Molecules **2004**, *9*, 989-1004

molecules
ISSN 1420-3049
http://www.mdpi.org

Review

Evaluation of Artificial Intelligence Based Models for Chemical Biodegradability Prediction

James R. Baker [1,*], Dragan Gamberger [2], James R. Mihelcic [1] and Aleksandar Sabljić [2]

[1] Department of Civil and Environmental Engineering, Michigan Technological University, 1400 Townsend Drive, Houghton, Michigan, USA
[2] Rudjer Boskovic Institute, POB 180, HR-10002 Zagreb, Croatia

* Author to whom correspondence should be addressed; email: jrbaker@mtu.edu

Received: 26 October 2004 / Accepted: 12 December 2004 / Published: 31 December 2004

Abstract: This study presents a review of biodegradability modeling efforts including a detailed assessment of two models developed using an artificial intelligence based methodology. Validation results for these models using an independent, quality reviewed database, demonstrate that the models perform well when compared to another commonly used biodegradability model, against the same data. The ability of models induced by an artificial intelligence methodology to accommodate complex interactions in detailed systems, and the demonstrated reliability of the approach evaluated by this study, indicate that the methodology may have application in broadening the scope of biodegradability models. Given adequate data for biodegradability of chemicals under environmental conditions, this may allow for the development of future models that include such things as surface interface impacts on biodegradability for example.

Keywords: Biodegradation, QSBR, environmental fate, QSAR, interfaces

Introduction

Biodegradation is an interfacial phenomenon influenced by a chemical's tendency to partition to various phases in the environment. Equilibrium partitioning between solid and liquid interfaces [1] strongly influences the biodegradability of chemicals in the presence of surfaces (e.g., soils and sediments). The resulting inaccessibility of solutes to microorganisms that are responsible for

degradation can limit biodegradation [2, 3]. Due to the need to predict the ultimate fate of chemicals in the environment, many methods have been developed for estimating or predicting a chemical's biodegradation potential. These methods have each been constructed and are utilized in different ways in an effort to manage the tradeoffs between model complexity, availability of input data, and model reliability. Model inputs include expert opinion assessment, physical property correlations, group contribution, and other qualitative and quantitative indicators of biodegradability.

Modeling techniques used include linear and nonlinear regression, chemometric analysis, neural networks and artificial intelligence. Each of these techniques has individual advantages and disadvantages and tradeoffs are managed such that all models have various limitations in their utility and predictive ability [4-6]. For example, individual models tend to have some level of chemical class specificity, either by design, or as an artifact of the breadth of the model training data set. Basic attributes such as model complexity, range of chemical structures and size of data set can be used to subjectively assess the general utility of specific models [5].

This paper presents a discussion of the various methods to estimate biodegradability and, more importantly an evaluation of an artificial intelligence technique based on inductive machine learning that allows consideration for physical properties and group contribution effects [7, 8]. The evaluation has been conducted using an independent, critically reviewed database of biochemical oxygen demand (BOD) values that has seen limited use in model development. The inductive machine learning approach allows for the development of models with simple logical rules that indicate important structural features for biodegradability and may provide for the elucidation of relevant factors in determining a chemical's availability in the environment in the presence of solid surfaces, and therefore its propensity to biodegrade. Factors such as acclimation and chemical concentration may also be incorporated in future inductive machine learning models to account for environmental variability and more reliably predict biodegradation.

In this study, the inductive machine learning approach is demonstrated as sound when evaluated against an independent, highly reviewed data set that is not related to its training set. While the development of reliable and realistic biodegradability QSARs will require data from different types of tests to better simulate actual environmental conditions [9], the inductive machine learning approach shows promise for incorporating important surface interface and other environmental impacts into future modeling efforts.

Data availability for environmental fate assessment of chemicals

There are literally hundreds of thousands of anthropogenic chemicals manufactured and ultimately released to the environment, either through their intended use or through accidental discharge. The ultimate disposition of these chemicals on the environment is important in assessing their short and long term impact on living systems, and ultimately, on human health. While new standards and requirements for testing and providing data for High Production Volume (HPV) chemicals have promise for improving data availability for new chemicals, the sheer number of chemicals currently in use makes individual testing and assessment impractical. For example, it has been reported that there are more than 100,000 compounds existing in the European Union as indicated by the contents of the ENECS database [10]. Furthermore, a recent study reviewed more than 10,000 pre-manufacture

notices submitted to the United States Environmental Protection Agency between 1995 and 2001 and was able to find only 305 chemicals with biodegradability data [11].

Relatively new requirements for screening tests in the European Union, Canada and Japan will undoubtedly improve the availability of data for biodegradability and other environmental fate parameters. Even with these requirements, however, information provided from these tests may not be sufficient to conduct risk assessments [4]. In addition, consistently measuring whether or not a chemical is likely to biodegrade and at what rate can be difficult. For example, analytically determined biodegradation half-lives have covered a wide range even when tested under similar conditions [12]. Even if the consistency of the results can be resolved, test conditions such as acclimation and test chemical concentration can produce results that are of potentially questionable relevance to a chemical's actual fate in the environment [9].

The development of models for predicting biodegradability has provided a number of useful tools for generally assessing the fate of various chemicals in the environment and even in helping to understand the mechanisms of degradation; however, work remains to be done for these tools to reach a level of general utility. While years of research in physical property modeling and structure activity relationships has resulted in the ability to predict many chemical properties with acceptable reliability from knowledge only of chemical structure, prediction of biodegradability among other properties still needs improvement [13]. Russom *et al.* [14] reported, for example, that for the BIOWIN package [15], the EU recommends only using a slow biodegradation output as confirmation that a substance is not readily biodegradable and recommends against relying on fast biodegradation outputs.

There are two frequently referenced broadly available data sources for biodegradation data, commonly referred to as the BIODEG and the MITI-I databases. BIODEG is a file of biodegradation data within the Environmental Fate Database [16] which is available commercially from Syracuse Research Corporation (Syracuse, N.Y., U.S.A., http://www.syrres.com/esc/). The MITI-I database is available directly from the Chemicals Evaluation Research Institute (Tokyo, Japan) and can be downloaded from http://www.cerij.or.jp/ceri_en/otoiawase/otoiawase_menu.html. These databases, in addition to the expert opinion survey conducted by Boethling and Sabljić [17], have been used extensively for model development and validation. These data sets are generally available and are regarded as of a high quality. It is notable that the two datasets do include some data that are contradictory for a small subset of overlapping chemicals in the BIODEG and MITI-I datasets [8]. Chemicals within these databases are generally classified as biodegradable or non-biodegradable or as fast or slowly biodegradable.

The BIODEG and MITI-I datasets are sufficiently unique that it is common for independent models to be generated based on each. Gamberger *et al*. [8], for example, created two different rules, each designed to best predict data from one or the other dataset. The commonly used BIOWIN model package recommended by the EU Risk Ranking Method [14] includes separate linear and non-linear models built from the MITI-I and the BIODEG data [11, 18]. It has been reported that due to cross correlations, it is possible to develop a model that fits the training set data well but is not reliable as a predictor for chemicals outside the training set [19]. Based on this fact and the extensive use of the BIOWIN and MITI-I data in model development, it would be useful to evaluate models on an independent data-set to see how they perform.

Another set of critically reviewed data for BOD that exists has been prepared by the American Institute of Chemical Engineers Design Institute for Physical Properties (DIPPR)® and is available commercially from EPCON International (http://www.epcon.com/Product22.htm). The DIPPR database includes 56 chemical properties for approximately 600 chemicals selected from U.S. Environmental Protection Agency regulatory lists [20]. Each BOD data point in the DIPPR database has been critically evaluated using a 10-point criteria system which utilizes five rating parameters as shown in Table 1. Data sources received a score between 0 and 2 for each parameter which were then totaled for all of the parameters. For chemicals that had multiple data points from multiple sources, only the highest rated data point was chosen for this study. For a complete discussion of the criteria and a summary of the BOD/ThOD data see [21]. As a critically evaluated data-set that has seen limited use for biodegradation model development, this data-set is ideal for evaluation of models and modeling approaches developed to-date.

Table 1. Evaluation Criteria used for BOD Data in the DIPPR database

Rating Parameter	Required for Highest Rating
Experimental Technique	Follow Standard Methods
Temperature	Maintained at 20 ºC
Seed Acclimation	Used acclimated seed
Concentration of Chemical Dilution	2-6 mg/L O_2 depletion
Internal Consistency	ThOD≥ BOD

Review of modeling efforts

There are a large number of correlations and models for biodegradability currently in the literature. For example, Raymond *et al*. [5] presented 41 correlations for various individual homologous series of chemicals and Loonen *et al*. [22] referred to an EU study that evaluated 84 individual models. Most models generate results that generally indicate propensity for biodegradability such as readily biodegradable, slowly biodegradable, or not readily biodegradable and typically do not produce quantitative results such as half lives or degradation rates. These semi-qualitative model outputs have been noted as useful for screening tools but lacking in utility for full scale fate modeling as environmental compartment models, or "box models", typically require at least compartmental half lives [4]. The fact that even consistent analytical results are difficult to obtain additionally suggests however, that screening level tools likely represent the finest level of detail that can be reasonably obtained given the complexity of the systems involved and the current level of understanding of biodegradation mechanisms. While the models constructed to-date certainly have utility, the continued development of models with predictable accuracy and that can reasonably account for multiple factors and provide insight into fundamental modes of action related to biodegradability, including interface phenomena, will require continued research.

A number of detailed reviews of modeling efforts are available [4, 5, 6, 23]. This work does not intend to repeat that work, but rather present a brief discussion of general modeling efforts to-date with a more detailed discussion and evaluation of an inductive machine learning method utilized by

Gamberger *et al*. [7]. The evaluation was conducted using a critically reviewed database that has seen limited use in model development and therefore should provide for reasonable independent assessment of the models' ability to predict the biodegradability of chemicals not included in the model training sets. This discussion also includes considerations of potential future directions related to interface considerations.

The types of approaches to modeling are generally categorized for the purposes of this study as; regression models, human expert system models, and machine learning models. Rorije [10] noted that the rule based artificial intelligence approach used by Gamberger *et al*. [7] cannot be compared in a straightforward fashion to other types of modeling approaches and as such, this method has seen limited review in the literature.

Regression models

Regression models consist of linear, multiple linear, and non-linear correlations of biodegradation rates with parameters including physical or chemical properties and/or molecular connectivity indices. Commonly used properties include molecular weight, solubility, and structural fragment or group contributions. Molecular connectivity indices have also been used that relate to branching, volume, and molecular weight as well as other factors. A number of previously published regression models are presented in Table 2.

Table 2. Examples of Published Biodegradation Models Representative of Common Modeling Approaches

Model Reference	**Training Data Set**	**Descriptors used**	**Modeling Technique Used**
Boethling and Sabljić [17]	Results of expert opinion survey	Molecular connectivity indices $^{2}X^{v}$ and $^{4}X_{pc}$, molecular weight, and number of chlorine atoms	Linear and multiple linear regression
Boethling *et al*. [29]	BIODEG and results of expert opinion survey	Molecular weight and calculated structural fragment/group contributions	Multiple linear and nonlinear regression
Howard *et al*. [15]	BIODEG	Structural fragment/group contributions	Linear and nonlinear regression
Huuskonen [19]	Results of expert opinion survey	Various atom-type electrotopological state indices	Multiple linear regression and artificial neural network
Loonen *et al*. [30]	Data measured using MITI-I protocol	Structural fragment/group contributions	Partial least squares discriminant analysis
Loonen *et al.* [22]	Data measured using MITI-I protocol	Structural fragment/group contributions	Partial least squares discriminant analysis
Cambon and Devilers [26]	Results of expert opinion survey	Structural features and molecular weight	Neural network
Gamberger *et al*. [7, 8]	BIODEG, expert opinion survey, and MITI-I	Structural features and molecular weight	Inductive machine learning
Klopman [31, 32]	BIODEG	Method uses machine learning techniques to determine relevant descriptors mathematically from data on activity and basic chemical structure.	Knowledge-based learning system
Rorije *et al*. [33] (model specific to anaerobic degradation)	Anaerobic degradation data from Environmental Fate Database EFDB [34]	Used Klopman method [32] to generate fragments important for anaerobic biodegradation.	Used Klopman [32] method

These models are attractive in their relative ease of development given reasonable availability of data and model inputs, but are generally limited to specific chemical classes. Additionally, while statistical measures can be undertaken to reduce the risk of chance correlations, their possibility remains. It has been reported, for example, that the significance of some variables may be difficult to rationalize given known factors that influence biodegradation [15]. The inability to rationalize the significance of some variables may suggest that they are the result of chance correlations.

Expert system/survey models

Human expert systems or survey models are designed to capture the collective wisdom of experts in the field of biodegradation in a process that results in identification of important structural features that stimulate or inhibit biodegradation. The models are constructed by conducting surveys of biodegradation experts regarding biodegradation potential of various chemicals. The survey results are correlated against structural fragments and other chemical properties to identify fragments and properties important for biodegradation. These correlations may be done using regression or other mathematical tools and so expert system models commonly also fit under other classifications, however the exclusive use of expert opinion information is a feature of the models unique enough to justify individual classification. The collection of expert opinions may lead to the potential identification of structural elements or other factors influencing biodegradability that may not be obvious to any individual expert. On the other hand, it has also been reported that the divergence of opinions of surveyed experts may indicate that biodegradation rates and pathways are not always obvious [17] and therefore warrant careful analysis and consideration prior to application.

Machine learning based models

Machine learning techniques include neural networks and inductive learning and utilize computers to process available data against chemical structural features and properties to elucidate important features and properties relevant for biodegradation. Neural network techniques were noted some time ago as a promising tool for summarizing biodegradability data [24] and have been described as attractive for developing robust models due to their ability to account for a variety of interacting factors that influence a chemical's biodegradability [25]. These models follow a similar logic to the expert system/survey models in that they seek to identify subtleties that are not initially obvious, but utilize computer and mathematical analysis to more rigorously identify the important structural features and properties. These techniques are attractive for modeling complex processes like biodegradation due to their dynamic nature and ability to modify their behavior in response to their environment, store experimental knowledge, and make that knowledge available for modeling [26]. Another advantage of using machine learning techniques is their ability to point out the importance of specific descriptors and relations among descriptors that are likely to stimulate further investigations into the specific mechanisms of biodegradation [8]. Similarly, understanding structural features discovered through machine learning analysis may be additionally helpful in designing chemicals with a higher propensity for degradability by including substituents that promote degradability and

removing substituents that inhibit degradability [27]. Examples of published machine learning modeling efforts are presented in Table 2.

Application of model batteries

In addition to the discrete use of individual models, it has been suggested in the past that a number of models can be used successively to evaluate confidence in the results. It is logical that if multiple models are run for the same chemical and produce conflicting results, then those results are potentially questionable. At a minimum, the user is faced with a decision about which one, among the conflicting models, is more accurate given comparably appropriate models (e.g. no class specificity or other issues with either model training set).

The general concept of utilizing multiple models and concluding that reliable results cannot be obtained given conflicting results has been suggested in previous studies [6, 15]. However this concept has more recently been rigorously evaluated. A recent study presented the use of model batteries selected through Bayesian analysis to improve the reliability of predictions or better qualify questionable predictions [28]. The model battery approach consists of selecting a series of models and qualifying confidence in the model results based on whether each of the models agrees or not. While not a fundamentally new approach to modeling, the battery test approach is a new method of formally assessing the reliability of the results obtained from various models or sequential combinations of models.

Gamberger et al. inductive machine learning artificial intelligence model

As described above, Gamberger *et al*. have developed inductive machine learning models for predicting biodegradation potential of organic chemicals. Two of these models have been selected for further analysis and are termed for this study "Rule A" [7] and "Rule B" [8]. Rule A was developed from the expert opinion data-set reported by Boethling and Sabljić [17] and Rule B was developed from MITI-I test data. These Rules use the structural descriptors noted in Table 3, but have different outcomes regarding the significance of those descriptors in biodegradability based on the nature of the data on which they were built. The MITI-I data has been reported to have a tendency to under-predict biodegradability and therefore classifies some compounds as non-degradable that are classified as degradable under other test conditions, such as those conditions that the chemicals in the BIODEG database were tested under [6]. This under-prediction has been reported in part to be potentially caused by the relatively high chemical concentration used in the MITI-I test which is higher than what is likely to be experienced in the environment, and may produce toxic effects on the test inoculum [4]. Based on these data differences, it is reasonable that two distinct models be developed, one as a general utility biodegradation model based on the Boethling and Sabljić [17] survey data (i.e. Rule A) and one which was developed to more closely predict the results of the MITI-I test (i.e. Rule B).

The inductive machine learning method involves describing each chemical with a number of structural descriptors as input variables. The structural descriptors used by Gamberger *et al*. are presented in Table 3. Binary output variables are assigned to each chemical with a 1 for fast biodegradability and 0 for slow biodegradability based on the training set data. Each chemical

represents a learning example and analysis is conducted to find individual rules that satisfy all of the learning examples. The simplest rule is assumed to have the greatest chance of being most correct against test data. Once the simplest rules are identified, they are further analyzed to determine if the exclusion of any single chemical can reduces the number of basic logical elements. If this occurs, that chemical is removed as a potential outlier or incorrect data point. Chemicals are removed in this manner until a simple non-reducible solution is obtained which is the rule that models the data best. Rules A and B are presented in common language format in Table 4 and in mathematical format in Table 5.

Table 3. Structural descriptors used in construction of Artificial Intelligence biodegradation models (from [7, 8])

Descriptor Designation	Rule A Descriptors	Rule B Descriptors
a	Presence of heterocyclic or anhydride groups	Presence of heterocyclic nitrogen atom
b	Presence of ester, amide, or anhydride groups	Presence of ester, amide, or anhydride groups
c	Number of chlorine atoms	Number of chlorine atoms
d	Bicyclic alkanes	Bicyclic alkanes
e	Chemical composed only of carbon, hydrogen, nitrogen, and oxygen atoms	Chemical composed of only carbon, hydrogen, nitrogen, and oxygen atoms
f	Presence of nitro group	Presence of nitro group
g	Number of rings	Number of rings
h	Presence of epoxy group	Presence of epoxy group
i	Primary alcohols and phenols	Primary alcohols and phenols
j	Molecular weight	Molecular weight
k	Number of all C-O bonds	Number of all C-O bonds
l		Number of tertiary amino groups
m		Number of quaternary carbon atoms
n		Number of C=C bonds
o		Number of aromatic amino groups
p		Number of acid groups
r		Number of ester groups

Table 4. Rules developed for inductive machine learning model by Gamberger *et al*.

Rule A [7]	Rule B [8]
A chemical will biodegrade fast if any of the following conditions is met: (a) chemicals with one or more C-O bonds and molecular weight below 180 (b) chemicals built of C,H,N, and O atoms but without a nitro group and having a number of rings equal to or smaller than the number of C-O bonds	A chemical will biodegrade fast if any of the following conditions is satisfied: (a) acyclic chemicals with one C-O bond, but without quaternary carbons (b) esters, amides, or anhydrides built of C, H, N, and O atoms, but without or with 2 C=C bonds (c) acyclic esters, amides, or anhydrides without quaternary carbons

Table 4. Cont.

(c) chemicals built of C,H, N, and O atoms but without a nitro group and their molecular weight must be in the range from 95 to 135	(d) esters, amides, or anhydrides built of C, H, N, and O atoms, having one ring or less but without quaternary carbons (e) acyclic chemicals built of C, H, N, and O atoms, but without either quaternary carbons or tertiary amino groups and without or with 2 C=C bonds (f) chemicals built of C, H, N, and O atoms, acyclic or with 1 ring, with at least one C-O bond, but without either quaternary carbons or tertiary amino groups and without or with 2 C=C bonds.

Table 5. Mathematical representation of two Rules developed by Gamberger *et al*. (See Table 3 for structural descriptors with letter designations)

Rule 1 [7]	**Rule 2 [8]**
Chemical will biodegrade fast if any of the following terms is satisfied: $(k \neq 0)\ (j < 180)$ $(e = 1)\ (f = 0)\ (g \leq k)$ $(e = 1)\ (f = 0)\ (95 < j < 135)$	Chemical will biodegrade fast if any of the following terms is satisfied: $(m = 0)\ (k = 1)\ (g = 0)$ $(b = 1)\ (n \neq 1)\ (e = 1)$ $(b = 1)\ (m = 0)\ (g = 0)$ $(b = 1)\ (m = 0)\ (e = 1)\ (g \leq 1)$ $(m = 0)\ (e = 1)\ (l = 0)\ (n \neq 1)\ (g = 0)$ $(m = 0)\ (e = 1)\ (l = 0)\ (n \neq 1)\ (k \neq 0)\ (g \leq 1)$

Rules A and B have been subject to review against the expert survey results of Boethling and Sabljić [17] and the BIODEG and MITI-I Data. Summaries of these evaluations have been reported in the literature [8] and are presented below in Tables 6 and 7.

Table 6. Results of Rule A when applied to Boethling and Sabljić [17] expert survey data and data from the BIODEG database [35].

Test Set	**Biodegradability indication**	**Number of correct predictions**	**Percent of correct predictions**
23 Chemicals from Boethling and Sabljić [17] expert survey	Readily Biodegradable	8/8	100%
	Slowly Biodegradable	14/15	93%
17 Chemicals selected from BIODEG database	Readily Biodegradable	9/9	100%
	Slowly Biodegradable	8/8	100%

Table 7. Results of Rule B when applied to MITI-I data test set [8]

Test Set	Biodegradability indication	Number of correct predictions	Percent of correct predictions
762 MITI-I data points	Fast Biodegradation	279/364	77%
	Slow Biodegradation	355/398	89%

With these positive results as an indication of the power of the method, an additional analysis was conducted with the critically reviewed DIPPR data set as an additional external check of the soundness of the method for predicting biodegradation. The results of this check are presented in the following section.

Results and Discussion

Evaluation of inductive machine learning model using 5-day Biochemical Oxygen Demand

As Table 2 shows, among the machine learning modeling efforts, the inductive machine learning models developed by Gamberger *et al*. [7, 8] are unique in that they are presented in an if-then-else format that is relatively simple to apply given basic understanding of a chemical's structure. In a recent comprehensive review of biodegradability prediction, Jaworska *et al*. [4] described the inductive machine learning approach as notable in that it was a simple system that could achieve results comparable with more complex models. This method takes advantage of the attractive attributes of machine learning in utilizing the power of a computer to analyze the complex interactions of various structural features and physical/chemical properties that stimulate or inhibit biodegradability but provides results that do not require a computer to utilize. Based on this ease of use, it would be useful to evaluate this model using a high quality data set that is independent of the model training set. Given the utilization of the BIODEG and MITI-I data in either the development of or previous efforts for validation of the inductive machine learning Rules, this study compared the results of the application of these Rules to the chemicals in the DIPPR database.

The DIPPR database includes experimental BOD and calculated ThOD data. In order to assess completion of biodegradation during the BOD test, BOD values are converted to a percentage of Theoretical (stoichiometric) Oxygen Demand (ThOD) from which the level of biodegradability is estimated. The ThOD was determined as described elsewhere [36]. For the purposes of this study, a BOD/ThOD value of less than or equal to 0.10 was considered to indicate that a chemical is not readily biodegradable and a value greater than 0.10 was considered to indicate that a chemical is biodegradable [21]. The DIPPR database contained quality BOD data and calculated ThOD values for 133 chemicals. 90 chemicals were classified as biodegradable (BOD/ThOD > 0.10) and 43 chemicals were classified as non-biodegradable (BOD/ThOD) < 0.10). Predictions following the inductive machine learning method results were compared to BOD/ThOD values in the DIPPR database and the results of the comparison are illustrated in Table 8.

Table 8. Results of comparison of inductive machine learning Rules A and B against DIPPR BOD/ThOD data

Results	**Rule A**	**Rule B**
Number of biodegradable chemicals correctly predicted	79/90	67/90
Percent of biodegradable chemicals correctly predicted	88%	74%
Number of non-biodegradable chemicals correctly predicted	21/43	26/43
Percent of non-biodegradable chemicals correctly predicted	49%	61%
Overall number correct	100/133	91/133
Overall percent correct	75%	68%

Both Rules performed reasonably well for predicting biodegradable chemicals but less well for predicting non-biodegradable chemicals. An analysis of the chemicals that were incorrectly predicted suggests that there may be some groups that are not adequately addressed in either Rule, perhaps as a result of the chemicals in the training sets for each Rule. For example, 17 of the incorrectly predicted chemicals had aromatic rings, including biphenyl and α-naphthylamine which were classified as non-biodegradable under both Rule A and B but had BOD/ThOD ratios greater than 0.10.

The number of compounds with rings that were predicted incorrectly as indicated by BOD/ThOD ratio may indicate that the modeling sets for Rule A and B did not contain enough compounds with multiple rings and positive biodegradability data to account for the importance of adjacent groups. Correction factors are sometimes used to compensate for interaction between individual functional groups [19] and models that don't account for interactions among fragments in multifunctional molecules may be somewhat simplistic [29]. The lack of data for compounds with multiple rings may have resulted in an inadequate accounting for the importance of functional group interactions that may promote biodegradability in some of these compounds in the inductive machine learning model.

With positive results of independent validation of both Rules against the DIPPR data, it would be useful to evaluate another commonly used model against the DIPPR data to see how they compare. An analysis was conducted of the linear and non-linear models in the BIOWIN package against the 133 chemical validation set which produced the results illustrated in Table 9.

Table 9. Results of comparison of BIOWIN against DIPPR BOD/ThOD data

Results	**Linear**	**Non-linear**
Number of biodegradable chemicals correctly predicted	78/90	79/90
Percent of biodegradable chemicals correctly predicted	87%	88%
Number of non-biodegradable chemicals correctly predicted	21/43	19/43
Percent of non-biodegradable chemicals correctly predicted	49%	44%
Overall number correct	99/133	98/133
Overall percent correct	74%	74%

When compared against the DIPPR data, the inductive machine learning Rules developed by Gamberger et al. provide very similar reliability results to both the linear and non-linear BIOWIN models. These results suggest that both methods are comparably robust. BIOWIN is used prominently by the U.S. Environmental Protection Agency in conducting pre-manufacture notice reviews [11] and is therefore considered a reasonably reliable model for predicting biodegradability. The inductive machine learning Rules have been shown to have comparable performance but the machine learning technique used to generate the Rules has a unique advantage in its ability to account for a number of diverse and even competing factors. This suggests that the method is likely to have additional application in potentially modeling broader aspects of biodegradability such as interactions with surfaces including microbial cell walls and other interfacial phenomena.

Conclusions: Utilization of Artificial Intelligence to introduce interface considerations

Limitations in available data and in the current level of understanding of how to represent various environmental factors influencing biodegradation currently limit what can be realistically done to develop environmentally relevant QSARs for biodegradability [9]. As soil and sediments are the principal sink for many hydrophobic organic substances [4], understanding interactions between solid surface/solute/microbe interfaces will be important to advance the predictability of existing QSARS. Boethling and Sabljić [17] for example used only data from tests that incorporated natural water and detrital sediment for evaluating the results of their expert opinion survey model as these conditions were reported as essential to reflect environmentally relevant biodegradation rates. A number of models developed to-date include a molecular weight cut-off based on the presumption that exceptionally large molecules cannot be transported across the cellular wall [8, 37]. It has been reported that models and handbook data tend to under-predict fate of persistent organic pollutants (POPs) but are more reliable for less persistent substances [38]. While there may be a number of factors contributing to this under-prediction of persistence it is very likely that interface effects and transport of heavily sorbed pollutants may limit bioavailability and subsequently biodegradability. Additionally, another study reported and verified a positive correlation between solubility and biodegradability [39]. This correlation may further suggest that solute concentration and therefore transport of sorbed or organic phase pollutants into the solute impacts biodegradation.

Artificial intelligence techniques have been noted as promising for their ability to allow for efficient consideration of large numbers of descriptors and modeling parameters [26] and their ability to account for interacting factors [25]. The artificial intelligence inductive machine learning model developed by Gamberger *et al*. has been shown in this study as fundamentally sound when evaluated against a critically reviewed external data-set and another commonly used biodegradability model. However, efforts to-date with these and other modeling techniques have focused on chemical structures and properties modeled against available biodegradability data. There have not been considerable efforts to-date to investigate and model important interfacial and other environmental conditions.

Given adequate data including such interfacial and environmental conditions, these models may be extendable to include surface interactions and other factors that could be analyzed in future modeling efforts. For example, when evaluating the biodegradation potential of chemicals discharged to the environment, models may take the total organic carbon content of the receiving waters and of the

sediments of those waters as well as other potential surfaces issues such as the relative oxidation level of sediment material and suspended clay particle content as inputs. Some models have been developed for specific environmental conditions, such as anaerobic degradation [33], but the authors are unaware of any studies that consider environmental compartment parameters such as surface interactions rigorously. This is likely due to lack of data, but if such data were available, artificial intelligence modeling techniques have been proven to be able to address chemical property and structural group contributions to biodegradability and have considerable promise for including environmental and interfacial considerations in realistic settings.

Acknowledgements

The authors would like to acknowledge the American Institute for Chemical Engineers (AIChE) Design Institute for Physical Property Data (DIPPR^{TM}) for use of BOD/COD data and the contributions of Eric Rupprecht who was M. Sc. student in Environmental Engineering at Michigan Technological University at the time of this study. A portion of this research that was conducted by Eric Rupprecht was funded by the AIChE/DIPPR under Projects 911 and 912 Environmental Health and Safety Compilation and Data Estimation Manual.

References

1. Baker, J.R.; Mihelcic, J.R.; Luehrs, D.C.; Hickey, J.P. Evaluation of estimation methods for organic Carbon normalized sorption coefficients. *Water Environ. Res.* **1997,** *69*, 136-145.
2. Mihelcic, J.R.; Leuking, D.R.; Mitzell, R.J.; Stapleton, J.M. Bioavailability of sorbed- and separate phase chemicals. *Biodegradation* **1993**, *4*, 141-153.
3. Mihelcic, J.R.; Pritschow, A.; Lueking, D.R. Uptake of dissolved and oil phase organic chemicals by bacteria. *Ground Water Monit. R.* Summer **1995**, 100-106.
4. Jaworska, J.S.; Boethling R.S.; Howard, P.H. Recent Developments in broadly applicable structure-biodegradability relationships, *Environ. Toxicol. Chem.* **2003**, *23*, 1710-1723.
5. Raymond, J.W.; Rogers, T.N.; Shonnard, D.R.; Kline, A.A. A review of structure-based biodegradation estimation methods. *J. Hazard. Mater.* **2001,** *B84*, 189-215.
6 Sabljić, A.; Peijnenburg W. Modeling lifetime and degradability of organic compounds in air, soil, and water systems - (IUPAC Technical Report). *Pure Appl. Chem.* **2001**, *73*, 1331-1348.
7. Gamberger, D.; Sekusak, S.; Sabljić, A. Modeling biodegradation by an example-based learning system. *Informatica* **1993,** *17*, 157-166.
8. Gamberger, D.; Sekusak, S.; Medven, Z.; Sabljić, A. Application of artificial intelligence in biodegradation modeling. In *Biodegradability Prediction*; Peijnenburg, W.J.G.M.; Damborsky J., Eds.; Kluwer Academic Publishers: Dordrecht, **1996**; pp. 41-50.
9. Cowan, C.E.; Federle, T.W.; Larson, R.J.; Feijtel, T.C.J. Impact of biodegradation test methods on the development and applicability of biodegradation QSARS. *SAR QSAR Environ. Res.* **1996**, *5*, 37-49.
10. Rorije, E.; Loonen, H.; Muller, H.; Klopman, G.; and Peijnenburg, W.J.G.M. Evaluation and application of models for the prediction of ready biodegradability in the MITI-I test, *Chemosphere* **1999**, *38*, 1409-1417.

11. Boethling R.; Lynch D.G.; Thom G.C. Predicting ready biodegradability of premanufacture notice chemicals. *Environ. Toxicol. Chem.* **2003**, *22*, 837-844.
12. Boethling, R.S.; Howard, P.H.; Beauman, J.A.; Larosche, M.E. Factors for intermedia extrapolation in biodegradability assessment. *Chemosphere* **1995,** *30*, 741-752.
13. Howard P.H., Celebrating QSARs. *Environ. Toxicol. Chem.* **2000**, *19*, 527-527.
14. Russom, C.L.; Breton, R.L.; Walker, J.D.; Bradbury, S.P. An overview of the use of quantitative structure-activity relationships for ranking and prioritizing large chemical inventories for environmental risk assessments. *Environ. Toxicol. Chem.* **2003,** *22*, 1810-1821.
15. Howard, P.H.; Boethling, R.S.; Stiteler, W.M.; Meylan, W.M.; Hueber, A.E.; Beauman, J.A.; Larosche, M.E. Predictive model for aerobic biodegradability developed from a file of evaluated biodegradation data. *Environ. Toxicol. Chem.* **1992,** *11*, 593-603.
16. Howard, P.H.; Heuber, A.E.; Mulesky, B.C.; Crisman, J.S.; Meylan, W.; Crosbie, E.; Gray, D.A.; Sage, G.W.; Howard, K.P.; LaMacchia, A.; Boethling, R.S.; Troast, R. BIOLOG, BIODEG, and FATE/EXPOS: New files on microbial degradation and toxicity as well as environmental fate/exposure of chemicals. *Environ. Toxicol. Chem.* **1986,** *5*, 977-988.
17. Boethling, R.S.; Sabljić, A. Screening-level model for aerobic biodegradability based on a survey of expert knowledge. *Environ. Sci. Technol.* **1989,** *23*, 672-679.
18. Tunkel J.; Howard P.H.; Boethling R.S.; Stiteler, W.; Loonen, H. Predicting ready biodegradability in the Japanese Ministry of International Trade and Industry test. *Environ. Toxicol. Chem.* **2000**, *19*, 2478-2485.
19. Huuskonen, J. Prediction of biodegradation from the atom-type electrotopolgical state indices. *Environ. Toxicol. Chem.* **2001**, *20*, 2152-2157.
20. American Institute of Chemical Engineers. *Design Institute for Physical Properties, Environmental Health and Safety Data Compilation, Project 911* (http://www.aiche.org/dippr/).
21. Rupprecht, E.M. *Assessment of a model which estimates chemical biodegradability from knowledge of chemical structure*. M.Sc. Thesis, Michigan Technological University **1998**.
22. Loonen, H.; Lindgren, F.; Hansen, B.; Karcher, W. Prediction of biodegradability from chemical structure. In *Biodegradability Prediction*, Peijnenburg, W.J.G.M.; Damborsky J., Eds.; Kluwer Academic Publishers: Dordrecht, **1996**; pp. 105-113.
23. Howard, P.H. Biodegradation. In *Handbook of Property Estimation Methods for Chemicals* Boethling, R. and Mackay, D. Eds; Lewis Publishers: Boca Raton FL, USA, **2000**.
24. Zitko, V. Prediction of biodegradability of organic-chemicals by an artificial neural network. *Chemosphere* **1991**, 23, 305-312.
25. Langenberg, J.H.; Peinjnenburg, W.J.G.M.; Rorije, E. On the usefulness and reliability of existing QSBRs for risk assessment and priority setting. *SAR QSAR Environ. Res.* **1996**, *5*, 1-16.
26. Cambon, B.; Devillers, J. New trends in structure-biodegradability relationships. *Quant. Struct-Act. Rel.* **1993,** *12*, 49-58.
27. Klopman, G; Tu, M. Structure-biodegradability study and computer automated prediction of aerobic biodegradation of chemicals. *Environ. Toxicol. Chem.* **1997**, *16*, 1829-1835.
28. Boethling, R.S.; Lynch, D.G.; Jaworska, J.S.; Tunkel, J.L.; Thom, G.C.; Webb, S. Using BIOWIN, BAYES, and Batteries to predict ready biodegradability. *Environ. Toxicol. Chem.* **2004,** *23*, 911-920.

29. Boethling, R.S.; Howard, P.H.; Meylan, W.; Stiteler, W.; Beauman, J.; Tirado, N. Group contribution method for predicting mobility and rate of aerobic biodegradation. *Environ. Sci. Technol.* **1994**, *28*, 459-465.
30. Loonen, H.; Lindgren, F.; Hansen, B.; Karcher, W.; Niemela, J.; Hiromatsu, K.; Takatsuki, M.; Peijnenburg, W.; Rorije, E.; Struijs, J. Prediction of biodegradability from chemical structure: modeling of ready biodegradation test data. *Environ. Toxicol. Chem.* **1999**, *18*, 1763-1768.
31. Klopman, G.; Wang, S. A computer automated structure evaluation (CASE) approach to calculation of partition coefficient. *J. Comput. Chem.* **1991**, *12*, 1025-1032.
32. Klopman, G. MULTICASE 1. A hierarchical computer automated structure evaluation program. *Quant. Struct-Act. Rel.* **1992,** *11*, 176-184.
33. Rorije, E.; Peijnenburg, W.J.G.M.; Klopman, G.; Structural Requirements for anaerobic biodegradation of organic chemicals: a fragment model analysis. *Environ. Toxicol. Chem.* **1998**, *17*, 1943-1950.
34. Howard, P.H.; Heuber, A.E.; Boethling, R.S. Biodegradation data evaluation for structure biodegradation relations. *Environ. Toxicol. Chem.* **1987**, *6*, 1-10.
35. Gamberger, D.; Sekusak, S.; Medven, Z.; Sabljić, A. Application of experts' judgment to derive structure-biodegradation relationships. *Environ. Sci. Pollut. Res.* **1996**, *3*, 224-228.
36. Baker, J.R., Milke, M.W., Mihelcic, J.R. Relationship between chemical and theoretical oxygen demand for specific classes of organic chemicals. *Water Res.* **1999**, *33*, 327-334.
37. Scow, K.M. In *Handbook of Chemical Property Estimation Methods*. Lyman, W.J.; Reehl, W.F.; Rosenblatt, D.H. Eds.; McGraw Hill: New York, **1982**; pp. 9-1 - 9-85.
38. Gouian, T.; Cousins, I.; Mackay, D. Comparison of two methods for obtaining degradation half-lives. *Chemosphere* **2004**, *56*, 532-535.
39. Klopman, G.; Balthasar, D.M.; Rosenkranz, H.S. Application of computer automated structure evaluation (CASE) program to the study of structure-biodegradation relationships of miscellaneous chemicals. *Environ. Toxicol. Chem.* **1993,** *12*, 231-240.