Atmospheric Chemistry: Rate Constants of the Gas-Phase Reactions Between Haloalkanes of Environmental Interest and Hydroxyl Radicals

C. Chiorboli,[a] R. Piazza,[a] M.L. Tosato[b] and V. Carassiti[a]

[a]Centro di Fotoreattivitá e Catalisi CNR, Dipartimento di Chimica dell'Universitá, via L. Borsari 46, 44100 Ferrara, Italy.

[b]Istituto Superiore di Sanitá, Structure Activity Research Group, viale Regina Elena 299, 00161 Roma, Italy.

Abstract

The rate constant (k_{OH}) of the gas-phase oxidation of hydrohalo-methanes and -ethanes by hydroxyl radical (OH) has been modelled in terms of eleven qualitative descriptors drawn from the chemical structure. The multivariate model was calibrated on a training set of 16 compounds selected by a statistical design out of the investigated range of structures. It was experimentally validated by comparing the predicted with the actual k_{OH}.s available for 15 additional compounds. The standard deviation of the prediction errors in the reactivity range of the series ($-14.77 \leq \log(k_{OH}) \leq 12.40$ in our study) was found to be ± 0.24.

1. INTRODUCTION

Several and sometimes poorly understood species and processes play a role in atmospheric chemistry and together ensure the maintenance of vital equilibria to the biotic systems. They also provide, to some extent, a sink for hazardous atmospheric pollutants which may modify such equilibria. For example, it is generally agreed that a dominant loss process for organics is their oxidation by OH radical [1]. The knowledge of the rate constant (k_{OH}) of this process is therefore essential in order to evaluate the atmospheric lifetimes, hence the environmental compatibility of pollutants.

To date, experimental k_{OH}.s are available for a limited number of organics. Nevertheless, such data, if properly utilized, may permit to develop empirical models, such as QSARs (Quantitative Structure-Activity Relationships) enabling to estimate the rate constant of non-tested compounds within a series of congeneric compounds.

In previous studies, QSARs based on statistical design and multivariate analysis have been developed in which the k_{OH} of haloalkanes was expressed in terms of

0010-8545/93/$24.00

structural descriptors of various type [2,3]. In this paper we report the results of additional studies focussed on the series of hydro-halogenated methanes and ethanes containing fluorine, chlorine and/or bromine atoms, a series of compounds which merit special attention. In fact, it includes, together with extremely persistent CFC.s (recently banned or restricted by international agreements), a number of possible, less persistent alternatives. A k_{OH} model based on eleven descriptors which can be directly drawn from the chemical structure has been developed and validated.

2. EXPERIMENTAL

2.1. Chemicals.

The chemicals addressed were the 16 C_1 and the 60 C_2 hydrohaloalkanes containing F, Cl, and/or Br atoms, included in the EINECS [4] list or in the revised Montreal protocol [5] list. A variety of halogenation patterns was represented by the 76 compounds.

2.2. Structural data

The multivariate characterization of the hydrohaloalkanes was carried out by means of eleven discrete variables. These were: six variables indicating the number of carbon, hydrogen, fluorine, chlorine, and bromine atoms and the number of all the halogens together present in the molecules (here-in-after referred to as "atom-counting" variables); plus five variables indicating the number of the following substituents at each HC- group present in the molecule: HC-$\boldsymbol{R_3}$, HC-$\boldsymbol{R_2 X}$, HC-$\boldsymbol{RX_2}$, HC-$\boldsymbol{X_3}$, and HC$\boldsymbol{H_2}$-$\boldsymbol{CX_3}$ (here-in-after referred to as "group-counting" variables). Table 1 provides examples of parametrization according to the atom- and group-counting variables.

2.3. Reactivity Data

The rate constant of the reaction with hydroxyl radical was known for 31 of the 76 compounds taken into consideration. All the parametrized compounds and the available reactivity data expressed in terms of $\log(k_{OH})$ (k = cm^3 $molecules^{-1}$ sec^{-1}), are listed in Table 2.

2.4. Statistical Methods

Principal Component Analysis (PCA) [6] and a modified version of Partial Least Squares (PLS) [7], named Generating Optimal Linear Partial least squares Estimation (GOLPE) [8,9] were the methods used in the data analysis.

Table 1. Atom and group counting variables and examples of parametrization.

	Atom-counting variables						Group-counting variables[a]				
No. of	C	H	Br	Cl	F	ΣX[b]	HC-R_3	HC-R_2X	HC-RX_2	HC-RX_3	HC-H_2CX_3
	(1)	(2)	(3)	(4)	(5)	(6)	(i)	(ii)	(iii)	(iv)	(v)
Examples[c]											
1 CF_3-CH_2F	2	2	0	0	4	4	0	0	1	0	0
2 CHF_2-CHF_2	2	2	0	0	4	4	0	0	2	0	0
3 CH_2Cl-$CHCl_2$	2	3	0	3	0	3	0	1	1	0	0
4 CH_3-CCl_3	2	3	0	3	0	3	0	0	0	0	1
5 $CHCl_2$-CCl_3	2	1	0	5	0	5	0	0	0	1	0

a) R = H or -CH; X = Cl, F, Br atoms or -CX_3 group.

b) ΣX = F + Cl + Br

c) Each CH was assigned as follows to the respective group: 1: $\overset{-}{C}F_3$ - $\overset{iii}{C}H_2F$; 2: $\overset{iii}{C}HF_2$ - $\overset{iii}{C}HF_2$; 3: $\overset{ii}{C}H_2Cl$ - $\overset{iii}{C}HCl_2$; 4: $\overset{v}{C}H_3$ - $\overset{-}{C}Cl_3$; 5: $\overset{-}{C}Cl_3$ - $\overset{iv}{C}HCl_2$;

The selection of the training set compounds was accomplished by a D-optimal design [10], applying the DESDOP routine available in the GOLPE [9] package. The design variables were the Principal Component extracted by means of a PCA from the original 11 structural descriptors.

The relationship between $\log(k_{OH})$ (the y variable) and the structural descriptors (the x variables) was developed using GOLPE. This method identifies the optimal model dimensionality as the one providing the lowest SDEP (Standard Deviation of Error of Prediction, eq. 1) [11], and the highest Q^2 (the complement to the fraction of unexplained variance over the total variance, eq. 2).

$$SDEP = \sqrt{\frac{\Sigma(y_{obs} - y_{pred})^2}{N}} \tag{1}$$

$$Q^2 = 1 - \frac{\Sigma(y_{obs} - y_{pred})^2}{\Sigma(y_{obs} - \bar{y})^2} = 1 - \frac{PRESS}{SS} \tag{2}$$

These parameters are computed at each model dimension using predicted y values (y_{pred}) for compounds held out of the model calibration. These predictions may slightly change depending on the number of held out compounds and on the number of random selection of the compounds included in each held out group.

Table 2. Reactivity data for the investigated compounds.[a]

			$\log(k_{OH})$					$\log(k_{OH})$	
V,T[a]	No	Name	obs.[b]	pred.(calc.)	V,T[a]	No	Name	obs.[b]	pred.(calc.)
V	1	CH_2Cl_2	-12,85	-13,25		39	$CHBr_2$-CHF_2		-13,54
	2	$CHBr_3$		-13,33		40	CF_3-CH_2Br		-13,92
V	3	CH_3Cl	-13,36	-13,22		41	CBr_3-CH_2Br		-13,01
	4	CH_2ClBr		-13,33	T	42	CH_2F-CH_3	-12,64	(-12,75)
T	5	$CHCl_3$	-12,99	(-13,08)		43	CF_2Br-$CHFBr$		-13,83
V	6	CHF_2Cl	-14,33	-13,85		44	CCl_3-$CHFCl$		-12,91
T	7	$CHFCl_2$	-13,52	(-13,47)		45	$CHCl_2$-$CFCl_2$		-12,91
	8	$CHBrCl_2$		-13,16	T	46	CF_3-CHF_2	-14,6	(-14,44)
	9	$CHBr_2Cl$		-13,24		47	$CHFCl$-$CHFCl$		-13,37
T	10	CH_3Br	-13,41	(-13,31)		48	CCl_2Br-CH_2Br		-12,85
	11	CHF_3		-14,23		49	CBr_3-$CHBr_2$		-12,93
V	12	CH_2ClF	-13,36	-13,63		50	$CHBr_2$-CH_2Br		-12,83
	13	CH_2Br_2		-13,41	T	51	CF_3-CH_2Cl	-13,79	(-13,84)
V	14	CH_3F	-13,77	-13,61		52	CH_2Cl-$CClF_2$		-13,45
	15	CHF_2Br		-13,93	V	53	CH_2F-CF_3	-14,07	-14,22
T	16	CH_2F_2	-13,96	(-14,01)	V	54	$CHFCl$-CF_3	-13,99	-14,06
T	17	$CHCl_2$-$CHCl_2$[c]	-12,6	(-12,61)	V	55	CH_3-CCl_2F[d]	-14,14	-14,16
V	18	$CHCl_2$-CH_2Cl	-12,48	-12,59	V	56	CH_2F-CH_2F	-12,96	-13,33
T	19	CH_2Cl-CH_2Cl	-12,66	(-12,57)	T	57	CF_3-CH_3	-14,77	(-14,92)
T	20	CH_2Cl-CH_3	-12,4	(-12,36)	V	58	CH_2F-CHF_2	-13,74	-13,74
	21	CH_2Br-CH_2Cl		-12,65	T	59	CHF_2-CHF_2	-14,28	(-14,14)
T	22	CH_2Br-CH_2Br	-12,6	(-12,73)		60	CH_2Br-CH_2F		-13,03
V	23	$CHCl_2$-CH_3	-12,59	-12,38		61	CCl_2F-$CHClF$		-13,29
	24	$CHBr_2$-$CHBr_2$		-12,93		62	CCl_3-CHF_2		-13,29
	25	CCl_3-$CHCl_2$		-12,52		63	CHF_2-$CFCl_2$		-13,67
T	26	CCl_3-CH_3	-13,92	(-13,77)		64	CHF_2-CF_2Cl		-14,06
V	27	CH_2Br-CH_3[c]	-12,46	-12,44		65	$CHCl_2$-$CHClF$		-12,99
V	28	CF_2Cl-CH_3	-14,45	-14,54		66	CH_2Cl-$CFCl_2$		-13,07
V	29	CHF_2-CH_3	-13,47	-13,15		67	CCl_3-CH_2F		-13,07
	30	CF_2Br-CH_2Br		-13,61		68	CHF_2-$CHCl_2$		-13,37
	31	CF_3-$CHClBr$		-13,75		69	CH_2F-$CFCl_2$		-13,45
T	32	CF_3-$CHCl_2$	-13,47	(-13,67)		70	CHF_2-$CHFCl$		-13,76
	33	CF_2Br-$CHFCl$		-13,75		71	CH_2F-CF_2Cl		-13,84
	34	CF_2Cl-$CHCl_2$		-13,29		72	CH_2Cl-$CHFCl$		-12,97
	35	CF_2Cl-$CHFCl$		-13,67		73	CH_2F-$CHCl_2$		-12,97
T	36	$CHBr_2$-CH_3[c]	-12,6	(-12,55)		74	CH_2Cl-CHF_2		-13,35
	37	CH_2Cl-CH_2F		-12,95		75	CH_2F-$CHFCl$		-13,35
	38	CCl_3-CH_2Cl		-12,69		76	CH_3-$CHFCl$		-12,77

[a] T and V indicate the training and validation set compounds, respectively. [b] From ref. 1, unless otherwise noted. [c] From Ref. 2. [d] From ref. 12.

3. RESULT AND DISCUSSION

3.1. Parametrization

As mentioned in section 2.2 (structural data), 11 discrete variables are used for describing the structural variations in the series. The choice of these variables was based on the consideration that all the hydrohaloalkanes react with the OH radical mainly *via* H-atom abstraction so that the rate constant certainly depends on the C-H bond energy which is neither easily measured, not easily computed. However, it is well known that the C-H bond energy depends on the number of hydrogens, carbons and halogens present in the molecules, and that it is also affected by the substituent groups at each carbon involved in the C-H bond cleavage.

Following this idea, each single compound was parametrized as reported in Table 1: first, all type of elements present in the molecules were identified and counted (variables 1-5); second, all the halogen atoms together were summed (variable 6); and third, five possible substituent groups at each C-H were identified and their number was counted (variables i - v).

3.2. Selection of a Training Set.

3.2.1. Design

The PC analysis of the 76x11 data matrix allowed to contract the original eleven variables to seven orthogonal "latent" dimensions accounting for the 98% of the total variance in the X data.

The seven PC.s were used as the design variables in the D-optimal selection of a training set among the 31 compounds with k_{OH} data. Sixteen compounds were selected in order to have an approximately equal number of compounds distributed between the training and the validation sets (see section 3.3 External Validation of the Model). The D-efficiency [10] of the design is 30%. Because of the use of a statistical design, the training set compounds (see Table 2, compounds labelled with T) ensure the best possible mapping of the x variables space. Furthermore, as it may be noted from their k_{OH} values, they are also well distributed within the y variable space.

3.2.2. Model Development

The model relating the $\log(k_{OH})$ with the eleven structural descriptors was developed by using the GOLPE [9] analysis. The optimal model dimension was identified on analysing the values of the SDEP and Q^2 parameters (equations 1 and 2, respectively) at each model expansion up to the sixth components. The parameters were calculated in two different way: first by the Leave One Out (L.O.O.) procedure, in which each compound, one at time, is held out of the training set, and its $\log(k_{OH})$ value is predicted by the model fitted to the remaining 15 compounds. These predictions were used in the calculation of SDEP and Q^2. Second, by the grouping procedure, in which the training set was initially sub-divided, in our application, into five groups (so that each group included three or four compounds) and thereafter the same procedure as above was applied, with the only difference that one group (instead of one compound) at a time was left out of the training set.

Whenever the grouping procedure is applied, the calculated SDEP and Q^2 values depend, as is obvious, on how the groups are formed. In order to overcome this pitfall, GOLPE includes a routine allowing to repeat the calculation of the two parameters several times, each time starting from different, randomly obtained, grouping of the training compounds. This permits to obtain average, stable values of SDEP and Q^2 together with their respective standard deviation. In our study, the process was repeated after each of 200 randomization of the training set compounds. Table 3 summarizes the results obtained. It also includes the fit parameters SDEC (Standard Deviation of Error of Calculation) and R^2 computed with equations 3 and 4, respectively, in which the term "y_{pred}" of eqs. 1 and 2, is replaced with the "calculated" y values provided by the model "fitted" to all the 16 training set compounds.

$$SDEC = \sqrt{\frac{\Sigma(y_{obs} - y_{calc})^2}{N}} \qquad (3)$$

$$R^2 = 1 - \frac{\Sigma(y_{obs} - y_{calc})^2}{\Sigma(y_{obs} - \bar{y})^2} \qquad (4)$$

Table 3. SDEP and Q^2 values obtained in the L.O.O. and in the grouping procedures

			L.O.O.		5 groups		
Components	SDEC	R^2	SDEP	Q^2	SDEP	SD_{SDEP}[a]	Q^2
1	0,396	0,724	0,541	0,483	0,551	0,035	0,463
2	0,147	0,962	0,351	0,782	0,396	0,068	0,723
3	0,112	0,978	0,255	0,885	0,302	0,077	0,839
4	0,11	0,979	0,231	0,906	0,273	0,072	0,868
5	0,109	0,979	0,245	0,894	0,279	0,067	0,862
6	0,109	0,979	0,254	0,886	0,287	0,068	0,854

[a] Standard deviation of the SDEP parameter.

The results may be commented as follows: a) both the L.O.O. and the "grouping" procedures indicate that the model with the optimal, and apparently rather good, predictive capacity is the four dimensional model (i.e., the one with the lowest SDEP and highest Q^2). However, the two parameters are not significantly different after the third and fourth model expansion. Therefore, since the fourth does not, in practice, increase the fraction of explained variance and, on the contrary, it may just introduce noise, the tridimensional model was selected as the optimal one. With this model: b) the SDEC and R^2 values (0.11 and 0.98, respectively, Table 3) show an excellent fit of the data to the model, as is also demonstrated by the agreement between actual and calculated $\log(k_{OH})$ values for the training compounds (see Table 2); however, fit parameters do not inform about the real predictive capacity of the model. This information is given by the predictivity parameters: c) the SDEP and Q^2 values are not as optimistic; in particular, they show that the grouping procedure gives, as expected, more conservative values than the L.O.O. procedure with respect to the predictive ability. However, the values resulting from the two procedures are not substantially different; this result suggests that the model should be rather stable, hence supports that the selected x variables can properly account for the variations of k_{OH}.

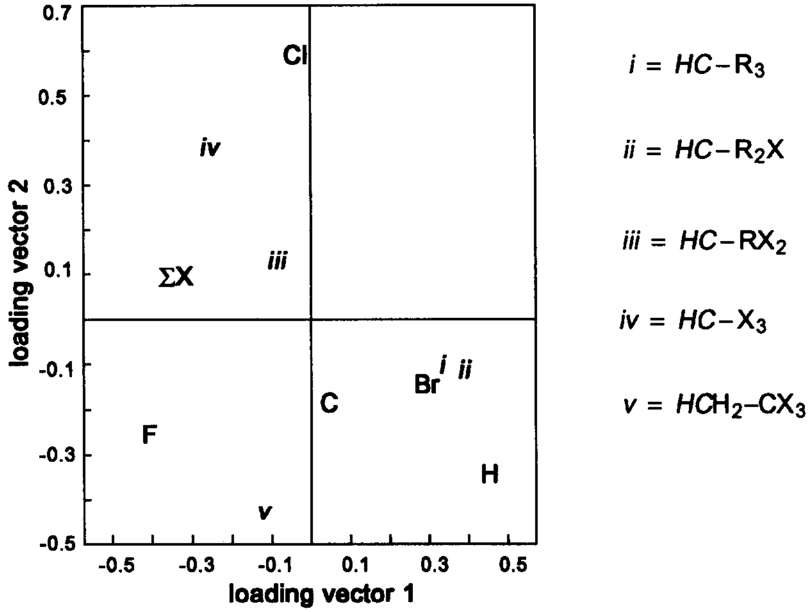

Figure 1. Loading plot with loading vector 2 *vs* loading vector 1 (symbols as in Table 2)

The relative importance of each descriptor (x variable) in explaining the reactivity towards the OH radical of the hydrohaloalkanes can be inferred from the analysis of the loading plot (figure 1) according to the two most important dimensions, i.e. the first and the second, together explaining 96% of the total variance. The plot shows that, in the first dimension, two groups of variables have opposite influence on the variation of log(k_{OH}): the number of fluorines (F), halogens (ΣX) are rate decreasing variables, whereas the number of hydrogens (H), bromines (Br), HC-R_3 (i) and HC-R_2X (ii) groups are rate increasing variables. In the second dimension three variables show to have a remarkable influence on the reactivity of hydrohaloalkanes towards the OH radical: the number of chlorines (Cl), and HC-X_3 (iv) groups enhance the reactivity, while the number of HCH_2-CX_3 (v) groups reduce it.

3.3. External Validation of the Model.

As already mentioned, while the fit parameters do not provide information about the predictive capacity of the model, the predictivity parameters do. However, the SDEP and Q^2 parameters estimate the predictivity only in terms of self consistency

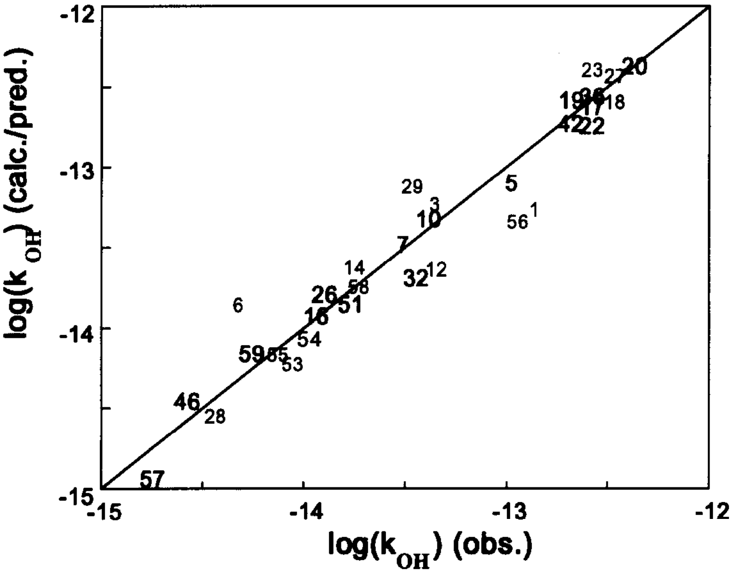

Figure 2. Plot of calculated/predicted vs. observed log(k_{OH}) data (training set compounds in bold).

of the model itself [9]. For this reason, in order to evaluate the actual model predictivity, it seemed appropriate to use an external validation set.

A comparison between the actual and the predicted $\log(k_{OH})$ values for the 15 compounds not included in the training set was, therefore, carried out.

The results obtained are plotted in figure 2 (where the actual *vs* the calculated $\log(k_{OH})$.s for the training set compounds are also included) and listed in Table 2.

The standard deviation and the Q^2 for the validation compounds were 0.24 and 0.87 respectively, indicating that a good agreement really exists between predicted and actual $\log(k_{OH})$ values. It seems worth noting that this agreement is better than that anticipated by the SDEP and Q^2 parameters provided by GOLPE according to the grouping procedure and it is practically the same as anticipated by L.O.O. procedure.

The results obtained in the external validation allowed to extend the predictions to the non tested 45 compounds in our data (the predicted $\log(k_{OH})$ values are reported in Table 2).

4. CONCLUSIONS

This study further supports the effectiveness of statistical design and multivariate analysis in the construction of locally valid QSAR models. These may, in fact, work, provided that they are developed for, and applied to series of chemicals with similar structure and the same mechanism of reactions. It also confirms that GOLPE enables to compute reliable parameters for identifying the optimal model and assessing its predictive capacity.

In our application a good k_{OH} model was obtained according to our validation process. This can mainly be attributed to 1) the homogeneity (structural and mechanistic) of the series; 2) the use of a statistical design in the selection of the training set; and 3) the adequacy of the utilized multivariate descriptors in accounting for the variation of k_{OH}.

Concerning the descriptors used, we note that they can be directly drawn from the chemical structure, enabling to estimate the k_{OH} of not yet synthesized compounds. The model therefore provides, among others, an effective tool for an early inclusion/exclusion from consideration of several possible substitutes for compounds such as those CFCs to be phased out of production and use. Further validation of the model, by testing additional compounds located into so far poorly investigated structural area, seems therefore worth conducting.

5. REFERENCES

1 R. Atkinson, *Chem. Rev.*, 86 (1985) 69.

2 M.L. Tosato; C. Chiorboli; L. Eriksson; J. Jonsson, *Sci. Total Environ.*, 109-110 (1991) 307.

3 M.L. Tosato; R. Piazza; C. Chiorboli; L. Passerini; A. Pino; G. Cruciani; G. Clementi, *Chemom. Intell. Lab. System.*, 16 (1992) 155.

4 European Inventory of Existing Chemical Sustances, Annex 1 to the CEC proposal for EEC Council Regulation on the *Evaluation and Control of the Risk of Existing Chemicals*, (1992).

5 Montreal Protocol Review Meeting, London, June 27-29, (1990).

6 S. Wold, K. Esbensen, P. Geladi *Chemom. Intell. Lab. System.*, 2 (1987) 37.

7 A. Höskuldsson, *J. Chemometrics*, 2 (1988) 211.

8 M. Baroni: Sviluppo e Confronto di Metodi Chemiometrici", Thesis, University of Perugia, (1989).

9 M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi, S. Clementi, *QSAR*, 00 (1992) 000 .

10 T.J. Mitchell, *Technometrics*, 16 (1974) 203.

11 G. Cruciani, M. Baroni, S. Clementi, G. Costantino, D. Riganelli, B. Skagerberg:, *J. Chemom.*, in press.

12 R. Liu, R.E. Huie, M.J. Kurilo, *J. Phys. Chem.*, 94 (1990) 3247.

Acknowledgments

This study has been supported by the CNR - ENEL (National Research Council - National Agency for Electric Energy). Project "Interaction of Energy System with Human Health and the Environment" ROMA, ITALY.

The provision from HIMONT S.p.A. (Ferruzzi Group) of a grant to R. Piazza is also gratefully acknowledged.