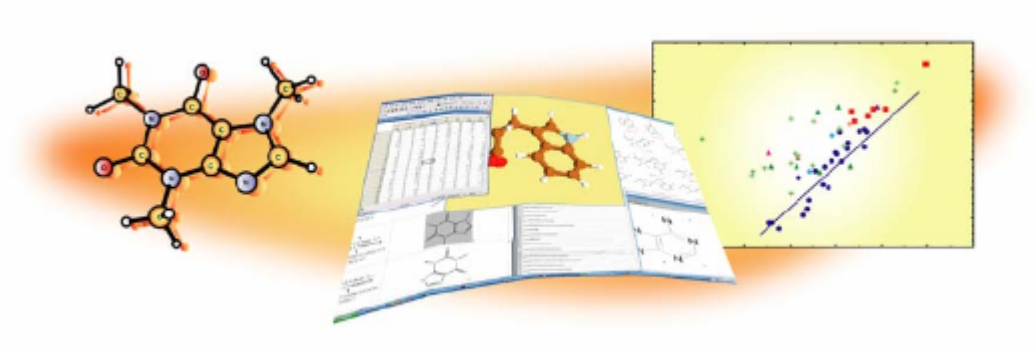
QSAR Biodegradation

ANJUMANARA ATHINA (155142), NICOLAS CHUANG HUI JIANG (152504), LIM WEI SEN (154151), ONG JIA JIE (153818)



Many regulatory laws resulting from the enactment of the United Nations Stockholm Convention in May 2004, together with the new REACH legislation, have promoted significant new activity in the assessment of Persistent, Bio accumulative and Toxic (PBT) substances. These are chemicals that have the potential to persist in the environment, accumulate within the tissues of living organisms and, in the case of chemicals categorized as RBTs, show adverse effects following long-term exposure. Under REACH, estimated data generated by (Q)SARS may be used both as a substitute for experimental data, and as a supplement to experimental data in weight-of-evidence approaches.

Aim: To use OSAR data from chemical compounds and build OSAR models to predict ready biodegradation of chemicals using different modeling methods and type of molecular descriptors.

Dataset Description

The QSAR biodegradation dataset was built in the Milano Chemometrics and QSAR Research Group (UniversitÃ degli Studi

Milano â€" Bicocca, Milano, Italy). The research leading to these results has received funding from the European Communityâ€™s Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n. 238701 of Marie Curie ITN Environmental Chemoinformatics (ECO) project. Biodegradation experimental values of 1055 chemicals were collected from the webpage of the National Institute of Technology and Evaluation of Japan (NITE)

	SpMax_L	J_Dz(e)	nHM	F01[N-N]	F04[C-N]	NssssC	nCb-	C%	nCp	nO	...	nCrt	C-026	F02[C-N]	nHDon	SpMax_B(m)	Psi_i_A	nN	SM6_B(m)	nArCOOR
0	3.919	2.6909	0	0	0	0	0	31.4	2	0	...	0	0	0	0	2.949	1.591	0	7.253	0
1	4.170	2.1144	0	0	0	0	0	30.8	1	1	...	0	0	0	0	3.315	1.967	0	7.257	0
2	3.932	3.2512	0	0	0	0	0	26.7	2	4	...	0	0	0	1	3.076	2.417	0	7.601	0
3	3.000	2.7098	0	0	0	0	0	20.0	0	2	...	0	0	0	1	3.046	5.000	0	6.690	0
4	4.236	3.3944	0	0	0	0	0	29.4	2	4	...	0	0	0	0	3.351	2.405	0	8.003	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1050	5.431	2.8955	0	0	0	2	0	32.1	4	1	...	2	0	6	1	3.573	2.242	1	8.088	0
1051	5.287	3.3732	0	0	9	0	0	35.3	0	9	...	0	0	3	0	3.787	3.083	3	9.278	0
1052	4.869	1.7670	0	1	9	0	5	44.4	0	4	...	0	4	13	0	3.848	2.576	5	9.537	1
1053	5.158	1.6914	2	0	36	0	9	56.1	0	0	...	0	1	16	0	5.808	2.055	8	11.055	0
1054	5.076	2.6588	2	0	0	0	4	54.5	0	0	...	0	2	0	0	4.009	2.206	0	9.130	0

Table 1: Samples of the dataset

Data Analysis

The dataset is analyzed in order to gain insight from the dataset. Data visualization technique such as scatter plot is used to analyze the relationship between two features. Examples of the plotting are given in Figure 1.

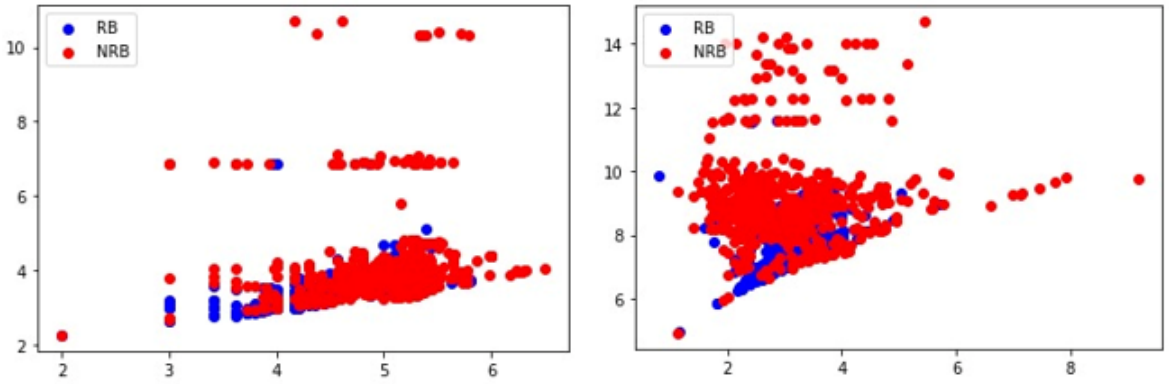


Figure 1: (Left) SpMax_L against SpMax_B(m), (Right) J_Dz(e) against SM6_B(m)

The bivariate analysis is performed to determine the relationship between the features and target. The bivariate analysis is performed by performing the feature selection using ANOVA, ANOVA is chosen since the features are numerical data and the target is categorical data. Six features are selected out of twenty features. Table 2 shows the selected features:

Feature	Type	Value/Statistics
SpMax_L	Numerical continuous	Range: 2 - 6.496 Mean: 4.7826 Std: 0.5469
J_Dz(e)	Numerical continuous	Range: 0.8039 - 9.1775 Mean: 3.0695 Std: 0.8313
SpMax_B(m)	Numerical continuous	Range: 2.267 - 10.695 Mean: 3.9182 Std: 0.9996
SM6_B(m)	Numerical Discrete	Range: 4.917 - 14.70 Mean: 8.629 Std: 1.2419
nHM	Numerical Discrete	Range: 0 - 12 Mean: 0.7166 Std: 1.4624
nN	Numerical Discrete	Range: 0 - 8 Mean: 0.6862 Std: 1.09038

Table 2: The selected features

Data Modeling

Two predictive models are built using Neural Network and Fuzzy Logic system. The ratio of the split is 80% training set and 20% test set. Another split is 8:2 for validation set. The split set is only used by Neural Network.

The parameters for the predictive models are given below.

Algorithm	Value/Statistics
Neural Network	3 layers: input, hidden, output binary classification
Fuzzy Logic system	Fuzzy rules

Table 3: Parameters of the predictive models

The result of the classification of each predictive models are given below.

```
0.8957345971563981
[[134  13]
 [  9  55]]
              precision    recall  f1-score   support

           0       0.94      0.91      0.92       147
           1       0.81      0.86      0.83        64

    accuracy                           0.90       211
   macro avg       0.87      0.89      0.88       211
weighted avg       0.90      0.90      0.90       211
```

Figure 2: Results of classification using Neural Network model.

```
0.3412322274881517
[[  4 695]
 [  0 356]]
              precision    recall  f1-score   support

           0       1.00      0.01      0.01       699
           1       0.34      1.00      0.51       356

    accuracy                           0.34      1055
   macro avg       0.67      0.50      0.26      1055
weighted avg       0.78      0.34      0.18      1055
```

Figure 3: Results of classification using Fuzzy Logic model.

For classification, we can do this by calculating the correct and incorrect classification and put them in a table called confusion matrix. The total number of correctly identified in yes class is true positive while the opposite is true negative. The number of instance which belongs to yes but misclassified as no is false negative and the opposite is false positive.

	Yes	No
Yes	True Positive (TP)	False Negative (FN)
No	False Positive (FP)	True Negative (TN)

A dataset contains values for 41 attributes, or the molecular descriptors and they are used to classify chemicals into 2 classes which are ready biodegradable and not ready biodegradable. Predictive models that we use to predict the target variable of the dataset are Neural Network and Fuzzy logic system. Neural Network

and Fuzzy logic system are both important methods in computational intelligence.

Neural Network is a series of algorithms which is similar to the ways of human brains functioning that can differentiate the relationship of data. Neural Network consists of three main layers which is input layer, processing layer and output layer. Fuzzy logic is method of reasoning that is similar to human reasoning which involves the if-else statement and the possibility of choosing between yes or no.

The main difference for neural network and fuzzy logic is that neural network is based on neurons in human brain to do computations, while the fuzzy logic is method of reasoning that resembles human reasoning and decision-making. Normally, fuzzy logic is used for pattern recognition and neural network is for performing predictions. Fuzzy logic is much simpler than neural network.

The formula to calculate precision is shown below:

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

For Neural Network, the precision for Not-ready biodegradable(NRB) is 0.94, while the precision for Ready biodegradable (RB) is 0.81. For the macro avg, it is 0.87 and for the weighted avg, it is 0.90.

For Fuzzy Logic system, the precision for Not-ready biodegradable(NRB) is 1.00, while the precision for Ready biodegradable (RB) is 0.34. For the macro avg, it is 0.67 and for the weighted avg, it is 0.78.

The formula to calculate recall is shown below:

$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

For Neural Network, the recall for Not-ready biodegradable(NRB) is 0.91, while the recall for Ready biodegradable (RB) is 0.86. For the macro avg, it is 0.89 and for the weighted avg, it is 0.90.

For Fuzzy Logic system, the recall for Not-ready biodegradable(NRB) is 0.01, while the recall for Ready biodegradable (RB) is 1.00. For the macro avg, it is 0.50 and for the weighted avg, it is 0.34.

The formula for calculating F1-score is:

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

For Neural Network, the F1-score for Not-ready biodegradable(NRB) is 0.92, while the recall for Ready biodegradable (RB) is 0.83. For the macro avg, it is 0.88 and for the weighted avg, it is 0.90.

For Fuzzy Logic system, the F1-score for Not-ready biodegradable(NRB) is 0.01, while the recall for Ready biodegradable (RB) is 0.51. For the macro avg, it is 0.26 and for the weighted avg, it is 0.18.

F1-score is a method to see the balance or equilibrium between the precision and recall. Overall, Neural Network has higher F1-score than the Fuzzy logic system.

The formula for calculating accuracy is:

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

For Neural Network, the accuracy is 0.90, while the accuracy of Fuzzy Logic system is 0.34. Accuracy of Neural Network is higher than the Fuzzy Logic system. The accuracy in fuzzy logic is low because we have to make our own fuzzy rule and we do not know the parameter's high, medium and low range for accurate prediction.

Not Ready Biodegradable (NRB)

Neural Network		**Fuzzy Logic system**
0.94	**Precision**	1.00
0.91	**Recall**	0.01
0.92	**F1-score**	0.01
0.90	**Accuracy**	0.34

Table 4: Not Ready Biodegradable (NRB)

For Not-Ready biodegradable class, Neural Network has higher recall and F1-score than Fuzzy logic system, while Fuzzy Logic has higher precision than Neural Network.

Ready Biodegradable (RB)

Neural Network		Fuzzy Logic system
0.81	**Precision**	0.34
0.86	**Recall**	1.00
0.83	**F1-score**	0.51
0.90	**Accuracy**	0.34

Table 5: Ready Biodegradable (RB)

For Ready biodegradable class, Neural Network has higher precision and F1-score than Fuzzy logic system, while Fuzzy Logic has higher recall than Neural Network. However, in overall, Neural Network has higher accuracy than Fuzzy Logic system.

Therefore, we can say that Neural Network is the best algorithm.