Removal efficiency calculated beforehand: QSAR enabled predictions for nanofiltration and advanced oxidation

D. Vries, B. A. Wols and P. de Voogt

ABSTRACT

The efficiency of water treatment systems in removing emerging (chemical) substances is often unknown. Consequently, the prediction of the removal of contaminants in the treatment and supply chain of drinking water is of great interest. By collecting and processing existing chemical properties of contaminants, QSARs (quantitative structure-activity relationships) for typical removal parameters can be constructed. Depending on the definition of the predicted endpoint, QSARs are (1) embedded in a process model suite, where they serve to predict a model parameter and the total, hybrid model predicts a removal rate or (2) used to directly predict, e.g., the removal rate, or a rejection coefficient for membrane systems. The different types of resulting prediction models, ranging from mechanistic (causal) to empirical (data-based), allow for hypothesis testing of current physico-chemical mechanisms and interactions between the contaminant, the type of water and the materials or energy (e.g. UV light) of the removal barrier. Two case studies illustrate this viewpoint and also pinpoint that, firstly, QSAR development, validation and residual analysis stress the linkage between the QSAR endpoints and process model predictions, and secondly, they lay bare the need to share data, algorithms and models.

Key words | modeling, nanofiltration (NF) process, priority compounds, QSARs, UV/H2O2 process, water treatment

INTRODUCTION

Due to developments in and production by, for example, chemical industries involved in ‘nanomaterials’, intensive agricultural industry, or medicine, trace amounts of possibly harmful substances have been found in water sources (Mompelat et al. 2009; Pal et al. 2010) and even in drinking water (Benotti et al. 2009; Eschauzier et al. 2012). These contaminants may pose health risks to consumers (Schríks et al. 2010). The traditional, individual approach for elucidating the behavior of contaminants in removal processes experimentally – for the large number of compounds that need an evaluation – would take years and would be prohibitively expensive due to the need for advanced measurement techniques. Prediction of the removal of these contaminants by models is thus of great interest. By collecting and processing existing chemical properties, and relating these to known removal rates through quantitative structure-activity relationships (QSARs), compound-specific prediction of removal characteristics can possibly be made. The trained QSAR can then be applied to another contaminant with similar chemical properties to obtain its, so-called, endpoint (i.e. a removal rate or other process parameter).

However, QSARs are not yet widely developed in water treatment research and, more importantly, only very few attempts have been made to (i) check the resolution of the endpoint prediction or (ii) assess the developed statistical relationships with validation data sets, and this latter aspect is crucial for meaningful mechanistic interpretation (Tropsha et al. 2003). An example of poor resolution is the use of percentage removal (Yangali-Quintanilla et al. 2009) instead of log
removal rates. In UV advanced oxidation processes where e.g. hydrogen peroxide is used to catalyze oxidative removal, externally validated QSARs have only been reported in a few papers (Kusic et al. 2009; Wang et al. 2009; Wols & Vries 2012). Related to both distinctiveness and validation is the risk of getting chance correlations when data sets containing a small number of compounds are used, especially if a high number of descriptors are available. To check over-fitting of data (Eriksson et al. 2003) and collinearity, cross-validation techniques are used while the application of an external data set is crucial to externally validate the prediction performance (Gramatica 2007).

Depending on the definition of the endpoint, QSARs are (i) embedded in a knowledge-based process model, where they serve to predict one or more model parameters (Benotti et al. 2009; Kusic et al. 2009; Wang et al. 2009; Wols & Vries 2012) while the total, hybrid model predicts a removal rate or (ii) used to directly predict the removal rate on the basis of data alone. An example of the latter is the prediction of the rejection coefficient for membrane systems as is done in the work of Yangali-Quintanilla et al. (2009, 2010). Furthermore, the set-up of QSARs can either be (i) designed by splitting the domain of chemical priority compounds in smaller clusters on the basis of a priori, mechanistic knowledge, or (ii) empirically designed by classifying the data by cluster analysis or unsupervised neural network techniques. The different types of resulting prediction models, ranging from mechanistic or causal to empirical (data-based, black box), allow for hypothesis testing of current physico-chemical mechanisms and interactions between the contaminant, the type of water and the materials or energy (e.g. UV light) of the removal barrier. In addition, calculations of removal efficiency, residual analysis and uncertainty propagation provide valuable information on whether the QSAR-based models are reliable enough for application in practice. Here, two case studies, i.e. a data-based QSAR and a QSAR knowledge-based process model illustrate the reliability tests, the potential of these models and their sensitivity to physico-chemical properties. Data-based and knowledge-based QSARs are developed for nanofiltration (NF) membrane rejection and removal by UV/H₂O₂ treatment.

**METHODS**

**Workflow**

The workflow to obtain the QSARs has been explained by Wols & Vries (2012). In short, the workflow comprises of the following.

1. **Data selection**, consisting of the following sequence of steps:
   (a) Selection of endpoints, where compounds with physicochemical properties are selected from literature data or from measured data.
   (b) Selection of relevant candidate descriptors, which are commonly calculated from different (quantum chemical) software packages.
   (c) Splitting of data set into training and test (validation) sets. Here, splitting of the data set is achieved by unsupervised neural network techniques, more specifically Kohonen Self Organizing Maps (Vesanto 2000).

2. **Construction of a QSAR model**, by a selection of candidate descriptors and subsequently choosing the best fitting descriptors. The endpoints predicted by QSARs can be evaluated by different error fitting criteria, for example $R^2$, $Q^2$, Akaike's information content, an $F$-value or other fitness parameters (Kubinyi 1994; Todeschini et al. 2004). Here the RQK-rule (Todeschini et al. 2004) is applied, and a multiple linear regression technique is used to obtain the statistical relations between descriptors and endpoints.

3. **Validation**.
   (a) Internal validation needs to be performed by cross-validation techniques such as leave-one-out (LOO) and leave-many-out. Permutation testing, also known as Y scrambling (Eriksson et al. 2003) shows the statistical significance of the permuted endpoint model, i.e. scrambled endpoints but with the same descriptors, compared to the significance of the (unpermuted) obtained QSAR.
   (b) External validation with the test data set as obtained in step 1 (c).

4. **Mechanistic interpretation** of the ‘optimal’ descriptors as obtained in the QSAR model.
Statistical evaluation

Evaluation of model simulations and predictions requires pre-specified criteria and statistical methods such as those specified in steps 2 and 3 of the workflow. Regarding the second step of the workflow, it is important to note that selecting the best fitting descriptor from a large pool of possible ones has the risk of selection based upon chance rather than on a mechanism of action depending on how many compounds (endpoints) are used and how many fitting descriptors. Due to the increasing number of candidate descriptors (the software program Dragon can calculate up to 4,885 descriptors, http://www.talete.mi.it/), the probability of chance correlations increases when searching through the large database of candidate descriptors. This issue was already addressed by Topliss & Costello (1972), but according to Kubinyi, little attention has been paid to this issue, or in his words: ‘the literature is now spoiled with, most probably, thousands of meaningless chance correlations’ (Kubinyi 2004). In that work, using the Selwood data set, Y-scrambling was performed for several numbers of selected descriptors. For each number of selected descriptors, 1,000 random models were developed for random endpoint values. By increasing the number of selected descriptors, the mean $R^2$ gradually increases ($R^2$ always increases with the number of descriptors), whereas the mean F-value decreases. Here, the effect of chance correlations have been studied using a similar Monte Carlo analysis of the coefficient of determination ($R^2$) by fitting a randomly generated endpoint for $n$ compounds with $x$ selected descriptors, and varying the ratio between $x$ and $n$. In our case, the database consisted of a pool of around 800 descriptors and 350 compounds. Results are shown in the Appendix (available online at http://www.iwaponline.com/ws/013/109.pdf).

Although the different types of QSAR models ask for different validation procedures, step 3 shares a common mathematical base. Specifically, all these QSAR models are obtained by some goal function which yields test statistics related to the residuals which can be compared with other (QSAR) models. The residuals may be QSAR endpoint data minus the measured (or reported) endpoint data, predicted process model data minus the measured (or reported) data or other, analogous variants. The output of a (water treatment) model may be defined as the removal, degradation or rejection efficiency of some compound usually expressed as log removal units or a percentage. The reliability of a QSAR, whether it predicts directly such a removal efficiency or is used as a parameter embedded in a process model, can be evaluated by examining the interval estimates of the predicted efficiencies. In the case of a knowledge-based QSAR (i.e. a QSAR embedded in a process model), the covariance of the QSAR endpoint can be propagated to obtain the interval estimates of removal efficiencies. Table 1 shows some different variants of models, their training functions and in bold typeface the type of case studies illustrated in this work. In this work, the model structure is typically a time-invariant dynamic model, dependent on states $x$ (e.g. concentrations of other compounds), inputs

<table>
<thead>
<tr>
<th>Type of model</th>
<th>Training function and model structure</th>
<th>Validation statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Mechanistic process model (not based on a QSAR)</td>
<td>$\min {l(e)} \text{ where } e = y_m - y; \ y = f(x, u, \theta_M; t)$</td>
<td>$[l(e), \theta_M, \theta_Q] \ (e.g.:) F$-statistic, FPE (final prediction error)</td>
</tr>
<tr>
<td>2a Knowledge based QSAR (error evaluation process model)</td>
<td>$\min {l(e)} \text{ where } e = y_m - y; \ y = f(x, u, \theta_M, \theta_Q; t); \ \theta_Q = g(\theta_D)$</td>
<td>$[l(e), \theta_M, \theta_Q] \ (e.g.:) Q_{LOO}^2, R^2$ and/or confidence intervals</td>
</tr>
<tr>
<td>2b Knowledge based QSAR (error evaluation QSAR endpoint)</td>
<td>$\min {l(e)} \text{ where } e = \theta_Q - \theta_Q; \ y = \theta_Q; \ \theta_Q = g(\theta_D)$</td>
<td>$[l(e), \theta_Q] \ (e.g.:) Q_{LOO}^2, R^2$</td>
</tr>
<tr>
<td>3 Data-based QSAR</td>
<td>$\min {l(e)} \text{ where } e = \theta_Q - \theta_Q; \ y = \theta_Q; \ \theta_Q = g(\theta_D)$</td>
<td>$[l(e), \theta_Q] \ (e.g.:) Q_{LOO}^2, R^2$ and/or confidence intervals</td>
</tr>
</tbody>
</table>

*aCase studies in this work are set in bold typeface.*

**Table 1** Overview of type of models, training function and validation statistics. See the main text for an explanation of used notation.
u (often considered as a constant, e.g. a UV lamp fluence rate) and parameters $\theta_Q$ (unknown parameter) and $\theta_M$ (known parameter). The training function is often the mean squared error of model errors ($\epsilon$). A QSAR is represented by the function $g$. The output $y$ of the model can either be a QSAR endpoint or the process model output.

**QSAR case studies**

A data-based QSAR and a QSAR knowledge-based process model have been trained and tested for NF membrane rejection and removal by UV/H$_2$O$_2$ treatment. Table 2 gives an overview of trained parameters used in the QSAR models.

**Experimental data**

The data-based QSARs for NF were trained with data measured by Verliefde (2008). The NF experiment was carried out with a virgin Desal HL membrane in a single 4040 membrane set-up under the following operating conditions: feed flow rate 1,500 L/h and recovery was set at 10%. The membrane rejection coefficient, i.e. the fraction of compounds which is separated from the filtrate due to the membrane, was calculated according to the stationary convection-diffusion transport model equations as introduced by Spiegler-Kedem and as used by many researchers, e.g. Verliefde et al. (2009). The logarithm of the membrane rejection was used as an endpoint in the QSAR. Instead of splitting up the data set, the QSAR model was validated with an external data set measured by Botton & Cornelissen (2012). Only the virgin membrane has been considered. The set-up was slightly different in the validation set: membrane units were 635 mm (25 inch) instead of 1016 mm (40 inch) and a Desal HL2521 TM membrane was employed. The feed flow rate was 350 L/h and the recovery was 9.7%. The logarithm of the membrane rejection was used as an endpoint. The complete data set is provided as supplementary material in Tables A.1 and A.2 in the Appendix (available online at [http://www.iwaponline.com/ws/013/109.pdf](http://www.iwaponline.com/ws/013/109.pdf)). The QSAR model has been validated by an external data set of 23 compounds. Some of these compounds were similar to the compounds in the training data, but also new compounds were introduced in the external data set. An overview of the experimental conditions used for the NF QSAR development is given in Table 3.

For the knowledge-based QSARs for UV/H$_2$O$_2$ approaches, separate QSAR models need to be developed for the hydroxyl radical rate constants, quantum yield and molar absorption. These three parameters largely define how well compounds are degraded by UV/H$_2$O$_2$. Literature values of hydroxyl radical rate constants, quantum yield and molar absorption were collected to develop the

<table>
<thead>
<tr>
<th>Table 2</th>
<th>QSAR case studies, their model functions and their respective parameters for the NF and UV/H$_2$O$_2$ case study</th>
</tr>
</thead>
<tbody>
<tr>
<td>NF</td>
<td>UV/H$_2$O$_2$</td>
</tr>
<tr>
<td>$f(\ldots ; t)$</td>
<td>$f(\ldots ; t)$</td>
</tr>
<tr>
<td>$\theta_Q$ (endpoint)</td>
<td>$\theta_Q$ (endpoint)</td>
</tr>
<tr>
<td>$\theta_M$</td>
<td>$\theta_M$</td>
</tr>
<tr>
<td>Knowledge-based Qsar</td>
<td>Knowledge-based Qsar</td>
</tr>
<tr>
<td>Data-based Qsar</td>
<td>[2] OH radical rate constant</td>
</tr>
<tr>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Membrane pore size, thickness of membrane</td>
<td>n.a.</td>
</tr>
<tr>
<td>n.a.</td>
<td>Half life degradation e.g.</td>
</tr>
<tr>
<td>n.a.</td>
<td>n.a.</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Table 3</th>
<th>Experimental conditions used for the development of the QSAR models for the NF process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>Reactor</td>
</tr>
<tr>
<td>Verliefde (2008)</td>
<td>Single 4040 membrane unit</td>
</tr>
<tr>
<td>Botton &amp; Cornelissen (2012)</td>
<td>3 parallel 2521 membrane units</td>
</tr>
</tbody>
</table>
knowledge-based QSARs. Table 4 shows the operational conditions of the measurement data. The knowledge-based QSAR model was used to predict the OH radical reaction rate constant, from which the OH radical-induced degradation was calculated as follows, using the process model described by Wols & Hofman-Caris (2012):

\[
\frac{N}{N_0} = \exp\left(-\ln(10)\frac{2\epsilon_{H_2O_2} \phi_{H_2O_2}[H_2O_2]}{U_f k_N dose}\right) \quad (1)
\]

and using the following definition for the UV dose:

\[
dose = E_p(0) \frac{1 - 10^{-aL}}{aL} \frac{T}{\ln(10)} U_f. \quad (2)
\]

where the definition and notation of terms in Equation (1) and (2) are further explained in Wols et al. (2011). In this work, it is important to note that \(k_N\) in Equation (1) represents OH radical reaction rate, which is calculated by the QSAR model. The parameter \(k_S\) representing all the reactions with background components (\(H_2O_2, HCO_3^-, CO_3^{2-}, DOC, etc.\)) was set to \(6.0 \times 10^4 \) 1/s. The hydrogen peroxide concentration was equal to 8.3 mg/L, whereas the UV dose in the reactor was set to 500 mJ/cm² for a ballast of 100%. Table 4 summarizes the operational conditions.

### RESULTS AND DISCUSSION

#### Assessment of NF data-based QSAR

The NF black-box, data-based QSAR model with the (logarithm of) the rejection coefficient (log \(y\)) defined as an end-point can be validated internally and externally with data under different flow conditions with high internal and external statistics (Figure 1). High prediction statistics were also obtained in Yangali-Quintanilla et al. (2009, 2010). The following QSAR is obtained:

\[
\log \theta_Q = 2.023 + a \cdot \theta_D
\]

with: \(a = [0.02616 - 0.894 0.09196 - 1.068]\) descriptors: \(\theta_D = [SsoO iddmn nAB COOH]^2\) training data: \(n = 40, R^2 = 0.884, Q_{LOO}^2 = 0.857, SD_{res} = 0.202\) test data: \(n = 23, Q_{ext}^2 = 0.784, SD_{res} = 0.258\).

The internal predictability of the QSAR model appears to be high, with a \(R^2\) as well as a \(Q^2\) above 0.8. This is also confirmed in Figure 2(d), where the \(R^2\) is significantly higher (0.88) than for the random permuted models (maximum of 0.3–0.4). Also, the residuals are well distributed (Figure 2(a–b)). The measured membrane rejections of those compounds were both well predicted by the QSAR model in terms of \(Q^2\) (\(Q_{ext}^2 = 0.78\)). Both the test data and the training data cover a wide domain of applicability (Figure 3). Furthermore, most of the predicted values of the rejection from the training and test data compare well on a logarithmic scale (95% confidence region), to the measured values (Figure 1). Computation of the variability of the results to a percentage scale per...
compound show that results should be interpreted with caution. Two compounds in both sets, sucrose and trimethoprim, are outside this domain due to high leverage, i.e. an influence on the model that is too large and would result in an unstable model (Eriksson et al. 2005). It appears that the combination of a very high value for SssO (electrotopological state index for the sum of (–O–) E-states (an E-state is an electrotopological-state index for atoms in a molecule: a description of the calculation procedure is given by Kier & Hall (1993))) and the absence of a COOH group are differentiating these compounds from the rest of the data set. Conversely, dikegulac for example, also has a very high value of SssO (see Appendix, available online at http://www.iwaponline.com/ws/013/109.pdf) but also exhibits a COOH group and falls within the domain of applicability. Bentazon and sulfamethoxazole show a too large deviation from measured data and are therefore considered as outliers (Eriksson et al. 2005).

The number of COOH groups (COOH) is an important descriptor in all QSARs found by the genetic algorithm and can be related to the hydrophobicity of the compound, which influences the membrane rejection. The other topological descriptors (SssO: sum of (– O –) E-states; Iddmm: mean information content on the distance degree magnitude; nAB: number of aromatic bonds) are important because they represent the size and shape of the molecule,

Figure 2 | (a) and (b) residual plots, (c) uncertainty in regression coefficients and (d) permutation testing.

Figure 3 | Domain of applicability is visualized by a Williams plot. The compounds in red have too large residuals (outliers) or too large leverage (unreliable predictions). (Note: The online version of this figure is in color: http://www.iwaponline.com/ws/toc.htm.)
which influences the way the molecules are passed through or blocked by the membrane.

**Prediction error evaluation of NF data-based QSAR**

While the validation statistics of the data-based QSAR are comforting at first sight, the confidence intervals when propagating the standard deviation of the QSAR endpoint are quite large. Closer inspection reveals that indeed, while the log rejection of the majority of compounds is about 1 to 1.5, the standard deviation of the QSAR is about 20–30% of the rejection logarithmic value. The magnitude of standard deviation on a logarithmic scale has profound effects on the calculated error bars for percentage removal, especially when removal is less than a unit log removal (below 90%), see e.g. the case of trimethoprim, paracetamol or sulfamethoxazole. Furthermore, assuming a 5% confidence level, the error bars of the interval estimates will then increase by a factor of 1.96, as is shown in Figure 4. Figure 5

![Figure 4](https://iwaponline.com/ws/article-pdf/13/6/1425/415223/1425.pdf)
depicts the difference between measured and predicted rejection with 95 and 99% confidence regions. This figure shows that, highly probably, the predicted rejection of sulfamethoxazole is uncertain, since the estimated 95 and 99% confidence regions are exceeded.

**Assessment of NF knowledge-based QSAR**

The NF knowledge-based QSAR model, where the partition coefficient of a rejection model (Verliefde et al. 2009) is defined as the endpoint, could not be validated for the range of pharmaceuticals. Further inspection of the process model revealed that the model output is dependent on several parameters, where two membrane parameter (i.e. $\theta_M$) values have to be estimated from retention and flux data. The many degrees of freedom did not yield unique values for $\theta_M$ and yielded QSARs with very low prediction statistics. Further results are therefore omitted.

**Assessment of UV/H₂O₂ knowledge-based QSAR: molar absorption and quantum yield**

The UV/H₂O₂ knowledge-based QSAR models, where the molar adsorption and quantum yield are defined as endpoints, have a moderate ($R^2 \approx 0.6$, $Q^2 < 0.5$; whereas an $R^2 > 0.7$ and $Q^2 > 0.5$ is considered as good, see Figure A.1 in Appendix, available online at [http://www.iwaponline.com/ws/015/109.pdf](http://www.iwaponline.com/ws/015/109.pdf), and respectively bad ($R^2 \approx 0.5$, $Q^2 < 0.4$) predictability when the training set includes all compounds. The prediction capability is slightly improved to an $R^2 = 0.60$, $Q^2 = 0.51$, $Q^2_{\text{ext}} = 0.58$ for the molar absorption constant when the training set is restricted to a subset of aromatic compounds. Compound structure classification does not yield better QSAR statistics for the quantum yield case. Therefore, the presentation of the QSAR results for the molar absorption and quantum yield are omitted.

**Assessment of UV/H₂O₂ knowledge-based QSAR: radical OH reaction constant**

The QSAR model for the prediction of the OH radical reaction constant (Wols & Vries 2012) has been set up by the
The same UV/H₂O₂ model and has a good prediction performance considering the validation statistics. The prediction of the removal degradation efficiency is calculated and further evaluated. Results of the removal degradation are shown in Figure 6. The model seems to explain the variability in compound degradation well, also the low value of metformin is reproduced.

The error bars representing the confidence interval estimates seem relatively large. Closer inspection of the residuals between measurements and model predictions reveal that, in fact, the degradation predictions are quite reliable (Figure 7), since the model residuals for the majority of compounds are within a region of 2 standard deviation units and considering the generally lower removal percentages compared to a NF application. Caution must be taken with the values predicted for certain compounds (paracetamol and phenazone), i.e. those with large uncertainty boundaries (red and blue colored compounds in Figure 7).
CONCLUSIONS

A knowledge- and data-based QSAR case study, with each case study involving the NF membrane rejection and removal by UV/H₂O₂ treatment, shows that the procedure to obtain QSAR models is heavily dependent on selected process conditions, availability of (compound) data and model structure.

Results are therefore a mixed bag, since they very much depend on the accuracy of the model containing the QSAR endpoint and other conditions (data availability for instance). More generally speaking, two aspects of QSAR development for use in water treatment applications are highlighted in this work. Firstly, it is noted that QSAR methodology should be used with care and only when results are validated internally and externally. Secondly, it is shown that translating the statistics to removal efficiency values has the advantage of assessing a QSAR calculation, and moreover, has the additional advantage that the effectiveness of treatment technologies can be estimated and compared beforehand and on a compound basis.

On a case-by-case basis, the following conclusions are made:

- A QSAR for a NF application with high predictability has been derived using the data-based approach ($R^2 > 0.7, Q^2_{\text{ext}} > 0.5$). This QSAR is only valid for Desal HL membranes because of the limitations of the data set used.
- For the UV/H₂O₂ process, moderately performing QSARs were derived for the quantum yield and molar absorption ($R^2 > 0.5$), that do not pass the external validation tests ($Q^2_{\text{ext}} < 0.5$). Possible explanations are (a) photochemical processes are more difficult to represent with these chemical descriptors alone and (b) the scarcity of compound data and literature values. Further work is needed to improve the reliability of the QSAR, by e.g. including more compound data. When considering the OH radical rate constant, the corresponding QSAR not only has high validation statistics, but residual analysis of the QSAR based degradation efficiency calculations showed that the degradation predictions were quite reliable as well.
- The knowledge-based QSAR models rely upon the process models. Since knowledge-based QSARs predict physico-chemical model constants, these QSARs are less prone to varying operating conditions and water
matrices, provided the process model is accurate enough. Improvement of process models is therefore recommended. For example, more reliable predictions in UV/hydrogen peroxide reactors can be established by using advanced (photo)chemical models in combination with computational fluid dynamics (CFD). Those types of models are currently being developed. Also, the process model for NF should be improved to apply knowledge-based QSARs.

This approach has been applied to UV/H$_2$O$_2$ and NF, but can also be used for other treatment processes, such as activated carbon filtration, river bank filtration, ozonation, reverse osmosis and ultrafiltration.

REFERENCES


First received 28 November 2012; accepted in revised form 6 February 2013. Available online 13 September 2013.