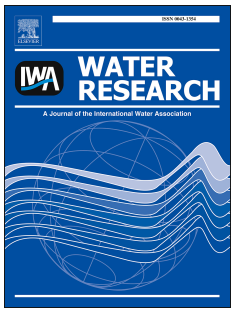
Accepted Manuscript



Predicting chloroform production from organic precursors

Tom Bond, Nigel Graham

PII: S0043-1354(17)30636-X

DOI: 10.1016/j.watres.2017.07.063

Reference: WR 13106

To appear in: *Water Research*

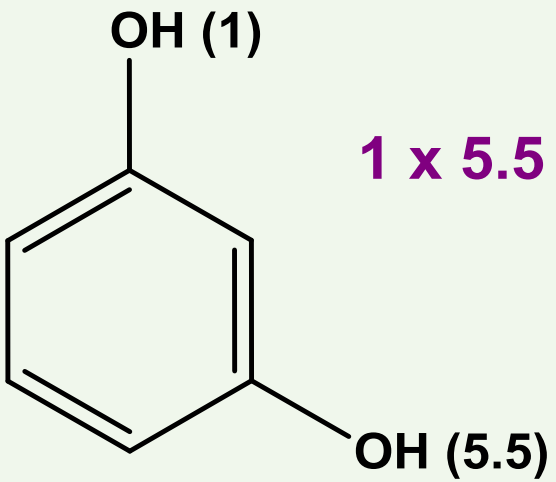
Received Date: 3 May 2017

Revised Date: 17 July 2017

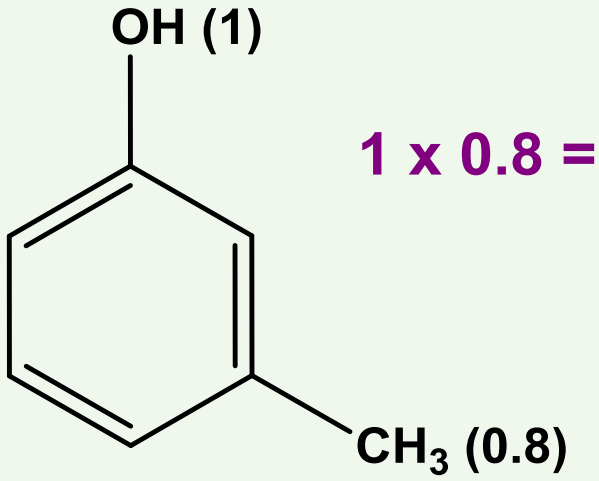
Accepted Date: 23 July 2017

Please cite this article as: Bond, T., Graham, N., Predicting chloroform production from organic precursors, *Water Research* (2017), doi: 10.1016/j.watres.2017.07.063.

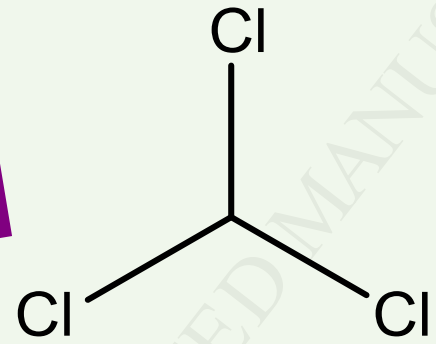
Aromatic score



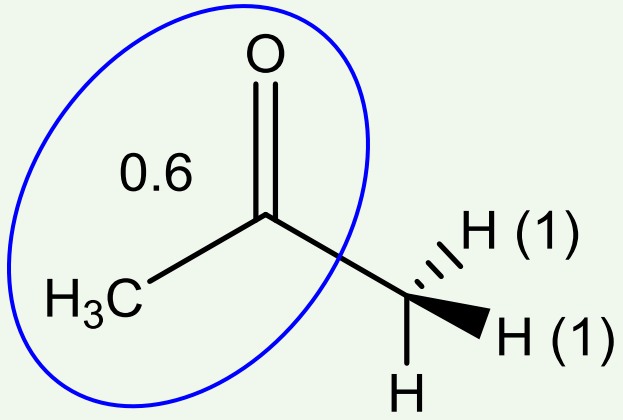
1 x 5.5 = 5.5



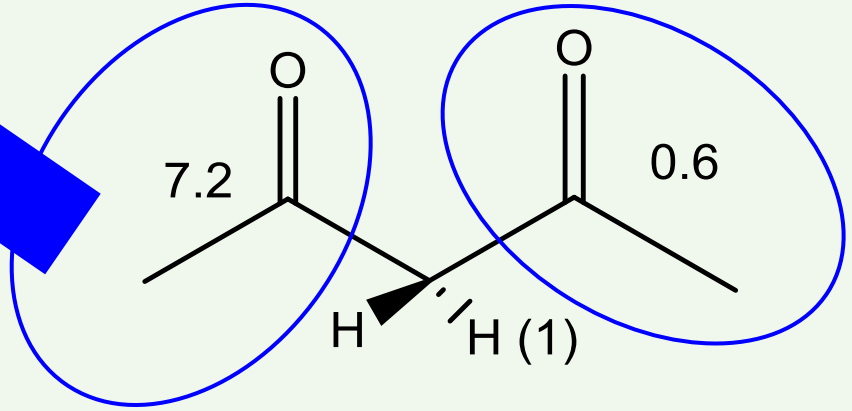
1 x 0.8 = 0.8



Enolizable score



0.6 x 1 x 1 = 0.6



0.6 x 1 x 7.2 = 4.32

Predicting Chloroform Production from Organic Precursors

Tom Bond[1]* and Nigel Graham[2]

1. Department of Civil and Environmental Engineering, University of Surrey, Guildford, GU2 7XH, UK

2. Department of Civil and Environmental Engineering, Imperial College London, London SW7 2AZ, UK

* Corresponding author, Tel: +44(0)1483684474, email: t.bond@surrey.ac.uk

Abstract

Quantitative methods which link molecular descriptors for recognized precursors to formation of drinking water disinfection byproducts are scarce. This study aimed to develop a simple mathematical tool for predicting chloroform (trichloromethane) yields resulting from aqueous chlorination of model organic precursors. Experimental chloroform yields from 211 precursors were collated from 22 literature studies from 1977 onwards. Nineteen descriptors, some established and others developed during this study, were used as inputs in a multiple linear regression model. The final model, calibrated using five-way leave-many-out cross-validation, contains three descriptors. Two novel empirical descriptors, which quantify the impact of adjacent substituents on aromatic and enolizable chlorine substitution sites, were the most significant. The model has r^2 = 0.91 and a standard error of 8.93% mol/mol. Experimental validation, using 10 previously untested precursors, showed a mean discrepancy of 5.3% mol/mol between experimental and predicted chloroform yields. The model gives insight to the influence that specific functional groups, including hydroxyl, chlorine and carboxyl, have on chloroform formation and the relative contributions made by separate substitution sites in the same molecule. It is anticipated that the detailed approach can be updated and extended as new experimental data emerges, to encompass additional precursors and groups of disinfection byproducts.

Keywords: trihalomethanes; THMs; disinfection byproducts; model compounds; QSAR

1 Introduction

The discovery, in the 1970s, that chloroform and other trihalomethanes are generated from chlorination of natural organic matter during water treatment (Rook 1974) surprised the scientific community. This breakthrough led to much research into disinfection byproducts and regulations for the four chlorinated and/or brominated trihalomethanes being introduced

in the USA by the decade's end. The limit for total trihalomethanes is currently 80 $\mu g \cdot L^{-1}$ in the USA, with five haloacetic acids, another group of halogenated disinfection byproducts, regulated at a total of 60 $\mu g \cdot L^{-1}$. Total trihalomethanes are also regulated in the EU at 100 $\mu g \cdot L^{-1}$. Initial concern about the trihalomethanes was based on results of a rodent bioassay which classified chloroform as a suspected human carcinogen (NCI 1976). Independently, epidemiological studies have shown that long-term consumption of chlorinated drinking water is associated with an enhanced risk of developing bladder cancer, although the underlying reasons remain obscure (Hrudey 2009). However, analysis of more recent toxicological evidence indicates that neither the trihalomethanes nor the haloacetic acids are plausible bladder carcinogens at typical drinking water concentrations (Hrudey 2009). Thus, other disinfection byproducts such as the halobenzoquinones (Zhao et al. 2012), various other halogenated aromatic byproducts (Zhang et al. 2008), nitrosamines, haloacetonitriles and haloacetamides (Shah and Mitch 2012) remain the focus of much current research attention, as these may be more toxicologically-significant.

The trihalomethanes and haloacetic acids are essentially viewed by regulators of drinking water quality as indicators of the total occurrence of chlorination disinfection byproducts (USEPA 2015). They also remain the most-studied disinfection byproducts, particularly the trihalomethanes. Model compounds have been heavily used to elucidate mechanistic formation pathways and precursor characteristics from the early days of disinfection byproduct research (Bond et al. 2012, Rook 1977). It was quickly appreciated that meta-substituted aromatic compounds are reactive chloroform precursors. For example, Rook (1977) reported that resorcinol (1,3-dihydroxybenzene) was converted into chloroform at a 85% mol/mol yield during chlorination, whereas yields from its regioisomers, catechol (1,2-dihydroxybenzene) and hydroquinone (1,4-dihydroxybenzene), were far lower at 0.5% and 1.5% mol/mol, respectively (see Figure S2 for a simplified mechanism for chloroform

production from resorcinol). This variation can be (qualitatively) explained by the two activating hydroxyl groups promoting electrophilic substitution reactions at the ortho- and para- positions of the aromatic ring. However, the presence of additional substituents complicates this pattern, as depending on their identity and position, chloroform yields can either be enhanced or suppressed. Thus, de Laat et al. (1982), reported chloroform yields for pyrogallol (1,2,3-trihydroxybenzene), 4-hydroxycatechol (1,2,4-trihydroxybenzene) and phloroglucinol (1,3,5- trihydroxybenzene) as 0.1, 15.5 and 93 % mol/mol, respectively. Certain aliphatic compounds, notably β-dicarbonyl compounds (Boyce and Hornig 1980, Dickenson et al. 2008), including 3-oxopentanedioic acid (Table 1), also act as reactive trihalomethane precursors. Formation of trihalomethanes from carbonyl compounds can be likened to the haloform reaction, used for the synthetic preparation of trihalomethanes from methyl ketones (Larson and Weber 1994). Its rate is controlled by the initial enolization of the organic precursor and the mechanism proceeds via electrophilic addition of chlorine at the carbon alpha to the carbonyl group (Figure S1).

Despite the extensive amount of research effort on this subject over the past ~40 years, quantitative methods to predict disinfection byproduct formation based on molecular descriptors are scarce. Hence, the aim of this study was to develop a simple mathematical method for predicting chloroform yields from model compounds, with the expectation that this will prove a useful screening tool for compounds which have not been tested experimentally. To achieve this, descriptors are required which quantitatively link reactive precursor functionalities to chloroform yields (and ultimately the key pathways leading from one to the other). A secondary aim was that such a mathematical framework would illuminate our knowledge of the characteristics of reactive precursors. Although quantitative structure activity relationship (QSAR) and quantitative structure−property relationship (QSPR) models are widely used in other fields, they have found limited use in disinfection byproduct

research, despite having the potential to streamline research efforts (Chen et al. 2015). While not focussed on disinfection byproducts, the paper by Luilo and Cabaniss (2010) is noteworthy as it details a QSPR, validated using literature on 201 organic compounds, for predicting chlorine demand based on eight molecular descriptors. The same authors subsequently developed a model for predicting chloroform formation from organic precursors (Luilo and Cabaniss 2011b), although this used a smaller subset of 117 model compounds.

2 Methods

2.1 Literature on chloroform formation

Chloroform formation data from 22 studies (Bond et al. 2009, Bond et al. 2014, Bond et al. 2016, Boyce and Hornig 1980, Boyce and Hornig 1983, Bull et al. 2006, Chaidou et al. 1999, Chang et al. 2011, de Laat et al. 1982, de Leer and Erkerlens 1988, Dickenson et al. 2008, Gallard and von Gunten 2002, Hong et al. 2009, Hureiki et al. 1994, Larson and Rockwell 1979, Navalon et al. 2008, Norwood et al. 1980, Rook 1977, Rule et al. 2005, Tawk et al. 2015, Tomita et al. 1981, Westerhoff et al. 2004) spanning the years 1977 – 2016 were collated and converted into units of % mol/mol where necessary. All studies measured chloroform formation from organic precursors under formation potential conditions, i.e. using excess chlorine. However, since there are no standard conditions for these tests, experimental conditions vary in the literature (Table S1). In this study only data collected at pH 7-8 using an excess of chlorine and contact times over 0.5 h were included. Thus, the modelling results only apply to these conditions, which are representative of full-scale drinking water chlorination. Median conditions from the studies included were pH = 7, contact time = 24 h, temperature = 20 °C and chlorine dose = 20 mol/mol. One important difference with the chlorination of natural waters is that model compound studies are typically undertaken in the absence of bromide, and thus chloroform is the only one of the four chloro- and/or bromo-

trihalomethanes monitored. In contrast, ambient bromide in natural waters leads to formation of varying amounts of brominated trihalomethanes.

For compounds tested in multiple studies, mean values were calculated and are given in Table S2. For example, chloroform yields from the well-studied precursor resorcinol have been reported 12 times, giving a mean value of 81.1% mol/mol and a standard deviation of 17.3 (Table S2). In total there were 69 compounds included with multiple chloroform yields. The mean of the standard deviations for these repeated compounds is 5.1% mol/mol. The final list of 211 precursors used for modelling, together with their chloroform yields, structures and alternative names, is given in Table S2.

2.2 Descriptor selection

Three descriptors used by Luilo and Cabaniss (2011b) to model chloroform formation and another three used by the same authors to model total organic halogen formation (Luilo and Cabaniss 2011a) were included. Respectively these are the carbonyl index (CI), the difference between the sum of strong electron-donating groups and the sum of carbonyls per carbon in each molecule (EDCORH), the number of 1,3-activated aromatic carbons (OTactC) (chloroform descriptors) and the number of phenolic groups per carbon (ArOH:C), the square root of the number of heteroatoms (sqHeA) and the log of the hydrogen to carbon ratio (log H:C) (total organic halogen descriptors). Hammett and Taft constants account for substituent effects in aromatic and aliphatic compounds, respectively, and have been used widely in the development of QSARs and linear free energy relationships. In this study they were used in a manner following that described by Lee and von Gunten (2012) and Gallard and von Gunten (2002), who showed that the sums of Hammett or Taft for organic compounds can be quantitatively linked to rate constants for reactions with aqueous chlorine. Taft/Hammett constants were taken from published sources (Hansch et al. 1995, Perrin et al. 1981). Four

descriptors were calculated by summing Taft constants for substituents around enolizable functionalities (Enol Taft and Enolizable Taft), alkenes (Alkene Taft) and amino acids (AmAc Taft). Hammett constants were summed to account for ortho-, meta- and para-interactions in aromatic compounds, as well as total interactions (HammettOrtho, Hammett Meta, Hammett Para and Hammett Sum). Finally, empirical constants were used to develop novel descriptors for five important precursor categories: alkenes (Alkene Score), enolizable aliphatics (Enolizable Score), aromatic ketones (Aromatic Ketone Score), beta-dicarbonyl (BDicarb Score) and aromatic compounds (Aromatic Score). These were calculated by giving each substituent around a potential chlorine substitution site a score, which were then multiplied together to give a combined score for the substitution site. Therefore, these descriptors quantify the influence of specific functional groups around a chlorine substitution site on chloroform formation.

For more complex molecules, scores for individual substitution sites were summed to obtain a total score for the whole molecule. Empirical substituent constants, used to derive a score for a substitution site, were selected to minimise regression residuals, in a similar fashion to some descriptors developed by Luilo and Cabaniss (2010). The 19 descriptors introduced above were used as inputs in a multiple linear regression model in SPSS, with no y-intercept, and with chloroform yield (% mol/mol) as the dependent variable, i.e.:

$$CHCl_3 = \sum_{j=1}^{M} \beta_j x_j$$

Significant descriptors were then identified by successively eliminating the least significant descriptor (that with the highest p value) until all remaining descriptors were significant ($p<0.05$), following previous work (Luilo and Cabaniss 2010, Luilo and Cabaniss 2011b). This gave four significant descriptors: Enolizable Score, SqHeA, OTactC and Aromatic Score. However, a correlation of r = 0.833 between Aromatic Score and OTactC indicated

these two descriptors were not independent (Table S5), so OTactC was removed as it had the higher p value. All other pairwise correlations between these descriptors had $r < 0.7$, indicating they are sufficiently independent to be used in multiple linear regression modelling (Eriksson et al. 2003). The three named descriptors were used in all subsequent modelling procedures and will now be described in more detail. The Enolizable Score multiplies empirical constants for a carbonyl group, denoted as R1, with those for the other substituents (R2 and R3) surrounding an enolizable site (Table 3). An alpha proton combined with a carbonyl group is defined as enolizable, although the proton is excluded from the calculation, as it is present in all enolizable precursors. This classification includes carboxylic acids, as well as aldehydes and ketones. One descriptor value is calculated for each enolizable group, even if multiple enolizable protons are present. Two carbonyls alpha (1,2-) or beta (1,3-) to each other constitute a single enolizable group. Two carbonyls gamma (1,4-) to each other, or more distant, constitute two separate enolizable groups. Assigning a single descriptor value to each enolizable group is reasonable as introduction of one chlorine alpha to an enolizable carbonyl promotes additional chlorine substitution at the same site (Larson and Weber 1994). However, multiple enolizable groups in the same molecule are given separate scores, which are then summed together. For example, for 3-acetylphenol (Table 1) the R1 group is COC_6H_5 (OH meta), while R2 and R3 are both H (Table 3), so the Enolizable Score is 1 x 1 x 1 = 1 (section 4.1 of the supplementary material). SqHeA is defined as the square root of the number of heteroatoms present in a precursor (Luilo and Cabaniss 2011a). For 3-acetylphenol there are two heteroatoms (Table 1), so SqHeA is 1.41 (Table S3). Finally, the Aromatic Score descriptor gives each substituent in an aromatic precursor an empirical substituent score (Table 4). These are then multiplied together to give a combined score for the whole ring, which accounts for interactions between the substituents. Benzene is given a baseline score of 1 and heterocycles 0.2. Depending on the substituents present and their interactions

these values can then either increase or decrease. The Aromatic Score for the ring in 3-acetylphenol is 1 (baseline score for benzene) x 1 (OH in C1 position) x 0.7 ($COCH_3$ in C3 position) = 0.7 (section 4.3 of the supplementary material). Scores for separate aromatic groups are summed to produce the final Aromatic Score value. Multiple detailed examples of how these three descriptors were calculated are given in the section 4 of the supplementary material.

2.3 Model calibration and validation

Initially, the complete dataset (n = 211) was randomly split into training data (n = 158) and an external validation dataset (n = 53). Then the training data was split five times into a calibration subset (n = 111) and an internal validation (cross validation) subset (n = 47), to facilitate leave-many-out cross-validation. For this step, stratified data splitting was used, so that each compound was used at least once for cross-validation (see the supplementary material, section 7). Multiple linear regression modelling was used to obtain a separate equation for each calibration subset of 111 compounds. Each of these five equations was used to predict chloroform yields of the respective cross-validation subsets, with their performance determined by comparing experimental and predicted values. The final predictive model was obtained by averaging coefficients from these five calibration datasets. Chloroform yields of the external validation dataset (n = 53) were predicted using the individual and average equations. The applicability domain was assessed by calculating standardised residuals of cross-validation and leverage for both training and external validation datasets. Using this approach data points can be defined as outliers if they have standardised residuals above +2.5 or below -2.5 (Luilo and Cabaniss 2010). The warning leverage was calculated using $3k'/N$ where k' is the number of descriptors plus one, and N is the number of compounds used to develop the model. Molecules in the training data and external validation which exceed the warning leverage indicate molecules which are: (i) excessively influential in determining the

model parameters, and (ii), are predicted due to over-extrapolation of the model (i.e. they fall outside the applicability domain).

2.4 Experimental validation

Experimental chloroform formation from compounds whose yields have not been reported previously in literature was measured using gas chromatography with electron capture detection (GC-ECD, Perkin Elmer Clarus 500 GC), a modified version of USEPA method 551.1(USEPA 1995) and a Restek Rxi-5 Sil MS column (30 m x 0.25 mm x 0.25 µM). All samples were prepared in duplicate at pH 7 (10 mM phosphate buffer). The method detection limit for chloroform was 0.2 $\mu g \cdot L^{-1}$. Precursor concentrations ranged from 1 to 10 µM (Table 5) and the chlorine dose was 20 mol/mol. All reagents were of at least analytical purity and chlorine concentrations were measured using the DPD-FAS titration method (APHA et al. 2005). After 24 h contact time the residual chlorine was quenched using sodium sulphite (APHA et al. 2005).

3 Results

3.1 Model Performance

Leave-many-out cross-validation produced five QSAR equations (Table S7), which were averaged to generate the final equation shown in Table 2. The average model has $r^2 = 0.91$ and a standard error of 8.93% mol/mol (Table S7). Respective values obtained from regression of chloroform formation across the complete dataset using the same three descriptors were 0.90 and 8.87 % mol/mol (Table S7). The similarity of these values indicates that the data splitting procedures and leave-many-out cross-validation did not introduce any meaningful bias into the average model.

For comparison, chloroform formation was also modelled by the three descriptors (CI, OTactC and EDCORH) used previously for this purpose by Luilo and Cabaniss (2011b),

240 although in the earlier case with a smaller dataset of 117 precursors. Using these same
241 descriptors and the 211 precursors included in the current study gave a model with $r^2 = 0.77$
242 and a standard error of 13.87% mol/mol (Table SI 11). The improved performance of the
243 current model highlights how the selected descriptors more accurately reflect how the
244 functionality around chlorine substitution sites relates to chloroform yields. The OTactC and
245 CI descriptors can only take a limited number of values (e.g. 0 or 1 for OTactC), whereas the
246 Aromatic Score and Enolizable Score descriptors can assume a range of values, depending on
247 the specific chemical identity of the substituents present (see below).

248 Mean statistical parameters for the fivefold leave-many-out cross-validation show the five
249 individual equations have high predictive power, since $r^2 > 0.6$; $q^2 > 0.5$; $0.85 < k$ or $k_0 < 1.15$
250 and $r^2\text{-}r_0^2/r^2 < 0.1$ (Table S8) (Golbraikh and Tropsha 2002). In this context q^2 is the predicted
251 r^2 value; with k/k_0 and r^2/r_0^2 the gradient and r^2 values with/without the y-intercept,
252 respectively. The latter two were obtained from plotting predicted versus chloroform
253 formation. The average model, when applied to the five individual cross validation datasets,
254 as well as to the training set (n = 158) also had high predictive power (Table S9; Figure 1).
255 Thus, it can be concluded that the predictive model is robust.

256 Chloroform yields for the 53 precursors in the external validation dataset were predicted
257 using the average model (Figure 1), as well as for the five individual equations derived during
258 cross validation (Table S10). The mean statistics from the five individual equations are also
259 indicative of high predictive power since they fulfil the criteria noted above. They also agree
260 well with equivalent statistics from the average equation (Table S10). Plotting predicted
261 chloroform yields against standardised residuals (Figure SI 6) shows that there is no pattern
262 amongst the compounds with higher residuals, again evidencing that the developed model is
263 appropriate for the dataset.

Based on the applicability domain analysis, compounds would be classified as outliers if: (i) they have standardised residuals above +2.5 or below -2.5 and (ii) they exceed the warning leverage of 0.095. Based on these criteria no precursors in either the training set or external validation dataset were outliers (Figure S1). Nonetheless, there are a number of compounds which fulfil one of these two conditions. In the training set there are four compounds which have standardised residuals above 2.5. These are phlorizin, arg-lys-glu-val-tyr, 3-oxohexanedioic acid and epigallocatechin gallate. However, in all these cases leverage values are well below the warning value and thus they are not considered influential in deciding the model descriptors. In the external validation dataset, tembotrione has a leverage above the warning value. Thus, its predicted chloroform yield was over-extrapolated, although its standardised residual still lies within the applicability domain (Figure S5). Other authors have reported improved performance when predicting log transformed disinfection byproduct formation using water quality parameters (Obolensky and Singer 2005). However, in this study a model generated with log transformed chloroform data did not improve upon that shown in Table 2 (see Table S6).

3.2 Experimental model validation

The 10 previously untested precursors comprised six phenols and four aliphatic carbonyls (Table 5). The latter include alkene, carboxylic acid and β-dicarbonyl functionalities (Table 5). Predicted chloroform yields, estimated using the coefficients in Table 2, ranged from 0.6% to 21.3% mol/mol, whereas experimental values were from 1.6 – 37.3 % mol/mol (Table 5). In general, experimental and predicted values compared well, with a mean difference of 5.3% mol/mol across the 10 precursors, which is lower than the standard error of 8.93% associated with use of the average model. Nonetheless, for 4-chloro-2-methylphenol, experimental and predicted chloroform yields differed by 16.0% (Table 5). This is a chlorinated phenol, a precursor category discussed in section 5.3.2.

3.3 Insights from the Model

3.3.1 Role of Significant Descriptors

The Aromatic Score and Enolizable Score descriptors are the most significant regarding the model's operation ($p = 0.000$ for both) and their role is to quantify reactivity in these two crucial precursor groups. While the contribution of SqHeA ($p = 0.033$) is less obvious, its negative coefficient (Table 2) indicates chloroform formation from more complex precursors would otherwise be typically slightly overestimated by the model. The absence of any descriptors involving Taft or Hammett constants highlights their limited utility in predicting trihalomethane formation. Since both have been quantitatively linked to rate constants of reactions between organic pollutants and chlorine (Gallard and von Gunten 2002, Lee and von Gunten 2012), in turn this indicates that the kinetics (of the initial reaction step) are not strongly correlated with trihalomethane formation. Since the mechanistic routes which lead to trihalomethane formation are complex and involve multiple steps this is perhaps not unexpected.

3.3.2 Aromatic Precursors

Hammett constants for the substituent OH do not reflect the potency of phenols as chloroform precursors, especially for meta-substituted compounds. The sigma meta value for OH is 0.12 and it is therefore less activating than CH_3 and NH_2, which have sigma meta values of -0.07 and -0.16, respectively. Chloroform yields for resorcinol (1,3-dihydroxybenzene), 3-aminophenol and meta-cresol (3-methylphenol) are 81.1±17.3 %, 13.9±5.9 % and 6.1±0.4 % mol/mol, respectively, which illustrates the difficulties associated with quantitatively linking Hammett constants to trihalomethane formation. Similarly, OTactC scores for the first two of these compounds are both 1. This descriptor, which applies to 1,3-activated aromatic carbons, can only take a value of 0 or 1 (Luilo and Cabaniss 2011b). All precursors ($n = 28$) with a

1,3-activated aromatic carbon have a OTactC value of 1, even though their experimental chloroform yields vary widely, from 4.6 to 98.0% mol/mol.

These comparisons explain why new empirical descriptors were developed during the course of this study. Constants for OH, CH_3 and NH_2 in the C3 (meta) position, as used when calculating the Aromatic Score descriptor, are 5.5, 0.8 and 1.3 (Table 4), which reflects the fact that precursors containing a resorcinol structure are more reactive chloroform precursors than the corresponding cresols or anilines. While aliphatic amine groups, e.g. in amino acids, are typically protonated under water treatment conditions (e.g. the pK_a for ammonium ($^{+}NH_4$) is 9.25), when calculating Aromatic Score and Enolizable Score values no attempt is made to distinguish between the protonated and non-protonated forms.

Another issue with using Hammett constants to explain the formation of halogenated products is that in water disinfection applications they are typically used in an additive manner (Gallard and von Gunten 2002, Lee and von Gunten 2012). Thus, their sum will increase/decrease with an increasing number of electron-withdrawing/electron-donating substituents around an aromatic ring. In turn, 4-hydroxycatechol (1,2,4-trihydroxybenzene) is more activated than resorcinol, whereas phloroglucinol (1,3,5-trihydroxybenzene) is less activated (in terms of the sum of their Hammett constants). However, this does not correlate with trihalomethane formation (Table 1).

In contrast, the Aromatic Score descriptor is a product, so can either increase or decrease as substituents are added to the aromatic ring: respective values for these three precursors are 1.80, 5.50 and 5.50 (Table 4). Another feature of the Aromatic Score descriptor is that it accounts for blocked meta interactions. Aromatic precursors where a resorcinol-type structure is bisected by an additional substituent tend to have low chloroform yields, as the preferred chlorine substitution site is already occupied. For example, pyrogallol (1,2,3-

trihydroxybenzene) generates 0.8±1.4% mol/mol of chloroform (Table 1). When calculating the Aromatic Score descriptors blocked meta interactions, such as in pyrogallol, were not considered.

There is one substituent which represents an exception to this rule: chlorine, as chlorine substitution can either occur where the existing chlorine is located, or elsewhere in the ring. Chlorinated phenols have variable chloroform yields: from 6.8±4.6% for 2,4,6-trichlorophenol to 98% for 4-chloro-1,3-benzenediol (Table S2). Aromatic Score values for a chlorine substituent ortho (C2 or C6) or para (C4) to a strongly-activating group (OH or NH_2) are higher than when more weakly-activating substituents are in the ortho- or para- positions (Table 4). This indicates that additional chlorine substitution proceeds at the carbon bonded to the existing chlorine. Nonetheless, there are several examples of meta-substituted chlorinated phenols which are highly reactive precursors, for example, 3,5-dichlorophenol, which generates 71.7% chloroform (Table S2) and chlorine has the second highest C3 Aromatic Score value of 2.2 (Table 4). Since both chlorine and OH are ortho/para directors (although they are respectively deactivating and activating) this indicates that chlorine substitution occurs elsewhere than the original chlorine group. For other precursors with multiple chlorine groups, e.g. 2,3,4,6-tetrachlorophenol, it is unclear where subsequent chlorination occurs. Predicted and experimental chloroform yields for 4-chloro-2-methylphenol differed by 15.6% (Table 5). This indicates that the empirical descriptor values used to calculate Aromatic Score values could be further optimised if additional experimental chloroform yields were available.

Another substituent which forms an exception in the model is carboxyl (-CO_2). There are a number of phenols containing a CO_2 group in addition to a resorcinol structure which generate similar amounts of chloroform to resorcinol itself. For instance, chloroform yields for 2,4-dihydroxybenzoic acid, 2,6-dihydroxybenzoic acid and 3,5-dihydroxybenzoic acid are

362 83.3±9.6%, 82.5±10.6% and 59.4±19.8%, respectively, versus 81.1±17.3% for resorcinol
363 (Table S2). Similarly, salicylic acid (2-hydroxybenzoic acid), 3-hydroxybenzoic acid and 4-
364 hydroxybenzoic acid all form a comparable amount of chloroform to phenol: from 3.0 to
365 6.9% for the three isomers, compared with 4.8±3.9% for phenol (Table S2). These
366 similarities are explained by decarboxylation of carboxylic acids during chlorination (Larson
367 and Rockwell 1979) and are dealt with by the model ignoring CO_2 groups in aromatic
368 precursors.

369 **3.3.3 Enolizable Precursors**

370 One helpful feature of the model is that, for molecules with multiple substitution sites, it
371 highlights which are primarily responsible for chloroform formation. Whereas benzaldehyde
372 is an unreactive precursor, with a chloroform yield of 0.1%, there are seven aromatic ketones,
373 all containing an acetophenone structure, including 3-acetylphenol (Table 1), with chloroform
374 yields from 10 – 45% (Table S2). Both –CHO and $–COCH_3$ are deactivating substituents
375 with respect to electrophilic aromatic substitution, as shown by respective Hammett meta
376 constants of 0.35 and 0.38. This indicates that chlorination occurs at the $–COCH_3$ group
377 rather than the aromatic ring. In turn, Aromatic Score values are also low - 0.05 for both
378 substituents in the C1 position - whereas the Enolizable Score R1 value for COC_6H_5 (no
379 substituents) is relatively high at 0.8 (Table 3). For 3-acetylphenol, the most reactive aromatic
380 ketone, the model suggests that both the aliphatic $–COCH_3$ group and the aromatic meta-
381 substituted moiety contribute significantly to its experimental chloroform yield of 45%. In
382 contrast, for acetophenone itself, which generates 10% mol/mol chloroform (Table 1), the –
383 $COCH_3$ group is the principal substitution site.

384 Cinnamic acid derivatives also contain aromatic and aliphatic functionalities, with both
385 assumed to be substitution sites. These precursors, and other aliphatic alkenes, are classified

as enolizable as there is a carbonyl (carboxylic acid) alpha to the alkene. Two descriptors – Alkene Taft and Alkene Score – were developed to quantify chloroform production from alkenes but neither are significant in the final model. This is because of the relatively small number of aliphatic alkenes in the dataset and their generally modest chloroform yields, of up to 23% for fumaric acid. While it is expedient for the model to treat these precursors as enolizable; it is likely that chlorination actually occurs via electrophilic addition to the alkene.

Another simplification in the model is how amino acids are treated. Chlorination of this group has been well studied (Bond et al. 2009, Hong et al. 2009, Hureiki et al. 1994) and proceeds via conversion of the amino group to either a nitrile or aldehyde (Figure S4). The proportion of the aldehyde and nitrile products depends on experimental conditions, but as both can be chlorinated (Wyman et al. 1964), the alpha carbon is assigned a composite score in the Enolizable Score descriptor. The low R1 value of 0.2 for amino acids reflects the fact that chloroform yields from alpha amino acids are generally modest. The two exceptions are L-tyrosine and L-tryptophan, where an activated aromatic ring can interact with the enolizable carbon by resonance, as evidenced by a high R2/R3 value of 5.6 (Table 3).

There is only one substituent which enhances chloroform formation to a greater extent, this being $COCH_3$, which has a R2/R3 value of 7.2 (Table 3). This only occurs in the R2/R3 position in certain β-dicarbonyls which are potent chloroform precursors, for example 3-oxopentanedioic acid (Table 1). Molecules with three carbonyl groups around an alpha proton, i.e. β-tricarbonyls, are extremely potent precursors. There are only two precursors in this subset: sulcotrione and tembotrione. This explains their high Enolizable Score, 5.89 for both, and chloroform formation, 91.0% and 99.0%, respectively (Table S2). Meanwhile, 3-oxopentanedioic acid has two separate β-dicarbonyl groups (and enolization sites) (Table 1), which explains its high Enolizable Score value of 4.68 (Table S4).

In contrast, R2/R3 values for other substituents are much lower, which reflects the low chloroform formation of most enolizable compounds. This includes monosaccharides, which the model assumes exist in the linear (aldehyde or ketone) form, rather than as the ether ring. Monosaccharides with multiple enolizable groups - maltose, maltotriose and maltopentaose - have higher chloroform formation, up to 18.9% for the latter. The model sums the contributions from the individual enolizable groups in these precursors.

4 Discussion: Future Research Directions

An empirical model of the type described in this study is only as comprehensive as the experimental data available. For some groups of precursors - phenols, amino acids and monosaccharides - there is extensive data on how the presence of various substituents affects chloroform production. The converse applies for other categories. One is the alkenes, which initially react with aqueous chlorine to produce chlorohydrins (Larson and Weber 1994) and eventually chloroform (Figure S3). It can be hypothesized that the presence of adjacent electron-donating groups will encourage the formation of halogenated products, by making the alkene more nucleophilic. Since the Enolizable (and Aromatic) Score descriptors are empirical this means they can be updated as new experimental data emerges.

Similarly, it is anticipated the detailed approach can serve as a starting point for predicting the formation of other disinfection byproducts. This includes the haloacetic acids and total organic halogen, for which modified descriptors may be developed. At present there is less data available for these groups than for the trihalomethanes. Nonetheless, it is clear that formation of trihalomethanes is accompanied by other groups of halogenated byproducts, with their relative yields dependent on the specific chemical functionality present and proceeding via common precursor structures. For example, β-dicarbonyls can generate significant concentrations of both trihalomethanes and haloacetic acids (Dickenson et al.

2008). Some of the high variability in chloroform yields between studies for 3-oxopentanedioic acid, 59.8±15.9% mol/mol (Table 1), can be attributed to selected experimental conditions favoring one halogenated product over another. As seen from their respective R2/R3 Enolizable Score values of 3.3, 7.2 and 4,6, the substituent COO^- promotes chloroform formation less than $COCH_3$ or $COCH_2CH_3$; with respect to haloacetic acid formation the opposite can be postulated. Something comparable applies to L-tyrosine and L-tryptophan (Table 1), which are also known to generate significant amounts of dichloroacetic acid, trichloroacetic acid and dichloroacetonitrile (Bond et al. 2009). The model is a screening tool for organic molecules whose chloroform yields have not been experimentally tested. It is not designed for use with bulk water quality parameters (e.g. total organic carbon and ultraviolet absorbance). Nonetheless, there is scope to use analyses which provide information about the specific chemical identity of aquatic organics (e.g. gas or liquid chromatography with mass spectrometry detection) to link the model to natural waters

5 Conclusions

This study details and validates a comprehensive mathematical framework to predict the amount of chloroform produced from reactions between aqueous chlorine and organic precursors. The key findings are as follows:

- The final model, calibrated using five-way leave-many-out cross-validation, has r^2 = 0.91 and a standard error of 8.93% mol/mol. It contains three descriptors, the two most significant, developed specifically for this study, empirically quantify the impact of adjacent substituents on aromatic and enolizable chlorine substitution sites.
- Experimental validation, using 10 previously untested precursors, showed a mean discrepancy of 5.3% mol/mol between experimental and predicted chloroform yields.

- Aromatic carboxyl groups are ignored by the model, which accounts for blocked meta interactions. For molecules with multiple substitution sites the model is helpful for evaluating which are primarily responsible for chloroform formation. Notably, the ketone side-group in acetophenone derivatives is a significant source of chloroform formation.

6 Acknowledgments

The first author acknowledges the support of the Imperial College Junior Research Fellowship scheme. Thanks also to Jineesha Mehta for helpful discussions.

7 References

APHA, AWWA and WEF (2005) Standard Methods for the Examination of Water and Wastewater, American Public Health Association, Washington, DC.

Bond, T., Goslan, E.H., Parsons, S.A. and Jefferson, B. (2012) A critical review of trihalomethane and haloacetic acid formation from natural organic matter surrogates. Environ. Technol. Reviews 1(1), 93-113.

Bond, T., Henriet, O., Goslan, E.H., Parsons, S.A. and Jefferson, B. (2009) Disinfection byproduct formation and fractionation behavior of natural organic matter surrogates. Environ. Sci. Technol. 43(15), 5982-5989.

Bond, T., Mokhtar Kamal, N.H., Bonnisseau, T. and Templeton, M.R. (2014) Disinfection by-product formation from the chlorination and chloramination of amines. J. Hazard. Mater. 278(0), 288-296.

Bond, T., Tang, S.C., Graham, N. and Templeton, M.R. (2016) Formation of disinfection byproducts during the preparation of tea and coffee. Environ. Sci.: Water Res.Technol. 2, 196-205.

480 Boyce, S.D. and Hornig, J.F. (1980) Formation of chloroform from the chlorination of
481 diketones and polyhydroxybenzenes in dilute aqueous solution. In Water chlorination:
482 Environmental Impact and Health Effects; vol. 3. Jolley, R.L., Brungs, W. and Cumming, R.
483 (eds), pp. 131-140., Ann Arbor Science, Ann Arbor, MI.
484 Boyce, S.D. and Hornig, J.F. (1983) Reaction pathways of trihalomethane formation from the
485 halogenation of dihydroxyaromatic model compounds for humic acid. Environ. Sci. Technol.
486 17(4), 202-211.
487 Bull, R.J., Reckhow, D.A., Rotello, V., Bull, O.M. and Kim, J. (2006) Use of Toxicological
488 and Chemical Models to Prioritize DBP Research. Report 91135, Awwa Research
489 Foundation, Denver, CO.
490 Chaidou, C.I., Georgakilas, V.I., Stalikas, C., Saraçi, M. and Lahaniatis, E.S. (1999)
491 Formation of chloroform by aqueous chlorination of organic compounds. Chemosphere
492 39(4), 587-594.
493 Chang, H., Chen, C. and Wang, G. (2011) Identification of potential nitrogenous organic
494 precursors for C-, N-DBPs and characterization of their DBPs formation. Water Res. 45,
495 3753 - 3764.
496 Chen, B., Zhang, T., Bond, T. and Gan, Y. (2015) Development of quantitative structure
497 activity relationship (QSAR) model for disinfection byproduct (DBP) research: A review of
498 methods and resources. J. Hazard. Mater. 299(0), 260-279.
499 de Laat, J., Merlet, N. and Dore, M. (1982) Chloration de composés organiques: demande en
500 chlore et réactivite vis-a-vis de la formation des trihalométhanes. Incidence de l'azote
501 ammoniacal. Chlorination of organic compounds: Chlorine demand and reactivity in
502 relationship to the trihalomethane formation. Incidence of ammoniacal nitrogen. Water Res.
503 16(10), 1437-1450.

de Leer, E.W.B. and Erkerlens, C. (1988) Pathways for the production of organochlorine compounds in the chlorination of humic materials. In *Biohazards of Drinking Water Treatment*, Larson, R. A., Ed. Lewis: Chelsea, MI., 1988.

Dickenson, E.R.V., Summers, R.S., Croué, J.-P. and Gallard, H. (2008) Haloacetic acid and trihalomethane formation from the chlorination and bromination of aliphatic β-dicarbonyl acid model compounds. Environ. Sci. Technol. 42(9), 3226-3233.

Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M. and Gramatica, P. (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ. Health Perspect. 111(10), 1361-1375.

Gallard, H. and von Gunten, U. (2002) Chlorination of phenols: kinetics and formation of chloroform. Environ. Sci. Technol. 36(5), 884-890.

Golbraikh, A. and Tropsha, A. (2002) Beware of q2! J. Molecular Graphics Modell. 20(4), 269-276.

Hansch, C., Leo, A. and Hoekman, D. (1995) Exploring QSAR. Hydrophobic, Electronic and Steric Constants, American Chemical Society, Washington, DC.

Hong, H.C., Wong, M.H. and Liang, Y. (2009) Amino acids as precursors of trihalomethane and haloacetic acid formation during chlorination. Arch. Environ. Contam. Toxicol. 56(4), 638-645.

Hrudey, S.E. (2009) Chlorination disinfection by-products, public health risk tradeoffs and me. Water Res. 43(8), 2057-2092.

Hureiki, L., Croué, J.-P. and Legube, B. (1994) Chlorination studies of free and combined amino acids. Water Res. 28(12), 2521-2531.

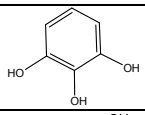
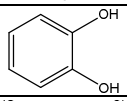
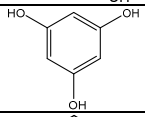
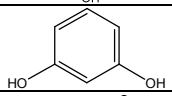
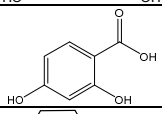
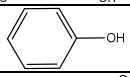
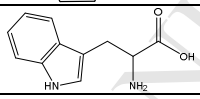
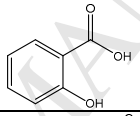
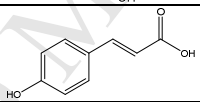
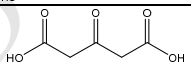
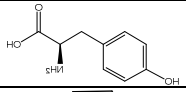
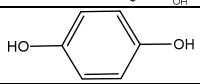
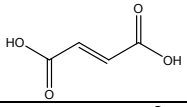
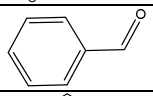
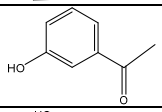
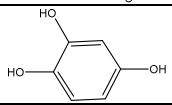
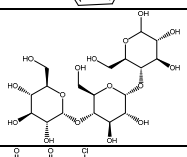
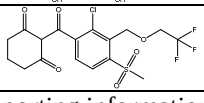
Larson, R.A. and Rockwell, A.L. (1979) Chloroform and chlorophenol production by decarboxylation of natural acids during aqueous chlorination. Environ. Sci. Technol. 13(3), 325-329.

529 Larson, R.A. and Weber, E.J. (1994) Reaction Mechanisms in Environmental Organic
530 Chemistry, Lewis Publishers, Ann Arbor, MI.
531 Lee, Y. and von Gunten, U. (2012) Quantitative structure–activity relationships (QSARs) for
532 the transformation of organic micropollutants during oxidative water treatment. Water Res.
533 46(19), 6177-6195.
534 Luilo, G.B. and Cabaniss, S.E. (2010) Quantitative structure−property relationship for
535 predicting chlorine demand by organic molecules. Environ. Sci. Technol. 44(7), 2503-2508.
536 Luilo, G.B. and Cabaniss, S.E. (2011a) Predicting total organic halide formation from
537 drinking water chlorination using quantitative structure–property relationships. SAR and
538 QSAR Environ. Res. 22(7-8), 667-680.
539 Luilo, G.B. and Cabaniss, S.E. (2011b) QSPR for predicting chloroform formation in
540 drinking water disinfection. SAR and QSAR Environ. Res. 22(5-6), 489-504.
541 Navalon, S., Alvaro, M. and Garcia, H. (2008) Carbohydrates as trihalomethanes precursors.
542 Influence of pH and the presence of Cl- and Br- on trihalomethane formation potential. Water
543 Res. 42(14), 3990-4000.
544 NCI (1976) National Cancer Institute. Report on Carciongenesis Bioassay of Chloroform.
545 NTIS PB-264018, Bethesda, MD.
546 Norwood, D.L., Johnson, J.D., Christman, R.F., Hass, J.R. and Bobenrieth, M.J. (1980)
547 Reactions of chlorine with selected aromatic models of aquatic humic material. Environ. Sci.
548 Technol. 14(2), 187-190.
549 Obolensky, A. and Singer, P.C. (2005) Halogen substitution patterns among disinfection
550 byproducts in the information collection rule database. Environ. Sci. Technol. 39(8), 2719-
551 2730.
552 Perrin, D.D., Dempsey, B. and Serjeant, E.P. (1981) pKa prediction for organic acids and
553 bases Chapman and Hall, New York.

554 Rook, J.J. (1974) Formation of haloforms during chlorination of natural water. Water
555 Treatment and Examination 23(2), 234-243.
556 Rook, J.J. (1977) Chlorination reactions of fulvic acids in natural waters. Environ. Sci.
557 Technol. 11(5), 478-482.
558 Rule, K., Ebbett, V. R. and Vikesland, P. J. (2005) Formation of chloroform and chlorinated
559 organics by free-chlorine-mediated oxidation of triclosan. Environ. Sci. Technol. 39(9),
560 3176-3185.
561 Shah, A.D. and Mitch, W.A. (2012) Halonitroalkanes, halonitriles, haloamides, and N-
562 nitrosamines: A critical review of nitrogenous disinfection byproduct formation pathways.
563 Environ. Sci. Technol. 46(1), 119-131.
564 Tawk, A., Deborde, M., Labanowski, J. and Gallard, H. (2015) Chlorination of the β-
565 triketone herbicides tembotrione and sulcotrione: Kinetic and mechanistic study,
566 transformation products identification and toxicity. Water Res. 76, 132-142.
567 Tomita, M., Manabe, H., Honma, K. and Hamada, A. (1981) Studies on trihalomethane
568 formation of model compounds by aqueous chlorination, In *Proceedings of the 8th*
569 *Symposium on Environmental Pollutants and Toxicology*, *vol. 28*, Sendai, Japan, 1981; pp
570 21-27.
571 USEPA (1995) Method 551.1. Determination of chlorination disinfection byproducts,
572 chlorinated solvents and halogenated pesticides/herbicides in drinking water by liquid-liquid
573 extraction and gas chromotography with electron capture detection. Revision 1.0, Cincinnati,
574 OH.
575 USEPA (2015) Stage 1 and Stage 2 Disinfectants and Disinfection Byproducts Rules. Fact
576 Sheet: Stage 2 Disinfectants and Disinfection Byproducts Rule
577 https://www.epa.gov/dwreginfo/stage-1-and-stage-2-disinfectants-and-disinfection-
578 byproducts-rules (Accessed 23 August 2016).

579 Westerhoff, P., Chao, P. and Mash, H. (2004) Reactivity of natural organic matter with
580 aqueous chlorine and bromine. Water Res. 38(6), 1502-1513.
581 Wyman, D.P., Kaufman, P.R. and Freeman, W.R. (1964) The chlorination of active hydrogen
582 compounds with sulfuryl chloride. II. esters, nitriles, nitro Compounds, and aldehydes. J. Org.
583 Chem. 29(9), 2706-2710.
584 Zhang, X., Talley, J.W., Boggess, B., Ding, G. and Birdsell, D. (2008) Fast Selective
585 Detection of Polar Brominated Disinfection Byproducts in Drinking Water Using Precursor
586 Ion Scans. Environ. Sci. Technol. 42(17), 6598-6603.
587 Zhao, Y., Anichina, J., Lu, X., Bull, R.J., Krasner, S.W., Hrudey, S.E. and Li, X.-F. (2012)
588 Occurrence and formation of chloro- and bromo-benzoquinones during drinking water
589 disinfection. Water Res. 46(14), 4351-4360.
590

Table 1: Selected chloroform precursors, taken from literature*

Name	Alternative name	Structure	Chloroform yield (% mol/mol)
Pyrogallol	1,2,3-trihydroxybenzene		0.8±1.4
Catechol	1,2-dihydroxybenzene		0.9±0.6
Phloroglucinol	1,3,5-trihydroxybenzene		74.6±22.4
Resorcinol	1,3-dihydroxybenzene		81.1±17.3
2,4-dihydroxybenzoic acid	β-resorcylic acid		83.3±9.6
Phenol			4.8±3.9
L-tryptophan			29.5±27.4
Salicylic acid	2-hydroxybenzoic acid		3.0±1.2
p-coumaric acid	4-hydroxycinnamic acid		1.3±0.6
3-oxopentanedioic acid	1,3-acetonedicarboxylic acid		59.8±15.9
L-tyrosine			16.0±16.4
Hydroquinone	1,4-dihydroxybenzene		1.7±1.1
Fumaric acid	(2E)-2-butenedioic acid		23.0
Benzaldehyde			0.1
3-acetylphenol	1-(3-hydroxyphenyl)ethanone		45.0
4-hydroxycatechol	1,2,4-trihydroxybenzene		15.5
Maltotriose			11.8
Tembotrione			99.0

*The full list, including references, is given in the supporting information (Table S2).

Table 2: Average model descriptors and standard errors obtained using leave-many-out cross-validation

	Descriptor (x_j)		
	Enolizable Score	**SqHeA**	**Aromatic Score**
Coefficient (β_j)	17.89	-0.78	14.13
Standard error	0.88	0.49	0.61

Table 3: empirical constants used to calculate the Enolizable Score descriptor

Carbonyl (R1)	Constant value	R2/R3*	Constant value
COO^-	0.1	H	1
CHO	0.1	OH	1.3
$COCH_3$	0.6	CH_2OH	1.1
$COCH_2CH_3$	0.8	OCH_3	1.1
$CONH_2$	0.05	CH_3	0.8
$COCOO^-$	0.05	C_6H_5	1
$COCOCH_2CH_3$	0.4	OC_6H_5	1
COCHOH	0.1	CH_2COO^-	1.5
$CONHCH_3$	0.4	$CONH_2$	1
COC_6H_5 (OH para and OCH3 meta)	0.7	CH_2CH_3	0.8
COC_6H_5 (no substituents)	0.8	$CH_2CH_2CH_3$	0.9
COC_6H_5 (OH ortho)	0.7	CH_2COCH_3	0.4
COC_6H_5 (3 x OCH3 in 3, 4 and 5)	0.8	CH_2NH_2	1.1
COC_6H_5 (OH para)	1.2	$NHCOCH_3$	0.8
COC_6H_5 (2 x OCH3 meta, OH para)	1	NH_2	1.1
COC_6H_5 (OH meta)	1	Inactivated aromatic ring	1
COC_6H_5 (2 x OH ortho and para, ether ortho)	1	Activated aromatic ring	5.6
$COCH_2COO^-$	0.6	COO^-	3.9
$COCH_2COCH_3$	0.3	$CH(OH)CH_3$	1.1
Amino acid	0.2	$NHCH_3$	1.1
		$C(OH)COO^-$	1.2
		CH_2CH_2OH	1.1
		$COCH_2CH_3$	4.9
		CH_2CHO	1.2
		$CH_2CH_2COCH_3$	1.2
		COC_6H_5	2 (R2) or 1.5 (R3)*
		$COCH_3$	7.2
		$CH(CH_3)_2$	0.9
		CH_2CONH_2	1
		SH	1.2
		$CH_2C_6H_5$	1
		$CH=CHNH_2$	0.8
		$CH=CH_2$	1.5
		$CH=(CH_3)(COO^-)$	1.8
		$CH=CHCOCH_3$	0.8
		$CH=CHCOO^-$	1.8

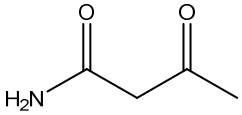
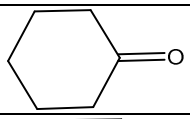
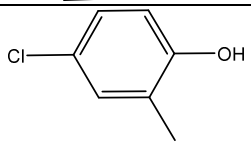
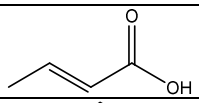
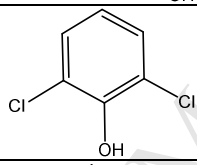
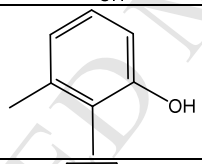
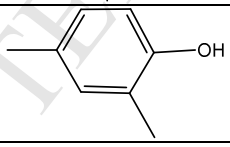
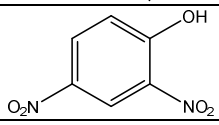
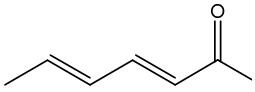
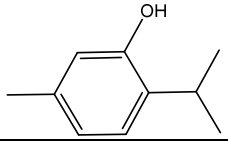
*R2 and R3 value are identical with one exception: where both are carbonyls, which only applies to sulcotrione and tembotrione.

Table 4: empirical constants used to calculate the Aromatic Score descriptor

Baseline values: 1 (benzene) or 0.1 (heterocycles)						
Substituent	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
OH	1	0.3	5.5	0.3	0.5	0.6
Cl	0.1	0.6/*0.9*	2.2	0.6/*1.4*	1.5	0.5/0.7
NH2	0.99	0.3	1.3	0.5	0.4	0.4
N(C2H5)2	1.1					
NO2	0.2	0.2	0.8	0.2		
OCH3	0.3	0.5	0.5	0.5	0.5	0.2
OCH2CH3	0.3	0.3	0.4	0.3	0.3	0.2
OCH=CH2	0.5	0.9	0.9	0.9	0.9	0.5
CH2COO-	0.2	0.3	0.3	0.3	0.3	0.15
CH2OH	0.2	0.4	0.4	0.3		0.3
CH3	0.3	0.85	0.8	0.8	0.8	0.5
CH2CH3	0.3	0.5	0.7	0.5	0.7	0.4
Large alkyl group	0.3	0.4	0.6	0.4	0.5	0.3
CH=CH2	0.25					
COO-	Ignore (see text)					
COOCH3	0.05	0.05	0.05	0.05	0.05	0.05
CHO	0.05	0.2	0.4	0.05	0.05	
COCH3	0.05	0.6	0.7	0.6	0.6	0.6
COCHC=CH2	0.1	0.8	0.8	0.8	0.8	0.8
OCH2COO-	0.05					
CONH2	0.2		0.3	0.05	0.3	
CSNH2	0.3					
NHCOCH3	0.2					
fused ring	0.3	0.8	0.8	0.8	0.8	0.5
Other aryl (not fused)	0.3	0.2	0.4	0.1		
C=O (as part of ring)	0.05	0.5	0.5	0.5	0.5	0.5
CH2CH(NH2)COO-	0.1	0.7	0.8	0.7		
CH2CH2CH2OH	0.3					
CH=NOH	0.3					
CH=CHCHO	0.2	0.4	0.5	0.3		
CH2CH2COO-	0.2	0.3	0.4	0.3	0.3	
CH=CHCOO-	0.2	0.4	0.5	0.3	0.3	
CH=CHCH2OH	0.05	0.4	0.5	0.3		
SO2CH3	0.2	0.3	0.5	0.4	0.4	0.3
I				1		
CN				1		

*Chlorine (Cl) ortho (C2 or C6) or para (C4) to a strongly-activating group (OH or NH_2) takes a higher values than in the presence of other (more weakly-activating) substituents. The higher values are shown in italics above.

Table 5: Experimental chloroform yields from previously untested precursors

Name (alternative name)	Structure	Precursor concentration (µM)	Experimental (and predicted) chloroform formation (% mol/mol)
Acetoacetamide (3-oxobutanamide)		1	7.6±0.7 (9.4)
Cyclohexanone		3	2.0±0.6 (10.5)
4-chloro-2-methylphenol (4-chloro-o-cresol)		1	37.3±0.4 (21.3)
Crotonic acid ((2E)-2-butenoic acid)		3	6.1±4.1 (1.6)
2,6-dichlorophenol		1	8.6±0.9 (7.6)
2,3-dimethylphenol (2,3-xylenol)		3	8.4±0.6 (13.4)
2,4-dimethylphenol (2,4-xylenol)		3	2.9±1.4 (10.9)
2,4-dinitrophenol		10	1.6±1.1 (0.6)
Sorbic acid ((2E,4E)-2,4-hexadienoic acid)		1	3.8±2.2 (1.6)
Thymol (2-isopropyl-5-methylphenol)		3	10.1±0.6 (5.0)

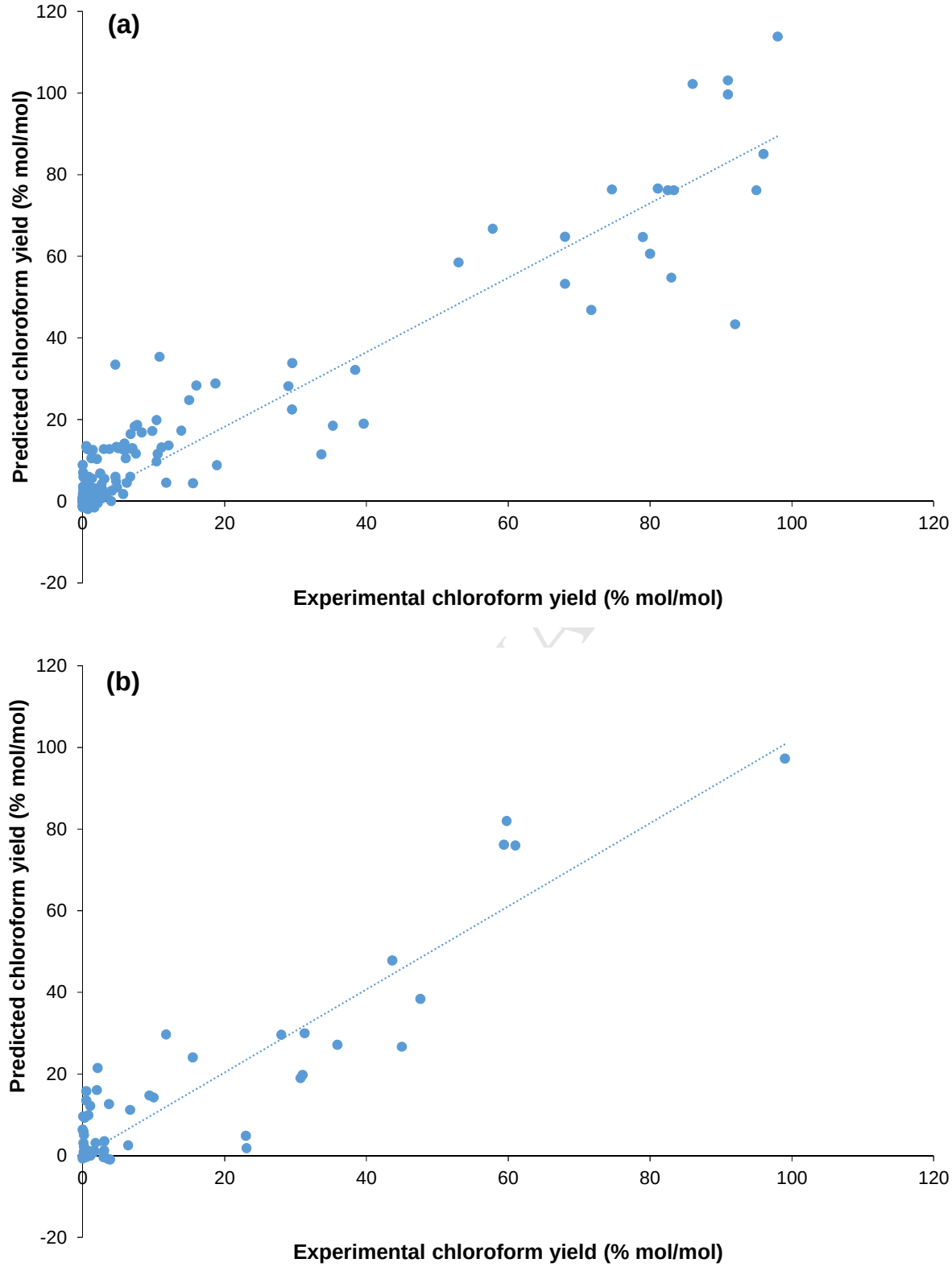


Figure 1: predicted versus experimental chloroform formation for (a) the training set (n = 158; r^2 = 0.91, standard error = 8.75) and (b) the external validation set (n = 53; r^2 = 0.90, standard error = 9.70) using the average model

Highlights

- Chloroform yields from aqueous chlorination of organic precursors were modelled.
- Novel descriptors were developed for the model.
- The final model has $r^2 = 0.91$ and a standard error of 8.93% mol/mol.
- Experimental validation was undertaken with previously unknown precursors.