JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Use of Quantitative Structural Analysis To Predict Fish Bioconcentration Factors for Pesticides

SCOTT H. JACKSON,*,† CHRISTINA E. COWAN-ELLSBERRY,§ AND GARETH THOMAS#

BASF Corporation, 26 Davis Drive, Research Triangle Park, North Carolina 27709; The Procter & Gamble Company, 11810 East Miami River Rosd, Cincinnati, Ohio 45252; and Centre for Chemical Management, Lancaster Environment Centre, Lancaster, United Kingdom LA1 4YQ

The focus of this research was to develop a model based solely on molecular descriptors capable of predicting fish bioconcentration factors (BCF). A fish BCF database was developed from high-quality, regulatory agency reviewed studies for pesticides based on the same laboratory protocol and the same fish species, *Lepomis macrochirus*. A commercially available software program was used to create a quantitative structure−activity relationship (QSAR) from 93 BCF studies based on unique molecules. An additional 16 molecules were used to test the accuracy of QSAR model predictions for a variety of pesticide classes. Regression of the measured versus predicted log BCF values yielded a regression coefficient of 0.88 for the validation data set. On the basis of the results from this research, the ability to predict BCF by a QSAR regression model is improved using a fully structurally derived model based solely on structural data such as the number of atoms for a given group (e.g., −CH3) or the local topology of each atom as derived from electron counts. Such descriptors provide insightful information on a molecule's potential BCF behavior in aquatic systems.

KEYWORDS: Fish; bioaccumulation; BCF; QSAR; structural analysis; pesticide

INTRODUCTION

Bioconcentration factor (BCF) data are currently needed for many chemical regulatory programs. New laws resulting from enactment of the United Nations Stockholm Convention Persistent Organic Pollutants Protocol (POPs) in May 2004 have led to significant new activity in the assessment of persistent, bioaccumulative, toxic substances (PBTs). Several committees, notably the Health and Environmental Sciences Institute (HESI) Bioaccumulation Assessment Project Committee, active in this work have explored the challenges involved with accurately assessing new and existing chemicals for their bioaccumulation potential. Before 2003, common methods for estimating or measuring BCF values were limited to regression models that often used the octanol−water partition coefficient (log K_{ow}) as the only input or expensive OECD 305 standard fish BCF guideline studies (*1*).

Some of the earliest attempts at BCF prediction have been simple regression relationships based on physically measured properties, which are not truly structural in nature. Therefore, one of the foci of the HESI committee was to better define the physical and structural factors that affect bioavailability, absorption, or metabolism of a chemical to more accurately predict bioconcentration in fish. A review paper by Barber (*2*) focused on fish bioconcentration models summarizes the many efforts to estimate BCF using simple regression relationships. Although the various methods described (*2*) are listed as being QSAR-based, none actually were structure−activity relationships unless adjustment factors for certain chemical moieties can be considered structural. Because current regression approaches for estimating BCF values inadequately characterize the potential of some chemical classes to bioaccumulate, many compounds could be misclassified as bioaccumulative using these methods if they are integrated into regulatory or product development evaluations. Therefore, it is important to better identify and characterize the dominant molecular descriptors driving bioconcentration and bioaccumulation. One of the goals for this project was to develop a QSAR model that could be used in the evaluation and screening of compounds molecularly similar to those pesticides used to develop the model in this work. The new model could then be used by industry or government agencies developing and reviewing new and existing chemicals. Predicting the BCF of a substance is a very important factor for deciding its safety in the environment. The earlier in the development process properties such as BCF are known, the sooner proactive decisions can be made about the use, development, and manufacture of a chemical. The goal of this work was to develop a fully structural model (without laboratory measurements) that will allow predictions of BCF.

† BASF Corporation.
§ The Procter & Gamble Company.
# Lancaster Environment Centre.

10.1021/jf803064z CCC: $40.75
Published on Web 01/12/2009

METHOD AND MATERIALS

To evaluate and develop structural relationships to predict BCF, a useful data set was required. Some of the desirable attributes of a data set were that the BCF factors were measured on the same test species following a standardized protocol. A data set that appeared to be an excellent choice for our purposes was the one developed for the support of pesticide registrations in the United States. The U.S. Environmental Protection Agency (U.S. EPA) Environmental Fate and Effects Division (EFED) has developed an online database that reports many results from regulatory environmental fate studies (*3*). The database contains detailed compound information for herbicides, insecticides, and fungicides including descriptive BCF data. This database was determined to be of great value for the development and evaluation of a structural BCF model because it contains high-quality data generated by following the same protocol and using the same test species.

The BCF studies were all conducted following EPA guideline 165-4, *Laboratory Studies of Pesticide Accumulation in Fish* (now OPPTS 850.1730) (*4*). The studies were typically conducted for 28 days unless a concentration plateau was reached earlier, which was an indication the study could be concluded sooner. Bluegill (*Lepomis macrochirus*) are the preferred species for the guideline studies; however, not all studies used bluegill fish. When a study did not use bluegill as the test species, it was not included in this work. The guideline protocol requires that the fish receive chemical exposure using a radiolabeled test substance under flow-through tank conditions. Because the entire tank volume is replaced many times daily, the animals are principally exposed to parent compound only. For each residue determination interval, the fish are separated into edible and viscera portions for determination of BCF. Whole fish BCF values are typically determined as well and were the end point used for this model development activity. The EPA database consisted of many compounds with BCF data potentially for this analysis. However, the subset we used for this work consists of 93 molecules (**Table 1**) with log BCF values ranging from −0.92 to 4.0. The criteria for the compounds used in our subset were (A) whole fish BCF values were measured, (B) bluegill was the test species, and (C) the test was conducted for about 28 days. The BCF values in the data set are plotted in **Figure 1** as an indication that the molecules selected were well distributed throughout the stated range.

To set a background for this work, we examined one of the standard methods commonly used to estimate fish BCF values, the Veith(*5*) regression (Log BCF = 0.76 log K_{ow} − 0.23). To evaluate the effectiveness of regression methods for predicting the fish BCF values for this data set, we chose the Veith bluegill regression to be representative of the effectiveness of similar BCF regression methods in general. The only input to this type of BCF regression is the compound-specific log K_{ow} value (Log *P*). For the evaluation of the Veith equation to occur, log K_{ow} values had to be obtained. The measured Log K_{ow} value for each compound (**Table 1**) was obtained from the FOOTPRINT database housed at the University of Hertfordshire (*6*). Using the measured log K_{ow} values in the Veith equation, an r^2 of 0.50 was achieved between the predicted and measured BCF values (**Figure 2**). Alternatively, a calculated log K_{ow} value was calculated for each molecule using MDL's QSAR (*7*) software (ver. 2.3) (**Table 1**) and was substituted in the Veith equation to determine if a calculated log K_{ow} enhanced BCF predictions. An r^2 of 0.48 was achieved using the calculated log K_{ow} values rather than the measured values (**Figure 3**). Statistical regression analysis of the predicted versus measured BCF values was performed using Origin software (*8*).

Results from the analysis presented in **Figures 2** and **3** indicate that simple regressions based on log K_{ow} are not capable of predicting BCF in a reliable fashion for the current generation of agricultural chemicals.

Because standard regression approaches based on measured or predicted log K_{ow} alone are inadequate for recently developed compounds, a more advanced method was explored for the prediction of fish BCF values. The project objective was to develop an improved BCF prediction model that would be a fully structural model based only on structural characteristics of the compound and would not require laboratory measurements to predict fish BCF values in **Table 1**. Although log *P* values will be included in the QSAR regression, the log *P* will also be a calculated value from the chemical structure.

The first step in performing the QSAR analysis was to obtain and enter structures for each molecule in the database we used for this evaluation. Once the structures were obtained, they were converted into SMILES codes. For this analysis, the structural SMILES code for each molecule was obtained directly from the National Center for Biotechnology Information Website (*9*). The SMILES codes were imported into MDL's QSAR (*7*) software (ver. 2.3). The QSAR software has a database information system capable of creating quantitative structure−activity relationships (QSAR) by calculating over 300 structural descriptors. The software has various multiple-regression analysis routines, can use a genetic algorithm, and is capable of performing principal component analysis as well.

One of the challenges in developing a structurally based QSAR is the selection of descriptors to use in the QSAR model beyond the calculated log *P* value. The selection of molecular descriptors can be divided into several distinct categories. The two broadest descriptor categories are two- and three-dimensional molecular property descriptors. The two-dimensional descriptors can be further broken down into five subcategories: simple connectivity descriptors, connectivity valences, molecular E-state's, kappa shape indices, and general molecular properties. The three-dimensional descriptors are limited to molecular measurements, such as dipole moment and shape/dimension features.

Model Development and Validation. In addition to calculated log *P*, which was always included, more than 250 molecular descriptors were added and removed in an iterative process until the optimum QSAR descriptors were obtained. In the estimation of fish BCF values, it was assumed that three-dimensional descriptors would be a dominating factor in developing a useful model because active interaction with membranes might dominate the process (*10*). However, in practice, structural molecule descriptors were added and removed from model development in an iterative process until a model optimum was achieved.

The model development process consisted of first identifying the best QSAR descriptors, and then several statistical regression methods were evaluated to determine if an optimized regression model could be achieved of the descriptors and the measured BCF values. A variety of regression routines were used including ordinary multiple regression, stepwise regression, all possible subsets regression, regression on principle components, and partial least-squares regression to develop the best predictive relationship. The regression methods were systematically tried until a best fit was obtained on the basis of the resulting model R^2 value and analysis of residuals. After the best regression model was determined, the regression was tested with a smaller data set of 16 compounds, the regression model validation set. The regression model validation set was not used in the initial development of the model (**Table 2**) but was only used to validate the new QSAR regression model. The process consisted of reading the validation set into the QSAR software, generating the required molecular descriptors, and predicting the BCF values. Regression analysis of the predicted versus measured BCF values was performed outside the QSAR software (*8*). For model validation, there are two methods typically used (*11*). The first method is referred to as cross-validation. In the cross-validation method, the molecules used to develop the model are rotated in and out of the regression in an attempt to validate the regression. The cross-validation approach is not considered to be robust, but it often gives an indication of the model's potential for prediction. The second method used is referred to as external-validation, in which a second independent data set is used in an attempt to validate the model. The external-validation method is considered to be more robust that the cross-validation method and is the method we used in this paper with the data set of 16 compounds (*11*).

RESULTS AND DISCUSSION

Model Development. The data used for the model development consisted of 93 molecules with log BCF values ranging from −0.92 to 4.0, which were well distributed throughout that range (**Table 1** and **Figure 1**). For development of this model, based on the descriptors selected, the ordinary multiple-

Table 1

no.	common name	meas Log BCF	pred Log BCF	Log *P* (meas)	Log *P* (calcd)	ABSQon	MaxNeg	xch10	xvp10	xvch9	SaasCacnt	SssNHacnt	SssOacnt	SdssSacnt
1	bensulide	2.9445	2.6061	4.2	3.6738	1.1729	−0.25578	0	0.63052	0	1	1	2	0
2	bentazon	0.5682	−0.13366	−0.46	1.6995	1.3332	−0.36307	0.0098209	0.035747	0	0	0	0	0
3	bifenthrin	3.7846	3.2773	5.0	6.1795	0.61167	−0.32958	0	0.1019	0	3	0	1	0
4	bitertanol	2.2405	1.8476	4.1	3.6582	1.1454	−0.3886	0	0.088576	0	3	0	1	0
5	bromacil	0.34242	0.89464	1.88	1.4754	1.1509	−0.40702	0	0	0	0	1	0	0
6	butralin	3.29	2.8596	4.93	3.5393	0.84592	−0.34031	0	0.011659	0	4	1	0	0
7	butylate	2.6128	1.6882	4.1	3.742	0.66702	−0.40904	0	0	0	0	0	0	0
8	imazapic	−0.92082	−0.88215	2.47	−0.10454	2.2219	−0.39143	0	0.040146	0	3	1	0	0
9	captan	2.1004	1.9538	2.5	2.2472	1.0193	−0.38963	0	0.11733	0.0062113	0	0	0	0
10	chlorothalonil	2.4216	2.7235	2.94	4.7878	0.40191	−0.20495	0	0	0	6	0	0	0
11	chlorpyrifos	3.4357	3.4095	4.7	4.757	2.0594	−2.9329	0	0.028931	0	4	0	3	0
12	clethodim	0.5563	1.8453	4.2	4.8713	1.2781	−0.40618	0	0.15118	0	0	1	1	0
13	clomazone	1.6021	1.9576	2.54	2.3877	0.78378	−0.39153	0	0.037692	0	2	0	1	0
14	coumaphos	2.7332	2.241	3.86	4.2432	1.2857	−0.38883	0.0092593	0.15153	0	3	0	4	0
15	cyclanilide	0.11394	1.0747	3.25	1.9562	1.3009	−0.42287	0	0.033976	0	3	1	0	0
16	cyclopropyl	2.2041	2.3696	na	2.8845	1.0916	−0.30701	0	0.018298	0	3	2	0	0
17	cyhalothrin	1.2788	2.8029	6.8	5.3743	1.1497	−0.35526	0	0.096121	0	3	0	2	0
18	cypermethrin	2.6702	2.883	5.3	5.722	1.1496	−0.35526	0	0.092871	0	3	0	2	0
19	cyproconazole	1.5315	1.6818	3.09	2.9741	0.80429	−0.38368	0	0.036266	0	2	0	0	0
20	cyprodinil	2.7086	2.1492	4.0	2.6796	0.80073	−0.27258	0	0.042114	0	4	1	0	0
21	DCPA, or dacthal	3.2774	2.3273	5.0	3.8022	1.3151	−0.38648	0	0	0	6	0	2	0
22	deltamethrin	2.8439	2.9249	4.6	5.8178	1.1497	−0.35526	0	0.10627	0	3	0	2	0
23	desmedipham	1.9912	1.7881	3.39	3.1601	1.7697	−0.39367	0	0.034213	0	3	2	2	0
24	dichlobenil	1.6021	2.291	2.7	2.8863	0.22002	−0.22002	0	0	0	3	0	0	0
25	dicofol	4	3.5842	4.3	6.3883	0.37389	−0.37389	0	0.049931	0	4	0	0	0
26	difenoconazole	2.5185	2.7261	4.2	3.9049	1.4572	−0.35523	0	0.11266	0	5	0	3	0
27	dimethenamid	1.7709	1.7511	1.89	2.7954	1.0228	−0.4118	0	0.013176	0	3	0	1	0
28	diquat dibromide	0.012837	−0.061715	−4.6	2.307	0.66233	−0.34899	0.027778	0.1438	0	0	0	0	0
29	disulfoton	2.6628	2.2605	3.95	3.1288	0.47007	−0.23504	0	0	0	0	0	2	0
30	endosulfan	3.4401	3.3222	3.13	3.6498	0.76462	−0.2633	0.0060141	0.26972	0	0	0	2	1
31	endothall	−0.45593	0.096311	1.91	−0.96127	1.6319	−0.37025	0	0	0	0	0	1	0
32	esfenvalerate	3.3784	2.9834	6.24	5.4222	1.1491	−0.35526	0	0.084764	0	5	0	2	0
33	ethalfluralin	3.0719	3.3432	5.11	4.524	0.83082	−0.32521	0	0.012591	0	4	0	0	0
34	etridiazole	2.2856	2.1672	3.37	3.0994	0.6772	−0.30593	0	0	0	2	0	1	0
35	fenamiphos	1.9494	2.5294	3.3	3.0085	0.70477	−0.2993	0	0.13292	0	3	1	2	0
36	fenarimol	2.0531	1.9969	3.69	3.7375	0.90855	−0.36848	0	0.061624	0	5	0	0	0
37	fenbuconazole	2.2304	1.8547	3.22	2.9923	0.62663	−0.23874	0	0.12273	0	3	0	0	0
38	fenitrothion	2.1106	2.2349	3.27	2.5925	0.99324	−0.26376	0	0	0	3	0	3	0
39	fenoxaprop-ethyl	2.7218	2.6447	4.28	4.031	1.6411	−0.32699	0	0.062743	0.002196	4	0	3	0
40	fenoxycarb	2.5038	2.2952	4.07	3.6847	1.5924	−0.37784	0	0.036129	0	2	1	3	0
41	fenpropathrin	3.0414	2.9266	6.04	5.9826	1.1497	−0.35526	0	0.078725	0	3	0	2	0
42	fipronil	2.5065	2.4606	3.75	2.3511	1.1008	−0.39703	0	0.079065	0	7	0	0	1
43	flumiclorac-pentyl	2.6335	2.692	4.99	4.5092	1.9398	−0.35581	0	0.16262	0.0069877	4	0	2	0
44	fluridone	1.6684	1.8718	1.87	3.6041	0.73747	−0.46391	0	0.067426	0	2	0	0	0
45	fluroxypyr-MHE	1.7932	1.3618	2.0	−0.55037	1.4883	−0.32454	0	0.0019567	0	5	0	1	0
46	flurprimidol	1.5453	1.6248	3.34	3.6509	1.2151	−0.37328	0	0.02034	0	3	0	1	0
47	flutolanil	2.1761	2.3267	3.17	4.0901	0.99217	−0.40389	0	0.047226	0	4	1	1	0
48	fomesafen	0.77815	1.4024	−1.2	3.473	1.5344	−0.33606	0	0.08244	0	6	1	1	0
49	fosthiazate	0.50515	0.93742	1.68	2.2317	0.76851	−0.40455	0	0.0375	0	0	0	1	0
50	glyphosate	−0.284	−0.35399	−3.2	−3.4102	1.4906	−0.35636	0	0	0	0	1	0	0
51	halofenozide	0.86332	1.6869	3.3	3.1488	1.1038	−0.4122	0	0.034878	0	3	1	0	0
52	hexythiazox	3.2041	2.2291	2.67	3.3477	1.1911	−0.39459	0	0.23605	0	2	1	0	0
53	hydramethylnon	3.1139	2.872	2.31	5.0944	0.99757	−0.31179	0	0.10386	0	4	2	0	0
54	imazamethabenz	0.4624	0.69518	1.78	2.1604	1.6437	−0.40183	0	0.040584	0	3	1	0	0
55	iprodione	1.8573	0.7199	3.1	2.6815	1.6174	−0.39832	0	0.05308	0	3	1	0	0

Table 1. Continued

no.	common name	meas Log BCF	pred Log BCF	Log *P* (meas)	Log *P* (calcd)	ABSQon	MaxNeg	xch10	xvp10	xvch9	SaasCacnt	SssNHacnt	SssOacnt	SdssSacnt
56	isoxaben	1.8451	2.215	3.94	4.0747	1.6489	−0.39835	0	0.058644	0	5	1	2	0
57	kresoxim-methyl	2.3424	2.4482	3.4	3.7566	1.5123	−0.37241	0	0.056481	0	4	0	3	0
58	lactofen	2.5798	2.5322	5.0	3.6983	1.7258	−0.32697	0	0.067893	0	6	0	3	0
59	L-cyhalothrin	3.3502	2.8029	6.8	5.3743	1.1497	−0.35526	0	0.096121	0	3	0	2	0
60	mepiquat chloride	0.30103	0.30104	−3.45	3.1766	0	−0.12421	0	0	0	0	0	0	0
61	metalaxyl	0.8451	1.2646	1.65	2.3374	1.6104	−0.40999	0	0.013819	0	3	0	2	0
62	methyl parathion	1.8513	2.1716	3.0	2.253	0.92075	−0.25281	0	0	0	2	0	3	0
63	metolachlor (stereoisomer)	1.8388	1.8675	3.4	3.1827	1.0241	−0.41892	0	0.02575	0	3	0	1	0
64	molinate	1.8573	1.4445	2.86	2.4656	0.66751	−0.40904	0	0.048412	0	0	0	0	0
65	norflurazon	1.4472	1.1883	2.45	2.8868	0.99792	−0.39997	0	0.038517	0	2	1	0	0
66	oryzalin	1.8797	2.2182	3.73	2.0778	1.7651	−0.34001	0	0.015462	0	4	0	0	0
67	pendimethalin	3.7076	2.8821	5.2	2.9122	0.84566	−0.34005	0	0.012224	0	5	1	0	0
68	phorate	2.6839	2.3327	3.86	3.4222	0.47007	−0.23504	0	0	0	0	0	2	0
69	phostebupirim	2.8573	2.2742	na	4.7655	1.1436	−0.26385	0	0.094496	0	2	0	3	0
70	prallethrin	3.0645	2.0552	4.49	4.735	1.0235	−0.41924	0	0.086408	0	0	0	1	0
71	prodiamine	3.1139	3.1033	4.1	3.7514	1.161	−0.3289	0	0.013391	0	5	0	0	0
72	prometryn	1.9294	2.2069	3.3	3.2598	1.0946	−0.30943	0	0.0083333	0	3	2	0	0
73	propachlor	1.5682	1.4523	1.6	2.2413	0.64948	−0.41857	0	0	0	1	0	0	0
74	propiconazole	2.0645	2.357	3.72	3.2841	1.1017	−0.34205	0	0.094823	0	3	0	2	0
75	pyridate	2.6665	2.77	0.5	5.8462	1.0009	−0.33562	0	0.17307	0	4	0	1	0
76	quinclorac	−0.38722	0.86763	−1.15	3.1023	1.0429	−0.39895	0.01134	0.014704	0	3	0	0	0
77	sethoxydim	1.3522	1.7863	1.65	4.0563	1.1634	−0.36067	0	0.21264	0	0	0	1	0
78	starlicide	1.9445	2.1256	na	2.4989	0.32995	−0.32995	0	0	0	3	0	0	0
79	tebuconazole	1.9956	1.6297	3.7	3.3698	0.80865	−0.38804	0	0.068188	0	2	0	0	0
80	tebufenozide	2.179	2.0637	4.25	4.2097	1.13	−0.42986	0	0.065333	0	5	1	0	0
81	tebuthiuron	0.41996	1.1248	1.79	2.1578	1.1578	−0.39095	0	0	0	2	1	0	0
82	temephos	3.3617	3.6928	4.95	5.7814	1.3359	−0.23834	0	0.35055	0	4	0	6	0
83	terbacil	0.77815	0.82132	1.89	1.377	1.1157	−0.40209	0	0	0	0	1	0	0
84	thiazopyr	2.415	2.0266	3.89	5.0907	1.3606	−0.40296	0	0.072479	0	5	0	1	0
85	thifensulfuron-methyl	−0.045757	−0.72743	−1.7	0.90575	2.7685	−0.3969	0	0.10049	0	5	2	2	0
86	thiobencarb	2.6138	1.8824	4.23	3.2642	0.66785	−0.407	0	0.093322	0	2	0	0	0
87	thiodicarb	0.75587	0.90617	1.62	2.6395	1.6856	−0.36869	0	0.016232	0	0	0	2	0
88	triadimenol	1.4314	1.4774	3.2	2.8793	1.1454	−0.3886	0	0.02466	0	2	0	1	0
89	triasulfuron	0.079181	0.11623	0.96	1.5412	2.42	−0.37428	0	0.087849	0	5	2	2	0
90	tribuphos	2.8633	2.964	5.52	3.8586	0	−0.093123	0	0.75467	0	0	0	0	0
91	tridiphane	3.0745	3.2964	4.3	5.063	0.35899	−0.35899	0	0	0	3	0	1	0
92	trifloxystrobin	2.734	2.4913	4.5	4.3748	1.569	−0.36597	0	0.049024	0	4	0	3	0
93	triflumizole	2.5944	2.2504	5.1	4.1921	0.92734	−0.36874	0	0.047515	0	3	0	1	0

no.	common name	SssNH	SssssNp	SdsssP	SHBint6	SHBint2Acnt	SHBint6Acnt
1	bensulide	2.5663	0	−2.4704	0	2	0
2	bentazon	0	0	0	0	1	0
3	bifenthrin	0	0	0	0	0	0
4	bitertanol	0	0	0	0	0	0
5	bromacil	2.6236	0	0	0	2	0
6	butralin	2.9668	0	0	0	2	3
7	butylate	0	0	0	0	0	0
8	imazapic	2.6281	0	0	0	2	0
9	captan	0	0	0	0	0	0
10	chlorothalonil	0	0	0	0	0	0
11	chlorpyrifos	0	0	−2.9185	0	0	0
12	clethodim	2.7346	0	0	0	0	0
13	clomazone	0	0	0	0	0	0

Table 1. Continued

no.	common name	SssNH	SssssNp	SdsssP	SHBint6	SHBint2Acnt	SHBint6Acnt
14	coumaphos	0	0	−2.865	0	0	0
15	cyclanilide	2.5914	0	0	0	2	0
16	cyclopropyl	6.5755	0	0	0	0	0
17	cyhalothrin	0	0	0	0	0	0
18	cypermethrin	0	0	0	0	0	0
19	cyproconazole	0	0	0	0	0	2
20	cyprodinil	3.2456	0	0	0	0	0
21	DCPA, or dacthal	0	0	0	0	0	0
22	deltamethrin	0	0	0	0	0	0
23	desmedipham	5.1311	0	0	0	2	2
24	dichlobenil	0	0	0	0	0	0
25	dicofol	0	0	0	0	0	4
26	difenoconazole	0	0	0	0	0	0
27	dimethenamid	0	0	0	0	0	0
28	diquat dibromide	0	0	0	0	0	0
29	disulfoton	0	0	−2.0229	0	0	0
30	endosulfan	0	0	0	0	0	0
31	endothall	0	0	0	0	2	0
32	esfenvalerate	0	0	0	0	0	0
33	ethalfluralin	0	0	0	0	2	9
34	etridiazole	0	0	0	0	0	0
35	fenamiphos	2.8596	0	−3.2976	0	1	0
36	fenarimol	0	0	0	0	0	2
37	fenbuconazole	0	0	0	0	0	0
38	fenitrothion	0	0	−2.7826	0	1	0
39	fenoxaprop-ethyl	0	0	0	0	0	0
40	fenoxycarb	2.5764	0	0	0	1	0
41	fenpropathrin	0	0	0	0	0	0
42	fipronil	0	0	0	0	1	0
43	flumiclorac-pentyl	0	0	0	0	0	0
44	fluridone	0	0	0	0	0	0
45	fluroxypyr-MHE	0	0	0	15.219	1	1
46	flurprimidol	0	0	0	0	0	0
47	flutolanil	2.4654	0	0	0	1	0
48	fomesafen	1.6072	0	0	0	4	0
49	fosthiazate	0	0	−3.0161	0	0	0
50	glyphosate	2.0621	0	−4.0964	26.503	4	5
51	halofenozide	2.6927	0	0	0	1	2
52	hexythiazox	3.0345	0	0	0	2	0
53	hydramethylnon	5.9404	0	0	0	1	0
54	imazamethabenz	2.7123	0	0	0	2	0
55	iprodione	2.5656	0	0	0	2	0
56	isoxaben	2.7202	0	0	0	1	0
57	kresoxim-methyl	0	0	0	0	0	0
58	lactofen	0	0	0	0	1	0
59	L-cyhalothrin	0	0	0	0	0	0
60	mepiquat chloride	0	1.0849	0	0	0	0
61	metalaxyl	0	0	0	0	0	0
62	methyl parathion	0	0	−2.7628	0	1	0
63	metolachlor (stereoisomer)	0	0	0	0	0	0
64	molinate	0	0	0	0	0	0
65	norflurazon	2.6626	0	0	0	0	0
66	oryzalin	0	0	0	0	4	13
67	pendimethalin	3.043	0	0	0	2	3
68	phorate	0	0	−2.0199	0	0	0
69	phostebupirim	0	0	−2.7998	0	0	0

Table 1. Continued

no.	common name	SssNH	SssssNp	SdsssP	SHBint6	SHBint2Acnt	SHBint6Acnt
70	prallethrin	0	0	0	0	0	0
71	prodiamine	0	0	0	0	2	13
72	prometryn	6.3551	0	0	0	0	0
73	propachlor	0	0	0	0	0	0
74	propiconazole	0	0	0	0	0	0
75	pyridate	0	0	0	0	0	0
76	quinclorac	0	0	0	0	1	0
77	sethoxydim	0	0	0	0	0	0
78	starlicide	0	0	0	0	0	0
79	tebuconazole	0	0	0	0	0	0
80	tebufenozide	2.7928	0	0	0	1	0
81	tebuthiuron	2.5371	0	0	0	2	0
82	temephos	0	0	−5.4536	0	0	0
83	terbacil	2.5299	0	0	0	2	0
84	thiazopyr	0	0	0	0	0	0
85	thifensulfuron-methyl	3.9173	0	0	0	5	0
86	thiobencarb	0	0	0	0	0	0
87	thiodicarb	0	0	0	0	0	0
88	triadimenol	0	0	0	0	0	0
89	triasulfuron	4.0754	0	0	0	5	0
90	tribuphos	0	0	−2.0584	0	0	0
91	tridiphane	0	0	0	0	0	0
92	trifloxystrobin	0	0	0	0	0	0
93	triflumizole	0	0	0	0	0	0

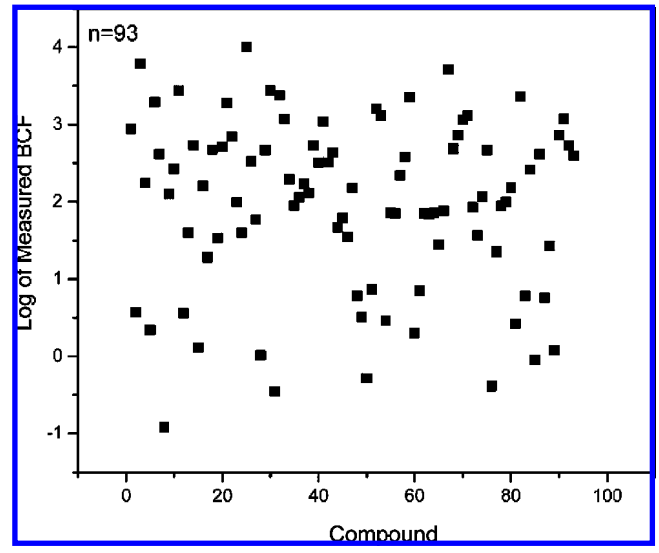

Figure 1. Distribution of log BCF values in the model development (training) data set.

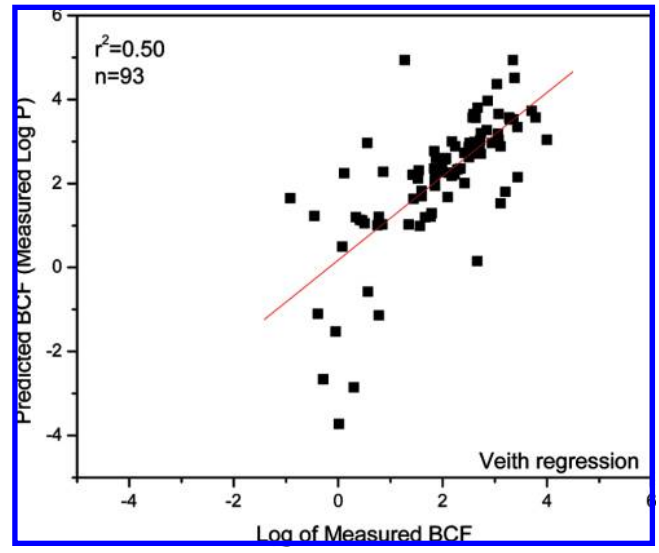

Figure 2. Veith bluegill model BCF predictions using laboratory-measured log K_{ow} values compared to the measured BCF values.

regression method provided the best solution as evidenced by evaluating regression coefficient values and residuals. The final regression relationship developed can be found in eq 1.

$$\begin{aligned}\text{Log BCF} = {} & 0.2432 \times \text{Log } P - 1.601 \times \text{ABSQon} - \\ & 0.9529 \times \text{MaxNeg} - 55.85 \times \text{xch10} + 1.392 \times \text{xvp10} + \\ & 168.7 \times \text{xvch9} + 0.09185 \times \text{SaasC\_acnt} - \\ & 2.861 \times \text{SssNH\_acnt} + 0.5661 \times \text{SssO\_acnt} + \\ & 0.7797 \times \text{SdssS\_acnt} + 1.078 \times \text{SssNH} - \\ & 1.886 \times \text{SssssNp} + 0.2769 \times \text{SdsssP} + \\ & 0.04637 \times \text{SHBint6} + 0.2835 \times \text{SHBint2\_Acnt} + \\ & 0.09498 \times \text{SHBint6\_Acnt} + 1.45606 \quad (1)\end{aligned}$$

We have included a calculated Log P as a descriptor in the QSAR model. This inclusion is because the simple partitioning of compounds toward either octanol or water has been a useful predictor of a molecule's tendency to move to tissue. However, using Log P alone is not predictive enough (**Figures 2** and **3**). K_{ow} or the Log P results have been viewed mechanistically as a predictor of a molecule's tendency either to partition to lipids or to be hydrophilic. However, beyond the correlation of Log P with lipophilicity, there are molecular reasons for such

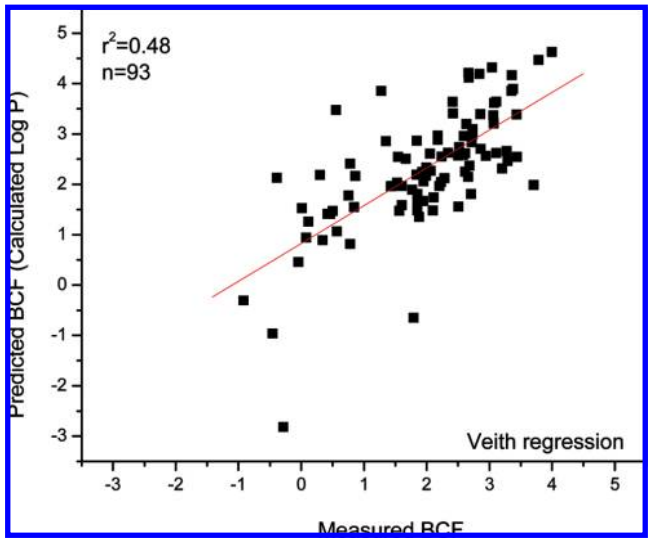

Figure 3. Veith bluegill model BCF predictions using structurally calculated log K_{ow} values compared to the measured BCF values.

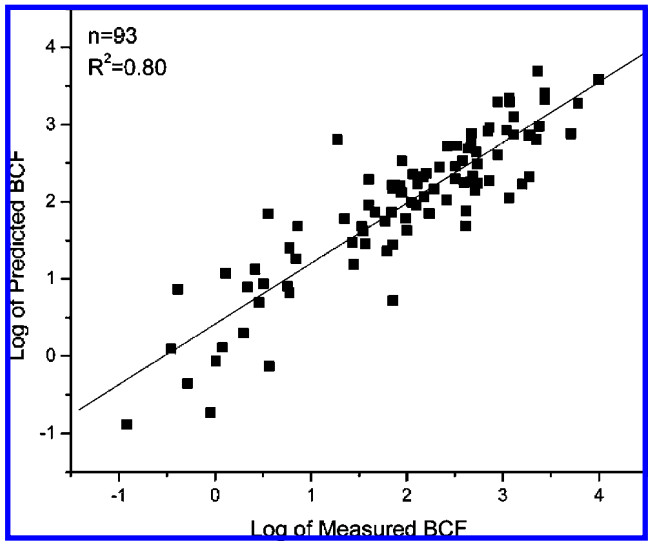

Figure 4. Predicted log BCF values using the structurally based model compared to the measured BCF values.

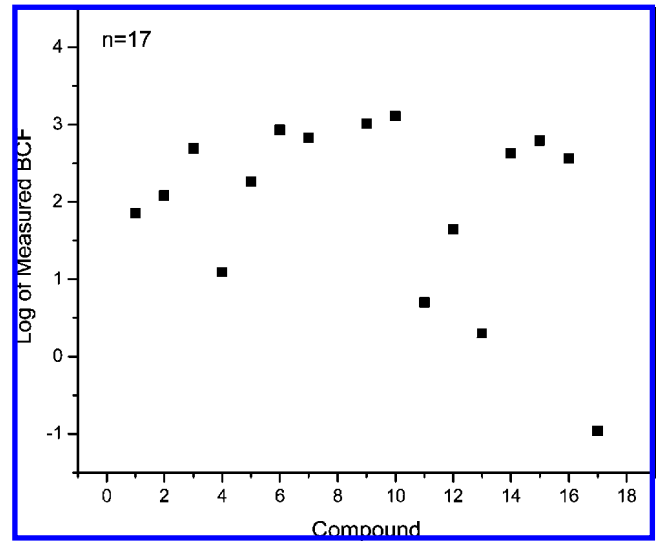

Figure 5. Distribution of log BCF values in the model validation data set.

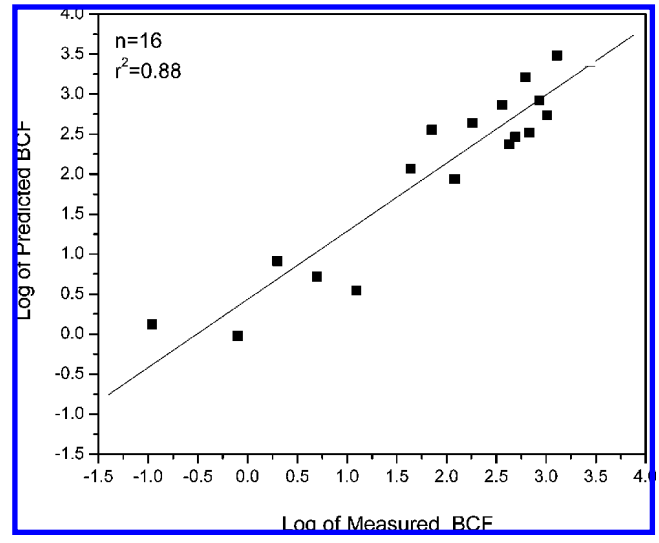

Figure 6. Predicted log BCF values using the structurally based model compared to the measured BCF values for the validation data set.

phenomena. The additional factors added to the QSAR model can provide insight into the molecular reasons for hydrophilic/lipophilic behaviors.

The resulting regression coefficient value for the eq 1 relationship was $R^2 = 0.80$, $n = 93$ (Origin Laboratories) (*8*). To determine if each model descriptor provided a contribution to the regression model, data were imported into Microsoft Office Excel 2003 Analysis ToolPak (*12*), and ANOVA was performed on the regression model. Results from the ANOVA indicate that all variables were significant at the 95% probability level ($P = 0.0429-0.0000000427$), and the strong $R^2 = 0.80$ was confirmed. Residuals from the regression relationship were analyzed and found to be normally distributed and followed the guidance for model selection as described by Aldworth and Jackson (*13*). A graphical presentation of the predicted to measured BCF results based on this model can be found in **Figure 4**.

Because QSAR software may require several different descriptors to describe, for example, the location of a nitrogen functional group on ring 2 of a three-ring benzyl structure and require descriptors to describe charge on the ring, the number of additional structural descriptors beyond log *P* may seem to be large compared to traditional BCF regression relationships. However, as the descriptive power of QSAR software increases, so do the number of descriptors that allow these software programs to describe small differences in structures that are important for determining activity. Furthermore, as the diversity of molecular structures increases, it is clear that in order to obtain a descriptive model for a large variety of different molecule types (linear molecules, one-, two-, or three-ring systems), a large number of molecular attributes (descriptors) will be required. Thus, if models are developed that are to be robust in predictive capability, then larger numbers of molecular descriptors will be required so that various structure−activity relationships can described.

One of the challenges in performing structural analysis is ensuring that the final model descriptors make sense and reasonably influence model predictions. In addition to the summary of model descriptors found in **Table 3**, a summary of molecular descriptors found in the regression QSAR (eq 1) is briefly given. The QSAR descriptors used in this model fall into two general categories. The descriptors can be classified first as simple counting descriptors such as how many amine groups are on the molecules. The first category of descriptors

Table 2

no.	name	Log BCF	Log *P*	ABSQon	MaxNeg	xch10	xvp10	xvch9	SaasC_acnt	SssNH_acnt	SssO_acnt	SdssS_acnt
1	boscalid	1.85	2.5555	4.2917	0.96897	−0.4328	0	0.080155	0	6	1	0
2	metconazole	2.08	1.9401	3.6361	0.80863	−0.38801	0	0.2446	0	2	0	0
3	pyraclostrobin	2.69	2.469	3.3564	1.4989	−0.35417	0	0.072418	0	5	0	3
4	diflufenzopyr	1.09	0.54545	2.6297	1.9525	−0.39448	0	0.025713	0	5	2	1
5	pyradiben	2.26	2.6401	5.8066	0.70786	−0.3873	0	0.25296	0	2	0	0
6	cyfluthrin	2.93	2.9217	5.4834	1.1475	−0.35308	0	0.095349	0	4	0	2
7	terbufos	2.83	2.5197	4.2555	0.47007	−0.23504	0	0	0	0	0	2
8	2,4-D	3.01	2.7355	2.3947	0.98095	−0.35252	0	0.003368	0	3	0	1
9	oxyfluorfen	3.11	3.4846	4.3575	0.85117	−0.31314	0	0.047475	0	6	0	2
10	imidacloprid	0.70	0.7196	1.1681	1.3851	−0.34365	0	0.040027	0	2	1	0
11	paclobutrazol	1.64	2.0668	6.4874	1.6072	−0.39089	0	0.21517	0	4	0	1
12	oxamyl	0.30	0.91363	0.68983	1.624	−0.39037	0	0	0	0	1	1
13	parathion	2.63	2.3778	2.8848	0.91415	−0.25281	0	0.037732	0	2	0	3
14	permethrin	2.79	3.2091	5.8166	0.96556	−0.35526	0	0.098988	0	3	0	2
15	phosalone	2.56	2.8653	3.9776	1.2974	−0.38792	0	0.34595	0.002196	3	0	3
16	imazapic	−0.96	0.12313	1.5656	1.85	−0.3917	0	0.032092	0	3	1	0

no.	name	SssNH	SssssNp	SdsssP	SHBint6	SHBint2_Acnt	SHBint6_Acnt
1	boscalid	0	2.8844	0	0	0	1
2	metconazole	0	0	0	0	0	0
3	pyraclostrobin	0	0	0	0	0	0
4	diflufenzopyr	0	4.3832	0	0	0	3
5	pyradiben	0	0	0	0	0	0
6	cyfluthrin	0	0	0	0	0	0
7	terbufos	0	0	0	−2.0762	0	0
8	2,4-D	0	0	0	0	15.234	1
9	oxyfluorfen	0	0	0	0	0	1
10	imidacloprid	0	2.1862	0	0	0	4
11	paclobutrazol	0	0	0	0	0	0
12	oxamyl	0	2.2102	0	0	20.622	1
13	parathion	0	0	0	−2.801	0	1
14	permethrin	0	0	0	0	0	0
15	phosalone	0	0	0	−2.4452	0	0
16	imazapic	0	2.6421	0	0	0	2

includes SaasC_acnt, SssNH_acnt, SssO_acnt, xch10 (which is a chi index), and SdssS_acnt. The second category of descriptors can be categorized as those that communicate a particular aspect of molecular or functional group charge. The descriptors that communicate charge could be further separated into two categories. Descriptors both count and provide charge information and simply convey "gross" molecular charge information. The descriptors in the latter group include the ABSQon and MaxNeg descriptors. Descriptors that convey count and valence (charge) information include all of the E-state descriptors SssNH, SssssNp, SdsssP, and SHBint6. E-state descriptors are used to help quantify electron accessibility at each structural component (e.g., bond) and represent the relative accessibility of those electrons to participate in intermolecular interactions. Two more variables that provided information on hydrogen bond interaction are the SHBint2_Acnt and SHBint6_Acnt descriptors. Two other variables that convey charge information are the chi index descriptors xvp10 and xvch9, which include not only counts but also the characteristics of the atom being described.

Because our regression relied on two-dimensional descriptors, many were required to properly describe the wide range of molecules and have predictive capability. We additionally tried regressing molecular weight as a variable, but that descriptor did not provide improvement to any of the models we developed. Furthermore, whereas Log *P* or octanol/water coefficients have shown positive correlation to BCF previously (*14*−*16*) and was included in the final QSAR, it alone as a single descriptor was not capable of predicting BCF for these compounds. If the process of bioconcentration or bioaccumulation can be described as an across the membrane phenomenon, then the fact that the majority of the QSAR model descriptors are relating to molecular valence or charge in some fashion is not surprising and might have been expected. Cell membranes are typically viewed as lipid bilayer structures. The lipid bilayer structure is normally an alkyl chain that serves to hold the membrane structure together by van der Waals forces as well as by other bonding phenomena (electrostatic forces). Although there are still many ideas of how nonpolar compounds move across membranes, it is well-known that polarity, van der Waals forces, and molecular charge all have a role.

Model Validation. To evaluate the ability of the regression to predict measured BCF values for classes of chemicals similar to those in the training database, an additional set of 16 molecules that were not used in the development of the original relationship were run through the regression via the QSAR software. The range of measured BCF values may be observed as a plot in **Figure 5**. The goal of this part of the work was to evaluate the ability of the QSAR regression model to predict the BCF of other compounds. If the QSAR model is able to predict the BCF values of the additional molecules, confidence is gained that the method could be an improvement over simple regression methods. Results from running the validation molecules through the regression model can be found in **Figure 6**. The model developed in eq 1 was able to predict Log BCF for the 16 molecules in the validation set, yielding an $r^2 = 0.88$.

On the basis of an examination of the methods proposed in the literature for predicting BCF, it is clear that most are simple regressions using a laboratory measure of Log *P*. The Log *P* values used in this QSAR regression were a molecularly calculated property. Log *P* does provide indirect information about a molecule's hydrophilic or lipophilic properties. These

Table 3

Descriptor	Definition
ABSQon	The sum of absolute values of the charges on the nitrogen and oxygen atoms in a molecule:
MaxNeg	The largest negative charge over the atoms in a molecule.
xch10	Simple 10th order chain chi index
xvp10	Valence 10th order path chi index
xvch9	Valence 9th order chain chi index
SaasC_acnt	Count of all (–– C(–) ––) groups in a molecule
SssNH_acnt	Count of all (– NH –) groups in a molecule.
SssO_acnt	Count of all (– O –) groups in a molecule
SdssS_acnt	Count of all (= S <) groups in a molecule
SssNH	Sum of all (– NH –) E-State values in a molecule
SssssNp	Sum of all (> N+ <) E-State values in a molecule
SdsssP	Sum of all (= P(–) <) E-State values in a molecule
SHBint6	Internal Hydrogen Bonding Index:SHBint6 = E-State (acceptor) * HE-State (donor),6 - numbe of skeletal bonds between donor and acceptor
SHBint2_Acnt	Count of internal hydrogen bonds with 2 skeletal bonds between donor and acceptor
SHBint6_Acnt	Count of internal hydrogen bonds with 6 skeletal bonds between donor and acceptor

hydrophilic or lipophilic properties can further be predictive of sorption to tissue, soil, or sediment.

The QSAR regression model presented in eq 1 is capable of predicting BCF on the basis of molecular structural properties for chemical classes examined in the data sets. Several simpler models were developed in an attempt to predict BCF as well. However, the simpler models provided a poorer fit and less confidence that such models would produce reasonably accurate BCF estimates for the range of molecules that might possibly be examined in regulatory review programs or during chemical development. Whereas eq 1 has many descriptors, it is clear from our analysis that many descriptors need to be included in order to have a robust method. One departure from many previous modeling approaches for predicting BCF is that this model is totally structurally based (there were no experimentally measured values for the compound descriptors). When the potential for any molecule to accumulate in an aquatic system is assessed, many factors that a BCF value cannot accurately describe need to be considered. The BCF values used to develop this method were based on constant-exposure, flow-through tank systems that contain only fish and laboratory-quality water (no sediment or plants, etc.). Depending on the type of compound assessed (pesticide, surfactant, etc.), exposure conditions like those in the standard guideline systems will not occur in natural systems. Therefore, this QSAR model provides a prediction based on these artificial conditions (constant exposure, flow-through tank). The model does not consider the many environmental factors that may greatly influence the estimation of BCF in the calculation of the end points relevant for field exposures. Factors such as dietary preference, variable environmental conditions (e.g., temperature, carbon black, or humic materials), and physiochemical properties such as hydrolysis, photolysis, or partitioning are not accounted for. Additional factors such as depuration and differential spatial/temporal exposure regimens to organisms as they naturally move through their environment are not considered either. Furthermore, it is evident that classifying compounds using static conditions or based on K_{ow} prediction approaches can be inadequate to identify substances with a potential to bioaccumulate in food webs (*17*, *18*). A further caution on the application of the QSAR model presented in eq 1 is the necessary understanding of the types of molecules used in the development of the model. Knowledge of the similarity of the new chemical under evaluation to the compounds used to develop the model must be understood before decisions based on the predicted BCF values can be made. Nevertheless, properly interpreted, the prediction of BCF by the model developed in this work can provide insightful information on the potential behavior of many new and existing classes of pesticides and similar chemistries in aquatic systems.

ACKNOWLEDGMENT

We thank Dr. Beate Escher, Dr. Annie Weisbrod, and Dr. Andy Goetz for their valuable contributions to the manuscript.

LITERATURE CITED

(1) OECD (Organization for Economic Cooperation and Development), 1996. Bioconcentration: Flow-through Fish Tests, 305, last updated 14 June 1996; Paris, France.

(2) Barber, M. C. A review and comparison of models for predicting dynamic chemical biocentration in fish. *Environ. Toxicol. Chem.* **2003**, *22*, 1963–1992.

(3) U.S. EPA Environmental Fate Database, http://cfpub.epa.gov/pfate/home.cfm.

(4) OPPTS Harmonized Test Guidelines, Series 850 Ecological Effects Test Guidelines, http://www.epa.gov/opptsfrs/publications/OPPTS_Harmonized/850_Ecological_Effects_Test_Guidelines/Drafts/.

(5) Veith, G. D.; Mace, K. J.; Petrocelli, S. R.; Carol, J. An evaluation of using partition coefficients and water solubilities to estimate bioconcentration factors for organic chemicals in fish In *Aq. Tox. STP 707*; Eaton, J. G., Parish, P. R., Hendricks, A. C., Eds.; American Society for Testing and Materials: Philadelphia, PA, 1980; pp 116−129.

(6) FOOTPRINT Pesticide Properties Database, http://sitem.herts.ac.uk/aeru/footprint/en/.

(7) *MDL QSAR*, Quantitative Structure−Activity Relationship Software; MDL Information Systems: San Leandro, CA, 2006.

(8) Origin 7.5: 2003 Scientific Graphing and Analysis Software, OriginLab Corp., Northhampton, MA.

(9) National Center for Biotechnology Information, http://pubchem.ncbi.nlm.nih.gov/.

(10) Bermudez-Saldana, J. M.; Cronin, M. T. D. Quantitative structure−activity relationships for the toxicity of organophosphous and carbamate pesticides to the tainbow trout *Onchorhyncus mykiss*. *Pest Manag. Sci.* **2006**, *62*, 819–831.

(11) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694–701.

(12) Microsoft Office Excel 2003 Analysis ToolPak, Microsoft Corp., Redmond,WA.

(13) Aldworth, J.; Jackson, S. H. A systematic approach for determining proper selection of regression models for analysis of environmental fate datasets *Pest Manag. Sci.* **2008**.

(14) De Wolf, W.; de Bruijn, J. H. M.; Seinen, W.; Hermens, J. L. M. Influence of biotransformation on the relationship between bioconcentration factors and octanol−water partition coefficients. *Environ. Sci. Technol.* **1992**, *26*, 1197–1201.
(15) Meylan, W. M.; Howard, P. H.; Boethling, R. S.; Aronson, D; Printup, H.; Gouchie, S. Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient. *Environ. Toxicol. Chem.* **1999**, *18*, 664–672.
(16) Kelly, B. C.; Gobas, F. A. P. C.; McLachlan, M. S. Intestinal absorption and biomagnificatioin of organic contaminants in fish, wildlife, and humans. *Environ. Toxicol. Chem.* **2004**, *23*, 2324–2336.
(17) Mcgeer, J. C.; Brix, K. V.; Skeaff, J. M.; Deforest, D. K.; Brigham, William, S. I.; Adams, J.; Green, A. Inverse relationship between bioconcentration factor and exposure concentration for metals: implications for hazard assessment of metals in the aquatic environment. *Environ. Toxicol. Chem.* **2003**, *22*, 1017–1037.
(18) Franke, C. How meaningful is the bioconcentration factor for risk assessment? *Chemosphere* **1996**, *32*, 1897–1905.

Received for review July 24, 2008. Revised manuscript received December 2, 2008. Accepted December 3, 2008.

JF803064Z