Predicting Biodegradability using Thermodynamic Descriptors

Jim Yu
Department of Environmental and Civil Engineering
Southern Methodist University
Dallas, Texas 75275

Yu Su, Margaret H. Dunham
Department of Computer Science and Engineering
Southern Methodist University
Dallas, Texas 75275-0122

Abstract. The ability to predict the biodegradability of a given compound is crucial in understanding the fate of this compound in the environment, which can have tremendous public health implications. Traditionally, the biodegradability prediction relies on the regression analysis utilizing quantitative structural relationships. This paper presents a novel prediction method utilizing neural networks and other physical/chemical/thermodynamical data besides structures. This method has shown to be successful at predicting biodegradability with better accuracy than some current popular methods.

1. INTRODUCTION

With hundreds of thousands of anthropogenic chemicals used and manufactured in the world, their discharge into the environment is inevitable.[1] Thus, understanding their behavior in the environment is critical for public and ecological health. There are only a few removal mechanisms (such as biodegradation, volatilization, and chemical oxidation/reduction) available in a natural environment to alleviate these contaminations.[2] Compounds that are not removed by natural mechanisms will ultimately accumulate in nature and are likely to pose human and/or ecological deleterious effects. Biodegradation is often the key removal mechanism for many organic chemicals; therefore, it is important to know whether biodegradation will occur and the rate of biodegradation of these chemicals.[2]

Although laboratory experiments have been conducted over the past thirty years to elucidate the biodegradability of many organic compounds (roughly 1,000), this knowledge in biodegradability is insufficient compared to the universe of possible organic pollutants (more than 100,000 recorded).[1] Additionally, biodegradation experiments are laborious and time consuming (it usually takes a month to conduct experiments), and it is impractical to conduct laboratory studies for all chemicals under different environmental conditions. Computational prediction tools can be useful in increasing the knowledge of the biodegradability of a compound.

There are three general approaches to biodegradation prediction modeling: utilization of regression analysis, expert opinions and artificial intelligence (AI).[3-5] The regression models have shown to possess the highest utility (currently being adapted by the Environmental Protection Agency). The majority of current regression models rely mainly on structure activity relationships in which statistical models (mostly regressions or Bayesian statistics) are applied based on expert knowledge regarding the biodegradability of organic compounds utilizing their given structure.[5-9] The expert opinions model is similar to the regression analysis approach in which statistical tools (often Bayesian statistics) were applied once experts defined the biodegradability of given compounds. These models ignore other thermodynamical data (such as boiling point, viscosity, and dipole moments), even though some have been shown to be important attributes in prediction models.[10] AI approaches have recently gained attention and have the potential to greatly improve prediction accuracy.[3-5] Noteworthy AI approaches include neural networks and inductive logic programming, which have improved the prediction accuracy of certain chemical classes, but general applications of such models are still lacking.[3, 4, 11]

The majority of the current predictions, highlighted above, rely heavily on quantitative structure activity relationship (QSAR). These relationships have been widely used to predict the chemical properties of a given compound such as its potential to sorp and to volatilize. These descriptors have worked well, but discrepancies between prediction and experimental values regarding their biodegradability are notable.[3, 5, 6] In addition, discrepancies between different computational models with the same input databases are striking. Utilizing the QSAR methods, prediction models ignore other physical/chemical properties (such as boiling point and viscosity) that are very often related to their structure and reactivity and could possibly improve the prediction values. Basu et.al. used not only QSAR but also viscosity in combination with Neural Networks to accurately predict biodegradability of mineral base oils.[10] Currently there are limited studies in developing prediction tools utilizing descriptors other than QSAR.

This article presents a novel approach in predicting biodegradability utilizing other physical chemical/thermodynamical properties that are often ignored in QSAR approaches as input attributes. A database ((http://lyle.smu.edu/cse/dbgroup/biodeg.html) of published physical/chemical properties (such as boiling point, viscosity, heat capacity) was constructed to further aid in the understanding of the potential importance of these attributes in predicting biodegradability. Classification techniques were first used to determine the most important physical/chemical properties as input parameters. Data mining tools (such as Weka[12]) were then used to determine the utility of such an approach. Experiments conducted using this data and our classification method show the benefit of both using the non-physical structure data and the use of traditional data mining classification techniques to address the biodegradation prediction problem.

2. METHODS

The overall goal of our investigation is to find the best accuracy in classifying compounds as biodegradable or not biodegradable. We used Weka[12] as the classification tool on a set of compound data containing many different types of attributes.

2.1 BDMINE

A database of exiting thermodynamic, physical, chemical and biodegradation data has been constructed utilizing existing databases (such as DIPPR and MITI) as well as literature data surveys. The comprehensive tool, BDMINE[13], provides online access to the data with export capabilities. Online documentation is readily available. Overall there are 140 compounds in which all 60 physical/chemical properties as well as biodegradability information are available. This set of data is used for the computational experiments conducted in this research.

2.2 Preprocessing

Real world data may be incorrect, incomplete or noisy. Thus, the KDD process[14] usually involves a first step of preparing data for the data mining step. Preprocessing may be applied to the data to improve the accuracy, efficiency and scalability of the classification process. Preprocessing usually includes data cleaning, data transformation and data reduction.[15]

- Data cleaning: Data cleaning refers to removing or smoothing the noisy data, filling in the missing values.
- Data transformation: In some cases, the data may need to be normalized, which refers scaling the values of a given attribute falling into a specified range, such as interval [0, 1] or [-1, 1]. Since most of the classifiers request numeric values as input, the data may need to convert from string type to numeric type or convert from continuous data to discrete data.
- Data reduction: data reduction reduces volume of data but still produces same or similar analytical result. Reducing volume means reducing rows or columns. Rows are reduced because the items may miss too many values. It is meaningless to classify those items. Attribute reducing is sometimes performed at the suggestion of domain experts based on their research. Another important reason is to remove the redundant or irrelevant attributes but keep the predication results.

Comment [??1]: Unclear

Other familiar data preprocess techniques include outlier detection/removal and data integration, which refers to linking, matching or integrating multiple data sources.

2.2 WEKA

Weka[12] (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms for data mining tasks. It also contains a popular suite of machine learning software written in Java and developed at the University of Waikato. Weka provides a sophisticated graphical user interface, but can also be accessed from the command line. Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. We used Weka (Version 3.6.1) as our software tool to provide biodegradation prediction for the compounds found in the BDMINE[13] dataset. The tools we use are for attribute selection and classification.

2.2.1 Attribute selection

Attribute selection is the process that selects the sub-attributes which preserve the original meaning of the records.[16] The purpose of doing attribute selection is to get the minimum number of attributes from the given dataset which still keeps the classification dependency relationship. From another point of view, the attribute selection gets rid of the irrelevant attributes, which do not preserve the character of the classes described by the original dataset. There are several reasons to do the attribute selection. First, the models generated from the subset are simpler and faster compared to the ones generated from the original dataset. Since the process of attribute selection removes the redundant attributes, this brings another advantage. The attribute selection process provides knowledge of which attributes are highly relevant to the class distribution.

The BDMINE dataset contains 60 attributes. Some of them are irrelevant or redundant attributes. Some of them hold too many missing values. Before doing predication, the attribute selection process needs to filter out those noise attributes that could bother the classification process and lower the predication accuracy. We briefly review the attribute selection processes provided by Weka. All were examined in our experiments to determine the subsets providing the best accuracy. Not all the attribute selections in Weka provide reasonable selected subsets of attributes from the original BDMINE. The improper attribute selections provide either too few or too many attributes from the BDMINE. Too few attributes means that only 1,2, or no attributes are selected and too many attributes means that either almost all attributes are selected or none of them are removed. Table 1 provides an overview of the attribute selection techniques from Weka that we used in our experiments. Table 2 shows the techniques from Weka that we used to search the subsets of attributes determined by the selection techniques. Those

three attribute selection techniques generate three subsets of attributes, which result in good classification accuracies when used as the inputs of the classification algorithm later.

TABLE 1. Attribute Selection Techniques in Weka

	Attribute selector	**Description**
1	ClassifierSubsetEval	This attribute selector evaluates subsets of attributes which behave the best for the provided class. To evaluate the proper subsets of the attributes, the attribute selector employs a selected classifier. For our research, we choose IBk as this classifier.
2	WrapperSubsetEval	WrapperSubsetEval "evaluates attribute sets by using a learning scheme".[17] The user selects a classifier as the learning algorithm and "cross validation is used to estimate the accuracy of the learning scheme for a set of attributes."
3	CfsSubsetEval	This evaluator algorithm "evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them." Subsets of features that are highly correlated with the class while having low intercorrelation are preferred.

TABLE 2. Search Strategies in Weka

	Attribute selector	**Description**
1	RankSearch	This method ranks the attributes by using an attribute evaluator. We chose ClassifierSubsetEval as the attribute evaluator in our research. After an evaluator is specified, a forward selection search is used to generate a ranked list. The parameter stepSize allows the algorithm to specify how many attributes (from the ranked list attributes) add to the selected attribute set at each iteration.
2	GeneticSearch	GeneticSearch performs a search for attribute subsets using a genetic algorithm.[18]
3	RandomSearch	This method performs a Random search in the space of attribute subsets. The method starts from a random point and reports the best subset found if no start set is supplied. If the method starts from a given set, the method searches randomly for subsets that are as good as or better than the start point with the same or fewer attributes.

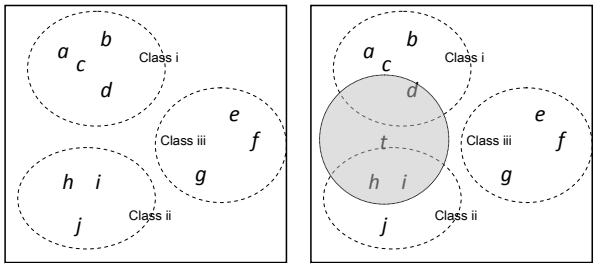
Comment [???]: 'were added' or 'to add'?

2.2.2 Classification

The objective of our work is to predict the biodegradability of a given compound. However, prediction of a precise value is extremely difficult and not very beneficial. Since the biodegradation process is largely based on an enzymatic process, there are many environmental factors (such as amount of nutrients, temperature and dissolved oxygen level) that influences biodegradability, the precise biodegradation percentage (i.e., 90% vs. 93%) has little informative value. Most existing techniques to predict biodegradation actually classify the compound into a category such as biodegradable or not biodegradable. We take this latter approach in our research. Classification assigns an object to one of a predetermined set of categories. It is perhaps the most popular data mining technique.[14] In our experiments we use two classes: biodegradable and not biodegradable. Examples of some popular classification techniques are K Nearest Neighbor, Naïve Bayes, and neural networks. These are the technique we use in Weka.

The K Nearest Neighbors algorithm (k-NN) classifies instances based on the closest training samples.[14] To make a decision to which class a new instance belongs, the K nearest items in the training set are considered. The new instance is placed in the class which holds the most items in those K nearest items. The distance measurement usually uses the Euclidean distance or Manhattan distance. Figure 1 shows a 3-NN example, which implies K = 3. To classify new instance *t*, the nearest three items in the training dataset need to be found. These three items are *d*, *h* and *i*. and the new instance *t* is placed into *class ii* since two of those items come from *class ii*.

Comment [??3]: This sentence is confusing



Training dataset Classify instance *t*

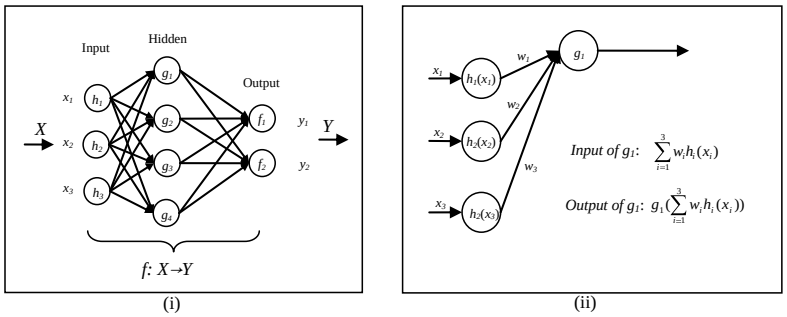
Figure 1. Classification of K Nearest Neighbors Algorithm

A naive Bayes classifier classifies instances based on the probability by Bayes' theorem with attribute independence assumptions. This approach places an instance in a class with the highest posterior probability where prior probabilities are derived from the training set. Advantages of Naive Bayes classifier are that it is simple and requires a small amount of training data to perform predictions. Since a

Naive Bayes classifier assumes that all attributes are strongly independent, it might not provide a high degree of accuracy because, in most of the cases, the attributes of the data somehow are related.

An Artificial Neural Network (usually called "neural network", NN) is a mathematical model or computational model, which simulates the functions and structures of biological neural networks.[19] A neural network model could be logically viewed as a mathematical function defined as $f: X \rightarrow Y$ or could be viewed as a directed graph with many nodes and directed arcs between them. Usually, a neural network is structured in three types of layers. One is the input layer, which is the interface of the NN to receive the input data. The output layer output the prediction results. There could be one or more hidden layers which are between input and output layers. Each node of NN corresponds to an activation function and each arc is associated with a weight. The weights in the NN may be determined in two ways.[14] The weight could be predetermined by domain experts or by the more common approach, adjusted via a training process. One node gets the input values from its incoming arcs and treats the summation of those values as input variable of its activation function. After computation, the node puts the result to the outgoing arcs. Each outgoing arc relays the value of product between the result and its weight to the next nodes. A neural network example is shown in Figure 2 (i). This neural network has three layers, one input layer, one hidden layer and one output layer. Figure 2 (ii) gives the computation of input and output of node g_1.

Comment [??4]: unclear



(i) (ii)

Figure 2. A Neural Network example

There are advantages using NNs for prediction. NNs provide high accuracy prediction rate if the appropriate training has been performed. Compared with other classifiers, NNs are robust, especially under high noise levels. But NNs are difficult to understand and the structures of NNs are complicated. The weights of NNs might not be converging. It requires domain experts carefully design the structures of NNs and wisely choose the training strategy.

2.2.3 K-fold Cross Validation

Cross validation[19] is "a technique for assessing how the results of a statistical analysis will generalize to an independent data set".[15] For a given dataset, a K-fold Cross Validation method divides the input dataset into k subsets. At each iteration time, one of the k subsets is used as the testing set and the rest of other k-1 subsets form a training set. The total number of trials is thus k. Average accuracy and error rates are computed across all k trials. Notice that no matter how the data gets divided, each item gets to be in a testing set exactly once and gets to be in a training set k-1 times. The advantage is that all items are used for both training and testing but each item is used for testing only once.

3. RESULTS

This section describes the results of performing the Weka attribute selection and subsequent classification on BDMine[13] data.

3.1 Preprocessing

The Aerobic HalfLife Complete Set[13] contains 142 records and 50 attributes. Our initial preprocessing step was performed to remove attributes and records that either would bias the results or hurt the results:

- We first removed attributes that are completely unrelated to our prediction problem and are arbitrarily assigned: *CAS* and *Cultivation Duration.*
- We then removed attributes that are directly related to biodegradation prediction: *Anaerobic Biodegradation and Aerobic Half life (hour).* Including this would have biased results.
- We also removed attributes that seem to be mainly related to the physical structure of the compound: *Formula, SMILES* and *Structure.* Recall that we want to investigate the effectiveness of prediction without direct knowledge of the physical structure of the compound.
- Creating a classification using attributes with many missy values, would be ineffective. Thus we removed attributes whose missing rate was equal or higher than 15%. Table 3 gives the attribute missing rate of BD Mine data and shows which attributes were removed.
- We also removed the records which do not have *BOD Biodegradability %* values since this is the classification target attribute and sorted the records by *BOD Biodegradability %* from smallest to largest since it could improve the classification models.

Comment [??5]: missing?

Table 3. Missing Rate of BDMine Attributes

Attributes	Removed	# values missing (out of 142)	% missing

CAS	**Y**	0	0%
Structure	**Y**	0	0%
Cultivation Duration (days)	**Y**	0	0%
BOD Biodegradability (%)		5	3.521127%
Activity Coefficient of Chemical in Water (unit less)		16	11.26761%
Activity Coefficient of Water in Chemical (unit less)	**Y**	74	52.11268%
Aerobic Half life (hour)	**Y**	0	0%
Autoignition Temperature (K)	**Y**	35	24.64789%
Biochemical O2 Demand (g O2/g chem)	**Y**	61	42.95775%
Bioconcentration Factor (unit less)		19	13.38028%
Critical Pressure (Pa)	**Y**	34	23.94366%
Critical Temperature (K)	**Y**	34	23.94366%
Critical Volume (m 3/kmol)	**Y**	33	23.23944%
Daphnia mag ,48h ,LC50 (mg /L)	**Y**	83	58.4507%
Dichromate Chemical O2 Demand (g O2/g chem)		1	0.704225%
Family		0	0%
Flash Point (K)	**Y**	40	28.16901%
Formula	**Y**	0	0%
Heat of Combustion (J/kmol)		18	12.67606%
Heat of Vaporization @ 25 C (J/kmol)	**Y**	44	30.98592%
Heat of Vaporization as f (T) (J/kmol)	**Y**	55	38.73239%
Heat of Vaporization at NBP (J/kmol)	**Y**	26	18.30986%
Ideal Gas Heat of Formation (J/kmol)		8	5.633803%
Liquid Density @ 25 C (kg /m 3)	**Y**	53	37.32394%
Liquid Density as f (T) (kg /m 3)	**Y**	38	26.76056%
Log Koc (cm 3/g OC)	**Y**	44	30.98592%
Log Kow (unit less)		0	0%
Lower Flammability Limit in Air (vol %in air)	**Y**	49	34.50704%
Melting Point (K)		2	1.408451%
Molecular Diffusivity in Air (m 2/s)		18	12.67606%
Molecular Diffusivity in Water (m 2/s)	**Y**	22	15.49296%
Molecular Weight (kg /kmol)		2	1.408451%
Normal Boiling Point (K)		9	6.338028%
Refractive Index (unit less)	**Y**	44	30.98592%
SMILES	**Y**	1	0.704225%
Surface Tension as f (T) (N/m)	**Y**	51	35.91549%
Surface Tension @ 25 C (N/m)	**Y**	51	35.91549%
Theoretical O2 Demand ,Carboceous (g O2/g chem)		1	0.704225%
Theoretical O2 Demand ,Combined (g O2/g chem)		0	0%
Thermal conductivity as f (T), liquid (W/m)	**Y**	51	35.91549%
Thermal conductivity as f (T), vapor (W/m)	**Y**	59	41.5493%
Upper Flammability Limit in Air (vol %in air)	**Y**	53	37.32394%
Vapor Pressure @ 25 C (Pa)		9	6.338028%
Vapor Pressure as f (T) (Pa)		0	0%
Vapor Viscosity as f (T) (Pa)	**Y**	55	38.73239%

Henrys Constant (k Pa mol /mol)		0	0%
Liquid Heat Capacity as f (T) (J/kmol K)	**Y**	53	37.32394%
Liquid Viscosity as f (T) (Pa s)	**Y**	46	32.39437%
Vapor Heat Capacity as f (T) (J/kmol K)	**Y**	142	100%
Anaerobic Biodegradation	**Y**	0	0%

As we discussed above, attribute *BOD Biodegradability %* is the classification target. Notice that *BOD Biodegradability %* is a continuous attribute, which ranges in [-0.03, 1.23]. To apply classification, we have to convert this range into two subranges. One will represent our "biodegradable" class while the other the "not biodegradable" class. We designed class A (not biodegradable) ranging in [-0.03, 0.73] and class B (biodegradable) ranging in (0.73, 1.23] since 0.73 is the point which maximizes the prediction performance of BDMine data. We then created a new attribute, *BOD Biodegradability class*, and appropriately initialized its values as A or B. The old *BODBiodegradability* attribute was then deleted. Note that the new attribute is used to define the desired classes during training.

3.2 Attribute Selection

Next, we needed to perform the attribute reduction. This step determines the best subsets of attributes which still preserve the classification properties. Not all the attribute selections provided by Weka work properly for the BDMINE. We chose three using the Weka attribute selection tools discussed earlier, which provide reasonable selected subsets of attributes from the original BDMINE. Table 4 gives the results of this step. Each selection/search algorithm combination resulted in slightly different results. We are thus given subsets of four or five attributes.

Table 4. BDMine Attribute Selection

Subset	Selection algorithm	Selected Attributes
1	ClassifierSubsetEval (IBk with KNN=3) + RankSearch (CfsSubsetEval with missing separate=true)	Aerobic Half life (hour), Family, Ideal Gas Heat of Formation (J/kmol), Log Kow (unit less) **(4 attributes)**
2	WrapperSubsetEval (IBk with KNN=3) + GeneticSearch	Activity Coefficient of Chemical in Water (unit less), Family, Ideal Gas Heat of Formation (J/kmol), Log Kow (unit less), Melting Point (K) **(5 attributes)**
3	CfsSubsetEval (missing	Activity Coefficient of Chemical in Water (unit less), Aerobic Half life (hour), Family, Ideal Gas Heat of

	separate=true)+GreedyStepwise	Formation (J/kmol), Log Kow (unit less) **(5 attributes)**

3.3 Classification

We now report on the results of applying the three chosen Weka classifiers (K Nearest Neighbors, Naive Bayes and Neural Network) using the three attribute subsets indicated in Table 4. Table 5 shows the results of the nine experiments performed. We report on the prediction accuracy based on experiments performed using 10 cross validation technique. Table 5 also gives the parameters used in Weka (default if not mentioned).

Table 5. Accuracy Rate of Different Classifiers

Algorithm	Classifier in Weka	Subset of attributes	Accuracy rate
K-NN	IBk (k = 3)	1	81.0219 %
		2	81.7518 %
		3	82.4818 %
Naive Bayes	NaiveBayes	1	79.562 %
		2	59.854 %
		3	57.6642 %
Neural Network	MultilayerPerceptron (nominalToBinaryFilter= false)	1	85.4015 %
		2	86.1314 %
		3	84.6715 %

The results show that the neural network has the best performance among the three classification techniques regardless of attributes used. K-NN is consistently second while NaiveBayes is always the worst. Compared with the other two, the accuracy rate of NaiveBayes has the widest range, from 57.6642 % to 79.562%. The reason is that the NaiveBayes classifier assumes the strong independence of attributes. It suggests that the attributes of the first subset preserve less relationship compared with the other two subsets.

Table 6. Neural Network Classification for Three Subsets of Attributes

Subset of attributes	Accuracy rate	ROC area	Confusion Matrix

1	85.4015 %	0.878	a b <-- prediction 96 8 | a = A 12 21 | b = B
2	86.1314 %	0.899	a b <-- prediction 94 10 | a = A 9 24 | b = B
3	84.6715 %	0.88	a b <-- prediction 97 7 | a = A 14 19 | b = B

Table 6 shows the accuracy rate, OCR area and Confusion Matrix for each of the NN classifiers. Both accuracy rate and ROC area suggest that the second subset of attributes are slightly better than the other two subsets. The following research will compare the predications made by NN with the BOWINS predications.

3.4 Comparing with BOWINS

The most popular biodegradation prediction program is the Biotransformation Probability Program for Windows (BIOWIN) developed by the U.S. Environmental Protection Agency. (U.S. EPA, http://www.epa.gov/oppt/exposure/docs.episuited1.htm). BIOWIN calculated the probability of rapid or slow biotransformation for a given chemical under aerobic conditions with mixed cultures of microorganisms. BIOWIN divided up the target compound into several fragments based on its structure. The BIOWIN program then used several approaches (such as expert surveys and multiple linear and non-linear regressions) to assign a biodegradability value to each fragment and summed the fragment values to assign an overall biodegradability to the compound. The 4 models used here to compare with our model are the linear probability model, non-linear probability model, Japanese MITI linear model, and Japanese MITI non-linear model. The developed linear and non-linear models used 186 chemicals as the dataset, while the MITI linear and non-linear models used approximately 900 discrete substances as the dataset to determine their regression.

If keeping the same definition of class A (not biodegradable) and class B (biodegradable) defined in section 3.1, Table 7 shows the predication accuracy rates made by BOWINS based on the same data used by NN classifiers. More precisely, we classify the outputs of BOWINS as class A if the prediction result is no greater than 0.73 or count it as class B if the prediction result is greater than 0.73.

Table 7. BOWINS prediction accuracy

	Prediction techniques	Accuracy rate
1	Linear Model Prediction	70.07 %
2	Non-Linear Model Prediction	65.69 %

3	MITI Linear Model Prediction	86.86 %
4	MITI Non-Linear Model Prediction	84.67 %

By comparing the accuracy rates with NN classifiers, which has the accuracy rate of near 86%, we conclude NN, MITI Linear Model Prediction and MITI Non-Linear Model provide similar prediction accuracy rates. On the other hand, BIOWIN models utilizing linear and non-linear model showed a much lower accuracy rate compared with our NN model. This result demonstrates the NN model using physical/chemical/thermodynamical data rather than structures can achieve high accuracy and is performing better than two current popular models.

3. CONCLUSIONS

The use of artificial intelligence rather than the traditional correlations derived from quantitative structural activity relationship to predict biodegradability of organic compounds has shown to be successful. The artificial intelligence method utilizes basic physical/chemical/thermodynamical data (other than structures) to perform prediction regarding biodegradability of a given compound and the prediction results were similar if not better in some cases than the popular current method (BIOWIN). The results encourage further exploration of using artificial intelligence for predicting the behavior of organic compounds using different physical/chemical/thermodynamical properties.

4. ACKNOWLEDGEMENTS

The authors wish to acknowledge the work and inspiration provided by Jennifer Lewis in an earlier class project that provided the inspiration for the approach used in this paper. The BDMINE project database was initially developed and tested by Pablo Legorreta. Of course this work would not have been possible without this data. Thanks Jennifer and Pablo!

References

(1) Boethling, R. S.; Lynch, D. G. Biodegradation of US premanufactured notice chemicals in OECD tests. *Chemosphere* **2007**, *66*, 715-722.

(2) Alexander, M. *Biodegradation and Bioremediation*, 2nd ed.; Academic Press: New York, 1994.

(3) Baker, J. R.; Gamberger, D.; Mihelcic, J. R.; Sabljic, A. Evaluation of artificial intelligence based models for chemical biodegradability prediction. *Molecules* **2004**, *9*, 989-1003.

(4) Klopman, G.; Tu, M. H. Structure-biodegradability study and computer-automated prediction of aerobic biodegradation of chemicals. *Environ. Toxicol. Chem.* **1997**, *16*, 1829-1835.

(5) Rorije, E.; Loonen, H.; Muller, M.; Klopman, G.; Peijnenburg, W. Evaluation and application models for the prediction of ready biodegradability in the MITI-I test. *Chemosphere* **1999**, *38*, 1409-1417.

(6) Raymond, J. W.; Rogers, T. N.; Shonnard, D. R.; Kline, A. A. A review of structure-based biodegradation estimation methods. *J. Hazard. Mater.* **2001**, *84*, 189-215.

(7) 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.

(8) Rorije, E.; Peijnenburg, W.; Klopman, G. Structural requirements for anaerobic biodegradation of organic chemicals: A fragment model analysis. *Environ. Toxicol. Chem.* **1998**, *17*, 1943-1950.

(9) Meylan, W.; Boethling, R.; Aronson, D.; Howard, P.; Tunkeli, J. Chemical structure-based predictive model for methanogenic anaerobic biodegradation potential. *Environ. Toxicol. Chem.* **2007**, *26*, 1785-1792.

(10) Basu, B.; Singh, M. P.; Kapur, G. S.; Ali, N.; Sastry, M. I. S.; Jain, S. K.; Srivastava, S. P.; Bhatnagar, A. K. Prediction of biodegradability of mineral base oils from chemical composition using artificial neural networks. *Tribol. Int.* **1998**, *31* (4), 159-168.

(11) Dzeroski, S.; Blockeel, H.; Kompare, B.; Kramer, S.; Pfahringer, B.; Laer, W. V. Experiments in predicting biodegradability. *Lect. Notes Comput. Sci.* **1999**, *1634*, 80-91.

(12) Weka 3.6.1: Data mining software. University of Waikato, Hamilton, New Zealand.

(13) BDMine. Computer model for predicting contaminant biodegradation. Southern Methodist University, 2008. http://lyle.smu.edu/~mhd/ida/biodeg.html

(14) Dunham, M. *Data Mining Introductory and Advanced Topics*, 1st ed.; Prentice Hall: New York, 2002.

(15) Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, CA, 2006.

(16) Caruana, R.; Freitag, D. Greedy attribute selection. In *Machine Learning*, Proceedings of the Eleventh International Conference on Machine Learning, New Brunswick, NJ, July 10-13, 1994; Cohen, W., Hirsh, H., Eds.; Morgan Kaufman: San Francisco, CA, 1994.

(17) Pentaho™ Data Mining: Tools for machine learning and data mining. Pentaho Community. http://wiki.pentaho.com/display/DATAMINING/WrapperSubsetEval (accessed Aug 2009).

(18) Goldberg, D.; Holland, J. Genetic algorithms and machine learning. *Machine Learning*. **1988**, *3* (2-3), 95-99.

(19) Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model

selection. In *IJCAI*, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montreal, Canada, Aug 20-25, 1995; Elsevier: New York, 1995; Vol. 2, p 1137-1145.