Group Contribution Method to Estimate the Biodegradability of Organic Compounds

Jhamb, Spardha; Hospital, Irene; Liang, Xiaodong; Pilloud, Francis; Piccione, Patrick M.; Kontogeorgis, Georgios M.

Published in:
Industrial and Engineering Chemistry Research

Link to article, DOI:
[10.1021/acs.iecr.0c03759](https://doi.org/10.1021/acs.iecr.0c03759)

Publication date:
2020

Document Version
Peer reviewed version

[Link back to DTU Orbit](#)

Citation (APA):
Jhamb, S., Hospital, I., Liang, X., Pilloud, F., Piccione, P. M., & Kontogeorgis, G. M. (2020). Group Contribution Method to Estimate the Biodegradability of Organic Compounds. *Industrial and Engineering Chemistry Research*, 59(47), 20916-20928. <https://doi.org/10.1021/acs.iecr.0c03759>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

A Group Contribution Method to Estimate the Biodegradability of Organic Compounds

*Spardha Jhamb^a, Irène Hospital^a, Xiaodong Liang^a, Francis Pilloud^b, Patrick M. Piccione^{c,d},
Georgios M. Kontogeorgis^{a*}*

^aDepartment of Chemical and Biochemical Engineering,
Technical University of Denmark, Building 229, Søltofts Plads 229, DK – 2800,
Kgs. Lyngby, Denmark.

^bSyngenta Crop Protection SA, Route de l’Ile au Bois, 1870 Monthey, Switzerland

^cSyngenta, Breitenloh 5, CH 4333, Münchwilen, Switzerland

^d Current address: F. Hoffmann-La Roche AG, Grenzacherstrasse 124, 4070 Basel, Switzerland

*Correspondence to Georgios M. Kontogeorgis,
KT Consortium, Department of Chemical and Biochemical Engineering,
Technical University of Denmark, Building 229, Søltofts Plads 229,
DK – 2800, Kgs. Lyngby, Denmark.

Email: gk@kt.dtu.dk; Telephone: +45 45 25 28 59; Fax: +45 45 88 22 58

ABSTRACT

The study of biodegradation of chemicals that are discharged into the water bodies is important to determine the risk to the environment that is associated with the chemical. The biodegradability of a chemical can be characterised by conventional oxidation parameters such as biochemical oxygen demand (BOD), chemical oxygen demand (COD) and theoretical oxygen demand (ThOD). Since the BOD₅ estimation method takes five days, other methods have been sought for measuring oxygen requirements of a sample. One of them is the chemical oxygen demand (COD) test which makes use of a strong oxidising agent under acidic conditions.

In this paper a combination of BOD₅ and COD experimental data is used to develop a quantitative model for the prediction of biodegradability. Furthermore, the criterion used to classify a substance as 'readily biodegradable' is the one specified by the European Chemical Agency (EChA)'s CLP (Classification, Labelling and Packaging) Regulation.

KEYWORDS: Biodegradation, Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), European Chemical Agency (EChA) CLP (Classification, Labelling and Packaging) Regulation, Group Contribution Model

1. Introduction

1.1 Defining Biodegradation

Biodegradation is a complex process consisting of many steps that critically depend on chemical structure, environmental conditions into which a chemical is released and the bioavailability of the chemical. Qualitatively, biodegradation is defined as the ability of living organisms to reduce or eliminate environmental hazards resulting from accumulations of toxic chemicals and other hazardous wastes¹.

In biological terms, this implies the breakdown of organic matter by micro-organisms e.g. bacteria and fungi, which use the organic matter as nutrients and decompose it into non-hazardous components: water, carbon dioxide and other minerals as per the following equation¹,

Under aerobic conditions:

Under anaerobic conditions:

Some microorganisms obtain energy from fermentation and anaerobic oxidation of organic carbon. Many anaerobic organisms (anaerobes) use nitrate, sulfate, and salts of iron (III) as practical alternatives to oxygen acceptor. For example, in the anaerobic reduction process of nitrates, sulfates, and salts of iron (III) the following reactions take place:

Briefly, biodegradation involves different stages. First, bio-deterioration occurs wherein, a change of mechanical, chemical and physical properties of the material takes place. The structure is weakened by external abiotic factors such as compression, light and temperature. Next, bio-fragmentation occurs wherein, the polymers bonds cleave. This is highly regulated by the presence of oxygen. Lastly, assimilation occurs wherein the products of biodegradation are integrated into microbial cells.

Biodegradation can happen under different environmental conditions that can enhance or reduce its rate. Important factors for its success are:

- The presence of metabolically capable micro-organisms
- A suitable environment (as the presence of oxygen)
- A temperature between 10 and 35 degrees: keeping a substance frozen or below the optimal operating temperature for microbial species, can prevent biodegradation
- The presence of water
- An appropriate level of nutrients and contaminants
- A favorable pH and alkalinity as microbial species that degrade the different compounds survive in a specific pH range. The optimum pH for hydrocarbons degradation is 7 and a globally suitable one is between 6 and 8

In this work, only the case of aerobic condition is used due to the availability of sufficient data for developing a quantitative model to predict the biodegradability of organic compounds. Therefore, from this point further, the term 'biodegradation' implies 'aerobic biodegradation'.

Quantifying Biodegradability

Biodegradability is a difficult property to model owing to the complexity of the biodegradation phenomenon as described above. Specifically, the biodegradability of a chemical can be quantitatively estimated by conventional oxidation parameters for bio-degradation according to the Organization for Economic Cooperation (OECD) report².

Biochemical Oxygen Demand (BOD): Biochemical oxygen demand (mg) is the amount of oxygen consumed by micro-organisms when metabolizing a test compound.

Chemical Oxygen Demand (COD): Chemical oxygen demand (mg) is the amount of oxygen consumed during oxidation of a test compound with hot, acidic dichromate. It provides a measure of the amount of oxidizable matter present.

Theoretical Oxygen Demand (ThOD): Theoretical oxygen demand (mg) is the total amount of oxygen required to oxidize a chemical completely. It is calculated using stoichiometry and is also expressed.

Total Organic Carbon (TOC): Total organic carbon of a sample is the sum of the organic carbon in solution and in suspension.

The BOD, COD and ThOD are all expressed as mg oxygen uptake per mg test compound. The TOC on the other hand is expressed as mg organic carbon per L test solution.

It is possible to determine whether a particular compound is biodegradable or not, using the oxidation numbers described above. Different criteria are used in literature regarding these numbers. The most used one is the oxidation number BOD₅, which represents the BOD calculated during 5 days of incubation of a sample containing the compound at 20 degree Celsius in the dark.

This involves calculating the consumed oxygen during that period to estimate the biodegradability of the sample. However, five days are rarely enough to reach a complete biodegradation expressed by the ‘ultimate biochemical oxygen demand: BOD_u’. The BOD₅, therefore represents only around 77 percent of the complete reaction.

According to the European Chemical Agency (EChA)'s CLP (Classification, Labelling and Packaging) Regulation in connection with degradation, a substance is considered to be rapidly biodegradable in the environment if one of the following three criteria are satisfied³:

(a) Readily biodegradable in biodegradation screening test

(b) Ratio of the five-day biochemical oxygen demand (BOD₅) to chemical oxygen demand (COD):

$BOD_5/COD \geq 0.5$

(c) Other convincing scientific evidence that the substance can be degraded in the aquatic environment to a level > 70 percent within a 28-day period; for instance, if the BOD₂₈ is greater than 70 percent of the ThOD, then the substance is considered to be readily biodegradable

A more complete way to define the range of biodegradability^{4,5} is as follows:

$BOD_5/COD > 0.5$: easily biodegradable

$BOD_5/COD = 0.4 - 0.5$: middle

$BOD_5/COD = 0.2 - 0.5$: slowly biodegradable

$BOD_5/COD < 0.2$: not biodegradable

Apart from the EU regulations, there are some other regulations to estimate the pollution of effluents according to different locations. These are briefly summarized below⁶:

- In Asia, BOD₅ value is directly used: between 5-15 mg/L means moderately polluted and a BOD₅ above 15 mg/L leads to severely polluted environment.
- In France, a BOD₅ < 100 mg/L and a COD < 300 mg/L are limited for effluent open waterways.
- In Germany, the COD needs to be inferior to 4 times the TOC to define the environment as acceptable.

However, the oxidation parameter TOC has recently come into attention as its experimental value can be obtained much faster. Correlations between the TOC, COD and BOD₅ parameters have been established so the use of TOC could accelerate the establishment of biodegradation.

1.2 Models for Prediction of Biodegradability

1.2.1 Quantitative Structure-Activity Relationship Models

Chemical descriptor based

Grammatica et al.⁷ proposed models to quantify the biodegradability of a chemical based on theoretical descriptors, where a genetic algorithm (GA) was used as the variable subset selections technique to search for the most relevant molecular descriptors. A diverse data set consisting of 71 alcohols, ketones and aromatic compounds, 15 anilines and phenols, 17 polychlorinated biphenyls and 43 heterogeneous compounds with BOD and percent of theoretical oxygen demand (TOD) data from the literature was used.

Each descriptor present in the compounds is denoted by an information bit equal to one if it is present in the regression model and equal to zero if it is absent from the model. A population constituted of zero/one strings is evolved by the GA. The developed regression models for BOD and %ThOD had a satisfactory performance ($R^2 = 82\text{--}84\%$ and $Q^2 = 78\text{--}82\%$)⁸.

Fragment based (Group Contribution)

The US-EPA (Environmental Protection Agency) has already developed seven group contribution models for assessing the biodegradability of a compound⁹. They have developed different mathematical models: Four models that predict the probability of biodegradation, two models that predict the time needed for a compound to achieve an either primary or ultimate biodegradation and one model based on anaerobic biodegradation. All these seven models are available via the software - EPI Suite¹⁰, which can be downloaded freely from the US-EPA website and they are briefly described below:

a) BIOWIN1 and BIOWIN2: These two models predict the probability of biodegradation using experimental values from the BIODEG database (National Information Services Corporation (NISC) U.S., 2000). A probability value of greater than 0.5 indicates that the compound is rapidly biodegradable. The development of these models makes use of 295 chemicals in the database and a linear and nonlinear regression is performed to estimate the parameters for BIOWIN1 and BIOWIN2 respectively.

For carrying out the regression, every compound in the set studied was defined as ‘rapidly biodegradable’ or ‘not rapidly biodegradable’ based on the experimental evaluation of biochemical oxygen demand (BOD), CO₂ production, loss of parent molecule properties etc. Then, depending on whether they were rapidly biodegradable, an indicator variable was assigned the value of 1 or 0 respectively.

Using 37 independent variables namely the frequency of structural fragments, f_i (number of times a substructure occurs in a molecule) and the molecular weight, M_w , the models to predict the probability for chemical j to biodegrade, Y_j are developed.

- The linear model is given by the following equation:

$$Y_j = a_0 + a_1f_1 + a_2f_2 + \cdots + a_{36}f_{36} + a_mM_w + e_i \quad (5)$$

- The non-linear model is expressed as follows:

$$Y_j = \frac{\exp(a_0 + a_1f_1 + a_2f_2 + \cdots + a_{36}f_{36} + a_mM_w + e_j)}{1 + a_0 + a_1f_1 + a_2f_2 + \cdots + a_{36}f_{36} + a_mM_w + e_j} \quad (6)$$

where,

- a_i = regression coefficient
- e_j = error term
- a_m = regression coefficient of the molecular weight
- M_w = molecular weight

b) BIOWIN3 and BIOWIN4: These two models predict the time frame for primary and ultimate biodegradation respectively. They are based on a survey conducted by a panel of 17 experts.

They studied the tendency of compounds to biodegrade by defining score numbers in terms of scale of reaction time: 5 = hours; 4 = days; 3 = weeks; 2 = months; 1 = longer. Then, an arithmetic mean was applied on these values reported by different scientists for each compound.

c) BIOWIN5 and BIOWIN6: Like BIOWIN1 and BIOWIN2, these models also predict the probability of biodegradation. However, the development of them made use of 884 compounds from the Japanese MITI (Ministry of International Trade and Industry) database.

The experimental data for these compounds is based on respirometry analytical methods which measures oxygen consumption. This test assigned the property of being readily biodegradable if the $BOD_{28} \geq 60\%ThOD$. This does not take into account the COD parameter. Also, the group contribution models used are based on 42 fragments instead of 36.

d) BIOWIN7: This is a screening-level model for the prediction of anaerobic ultimate biodegradation in sludge digesters. It was developed from a data set of 169 chemicals tested for methanogenic anaerobic biodegradation using the serum bottle method. There are controversies on the validity of aerobic models for anaerobic conditions². Regarding the molecular fragments used in BIOWIN models, the same coefficient for the different models were found. However, there are other rules which apply to anaerobic conditions specifically as unsaturated (olefinic) hydrocarbons degrade much faster than saturated hydrocarbons².

It is worth noticing that the ratio used by BIOWIN5 and BIOWIN6 of $BOD_{28} \geq 60\%TOD$ is one of the possible parameters to define a compound as readily biodegradable. Many others, as Dissolved Oxygen Content (DOC) and CO_2 production can also be used¹¹.

Their method has different advantages: even though it lacks precision, it still provides quantitative or semi-quantitative estimates of biodegradation rate for use in chemical ranking schemes and also estimates the probability of rapid biodegradation. Still, the fragment constant approach to biodegradability modeling is simplistic and does not, for example, take into account the possible interactions among fragments in multifunctional molecules. Moreover, the sum of the coefficients of the regressed equations have low reliability as the number of compounds used to estimate them are very few. For example, very few compounds which contain the aromatic F, N-nitroso and aliphatic Br fragments are used in the regression. The ProCAMD tool in ICAS¹² enables the user to make use of some of the above biodegradation models (BIOWIN1 - BIOWIN4). The performance of these models is satisfactory with a $R^2 \approx 0.7$.

1.2.2 Inductive Machine Learning Models

Artificial intelligence techniques to model biodegradability were presented by Gamberger et al.¹³ and Gamberger et al.¹⁴. They derived rules using inductive machine learning methods based on

structural requirements in order to predict slow and fast biodegradation. A total of 341 compounds comprising of 293 substances from the *BIODEG* database and a set of expert judgements for 48 chemicals is used as the dataset to develop the three simple structure-based rules. This makes use of an example-based learning system.

The model was able to correctly classify 80% of readily biodegradable predictions and 90% of not readily biodegradable predictions. The defined rules are based on combinations of structural groups that allow neighboring structures to be considered.

2. Development of a Group Contribution Model to Predict Biodegradability

The group contribution models are an efficient method for property prediction due to their simplicity and computational efficiency. Although, the US-EPA has developed seven group contribution models for biodegradability prediction, they do not cover an exhaustive range of organic compounds due to the small number of molecular descriptors used. Therefore, in this work, a GC model that can be used for a more comprehensive range of compounds, is developed.

2.1 Methods and Tools Used for Model Development

A systematic methodology based on the property-data-model-analysis is used¹⁵ to develop a GC-model. The methodology includes three steps. First, a trend analysis of the property values (obtained through a literature review or by conducting experiments) is performed for various classes of compounds as a function of their molecular weight. Second, the most appropriate form of the property model equation, $f(x)$, based on the observed trends is selected. Lastly, parameter regression using all the available datapoints in order to obtain model parameter values (group contributions), is performed. This in turn can be used to predict the values of the physical property.

After the model has been developed, an analysis of prediction errors (that is, the difference between the experimental value and the predicted value) for all compounds in the dataset is made. The model performance is evaluated by the statistical parameters such as coefficient of determination (R^2), standard deviation (SD), average absolute error (AAE) and average relative error (ARE). If the calculated model performance statistics are acceptable, then the estimated GC-model parameters are considered as final. Otherwise, the steps in the procedure described is repeated until the desired correlation statistics are achieved.

While developing property models, very often, the experimental dataset is divided into training and validation sets. However, while developing a GC-model, it is better to use all the available experimental datapoints in the parameter estimation step as it results in lower uncertainties of estimated model parameters and hence lower uncertainties (better reliability) of the predicted property values¹⁵. The difference in the change in predictivity of the model on including the validation dataset for parameter estimation, has been highlighted in this work.

2.1.1 The Marrero and Gani (MG) Method

In the MG GC-method, the property estimation is performed at three levels. The first level has a large set of simple and monofunctional groups that represent the complete organic compound. The second level of estimation involves polyfunctional, polar/nonpolar compounds of medium size ($C = 3-6$); aromatic or cycloaliphatic compounds with only one ring and several substituent groups that provide better description of proximity effects and can differentiate among isomers. The third level of estimation includes groups that provide more structural information for molecular fragments of compounds, whose description is insufficient through the first and second level groups.

The MG GC-model has the form¹⁶,

$$f(X) = \sum_i N_i C_i + w \sum_j M_j D_j + z \sum_k O_k E_k \quad (7)$$

The function $f(X)$ is a function of property X and it may contain additional adjustable model parameters (universal constants) depending on the property involved. In Eq. (7), C_i is the contribution of the first-order group of type- i that occurs N_i times. D_j is the contribution of the second-order group of type- j that occurs M_j times. E_k is the contribution of the third-order group of type- k that has O_k occurrences in a component. Eq. (7) is a general model applicable to all properties and the definition of $f(X)$ is specific for each property X .

An example showing the molecular structure representation of the compound: N-[(2-Hydroxyphenyl)methyl]-4-phenylbenzamide using first-order, second-order and third-order groups of the MG GC-method is given in Table 1.

Table 1. Molecular structure representation of N-[(2-Hydroxyphenyl)methyl]-4-phenylbenzamide using the Marrero-Gani (MG) groups

First-order groups ^a / their occurrences	Second-order groups ^a / their occurrences	Third-order groups ^a / their occurrences
aCH / 13	aC-CH _n -NH _m (n in 1..2; m in 0..2) / 1	aC-aC (different rings) / 1
aC / 2	AROMRINGS ₁ S ₂ /1
aC-CH ₂ / 1	AROMRINGS ₁ S ₄ / 1
aC-OH / 1
aC-CONH / 1

^aThe marked circles in the drawing of the structure of N-[(2-Hydroxyphenyl)methyl]-4-phenylbenzamide show the listed groups in respective symbols.

2.1.2 Data Sets

Based on EChA's second criterion of biodegradability, the need for a model to predict the ratio of the oxidation indicators, BOD₅ and COD, is recognized. In order to develop a GC model for this, experimental data to estimate the parameters of the model, is required. In this study, the BOD₅ and COD values for organic compounds measured in mg O₂.L⁻¹ were collected. These are compiled as four databases (Database 1 - Database 4) explained below.

Database 1: Simple Acyclic Compounds and Aromatics

BOD₅ and COD experimental data to estimate the parameters of the model for simple acyclic compounds and aromatics were collected. The data is collected from various handbooks and scientific publications¹⁷⁻²⁰ containing environmental data. This database is available in the Supporting Information (Table S1).

Database 2: Benzene Ring - Containing Aromatics

The second database comes from the work of Cvetnic et al.²¹ about the prediction of biodegradability of aromatics in water using QSAR modeling. It makes use of experimental values, obtained from colorimetric method using HACH DR2800 spectrophotometer. Also, its particularity is that it is focused on aromatics containing a benzene ring. This type of structure is the most widespread chemical structure in nature. Therefore, the number, type and position of its substituents are analyzed. Here, 36 single benzene rings compounds are part of the study.

Their biodegradability as the ratio BOD₅/COD has been calculated and then, a UV-C/H₂O₂ treatment has been conducted on these compounds. As a response, the half-life of parent pollutant has been measured each time. It is worth noting that this definition has never been previously

mentioned in the literature survey, but this value could be used to analyze the trend of biodegradability of this specific type of compound. The experimental conditions for these different ratios were a temperature of 25 +/- 0.2 °C and a time of treatment from 8 to 75.5 min (corresponding to reaching of half-life ($t_{1/2}$) for each molecule) with their model solutions ($c_0=1$ mM and $V=1.4$ L).

The database consisting of these compounds is available in Supporting Information (Table S2).

Database 3: Acyclic Compounds, Aromatics and Complex Pesticides

This database comes from a scientific publication²² which is partly an extract from the “ecoinvent” database (starting from the 45th compound) and partly comes from their "in house" analysis which was made to develop molecular structure-based models of chemical inventories using neural networks. The “ecoinvent” database is a Swiss organization which gathers information on thousands of compounds since 2003. This information is used by companies to make more rational environmental choices.

In this database, (database – 3) ten herbicides/pesticides are present. These compounds contain complex chemical formulas and are important compounds in the analysis of biodegradability as they are directly incorporated in the soils. For example, atrazine which is an herbicide widely used in the United States and Australia, was banned in Europe in 2003 for its toxicity²³.

This database also includes compounds part of BTEX which is a term used in the petrochemical industry and petroleum refining to describe mixtures of benzene, toluene, and the three xylene isomers, all of which are aromatic hydrocarbons. Those compounds are widely used in the industry and can be very toxic to the environment. Their analysis is, with thus essential, to deal current environmental issues.

This database also contains two compounds containing fluorine, which was not present in the previous database. Also, there are seven amine compounds, which are compounds widely analyzed for their biodegradability as they are considered to be much less degradable than amides. This database is available in Supporting Information (Table S3).

Database 4: Deicers

This database comes from the work by Corsi, Mericas & Bowman²⁴ and is available in Supporting Information (Table S4). This study is about the oxygen demand of aircraft and airfield pavement deicers and alternative freezing point depressants. Here, BOD₅ and COD values are given and thus the ratio has been calculated for each compound.

Globally, it was found that ethylene glycol-based products had lower degradation rates than propylene glycol-based products and that acetate-based products had greater percentage degradation than glycols.

Database 5: Database from Syngenta AG

This database comprises of 45 compounds with BOD₅ and COD values measured by Syngenta AG. The compounds in this database are complex agrochemicals. Initially, this database is used only as the validation set for the developed model. However, at a later stage, it is also used as a training set for the model.

2.1.3 Selection of Suitable Property Model Function

An appropriate function $f(X)$ for the ratio of the two oxidation parameters which is modelled in this work was chosen by studying the behaviour of the experimental values for organic compounds with the increasing number of carbon atoms for each class / family of compounds. The ThOD shows an increasing trend with the increase in carbon number as seen from Figure 1.

Figure 1. Trend of ThOD with increasing Number of Carbon Atoms

This trend is expected and can be calculated from stoichiometry. As an example, the complete oxidation of primary alcohols is considered. This is given by the equation shown below,

After full combustion one mole of this alcohol will generate n mole CO_2 and $(n+1)$ mole H_2O . In total, one requires n and $(n+1)/2$ mole O_2 to make these products, respectively. The oxygen atom of the alcohol means that we have $0.5 O_2$ to start with. So, the total oxygen required for the full combustion is $1.5n$ mole.

Therefore, ThOD = mass O₂ / mass alcohol

$$= 1.5 n * 32 / (12n + 2n + 2 + 16)$$

$$= 24n / (7n + 9)$$

This function has an asymptotic limit of $24/7 \sim 3.4$, which is in line with Figure 1. From the entire dataset, it was possible only to identify seven families and classify the data accordingly. Apart from the trend analysis of the ratio of BOD₅ to COD, for which the model is developed, the trend analysis for BOD₅ and COD has been individually carried out as well, in order to determine if their trend is similar to that of ThOD, which is calculated from the stoichiometry of an oxidation reaction. However, a monotonic trend is not followed by the individual oxidation parameters, BOD₅ and COD (Figure 2 – Figure 3) as well as for the ratio of these oxidation parameters (Figure 4). It has been reported that, in the case of n-alkanes, the ones with longer chains have low solubility and consequently low bioavailability compared to the mid-sized n-alkanes. Moreover, although short-chain n-alkanes have higher aqueous solubility, they are toxic to microbes or cells, and the mechanisms of this toxicity are associated with their uptake and dissolution in the cell membrane²⁵. Thus, though short chain n-alkanes are more bioavailable than mid-sized or long chain n-alkanes, they are associated with cell toxicity and hence inhibition of microbial growth resulting in lower BOD values. This explains the observed trend in the property with increase in the chain length for other classes of organic compounds as well.

Figure 2. Trend of BOD₅ with increasing Number of Carbon Atoms

Figure 3. Trend of COD with increasing Number of Carbon Atoms

Figure 4. Trend of Ratio of BOD_5 to COD with increasing Number of Carbon Atoms

Nevertheless, an appropriate function that can model the trend for most organic classes of chemicals was chosen, considering the fact that the ratio of BOD₅ to COD, denoted by ε will always be positive, and could have the form shown below:

$$\varepsilon = \left(\sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k \right)^2 + \varepsilon_0 \quad (8)$$

$$\Rightarrow (\varepsilon - \varepsilon_0)^{1/2} = \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k \quad (9)$$

In Eq.8 and Eq.9, ε is the estimated ratio of the BOD₅ to COD and ε_0 is the universal model parameter.

2.1.4 Parameter Estimation

All the group contribution parameters (C_i , D_j and E_k) together with the universal constant (ε_0) are estimated simultaneously ($p \equiv [C_i, D_j, E_k, \varepsilon_0]$) using a special case of maximum likelihood theory – the least squares method. This is done by minimizing the cost function shown in Eq. 10 wherein y denotes the available experimental data and $\varepsilon(p)$ is the model function shown in Eq. 9. This method assumes that the measurements are independent and identically distributed with white measurement errors having a known standard deviation, σ (Gaussian). The objective function, $S(y,p)$ to be minimized is,

$$S(y, p) = \frac{(y - \varepsilon(p))^2}{\sigma^2} \quad (10)$$

This minimization is carried out in MATLAB using the `fminunc` and `lsqnonlin` functions. The gradient of $S(y, p)$ with respect to p is analytically calculated and supplied in order to define the search direction for parameter estimation using the `fminunc` function.

2.2 Model Performance Statistics

Test of databases one by one

As a comparison, the first three of the four databases have been tested separately since the measurements are carried out by different research groups / organizations. This check is not performed for Database – 4 as the number of groups representing the compounds in the database is more than the number of compounds. Moreover, a combined database comprising of compounds from Databases 1 – 4, is also tested. Table 2 presents the different performance statistics for each database.

Table 2 Model Performance Statistics when Parameter Regression is Performed for Each Database Separately

Database	R ²	AAD	Number of Compounds	Number of Groups
Database – 1	0.95	55.7%	121	90
Database – 2	0.72	33.7%	36	25
Database – 3	0.78	324.8%	102	94
Combination of the Databases 1 - 4 (after removing the common compounds)	0.69	133%	232	134

The corresponding parity plots are shown in Figure 5 – Figure 8.

Figure 5. Parity Plot for Database – 1

Figure 6. Parity Plot for Database – 2

Figure 7. Parity Plot for Database – 3

Figure 8. Parity Plot for the Combined Database (1-4)

Test of pairs of databases

As the next step, following the results in the previous section, some databases have been mixed in order to further analyze the results and identify the best combinations and the following performances are obtained, as shown in the first four columns of Table 3. The parity plots for each of the three cases are presented in Figure 9 – Figure 11.

Table 3. Model Performance Statistics when the Parameter Regression is Performed for Different Pairs of Databases

Database	R ²	AAD	Number of Compounds	Number of Groups	Number of Outliers	R ² after Removing Outliers
Database - 1 + Database - 3	0.74	112%	192	124	17	0.85
Database - 2 + Database - 3	0.91	16%	106	99	9	--
Database - 1 + Database - 2	0.9	30%	140	99	10	0.95

Moreover, for each of these combinations shown in Table 3, the number of outliers using the definition of Cook's distance²⁶, which shows the influence of each observation on the modelled values, have been determined. According to this definition, an observation is classified as an outlier if it exceeds three times of the mean Cook's distance. The performance of the model has been calculated again with those outliers removed. Sometimes, for small databases after deleting the outliers, the number of groups can become more than the number of compounds. This is the case for the 'Database - 2 + Database - 3' combination and hence the parameters cannot be estimated in that case.

Figure 9. Parity Plot for Database – 1 + Database – 3

Figure 10. Parity Plot for Database – 2 + Database – 3

Figure 9. Parity Plot for 'Database - 1 + Database - 2'

The performance of the mix of two databases looks better compared to the performance of the total database. The combination 'Database - 2 + Database - 3' has a good regression coefficient and average absolute deviation but fewer compounds in the region from 0.4 – 0.6. Most of the compounds in this combination of datasets are either highly biodegradable, wherein the ratio is greater than 0.8 or highly non-biodegradable, wherein the ratio is less than 0.2.

However, the combination 'Database - 1 + Database - 2', also has a good performance and the compounds in this dataset can be represented with the same number of groups as 'Database - 2 + Database - 3'.

Overall, the best performance is obtained with 'Database - 1 + Database - 2' after the removal of 12 outliers. The mix 'Database - 1 + Database - 3' is less appealing. This could be due to a

quality issue in the Database 3. Some of the compounds from Database 3 have high ratios for BOD₅ to COD. For instance, monochlorobenzene, tetrahydrofuran, diuron, metolachlor and propachlor have a ratio of greater than 0.9 suggesting that they are readily biodegradable. This is not expected given that these are aromatic and halogenated compounds. Therefore, the parameter regression performed using ‘Database – 1 + Database – 2’ is considered as the final recommended one and it is used further to predict the biodegradability of other organic compounds.

2.3 Model Validation

The developed model is then further validated. For this, a database containing 45 compounds provided by Syngenta AG was used. However, only 10 out of these 45 compounds could be represented by the groups used to develop the model. For 7 out of these 10 compounds the predictions were quite good (AAE $\leq$ 0.1). While, for the remaining three compounds the AAE was 0.38, 0.46 and 1.38.

Furthermore, the model is redeveloped by including the compounds from the Syngenta AG database in the parameter estimation step. The parameters for this model are available in Table 4.

Table 4. Group Contribution Model Parameters

Groups	Parameters
Universal Constant	
ϵ_0	-0,0023
1st Order Groups	
CH ₃	0,4687
CH ₂	-0,0261
CH	-0,4215
C	-1,4598
CH ₂ =CH	-0,2305
CH=CH	-1,9589
CH ₂ =C	-0,7647
CH=C	-1,4002
CH#C	-0,1847

aCH	0,1025
aC	0,0893
aC	0,0219
aC	0,2528
aN	-0,1759
aC-CH3	0,3047
aC-CH2	-0,1398
aC-C	-0,7655
aC-CH=CH2	0,1482
OH	0,3620
aC-OH	0,2026
COOH	0,4134
aC-COOH	0,3057
CH3CO	0,5140
CH2CO	0,0696
CCO	0,8638
aC-CO	-0,7032
CHO	0,1174
aC-CHO	0,3934
CH3COO	0,2852
CH2COO	0,0126
aC-COO	-0,0178
COO	-0,1714
CH3O	0,2643
CH2O	-0,1078
CH-O	-1,0281
aC-O	-0,4774
CH2NH2	-0,0017
CHNH2	0,0916
CH2NH	-1,0746
CHNH	0,1231
CH2N	-1,3474
aC-NH2	0,1886
aC-NH	0,1770
NH2	-0,3386
CH=N	-0,4524
C=N	0,4633
CN	0,4759
aC-NO2	-0,1081
CONH2	0,4301
aC-CONH2	-0,2204
aC-NH(CO)H	0,5000
aC-CONH	0,0066
aC-NHCONH	-0,2238
NHCO	-0,5998
CH2Cl	-0,1549

CHF2	-0,0989
CF3	1,4849
aC-Cl	-0,0082
aC-F	-0,1338
aC-Br	-0,2183
Br	0,3023
F	0,1669
Cl	-0,3243
OCH2CH2OH	0,0497
OCH2CHOH	-0,7216
aC-SO2	-0,1467
C2H3O	-0,3403
CH2(cyc)	-0,0039
CH(cyc)	0,0198
C(cyc)	-0,2749
CH=CH(cyc)	0,0703
CH=C(cyc)	-0,4184
C=C(cyc)	-0,0560
NH(cyc)	0,1912
N(cyc)	-0,1442
CH=N(cyc)	0,2453
C=N(cyc)	0,5724
O(cyc)	0,1994
CO(cyc)	-0,0062
>NH	0,0417
-O-	0,0126
>CO	0,2190
CH-N	-0,0989
aC=O	0,5174
aN-	0,5174
Na	-0,0095
2nd Order Groups	
(CH3)2CH	-0,0462
(CH3)3C	-0,5713
CH3-CHm=CHn (m.n in 0..2)	0,4210
CH2-CHm=CHn (m.n in 0..2)	0,8618
CH3COCH2	-0,0025
CHCOOH	0,5439
CHOH	0,0492
COH	-0,0379
CHm(OH)CHn(OH) (m.n in 0..2)	0,0277
CHm(OH)CHn(NHp) (m.n.p in 0..2)	0,1976
CHm(NHn)-COOH (m.n in 0..2)	0,0916
HOOC-CHn-CHm-COOH (n. m in 1..2)	0,0830
HO-CHn-COOH (n in 1..2)	0,0007
CHm-O-CHn=CHp (m.n.p in 0..3)	0,8638

CHm=CHn-CN (m,n in 0..2)	0,4759
CHm=CHn-COOH (m,n in 0..2)	1,0097
aC-CHn-O- (n in 1..2)	0,0155
aC-CHn-OH (n in 1..2)	0,0687
aC-SO2-OH	-0,3174
aC-C(CH3)3	-0,7954
aC-CF3	-0,0644
(CHn=C)cyc-CO- (n in 0..2)	0,2086
(CHn=C)cyc-CH3 (n in 0..2)	0,0305
CHcyc-CH3	0,1851
CHcyc-CH2	0,0532
CHcyc-OH	-0,2409
CHcyc-NH2	0,3030
CHcyc-NH-CHn (n in 0..3)	0,2077
CHcyc-CO	-0,0383
CHcyc-COO	-0,0095
>Ncyc-CH3	-0,4166
>Ncyc-CH2	-0,3503
AROMRINGs1s2	-0,1700
AROMRINGs1s3	-0,0823
AROMRINGs1s4	-0,2978
AROMRINGs1s2s3	-0,7233
AROMRINGs1s2s4	-0,6885
AROMRINGs1s3s5	-0,9197
AROMRINGs1s2s3s5	-0,6973
PYRIDINEs2s3	-0,0057
PYRIDINEs2s4	0,1333
PYRIDINEs3s5	-0,1831
PYRIDINEs2s3s6	0,0987
aC-SO2NHn (n>=0;n<3)	-0,0191
aC-SO2NHn (n>=0;n<3)	0,2333
3rd Order Groups	
aC-(CHn=CHm)cyc (fused rings) (n,m in 0..1)	0,0305
aC-CHncyc (different rings) (n in 0..1)	0,3030
CH multiring	0,0698
C multiring	-0,2749
aC-CHm-aC (different rings) (m in 0..2)	0,5350
aC-CO-(C=CHn)cyc (different rings) (n in 0..1)	0,3430
aC-CO-CHncyc (different rings) (n in 0..1)	-0,0689
aC-NHncyc (fused rings) (n in 0..1)	0,0213
aC-NH-aC (different rings)	0,1770
aC-(N=CHn)cyc (fused rings) (n in 0..1)	0,1110
AROM.FUSED[2]	-0,0001
AROM.FUSED[2]s1	-0,2849
PYRIDINE.FUSED[2]	0,3116

A comparison of the group representation in the model based on ‘Database – 1 + Database – 2’ and the new model based on ‘Database – 1 + Database – 2 + Syngenta AG compounds’ is shown in Table 5.

Table 5. Comparison of Group Representation and Performance Statistics between Model based on ‘Database – 1 + Database – 2’ and Model based on ‘Database – 1 + Database – 2 + Syngenta AG compounds’

Parameters	Representation	Model based on Database - 1 + Database – 2 + Syngenta AG Database	Model based on Database – 1 + Database – 2
		Number	Number
1 st Order Group Contribution	C_i	86	58
2 nd Order Group Contribution	D_j	45	28
3 rd Order Group Contribution	E_k	13	5
Universal Constant	ε_0	1	1
Model Performance Statistics	AAE	0.04	0.05
	R^2	0.90	0.92

3. Application Examples

In this section, two examples are provided where the developed GC model based on parameters regressed using the combined dataset comprising of ‘Database – 1 + Database – 2 + Syngenta AG compounds’ is employed. The performance of the model for determining the ratio of BOD₅ to COD for these two compounds and the pertinent calculations are presented in Table 6.

Table 6. Biodegradability Calculation for Two Compounds Using the Developed Model

Compound – 1			
(Acyclic Compound)	First Order Groups	Occurrences (N_i)	Contribution (C_i)
isopropanol (CAS Number: 67-63-0) $(BOD_5 / COD)_{\text{expt}} = 0.7059$	CH ₃	2	0.4687
	CH	1	-0.4215
	OH	1	0.362
	Second Order Groups	Occurrences (O_k)	Contribution (E_k)
	(CH ₃) ₂ CH	1	-0.0462
	CHOH	1	0.0492
	BOD ₅ / COD =		0.7737
	Absolute Error =		0.07

Compound – 2			
(Aromatic Compound)	First Order Groups	Occurrences (N_i)	Contribution (C_i)
aCH	3	0.1025
	aC-OH	1	0.2026
	aC-Cl	2	-0.0082

2,4-dichlorophenol (CAS Number: 120–83–2) (BOD ₅ / COD) _{expt} = 0.04	Second Order Groups Occurences (O_k) Contribution (E_k)		
	AROMRINGs1s2s4	1	-0.6885
		BOD ₅ / COD =	0.0357
		Absolute Error =	4.31E-03

Besides predicting the value of this ratio, the model is also able to correctly classify 96% of the compounds in the database ‘Database – 1 + Database – 2 + Syngenta AG compounds’ as 'readily biodegradable' or 'not readily biodegradable'. A ratio of greater than 0.5 suggests biodegradability of the compound and can serve as a first screening test for the biodegradability of a specific chemical.

4. Conclusion

A group-contribution (GC) based property model has been developed with the objective of providing reliable and quick estimation of the biodegradability for a wide range of organic compounds. The model can be used for applications in sustainable chemical product design and for substituting hazardous chemicals with their greener counterparts. The prediction of the ratio of BOD₅ to COD using the developed GC model is satisfactory ($R^2 = 0.92$, AAE = 0.05). However, the quality of the data compiled in the form of the four databases from various sources in the literature, which has been used to develop the model, is unknown. This poses a limitation on developing such models that are highly dependent on the data.

Motivated by the results obtained in this work, future work could focus on the property modelling combining machine learning algorithms with GC towards further improving the

prediction accuracy of the model. Machine learning algorithms could help in handling the exceptional cases wherein the modelled value of the property deviates significantly from the observed (or experimental) values. Herein, the structural group information could be provided for training the machine learning-based model, thus giving rise to hybrid models that could have an enhanced performance.

SUPPORTING INFORMATION

I. Databases to Develop the Model for Biodegradation

- Table S1. Database – 1: Simple Acyclic Compounds and Aromatics
- Table S2. Database – 2: Benzene Ring - Containing Aromatics
- Table S3. Database – 3: Acyclic Compounds, Aromatics and Complex Pesticides
- Table S4. Database – 4: Deicers

II. List of Abbreviations

AUTHOR INFORMATION

Corresponding Author

Georgios M. Kontogeorgis. Phone: +45 45 25 28 59, e-mail: gk@kt.dtu.dk.

Notes

The authors declare no competing financial interest.

ABBREVIATIONS

The list of abbreviations are available in Section B of the Supporting Information.

ACKNOWLEDGEMENT

The authors are grateful to Syngenta AG for providing useful data which were considered in the models' development and validation.

REFERENCES

- (1) Speight, J. G. *Environmental Organic Chemistry for Engineers*; 2017.
- (2) Organization for Economic Cooperation and Development (OECD). Test No. 301: Ready Biodegradability. In *OECD Guidelines for Testing of Chemicals, Section 3*; 2018, DOI 10.1787/9789264070349-en.
- (3) Wik, A.; Montero, R. M.; Dilhac, B. How to bring your registration dossier in compliance with REACH - ECHA (accessed Sep 8, 2019).
- (4) Jördening, H. J.; Winter, J. *Environmental Biotechnology: Concepts and Applications*; 2005, DOI 10.1002/3527604286.
- (5) Quayle, W. C.; Fattore, A.; Zandona, R.; Christen, E. W.; Arienzo, M. Evaluation of Organic Matter Concentration in Winery Wastewater: A Case Study from Australia. *Water Sci. Technol.* **2009**, 60 (10), 2521–2528, DOI 10.2166/wst.2009.688.
- (6) General Electric Power and Water. *Correlating Total Organic Carbon (TOC) to Biochemical (BOD5) and Chemical Oxygen Demand (COD)*; 2015.
- (7) Grammatica P, Pavan M, Consolaro F, Connsonni V, T. R. QSAR Modeling of the Biodegradation by Holistic Molecular Descriptors. In *9th International Workshop on QSARs, Bourgas, Bulgaria*; 2000; p 36.
- (8) Pavan, M.; Worth, A.; Netzeva, T. I. *Review of QSAR Models for Bioconcentration*; 2006.
- (9) Boethling, R. S.; Howard, P. H.; Meylan, W.; Stiteler, W.; Beauman, J.; Tirado, N. Group Contribution Method for Predicting Probability and Rate of Aerobic Biodegradation. *Environ. Sci. Technol.* **1994**, 28 (3), 459–465, DOI 10.1021/es00052a018.

- (10) U.S. Environmental Protection Agency. EPI Suite™-Estimation Program Interface (accessed Sep 3, 2019).
- (11) C.J. van Leeuwen, T. G. V. *Risk Assessment of Chemicals: An Introduction*, Second Edi.; Springer, 2009; Vol. 15, DOI 10.1080/10807030802616012.
- (12) Gani, R.; Hytoft, G.; Jaksland, C.; Jensen, A. K. An Integrated Computer Aided System for Integrated Design of Chemical Processes. *Comput. Chem. Eng.* **1997**, *21* (10), 1135–1146, DOI 10.1016/S0098-1354(96)00324-9.
- (13) Gamberger, D.; Sekuak, S.; Sabljic, A. Modelling Biodegradation by an Example-Based Learning System. *Informatica* **1987**, *17*, 157–166.
- (14) Gamberger, D.; Horvatic, D.; Sekusak, S.; Sabljic, A. Applications of Experts' Judgement to Derive Structure-Biodegradation Relationships. *Environmental Science and Pollution Research*. 1996, pp 224–228, DOI 10.1007/BF02986965.
- (15) Hukkerikar, A. S.; Sarup, B.; Ten Kate, A.; Abildskov, J.; Sin, G.; Gani, R. Group-Contribution+(GC+) Based Estimation of Properties of Pure Components: Improved Property Estimation and Uncertainty Analysis. *Fluid Phase Equilib.* **2012**, *321*, 25–43, DOI 10.1016/j.fluid.2012.02.010.
- (16) Marrero, J.; Gani, R. Group-Contribution Based Estimation of Pure Component Properties. *Fluid Phase Equilib.* **2001**, *183–184*, 183–208, DOI 10.1016/S0378-3812(01)00431-9.
- (17) Bridié, A. L.; Wolff, C. J. M.; Winter, M. BOD and COD of Some Petrochemicals. *Water Res.* **1979**, *13* (7), 627–630, DOI 10.1016/0043-1354(79)90011-3.
- (18) Haandel, A; Lubbe, J. *Handbook of Biological Wastewater Treatment*; IWA Publishing, 2012.

- (19) Pitter, P., Chudoba, J. *Biodegradability of Organic Substances in the Aquatic Environment*; CRC Press: Boca Raton, Florida, 1990.
- (20) Verschueren, K. Handbook of Environmental Data on Organic Chemicals. *Soil Sci.* **1985**, *139* (4), 376, DOI 10.1097/00010694-198504000-00014.
- (21) Cvetnic, M.; Juretic Perisic, D.; Kovacic, M.; Kusic, H.; Dermadi, J.; Horvat, S.; Bolanca, T.; Marin, V.; Karamanis, P.; Loncaric Bozic, A. Prediction of Biodegradability of Aromatics in Water Using QSAR Modeling. *Ecotoxicol. Environ. Saf.* **2017**, *139*, 139–149, DOI 10.1016/j.ecoenv.2017.01.031.
- (22) Wernet, G.; Hellweg, S.; Fischer, U.; Papadokonstantakis, S.; Hungerbühler, K. Molecular-Structure-Based Models of Chemical Inventories Using Neural Networks. *Environ. Sci. Technol.* **2008**, *42* (17), 6717–6722, DOI 10.1021/es7022362.
- (23) Lin, Z.; Zhen, Z.; Liang, Y.; Li, J.; Yang, J.; Zhong, L.; Zhao, L.; Li, Y.; Luo, C.; Ren, L.; et al. Changes in Atrazine Speciation and the Degradation Pathway in Red Soil during the Vermiremediation Process. *J. Hazard. Mater.* **2019**, *364*, 710–719, DOI 10.1016/j.jhazmat.2018.04.037.
- (24) Corsi, S. R.; Mericas, D.; Bowman, G. T. Oxygen Demand of Aircraft and Airfield Pavement Deicers and Alternative Freezing Point Depressants. *Water. Air. Soil Pollut.* **2012**, *223* (5), 2447–2461, DOI 10.1007/s11270-011-1036-x.
- (25) Nzila, A. Current Status of the Degradation of Aliphatic and Aromatic Petroleum Hydrocarbons by Thermophilic Microbes and Future Perspectives. *International Journal of Environmental Research and Public Health.* 2018, DOI 10.3390/ijerph15122782.
- (26) Cook, R. D. Cook's Distance. In *International Encyclopedia of Statistical Science*;

FOR TABLE OF CONTENTS ONLY