Bull. Environ. Contam. Toxicol. (2001) 66:755–761
© 2001 Springer-Verlag New York Inc.
DOI: 10.1007/s00128-001-0073-3

Bulletin of Environmental Contamination and Toxicology

Quantitative Structure–Property Relationships for Octanol-Air Partition Coefficients of PCDD/Fs

J. W. Chen,[1,2,*] X. Quan,[1] Y. Z. Zhao,[1] F. L. Yang,[1] K.-W. Schramm,[2] A. Kettrup[2,3]

[1] School of Environmental Science and Technology, Dalian University of Technology, Zhongshan Road 158-129, Dalian 116012, People's Republic of China
[2] Institute of Ecological Chemistry, GSF-National Research Center for Environment and Health, D-85764, Neuherberg, Munich, Germany
[3] Department of Ecological Chemistry and Environmental Analytics, Technical University of Munich, 85350 Freising-Weihenstephan, Germany

Received: 26 February 2001/Accepted: 16 March 2001

Polychlorinated dibenzo-*p*-dioxins and dibenzo-*p*-furans (PCDD/Fs) are typical persistent pollutants with high toxicity (Nebert, 1989). Recent studies (Younes 1999; Fossi *et al.* 1999) reveal that most PCDD/Fs are endocrine disrupting chemicals. The octanol-air partition coefficient (K_{OA}) is recognized as a key descriptor of chemicals partitioning between the atmosphere and organic phases (Harner *et al.* 2000). Recently, K_{OA} based approaches have been successfully employed to model surface-air partitioning of persistent organic pollutants of aerosols, soil and vegetation (Harner *et al.* 2000). However it is difficult to comprehensively determine the K_{OA} for all PCDD/Fs because of large expenditures of money and time. Thus the development of quantitative structure-property relationship (QSPR) models for K_{OA} is very important.

As quantum chemical descriptors can be easily obtained by computation, can clearly describe defined molecular properties, and are not restricted to closely related compounds, the development of QSPR models in which quantum chemical descriptors are used is of great importance. According to the present chemometric theory, as many relevant data as possible should be considered in QSPR studies because this increases the probability of a good characterization of compounds (Kaliszan, 1993). As a consequence of the increase of the number of descriptors, the problem of intercorrelation of independent variables (multicollinearity) will increase. Especially when the number of independent variables is equal to or greater than the number of compounds in the training set, regression analysis (a method that was frequently used in QSPR studies) will not be useful. To overcome these problems, the partial least squares (PLS) method, a widely used chemometric method first developed by Wold *et al.* (1984), will be used in this study.

MATERIALS AND METHODS

Recently, Harner *et al.* (2000) have determined logK_{OA} values for 10 PCDD and 1 PCDFs at 298 K using a generator column method. These 11 PCDD/Fs constitute the training set of the study. The logK_{OA} values are reproduced in Table 1. In addition, some other PCDD/Fs (mainly those with chlorines substituted in the 2,3,7,8 positions) were selected randomly in the study. The 2,3,7,8- substituted PCDD/Fs are thought to pose a risk to human health due to their toxicity, carcinogenic potency, and potential effects on animal reproductive and immunological systems (Safe, 1986). Their logK_{OA} will be predicted based on the models obtained. The predicted values may be useful in exposure assessment of the PCDD/Fs.

[*] *Present address:* Institute of Ecological Chemistry, GSF-National Research Center for Environment and Health, D-85764, Neuherberg, Munich, Germany
Correspondence to: J. W. Chen

PM3 (Stewart 1989a; 1989b) Hamiltonian contained in the quantum chemical computation software MOPAC (Ver. 6.0, *Stewart, J. J. P., 1990, Frank J. Seiler Research Laboratory, U. S. Air Force Academy, Co 80840*) was used to compute quantum chemical descriptors of the PCDD/Fs. The PM3 was selected because it is a recently developed semi-empirical molecular orbital algorithm and the computational time is much shorter than needed by *ab initio* methods. MOPAC was run with the following keywords: PM3, ESP, POLAR, DIPOLE, PRECISE.

A total of 12 MOPAC derived descriptors reflecting the overall character of the PCDD/F molecules were selected in this study. They are molecular weight (*Mw*), average molecular polarizability (α), dipole moment (μ), final heat of formation (*HOF*), total energy (*TE*), electronic energy (*EE*), core-core repulsion energy (*CCR* , *CCR* = *TE* - *EE*), energy of the highest occupied molecular orbital (E_{homo}), energy of the lowest unoccupied molecular orbital (E_{lumo}), the largest negative atomic charge on a carbon atom (Q_C^-), the most positive net atomic charges on a hydrogen atom (Q_H^+), and the most positive net atomic charges on a chlorine atom (Q_{Cl}^+). The values for some of the molecular

Table 1. The PCDD/Fs and their logK_{OA} values (298 K)*

No	Compounds	*Obs.*	*Pred.*	*SE*	*Obs.* (RTI)
1	Dioxin		7.08	±0.13	
2	1-CDD	7.86	7.77	±0.11	
3	2-CDD		7.78	±0.09	
4	2,3-D_2CDD		8.57	±0.07	8.50
5	2,7-D_2CDD	8.36	8.45	±0.07	8.48
6	2,8-D_2CDD	8.36	8.46	±0.07	8.48
7	1,2,4-T_3CDD		9.22	±0.06	8.97
8	2,3,7-T_3CDD	9.14	9.21	±0.05	9.42
9	1,2,3,4-T_4CDD	9.70	9.84	±0.06	9.64
10	1,2,3,7-T_4CDD		9.89	±0.05	9.94
11	1,3,6,8-T_4CDD		9.76	±0.05	9.38
12	2,3,7,8-T_4CDD	10.05	9.85	±0.05	9.95
13	1,2,3,4,7-P_5CDD	10.67	10.47	±0.06	10.42
14	1,2,3,7,8-P_5CDD	10.57	10.51	±0.06	10.46
15	1,2,3,4,7,8-H_6CDD	11.11	11.07	±0.07	10.95
16	1,2,3,6,7,8-H_6CDD		11.04	±0.07	10.97
17	1,2,3,7,8,9-H_6CDD		11.05	±0.07	11.01
18	1,2,3,4,6,7,8-H_7CDD	11.42	11.61	±0.09	11.45
19	O_8CDD		12.15	±0.12	12.05
20	Dibenzofuran		7.19	±0.13	
21	2,8-D_2CDF		8.54	±0.13	8.36
22	1,2,7,8-T_4CDF		9.97	±0.11	9.78
23	2,3,7,8-T_4CDF	10.02	10.03	±0.13	9.82
24	1,2,3,8,9-P_5CDF		10.58	±0.11	10.48
25	2,3,4,7,8-P_5CDF		10.72	±0.12	10.37
26	1,2,3,4,7,8-H_6CDF		11.33	±0.13	10.77
27	1,2,3,6,7,8-H_6CDF		11.29	±0.13	10.78
28	1,2,3,4,8,9-H_6CDF		11.22	±0.12	10.95
29	1,2,3,7,8,9-H_6CDF		11.25	±0.13	10.93
30	2,3,4,6,7,8-H_6CDF		11.30	±0.13	10.87
31	1,2,3,4,6,7,8-H_7CDF		11.90	±0.14	11.17
32	1,2,3,4,7,8,9-H_7CDF		11.88	±0.14	11.43
33	O_8CDF		12.46	±0.15	11.90

* *Obs.*: Observed values (298 K) determined by Harner *et al.* (2000) using a generator column method; *Pred.*: Predicted values by model (2) of this study, *SE*: Standard errors of the predicted values; *Obs.*(RTI): logK_{OA} values determined semi-emperically by retention time indices (RTI) using eq. 2 of Harner *et al.* (2000) (the data was provided by Dr. Harner).

descriptors are listed in Table 2. The compound numbers in Table 2 correspond to those in Table 1. The unit of *HOF* is kilocalories, and units of energy, charge, dipole and polarizability are electron volts (eV), atomic charge units (a.c.u) and atomic units (a.u.) respectively.

Table 2. Selected quantum chemical descriptors of the PCDD/Fs

No	α	Mw	TE	EE	CCR	E_{lumo}	Q_{Cl}^{+}
1	110.156	184.194	-2130.603	-11891.038	9760.435	-0.178	
2	119.914	218.639	-2431.938	-13536.608	11131.670	-0.298	0.102
3	122.345	218.639	-2432.001	-13399.609	10967.608	-0.368	0.073
4	134.107	253.084	-2733.349	-15066.905	12333.557	-0.533	0.101
5	134.935	253.084	-2733.395	-14957.399	12224.004	-0.522	0.076
6	134.897	253.084	-2733.396	-14958.712	12225.316	-0.527	0.076
7	143.019	287.529	-3034.614	-17081.506	14046.892	-0.603	0.134
8	147.097	287.529	-3034.741	-16674.459	13639.717	-0.665	0.103
9	154.804	321.974	-3335.921	-19000.732	15664.811	-0.709	0.139
10	157.106	321.974	-3336.026	-18660.313	15324.287	-0.745	0.137
11	155.775	321.974	-3336.040	-18690.131	15354.091	-0.722	0.115
12	159.537	321.974	-3336.084	-18441.711	15105.627	-0.785	0.106
13	167.900	356.419	-3637.310	-20736.372	17099.062	-0.822	0.141
14	169.555	356.419	-3637.369	-20492.895	16855.526	-0.854	0.139
15	180.536	390.865	-3938.653	-22625.924	18687.271	-0.922	0.142
16	179.648	390.865	-3938.653	-22609.132	18670.479	-0.92	0.139
17	179.086	390.865	-3938.652	-22650.455	18711.803	-0.916	0.142
18	190.133	425.310	-4239.935	-24847.674	20607.739	-0.981	0.146
19	200.645	459.755	-4541.216	-27151.547	22610.331	-1.037	0.146
20	106.918	168.195	-1837.164	-10106.423	8269.259	-0.477	
21	130.157	237.085	-2439.969	-13066.484	10626.515	-0.785	0.068
22	151.973	305.975	-3042.640	-16650.267	13607.627	-1.041	0.115
23	157.256	305.975	-3042.669	-16417.580	13374.911	-1.079	0.105
24	160.504	340.420	-3343.879	-18742.164	15398.286	-1.138	0.126
25	168.124	340.420	-3343.956	-18340.863	14996.908	-1.169	0.137
26	178.161	374.865	-3645.271	-20451.212	16805.942	-1.273	0.14
27	177.322	374.865	-3645.272	-20439.361	16794.089	-1.263	0.137
28	172.240	374.865	-3645.163	-20756.808	17111.646	-1.235	0.139
29	175.059	374.865	-3645.214	-20603.679	16958.465	-1.264	0.13
30	178.805	374.865	-3645.240	-20352.829	16707.590	-1.254	0.14
31	188.909	409.310	-3946.554	-22532.548	18585.994	-1.351	0.144
32	186.811	409.310	-3946.506	-22679.458	18732.952	-1.351	0.142
33	198.324	443.755	-4247.787	-24845.576	20597.789	-1.429	0.144

Simca (Simca-S Version 6.0, *Umetri AB & Erisoft AB*) software was used to perform the PLS analysis. The conditions for the computation were based on the default values of the software. The criterion used to determine the model dimensionality - the number of significant PLS components - is cross validation (CV). With CV, when the fraction of the total variation of the dependent variables that can be predicted by a component, Q^2, for the whole data set is larger than a significance limit (0.097), the tested PLS component is considered significant. When the cumulative Q^2 for the extracted components, Q^2_{cum}, is larger than 0.5, the model is considered to have a good prediction ability. Model adequacy was mainly measured as the number of PLS principal components (k), Q^2_{cum}, the correlation coefficient between observed values and fitted values (R), and the significance level (p).

RESULTS AND DISCUSSION

PLS analysis for the 11 PCDD/Fs in the training set, with logK_{OA} as dependent variables and the 12 quantum chemical descriptors as independent variables, resulted in QSPR model (1). The results of the model are listed in Table 3. In Table 3, $R^2_{\mathrm{X(adj)(cum)}}$ and $R^2_{\mathrm{Y(adj)(cum)}}$ stand for cumulative variance of all the X's and Y's, respectively, explained by all extracted components. So it can be concluded from Table 3 that 1 PLS principal component was selected in model (1), and the PLS principal components explained 58.7% of the variance of the independent variables, and 98.0% of the variance of the dependent variable.

Table 3. Model fitting results

Models	k	$R^2_{\mathrm{X(adj)(cum)}}$	$R^2_{\mathrm{Y(adj)(cum)}}$	Q^2_{cum}	R	p
(1)	1	0.587	0.980	0.969	0.991	4.080×10^{-9}
(2)	2	0.950	0.984	0.981	0.994	7.602×10^{-10}
(3)	2	0.999	0.978	0.975	0.991	3.439×10^{-9}
(4)	1	0.993	0.971	0.970	0.987	2.085×10^{-8}

VIP (Variable Importance in the Projection) is a parameter that shows the importance of a variable in a model. Terms with a large value of *VIP*, larger than 1, are the most relevant for explaining the dependent variable. As indicated by the *VIP* values of model (1) listed in Table 4, the descriptors α, Mw, TE, EE, CCR, E_{lumo} and Q_{Cl}^+ are more significant than the other 5 descriptors in governing the logK_{OA} values of the PCDD/Fs. Although the PLS method offers the advantage of handling data sets where the number of independent variables is greater than the number of observations, it can be seen that considerable worse predictions are obtained if many irrelevant descriptors are included in the PLS model (Luco, 1999). So it is necessary to perform a PLS analysis that includes the 7 significant descriptors only. Such a PLS analysis resulted in model (2). The *VIP* values of model (2) (Table 4) showed that the descriptors E_{lumo} and Q_{Cl}^+ were less significant than the remaining 4 descriptors. A new PLS analysis with exclusion of E_{lumo} and Q_{Cl}^+ resulted in model (3). The *VIP* values (Table 4) of model (3) indicated that α and Mw were two most significant descriptors in governing the logK_{OA} values of the PCDD/Fs. Again it would be interesting to perform a PLS analysis with the inclusion of the two descriptors only. Such an analysis resulted in Model (4).

As can be seen from Table 3, the statistics $R^2_{\mathrm{X(adj)(cum)}}$, $R^2_{\mathrm{Y(adj)(cum)}}$, Q^2_{cum} and R of model (2) are higher than those of model (1), and the significance level (p) of model (2) are smaller than the p value of model (1). So model (2) are more statistically significant than model (1), as a result of removing "noisy" descriptors. It was also because of removing redundant descriptors that $R^2_{\mathrm{X(adj)(cum)}}$ of model (2) increased significantly

Table 4. *VIP*s (Variable Importance in the Projection) and pseudo-regression coefficients (Unscaled)

	Model (1)			Model (2)	
Variables	*VIP*	*Coefficients*	Variables	*VIP*	*Coefficients*
α	1.230	$7.107\text{x}10^{-3}$	α	1.048	$1.068\text{x}10^{-2}$
Mw	1.219	$2.355\text{x}10^{-3}$	*Mw*	1.037	$2.881\text{x}10^{-3}$
TE	1.201	$-2.629\text{x}10^{-4}$	*TE*	1.023	$-2.760\text{x}10^{-4}$
EE	1.178	$-4.046\text{x}10^{-5}$	*EE*	1.006	$-3.353\text{x}10^{-5}$
CCR	1.172	$4.778\text{x}10^{-5}$	*CCR*	1.002	$3.748\text{x}10^{-5}$
E_{lumo}	1.101	$-5.819\text{x}10^{-1}$	E_{lumo}	0.981	-1.603
Q_{Cl}^{+}	1.047	4.770	Q_{Cl}^{+}	0.895	3.904
E_{homo}	0.934	-1.234	*Constant*		3.436
Q_{C}^{-}	0.869	-9.358			
HOF	0.717	$-6.446\text{x}10^{-3}$			
Q_{H}^{+}	0.499	1.001x10			
μ	0.313	$-2.041\text{x}10^{-1}$			
Constant		-9.128			
	Model (3)			Model (4)	
Variables	*VIP*	*Coefficients*	Variables	*VIP*	*Coefficients*
α	1.053	$4.812\text{x}10^{-1}$	α	1.004	$2.820\text{x}10^{-2}$
Mw	1.002	$8.422\text{x}10^{-3}$	*Mw*	0.995	$9.342\text{x}10^{-3}$
CCR	0.982	$-9.869\text{x}10^{-5}$	*Constant*		2.386
TE	0.981	$-2.628\text{x}10^{-4}$			
EE	0.980	$6.296\text{x}10^{-5}$			
Constant		1.361			

over model (1). By the same comparison of statistics $R^2_{\text{Y(adj)(cum)}}$, Q^2_{cum}, *R* and *p* listed in Table 3, it can be concluded that model (4) is less significant than model (3), and model (3) is less significant than model (2). This implies that the descriptors E_{lumo}, Q_{Cl}^{+}, *CCR*, *TE* and *EE* contain some necessary molecular structural information relevant to $\log K_{\text{OA}}$, so these descriptors should not be removed from the models.

Therefore model (2) is the best one. As indicated by *R* and *p* values of model (2) listed in Table 3, for the 11 PCDD/Fs under study, the correlation between observed and predicted $\log K_{\text{OA}}$ values is very significant (Figure 1). As the cross-validated Q^2_{cum} values of model (1) is remarkably larger than 0.50, model (2) is surely stable and has a good prediction ability. Based on model (2), $\log K_{\text{OA}}$ for the other PCDD/Fs were predicted, as listed in Table 1. As shown by Table 1 and Figure 2, the predicted values were consistent with the corresponding $\log K_{\text{OA}}$ values determined semi-emperically by retention time indices (RTI) using eq. 2 of Harner *et al.* (2000). So model (2) has been validated on the basis of predictions for PCDD/Fs not included in the training set.

The pseudo-regression coefficients of the independent variables and constants transformed from PLS results for the 4 QSPR models were also listed in Table 4. From the positive and negative symbols of the coefficients of the independent variables, one can evaluate the effects of each independent variable on the dependent variables. Based on the unscaled coefficients and constants, QSPR equations like those obtained from multiple regression analysis can be obtained.

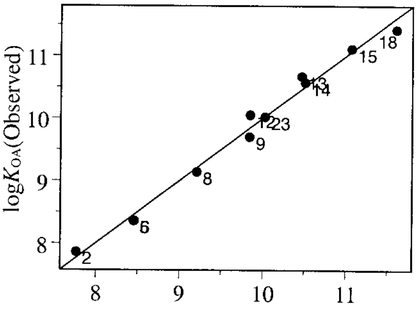

Figure 1. Plot of observed logK_{OA} values versus those predicted by model (2)

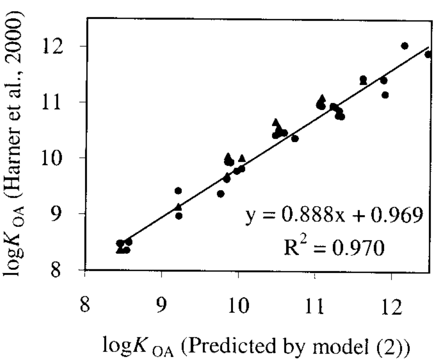

Figure 2. Plot of logK_{OA} values predicted by model (2) and those determined semi-emperically by retention time indices (RTI) using eq. 2 of Harner *et al.* (2000) •Determined semi-emperically by RTI; ▲Determined by generator column methods

From the data in Table 4, one may conclude the following: (I) Increasing α, Mw, CCR and Q_{Cl}^{+} values of the PCDD/Fs leads to increasing logK_{OA} values, while increasing TE, EE and E_{lumo} values of the PCDD/Fs leads to decreasing logK_{OA}. This is because these descriptors are inter-correlated. α correlates with Mw, CCR and Q_{CL}^{+} positively, and correlates with TE and EE negatively. Their correlation coefficients were listed in Table 5. α is the most significant descriptor in governing logK_{OA} of the PCDD/Fs. The increase of logK_{OA} with α is reasonable since intermolecular dispersive forces are in direct proportion to the product of α of two interactional molecules, and PCDD/Fs with great α value may have great intermolecular dispersive forces with octanol molecules, which favor to partition into octanol phase. So the more chlorines in PCDD/F molecules, the greater the α and Mw, and the greater the logK_{OA} values. (II) Increasing E_{lumo} values of the PCDD/Fs leads to decreasing logK_{OA} values. E_{lumo} measures the ability of a molecule to accept electrons in intermolecular interactions. So it can be concluded that the lower the E_{lumo} values, the greater the tendency of PCDD/F molecules to accept electrons in intermolecular interactions, the greater the

Table 5. Correlation coefficients between some descriptors ($p<0.05$)

	α	Mw	TE	EE	CCR	E_{lumo}	Q_{CL}^{+}
α	1						
Mw	0.995	1					
TE	-0.980	-0.991	1				
EE	-0.965	-0.983	0.997	1			
CCR	0.962	0.981	-0.996	-0.999	1		
E_{lumo}	-0.825	-0.794	0.708	0.671	-0.664	1	
Q_{CL}^{+}	0.847	0.862	-0.863	-0.853	0.851	-0.653	1

intermolecular (covalent) interactions between PCDD/F and octanol molecules, and thus the greater the logK_{OA} values. (III) Increasing Q_{Cl}^{+} values of the PCDD/F molecules leads to increasing logK_{OA} values, which implies possible intermolecular electrostatic interactions between PCDD/F molecules and octanol molecules, with the chlorines in PCDD/F molecules to accept electrons and the oxygen atoms in octanol molecules to donate electrons.

Acknowledgments: The study was supported by the Teaching and Research Award Program for Outstanding Young Teachers in Higher Education Institutions of MOE (TRAPOYT), P. R. China, and the State Key Laboratory of Freshwater Ecology and Biotechnology (2000FB12), P. R. China. The research results were attained with the assistance of the Alexander von Humboldt (AvH) Foundation and we thank the AvH foundation. We thank Dr. Tom Harner from Meteorolgical Service of Canada, Environment Canada, for providing us the data and reviewing the manuscript.

REFERENCES

Fossi MC, Casini S, Marsili L (1999) Nondestructive biomarkers of exposure to endocrine disrupting chemicals in endangered species of wildlife. Chemosphere 39:1273-1285

Harner T, Green NJL, Jones KJ (2000) Measurements of octanol-air partition coefficients for PCDD/Fs: A tool in assessing air-soil equilibrum status. Environ Sci Technol 34: 3109-3114

Kaliszan R (1993) Quantitative structure retention relationships applied to reversed-phase high-performance liquid chromatography. J Chromatogr 656A:417-435

Luco JM (1999) Prediction of the brain-blood distribution of a large set of drugs from structurally derived descriptors using partial least-squares (PLS) modeling. J Chem Inf Comput Sci 39: 396-404

Nebert DWC (1989) The Ah locus: genetic differences in toxicity cancer mutations and birth defects. CRC Crit Rev Toxicol 20:153-174

Stewart JJP (1989a) Optimization of parameters for semiempirical methods Ⅰ. Method . J Comp Chem 10: 209-220

Stewart JJP (1989b) Optimization of parameters for semiempirical methods Ⅱ. Applications. J Comp Chem 10: 221-264

Wold S, Wold H, Dunn W J III (1984) Report UMINF-83, Department of Chemistry, University of Umeå, Sweden

Safe SH (1986) Comparative toxicology and mechanism of action of polychlorinated dibenzo-*p*-dioxins and dibenzofurans. Ann Rev Pharmacol Toxicol 26: 371-399

Younes M (1999) Specific issues in health risk assessment of endocrine disrupting chemicals and international activities. Chemosphere 39:1253-1257