Atmos. Chem. Phys. Discuss., 13, C7696–C7702, 2013
www.atmos-chem-phys-discuss.net/13/C7696/2013/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Structure-activity relationship for the estimation of OH-oxidation rate constants of carbonyl compounds in the aqueous phase" *by* J. F. Doussin and A. Monod

J. F. Doussin and A. Monod

jean-francois.doussin@lisa.u-pec.fr

Received and published: 4 October 2013

First of all, we want to warmly thank the reviewers (and the readers who sent direct e-mails) for their positive feedbacks on our work and for their comments. Most ofthe requests for correction or improvement have been taken into account which has certainly helped improving the present paper. Discussion with Reviewer #1 "P. 15951, l. 3: 'Carbonyls': I would like to suggest to always address the compounds in question as 'carbonyl compounds' instead of only 'carbonyls' because in chemistry the latter term is often used to describe metal complexes of CO." It has been replaced through the whole document

"P 15960, l.15: Setting up a SAR estimation for electron transfers might be a good future project, especially for radical other than OH" We totally agree with this statement while, once again, we are afraid that the experimental database for organic compounds in water may be scarce. For the OH radical electron transfer with oxoacids, we have tentatively included in the paper a small remark which indicates "It must be pointed out here that the introduction of a k(C(O)COO) = 2.1 $\times$ 108 M 1 s-1 for reaction 2 and the corresponding use of the factors F(CH3), F(COOH) and F(COO-) would reconcile the calculated and the experimental data for pyruvate, ketomalonate and ketomalonate dianion, respectively. This value is close to 3.9 $\times$ 108 M 1 s-1 which is the rate constant for (. . .) the electron transfer between the carbonate ions and the OH radical".

"P 15964, l. 14ff: As nice as the method is, it might be good to call for some care in its application (. . .) Furthermore, in the 'Conclusions' section I am missing an outlook to other aqueous phase radicals. " A remark in this direction has been added in the conclusion section.

âĂČ Discussion with Reviewer #2

"The model terms are used to predict OH rate constants, which are compared with the input data. As this calibration and testing activity on the same compounds seems like a circular process (albeit a necessary one), it is unsurprising that reasonable agreement is obtained, and also unsurprising that the model outperforms other SAR models on its native dataset." The authors fully agree with the fact that adopting a methodology that would comprises a "training dataset" and a "validation dataset" would have been more robust (two steps process). This approach can only be applied when the database is sufficiently large. This is why significant efforts have been made to produce the largest database possible. To our best knowledge, no kinetic study has been disregarded. Nevertheless, the available dataset contains only 30 different species (i.e. 9 aldehydes, 7 ketones and 14 polyfunctionals). This number (added to the diversity of structures concerned and the diversity of experimental data for some compounds) has been considered too small to allow the splitting of the dataset into a "training subset"

and a “validation dataset”. In fact, it would have been a paradox ending up with a too small number of data to allow a good training and then declaring the validation failed due to that. On the contrary, we have decided to keep all the available data to allow the best “training” possible. While one cannot deny a sort of circularity, the resulting process may sound more circular than it is really. In fact, one must keep in mind that for many species, the value of the calculated rate constant is significantly constrained by the parameters of the first version of this SAR (the 2008 version), and these parameters have not been changed . In addition, only five new descriptors have been added. This means that their values must allow verifying 30 equations. The obtained values must be rationalisable in term of chemical activity. A priori, it was not obvious that the structure of these molecules carries enough chemical information to describe their reactivity. The fact that a set of only five additional parameters allowed the extension of this SAR seems to demonstrate that to some extent. “It is (. . .) also unsurprising that the model outperforms other SAR models on its native dataset. The authors should justify to what extent these comparisons between SAR models are fair, given the overlap or lack of overlap of data used to build and test each model. “ In our mind, the match percentages and other statistic parameters we provide are only quantitative means to describe the performance of the SAR we propose. We would be sorry if the paper would give the impression of any kind of contest between SAR models (we have carefully reviewed the paper to track any sentence that could mislead the reader in this direction). Going too deeply into the comparison between the four SARs we mention would probably fail due to methodological differences. Concerning Monod et al (2005) SAR, it is discussed in the manuscript that indeed the “training dataset” was extremely limited (i.e. 8 compounds for 8 parameters determined) and that this is certainly the reason of its poor performance. Concerning Ervens et al (2003) SAR, the dataset is limited to the compounds for which both the BDE and the kOH are available. For Minakata et al (2009), the dataset used is certainly very close to ours as it has been built upon an extensive literature search of the kOH experimental values (as we did) and as all the references we used but one (Gligorowski et al, 2009) were available when they submitted the

paper. The difference in the dataset is hence rather methodological as they indicate “When several rate constants were reported for the same compound, an average value was used or the most reasonable rate constant was selected (. . .)”. In our case, we have always used all the experimental determinations for the “training process”. As it can be seen It is not easy to distinguish what in the performance of the SAR is coming from the extent of dataset used, the selection principle, and available rate constants at the time of the SAR building or the physico-chemical basis of the SAR. . . and this was not really our purpose. Rather, we wanted to provide a description of the various SARs with common basic methodology (i.e comparing the calculated values with all the available experimental rate constants). “Abstract / p. 15959 line 6: The authors should clarify what compounds they are including in the reported percentages of compounds where model predictions match measurements within a certain tolerance. In the abstract, they appear to be reporting a number based on results for all compounds used to develop the model (alcohols, acids, bases, alkanes, carbonyls, multifunctional), giving 58% of compounds matching within 20%. However, the more important result to include in the abstract is for the carbonyls and multifunctionals alone, the focus of this study. (On p. 15959 it is reported that only 41% of these compounds have predictions that match the measurement data within +/-20%.) It seems to me that the success of the previous version of the model is inflating the perceived success of the current extension when non-carbonyl compounds are included. ”

Our idea was more taking the resulting SAR as a whole and not inflating performance. Nevertheless, as the focus of the paper is about carbonyls, the reviewer is right and for coherence purpose, this has been added in the abstract. Nevertheless, while the reviewer is right for the “20% match area” where one obtains 41% for the carbonyls dataset in place of 58% for the whole dataset, for the “40% match area”, the difference in performances is not as large : 72% versus 78%.

Abstract / p. 15959 line 6: It would be helpful to give these types of results for ketones + aldehydes and for polyfunctional separately so that the reader can best judge the

utility of the method. The available database contains only 30 different species (i.e. 9 aldehydes, 7 ketones and 14 polyfunctionals). Providing match percentage for the whole dataset is barely correct as 30 is a small number and we were even more reluctant to provide similar percentages for each class of compounds as it does not seem meaningful for such small populations.

"p. 15959 line 15: Why is a limited diversity of structures a problem for model accuracy if the same structures used to develop the model are then used to validate it? Wouldn't this actually improve the model performance on these structures?" Again, the idea is to keep the maximum chemical diversity in the "training dataset" to have the best possible descriptors determination. To be more precise one can say that a limited diversity of structures available during the training is a problem for the future accuracy of the models.

"p. 15954 line 21: Cyclic descriptor terms are mentioned here, yet I can find no other mention or listing of such terms in the manuscript. Were these terms part of the previous model?" Yes, cyclic descriptors were part of the previous version of this SAR. As all the other parameters from Monod and Doussin, 2008, they have not been updated. Figure 2: It appears that the rate constants for ketones and aldehydes are slightly but systematically underpredicted, while those for polyfunctionals are overpredicted. The manuscript should confirm and comment on these effects, if they are real.

These effects are partly real: additional statistics performed on these data (see Table below – statistic for individual chemical families - In bold are indicated when the correlations are significantly different from the 1:1 slope) show that ketones are well predicted, aldehydes and polyfunctionals are slightly underpredicted. However, these statistics are based on a few number of data ($\leq$ 25 points), which are very sensitive to the dispersion of experimental data, this dispersion being increasingly important from ketones, aldehydes to polyfunctionals.

Furthermore, we want to emphasize the fact that the parameters of this SAR were

determined on the whole database, and not for each chemical family separately. Therefore, the slight underpredictions observed for aldehydes and polyfunctionals cannot be interpreted in terms of misdetermination of parameters. Rather, they should be interpreted in terms of the need for further experimental data, especially for aldehydes and polyfunctional compounds, which are the most tricky compounds to experimentally handle. As an illustration of the sensitivity of the correlation values taken chemical family by chemical family, it must be shown that disregarding a single experimental data point (obviously different from previous determinations) such as the latest values for 3-Hydroxybutan-2-one would bring the slope of the correlation line from 0.68 to 0.79.

Please also note the supplement to this comment:
http://www.atmos-chem-phys-discuss.net/13/C7696/2013/acpd-13-C7696-2013-supplement.pdf

---

Interactive comment on Atmos. Chem. Phys. Discuss., 13, 15949, 2013.

	intercept	slope	Statistics		
	value	value	R^2	Number of points	Degree of freedom
All	(-4.4±14.0) x 10^7	1.01±0.06	0.77	69	67
Ketones	(-1.5±2.0) x 10^8	1.05±0.10	0.85	21	19
aldehydes	**(8.6±5.7) x 10^8**	**0.75±0.20**	0.36	25	23
Polyfunctionals	(2.5±11.2) x 10^7	**0.68±0.10**	0.66	27	25

Fig. 1. statistic for individual chemical families - In bold are indicated when the correlations are significantly different from the 1:1 slope