle plane (*z* = 0 Å).

the multitude of modes is obtained as a weighed average of the fields of individual modes.[12] The original procedure is almost forgotten, but its reincarnations are occasionally published due to its intuitive appeal. For this reason, we would like to demonstrate that the weighed-field approach lacks the physical basis.

In CoMFA, lnK_{ij} of an individual mode is described as a linear combination of the interaction energies X_{ijk} as seen in eq 3 after logarithmization. In the weighed-field approach, the overall association constant is correlated as lnK_i with the weighed averages of individual fields, i.e., lnK_i is expressed as the weighed average of lnK_{ij}. Taking the antilogarithm, we get

$$K_i = K_{i1}^{p_1} \times K_{i2}^{p_2} \times ...K_{im}^{\;p_m} \qquad (8)$$

where *p* are the prevalences (weights) of individual modes. Equation 8 is in contradiction with kinetic and thermodynamic analyses[21−25] showing that the overall association constant is the sum of the partial association constants as given in eq 2. Moreover, the prevalences (equal to K_{ij}/K_i) cannot be estimated a priori, because they depend on the regression coefficients C_k in a nonlinear way as shown in eqs 3 and 4.

The weighed-field approach was examined by running the analysis for all 31 PCDFs with precise pE_{50} values (Table 1). The weighed field values were calculated for the mode prevalences generated by the four-mode approach. The weighed-fields approach (Table 2), when compared to the results of the one-mode approach obtained from full data set (data not shown), exhibits better fit (SSE, R^2) but is weaker in predictions as indicated by q^2 and prediction of affinities for the test set compounds **1, 14,** and **15** (structures in Table 1) with semiquantitative pEC_{50}. Moreover, the input prevalences of the binding modes as provided by the four-mode approach are not reproduced when calculated back after the analysis using eq 3 with optimized regression coefficients *C* (the correlation between the input and output prevalences has R^2 = 0.036).

CONCLUSIONS AND OUTLOOK

The presented multimode CoMFA procedure provides statistically better descriptions and predictions than both one-mode and weighed-field approaches. The better predictive ability is not a consequence of increased flexibility of the model due to a higher number of optimized regression coefficients. Just the opposite is true: the used forward-selection approach reduces the number of variables about 20 times as compared with the current CoMFA approaches.

The prevalences of individual binding modes depend on the regression coefficients *C* and, consequently, are optimized by the multimode procedure. We demonstrated that, in the studied data set, the procedure effectively selects one to four binding modes out of the ensemble of 16 modes for each ligand. The feature makes multimode CoMFA a useful tool for objective selection of binding modes. This ability alleviates one of the CoMFA bottlenecks—the alignment problem,[42] significantly reduces the subjective input into the alignment procedure and, thus, promotes future automatization of the CoMFA analyses. The procedure can be recommended for cases where an exhaustive and objective evaluation of multimode binding is desired.

ACKNOWLEDGMENT

Although the research described in this article has been funded in parts by the U.S. EPA (R82-6652-011) and NIH NCRR (1 P20 RR15566, 1 PP20 RR 16471), it has not been subjected to the agencies' peer and policy reviews and therefore does not necessarily reflect the views of the agencies and no official endorsement should be inferred.

Supporting Information Available: The Tripos mol2 file containing optimized geometries and calculated charges of the PCDF congeners, aligned in the used superposition of 16 binding modes and a table summarizing all resulting binding site models by classification of individual grid points as belonging to the Core (marked as c), Interface (marked by the regression coefficients), and Bulk (no mark). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Mattos, C.; Ringe, D. Multiple binding modes. In *3D-QSAR in Drug Design: Theory, Methods, and Applications*; Kubinyi, H., Ed.; Escom: Leiden, 1993; pp 226−254. Balaz, S.; Hornak, V. Multiple binding modes in three-dimensional quantitative structure−activity relationships. In *QSAR in Environmental Sciences*; Walker, J., Ed.; Society for Environmental Toxicology and Chemistry: Pensacola (in press). de la Paz, P.; Burridge, J. M.; Oatley, S. J.; Blake, C. C. F. Multiple modes of binding of thyroid hormones and other iodothyronines to human plasma transthyretin. In *The Design of Drugs to Macromolecular Targets*; Bedell, C. R., Ed.; John Wiley and Sons: Chichester, 1992; pp 119−172. Arevalo, J. H.; Hassig, C. A.; Stura, E. A.; Sims, M. J.; Taussig, M. J.; Wilson, I. A. Structural analysis of antibody specificity. Detailed comparison of five Fab'-steroid complexes. *J. Mol. Biol.* **1994**, *241*, 663−690.

(2) Mattos, C.; Rasmussen, B.; Ding, X.; Petsko, G. A.; Ringe, D. Analogous inhibitors of elastase do not always bind analogously. *Nat. Struct. Biol.* **1994**, *1*, 55−58.

(3) Lemmen, C.; Zimmermann, M.; Lengauer, T. Multiple molecular superpositioning as an effective tool for virtual database screening. *Persp. Drug Discov. Des.* **2000**, *20*, 43−62.

(4) Diana, G. D.; Kowalczyk, P.; Treasurywala, R. C. O.; Pevear, D. C.; Dutko, F. J. CoMFA analysis of the interactions of antipicornavirus compounds in the binding pocket of human rhinovirus-14. *J. Med. Chem.* **1992**, *35*, 1002−1008.

(5) Oprea, T. I.; Waller, C. L.; Marshall, G. R. Three-dimensional quantitative structure−activity relationship of Human Immunodeficiency Virus (I) protease inhibitors. 2. Predictive power using limited exploration of alternate binding modes. *J. Med. Chem.* **1994**, *37*, 2206−2215.

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

(7) Klebe, G.; Abraham, U. On the prediction of binding properties of drug molecules by Comparative Molecular Field Analysis. *J. Med. Chem.* **1993**, *36*, 70−80.

(8) Crippen, G. M. Intervals and the deduction of drug binding site models. *J. Comput. Chem.* **1995**, *16*, 486-500. Jain, A. N.; Koile, K.; Chapman, D. Compass: Predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.* **1994**, *37*, 2315−2327.

(9) To illustrate the magnitude of this number, let us consider a small set of 30 ligands, each binding in two modes. A systematic evaluation requires 2^{30} one-mode 3D-QSAR analyses, which would be done in about 34 years by a fast method consuming only 1 s per analysis.

(10) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(11) Kim, K. H.; Martin, Y. C. Direct prediction of dissociation constants (pK_a's) of clonidine-like imidazolines, 2-substituted imidazoles, and 1-methyl-2-substituted imidazoles from 3D-structures using a Comparative Molecular Field Analysis (CoMFA) approach. *J. Med. Chem.* **1991**, *34*, 2056−2060.

(12) Cramer, R. D., III; Wold, S. Comparative Molecular Field Analysis (CoMFA). U.S. Patent 5,307,287, 1994.

(13) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B. Q.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.

(14) Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure−activity relationships and quantitative structure−property relationship. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.

(15) Vedani, A.; McMasters, D. R.; Dobler, M. Genetic algorithms in 3D-QSAR: The use of multiple ligand orientations for improved predictions of toxicity. *Altex* **1999**, *16*, 142−145. Vedani, A.; McMasters, D. R.; Dobler, M. Multi-conformational ligand representation in 4D-QSAR: Reducing the bias associated with ligand alignment. *Quant. Struct. Act. Relat.* **2000**, *19*, 149−161.

(16) Vedani, A.; Briem, K.; Dobler, M.; Dollinger, H.; McMasters, D. R. Multiple-conformation and protonation-state representation in 4D-QSAR: The neurokinin-1 receptor system. *J. Med. Chem.* **2000**, *43*, 4416−4427.

(17) Safe, S.; Sawyer, T.; Mason, G.; Bandiera, S.; Keys, B.; Romkes, M.; Piskorska-Pliszczynska, J.; Zmudzka, B.; Safe, L. Polychlorinated dibenzofurans: Quantitative structure activity relationships. *Chemosphere* **1985**, *14*, 675−683.

(18) Mason, G.; Sawyer, T.; Keys, B.; Bandiera, S.; Romkes, M.; Piskorska-Pliszczynska, J.; Zmudzka, B.; Safe, S. Polychlorinated dibenzofurans (PCDFs): Correlation between in vivo and in vitro structure−activity relationships. *Toxicology* **1985**, *37*, 1−12. Safe, S.; Sawyer, T.; Robertson, L.; Safe, L.; Parkinson, A.; Thomas, P. E.; Ryan, D. E.; Reik, L. M.; Levin, W. PCBs: Structure−function relationships and mechanism of action. *Environ. Health Perspect.* **1985**, *60*, 47−56. 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*, 51−88. Safe, S. H. Comparative toxicology and mechanism of action of polychlorinated dibenzo-p-dioxins and dibenzofurans. *Annu. Rev. Pharmacol. Toxicol.* **1986**, *26*, 371−399.

(19) Long, G.; McKinney, J.; Pedersen, L. Polychlorinated dibenzofuran (PCDF) binding to the Ah receptor(s) and associated enzyme induction. Theoretical model based on molecular parameters. *Quant. Struct. Act. Relat.* **1987**, *6*, 1−7. Sulea, T.; Kurunczi, L.; Simon, Z. Dioxin-type activity for polyhalogenated arylic derivatives. A QSAR model based on MTD-method. *SAR QSAR Environ. Sci.* **1995**, *3*, 37−61. Vedani, A.; McMasters, D. R.; Dobler, M. Genetic algorithms in 3D-QSAR: Predicting the toxicity of dibenzodioxins, dibenzofurans and biphenyls. *Altex* **1999**, *16*, 9−14. Mekenyan, 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. Todeschini, R.; Gramatica, P. 3D-modelling and prediction by WHIM descriptors. 6. Application of WHIM descriptors in QSAR studies. *Quant. Struct. Act. Relat.* **1997**, *16*, 120−125.

(20) Waller, C. L.; McKinney, J. D. Three-dimensional quantitative structure−activity relationships of dioxins and dioxin-like compounds: Model validation and Ah Receptor characterization. *Chem. Res. Toxicol.* **1995**, *8*, 847−858. Waller, C. L.; McKinney, J. D. Comparative Molecular Field Analysis of polyhalogenated dibenzo-p-dioxins, dibenzofurans, and biphenyls. *J. Med. Chem.* **1992**, *35*, 3660−3666.

(21) Balaz, S.; Hornak, V.; Haluska, L. receptor mapping with multiple binding modes: binding site of PCB-degrading dioxygenase. *Chemom. Intell. Lab. Sys.* **1994**, *24*, 185−191.

(22) Wang, J.; Szewczuk, Z.; Yue, S. Y.; Tsuda, Y.; Konishi, Y.; Purisima, E. O. Calculation of relative binding free energies and configurational entropies: A structural and thermodynamic analysis of the nature of nonpolar binding of thrombin inhibitors based on hirudin55−65. *J. Mol. Biol.* **1995**, *253*, 473−492.

(23) Hornak, V.; Balaz, S.; Schaper, K. J.; Seydel, J. K. Multiple binding modes in 3D-QSAR: Microbial degradation of polychlorinated biphenyls. *Quant. Struct. Act. Relat.* **1998**, *17*, 427−436.

(24) Smith, W. R.; Missen, R. W. *Chemical Reaction Equilibrium Analysis: Theory and Algorithms*; John Wiley and Sons: New York, 1982.

(25) Jullien, L.; Proust, A.; LeMenn, J. C. How does the Gibbs free energy evolve in a system undergoing coupled competitive reactions? *J. Chem. Educ.* **1998**, *75*, 194−199.

(26) In our application example, however, no conformational energy term was needed because the studied compounds do not have rotatable bonds.

(27) Balaz, S.; Lukacova, V. Method for drug design using Comparative Molecular Field Analysis extended for multi-mode ligand binding and disposition. U.S. Patent, 2001 (pending).

(28) Otherwise, for a small number of *C*, no linearization would be necessary and nonlinear regression analysis could be adopted using directly eq 4.

(29) Henry, E. C.; Gasiewicz, T. A. Transformation of the aryl hydrocarbon receptor to a DNA-binding form is accompanied by release of the 90 kDa heat-shock protein and increased affinity for 2,3,7,8-tetrachlorodibenzo-p-dioxin. *Biochem. J.* **1993**, *294*, 95−101.

(30) *Sybyl*, version 6.9. Tripos Inc., St. Louis, MO, 2002.

(31) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. Jr. General atomic and molecular electronic structure system. *J. Comput. Chem.* **1993**, *14*, 1347−1363.

(32) Rabinowitz, J. R.; Little, S. B.; Gifford, E. M. Interactions between chlorinated dioxins and a positively charged molecular probe: new molecular interaction potential. *J. Comput. Chem.* **1998**, *19*, 673−684.

(33) Leon, L. A.; Notario, R.; Quijano, J.; Sanchez, C. Structures and enthalpies of formation in the gas phase of the most toxic polychlorinated dibenzo-p-dioxins. A DFT study. *J. Phys. Chem. A* **2002**, *106*, 6618−6627.

(34) Pastor, M.; Cruciani, G.; Clementi, S. Smart Region Definition: A new way to improve the predictive ability and interpretability of three-dimensional quantitative structure−activity relationships. *J. Med. Chem.* **1997**, *40*, 1455−1464. *GOLPE*, version 4.5. Multivariate Infometric Analysis Srl., Perugia, Italy, 1999.

(35) The number of included variables or the number of compounds in the set, whichever is smaller.

(36) Press: W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes. The Art of Scientific Computing*; Cambridge University Press: Cambridge, 1986.

(37) The leave-one-out *q* values for the training set are not used to assess predictive abilities of the models. Predictions for the test set that did not come into contact with calibration are used for this purpose.

(38) Standard deviation SD must exceed certain limit. Variability must not be caused by one or two extreme X_{ijk} values that substantially differ from the rest (tested as a difference between mean and median values).

(39) Due to the duration of the optimization procedure, predictive ability was examined using a single test set, rather than a series of smaller test sets.

(40) Thibaut, U.; Folkers, G.; Klebe, G.; Kubinyi, H.; Merz, A.; Rognan, D. Recommendations for CoMFA studies and 3D QSAR publications. In *3D QSAR in Drug Design, Vol. 1*; Kubinyi, H., Ed.; Escom: Leiden, 1993; pp 711−716.

(41) Since all ligands are planar and aligned in the skeleton plane in all modes, the resulting fields exhibit symmetry about this plane. The analyzed ligand clusters are symmetrical about the *y*-axis (two modes) and both x- and *y*-axes (four and, partially, 16 modes). Consequently, the multimode procedure produces equivalent maps, two and four for the two-mode and the four-mode/16-mode approaches, respectively, all rotated by 180° about the corresponding symmetry axes. Each multimode approach arrives at one of the maps depending upon the initial estimates and other details in the optimization procedure.

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

CI034100A