Prediction of degradability of micropollutants by sonolysis in water with QSPR - a case study on phenol derivates

Judith Glienke, Willy Schillberg, Michael Stelter, Patrick Braeutigam

A R T I C L E   I N F O

Keywords: Quantitative structure-property relationship, Advanced Oxidation Processes, Organic micropollutants, Machine learning, Sonolysis, Prediction

ABSTRACT

The increasing quantity and variety of organic contaminants discharged into surface and groundwater increase the necessity of additional and suitable water treatment methods, which can be incorporated into existing wastewater treatment plants. The huge variety of micropollutants and local variability of the composition of the organic load or matrix effects paired with multiple possible degradation processes lead to the requirement of a recommendation tool for the best possible water treatment method under given local conditions. Due to the diversity of physicochemical properties of micropollutants, such predictions are challenging. In this study, a quantitative correlation between the structural properties of certain micropollutants and their degradability using high-frequency sonolysis has been investigated. Therefore, Quantitative Structure-Property Relationship (QSPR) has been applied on a set of phenol derivatives. To obtain the kinetic data, all experiments have been conducted in standardized, constant conditions for all 32 investigated phenol derivates. QSPR modelling was then executed using the software PaDEL for descriptor calculation and the software QSARINS for the overall modelling process including genetic algorithm (GA) and multiple linear regression (MLR). The final model consisting of 5 molecular descriptors was selected using a multi-criteria decision-making method based on extensive statistical parameters. The predictive power and robustness of the model was evaluated by means of internal cross validation and external validation using an independent validation set. The final selected model showed very good values for regression abilities, predictive power as well as stability (R^2_adj = 0.9455, CCC_adj = 0.9777, Q^2_tr = 0.9285, CCC_ext = 0.9797 and Q^2_ext-F1 = 0.9711). The applicability domain of the QSPR model was defined based on the Williams plot and Insubria plot. The five OECD principles for the application of QSPR/QSAR modelling in industry and regulation were fulfilled in the whole process to the best of our knowledge, including the collection of the underlying experimental data as well as the entire modelling process. The final QSPR model included the molecular polarity and occurrence of hydrogen bonds as major influences on the reaction rate constants in accordance with previous studies. Nevertheless, potential biases in the selection of these descriptors due to the small size of the dataset were highlighted.

1. Introduction

In recent years, an increasing quantity and variety of organic micropollutants in wastewater but also in surface and ground water have been observed [1–3]. Such pollutants, including pharmaceutical active compounds, personal care products, industrial chemicals,
pesticides, herbicides, and endocrine disrupting chemicals (EDCs), cause a big challenge to the wastewater treatment facilities [4]. Although micropollutants are present in low concentrations and only account for a minor part of dissolved organic matter (DOM), their potential toxicity paired with persistence and bioaccumulation can have disturbing effects on local water environments and biosystems [1,3]. This also includes phenol and its derivates, which are widely used in the industry but are suspected to have carcinogenic properties [5]. Due to their risk they provide not only to the environment but also possibly to human health, the occurrence, control, and elimination of these micropollutants are getting more and more attention in science [1,6-8].

With ongoing emission of new chemicals, conventional wastewater treatment plants more often become unable to eliminate all observed dissolved organic micropollutants completely. This is why degradation methods need to get more and more adjusted and diverse [6,9]. Therefore, additional water treatment methods specialized on the removal of organic micropollutants are highly investigated by different research groups and companies around the world. Promising approaches are ozonation and different advanced oxidation processes (AOPs) like electrochemical oxidation, Fenton reaction, photocatalysis or cavitation-based processes [10-14].

Ultrasound-assisted degradation methods are of particular interest as previous studies demonstrated high effectiveness in removing chlorinated organic compounds [15]. When a liquid is exposed to acoustic ultrasound waves, formation of liquid voids and cavitation bubbles can take place. As these microbubbles expand during the cycles of the pressure waves, they ultimately reach a critical size and undergo implosions, producing high local temperatures (up to several 1000 K) and high pressure (up to several 100 atm) [16]. These effects create the foundation of the hot spot theory [17]. Under such extreme conditions, water molecules can be thermally dissociated to H and OH radicals [18]. Organic contaminants can be trapped as gas molecules in the cavitation bubble where they undergo thermal fragmentation by pyrolysis. Otherwise, they can be solved in the bulk near the bubbles where they can react with reactive radicals. Phenol and phenol derivates are typically transformed into inorganic ions and short-chain organic compounds, which eventually are being mineralized to CO₂, CO₂H, and H₂O [19-22]. Exact degradation pathways, effects and influences of different parameters have not been clearly enlightened yet due to the high complexity of sonochemical reactions [15].

All AOP have in common, that reactive oxygen species (ROS) like hydroxyl radicals (OH) are generated in-situ [11,23]. Advantages of AOP include rapid oxidation reaction rate due to high oxidation potentials of ROS, and small to none secondary pollution problems [24,25]. Several AOP, such as photocatalysis, also distinguish themselves by high mineralization efficiencies. Cavitation based degradation however are most effective in degrading nonpolar compounds but tend to generate more hydrophilic products, resulting in their accumulation and lower overall mineralization [26,27]. A recommendation or even prediction of the best elimination approach for different pollutants would be highly supportive. However, accumulating extensive experimental data to characterize the activity and behaviour of all relevant substances in every possible degradation method is highly cost and time consuming [28-30].

The calculation of mathematical models to predict the chemical behaviour of chemicals without their experimental examination is a promising alternative. Quantitative Structure-Property Relationship (QSPR) or Quantitative Structure-Activity Relationship (QSAR) is used to correlate different biological, chemical or physical properties of a compound with its physicochemical characteristics [31]. These studies are not supposed to substitute experimental studies completely but to reduce their extent and gain a deeper and broader understanding of existing systems [32]. The use of such models is also suggested in the European legislation of chemicals REACH [33] to reduce the application of animal tests but also the general experimental costs. Applied to ultrasound-assisted degradation, the tool of QSAR modelling can contribute to a broader understanding of reaction pathways by identifying important molecular sites, which are the most relevant for the degradation.

QSAR modelling is based on the assumption that a molecules intrinsic properties like the melting point or the biological activity are defined by its molecular geometric, steric and electronic properties [34]. Two main principles are applied during the use of QSAR models. (1) Chemicals with similar structures show resemblances in molecular behaviour under related environmental conditions. (2) Variable behaviours are a result of differences in structure and composition [35]. The chemical structure is thereby translated into numerical parameters, so called molecular descriptors [36,37].

To implement QSAR models in industry and regulation, the Organization of Economis Co-operation and Development (OECD) drafted five principles of good practice QSAR models have to meet. According to the OECD, reliable QSAR models have to satisfy the following requirements: (1) a defined endpoint, (2) an unambiguous algorithm, (3) a defined domain of application, (4) appropriate measures of goodness-of-fit, robustness, and predictivity, and (5) a mechanistic interpretation, if possible [38,39].

Although QSAR modelling has already been applied in water research for a few years, many studies are not consistent with all requirements mentioned above. A fundamental issue is that many studies collect their underlying experimental data from different publications using different experimental setups and varying reaction conditions. Although this may be necessary to obtain a sufficiently amount of data, this may also lead to data heterogeneity, resulting in a potentially compromised integrity of the model [35,40,41]. This may interfere with the first OECD principle, which proposes that QSAR models based on defined endpoint values gained through standardized test protocols have the greatest potential for accordance between experimental values and calculated ones [38,42]. Therefore, collecting data from different resources should be avoided if possible [43]. This can also be seen in previous studies of the degradation of phenol derivates using sonolysis. By applying different experimental setups, parameters and matrices, different reaction rate constants for the same compound can be found in literature [20,44]. For example, Cost et al. [45] and Tauber et al. [46] found different reaction rate constants for the sonolytic degradation of 4-nitrophenol not only within their own study but also compared with one another due to a different experimental basis. Because of the inconsistency in data curation, combining results of different studies would not lead to datasets suitable for reliable QSAR modelling.

Also, according to Awa et al. [32], many developed QSAR models in wastewater treatment lack sufficient but necessary validation techniques. Further use of extensive validation methods and other statistical quality assurances should be introduced. Cherkasov et al. [47], Crinio et al. [43], Dearden et al. [40], and Gramatica et al. [41] also summarized substantial problems found in recent QSAR/QSAR models, addressing not only inconsistences of sufficient validation and inadequate or missing statistic measures, but also problems in data curation, data splitting and missing pre-treatments, such as descriptor normalisation and redundancy in information.

In this study the degradation of a group of phenol derivates using high-frequency sonolysis is investigated using a QSAR workflow based on Gramatica et al. [48]. To the best of our knowledge, this is the first QSAR model applied to sonolysis as Advanced Oxidation Process in wastewater treatment and in water research the first model evaluated with extensive amounts of statistical methods, e.g. multiple validation methods, tests for chance correlation and multicollinearity. To address some problems mentioned in previous publications [40,41,43,47], the experimental data was obtained with a standardized setup and protocol under reproducible condition to ensure the homogeneity of the experimental dataset. The introduced workflow possesses a variety of statistical methods to ensure the stability and reliability of the obtained model to reduce many problems as mentioned before. Overall, this study gives an overview of necessary statistical values and methods, which should
be included in further QSPR modelling in water research to meet all OECD principles.

2. Material and methods

2.1. Reagents and materials

All sources of chemicals, including CAS-numbers, molecular weight, structures, SMILES-codes and purity, are described in Tables S3 and S4. All chemicals were used as received and possessed a purity > 97%. Reaction solutions were prepared using freshly filtered ultrapure water (σ ≤ 0.055 µS/cm, TOC < 5 ppb; GenPure Pro, Fisher Scientific).

2.2. Experimental data

The reaction rate constant k of the degradation with sonolysis was obtained for 32 phenol derivates in a standardized experimental setup. All experiments were performed in 400 mL of a 10 µmol/L micropollutant solution in a triple determination.

The reaction took place in a cylindrical quartz glass reactor (Meinhardt Ultrasonics, Germany) with flange-mounted ultrasonic transducer (E/805/T, Meinhardt Ultrasonics), connected to an ultrasonic power generator (K 8, Meinhardt Ultrasonics) (Fig. 1). An ultrasound frequency of 860 kHz was used with a power of around 90 W. The system was kept at a constant temperature of 23 °C ± 1 °C through a water-cooling system. A CAD (computer-aided design) drawing concerning the reactor geometry in 3D can be found in the Supplement Material.

To determine the reaction kinetics, samples were taken at specific intervals during 30 min. To stop the reaction and to make sure, no possible polymerization reactions took place, 800 µL of the sample was blended with 200 µL of a 5 mmol/L sodium thiosulfate solution. The reproducibility of the experimental setup has been ensured with an ANOVA test. Results can be found in Text S2, Figure S2 and Tables S1 and S2.

For analysis, the concentration of micropollutants were measured using a high-performance liquid chromatography (HPLC) (LC2000, Jasco). The system included a fluorescence detector (FP-2020Plus, Jasco), a multiwavelength detector (MD-2010Plus, Jasco), an auto sampler (AS-2055Plus), a 100 µL injection loop and a RP C18 column (Dr. Maisch GmbH Kromasil 100 C18 10 mm*4.6 mm, 5 µm & 250 mm*4.6 mm, 5 µm) tempered at 40 °C. All HPLC methods are described in Table S5, including dissolvent ratio, retention time, detector, wavelength, and injection volume.

Before starting the modelling step, the underlying dataset was further analysed to ensure good quality. This step is not only useful for optimizing someone’s own modelling process, but also to minimize further problems and misleading results. Therefore, the profile and distribution of the dataset compounds was displayed. This included both the distribution of the experimental endpoint but also the structural

2.3. QSPR modelling process

To ensure the reliability of the modelling process in this study, all five OECD-principles were carefully respected. A list of the five principles with the respected modelling part is summarized in Table 1.

A flowchart of the entire QSPR modelling process is shown in Fig. 2. Some detail information can be also found in the Supplement Material (Text S3-S8).
domain inspected by principal component analysis (PCA) of the molecular descriptors to search for potential clustering of compounds or possible outliers [41].

2.3.3. Dataset splitting

The dataset was split into a training set and a validation set to ensure a later external validation of the selected QSPR model. To guarantee the significance of the later external validation, the two sets have to represent the whole dataset. Due to the fact that random splitting showed a huge diversity in model quality in trial models depending on the selected subsets, the splitting for this study was performed structure based on the previously performed PCA. After splitting, another round of filtering constant, near constant (>80%) and highly bivariate correlated (>95%) descriptors on basis of the training set was executed.

2.3.4. Model building and descriptor selection

In this study, multiple linear regression was used as the model approach with the following equation (2).

$$\hat{Y} = b_0 + \sum_{i=1}^{J} (b_j \times X_j) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_JX_J + u$$

The by QSARINS software implemented OLS technique thereby minimizes the sum of squares of the difference between experimental endpoint and its calculated value.

After setting up the dataset and its splitting, the descriptor selection step was executed. Therefore, all possible combinations of descriptors up to a number of 3 descriptors were calculated for the training set. This ensures the calculation of all low dimensional models before increasing the number of variables to obtain maximal results.

On that basis, higher dimensional models were obtained by applying a genetic algorithm to find the best model combination of a higher number of descriptors. The fitness function, on which the algorithm optimized the model quality, was selected to be $Q^2_{\text{loo}}$, which is
calculated following equation (3).

\[ Q^2 = 1 - \frac{\text{PRESS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n} (y_{i,\text{cv}} - \bar{y}_{i,\text{cv}})^2}{\sum_{i=1}^{n} (y_{i,\text{cv}} - \bar{y})^2} \]  

(3)

with \( \text{PRESS} \) as predictive residual sum of squares, \( \text{TSS} \) as total sum of squares, \( \bar{y}_{i,\text{cv}} \) as the predicted endpoint, \( y_i \) as the experimental endpoint and \( \bar{y} \) as the mean of the experimental endpoints.

The maximal number of descriptors was set as 15 of the number of compounds in the training set to \( n_{\text{feature,max}} = 5 \). Only models including descriptors with a significance level \( \leq 0.05 \) were retained. Best modelling output with minimal computational power was found with the GATS-parameters set to a population size of 200, 100 generations per size and mutation rate of 20%.

2.3.5. Model analysis and internal validation

After the modelling step, for each number of descriptors 10 final models were stored by the program with different statistical values for regression ability and for the internal LOO-cross validation applied during the model building. To select the best and final model, a few analysis steps were done. First of all, the QUIK rule (\( Q^2 \) Under Influence of K) was applied with a critical value of 0.05 to filter the models, where the correlation between the block of descriptors and the response is too allowed to calculate the reaction rate constant for each compound. A full list of all experimental data as well as degradation curves for all examined phenol derivates differ in value reaction rate constants for the examined phenol derivates. The analytical data obtained through HPLC analysis allowed to calculate the reaction rate constant for each compound. A full list of all experimental data as well as degradation curves for all experiments are given in Table S6 and Figure S3. As seen in Fig. 3 the reaction rate constants for the examined phenol derivates differ in value in the range from 0.01142876 1/min to 0.033556363 min \(^{-1} \), with the highest value for 4-hexylbenzene-1,3-diol, whereas 2,5-dihydroxybenzoic acid shows the lowest rate constant. Standard variations are all below 10%, indicating good reproducibility in accordance with the results of the executed one-factor ANOVA (Text S2, Tables S1 and S2, Figure S2).

The distribution of the resulting reaction rate constants is shown in Fig. 4. Even a perfect normal distribution could not be achieved, the endpoint distribution still seemed to be suitable for modelling purposes. In 2000 iterations, 30% of the training set were selected randomly for the internal validation set.

2.3.6. External validation, model selection and interpretation

In addition to the internal validation via Leave-One-Out cross validation, an internal validation through Leave-Many-Out cross validation was applied. In 2000 iterations, 30% of the training set were selected randomly and the Y-scrambling and Y-randomization were applied with 2000 iterations. For the remaining models, the relative importance of the model descriptors was inspected by counting their occurrence among the best 10 descriptors was inspected by counting their occurrence among the best 10 descriptors included the verification of the relevance of each descriptor by looking at its standardized coefficients and checking the confidence intervals for each descriptor.

3. Results and discussion

3.1. Experimentally derived \( k \) values and calculated descriptors

Sonolytic degradation experiments were performed with 32 different phenol derivates. The analytical data obtained through HPLC analysis were kept exactly how they were set in the original modelling process (chapter 2.3.4) with an iteration of 25 for the permutation and randomization, respectively.

After the final model was selected and analysed, an interpretation of the model equation with its included descriptors was performed. This included the verification of the relevance of each descriptor by looking at its standardized coefficients and checking the confidence intervals for each descriptor.

3.2. QSAR model for \( k_{US} \)

3.2.1. Statistical parameters of the final QSAR model

Based on these determined \( k \) values and the normalized descriptors, QSAR modelling and model selection was executed as described in chapter 2.3.4. The final QSAR model selected via MCDM consists of the following equation for the unstandardized coefficients:

\[ k_{\text{pred}} = 0.0220 - 0.2190 \times \text{ATSC4i} - 0.6174 \times \text{GATS2} + 0.1777 \times \text{GATS5} + 0.8494 \times \text{CrippenLogP} - 0.4390 \times \text{maxHBa} \]

The model equation with standardized coefficients is as followed:

\[ k_{\text{pred}} = 0.0220 - 0.0034 \times \text{ATSC4i} - 0.0110 \times \text{GATS2} + 0.0036 \times \text{GATS5} + 0.0169 \times \text{CrippenLogP} - 0.0064 \times \text{maxHBa} \]
The coefficients, standardized coefficients, the confidence intervals (Co. int. 95%) and the p-values of the model descriptors and intercept.

The ratios of confidence interval and descriptor coefficient as well as the p-values are below the critical values, whereas the model does not have to be considered suspect.

The coefficients, standardized coefficients, the confidence intervals (Co. int. 95%) and the p-values of the descriptors are listed in Table 2. The ratios of confidence interval and descriptor coefficient as well as the p-values are below the critical values, whereas the model does not have to be considered suspect.
The intercorrelation matrix of the model descriptors is listed in Table S8. The binary correlation between two descriptors is always below 0.7, which is stated as the critical value in most research papers for high collinearity [54]. With the applied value of the QUIK test higher than 0.050, severe multicollinearity of the model descriptors can be precluded with high probability.

All other statistical values obtained during the whole modelling process, including fitting criteria, internal validation criteria, and external validation criteria are listed in Table 3. With high values for $R^2$ and $R^2_{\text{ext}}$, there is an almost linear regression between predicted and experimental endpoint values seen in Fig. 5 with simultaneous low residuals (Figure S4). This is also supported by the high values of the lack of fit (LOF), and the concordance correlation coefficients for the training set and external validation set, $\text{CCC}_{\text{tr}}$ and $\text{CCC}_{\text{ext}}$, which are statistical variables for evaluating the regression ability and correctness of linear regression. With good quality of the internal and external validation, an over-fitting of the model can be rejected. Overall, the model shows very good performance in regression, predictability, and robustness.

### 3.2.2. Applicability domain

The Williams plot of the diagonal hat elements vs. standardized residuals by the model equation with a calculated critical HAT value of $h^* = 0.6923$ are shown in Fig. 6. For the model equation there is neither a structural outlier nor an outlier with high standardized residuals. Therefore, one can say, that both data subsets are inside the applicability domain and can be predicted with high precision. The Insubria graph of the diagonal hat elements vs. the predictions by the model equation is also shown in Fig. 6. This graph verifies the applicability domain for external applied compounds outside the training set without experimental data.

For future prediction application of the model equation, this applicability domain has to be considered, because every external compound has to fall within these boundaries. The model can be applied on molecules inside this defined applicability domain. This includes not only the structural domain based on the 5 model descriptors, but also the predicted endpoint value range, which is not included in the Williams plot but in the Insubria graph. The reliable prediction zone is therefore defined by the critical leverage value $h^*$ and the minimal and maximal values of the predicted endpoint from the training set. In this study, all compounds of the external validation set were found to be inside this domain.

### 3.2.3. Additional validation and test for chance correlation

Figure 7 shows the LMO plot, the Y-scramble plot as well the graph for the random responses.

All these three statistical tests show very good results. The $Q^2$-values for the LMO-cross validation are almost all relatively high. Therefore, one can say, that the model results are not entirely dependent on the whole training set. Hence the stability of the final model can be evaluated as good.

The Y-scrambling, a tool for showing possible chance correlation between the descriptors and the endpoint, resulted in very low values for $R^2$ and $Q^2$. The same effect was achieved by the Y-randomization. A possibility of mathematical chance correlation in the final model between the endpoint and the five model descriptors can therefore be seen as very low based on the results.

This is also consistent with the result of the permuted endpoint randomization test on the entire dataset. The calculated probability for chance correlation resulted in a p-value for $R^2$ of 0.0067 and for $Q^2$ of 0.0043. For the randomized endpoint randomization test, the calculated
probability for chance correlation resulted in a p-value for $R^2$ of 0.0004 and for $Q^2$ also of 0.0004. With the widely used acceptance level of probability of chance correlation $\leq 0.05$, the probability for mathematical chance correlation can be dismissed for the original final model. This is especially important, because of the small number of molecules used to develop the model due to higher probability of chance correlation with small datasets combined with large size of descriptor pool.

3.2.4. Mechanistical interpretation

Mechanistical interpretations of QSPR models are quite difficult, as multiple aspects have to be considered. This is why the OECD stated that it only has to be executed if possible [38]. Even though the obtained model is mathematically not overfitted, some potential biases in the descriptor selection due to the small size of the underlying dataset have to be highlighted. Detailed reasons are given below.

Five descriptors were selected for the final model. Compared with the relative frequency of descriptors in the best 10 models calculated during the whole modelling process, these descriptors are also present in a lot of other good models (Fig. 8).

ATSC4i is included in 30% of the 10 best models, and GATS5s and CrippenLogP are the most often selected descriptors with a 100% and 90% including rate, respectively. This indicates the high importance of the structural information included in these descriptors and their major influence on the endpoint. The descriptor ATSC4i is the centred Broto-Moreau autocorrelation of lag 4 weighted by first ionization potential. Therefore, this descriptor represents the autocorrelation over the topological distance 4 of the first ionization potential as the atomic property [55]. It is calculated as the sum of all products of the first ionization potential of atoms with a topological distance of 4. The descriptor has a negative influence on the value of the calculated reaction rate constant in this model.

The descriptor GATS2e is the Geary autocorrelation of lag 2 weighted by Sanderson electronegativity. It represents 2D-autocorrelation classes of descriptors. This descriptor is related to the topology of the structure or its parts that are in association with a given physicochemical property, in this case the Sanderson electronegativity [55]. According to the model equation, an increase in the GATS2e descriptor’s value decreases the reaction rate constant.

The descriptor GATS5s, which is the Geary autocorrelation of lag 5 weighted by intrinsic state, has a positive influence on the value of the reaction rate constant in the model of this study.

The descriptor CrippenLogP describes the logP calculated based on the approach of Crippen, which assigns additive contributions to the molecular logP value by the individual atoms of the molecule [56]. According to the model equation, higher values of the descriptor result in a higher calculated reaction rate constant. This effect is visible for example for 4-alkylbenzene-1,3-diol, which has the highest value for the CrippenLogP descriptor (Table S7), as it is the most nonpolar molecule in the dataset due to the long alkyl-chain. Simultaneously, it showed the highest reaction rate constant in the experimental investigation. At the same time, as seen in the experimental data (Fig. 3), the benzoic acids in the dataset have the lowest reaction rate constants with concurrently high values for molecular polarity (Table S7). Overall,
molecules with an increased molecular polarity showed a lower reaction rate constant. This effect is also observed in other experimental studies. For example, Xu et al. [57] showed, that in a group of phthalate acid esters, molecules with higher polarity (lower logP) were degraded slower with sonolysis. Likewise, Wu et al. [58] found a strong correlation between the logP value and the reaction rate constant of sonolysis, stating a molecule with higher logP value degrades faster. The obtained QSPR model in this study seems to support the observations, that the molecular polarity is indeed responsible for the decreased reactivity of these molecules. Due to the small size of the dataset in this study, this interpretation could however be misleading to an extent, as one cannot be entirely sure if this descriptor is only selected due to a bias in the underlying dataset. A low logP value seems to be equivalent to the sole presence of a carboxyl group, in this dataset even equivalent to a benzoic acid group. The later possesses a negative mesomeric effect due to its electron-withdrawing abilities, which leads to a decreased reactivity of the molecule. The activation energy for secondary substitutions, for example with ROS, is elevated and therefore a degradation with ROS would be decelerated. These two effects, sole polarity on the one hand, and presence of a benzoic acid on the other, might not be distinguished by the model and the real driving force behind the selection of that descriptor cannot be resolved completely. To prove which of these effects is influencing the reaction rate constant in sonolytic degradation, such biases have to be eliminated first. Therefore, a significantly larger dataset is needed in the future, where all key functional groups are represented in various contexts and combinations to reduce the occurrence of such biases. A selection of polar molecules without substituents with mesomeric effects as well as carboxyl groups substituted not directly onto the phenyl ring could provide clarity for that specific case.

The descriptor maxHba is dimensionless and represents the maximum E-states for (strong) hydrogen bond acceptors. With a negative coefficient in the model equation, an increase of the maxHba descriptor decreases the calculated endpoint of this QSPR model. Strong hydrogen bond acceptors like free pairs of electrons of oxygen atoms or ions therefore contribute negatively to the overall reactivity of the molecule during sonolysis. This could be due to an increased intermolecular interaction with other micropollutant molecules or water, which decreases the overall reactivity of the compound. The high significance of hydrogen bonds in stabilizing phenol derivatives was also examined by Mammino et al. [59]. The dihydroxybenzoic acid phenols in the dataset show the lowest reaction rate constants with simultaneous low values for the CrippenLogP and large values for the maxHba descriptor (Table S7), which could go along with the assumption, that acid groups paired with multiple hydroxy groups lead to higher polarity as well as stronger intermolecular hydrogen bonds. In the dataset, the phenol derivatives with nitro groups have also high values for the maxHba descriptor (Table S7). This could be explained with the meaning of the descriptor maxHba, because nitro groups as hydrogen bond acceptors result in a decrease of overall reactivity for these molecules as well. But there occurs the same problem as with the CrippenLogP descriptor. In all molecules containing strong hydrogen bond acceptors, these functional groups are directly substituted to the phenyl ring. Both functional groups (–COOH and –NO2) are fragments with negative mesomeric effect, resulting in a stabilization of the molecule. Therefore, the presence of hydrogen bond acceptors in this dataset is equivalent with the presence of a negative mesomeric effect. Thus, the selection of this descriptor might have been biased, as the model can again not distinguish between the stabilizing influence of hydrogen bond acceptors or of the negative mesomeric effect. This could only be differentiated with a larger dataset, including the functional groups substituted not directly to the phenyl ring and molecules with substituents functioning as strong hydrogen bond acceptors without negative mesomeric effect.

Due to the experimental background of the underlying dataset, the obtained model is only valid for sonolytic degradation at 860 kHz. As chemical and mechanical effects in sonolytic degradation are highly dependent on the frequency, different degradation pathways are more or less relevant due to for example different ROS concentrations, cavitation bubble sizes and lifetimes and other mechanical processes. Hence, models for other frequencies could be based on different descriptors, as polarity, hydrogen bonds, volatility and other molecular properties could have a varied impact on the degradation abilities of a compound.

This study showed, that the mechanistical interpretation of QSPR models is definitely not trivial. To obtain an extensive interpretation without biases, a large and diverse pool of chemicals, including all important functional groups in various contexts and combinations, is necessary as an underlying dataset in the future. Although not every descriptor could be fully interpreted, the two descriptors CrippenLogP and maxHba allowed a connection of the experimental observations and the results of the QSPR modelling. It was shown that despite the complexity of sonolytic degradation mechanisms, assumptions about the degradation ability of different phenol derivatives were possible. Due to the fact that sonolytic degradation does not follow only one reaction pathway, but is an interaction of multiple simultaneous reactions, QSPR is especially useful as it can contribute to mechanistical clarification.

4. Conclusions

In this study, a QSPR model to predict the reaction rate constant of sonolysis was developed for 32 phenol derivatives. The dataset was obtained using a standardized experimental setup with fixed parameters to eliminate the negative influence of anomalies inside the data on the modelling results. All five OECD principles for QSAR/QSPR models were carefully respected to the best of our knowledge. 1D and 2D descriptors calculated with PaDEL software were used as a basis and normalization was performed to eliminate the influence of different scales. The whole modelling process was executed with the software QSARINS.

Constant, near constant and correlated descriptors were excluded from further processing. The dataset was split into training set and external validation set in consideration of the chemical structure. 5 relevant descriptors were selected using genetic algorithm applying $Q_{\text{loo}}^2$ as fitness function. The final model was selected from a pool of generated models using a multi-criteria decision-making tool, finding the best model with an agreement of optimal values for all calculated statistical parameters. The final model shows good performance in regression, predictability, and robustness with $R_{\text{adj}}^2 = 0.9455$, $CC_{\text{GRR}} = 0.9777$, $Q_{\text{loo}}^2 = 0.9285$, $CC_{\text{ext}} = 0.9797$ and $Q_{\text{ext-F1}}^2 = 0.9711$. It was shown that a very homogenous dataset with a standardized experimental background leads to good modelling results, even with a very small number of compounds.

For sufficient model evaluation, multiple internal and external statistical tests were performed, and statistical parameters were calculated. This included not only an external validation for evaluation of the predictability of the model, but also internal cross validation to evaluate the stability. To remove the possibility of chance correlation between endpoint and descriptors, Y-scrambling and Y-randomisation was carried out with good results. Probability of chance correlation could also be dismissed for the final model through permuted and randomized response modelling. This was especially important due to the small size of the underlying dataset. The interpretation of selected descriptors was carried out as best as possible. An influence of the polarity and the occurrence of hydrogen bonds on the degradation ability of phenol derivatives was calculated by the obtained QSPR model, which goes along with previous experimental studies. However, due to the small size of the dataset, a biased selection of these descriptors could not be excluded. This is because all molecules with high polarity and strong hydrogen bond acceptors included substituents on the phenyl ring with negative mesomeric effect, which could also be responsible for the stabilization of these compounds. Larger datasets, including multiple variations of functional groups, are necessary to eliminate such biases and to make an extensive interpretation possible.

For future application of QSPR modelling in wastewater treatment, we suggest implementing not only the necessary internal and external
Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We express our gratitude to Prof. Paolo Gramatica, University of Insubria, Varese, Italy for providing access to QSARINS software and her work in QSAR-modelling. We also thank Prof. Dr. Emma Schymanski, University of Luxembourg, Luxembourg, Dr. Thomas Bocklitz, Leibniz Institute of Photonic Technology (IPHT) and Friedrich-Schiller-University, Jena, Germany, and Prof. Dr. Christoph Steinbeck, Friedrich-Schiller-University, Jena, Germany, for their support. The authors also thank Tobias Holzl for the CAD drawing of the reactor.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ultsonch.2021.105867.

References

[1] M. Gavrilescu, K. Demnerová, J. Amand, S. Agathos, F. Fava, Emerging pollutants in the environment: present and future challenges in bioremediation, ecological risks and bioremediation, N Biotechnol. 32 (1) (2015) 147–156, https://doi.org/10.1016/j.jbt.2014.01.001.
[2] A.S. Oberoi, Y. Jia, H. Zhang, S.K. Khanal, H. Lu, Insights into the fate and removal of antibiotics in engineered biological treatment systems: a critical review, Environ. Sci. Technol. 53 (13) (2019) 7234–7264, https://doi.org/10.1021/acs.est.8b01131.
[3] R.P. Schwarzenbach, B.J. Escher, K. Fenner, T.B. Hofstetter, C.A. Johnson, U. Gunten, B. Wehrli, The challenge of micropollutants in aquatic systems, Science (80-) 313 (5790) (2006) 1072–1077, https://doi.org/10.1126/science.1127291.
[4] N.S. Thomaidis, A.G. Asimakopoulou, A. Bletsou, Emerging contaminants: a tutorial mini-review, Glob. NEST J. 14 (2013) 72–79, https://doi.org/10.30955/gn.009023.
[5] J.-H. Kim, K.-K. Oh, S.-T. Lee, S.-W. Kim, S.-I. Hong, Biodegradation of phenol and chlorophenols with defined mixed culture in shake-flasks and a packed bed reactor, Process Biochem. 37 (12) (2002) 1367–1373, https://doi.org/10.1016/S0032-9592(02)00007-9.
[6] T. Deblonde, C. Cosus-Leguille, P. Hartemann, Emerging pollutants in wastewater: a review of the literature, Int. J. Hyg. Environ. Health 214 (6) (2011) 442–448, https://doi.org/10.1016/j.ijheh.2011.08.002.
[7] U.G. Akpan, B.H. Hameed, Photocatalytic degradation of 2,4-dichlorophenoxy-acetic acid by Ga-Ce–W–TiO2 composite photocatalyst, Chem. Eng. J. 173 (2) (2011) 369–375, https://doi.org/10.1016/j.cej.2011.07.009.
[8] J. Benner, D.E. Helbling, H.-P. Kohler, J. Wittelbe, E. Kaiser, C. Prase, T.A. Ternes, C.N. Albers, J. Amand, B. Horemans, D. Springel, E. Walravens, N. Boon, Is biological treatment a viable alternative for micropollutant removal in drinking water treatment processes? Water Res. 47 (16) (2013) 5955–5976, https://doi.org/10.1016/j.watres.2013.07.015.
[9] B.D. Blair, J.P. Crago, C.J. Hedman, R.J.F. Treguer, C. Magruder, L.S. Royer, R. D. Kläper, Evaluation of a model for the removal of pharmaceuticals, personal care products, and hormones from wastewater, Sci. Total Environ. 444 (2013) 515–521, https://doi.org/10.1016/j.scitotenv.2012.11.103.
[10] P. Braeutigam, M. Franke, R.J. Schneider, A. Lehmann, A. Stolle, B. Ondruschka, Degradation of carbamazepine in environmentally relevant concentrations in water by Hydrodynamic-Acoustic-Cavitation (HAC), Water Res. 46 (7) (2012) 2469–2477, https://doi.org/10.1016/j.watres.2012.02.013.
[11] M. Coba, G. Farinelli, A. Tiraferri, M. Minella, D. Vione, Advanced oxidation processes in the removal of organic substances from produced water: potential, configurations, and research needs, Chem. Eng. J. 411 (2021) 128668, https://doi.org/10.1016/j.cej.2021.128668.
[12] D. Ma, H.Y. Ci, L. Lai, X. Xiu, X. Hua, Z.A. An, L. Li, Y. Fu, B. Li, M. Zhang, L. Qin, S. Liu, L. Yang, Critical review of advanced oxidation processes in organic wastewater treatment, Chemosphere 275 (2021) 130104, https://doi.org/10.1016/j.chemosphere.2021.130104.
[13] M. Sillanpää, M.C. Nelli, A. Matilainen, Advanced oxidation processes for the removal of natural organic matter from drinking water sources: a comparative review, J. Environ. Manage. 208 (2018) 56–76, https://doi.org/10.1016/j.jenvman.2017.12.009.
[14] W.T. Vieira, M.B. de Faria, M.P. Spazoloni, M.G.C. da Silva, M.G.A. Vieira, Latest advanced oxidative processes applied for the removal of endocrine disruptors from aqueous media – A critical report, J. Environ. Chem. Eng. 9 (4) (2021) 105748, https://doi.org/10.1016/j.jece.2021.105748.
[15] Y.i. Jiang, C. Petrier, T.D. Waite, Sonolysis of 4-chlorophenol in aqueous solution: Effects of substrate concentration, aqueous temperature and ultrasonic frequency, Ultrason. Sonochem. 13 (5) (2006) 415–422, https://doi.org/10.1016/j.ultsonch.2005.07.003.
[16] P. Pies, T. Kondo, Free radical formation induced by ultrasound and its biological implications, Free R可以从上述文本中提取出以下信息：

- 超声波和声化学在环境中的应用
- 水处理中化学物质降解
- 环境中微污染物的处理
- 超声波和声化学在饮用水中的应用
- 超声波和声化学在有机污染物中的应用
- 超声波和声化学在水处理中的应用
- 超声波和声化学在水体中化学物质降解
- 超声波和声化学在水体中化学物质降解
