How do water matrices influence QSPR models in wastewater treatment?–A case study on the sonolytic elimination of phenol derivates

Judith Glienke1,2, Michael Stelter1,2,3, Patrick Braeutigam1,2,3*

1 Institute of Technical Chemistry and Environmental Chemistry, Friedrich Schiller University Jena, Jena, Germany, 2 Center of Energy and Environmental Chemistry (CEEC Jena), Friedrich Schiller University Jena, Jena, Germany, 3 Fraunhofer IKTS, Fraunhofer Institute for Ceramic Technologies and Systems, Hermsdorf, Germany

*patrick.braeutigam@uni-jena.de

Abstract

As the demand of freshwater increases with simultaneously aggravated climatic challenges, the development of efficient and effective water purification methods is of high importance. Qualitative Structure-Property Relationships (QSPRs) can support this process by calculating a correlation between the molecular structure and the degradability of water pollutants in a defined removal procedure, expressed by the kinetic constant of their removal. This can help to receive more mechanistical interpretation of the underlying process, but also to reduce experimental costs and time. As most QSPR models in wastewater treatment research are based on experimental data using ultrapure water as reaction solutions, it is still unknown to which extent QSPR models for different water matrices differ from each other with regard to selected descriptors and performance. Therefore, in this study the sonolytic degradation of 32 phenol derivates was investigated for three different water matrices (NaCl, Glucose, NaCl+Glucose) and compared to a previous study in ultrapure water. With only very few exceptions, the addition of water additives reduced the degradability of the target analytes. Based on these four datasets, QSPR modelling, respecting all five OECD principles for reliable QSPR models, were performed using numerous internal and external validations as well as statistical quality assurances to ensure good regression abilities as well as stability and predictivity. As the final four models were compared, it was observed that the descriptor selection and model calculation were highly impacted by the water additives. This was also confirmed when the descriptor pools of the best 10 models for each water composition were compared, as the descriptor pools were also highly dissimilar, indicating a shift in structural importance when changing the water composition. It could be shown that water matrices significantly influence the results of QSPR modelling even at very low concentrations of the matrix components.
Introduction

Freshwater availability in the world continues to further decrease due to growing water demands and usage as well as more and more challenging climatic conditions such as long periods of drought [1]. This means that the purification of used water is of great importance for a sustainable use of the limited water body. Therefore, the remediation of industrial and municipal wastewaters, including the removal of dissolved pollutants, is becoming an increasing research field. Organic micropollutants, characterized by low concentrations between μmol/L and nmol/L and a wide-ranged toxicological profile, are a group of very diverse anthropogenic chemicals, including pesticides, herbicides, pharmaceutical and personal care products, hormones and industrial chemicals [2]. Due to their mostly low biodegradability, they possess harmful risks to terrestrial and aquatic lives even at trace concentrations [3, 4]. Conventional water treatment methods are not specialized in dealing with these chemicals and mostly cannot sufficiently eliminate them [5, 6]. To reduce the potential risk for environmental and human health, additional water treatment technologies are required. One investigated ultrasound-assisted concept is the use of acoustic cavitation as an advanced oxidation process (AOP). This sonochemical approach is based on the fact that when a liquid is exposed to sufficiently large ultrasound waves, the formation of liquid voids or cavitation bubbles can occur [7]. As these microbubbles, filled with small amounts of dissolved gases and vapour from the surrounding bulk, expand during the cycles of the pressure waves, they will reach a critical size and ultimately undergo violent implosions, releasing high temperatures (up to several thousand K) and pressures (up to several 100 atm) at local hot spots [8]. Under these conditions, gas and water molecules inside the cavitation bubbles are thermally dissociated and form a variety of short-lived reactive radicals such as H• and OH•. Organic micropollutants like substituted phenol derivates can be degraded by sonolysis either directly through activating thermal decomposition reactions, or indirectly via the reaction with these generated oxidative species [7].

Among the large variety of organic pollutants, phenol and phenol derivates are used as industrial chemicals in petrochemical, chemical and pharmaceutical industries [1, 9]. Chlorinated and nitro-substituted phenols, used as pesticides and antibiotics, are among the most toxic phenol derivates [1, 10]. This is why the maximum allowable concentration of phenol and its derivates in effluent streams is set to less than 1 ppm in most countries [11]. The application of cavitation for the degradation of phenolic compounds has been investigated by numerous experimental studies with a focus on identification and optimization of experimental parameters, such as the pH, pollutant concentration, ultrasound frequency and power intensities [7]. Phenol is thereby often used as a model compound for water treatment studies [1].

As the range of substituted phenol derivates is quite large, a correlation between the structural properties of a molecule and the degradability by a treatment method like sonolysis is difficult. A promising approach is mathematical modelling, including machine learning, to predict a physical property of a substance solely based on its physicochemical characteristics, which is called QSPR modelling (Quantitative Structure-Property-Relationship) [12, 13]. The use of such predictive models is suggested by the European legislation of chemicals REACH, as it is a possibility to reduce general experimental cost and time [14]. The OECD (Organization of Economic Co-operation and Development) defined five fundamental principles, which has to be met for QSPR-models used in industry and regulation, which include (1) a defined end-point, (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 [15].
During the last few years, QSPR modelling has been applied more and more in wastewater treatment research and development, for example for micropollutant removal methods like ozonation, photocatalysis and adsorption [13]. As the use of QSPR modelling is still an emerging and developing research field in water treatment, a lot of previous studies fail to meet all five OECD guidelines and are facing different problems like insufficient model validation and inhomogeneous data quality [13]. A very pressing problem is that most previous studies are based on experimental data derived under unrealistic conditions like ultrapure water, hence they neglect the effects of water matrices on the chemical behaviour. As multiple previous experimental studies on phenol derivates investigated, different molecules behave differently when water matrix components are added to sonolytic degradation experiments. For example, Uddin and Okitsu showed in their study, that the addition of sodium chloride had no effect on the sonolytic degradation of hydroquinone, whereas the decomposition of 1,4-benzoquinone was affected negatively by the presence of NaCl [16]. Another study by Xiao et al. showed the different effect of added radical scavengers on ciprofloxacin and ibuprofen [17]. Even if the degradability of both compounds were decreased, ibuprofen was affected to a much lesser extent.

Such results show that the influence of one matrix component is not equal for each substance. To this day it is still unknown to which extent QSPRs based on datasets for different water matrices differ from each other and how strongly water additives influence the descriptor selection. The influence could vary between a potential change only in descriptor coefficients in the model equation with a selection of the same descriptors to an exchange of some of the descriptors or even to a total selection of entirely new descriptors. To the best of our knowledge, no previous study deals with the influence of specific matrix components on the results of QSPR modelling in wastewater treatment, especially high frequency sonolysis. Therefore, we conducted sonolytic degradation experiments with each sodium chloride and glucose and their combination for a set of 32 phenol derivates and compared the obtained three sets of kinetic data with the results from our previous study conducted in ultrapure water [18]. We then calculated a QSPR model for each of the four datasets to compare the resulting models and the selected descriptors to see if the added matrix components significantly changed the modelling outcome.

Material and methods

Reagents and materials

All sources of chemicals, including CAS-numbers, molecular weight, structures, SMILES-codes, and purity, are described in Tables A and B in S1 Text. All chemicals were used as received and possessed a purity > 95%. Reaction solutions were prepared using freshly filtered ultrapure water (σ ≤ 0.055 μS/cm, TOC < 5 ppb; GenPure Pro, Fisher Scientific).

Experimental data

The reaction rate constants for 32 phenol derivates were obtained using a standardized experimental setup described in detail by Glienke et al. [18]. The 32 phenolic structures thereby included different combinations of hydroxyl-, carboxyl-, nitro- and alkyl-groups as well as chloride and bromide. Most of these compounds, such as nitrophenols and alklyphenols are important industrial chemicals, but also hazardous substances widely found in real wastewater samples [19, 20]. The setup included a cylindric quartz glass reactor (Meinhardt Ultrasonics, Germany) with a flange-mounted ultrasonic transducer (E/805/T, Meinhardt Ultrasonics), connected to an ultrasonic power generator (K 8, Meinhardt Ultrasonics). The reaction solution with a volume of 400 mL was kept at a constant temperature of 23°C ± 1°C through a water-cooling system. The ultrasound frequency was set to 860 kHz with a power of around 90 W.
For each micropollutant, tested individually at a concentration of 10 μmol/L, 3 different simple water matrices were investigated. The kinetic data in the form of pseudo first order rate constants for the phenol derivatives in ultrapure water were taken from our previous study [18]. For the second dataset, kinetic data was obtained with the addition of 300 mg/L sodium chloride, with a resulting chloride concentration of 180 mg/L (≈ 0.5 mmol/L). The limit value for chloride is set to 250 mg/L by the WHO’s Guidelines for Drinking-Water Quality and is seen as the drinking water standard in many countries like Canada, the United States and the European Union [21]. Previous studies calculated an average chloride concentration in greywater of around 180 mg/L [22], which is why we set the investigated concentration to this value. The third dataset was obtained with the addition of glucose at a concentration of 50 μmol/L as a summed representative of very low residual concentrations of different sugars found in wastewater effluents [23]. Ultimately, the fourth dataset was derived with a combinational water matrix containing both sodium chloride and glucose with their respective concentrations.

To determine the reaction kinetics, triple determinations were conducted. Samples were taken at specific intervals during 30 min. To any possible polymerization reactions [24] of phenolic intermediates, 800 μL of the sample was blended with 200 μL of a 5 mmol/L sodium thiosulfate solution immediately after sample taking. In-detail results of the degradation experiments can be found in Fig 1 as well as in Fig A and Tables D–G in S1 Text.

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 autosampler (AS-2055Plus), a 100 μL injection loop and a RP C18 column (Dr. Maisch GmbH Kromasil 100 C18 10mm × 4.6mm, 5 μm & 250mm × 4.6mm, 5 μm) tempered at 40˚C. All HPLC methods are described in Table C in S1 Text, including dissolvent ratio, retention time, detector, wavelength, and injection volume.

Due to the low concentration range of the analyte and the constant production of reactive oxygen species during the whole reaction, a reaction rate of pseudo-first order was assumed and the kinetic constant was calculated using the following Eq (1) [25, 26]:

\[
\ln \left( \frac{c_0}{c_t} \right) = k \times t
\]

With \(c_0\) as start concentration, \(c_t\) as concentration of the analyte at time \(t\) and \(k\) as the reaction rate constant.

The average value of the triple determination was calculated and used as the dataset basis for the subsequent modelling.

QSPR modelling process

The QSPR modelling process mainly followed the workflow described in-depth by Glienke et al. [18]. To ensure the reliability of the modelling process, all five OECD-principles were carefully respected throughout the modelling. The QSPR modelling was conducted with the same parameters and settings for all four datasets, to ensure comparability of the results. Additional information about the process can also be found in the Supplement Material in Texts D–G in S1 Text.

Molecular descriptors

Due to the small dataset size of 32 molecules in the dataset, only one- and bi-dimensional descriptors, as well as PubChem fingerprints, substructure fingerprints and substructure fingerprint counts were calculated using the PaDEL-descriptor software version 2.21 [27].
Descriptors with missing values were deleted and the remaining descriptors were filtered for
descriptors with pair-wise correlation >95%, constant and near constant (>80%) descriptors
during the implementation step, to avoid later mathematical problems due to redundant infor-
mation and high collinearity within the descriptor pool [28]. A list of the descriptors imple-
mented into the QSARINS software version 2.2.4 [http://www.qsar.it] [29, 30] for further

Fig 1. Reaction rate constants of the degradation experiments for 32 phenol derivates. Experiments were carried
out in ultrapure water ($k_{\text{pure}}$), with 300 mg/L sodium chloride ($k_{\text{NaCl}}$), with 50μmol/L glucose ($k_{\text{Glucose}}$), and with
sodium chloride and glucose ($k_{\text{NaCl/Glucose}}$).

https://doi.org/10.1371/journal.pwat.0000201.g001

Descriptors with missing values were deleted and the remaining descriptors were filtered for
descriptors with pair-wise correlation >95%, constant and near constant (>80%) descriptors
during the implementation step, to avoid later mathematical problems due to redundant infor-
mation and high collinearity within the descriptor pool [28]. A list of the descriptors imple-
mented into the QSARINS software version 2.2.4 [http://www.qsar.it] [29, 30] for further
analysis and modelling are listed in the supplement material in Text E in S1 Text. After implementation into QSARINS software, the descriptor values were normalized to eliminate the influence of different descriptor value scales on the model calculation.

**Dataset splitting**

The dataset was split into a structural representative training set and a validation set to ensure a later external validation of the selected QSPR model. Therefore, the splitting was performed based on a principal component analysis of the entire descriptor pool in a ratio of 4:1 (Text F in S1 Text). After the splitting, constant, near constant (>80%) and highly bivariate correlated (>95%) descriptors on basis of the training set were filtered out of the descriptor pool.

**Model building and validation**

Based on the selected training set, QSARINS software uses multiple linear regression (MLR) as the underlying mathematical approach. First, all possible combinations of 3 descriptors out of the descriptor pool were calculated to ensure the best low dimensional model results. After that, higher dimensional models up to a maximum descriptor number of 5 were calculated using a genetic algorithm. This maximum descriptor number was set, because the amount of descriptors should not exceed 1/5 of the amount of compounds in the underlying training set [31]. Q^2_{loo} was selected as the fitness function and the algorithm parameters were set to a population size of 400, 500 generations per size and a mutation rate of 20%. The significance level for model descriptors were set to ≤ 0.05. The best 20 models of each number of descriptors were saved, although only models with descriptors of a critical value for the QUIK rule (Q^2 Under Influence of K) of 0.05 were retained. The later was set, because the correlation between the block of descriptors and the response should not be too similar to the inter-correlation between the model descriptors [32]. Internal validation was carried out with a leave-one-out cross validation (LOO-CV) as well as a leave-many-out cross validation (LMO-CV). External validation was performed by applying the model equations to the validation set. More details can be found in the supplement material in Text G in S1 Text. The overall best model, regardless of the number of selected descriptors, was selected via multi-criteria decision-making (MCDM) with regard to all previous calculated statistical values to find the best overall performing model with good abilities not only in regression, but also in stability and predictive ability.

**Results and discussion**

**Experimentally derived k values and calculated descriptors**

Sonolytic degradation experiments were performed for 32 phenol derivatives in three different water matrices (NaCl, glucose, NaCl + glucose). For comparison, kinetic data for the degradation of these substances in ultrapure water (k_{pure}) was retrieved from our previous study using the same experimental setup [18]. A full overview of the four datasets with the kinetic constants, standard variation, and the percentage variation to k_{pure} are given in the supplement material in Tables D–G in S1 Text.

The kinetic data of the four datasets are shown in **Fig 1**

The values for the rate constants k displayed in **Fig 1** are sorted by increasing values for k_{pure}. The values of k_{pure} vary between 0.01143 1/min for 2,5-dihydroxybenzoic acid and 0.03356 1/min for 4-hexylbenzene-1,3-diol. For the experiments with added NaCl, the rate constant ranges between 0.06830 1/min for 2,5-dihydroxybenzoic acid to 0.02800 for 4-methylbenzene-
1,2-diol. For glucose as water additive, $k_{\text{glucose}}$ values vary between 0.00912 1/min for 2,5-dihydroxynbenzoic acid and 0.02637 for 4-hexylbenzene-1,3-diol. Finally, for the water matrix with NaCl and glucose, the range of the rate constants lay between 0.01065 1/min for 2,3-dihydroxybenzoic acid and 0.02733 for 3,5-dichlorobenzene-1,2-diol. Overall, it can be observed that the general trend following the sorting for increasing values of $k_{\text{pure}}$ (higher $k$ values from left to right) is still dominant for the single component matrices, but a few major outliers can be perceived. For the two-component water matrix, the sorting gets even more mixed up. Hence, the influence of specific water additives is unique for every chemical structure.

To get a little bit more insight, the change in the kinetic constant due to the addition of water matrices were observed. The variation of the kinetic values for the three water matrices to the values in ultrapure water, calculated as $\Delta k = k_{\text{matrix}} - k_{\text{pure}}$ are displayed in the supplement material in Fig B in S1 Text. The percentage variations of the kinetic constants with water additives to the values in ultrapure water are given in Fig 2.

It is noticeable that the influence of matrix compositions on structural related compounds such as phenol derivates is not equal for all substances, but rather large discrepancies can be observed. Qualitatively, a straightforward trend and interpretation for the whole dataset cannot be identified.

The addition of sodium chloride or glucose decreased the kinetic constant for almost all phenol derivates. For the addition of glucose, this goes along with the assumption that an additional hydrophilic organic matrix decreases the degradability as it functions as competitive reactants for reactive oxygen species (ROS) in the bulk liquid, which leads to a decreased probability of a reaction between organic micropollutants and ROS [33]. The addition of sodium chloride has also lead to an inhibition of degradability in previous studies, as chloride ions can react with hydroxyl radicals to form a chloride radical and a hydroxyl anion, scavenging the highly reactive hydroxyl radicals in the liquid [34, 35]. Additionally, chloride and sodium radicals formed in cavitation bubble collapses can also further scavenge other ROS [36]. Formed chloride radicals can subsequently react with chloride anions or hydroxyl radicals to form dichloride anion radicals (Cl$_2^-$) or ClHO$^-$, respectively. As the redox potentials of these oxidizing species are lower than those of hydroxyl radicals, the formation of these radicals can result in a reduced degradation of organic species [37, 38].

Interestingly, the influence of the two-component matrix is not a sum of the single-component influences. For 2,5-dihydroxybenzoic acid for example, the variation to $k_{\text{pure}}$ for NaCl+-glucose (-15.42%) is in between the influence of the single NaCl and single glucose matrix (-6.52% and -20.20%, respectively). Other than that, the percentage variation for NaCl+glucose (-30.15%) is higher than the influence of glucose (-25.86%) or NaCl (-11.74%), but not as high as the sum of these two single influences. Therefore, the influence of more complex water compositions cannot simply be predict based on experiences of more simple matrices.

Even though a clear trend cannot be seen, a few interesting aspects can be observed. 4-nitrobenzene-1,2-diol for example is among the most influenced compounds in the datasets, as all three investigated water compositions decrease the degradability of this compound drastically. However, for 4-nitrophenol, the addition of NaCl does not have such a large impact, whereas the presence of glucose in the water severely reduces its rate constant. For 2-nitrophenol, the reduction of the kinetic constant is almost equal for all three water compositions. This example shows that small structural differences lead to a completely different impact of water additives on the degradation ability. This can be seen for example also for 4-methylphenol and 4-bromophenol, as the degradability of the first compound is hardly affected of any of the investigated water matrices, whereas the second-named molecule shows high differences in the kinetic constants.
Exceptions in this investigated set of phenol derivates are 4-methylbenzene-1,2-diol and 3,5-dichlorobenzene-1,2-diol. For these two compounds, the degradability increased with the addition of sodium chloride and glucose, respectively. This enhancement with salt was seen in experimental studies before. Seymour and Gupta observed a salt-induced enhancement of the sonolytic degradation of different phenol derivates [39]. On the one hand, this was explained with a salting-out effect, in which the dissolved salt increases the hydrophilicity of the liquid.
driving more nonpolar compounds towards the bubble-liquid interface, increasing their sono-
lytic degradation. This effect should probably be the strongest for hydrophobic compounds.
The same applies for the addition of glucose, as glucose is very hydrophilic, the assumption
would be that glucose has a lower effect on hydrophobic compounds, as they would be nearer
to the bubble interface and therefore would not compete with glucose for ROS as much as
hydrophilic compounds in the bulk liquid [34]. In this dataset however, the enhancement
effect is not present for the most nonpolar compounds (such as 4-hexylbenzene-1,3-diol) and
a salting-out effect is usually observed with much higher salt concentrations as used in this
study. Therefore, this explanation appears to be not applicable for the results in this dataset.
On the other hand, an enhancement of sonolytic degradability could be explained through a
decreased vapor pressure and increased surface tension due to salt ions, but this effect would
be very small with such a low salt concentration [39]. A different explanation for these anom-
ali es could be, that 4-methylbenzene-1,2-diol as well as 3,5-dichlorobenzene-1,2-diol could
have a higher reactivity towards chloride radicals and degradation intermediates of glucose,
introducing another degradation pathway in comparison to an ultrapure water environment.

For the simultaneous addition of sodium chloride and glucose, the kinetic constants
increased for four compounds, including benzene-1,2-diol and 4,5-dichlorocatechol in addi-
tion to the two substances discussed before. For benzene-1,2-diol however, the percentage
increase of the kinetic constant in a water matrix of NaCl and glucose is within the standard
variation of the triple determination. Hence, the effect seems negligible for this compound.
For all other compounds, the degradability was again decreased with additional water matrices
compared to the kinetics in ultrapure water. As seen with the single matrix compound experi-
ments, a qualitative discussion about the reasons behind these differences in enhancement and
inhibition is not trivial, as the theoretical background of matrix effects on sonochemistry and
sonolytic degradation are not well understood in detail for the used matrix components.

Generally, our data shows that the influence of matrix effects on the degradability of phenol
derivatives cannot easily be interpreted qualitatively. Additionally, the combination of differ-
ent matrix compositions does not seem additive. This makes a qualitative prediction of the
influence of water matrices on the degradability even more complex, as one cannot simply
investigate single component matrices and then make additive assumptions for more complex
water compositions. This could also be a problem for the application of QSRR models in real
wastewater systems, as most models are based on a dataset derived from experiments using
ultrapure water or real wastewater in one specific treatment plant. As seen in Fig 1, the four
experimental datasets differ strongly, and the sequence by value of the molecules is mixed up
by the addition of water additives. Due to the significantly different values of the experimental
endpoint, an influence on the selection of independent variables for the QSRRs is expectable,
as descriptor values stay the same with simultaneous changes in endpoint values. However,
because of the lack of previous studies, the extent of this impact is still unknown.

**QSPR models for k_{pure} k_{NaCl} k_{Glucose} k_{NaCl/Glucose}**

**Final QSPR models and descriptor comparison.** Therefore, we calculated QSRR models
for all four water matrices, to compare the modelling outcome as well as the selected descrip-
tors. QSRR modelling was carried out as described previously in the Material and Methods
section individually for all four datasets. The four final models were selected via MCDM. Their
most important statistical values are displayed in Table 1. More calculated statistical parame-
ters and intercorrelation matrices of the model descriptors can be found in Tables L–S in S1
Text. Additionally, a graphical display of the regression abilities, applicability domains and sta-
tistical tests can be found in the supplement material (Figs C–N in S1 Text).
### Table 1. Calculated QSPR models for four different water compositions and a few selected statistical parameters.

<table>
<thead>
<tr>
<th>Water Matrix Dataset</th>
<th>Unstandardized Model Equation</th>
<th>Selected statistical parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ultrapure water</td>
<td>( k_{\text{pred}} = 0.0219 - 0.0079 \times \text{MATS}1v + 0.0137 \times \text{VP} - 6 - 0.004 \times \text{maxHBa} )</td>
<td>( R^2 = 0.9333, R^2_{\text{adj}} = 0.9247, \ LOF = 0.0000, CCC_{\text{ext}} = 0.9665, Q^2_{\text{loo}} = 0.9135, R^2_{\text{tr}} = 0.9262, CCC_{\text{tr}} = 0.8930, Q^2_{\text{tr}-\text{loo}} = 0.8122, \ MCDM = 0.8749 )</td>
</tr>
<tr>
<td>NaCl</td>
<td>( k_{\text{NaCl, pred}} = 0.0181 + 0.0144 \times \text{ATSC}5i + 0.0078 \times \text{MATS}4v - 0.0136 \times \text{MATS}5i - 0.0169 \times \text{minHaaCH} )</td>
<td>( R^2 = 0.7499, R^2_{\text{adj}} = 0.7044, \ LOF = 0.0000, CCC_{\text{ext}} = 0.8570, Q^2_{\text{loo}} = 0.6448, R^2_{\text{tr}} = 0.9010, CCC_{\text{tr}} = 0.9446, Q^2_{\text{tr}-\text{loo}} = 0.9107, \ MCDM = 0.7859 )</td>
</tr>
<tr>
<td>Glucose</td>
<td>( k_{\text{Glucose, pred}} = 0.0172 - 0.0136 \times \text{AATS}4s + 0.0145 \times \text{MATS}4v - 0.0101 \times \text{GATS}3s + 0.0036 \times \text{VR}_2Dzs + 0.0034 \times \text{SpMin}_8\text{Bhs} )</td>
<td>( R^2 = 0.8402, R^2_{\text{adj}} = 0.8022, \ LOF = 0.0000, CCC_{\text{ext}} = 0.9132, Q^2_{\text{loo}} = 0.7438, R^2_{\text{tr}} = 0.9770, CCC_{\text{tr}} = 0.9584, Q^2_{\text{tr}-\text{loo}} = 0.9180, \ MCDM = 0.8380 )</td>
</tr>
<tr>
<td>NaCl + Glucose</td>
<td>( k_{\text{NaCl/Glucose, pred}} = 0.0204 + 0.0115 \times \text{MATS}3s - 0.0122 \times \text{GATS}4e - 0.0070 \times \text{hmax} + 0.0047 \times \text{JGI4} )</td>
<td>( R^2 = 0.8061, R^2_{\text{adj}} = 0.7708, \ LOF = 0.0000, CCC_{\text{ext}} = 0.8926, Q^2_{\text{loo}} = 0.7090, R^2_{\text{tr}} = 0.9690, CCC_{\text{tr}} = 0.9765, Q^2_{\text{tr}-\text{loo}} = 0.9451, \ MCDM = 0.8334 )</td>
</tr>
</tbody>
</table>

As seen in Table 1, the four final QSPR models and the selected descriptors are quite different. For a better understanding, the selected descriptors are displayed in Table 2 with their meaning and influence on the reaction rate constant.

Nevertheless, it could be useful to compare not only the overall best final model, but the descriptor frequency of the respective best ten models, to get a little bit more insight in the most important descriptors for calculating several good models rather than one specific equation. This reduces the risk of overlooking trends in the descriptor selection that are not displayed in one single model by chance. Therefore, the descriptor frequencies for the best 10 models for all four datasets are shown jointly in Fig O in S1 Text. All descriptors for all four datasets that occur in more than 10% of the models for one dataset or that occur in models for different water matrices are shown in Fig 3.

The selected descriptors are quite diverse for each dataset. 51 of the overall 56 descriptors are only present in the best 10 models of one water matrix, leaving only 5 descriptors which are selected in good models for more than one water composition. This indicates that water

### Table 2. List of selected QSPR model descriptors for all four finale QSPR models. Influence on \( k_{\text{pred}} \). Sign of descriptor coefficient in the model equation.

<table>
<thead>
<tr>
<th>QSPR model</th>
<th>Descriptor Symbols</th>
<th>Definition of the Descriptor</th>
<th>Influence on ( k_{\text{pred}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>QSPR model_pure</td>
<td>MATS1v</td>
<td>Moran autocorrelation of lag 1 weighted by van der Waals volumes</td>
<td>negative</td>
</tr>
<tr>
<td></td>
<td>VP-6</td>
<td>Valence path, order 6</td>
<td>positive</td>
</tr>
<tr>
<td></td>
<td>maxHBa</td>
<td>Maximum E-states for (strong) hydrogen bond acceptors</td>
<td>negative</td>
</tr>
<tr>
<td>QSPR model_NaCl</td>
<td>ATSC5i</td>
<td>Centered Brutto-Moreau autocorrelation of lag 5 weighted by first ionization potential</td>
<td>positive</td>
</tr>
<tr>
<td></td>
<td>MATS4v</td>
<td>Moran autocorrelation of lag 4 weighted by van der Waals volumes</td>
<td>positive</td>
</tr>
<tr>
<td></td>
<td>MATS5i</td>
<td>Moran autocorrelation of lag 5 weighted by first ionization potential</td>
<td>negative</td>
</tr>
<tr>
<td></td>
<td>minHaaCH</td>
<td>Minimum atom-type H E-state: CH:</td>
<td>negative</td>
</tr>
<tr>
<td>QSPR model_Glucose</td>
<td>AATS4s</td>
<td>Average Brutto-Moreau autocorrelation of lag 4 weighted by I-state</td>
<td>negative</td>
</tr>
<tr>
<td></td>
<td>MATS4e</td>
<td>Moran autocorrelation of lag 4 weighted by Sanderson electronegativities</td>
<td>positive</td>
</tr>
<tr>
<td></td>
<td>GATS3s</td>
<td>Geary autocorrelation of lag 3 weighted by I-state</td>
<td>negative</td>
</tr>
<tr>
<td></td>
<td>VR2_Dzs</td>
<td>Normalized Randic-like eigenvector-based index from Barysz matrix weighted by I-state</td>
<td>positive</td>
</tr>
<tr>
<td></td>
<td>SpMin8_Bhs</td>
<td>Smallest absolute eigenvalue of Burden modified matrix weighted by relative I-state</td>
<td>positive</td>
</tr>
<tr>
<td>QSPR model_NaCl_Glucose</td>
<td>MATS3s</td>
<td>Moran autocorrelation of lag 3 weighted by I-state</td>
<td>positive</td>
</tr>
<tr>
<td></td>
<td>GATS4e</td>
<td>Geary autocorrelation of lag 4 weighted by Sanderson electronegativities</td>
<td>negative</td>
</tr>
<tr>
<td></td>
<td>hmax</td>
<td>Maximum H E-state</td>
<td>negative</td>
</tr>
<tr>
<td></td>
<td>JGI4</td>
<td>Mean topological charge index of order 4</td>
<td>positive</td>
</tr>
</tbody>
</table>
additives indeed have a huge impact on the importance and selection of molecular descriptors for QSPR modelling. As not all of these descriptors can be discussed in detail, a few interpretational examples are given below.

One interesting example is the importance of the LogP value as a measure of molecular polarity for the calculation of the kinetic constant in sonolytic degradation. Based on the degradation mechanisms, the polarity of a compound is highly important for the pollutant elimination by acoustic cavitation, as more nonpolar compounds are near the bubble-liquid interface and can undergo faster reaction with ROS and additional processes like pyrolysis [34]. Within the descriptor pool, this structural information is represented by CrippenLogP and XLogP. The former is based on an atom-based calculation, whereas the latter uses the group contribution method. CrippenLogP is very prominent in the best 10 models for ultra-pure water, as 60% of these models contain that descriptor. With the addition of NaCl or Glucose, the importance of the polarity of a compound was expected to be even more relevant. As described before, dissolved NaCl as an inorganic salt increases the hydrophilicity of the liquid, driving more nonpolar compounds towards the bubble-bulk interface, supporting the sonolytic degradation [39]. Glucose on the other hand as a hydrophilic compound will be primary

Table 3. Predicted kinetic constants calculated by all four QSPR models for the four water matrices for 4-butylbenzene-1,2-diol as an external validation molecule (with the percentage variation of the predicted values to the experimental value). The variation of the predicted value for one water matrix on the experimental value for the same water matrix are highlighted.

<table>
<thead>
<tr>
<th>4-butylbenzene-1,2-diol</th>
<th>k_{pred,pure} = 0.0288 min^{-1}</th>
<th>k_{exp,pure} = 0.02845 min^{-1}</th>
<th>k_{exp,NaCl} = 0.02383 min^{-1}</th>
<th>k_{exp,Glucose} = 0.02143 min^{-1}</th>
<th>k_{exp,NaCl/Glucose} = 0.02111 min^{-1}</th>
</tr>
</thead>
<tbody>
<tr>
<td>k_{exp,pure}</td>
<td>-1.23%</td>
<td>-20.86%</td>
<td>-34.39%</td>
<td>-36.43%</td>
<td></td>
</tr>
<tr>
<td>k_{exp,NaCl}</td>
<td>14.94%</td>
<td>-1.55%</td>
<td>-12.93%</td>
<td>-14.64%</td>
<td></td>
</tr>
<tr>
<td>k_{exp,Glucose}</td>
<td>19.16%</td>
<td>3.48%</td>
<td>-7.33%</td>
<td>-8.85%</td>
<td></td>
</tr>
<tr>
<td>k_{exp,NaCl/Glucose}</td>
<td>24.08%</td>
<td>9.36%</td>
<td>-0.79%</td>
<td>-2.32%</td>
<td></td>
</tr>
</tbody>
</table>

https://doi.org/10.1371/journal.pwat.0000201.t003
dissolved in the liquid. More nonpolar compounds, which are situated nearer the interface would be less affected by competing reactions [34]. Therefore, it would be assumed that the LogP would stay as important or would be even more important for the calculation of QSPR models for these water matrices. However, the LogP is only present in 20% of the QSPR model\textsubscript{NaCl} and the QSPR model\textsubscript{NaCl/Glucose} in the form of the descriptor XLogP and CrippenLogP, respectively. For the QSPR model\textsubscript{Glucose}, neither of the two descriptors are amongst the descriptors selected in the best 10 models. It seems that other structural properties are more important for describing the kinetic constant when the water composition is changed. However, it is important to note that these findings should not be generalized beyond the structural range of the investigated molecules. Due to the very small size of the dataset used in this study, possible dataset anomalies and limited applicability domains of the models have to be kept in mind when deriving qualitative statements from descriptor interpretations. In addition, it could be also possible that a portion of the structural information of the polarity could be present in other, non-interpretable descriptors. To clarify and strengthen this observation, an investigation of larger datasets should be conducted in the future.

Another example is the descriptor JGI4, which is the mean topological charge index of order 4. It therefore measures the charge transfer between pairs of atoms separated by four bonds, and consequently, the global charge transfer in the molecule (e.g. dipol moment) [40, 41]. The positive coefficient for example in the QSPR model\textsubscript{NaCl/Glucose} (Table 2) indicates, that a specific charge distribution enhances the sonolytic degradability of a molecule. JGI4 is present in 10% of the best 10 models for ultrapure water and in 20% of the ones for NaCl/Glucose, but not at all for the models for the single matrix components.

In addition, the noticeable increased presence of specific structural properties can be seen in Fig 3. The occurrence of autocorrelation descriptors calculated using a function of the San- derson electronegativities (\(e\)), and the I-state (\(s\)) is highly significant for the calculation of good models on basis of the datasets for glucose and NaCl/glucose water matrices, represented by descriptors like GATS4e and MATS4 as well as GATS3s, MATS3s and AATS4s, respectively. The intrinsic state as one underlying function for autocorrelation descriptors reflects for an atom the possible partitioning of non-\(\sigma\) electrons influence along the path starting from the considered atom. Less partitioning of the electron influence leads to a higher availability of valence electrons for intermolecular interactions [40, 42]. With the addition of glucose to the experimental solution, the importance of electronegativities and intrinsic states seem to be expanded regarding the degradability of a compound.

In general, Fig 3 shows that the descriptor pool of the best 10 models of all four modelling processes are quite diverse. It is seen that the addition of water matrices changes the dataset significantly enough to highly impact the descriptor selection. Otherwise, more descriptors would be equal for all four datasets. This not only leads to different final QSPR models, but also totally changes the descriptors most important for calculating the kinetic constant. Even though a generality of this observation should be questioned due to the very limited dataset size in this study, the impact of water matrices on QSPR modelling outcomes should be considered in future modelling approaches and water matrices tailored on the application purpose should be used for obtaining the underlying dataset. Larger datasets for single matrix components as well as more complex compositions are needed to fully understand this impact and make more ubiquitous statements.

**Prediction on external molecules by the four QSPR models**

Based on the descriptor comparison, the influence of water matrices on the resulting QSPR models could be discussed. However, based on that qualitative interpretation, an assessment of
to what extent this impacts the predictive accuracy of external molecules is not possible. To quantify this impact based on an exemplary application, the predicted endpoints for 4-butylbenzene-1,2-diol (from the external validation set) calculated by all four QSPR models are compared to the experimental values in all four water matrices (Table 3). This was executed to see how large the variations for the predicted values on the experimental values in other water matrices is, to get more inside in the practical application of QSPR models and to see, to what extent a model calculated for a specific water composition can be applied to other water matrices.

As 4-butylbenzene-1,2-diol is not part of the training set, the molecule did not have any influence on the model calculation. As seen in the highlighted values in Table 3, all QSPR models can fairly precisely predict the experimental value for its specific water matrix, as predicted values correspond well with the experimental value for the specific water composition.

However, it can be shown that the predicted endpoints of one QSPR model have very high variations to the experimental endpoints for other water matrices. This can be seen for example for $k_{\text{pure}}$. This calculated kinetic constant equals the experimental value derived in ultrapure water very well with a deviation of around 1%. However, it differs highly from the experimental data obtained with water additives, with deviations of around 21%- 36%. The prediction of the degradability with water additives based on the QSPR model for ultrapure water therefore would not lead to an accurate prediction and should not be executed. The same phenomenon can be seen for the model $\text{NaCl}$.

However, it can also be seen that the predicted value of one QSPR model sometimes resonates good with an experimental value for a different water composition. This can be seen for example for $k_{\text{NaCl/Glucose}}$. This predicted value corresponds well with $k_{\text{Glucose}}$, so the prediction of the kinetic constant for 4-butylbenzene-1,2-diol in the same water matrix works well. But as also seen in Table 3, the experimental value for glucose as the only matrix component is also predicted accurately. However, this incident only occurs by chance due to two reasons. First, the QSPR model $\text{NaCl/Glucose}$ predicts $k_{\text{Glucose}}$ a little bit too high. Additionally, the difference of the experimental values for glucose and NaCl/glucose as water additives is very small, so this higher predicted value corresponds well with $k_{\text{Glucose}}$ by chance.

Based on these observations, it can be stated that the prediction of the kinetic constant is indeed only reliable if made by a QSPR model that implies the exact water composition as desired for the application even with very low concentrations of water additives. The accurate prediction of the degradability in other water compositions might be only possible due to serendipity. This can for example occur if the degradability of a compound is not severely affected by the addition of other water compounds (small differences in experimental endpoint values) or due to higher residuals of the prediction of that compound (smaller or larger predicted value). Further studies on more and more complex water compositions are needed to see if the impact on water matrix variations is still as high for highly complex mixtures.

As the underlying data and structural range was quite small, more work has to be done to get more insight in the influence of water matrices on the calculation and applicability of QSPR models. An increase in reliable dataset sizes for different water compositions is crucial to give more universally valid statements on the degree of that impact as well as deeper mechanical interpretation. Based on our results, it must be assumed that the impact is large enough to compromise the predictive ability of a QSPR model applied on other water compositions and that QSPR models calculated based on data for one specific water composition should not be used to predict the property of compounds in other, more complex water matrices, because their accurate prediction would be only possible by chance.

Additionally, these results highlight the need for homogeneous datasets in environmental chemistry. As no big datasets are available at the moment, a lot of QSPR studies derive their
data as a collection from different experimental studies. As this study shows, even small variations in the water composition led to very different experimental data as the degradability of some compounds varied by more than 30% when the water composition was changed compared to experiments in ultrapure water. This shows that if data is mixed from different experimental setups and studies, even small differences in the water composition can lead to a later compromised model, as changes in the experimental endpoint are not entirely due to the chemical structure. Therefore, an in-depth validation of data quality and experimental origin is crucial.

This problem is intensified by the inhomogeneous water matrices of real wastewater, as influent and effluent water compositions in wastewater treatment plants vary both locally and temporarily. As shown in this study, QSPR models for applications in real wastewater have to account for the water matrix, as models based on data from experiments in ultrapure water are probably not reliable for the prediction in complex water compositions. Therefore, a model wastewater composition should be defined by the scientific community, to reduce that problem of applicable QSPR models on real wastewater problems. This could help to standardize experimental setups and parameters (including the exact water compositions for model wastewaters), which could lead to larger experimental datasets of good quality, that are urgently needed to achieve reliable QSPR models with larger applicability domains and more extensive mechanistical interpretation.

Conclusions

In this study, four QSPR models were calculated for a set of 32 phenol derivatives using four different water matrices to evaluate the influence of water compositions on the modelling outcome as well as the overall descriptor selection. The dataset for ultrapure water was obtained from a previous study. The same standardized experimental setup was used to determine the kinetic constants for the organic micropollutants with three different water additives (NaCl, Glucose, NaCl+Glucose). The degradability of almost all compounds were drastically decreased with the addition of any of the water additives with only a very few exceptions.

1D and 2D descriptors as well as a set of selected Fingerprints were calculated using the PaDEL software and descriptor filtering and normalization were executed as data pre-treatment. During the subsequent QSPR modelling process carried out with the software QSAR-INS, all five OECD principles were carefully respected, and all parameters were kept constant for all four executed model calculations.

One final model was obtained for each water matrix. The models generally showed good regression abilities, as well as stability and predictivity, secured with internal and external validations. It was observed that the models varied widely, as the descriptor selection was unique for all four equations.

For a more general comparison, the frequency of the model descriptors for the best 10 models for each water matrix were compared. It could be seen that the descriptor pools were also highly different, leading to the conclusion, that water matrices significantly impact the descriptor selection and therefore the outcome of the modelling process even at very low concentrations. It was shown that the prediction of the kinetic constant is only reliable if made by a QSPR model that implies the exact water composition as desired for the application. The accurate prediction of the degradability in other water compositions might be only possible due to serendipity. Larger datasets and future studies on numerous water additives are necessary to further validate this statement and give more general recommendations about the applicability of QSPR models on different water compositions.
Supporting information
S1 Text. Supplementary material.
(DOCX)

Acknowledgments
We express our gratitude to Prof. Paolo Gramatica, University of Insubria, Varese, Italy for providing access to QSARINS software. 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.

Author Contributions
Conceptualization: Judith Glienke.
Data curation: Judith Glienke.
Formal analysis: Judith Glienke.
Investigation: Judith Glienke.
Methodology: Judith Glienke.
Project administration: Judith Glienke.
Resources: Michael Stelter, Patrick Braeutigam.
Supervision: Michael Stelter, Patrick Braeutigam.
Validation: Judith Glienke.
Visualization: Judith Glienke.
Writing – original draft: Judith Glienke.
Writing – review & editing: Judith Glienke.

References


