CrossMark

RESEARCH ARTICLE

The conformation-independent QSPR approach for predicting the oxidation rate constant of water micropollutants

Erlinda V. Ortiz[1] · Daniel O. Bennardi[2] · Daniel E. Bacelo[3] · Silvina E. Fioressi[3] ·
Pablo R. Duchowicz[4]

Abstract In advanced water treatment processes, the degradation efficiency of contaminants depends on the reactivity of the hydroxyl radical toward a target micropollutant. The present study predicts the hydroxyl radical rate constant in water (k_{OH}) for 118 emerging micropollutants, by means of quantitative structure-property relationships (QSPR). The conformation-independent QSPR approach is employed, together with a large number of 15,251 molecular descriptors derived with the PaDEL, Epi Suite, and Mold2 freewares. The best multivariable linear regression (MLR) models are found with the replacement method variable subset selection technique. The proposed five-descriptor model has the following statistics for the training set: $R^2_{train} = 0.88$, $RMS_{train} = 0.21$, while for the test set is $R^2_{test} = 0.87$, $RMS_{test} = 0.11$. This

QSPR serves as a rational guide for predicting oxidation processes of micropollutants.

Introduction

Determining the presence of organic micropollutants in the aquatic environment at the low-nanogram per liter level is considered of great concern, as the risk these compounds have on the environment and human health is still inconclusive (Luo et al. 2014; Sudhakaran et al. 2012).

Although conventional treatment processes have observed insufficient removals of many micropollutants from drinking water, advanced technologies have shown great abilities to degrade or remove many of these micropollutants. In particular, hydroxyl radical-based advanced oxidation processes (AOPs) are effective means for degrading micropollutants during drinking water treatment (Bagheri and Mohseni 2015; Jin et al. 2012). Hydroxyl radicals react relatively non-selectively with organic contaminants (Sudhakaran and Amy 2013) and can be generated by various combinations of reactants such as UV/H_2O_2, O_3/H_2O_2, Fenton/photo-Fenton, and UV/TiO_2.

During oxidative water treatments, the transformation efficiency of micropollutants depends on the reactivity of the hydroxyl radical toward a target micropollutant. Therefore, rate constants are needed to predict the extension to which micropollutants are degraded after a specified treatment duration (Jin et al. 2012; Lee and Gunten 2012). Nowadays, kinetic data are available for a large number of compounds for the reaction with hydroxyl radicals (Buxton et al. 1988). However, due to the

✉ Pablo R. Duchowicz
pabloducho@gmail.com

1 IMCoDeG (CONICET), Facultad de Tecnología y Ciencias Aplicadas, Universidad Nacional de Catamarca, Maximio Victoria 55, Catamarca, Argentina

2 Cátedra de Química Orgánica, Facultad de Ciencias Agrarias y Forestales, Universidad Nacional de La Plata (UNLP), 60 y 119, B1904AAN La Plata, Argentina

3 Departamento de Química, Facultad de Ciencias Exactas y Naturales, Universidad de Belgrano, Villanueva 1324, CP 1426 Buenos Aires, Argentina

4 Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), CONICET, UNLP, Diag. 113 y 64, C.C. 16, Sucursal 4, 1900 La Plata, Argentina

complexity of analytical methods and the high experimentation cost, there is still a gap for kinetic data of emerging micropollutants (Jin et al. 2012; Sudhakaran et al. 2012).

Predictive models from the quantitative structure-property relationships (QSPR) theory (Benfenati 2013; Hansch and Leo 1995; Puzyn et al. 2010; Roy 2015) are a fast and cost-effective alternative to experimental evaluation. By means of the QSPR technique, an experimental property of a chemical compound, i.e., the reaction rate constant with hydroxyl radical (k_{OH}), can be predicted through the knowledge of its chemical structure. The structure is quantified by means of a set of suitable molecular descriptors, in other words, numerical quantities carrying specific information on the constitutional, topological, geometrical, hydrophobic, and/or electronic aspects (Diudea 2001; Katritzky and Goordeva 1993; Todeschini and Consonni 2009). Therefore, a set of descriptors is statistically correlated to the experimental property, resulting in a mathematical model that can be used to find out useful structure-property parallelisms.

The application of the QSPR theory to water treatment studies is relatively new, and only a few QSPR models have been proposed using different approaches for ozonation, chlorination, AOPs, membrane filtration, and activated carbon adsorption (Borhani et al. 2016; Delgado et al. 2012). In some k_{OH} modeling studies of organic compounds, Hammett-type relationships have been applied; the main disadvantage of such method is only applicable to substituted aromatic compounds with known substituent constants (Lee and von Gunten 2012; Peres et al. 2010; Zimbron and Reardon 2005). A few QSPR studies based on descriptor selection techniques involve advanced statistical methods, like principal component analysis (PCA) and genetic algorithms (Kusic et al. 2009; Sudhakaran and Amy 2013; Toropov et al. 2012). The application of group contribution methods for k_{OH} prediction (Minakata et al. 2009; Monod and Doussin 2008) faces the problem of availability of data for all possible functional groups and the assumption of additivity of rate constants. All such established QSPR would not be specific to the new emerging micropollutants having diverse molecular structures.

In a recent QSPR study reported by Jin et al. (Jin et al. 2015), the authors predict the reaction rate constant with hydroxyl radical (k_{OH}) in water of 118 emerging micropollutants, composed of endocrine disruptor chemicals (EDCs) and pharmaceuticals and personal care products (PPCPs). A special attention is paid on model validation, applicability domain analysis, and mechanistic interpretation. The dataset is randomly split into a training set of 89 compounds for model calibration and a validation set with 29 compounds for testing its predictive capability. The multivariable linear regression (MLR) method analyzes 951 0D-3D Dragon molecular descriptors through the forward stepwise procedure. The best QSPR established includes seven descriptors related to the

electronegativity, polarizability, and double bonds of the compounds. With outliers identified and removed, the final model fits the training set very well and shows good robustness and internal predictivity.

In this work, we report a new alternative QSPR model for the k_{OH} parameter in the same molecular set studied by Jin et al. (2015), however, using the conformation-independent QSPR approach which does not consider the conformational representation of the chemical structure, by only relying on its constitutional and topological representations. It is to be noticed that this method cannot be defined as "geometry independent," because also the 2D descriptors used depend on the geometry (the molecular graph), not on the conformation: the two concepts are different.

The conformation-independent QSPR approach emerges as an important alternative methodology where the conformational representation of the chemical compounds is not considered (Aranda et al. 2016; Duchowicz et al. 2014; Duchowicz et al. 2012; Duchowicz et al. 2015). Therefore, neither experimental information on the X-ray crystal structure of the compound conformations is required nor the geometry optimization calculation of their molecular structure. The exclusion of such 3D-structural aspects avoids ambiguities due to the existence of compounds in various conformational states, which would also lead to the loss of predictive capability of the QSPR model.

Materials and methods

Experimental dataset

The 118 experimental k_{OH} [$M^{-1}s^{-1}$] values of emerging micropollutants collected from the literature (Jin et al. 2015) range in the interval (5.4 10^7, 1.7 10^{10}) and for modeling purposes are converted into decimal logarithmic units (logk_{OH}). The micropollutants include many EDCs and PPCPs and are highly heterogeneous with different chemical classes, such as phenols, polycyclic aromatic hydrocarbons, alkanes, halogenated aromatic compounds, and organophosphorus compounds. The complete list of compounds studied here appear in Table 1S as Supplementary material.

Structural representation and molecular descriptor calculation

The molecules are drawn in MDL mol (V2000) format with ACDLabs ChemSketch freeware (2016). All file format conversions are performed with Open Babel for Windows (2016).

The conformation-independent molecular descriptors are computed as follows. First of all, we use the Pharmaceutical Data Exploration Laboratory (PaDEL) freeware version 2.20 (2016), because it has the advantage that it is a freely available

and open source program. PaDEL allows us to calculate 1444 0D-2D descriptors and nine fingerprint types (13,020 bits) (Yap 2011), with molecules in MDL mol (V2000) format. Additional molecular descriptors are also calculated with the molecular descriptors from 2D structures (Mold2) freeware (Hong et al. 2008), which generates 779 1D and 2D structural variables with molecules in MDL sdf format. Finally, eight semiempirical descriptors are calculated (in decimal logarithmic units) from the Estimation Program Interface (EPI Suite 4.11, 2016) freeware modules, with molecules in SMILES format. EPI Suite uses a series of group contribution factors for calculating: (i) the octanol/water partition coefficient $\log K_{ow}EPI$; (ii) the water solubilities $\log S_{w1}EPI$ and $\log S_{w2}EPI$: the second parameter is based on $\log K_{ow}EPI$; (iii) the Henry's law constant at 25 °C $\log K_H EPI$; (iv) the soil sorption partition coefficients $\log K_{oc1}$ and $\log K_{oc2}$: the first parameter is based on the first order molecular connectivity index, while the second one is based on $\log K_{ow}EPI$; (v) the octanol-air partition coefficient $\log K_{oa}$, based on the ratio $K_{ow}EPI$ and the dimensionless $K_H EPI$; and (vi) the bioconcentration factor $\log BCF$.

Therefore, the total number of non-conformational molecular descriptors explored in this work is 15,251. It is our intention to capture, with such a great number of descriptors, the most relevant structural characteristics affecting the studied property.

Model development

Molecular descriptor selection in MLR

The 15,251 non-conformational molecular descriptors calculated with PaDEL, Mold2, and EPI Suite are analyzed in order to remove the "collinear" descriptors, that is to say, the linearly dependent pairs are identified, and only one variable from each pair is kept for further analysis. This process leads to a set containing 2899 linearly independent 0D-2D descriptors.

We employ the replacement method (RM) technique (Duchowicz et al. 2006) in order to generate MLR models on the training set, by searching in a pool having $D = 2899$ descriptors for optimal subsets containing d descriptors (d is much lower than D), with smallest values for the standard deviation (S_{train}) or the root mean square error (RMS_{train}).

The main idea behind the RM is that one can approach the minimum of S_{train} by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a set of d descriptors. In other words, we should find the global minimum of $S_{train}(d)$ in a subspace of $D\,!\,/[d\,!\,(D-d)!]$ points d, where D represents the total number of available descriptors. The quality of the results achieved with this technique approaches that obtained by performing an exact (combinatorial) full search of molecular descriptors although, of course, requires much less computational work. This

technique has been shown to compare favorably with the genetic algorithm approach (Morales et al. 2006).

Table 2S includes a list of mathematical equations involved in the present study. All the MATLAB-programmed (Matlab 7.6.0.324, 2008) algorithms used in our calculations are available upon request.

Model validation

The best choice for validating a QSPR model consists on testing its ability to predict the property of compounds not considered during the model development, by comparing such predictions with the experimental values. For this purpose, the complete molecular set is split into training (train) and test (test) sets. The training set is used to calibrate the model and to obtain its optimized parameters, while the test set includes compounds "never seen" during the calibration and demonstrates the true predictive power.

It is known that randomly splitting the compounds into the training and test sets does not lead to a rational selection, as such sets should have similar structure-property relationships. For this purpose, the split of the dataset is carried out by means of the balanced subset method (BSM) (Rojas et al. 2015); a procedure developed in our group that ensures that the training set is representative of the test set. The BSM is based on the k-means cluster analysis (k-MCA) method (Kaufman and Rousseeuw 2005): the essence of k-MCA is to create k-clusters or groups of compounds, in such a way that compounds in the same cluster are very similar in terms of distance metrics (i.e., Euclidean distance), and compounds in different clusters are very distinct.

In addition, we also partition the dataset with the property sampling method (PSM) (Leonard and Roy 2006) as done by Jin et al. (2015), in order to compare such results with those obtained with our BSM. In the PSM, the compounds are sorted according to their descending experimental property values, then taking one compound out of every four compounds. Compounds taken out are used as the test set and the remaining ones as the training set.

The linear regression models are theoretically validated through leave-one-out cross-validation (loo) (Golbraikh and Tropsha 2002). According to the specialized literature, the loo explained variance (R^2_{loo}) should be greater than 0.5 for a validated model, although this is a necessary but not sufficient condition for its predictive power. A more rigorous leave-30%-out cross-validation (l30%o) is also practiced on the obtained linear model (200,000 cases).

We also validate our established QSPR models with the newly proposed mean absolute error (MAE)-based criteria (Roy et al. 2016). These authors provide some useful guidelines for determining the quality of predictions based on the MAE parameter and its standard deviation computed from the test set predictions after omitting 5% high residual data points,

in order to obviate the influence of any rarely occurring high prediction errors that may significantly affect the quality of predictions for the whole external test set. It is considered that an error of 10% of the training set range should be acceptable, while an error value more than 20% of the training set range should be a very high error.

Finally, we scramble the experimental property values with Y-randomization (Rücker et al. 2007) and 10,000 cases, as a way of checking that the model is not a result of chance correlation when RMS^{rand} is greater than RMS_{train}.

Applicability domain

A predictive QSPR model is only able to predict molecules falling within its applicability domain (AD), so that the predicted property is not a result of substantial extrapolation (unreliable prediction) (Gramatica 2007; Roy et al. 2015; Jaworska et al. 2005). The AD definition is dependent on the model's descriptors and the experimental property. In this way, through the AD definition, it is possible to confirm the validity of the developed models for the training and test set compounds and to identify outliers.

In this work, we determine the AD through two alternative methodologies. The first one is based on the well-known leverage approach (Eriksson et al. 2003), where a test set compound i must have a calculated leverage h_i smaller than the warning leverage h^*. The second one is based on a simple standardization approach (Roy et al. 2015): a given test set compound i having d standardized descriptor values s_{ik}, $k = 1, \ldots, d$ must have a maximum value $s_{ik}^{max} \leq 3$. In the case that $s_{ik}^{max} > 3$ and its minimum value $s_{ik}^{min} < 3$, then the s_i^{new} parameter has to be calculated and must fulfill the condition: $s_i^{new} = \langle s_i \rangle + 1.28 \cdot \sigma_{s_i} \leq 3$, where $\langle s_i \rangle$ is the mean of s_{ik} values for i and σ_{s_i} is the standard deviation for such values.

Importance of model descriptors

In order to find out the relative importance of the jth descriptor in the linear model, the regression coefficients are standardized (b_j^s, see Table 1S). The larger is the absolute value of b_j^s, the greater is the importance of such descriptor (Draper and Smith 1981).

Results and discussion

We split the dataset into training and test sets with the first partitioning method, the PSM as originally used by Jin et al. (2015). About 25% of the total dataset is used for the validation set (29 molecules). Then, the most representative molecular descriptors of the training set are searched through the RM technique. The best MLR models based on 1–7 structural features are listed in Table 1, while a brief description of the descriptor's meanings is supplied in Table 3S.

From Table 1, it is appreciated that the RMS_{train} parameter continuously improves with the addition of molecular descriptors to the linear equation, a typical behavior in variable subset selection, but RMS_{test} does not significantly improve beyond the number of four descriptors. In order to keep the model's size as small as possible, we select such model as the best linear regression QSPR found with the PSM partition. This non-conformational four-descriptor model has an acceptable statistics: $R_{train}^2 = 0.89$, $R_{test}^2 = 0.77$, $RMS_{train} = 0.17$, and $RMS_{test} = 0.34$ and can also be favorably compared to the previous reported conformational-dependent seven-descriptor model, achieved with the forward stepwise procedure (Jin et al. 2015): $R_{train}^2 = 0.81$, $R_{test}^2 = 0.79$, $RMS_{train} = 0.22$, and $RMS_{test} = 0.31$. In terms of model's size, our equation has a better quality.

It is our intention now to improve the present QSPR study by applying the second partitioning method of BSM. We apply the k-MCA-based procedure for splitting the dataset into $N_{train} = 89$ and $N_{test} = 29$ compounds (refer to Table 1S), thus ensuring a design of balanced molecular sets. The cluster centroid locations in terms of descriptor values are provided as a 87×117 matrix in the c1.txt file (Supplementary material).

The results shown in Table 1 clearly reflect that the new training set of 91 compounds from the BSM represents better the dataset than the training set from PSM, as RMS_{test} is lower than RMS_{train} for BSM contrary to PSM. This means that the training set molecules are able to explain the structure-property behavior for the test set molecules. In Table 1, the selected QSPR involves the following five descriptors:

$$\log k_{OH} = -0.43\ AATS0e + 1.50\ SIC5 - 0.91\ PC200 - 0.33\ KC3738 - 0.03\ AC.CX7 + 11.91 \tag{1}$$

$N_{train} = 89$, $R_{train}^2 = 0.88$, $RMS_{train} = 0.21$
$R_{ij\ max}^2 = 0.11$, $o\ 2.5 = 0$, $R_{rand}^2 = 0.37$, $RMS^{rand} = 0.48$
$R_{loo}^2 = 0.85$, $RMS_{loo} = 0.24$, $R_{l30\%o}^2 = 0.74$, $RMS_{l30\%o} = 0.32$
$N_{test} = 29$, $R_{test}^2 = 0.87$, $RMS_{test} = 0.11$

In this equation, R_{ijmax} denotes the maximum correlation coefficient between descriptor pairs, while $o2.5$ indicates the

number of outlier compounds in the training set having a residual (difference between experimental and calculated property) greater than 2.5 times RMS_{train}.
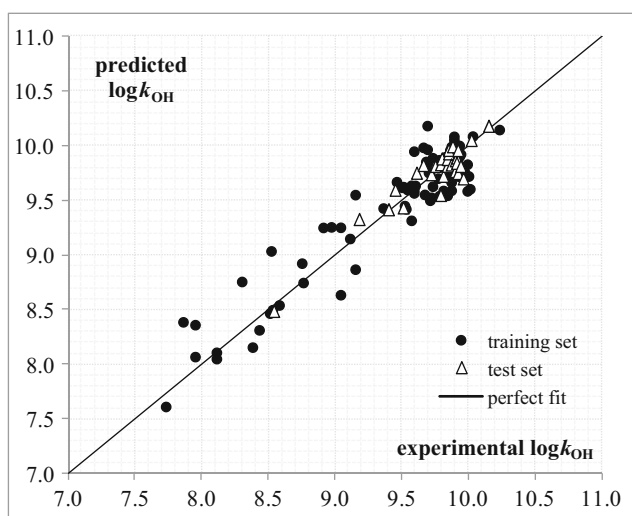
Equation 1 obtained with the BSM partitioning leads to an improved QSPR model for predicting $\log k_{OH}$; a plot for the predictions as a function of the experimental values is

Table 1 The multidimensional QSPR obtained with dataset partitions based on a the property sampling method and b the balanced subset method. The best models are in bold

PSM splitting

d	Descriptors	R^2_{train}	R^2_{test}	RMS_{train}	RMS_{test}
1	BIC5	0.62	0.77	0.31	0.33
2	BIC5; K2925	0.74	0.81	0.26	0.30
3	Sub295; K2925; A.CC3	0.86	0.74	0.19	0.36
4	**Sub295; K2925; A.CC3; D473**	**0.89**	**0.77**	**0.17**	**0.34**
5	Sub295; K2925; A.CC3; KC3587; AC.OX3	0.91	0.76	0.15	0.34
6	Sub295; K2925; A.CC3; A.CO4; AC.OX3; AC.OX4	0.93	0.76	0.13	0.34
7	Sub295; K312; K2925; A.CC3; A.CO4; AC.OX3; AC.OX4	0.94	0.76	0.12	0.34
7	Published seven-descriptor model (Jin et al. 2015)	0.81	0.79	0.22	0.31

BSM splitting

| d	Descriptors	R^2_{train}	R^2_{test}	RMS_{train}	RMS_{test}
1	BIC5	0.67	0.67	0.35	0.20
2	SIC5; D146	0.77	0.84	0.29	0.13
3	AATS0e; SIC5; KC3738	0.83	0.84	0.25	0.12
4	SIC5; KC3738; D146; D361	0.87	0.84	0.22	0.12
5	**AATS0e; SIC5; PC200; KC3738; AC.CX7**	**0.88**	**0.87**	**0.21**	**0.11**
6	AATS0e; SIC5; K2925; KC3738; AC.OX5; D361	0.91	0.88	0.18	0.11
7	Mse; SIC5; MACCS124; K2925; A.PCl4; KC3738; D361	0.93	0.89	0.16	0.11

provided in Fig. 1. The dispersion plot of residuals in Fig. 2 tends to obey a random pattern around the zero line, suggesting that Eq. 1 predicts the whole dataset without systematic errors or residual bias.

The QSPR represented by Eq. 1 has an acceptable predictive power on the external test set of 29 water micropollutants according to R^2_{test} and RMS_{test} parameters. Such model approves the internal validation process of leave-one-out cross-validation through the exclusion of one molecule at a time.

The Y-randomization technique demonstrates that the model has $RMS_{train} < RMS^{rand}$ and $R^2_{rand} < R^2_{train}$ and that a valid structure-logk_{OH} relationship is established. The recommended external validation criteria (Golbraikh and Tropsha 2002) to assure predictive capability are also achieved: $1-R^2_0/R^2_{test}$ $(0.0046) < 0.1$ or $1-R'^2_0/R^2_{test}(0.0058) < 0.1$; $0.85 \leq k(1.0005) \leq 1.15$ or $0.85 \leq k'(0.9994) \leq 1.15$; $R^2_m(0.82) > 0.5$.

The prediction performance of our QSPR model on the test set is found to be "good" by the MAE-based criteria, as it is satisfied the following condition: $MAE \leq 0.1 \, tr$ and $MAE +$

Fig. 1 Predicted and experimental logk_{OH} values according to the QSPR of Eq. 1
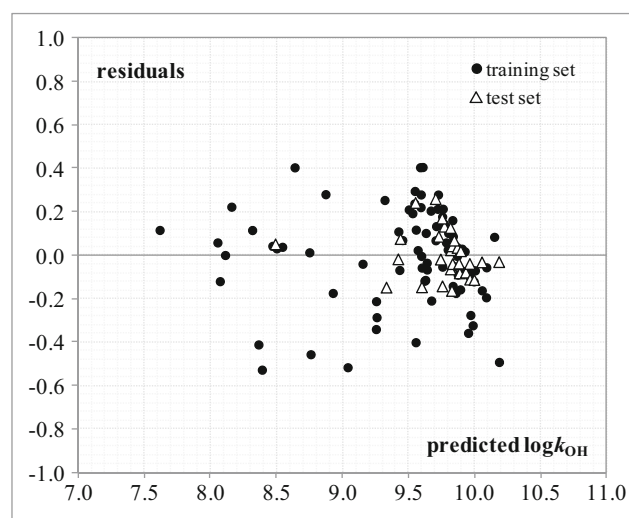
Fig. 2 Dispersion plot of residuals for Eq. 1

$3\sigma \leq 0.2$ *tr*, where *tr* is the training set response range (2.5 logarithmic units) and the σ value denotes the standard deviation of the absolute error values for the test set data. In present case, our QSPR satisfies this condition not only for the whole test set ($MAE = 0.0892$ $\sigma = 0.0661$) but also after omitting 5% high residual data points ($MAE(95\%) = 0.0771$, $\sigma(95\%) = 0.0495$).

The five conformation-independent molecular descriptors appearing in the proposed quantitative structure-OH oxidation rate relationship are readily calculated from the molecular structure, and such variables belong to different classes (Diudea 2001; Katritzky and Goordeva 1993; Todeschini and Consonni 2009):

- An autocorrelation of the topological structure descriptor: *AATS0e*, the average Broto-Moreau autocorrelation-lag 0/ weighted by Sanderson electronegativities. The structural variables introduced by Broto-Moreau are bidimensional autocorrelations between atom pairs (i, j) in a molecule, with the main purpose of capturing the degree of interaction between them. The nature of atoms is considered through a given property as atomic weight (w), i.e., atomic mass, polarizability, electronegativity, or volume. These indices are calculated from the graph by summing products of terms $w_i \cdot w_j$ including terminal atomic contributions in all the paths of a prescribed length (lag). For the case of *AATS0e*, the variable considers in its calculation the atomic composition and the Sanderson electronegativity.

- An information content descriptor: *SIC5*, the structural information content index (neighborhood symmetry of five-order). The information theory descriptors measure the molecular complexity as the diversity of elements present in the structure, such as atoms, bonds, cycles, symmetry, and centricity. In the present case, the descriptor expresses the five-order neighborhood symmetry for all the vertexes in the chemical graph.

Also, the next descriptors have a straightforward structural interpretation:

- A 2D atom pair fingerprint descriptor: *AC.CX7*, the count of C-X at topological distance 7, where X is an halogen atom (Cl, Br, and I).
- A Pubchem fingerprint descriptor: *PC200*, the presence of greater than or equal to 4 (saturated or aromatic) carbon-only ring size 6.
- A Klekota-Roth fingerprint descriptor: *KC3738*, indicating the count of the SMARTS pattern CCCCOC=O.

The molecular descriptors of Eq. 1 have positive numerical values, and thus, the sign of their regression coefficients in the linear model determines whether their contributions increase or decrease the predicted $\log k_{OH}$ values. High numerical values of *SIC5* and low values for *AATS0e*, *PC200*, *KC3738*, and *AC.CX7* tend to lead to high predicted $\log k_{OH}$ values. After standardization, the most important descriptors for predicting the hydroxyl radical rate constants of organic micropollutants are *SIC5* ($b_j^s = 0.67$) and *AATS0e* ($b_j^s = 0.45$), whose numerical values change most in accordance with the numerical variations of the experimental property (Table 2). The remaining descriptors *KC3738* ($b_j^s = 0.25$), *PC200* ($b_j^s = 0.22$), and *AC.CX7* ($b_j^s = 0.11$) complement each other inside the linear equation and have a comparable relevance. Some additional model properties of the selected molecular descriptors are provided in Table 2.

It is possible to draw a mechanistic interpretation of the descriptors participating in the QSPR model of Eq. 1. For instance, the *AATS0e* descriptor has a straightforward correlation to the mean atomic Sanderson electronegativity (*Mse*) with an intercorrelation coefficient $R_{ij} = 0.9995$. This is in line with the observations of Jin's work (Jin et al. 2015), which state that a molecule with a high electronegativity requires a high energy to remove its electrons, thereby making it difficult the OH radical induced electron transfer and impeding oxidation by OH radicals. In other words, *AATS0e* is negatively correlated to $\log k_{OH}$ in Eq. 1. The *AC.CX7* descriptor involves the presence of halogen atoms (Cl, Br, and I) and thus considers the presence of electron withdrawing groups, making the C

Table 2 Model properties of the selected molecular descriptors

Descriptor	Regression coefficient	Standard error	Standardized coefficient	Probability level	VIF	Correlation to $\log k_{OH}$ (R^2_{train})
Intercept	11.91	0.340	0.00	< 0.0001	–	–
AATS0e	− 0.43	0.039	− 0.45	< 0.0001	1.03	0.36
SIC5	1.50	0.089	0.67	< 0.0001	1.01	0.65
PC200	− 0.91	0.160	− 0.22	< 0.0001	1.00	0.01
KC3738	− 0.33	0.050	− 0.25	< 0.0001	1.00	0.02
AC.CX7	− 0.03	0.011	− 0.11	0.0059	1.00	6.10^{-3}

Table 3 Performance of Jin's model (Jin et al. 2015) and proposed model in the PSM and BSM-based splittings. The best models are in bold

QSPR model	R^2_{train}	R^2_{test}	RMS_{train}	RMS_{test}
PSM splitting				
Jin's seven-descriptor model	0.81	0.79	0.22	0.31
Proposed model (Eq. 1)	0.86	0.92	0.19	0.26
BSM splitting				
Jin's seven-descriptors model	0.81	0.48	0.26	0.22
Proposed model (Eq. 1)	0.88	0.87	0.21	0.11

atom electrophilic and prone to attack by nucleophiles. Therefore, compounds posing halogens are less likely to be attacked by OH radicals, which are electrophiles. In this way, $AC.CX7$ is negatively correlated to $\log k_{OH}$ in Eq. 1 and this observation also agrees with the reported result (Jin et al. 2015).

The remaining $SIC5$, $PC200$, and $KC3738$ contributing descriptors clearly reflex the importance of the topological structure description for the molecules under study.

The model's squared correlation matrix is provided in Table 4S, showing the absence of high correlations between descriptors pairs. We also calculate the variance inflation factor (VIF), a parameter that measures the multicollinearity among descriptors. A VIF of 1 for a specific descriptor means that there is no correlation between this descriptor and all the remaining descriptors of the model, and a VIF exceeding 10 indicates that multicollinearity is a problem in the dataset (Roy and Roy 2009). From Table 2, it is demonstrated that the VIF parameter for each descriptor of Eq. 1 is near to 1. The numerical descriptor values are given in Table 5S.

Now we demonstrate that the proposed QSPR of Eq. 1 is generalizable and useful for application, that is to say, our model is not determined only by the training set composition due to the specific dataset partitioning of BSM. In Table 3, Eq. 1 derived with the BSM splitting is applied to the PSM splitting of Jin's work (Jin et al. 2015) and vice versa. It is clearly appreciated that in both the PSM and the BSM splittings, our proposed QSPR has a better predictive capability on the test set according to the R^2_{test} and $RMSE_{test}$ parameters.

In addition, we perform 100 different random splittings and recalculate the statistics of the model obtained by Jin et al. and the one proposed by us in the present work. As shown by Table 6S, Eq. 1 leads to R^2_{test} for the 100 external test sets ranging from 0.58 to 0.95 and RMS_{test} ranging from 0.15 to 0.26, while for the Jin's seven-descriptor model, R^2_{test} ranges from 0.35 to 0.94 and RMS_{test} ranges from 0.17 to 0.40. These findings suggest that the final model of Eq. 1 has a better stability of its predictive ability than the previous reported

model. The good predictivity of our QSPR model on the test set does not result by chance, and the molecular descriptors involved in Eq. 1 work satisfactorily on the different training-test set partitionings.

The analysis of the AD for the QSPR of Eq. 1 reveals that five training set compounds have high leverages exceeding the h^* limit (0.202) such as: **54**, **38**, **101**, **117**, and **75**. However, it is found that all the 29 test set compounds belong to the AD. The Williams plot (standardized residuals as function of h_i values) for Eq. 1 is provided in Fig. 3, while the leverages for the 118 compounds are reported in Table 1S. It is known the fact that a compound with a high leverage would reinforce the model if the compound is in the training set (good leverage), but such a compound in the test set could have unreliable predicted data, the result of substantial extrapolation of the model (bad leverage) (Gramatica 2007; Jaworska et al. 2005). Thus, the five training set compounds lying outside the AD reinforce Eq. 1, while their high h_i values may be purely attributed to the structurally heterogeneous dataset being studied. The results obtained with the leverage approach coincide with the ones obtained by using the
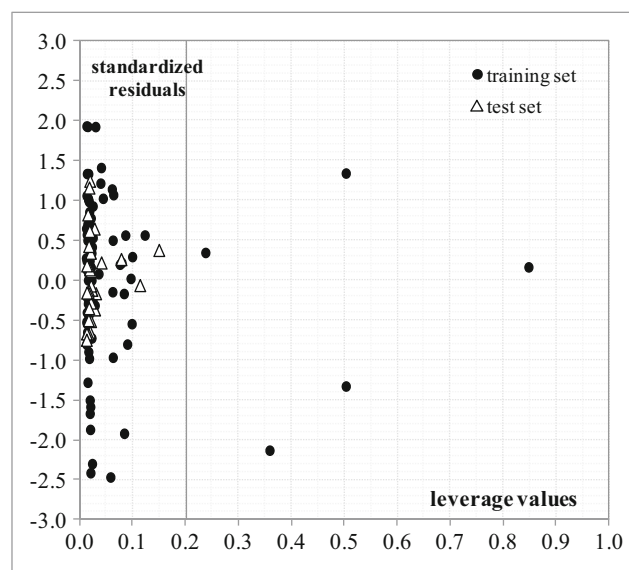

Fig. 3 Williams plot for Eq. 1

standardization approach, as the two conditions $s_{ik}^{max} \leq 3$ or $s_i^{new} \leq 3$ are followed by all the test set compounds. Thus, the predicted $\log k_{OH}$ values for the test set compounds can be considered as reliable.

Finally, it is noteworthy that an improved result is achieved in the published seven-descriptor model (Jin et al. 2015) when removing compounds **51** (Dalapon) and **54** (di(2-ethylhexyl) phthalate). The removal of such two high residual compounds in the reported QSPR leads to $N_{train} = 88$ and $N_{test} = 28$, and the following quality: $R_{train}^2 = 0.86$, $R_{test}^2 = 0.87$, $RMS_{train} = 0.18$, and $RMS_{test} = 0.26$. In the present study, we consider that our proposed QSPR of Eq. 1 represents an improvement over the reported Jin's models due to the next main reasons:

i. Equation 1 has a better predictive capability on the test set; $R_{test}^2 = 0.87$, $RMS_{test} = 0.11$, compared to the published $R_{test}^2 = 0.87$, $RMS_{test} = 0.26$.
ii. Outlier compounds are not removed by Eq. 1, contrary to the published result that removes two compounds.
iii. Equation 1 is a simpler model as it involves five descriptors instead of seven from Jin's work.
iv. Our linear QSPR only involves molecular descriptors that do not depend on the three-dimensional conformations of the organic micropollutants, making easier its application for predicting k_{OH} values for new compounds.

The conformation-independent QSPR approach employed in this work has as main advantage that the only experimental input required for designing the QSPR models is the experimental property being analyzed (Jagiello et al. 2016; Aranda et al. 2016; Duchowicz et al. 2014; Duchowicz et al. 2012), the experimental k_{OH} values in the present case. No further experimental information is needed, such as, i.e., the experimental X-ray crystal structure for each compound's conformation. It is well-known that additional more accurate and specific experimental information is always required when a microscopic and more sophisticated type of modeling methodology is involved in the study of a considered property (Jagiello et al. 2016). However, such specific experimental details are usually unavailable for any chemical system under study. In this work, we achieve our goal by predicting the hydroxyl radical rate constant in water without the need of additional experimental information in an approach that considers only constitutional and topological representations of the chemical structures.

Conclusions

The removal efficiency of contaminants from drinking water, distribution systems, and tap water can be assessed with the knowledge of their susceptibility toward oxidation in water treatment processes, in other words, by predicting the hydroxyl radical rate constant of organic micropollutants. Such experimental kinetic data, which are considered useful for the water industry when screening contaminants for their susceptibility to AOPs, are often unavailable for many emerging micropollutants. It is possible to assess the feasibility of an AOP for a specific compound in a specific natural water, by combining the k_{OH} predictions with the $R_{OH,UV}$ model (Rosenfeldt and Linden 2007). In addition, it is possible to estimate the removal of contaminants in natural water by ozonation treatment using the R_{ct} model (Elovitz and von Gunten 1999), which requires rate constants for each micropollutant screened as input.

In terms of model's size and conformation independence, our established QSPR models outperform previous published results. In addition, the proposed models are obtained after the simultaneous analysis of a large number of molecular descriptors calculated through freely available softwares like PaDEL, Mold2, and Epi Suite. The consideration of the constitutional and topological aspects of the molecular structures in the conformation-independent QSPR approach achieves once more acceptable results, and new results on other physicochemical and biological properties of interest will be published soon elsewhere.

References

ACD/ChemSketch (2016) http://www.acdlabs.com

Aranda JF, Garro Martinez JC, Castro EA, Duchowicz PR (2016) Conformation-independent QSPR approach for the soil sorption coefficient of heterogeneous compounds. Int J Mol Sci 17:1247–1255

Bagheri M, Mohseni M (2015) A study of enhanced performance of VUV/UV process for the degradation of micropollutants from contaminated water. J Hazard Mater 294:1–8

Benfenati, E. (2013) Theory, guidance and applications on QSAR and REACH, Orchestra, http://ebook.insilico.eu/insilico-ebook-orchestra-benfenati-ed1_rev-June2013.pdf

Borhani TNG, Saniedanesh M, Bagheri M, Lim JS (2016) QSPR prediction of the hydroxyl radical rate constant of water contaminants. Water Res 98:344–353

Buxton GV, Greenstock CL, Helman WP, Ross AB (1988) Critical review of rate constants for reactions of hydrated electrons, hydrogen-atoms and hydroxyl radicals (OH/O$^-$) in aqueous solution. J Phys Chem Ref Data 17:513–886

Delgado LF, Charles P, Glucina K, Morlay C (2012) QSAR-like models: a potential tool for the selection of PhACs and EDCs for monitoring

purposes in drinking water treatment systems—a review. Water Res 46:6196–6209

Diudea MVE (2001) QSPR/QSAR studies by molecular descriptors. Nova Science Publishers, New York

Draper NR, Smith H (1981) Applied regression analysis. John Wiley&Sons, New York

Duchowicz PR, Castro EA, Fernández FM (2006) Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. MATCH Commun Math Comput Chem 55:179–192

Duchowicz PR, Comelli NC, Ortiz EV, Castro EA (2012) QSAR study for carcinogenicity in a large set of organic compounds. Curr Drug Safe 7:282–288

Duchowicz PR, Bennardi DO, Baselo DE, Bonifazi EL, Rios-Luci C, Padrón JM, Burton G, Misico RI (2014) QSAR on antiproliferative naphthoquinones based on a conformation-independent approach. Eur J Med Chem 77:176–184

Duchowicz PR, Fioressi SE, Bacelo DE, Saavedra LM, Toropova AP, Toropov AA (2015) QSPR studies on refractive indices of structurally heterogeneous polymers. Chemom Intell Lab Syst 140:86–91

Elovitz MS, von Gunten U (1999) Hydroxyl radical ozone ratios during ozonation processes. I The R_{ct} concept. Ozone Sci Eng 21:239–260

Epi Suite 4.11 (2016) U.S. EPA: https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface

Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 111:1361–1375

Golbraikh A, Tropsha A (2002) Beware of q2! J Mol Graph Model 20:269–276

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

Hansch C, Leo A (1995) Exploring QSAR. fundamentals and applications in chemistry and biology. American Chemical Society, Washington, D. C

Hong H, Xie Q, Ge W, Qian F, Fang H, Shi L, Su Z, Perkins R, Tong W (2008) Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. J Chem Inf Model 48:1337–1344

Open Babel for Windows (2016) http://openbabel.org/wiki/Category:Installation

Jagiello K, Grzonkowska M, Swirog M, Ahmed L, Rasulev B, Avramopoulos A, Papadopoulos MG, Leszczynski J, Puzyn T (2016) Advantages and limitations of classic and 3D QSAR approaches in nano-QSAR studies based on biological activity of fullerene derivatives. J Nanopart Res 18:256. https://doi.org/10.1007/s11051-016-3564-1

Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) QSAR applicability domain estimation by projection of the training set in descriptor space: a review. ATLA Altern Lab Anim 18:256

Jin X, Peldszus S, Huck PM (2012) Reaction kinetics of selected micropollutants in ozonation and advanced oxidation processes. Water Res 46:6519–6530

Jin X, Peldszus S, Huck PM (2015) Predicting the reaction rate constants of micropollutants with hydroxyl radicals in water using QSPR modeling. Chemosphere 138:1–9

Katritzky AR, Goordeva EV (1993) Traditional topological indices vs. electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. J Chem Inf Comput Sci 33:835–857

Kaufman L, Rousseeuw PJ (2005) Finding groups in data: an introduction to cluster analysis. Wiley, New York

Kusic H, Rasulev B, Leszczynska D, Leszczynski J, Koprivanac N (2009) Prediction of rate constants for radical degradation of aromatic pollutants in water matrix: a QSAR study. Chemosphere 75:1128–1134

Lee Y, Gunten U v (2012) Quantitative structure-activity relationships (QSARs) for the transformation of organic micropollutants during oxidative water treatment. Water Res 46:6177–6195

Lee Y, von Gunten U (2012) Quantitative structure–activity relationships (QSARs) for the transformation of organic micropollutants during oxidative water treatment. Water Res 46:6177–6195

Leonard JT, Roy K (2006) On selection of training and test sets for the development of predictive QSAR models. QSAR Comb Sci 25:235–251

Luo Y, Guo W, Ngo HH, Nghiem LD, Hai FI, Zhang J, Liang S, Wang XC (2014) A review on the occurrence of micropollutants in the aquatic environment and their fate and removal during wastewater treatment. Sci Total Environ 473-474:619–641

Matlab 7.0. (2008) The MathWorks Inc., Masachussetts, USA. http://www.mathworks.com

Minakata D, Li K, Westerhoff P, Crittenden J (2009) Development of a group contribution method to predict aqueous phase hydroxyl radical (OH) reaction rate constants. Environ Sci Technol 43:6220–6227

Monod A, Doussin JF (2008) Structure-activity relationship for the estimation of OH-oxidation rate constants of aliphatic organic compounds in the aqueous phase: alkanes, alcohols, organic acids and bases. Atmos Environ 42:7611–7622

Morales AH, Duchowicz PR, Cabrera Pérez MA, Castro EA, Cordeiro MNDS, González MP (2006) Application of the replacement method as a novel variable selection strategy in QSAR. 1. Carcinogenic potential. Chemom Intell Lab Syst 81:180–187

PaDEL (2016). http://www.yapcwsoft.com

Peres JA, Dominguez JR, Beltran-Heredia J (2010) Reaction of phenolic acids with fenton-generated hydroxyl radicals: Hammett correlation. Desalination 252:167–171

Puzyn T, Leszczynski J, Cronin MTD (2010) Recent advances in QSAR studies: methods and applications: challenges and advances in computational chemistry and physics. Springer Science&Business Media B.V, Netherlands

Rojas C, Duchowicz PR, Tripaldi P, Pis Diez R (2015) Quantitative structure-property relationship analysis for the retention index of fragrance-like compounds on a polar stationary phase. J Chromatogr A 1422:277–288

Rosenfeldt EJ, Linden KG (2007) The ROH, UV concept to characterize and the model UV/H_2O_2 process in natural waters. Environ Sci Technol 41:2548–2553

Roy K (2015) Quantitative structure-activity relationships in drug design, predictive toxicology, and risk assessment. IGI Global, New York

Roy K, Roy PP (2009) Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FAMLR, PLS, GFA, G/PLS and ANN techniques. Eur J Med Chem 44:2913–2922

Roy K, Kar S, Ambure P (2015) On a simple approach for determining applicability domain of QSAR models. Chemom Intell Lab Syst 145:22–29

Roy K, Das RN, Ambure P, Aher RB (2016) Be aware of error measures. Further studies on validation of predictive QSAR models. Chemom Intel Lab Syst 152:18–33

Rücker C, Rücker G, Meringer M (2007) Y-randomization and its variants in QSPR/QSAR. J Chem Inf Model 47:2345–2357

Sudhakaran S, Amy GL (2013) QSAR models for oxidation of organic micropollutants in water based on ozone and hydroxyl radical rate constants and their chemical classification. Water Res 47:1111–1122

Sudhakaran S, Calvin J, Amy GL (2012) QSAR models for the removal of organic micropollutants in four different river water matrices. Chemosphere 87(2):144–150

Todeschini R, Consonni V (2009) Molecular descriptors for chemoinformatics (methods and principles in medicinal chemistry). Wiley-VCH, Weinheim

Toropov AA, Toropova AP, Rasulev BF, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2012) Coral: QSPR modeling of rate constants of reactions between organic aromatic pollutants and hydroxyl radical. J Comput Chem 33:1902–1906

Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474

Zimbron JA, Reardon KF (2005) Hydroxyl free radical reactivity toward aqueous chlorinated phenols. Water Res 39:865–869