

Doctoral Thesis

Comprehensive evaluation of  
prediction for residual pesticides in  
foods by quantitative structure-  
property relationship

SERINO Takeshi

Program of Information Science and Engineering  
Graduate School of Science and Technology  
Nara Institute of Science and Technology

Supervisor: Prof. Shigehiko Kanaya  
Computational Systems Biologoy Lab

Submitted on 31/07/2020

A Doctor's Thesis  
submitted to Graduate School of Science and Technology,  
Nara Institute of Science and Technology  
in partial fulfillment of the requirements for the degree of  
Doctor of ENGINEERING

Takeshi Serino

Thesis Committee:

Professor Shigehiko Kanaya  
(Supervisor, Division of Information Science)  
Professor Keiichi Yasumoto  
(Co-supervisor, Division of Information Science)  
Associate Professor MD.Altaf-Ul-Amin  
(Co-supervisor, Division of Information Science)  
Associate Professor Naoaki Ono  
(Co-supervisor, Division of Information Science)  
Assistant Professor Ming Huang  
(Co-supervisor, Division of Information Science)

# Comprehensive evaluation of prediction for residual pesticides in foods by quantitative structure-property relationship\*

Takeshi Serino

## Abstract

Residual pesticides in foods are routinely analyzed by Mass Spectrometer combined with chromatograph, such as LC-MS and GC-MS. Over 700 pesticides are listed in the latest Japan Positive List by the Ministry of Human Labor and Warfare for ensuring the food safety. As various pesticides are used for production of various crops, there are technical challenges in simultaneous residual pesticides analysis by GC-MS and LC-MS due to the vast chemical diversity of pesticides. Food analysis laboratories (Food Labs) need to inject multiple samples actually for method development. For addressing this issue, I evaluated two predictions by the quantitative structure property relationship, 1. residual pesticides recovery rate and 2. classification of pesticides amenability between LC-MS and GC-MS. Both predictions utilize the existing validation reports of food analysis. Molecular descriptors (MDs) of pesticides were used as the explanatory variables for prediction, obtained using canonical SMILES strings. The caret package of R program was used for machine learning with various algorithms. Selection of highly correlated MDs was also investigated by the correlation analysis and cluster analysis. The procedures developed by this study enable Food Labs to predict the recovery rate and LC/GC amenability of pesticides without actual sample injections, contributing to the reduction of labors and cost for method development.

## Keywords:

Pesticides, GC-MS, LC-MS, Molecular Descriptors, Machine Learning, QSPR

---

\*Doctoral Dissertation, Graduate School of Information Science, Nara Institute of Science and Technology, July 31, 2020.

# 定量的構造物性相関による食品中残留農薬の 網羅的な予測評価\*

芹野 武

## 内容梗概

食品中の残留農薬分析は LC-MS や GC-MS などのクロマトグラフィーと質量分析計を組み合わせたシステムで測定されている。最新の日本の厚生労働省のポジティブリストによると、食の安全を担保することを目的として、700 以上の農薬が一斉分析法に掲載されている。多くの種類の農薬が様々な作物で使用されており、多様な化学的特性により GC-MS や LC-MS による一斉分析において技術的な問題が存在する。食品中の残留農薬分析を行う食品分析ラボでは、これらの問題を、コストがかかる実サンプルの測定を何度も行い分析メソッドを修正・開発することで現状対応している。本研究ではこれらの問題解決の手法として、定量的構造物性相関 (QSPR) によって 1. 残留農薬の回収率の回帰予測 および 2. 農薬の分類予測を行い、それぞれのパフォーマンスを評価した。2つの予測では既存の食品分析の検証されたデータを用いた。農薬の化学的な特性を分子レベルで表す分子記述子 (MD) を予測の説明変数として用いた。224 種類の分子記述子が農薬の canonical SMILES によって得られた。R プログラムの caret パッケージにより様々なアルゴリズムで機械学習を適用した。強い相関のある分子記述子について相関分析とクラスタ分析で検討を行った。本研究の QSPR によって開発された予測手法は、実際の測定がなくても未知の残留農薬の回収率の予測や、分析手法を GC-MS から LC-MS に測定を変更する際の予備検討などで活用することが可能で、食品分析ラボのメソッド開発の工数やコストの削減に貢献することが期待される。

## キーワード

農薬, GC-MS, LC-MS, 分子記述子, 機械学習, QSPR

\*奈良先端科学技術大学院大学 情報科学研究科 博士論文, 2020 年 7 月 31 日.

# Contents

<b>1. Introduction</b>	<b>1</b>
1.1 Pesticides . . . . .	1
1.2 Pesticide Analysis for food safety . . . . .	1
1.3 Molecular descriptors . . . . .	3
1.4 Machine Learning methods . . . . .	4
1.5 Dissertation outline . . . . .	5
<b>2. Pesticide recovery rate prediction for seven crops</b>	<b>6</b>
2.1 Introduction . . . . .	6
2.2 Materials and Methods . . . . .	6
2.2.1 List of pesticides with recovery rate . . . . .	6
2.2.2 Molecular descriptors of the pesticides . . . . .	7
2.2.3 Regression analysis of pesticides recovery rate prediction by machine learning . . . . .	10
2.3 Results and Discussion . . . . .	12
2.3.1 Pesticide recovery rate distribution . . . . .	12
2.3.2 Machine Learning results for execution time and prediction error . . . . .	14
2.3.3 Generalization performance for prediction . . . . .	17
2.4 Conclusion of Chapter 2 . . . . .	22
<b>3. Selection of optimum molecular descriptors for recovery rate pre- diction using graph clustering tool</b>	<b>23</b>
3.1 Introduction . . . . .	23
3.2 Materials and Methods . . . . .	23
3.2.1 Correlation analysis and cluster analysis . . . . .	23
3.2.2 Machine learning by ‘caret’ package . . . . .	27
3.3 Results and Discussion . . . . .	28
3.3.1 Correlation analysis among molecular descriptors . . . . .	28
3.3.2 Selection of molecular descriptors by the clustering tool - DP Clus . . . . .	28

3.3.3	Selection of molecular descriptors for machine learning after cluster analysis by DP Clus . . . . .	50
3.3.4	Comparison of machine learning performance between with and without selection of molecular descriptors . . . . .	54
3.4	Conclusion of Chapter 3 . . . . .	60
<b>4.</b>	<b>Optimum molecular descriptors selection by hierarchical cluster analysis</b>	<b>62</b>
4.1	Introduction . . . . .	62
4.2	Materials and Methods . . . . .	62
4.2.1	Correlation analysis among molecular descriptors . . . . .	62
4.2.2	Machine learning by ‘caret’ package . . . . .	66
4.3	Results and Discussion . . . . .	66
4.3.1	Correlation analysis among molecular descriptors . . . . .	66
4.3.2	Selection of molecular descriptors by the clustering tool - hierarchical cluster analysis . . . . .	68
4.3.3	Comparison of PE among MD6, MD4, MD2 and MD0 for machine learning method category . . . . .	84
4.4	Conclusion of Chapter 4 and pesticide recovery prediction by regression model . . . . .	87
<b>5.</b>	<b>Classification of pesticides amenability between LC or GC</b>	<b>89</b>
5.1	Introduction . . . . .	89
5.2	Materials and Methods . . . . .	89
5.2.1	Preparation of the pesticide list . . . . .	89
5.2.2	Molecular descriptors of the pesticides . . . . .	90
5.2.3	Classification of pesticides by the machine learning . . . . .	93
5.3	Results and Discussion . . . . .	96
5.3.1	Classification Performance (Accuracy of CV10 resample) . . . . .	96
5.3.2	Execution time (ET) . . . . .	96
5.3.3	Total performance - both Accuracy and Execution Time . . . . .	96
5.3.4	Pesticides which were not accurately predicted by the xgb-DART . . . . .	102
5.4	Conclusion of Chapter 5 . . . . .	109

<b>6. Classification of pesticides amenability between LC or GC with graph clustering tool</b>	<b>110</b>
6.1 Introduction . . . . .	110
6.2 Materials and Methods . . . . .	110
6.2.1 Correlation analysis and cluster analysis . . . . .	110
6.2.2 Machine learning by ‘caret’ package . . . . .	113
6.3 Results and Discussion . . . . .	114
6.3.1 Correlation analysis results among molecular descriptors .	114
6.3.2 Selection of molecular descriptors by the clustering tool - DP Clus . . . . .	114
6.3.3 Molecular descriptors for machine learning . . . . .	136
6.3.4 Comparison of machine learning performance between with and without selection of molecular descriptors . . . . .	140
6.4 Conclusion of Chapter 6 . . . . .	146
<b>7. Classification of pesticides amenability between LC or GC with hierarchical cluster analysis</b>	<b>147</b>
7.1 Introduction . . . . .	147
7.2 Materials and Methods . . . . .	147
7.2.1 Correlation analysis among molecular descriptors . . . . .	147
7.2.2 Machine learning by ‘caret’ package . . . . .	151
7.3 Results and Discussion . . . . .	152
7.3.1 Correlation analysis results among molecular descriptors .	152
7.3.2 Selection of molecular descriptors by the clustering tool - hierarchical cluster analysis . . . . .	152
7.3.3 Comparison of Accuracy among MD6, MD4, MD2 and MD0 for machine learning method category . . . . .	168
7.4 Conclusion of Chapter 7 and pesticide classification prediction models in caret package . . . . .	171
<b>8. Concluding remarks</b>	<b>173</b>
<b>URLs</b>	<b>177</b>
<b>References</b>	<b>178</b>

<b>Achievements</b>	<b>182</b>
<b>Acknowledgements</b>	<b>184</b>
<b>Appendix</b>	<b>185</b>
<b>A. Additional information for regression analysis</b>	<b>185</b>
A.1 Pesticides and Canonical SMILES for regression analysis . . . . .	185
A.2 Average of log PE(Av log PE) and Average of log Execution Time(Av log ET). Regression methods(Method) are ordered according to Av log PE. . . . .	190
<b>B. Additional information for classification analysis</b>	<b>191</b>
B.1 Pesticide and Canonical SMILES for classification analysis . . . . .	191
B.2 Accuracy and log Execution Time(Av log ET). Classification meth- ods (Method) are ordered according to accuracy. . . . .	195
<b>C. Specification of PC and Software programs</b>	<b>196</b>



## List of Figures

1.1	Outline of this dissertation . . . . .	5
2.1	Sample preparation procedure of Japan Positive List . . . . .	7
2.2	Distribution of Pearson correlations between recovery rates and molecular descriptors(left) where X-axis represents Pearson correlations; while Y-axis represents the count of descriptors, and Distribution of Pearson correlations between experimental and predicted recovery rates in 89 regression models(right) where X-axis represents Pearson correlations; while Y-axis represents the count of regression models. . . . .	13
2.3	Dendrogram and heat map of hierarchical cluster analysis of pesticide recovery for seven crops . . . . .	14
2.4	Distribution of log PE and log Execution Time for seven crops. Averages of log PE and log Execution Time corresponding to all regression models are shown along the crop names. “Av” in row represents averages of log PE- and Log Execution Time-averages for seven crops. . . . .	16
2.5	Averages of log PE for seven crops corresponding to a number of regression methods . . . . .	17
2.6	Relationships between average of rank(X-axis) and PE <sub>k</sub> (Y-axis) for seven crops. The range of PE <sub>k</sub> is set between 0 and 1 because <i>k</i> th regression method with <i>PE</i> larger than 1 means no prediction performance according to Eq. 2.3 . . . . .	20
2.7	Box plots for log PE and average of log(Execution Time) for seven crops corresponding to 11 categories of regression methods. . . . .	21
2.8	Relationships between average of log(PE) and average of log(Execution Time) for regression methods. Ordinary learning and Ensemble learning methods corresponds to blue and red circles, respectively. . . . .	22
3.1	Process chart of selecting the optimum MDs . . . . .	24
3.2	Distribution of correlation coefficient of MD . . . . .	28
3.3	Cluster analysis result by DP Clus . . . . .	29
3.4	Relationship of dependent clusters by DP Clus . . . . .	29
3.5	Decision tree to select the MD(s) from each cluster . . . . .	31

3.6	Cluster 1: 26 molecular descriptors correlates as shown in green lines. Eight other clusters are connected by the red lines. . . . .	32
3.7	Cluster 2: 11 molecular descriptors correlates as shown in green lines. Three other clusters are connected by the red lines. . . . .	33
3.8	Cluster 3: Seven molecular descriptors correlate as shown in green lines. Two other clusters are connected by the red lines. . . . .	34
3.9	Cluster 4: Six molecular descriptors correlate as shown in green lines. Three other clusters are connected by the red lines. . . . .	34
3.10	Cluster 5: Six molecular descriptors are correlated as the green line.	35
3.11	Chemical structure of Nereistoxin oxalate, the pesticide with two molecules . . . . .	36
3.12	Cluster 6: Five MDs correlated as green lines. Other five clusters are connecting with red lines. . . . .	36
3.13	Cluster 7: Five MDs are correlated as green lines. Cluster 7 connects with the other four clusters as shown in red lines . . . . .	37
3.14	Cluster 8: Five MDs are correlated as green lines. Cluster 8 connects with the other four clusters as shown in red lines . . . . .	38
3.15	Chemical structure of $\alpha$ -Endsulfan . . . . .	38
3.16	Cluster 9: Four MDs correlate as shown in green lines. Cluster 9 connects with cluster 28 in red lines. . . . .	39
3.17	Cluster 10: Five MDs are correlated as shown in green lines. Cluster 10 connects with the other seven clusters as red lines. . . . .	39
3.18	Cluster 11: Three MDs are correlated in green lines. . . . .	40
3.19	Cluster 12: Five MDs are correlated as shown in green lines. Cluster 12 connects with cluster 4 with red lines. . . . .	41
3.20	Cluster 13: Three MDs correlate as shown in green lines. Cluster 13 connects with three clusters in red lines. . . . .	41
3.21	Cluster 14: Three MDs correlate as shown in green lines. . . . .	42
3.22	Cluster 15: Two MDs correlated with green line, and MDEO.22 is connecting with the khs.ssO of cluster 17 . . . . .	42
3.23	Cluster 16: Four MDs correlate as shown in green lines. Cluster 16 connects with nine clusters in red lines. . . . .	43

3.24	Cluster 17: Two MDs correlate as green line. Both MDs connect to the other clusters. . . . .	43
3.25	Cluster 18: Two MDs correlated as in green line. . . . .	44
3.26	Cluster 19: Two MDs correlate as shown in green line. . . . .	44
3.27	Chemical structure of Propyzamide (left) and Flumioxazin (right). (-C≡CH) is colored in red . . . . .	45
3.28	Cluster 20: Three MDs correlated as shown in green lines. Cluster 20 connected with the cluster 3 and cluster 16 as shown in red lines	45
3.29	Cluster 21: Two MDs correlate by green line. . . . .	46
3.30	Scatter plot of 248 pesticides, AlogP vs. AlogP2 . . . . .	46
3.31	Cluster 22: Three MDs correlate in green line. Two MDs connected to the other clusters in red lines. . . . .	47
3.32	Cluster 23: Four MDs correlated as shown in green lines. Cluster 23 connects with the other clusters as shown in red lines. . . . .	48
3.33	Cluster 24: Two MDs correlate as shown in green line. . . . .	48
3.34	Cluster 25: Two MDs correlate as shown in green line. . . . .	48
3.35	Cluster 26: Two MDs correlated as shown in the green line. . . .	49
3.36	Cluster 27: Tree MDs correlate as shown in green lines. Cluster 27 connects to the cluster 10 as shown in red line. . . . .	49
3.37	Cluster 28: Two MDs correlate as shown in green line. Both MDs connect to the other clusters by red lines. . . . .	50
3.38	Cluster diagram for dependent clusters . . . . .	52
3.39	Comparison of Prediction Error by machine learning method on with and without selection of molecular descriptors by correlation analysis and cluster analysis. Plots on the rectangular line shows no difference by selection of molecular descriptors. . . . .	54
3.40	Boxplot of Prediction Error comparison of the molecular descriptors before selection(left) and after selection(right). . . . .	57
3.41	Boxplot of Execution Time comparison of the molecular descriptors before selection(left) and after selection(right). . . . .	58
4.1	Selection of MD-MD pair from the dendrogram . . . . .	63
4.2	Selection of MD-MD pair from the dendrogram . . . . .	65
4.3	Process chart of selecting the optimum MDs . . . . .	66

4.4	Distribution of correlation coefficient of MD . . . . .	68
4.5	Dendrogram of 118 MDs according to the similarity of 248 pesticides	69
4.6	Dendrogram of 87 MDs for NbClust . . . . .	69
4.7	Histogram of the result of optimum number of cluster according to the gap function . . . . .	70
4.8	Dendrogram of 87 MDs at k=2 . . . . .	71
4.9	Dendrogram of 87 MDs at k=3 . . . . .	72
4.10	Dendrogram of 87 MDs at k=14 . . . . .	72
4.11	Log(PE) and Log(ET) for 89 methods using the MD4 . . . . .	75
4.12	Log(PE) comparison between MD4 and MD0 . . . . .	78
4.13	Log(PE) comparison between MD4 and MD2 . . . . .	79
4.14	Log(PE) and Log(ET) for 89 methods using the MD6 . . . . .	81
4.15	Log(PE) comparison between MD6 and MD0 . . . . .	82
4.16	Log(PE) comparison between MD6 and MD2 . . . . .	83
4.17	Log(PE) comparison between MD6 and MD4 . . . . .	84
4.18	Comparison of Log(PE) comparison by learning method category among MD6, MD4, MD2 and MD0 . . . . .	85
4.19	ET of SBC and xgbLinear for MD6, MD4, MD2 and MD0 with 5 replicates . . . . .	87
5.1	Venn diagram to describe the number of pesticides by the list(FDA or EURL) and the technology used for analysis(L:LC-MS and G:GC- MS). . . . .	90
5.2	Accuracy of classification(CV10 resample) for 119 machine learn- ing methods . . . . .	97
5.3	Execution time for 119 machine learning methods . . . . .	98
5.4	Accuracy and Execution time for 119 machine learning methods in full scale . . . . .	99
5.5	Accuracy and Execution time in expanded view . . . . .	100
5.6	Chemical structure of pesticides wrongly classified by xgbDART (1 of 5) . . . . .	104
5.7	Chemical structure of pesticides wrongly classified by xgbDART (2 of 5) . . . . .	104

5.8	Chemical structure of pesticides wrongly classified by xgbDART (3 of 5) . . . . .	105
5.9	Chemical structure of pesticides wrongly classified by xgbDART (4 of 5) . . . . .	105
5.10	Chemical structure of pesticides wrongly classified by xgbDART (5 of 5) . . . . .	106
5.11	Decision tree of ‘rpart’ to classify 194 pesticides using 176 MDs .	107
5.12	Comparison of XLogP between all 194 pesticides and 35 pesticides wrongly classified by xgbDART . . . . .	108
6.1	Process chart of selecting the optimum MDs . . . . .	111
6.2	Distribution of correlation coefficient of MD . . . . .	115
6.3	Cluster analysis result by DP Clus . . . . .	116
6.4	Relationship dependent clusters by DP Clus . . . . .	116
6.5	Decision tree to select the optimum MDs from each cluster . . . .	118
6.6	Cluster 1: 25 molecular descriptors are correlated as shown in green lines. Cluster 1 connects with the other nine clusters as in red lines.	119
6.7	Cluster 2: 16 MDs correlated as shown in the green lines. Cluster 2 connects with other seven clusters as shown in red lines. . . . .	120
6.8	Cluster 3: 11 MDs correlated as shown in green lines. Cluster 3 connects to other four clusters as shown in red lines. . . . .	121
6.9	Cluster 4: Nine MDs correlate as shown in green lines. Cluster 4 connects with the other three clusters as shown in red lines. . . .	122
6.10	Cluster 5: Nine MDs correlate as shown in green lines. Cluster 5 connects with other two clusters as shown in red lines. . . . .	123
6.11	Cluster 6: Four MDs correlated as shown in green lines. . . . .	124
6.12	Cluster 7: Four MDs correlate as shown in green lines. Cluster 7 connects with other two clusters as shown in red lines. . . . .	124
6.13	Cluster 8: 15 MDs correlate as shown in green lines. Cluster 8 connects with other five clusters as shown in red lines. . . . .	125
6.14	Cluster 9: Six MDs correlate as shown in green lines. Cluster 9 connects with other seven clusters as shown in red lines. . . . .	126
6.15	Cluster 10: Three MDs correlate as shown in green lines. . . . .	127
6.16	Cluster 11: Three MDs correlate as shown in green lines. . . . .	128

6.17	Chemical structure of pesticides with the unique value of MDs for cluster 11. . . . .	128
6.18	Cluster 12: Three MDs correlate as shown in green lines. . . . .	129
6.19	Cluster 13: Eight MDs correlate as shown in green lines. Cluster 13 connects with other six clusters as shown in red lines. . . . .	130
6.20	Cluster 14: Three MDs correlate as shown in green lines. Cluster 14 connects with other two clusters as shown in red lines. . . . .	131
6.21	Cluster 15: Two MDs correlate as shown in green lines. Cluster 15 connects with other two clusters as shown in red lines. . . . .	131
6.22	Cluster 16: Two MDs correlate as shown in green lines. Cluster 16 connects with cluster 7 as shown in red lines. . . . .	132
6.23	Cluster 17: Two MDs correlate as shown in green lines. . . . .	132
6.24	Cluster 18: Two MDs correlate as shown in green lines. Cluster 18 connects with cluster 3 as shown in red lines. . . . .	133
6.25	Cluster 19: Four MDs correlate as shown in green lines. Cluster 19 connects with three clusters as shown in red lines. . . . .	133
6.26	Cluster 20: Two MDs correlate as shown in green lines. . . . .	134
6.27	Cluster 21: Two MDs correlate as shown in green lines. Cluster 21 connects with two clusters as shown in red lines. . . . .	134
6.28	Cluster 22: Two MDs correlate as shown in green lines. . . . .	135
6.29	Cluster 23: Two MDs correlate as shown in green lines. . . . .	135
6.30	Cluster 24: Two MDs correlate as shown in green lines. . . . .	135
6.31	Cluster 25: Two MDs correlate as shown in green lines. Cluster 25 connects with cluster 7 as shown in red line. . . . .	136
6.32	Cluster diagram for dependent clusters . . . . .	138
6.33	Comparison of accuracy by machine learning method on with and without selection of molecular descriptors by correlation analysis and cluster analysis. Plots on the rectangular line shows no difference by selection of molecular descriptors. . . . .	141
6.34	Boxplot of accuracy comparison of the molecular descriptors before selection(left) and after selection(right). . . . .	144
6.35	Boxplot of Execution Time comparison of the molecular descriptors before(left) and after selection(right). . . . .	145

7.1	Process chart of selecting the optimum MDs . . . . .	148
7.2	Selection of MD-MD pair from the dendrogram . . . . .	150
7.3	Process chart of selecting the optimum MDs . . . . .	151
7.4	Distribution of correlation coefficient of MD . . . . .	152
7.5	Dendrogram of 130 MDs according to the similarity of 194 pesticides	153
7.6	Dendrogram of 128 MDs . . . . .	154
7.7	Histogram of the result of optimum number of cluster according to the gap function . . . . .	155
7.8	Dendrogram of MDs at k=13 . . . . .	156
7.9	Accuracy and Log(ET) of 119 machine learning methods with MD4. Plot in red is ordinary method and blue is ensemble method.	158
7.10	Accuracy and Log(ET) of 119 machine learning methods with MD4 in expanded view. Plot in red is ordinary method and blue is ensemble method. . . . .	159
7.11	Accuracy comparison between MD4 and MD0. Plot in red is ordi- nary method and blue is ensemble method. . . . .	160
7.12	Accuracy comparison between MD4 and MD2. Plot in red is ordi- nary method and blue is ensemble method. . . . .	161
7.13	Dendrogram of MDs at k=3 . . . . .	162
7.14	Accuracy and Log(ET) for 119 methods using the MD6. Plot in red is ordinary method and blue is ensemble method. . . . .	163
7.15	Accuracy and Log(ET) for 119 methods using the MD6 (expanded view). Plot in red is ordinary method and blue is ensemble method.	164
7.16	Accuracy comparison between MD6 and MD0. Plot in red is ordi- nary method and blue is ensemble method. . . . .	165
7.17	Accuracy comparison between MD6 and MD2. Plot in red is ordi- nary method and blue is ensemble method. . . . .	166
7.18	Accuracy comparison between MD6 and MD4. Plot in red is ordi- nary method and blue is ensemble method. . . . .	167
7.19	Comparison of Accuracy by learning method category among MD6, MD4, MD2 and MD0 . . . . .	169
7.20	ET of best machine learning method of MD6, MD4, MD2 and MD0 with 5 replicates . . . . .	171

## List of Tables

2.1	Pesticide recovery rate summary . . . . .	7
2.2	Summary of molecular descriptors obtained by SMILES for chemicals	8
2.3	Regression methods in caret evaluated in this study . . . . .	11
2.4	Tuning parameters of xgbLinear for pesticide recovery prediction.	19
3.1	DP Clus parameters . . . . .	25
3.2	Group of MDs for optimum selection using DP Clus . . . . .	27
3.3	Molecular descriptors comparison between Nereistoxin and average of other 247 pesticides . . . . .	35
3.4	Molecular descriptors comparison between $\alpha$ -Endosulfan and av- erage of other 247 pesticides . . . . .	38
3.5	Molecular descriptors comparison among Propyzamide, Flumiox- azin and average of other 246 pesticides . . . . .	44
3.6	Candidates of MD2(29 MDs from 28 clusters) . . . . .	51
3.7	Combination of MDs with the $r \geq 0.7$ . . . . .	51
3.8	Final MDs selected by cluster analysis . . . . .	53
3.9	Summary of molecular descriptors selected by the correlation anal- ysis and cluster analysis . . . . .	53
3.10	Top 20 method for MD2 > MD0(Worse PE by selecting the molec- ular descriptors sorted by the Prediction Error difference) . . . . .	55
3.11	Top 20 method for MD2 < MD0(Better PE by selecting the molec- ular descriptors sorted by the Prediction Error difference) . . . . .	56
3.12	Top 20 method for MD2 < MD0 (Shorter ET in MD2 by selecting the molecular descriptors sorted by the Execution Time difference)	59
3.13	Top 20 method for MD2 > MD0 (Longer ET in MD2 by selecting the molecular descriptors sorted by the Execution Time difference)	60
3.14	Tuning parameters of xgbLinear (eXtreme Gradient Boosting Lin- ear) for pesticide recovery prediction (MD2). . . . .	60
4.1	Hierarchical Cluster Analysis parameters . . . . .	64
4.2	Group of MDs for optimum selection using hierarchical cluster tree	67
4.3	NbClust parameters . . . . .	70
4.4	Molecular Descriptor group in this chapter . . . . .	73



4.5	Molecular Descriptors selected by the correlation analysis and cluster analysis (k=14) . . . . .	74
4.6	Summary of molecular descriptors selected by the correlation analysis and cluster analysis (k=14) . . . . .	74
4.7	Molecular Descriptors selected by the correlation analysis and cluster analysis (k=14) . . . . .	76
4.8	Tuning parameters of xgbLinear for pesticide recovery prediction (MD4). . . . .	77
4.9	Molecular Descriptors selected by the correlation analysis and cluster analysis (k=3) . . . . .	80
4.10	Summary of molecular descriptors selected by the correlation analysis and cluster analysis (k=3) . . . . .	80
4.11	Tuning parameters of xgbLinear for pesticide recovery prediction (MD6). . . . .	81
4.12	Results of machine learning method category among the MD groups	86
5.1	Summary of molecular descriptors obtained by SMILES for chemicals	91
5.2	Classification methods in caret evaluated in this study . . . . .	95
5.3	Top 20 method for classification of pesticides sorted by accuracy .	101
5.4	Tuning parameters of xgbDART for pesticide classification prediction. . . . .	102
5.5	Chemical characteristics of wrongly classified 35 pesticides by xgbDART with 176 MDs . . . . .	103
5.6	Tuning parameters of 'rpart' for decision tree to classify 194 pesticides by 176 MDs . . . . .	103
5.7	XLogP of LC-MS pesticides (All 110 pesticides and 14 pesticides wrongly classified) . . . . .	108
5.8	XLogP of GC-MS pesticides (All 84 pesticides and 21 pesticides wrongly classified) . . . . .	109
6.1	DP Clus parameters . . . . .	112
6.2	Group of MDs for optimum selection . . . . .	114
6.3	Candidates of MD2(26 MDs from 25 clusters) . . . . .	137
6.4	Combination of MDs with the $r \geq 0.7$ . . . . .	137
6.5	Final MDs selected by cluster analysis . . . . .	139

6.6	Summary of molecular descriptors selected by the correlation analysis and cluster analysis . . . . .	139
6.7	Top 20 method for MD2 < MD0(Worse accuracy by selecting the molecular descriptors) . . . . .	142
6.8	Top 20 method for MD2 > MD0(Better accuracy by selecting the molecular descriptors) . . . . .	143
6.9	Tuning parameters of xgbDART for pesticide classification prediction. (MD2) . . . . .	146
7.1	Hierarchical Cluster Analysis parameters . . . . .	149
7.2	Group of MDs for optimum selection using hierarchical cluster tree	153
7.3	NbClust parameters . . . . .	154
7.4	Molecular Descriptor group in this chapter . . . . .	156
7.5	Molecular Descriptors selected by the correlation analysis and cluster analysis (k=13) . . . . .	157
7.6	Summary of molecular descriptors selected by the correlation analysis and cluster analysis (k=13) . . . . .	158
7.7	Tuning parameters of xgbDART for pesticide classification prediction. (MD4) . . . . .	162
7.8	Summary of molecular descriptors selected by the correlation analysis and cluster analysis (k=3) . . . . .	163
7.9	Tuning parameters of xgbDART for pesticide classification prediction.(MD6) . . . . .	168
7.10	Results of machine learning method accuracy category among the MD groups . . . . .	170
A.1	Pesticides and Canonical SMILES for regression analysis . . . . .	185
A.2	Average of log PE(Av log PE) and Average of log Execution Time(Av log ET). Regression methods(Method) are ordered according to Av log PE. . . . .	190
B.1	Summary of molecular descriptors obtained by SMILES for classification.Technology L is analyzed by LC-MS, G is GC-MS. List of E is EURL list, F is FDA list and Both is both EURL and FDA list.	191
B.2	Accuracy and log Execution Time (Av log ET). Classification methods (Method) are ordered according to the accuracy. . . . .	195

C.1	Specification of the PC and Software programs . . . . .	196
-----	---	-----

# 1. Introduction

## 1.1 Pesticides

The history of pesticides can be traced back to 1500 B.C. for expelling the fleas from the house and there are records showing that by 900 A.D. arsenic sulfides were used in China to control garden insects [1]. In Japan, a document entitled Family Traditions on the Killing of Insects was found in 1600s, that records that noxious insects can be exterminated using a mixture of five types of ingredients [2]. Until the mid 1930's, pesticides were mainly natural origin or inorganic compounds. In the late 19th century, various products began to be introduced such as lime sulphur, Bordeaux mixture and other copper agents as well as pyrethrum insect powder, nicotine and other natural products [1]. In spite that pesticides have contributed to economic growth of Japan by enhancing the crop production, incidents related to the pesticides(toxicity and abuse of use) occurred, the needs of safety of pesticides manifested among the society in Japan [2]. In 1971, the Agricultural Chemicals Regulation Act was amended in Japan with the additional requirement of "protecting human health and conservation of living environment" as considerations to ensure "safety of use". Long-term toxicity testing and environmental impact evaluation(residual in the crops) were also required for registering pesticides. Requirements for testing methods of pesticides before commercialization became stricter with the introduction of good laboratory practice for pesticides [2].

## 1.2 Pesticide Analysis for food safety

Today, residual pesticides in foods are routinely analyzed by Mass Spectrometer(MS) combined with chromatography system, such as liquid chromatograph(LC) and gas chromatography(GC), commonly expressed in LC-MS and GC-MS. In 1980s, GC-MS performance evolved drastically along with the evolution of semiconductor for data processing, because huge amount of commands for operating the Mass Spectrometer are required. GC-MS was widely adopted for the analysis of pesticides in the river water in response to the social concerns of the pollutant of pesticides near golf course in 1989 by the environmental analysis laboratories

[2]. GC-MS application also expanded to the other compounds of environmental analysis in addition to the pesticides by the regulation. GC-MS is also used for residual pesticides in foods, but the application is limited to small number of pesticides at that time because the pesticide regulation was “Negative List”, i.e. there was no need to analyze the other pesticides because the usage of pesticide is rigidly regulated at that time among the laboratories of food analysis (Food Labs). Developments in food safety and quality continued throughout the 20th century with by domestic food regulations for each country. However, the globalized economy changed the situation of the food safety, and each country’s individual regulation was often conflicting and contradictory at global trades. The Codex Alimentarius Commission was found in 1963 by FAO and WHO to develop food standards, guidelines and related documents which are published in the Codex Alimentarius [3]. In 2000s’ evolutions of internet and global logistics accelerated the global food trade dramatically. Residual pesticides in the imported foods were global concerns. In response to this issue, Codex took the initiatives of global food regulation, changing the regulations of residual pesticides in food from “Negative list” to “Positive list”, and set the minimum residual level(MRL) at 10ppb for most pesticides with several exceptions. There are guidelines for ensuring the data quality of pesticide residues in food to describe the method validation and analytical quality control requirements to support the validity of data used for checking compliance with MRLs, enforcement actions, or assessment of consumer exposure to pesticides.[4, 5]. In Japan, Ministry of Human Warfare and Labors(MHLW) released the new regulation of residual pesticides in foods at MRL 10ppb, changed from negative list to Japan Positive List(JPL) as the harmonization to the global movement. MHLW also released simultaneous analysis of multi class pesticides method called “JPL method” for LC-MS and GC-MS in 2006. As of May 2020, the number of pesticides in the JPL method using LC-MS and GC-MS is over 700 [6]. There are challenges in simultaneous analysis of multi class pesticides technically by GC-MS and LC-MS because of the variety of chemical natures of pesticides and residual compounds in the sample matrix of crops(other compounds in the sample than pesticides) after sample preparation of JPL method. The sample preparation of JPL method tries to remove the matrices while minimizing the loss of residual pesticides, but removal of sample

matrix is not perfect. There are chemical interactions between the pesticides and sample matrices called "Matrix Effect" [7, 8], that alters the recovery rate of pesticides in detection by GC-MS and LC-MS. In additions, pesticides also interact with the active surface sites of flow path in GC and LC system chemically, that also alter the recovery rate of pesticides. The degree of Matrix Effect depends on the chemical nature of pesticides, such as presence of polar chemical functional group(s), molecular shapes and so on. The other issues of pesticide analysis is the limitation of LC-MS ionization. Mass Spectrometer requires the pesticide to be "Ionized", but ElectroSpray Ionization(ESI) is not the perfect ionization technology for whole pesticides [9]. Additionally, there is the technical limitation of separation by liquid chromatography. "Reversed Phase" which can separate wider range of compounds with the development of LC column by the bonded stationary phases [10] is used for most LC-MS analysis, but reversed phase separation is challenging in separating the non-polar pesticides by LC. Thus, neither GC-MS nor LC-MS cannot analyze whole pesticides by single technology. The classification of pesticide amenability between LC or GC are frequent concerns among many food analysis chemists in the world. The initial cost and analytical skill especially sample preparation are still the barrier for Food Labs to switch from GC-MS to LC-MS or additional introduction of new instruments to the laboratory. Conventional instrument initial cost of GC-MS is around 100k USD and LC-MS is around 250k USD, Food Lab managers need to consider the needs of food analysis, available instruments and the cost for managing the technical difficulties. Chemists in Food Labs want to manage the pesticide analysis with the currently using instruments. Thus, prediction of residual pesticide response is still demanding for food analysis laboratories to manage the analytical data quality and method development in advance for unknown pesticides [11, 12].

### 1.3 Molecular descriptors

In this study, I used the approach of quantitative structure-property relationship(QSPR) for prediction of pesticides in foods using the validated report of residual pesticides in foods by GC-MS and LC-MS. QSPR consists of two steps, obtaining the molecular descriptor and establishing the relationship between molecular descriptors and properties. The first step is to calculate the chemi-

cal property of pesticides in a computer-friendly format, i.e. numerical format called “molecular descriptors”, and second step is to establish the relationship between molecular descriptors and the pesticide property(e.g. recovery ratio of pesticides in Chapter 2 and 3, classification of pesticides in Chapter 4 and 5) [13]. Various molecular descriptors(MDs) have been developed such as molecular weight, Water-Octanol coefficient(LogP) [14], number of hydrogen bond acceptors/donors [15], various topological descriptors based on the shape of molecule based on graph theory such as BCUT descriptors [16]. The rcdk package of R program can retrieve the MDs from the SMILES strings [17]. SMILES is the abbreviation of Simplified Molecular Input Line Entry System, which was created by David Weininger in 1986. SMILES is the descriptor of molecule in 2D format, representing with the atoms and chemicals bonds, without notation of hydrogens. [18] SMILES of pesticides can be obtained from the PubChem website or can be generated by Online tools from the chemical structure. 224 MDs were obtained by the rcdk package, including the constitutional, topological, geometric and electronic descriptors which the characteristics of each pesticide in unique numeric value.

## 1.4 Machine Learning methods

In order to create the prediction model using the molecular descriptors(MDs), machine learning methods available in the caret package of R program are used for classification and regression of pesticides [19]. In this study, MDs are used as the explanatory variables of prediction model. Before building the prediction model by caret, MDs need to be cleaned up because some MDs retrieved by the rcdk are meaningless value such as no variance across the pesticides which can cause the issue in deploying the machine learning. Removal such MDs is essential step for successful prediction by machine learning. In this dissertation, comparisons of removal of MDs are discussed for 89 regression analysis methods of pesticide recovery rate in Chapter 2 and 3, and 119 classification methods of amenability of LC-MS or GC-MS in Chapter 4 and 5.

## 1.5 Dissertation outline

The outline of this dissertation is described in Figure 1.1. Two main studies for the pesticide data set acquired by GC-MS and LC-MS are included on this dissertation, i.e. regression analysis and classification analysis. In Chapter 2, I evaluated the machine learning methods of regression analysis of pesticide recovery for seven crops. In Chapter 3 and Chapter 4, I developed the procedure to optimize the selection of MDs for pesticides recovery rate prediction using the correlation analysis and clustering tools. I developed the prediction method to classify the pesticides amenability between LC- and GC- using the MDs in Chapter 5. In Chapter 6 and Chapter 7, the similar procedure of selecting the optimum MDs developed in Chapter 3 and Chapter 4 is applied on the classification model of Chapter 5, and evaluate the performance of classification by selecting the MDs.

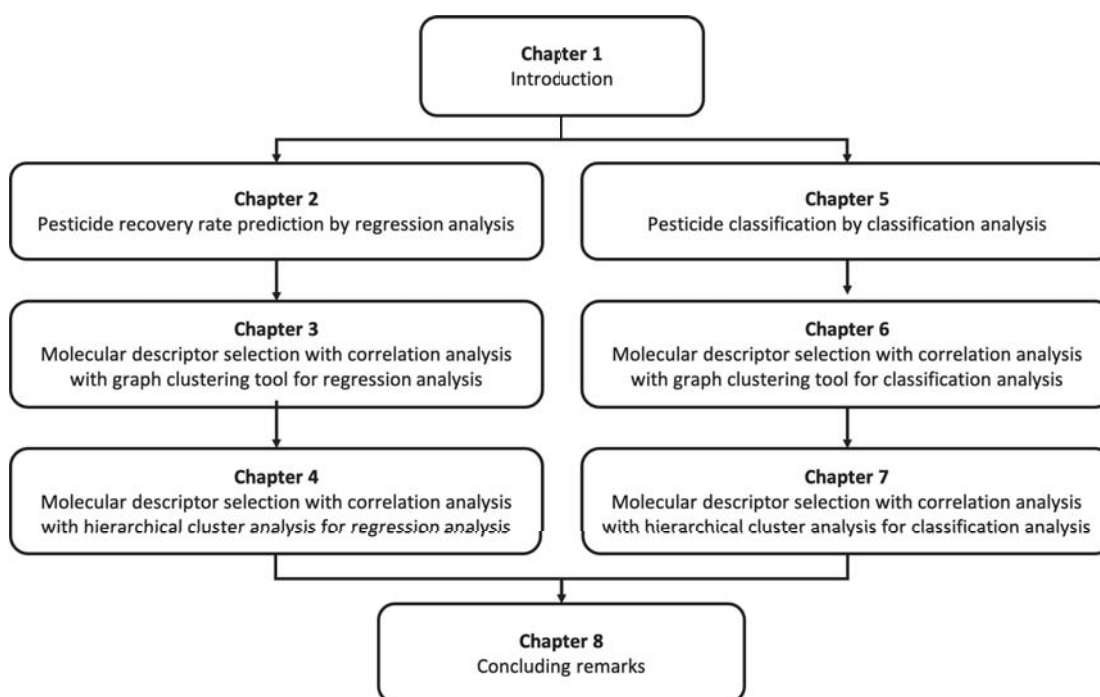


Figure 1.1. Outline of this dissertation



## 2. Pesticide recovery rate prediction for seven crops

### 2.1 Introduction

In this chapter, I develop the procedure for building the data set of pesticide recovery rate prediction by the regression analysis of machine learning using the data of pesticide analysis study on the recovery rate of seven crops with the sample preparation procedure of multi residue method of pesticides for Japan positive List, which was released by Ministry of Health, Labour and Welfare(MHLW), Japan [6].

### 2.2 Materials and Methods

#### 2.2.1 List of pesticides with recovery rate

The data sets of 305 pesticide recovery ratio (50 ppb standard addition and recovery ratio was calculated with the solvent standard calibration curve of 20 ppb, 50 ppb, 100 ppb and 200 ppb) of seven crops (spinach, brown rice, apple, orange, soybean, cabbage and potato) were gathered from the validation report of JPL [20]. The data were acquired by the procedure according to the Japan Positive List(JPL) of MHLW [6] as shown in the Figure 2.1. The sample preparation process includes the extraction and clean-up processes. As the grains, beans, nuts and seeds contain huge amount of fatty acids, additional clean-up steps for removal were applied. In the present study, brown rice and soybean were applied to the extraction procedure of “grains, beans, nuts and seeds”, and spinach, apple, orange and potato were applied to that of “fruits, vegetables, herbs, tea and hops” of JPL method. 50 ppb of 305 pesticides were spiked to the sample and then analyzed by GC-MS in SIM/Scan mode. The recovery rate of pesticides was calculated by a ratio of peak area of 50 ppb spiked in the sample to that in the solvent standard calibration curve, calculated by the Equation 2.1.

$$\text{Recovery rate} = \frac{\text{Concentration calculated by calibration curve in ppb}}{\text{Pesticide amount injected on sample (50 ppb)}} \times 100(\%) \quad (2.1)$$

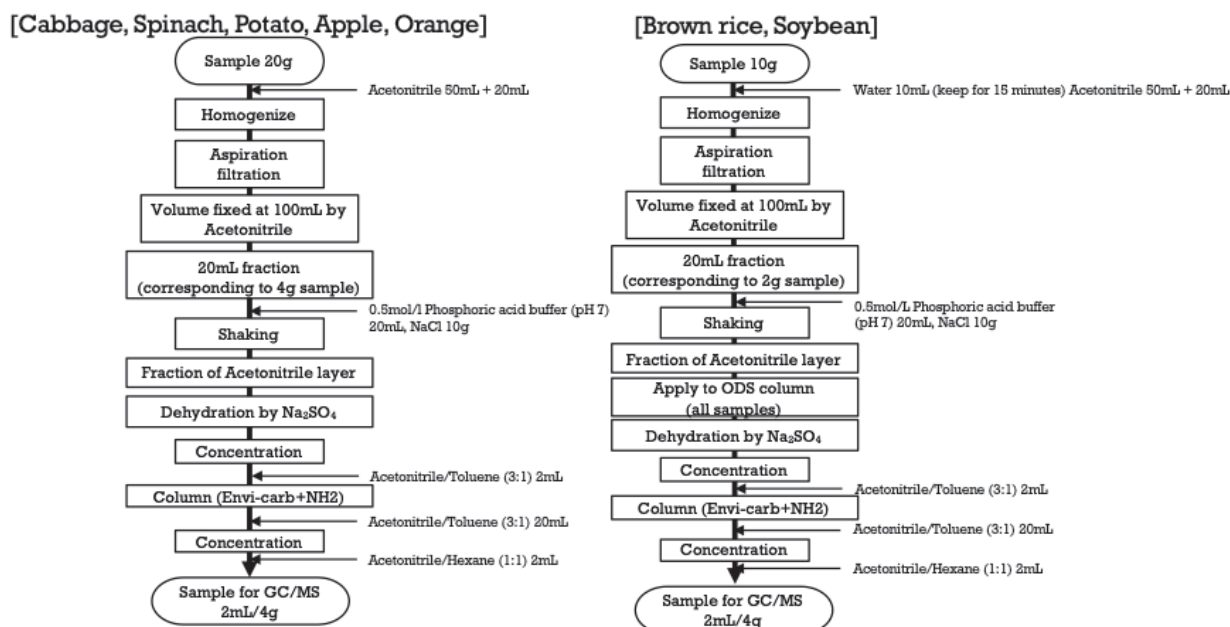


Figure 2.1. Sample preparation procedure of Japan Positive List

The summary of pesticide recovery rate for each crop is summarized in the table 2.1.

Table 2.1. Pesticide recovery rate summary

Summary	Spinach	Apple	Brown Rice	Orange	SoyBean	Cabbage	Potato
Max.	201	532	204	188	232	254	187
3rd Qu.	88	102	101	103	96	106	97
Mean	81	92	83	89	78	93	85
Median	84	96	92	98	87	99	91
1st Qu.	79	87	79	89	67	91	82
Min.	0	0	0	-595	0	0	0

### 2.2.2 Molecular descriptors of the pesticides

For better prediction of recovery rate using the molecular descriptors obtained by the canonical SMILES strings, 248 unique pesticides are selected for machine

learning by removing the pesticides that have geometric isomer(s) which have different recovery ratio respectively while the same canonical SMILES strings. Canonical SMILES strings were obtained from the PubChem website and added to the data set. The chemical structure of canonical SMILES strings are converted to the 178 molecular descriptors (Table 2.2) for each of the 248 pesticides by rcdk package. Table A.1 shows the name and canonical SMILES of 248 pesticides. Each molecular descriptor was standardized for comparable as expressed by the Equation 2.2 (z-score), where  $z_i$  is the standardized value to be used for machine learning,  $x_i$  is the raw value from rcdk,  $\mu_i$  is the average of 248 pesticides and  $\sigma_i$  is the standard deviation of 248 pesticides for  $i$ th molecular descriptor.

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \quad (2.2)$$

Table 2.2: Summary of molecular descriptors obtained by SMILES for chemicals

Descriptor Class	Descriptor (Description)
ALOGP Descriptor (2)	ALogP (Ghose-Crippen LogKow), ALogP2 (Square of ALogP)
APol Descriptor (1)	APol (Sum of the atomic polarizabilities (including implicit hydrogens))
Aromatic Atoms Count Descriptor (1)	naAromAtom (Number of aromatic atoms)
Aromatic Bonds Count Descriptor (1)	nAromBond (Number of aromatic bonds)
Atom Count Descriptor (2)	nAtom (Number of atoms), nB (Number of binding)
Autocorrelation Descriptor Charge (5)	ATSc1, ATSc2, ATSc3, ATSc4, ATSc5 (ATS autocorrelation descriptor, weighted by charges)
Autocorrelation Descriptor Mass (5)	ATSm1, ATSm2, ATSm3, ATSm4, ATSm5 (ATS autocorrelation descriptor, weighted by scaled atomic mass)
Autocorrelation Descriptor Polarizability (5)	ATSp1, ATSp2, ATSp3, ATSp4, ATSp5 (ATS autocorrelation descriptor, weighted by polarizability)
BCUT Descriptor (6)	BCUTw.1l (nhigh lowest atom weighted BCUTS), BCUTw.1h (nlow highest atom), BCUTc.1l (nhigh lowest partial charge), BCUTc.1h (nlow highest partial charge) BCUTp.1l (nhigh lowest polarizability), BCUTp.1h (nlow highest polarizability)
BPolDescriptor (1)	bpol (Sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule (including implicit hydrogens))
Carbon Types Descriptor (9)	C1SP1 (Triply bound carbon bound to one other carbon), C2SP1 (Triply bound carbon bound to two other carbons), C1SP2 (Doubly bound carbon bound to one other carbon), C2SP2 (Doubly bound carbon bound to two other carbons), C3SP2 (Doubly bound carbon bound to three other carbons), C1SP3 (Singly bound carbon bound to one other carbon), C2SP3 (Singly bound carbon bound to two other carbons), C3SP3 (Singly bound carbon bound to three other carbons), C4SP3 (Singly bound carbon bound to four other carbons)
Chi Chain Descriptor (10)	SCH.3-7 (Simple chain, orders 3-7), VCH.3-7 (Valence chain, orders 3-7)
Chi Cluster Descriptor (8)	SC.3-6 (Simple cluster, orders 3-6), VC.3-6 (Valence cluster, orders 3-6)
Chi Path Cluster Descriptor (6)	SPC.4-6 (Simple path cluster, orders 4 to 6), VPC.4-6 (Valence path cluster, orders 4-6)
Chi Path Descriptor (16)	SP.0-7 (Simple path, orders 0-7), VP.0-7Valence path, orders 0-7)

*Continue to next page*

*Continue from previous page*

Descriptor Class	Descriptor (Description)
Eccentric Connectivity Index Descriptor (38)	ECCEN (A topological descriptor combining distance and adjacency information), khs.sCH3 (Count of atom-type E-State: -CH3), khs.dCH2 (=CH2), khs.ssCH2 (-CH2-), khs.tCH (#CH), khs.dsCH (=CH-), khs.aaCH (:CH:), khs.sssCH (>CH-), khs.tsC (#C-), khs.dssC (=C<), khs.aasC (:C:-), khs.aaaC (::C:), khs.ssssC (>C<), khs.sNH2 (-NH2), khs.ssNH (-NH2-+), khs.aaNH (:NH:), khs.tN (#N), khs.sssNH (>NH-+), khs.dsN (=N-), khs.aaN (:N:), khs.sssN (>N-), khs.ddsN (-N<<), khs.aasN (:N:-), khs.sOH (-OH), khs.dO (=O), khs.ssO (-O-), khs.aaO (:O:), khs.sF (-F), khs.ssssSi (>Si<), khs.dsssP (->P=), khs.dS (=S), khs.sss (-S-), khs.aas (aSa), khs.dssS (>S=), khs.ddssS (>S==), khs.sCl (-Cl), khs.sBr (-Br)
Fragment Complexity Descriptor (1)	fragC (Complexity of a system)
Ghose Crippen Molecular Refractivity Descriptor (1)	AMR (Molar refractivity)
H Bond Acceptor Count Descriptor (1)	nHBacc (Number of hydrogen bond acceptors)
H Bond Donor Count Descriptor (1)	nHBDon (Number of hydrogen bond donors)
KappaShape Indices Descriptor (3)	Kier1-3 (First, Second, Third kappa ( $\kappa$ ) shape indexes)
Largest Chain Descriptor (1)	nAtomLC (Number of atoms in the largest chain)
Longest Aliphatic Chain Descriptor (1)	nAtomLAC (Number of atoms in the longest aliphatic chain)
Mannhold LogP Descriptor (1)	MLogP (Mannhold LogP)
MDEDescriptor (19)	MDEC.11 (Molecular distance edge between all primary carbons), MDEC.12 (between all primary and secondary carbons), MDEC.13 (between all primary and tertiary carbons), MDEC.14 (between all primary and quaternary carbons), MDEC.22 (between all secondary carbons), MDEC.23 (between all secondary and tertiary carbons), MDEC.24 (between all secondary and quaternary carbons), MDEC.33 (between all tertiary carbons), MDEC.34 (between all tertiary and quaternary carbons), MDEC.44 (between all quaternary carbons), MDEO.11 (between all primary oxygens), MDEO.12 (between all primary and secondary oxygens), MDEO.22 (between all secondary oxygens), MDEN.11 (between all primary nitrogens), MDEN.12 (between all primary and secondary nitrogens), MDEN.13 (between all primary and tertiary nitrogens), MDEN.22 (between all secondary nitrogens), MDEN.23 (between all secondary and tertiary nitrogens), MDEN.33 (between all tertiary nitrogens)
Petitjean Number Descriptor (1)	PetitjeanNumber (Petitjean number)
Rotatable Bonds Count Descriptor (1)	nRotB (Number of rotatable bonds, excluding terminal bonds)
Rule Of Five Descriptor (1)	LipinskiFailures (Number failures of the Lipinski's Rule Of 5)
TPSA Descriptor (19)	TopoPSA (Topological polar surface area)
VAdjMat Descriptor (1)	VAdjMat (Vertex adjacency information (magnitude))
Weight Descriptor (1)	MW (Molecular weight)
Weighted Path Descriptor (5)	WTPT.1 (Molecular ID), WTPT.2 (Molecular ID / number of atoms), WTPT.3 (Sum of path lengths starting from heteroatoms), WTPT.4 (Sum of path lengths starting from oxygens), WTPT.5 (Sum of path lengths starting from nitrogens)
Wiener Numbers Descriptor (2)	WPATH (Weiner path number), WPOL (Weiner polarity number)
XLogP Descriptor (1)	XLogP (XLogP)
Zagreb Index Descriptor (1)	Zagreb (Sum of the squares of atom degree over all heavy atoms i)
Petitjean Shape Index Descriptor (1)	topoShape (Petitjean topological shape index)

*Continue to next page*

*Continue from previous page*

Descriptor Class	Descriptor (Description)
Others (17)	nAcid (Acidic group count descriptor), nBase (Basic group count descriptor), nSmallRings (the number of small rings from size 3 to 9), nAromRings (the number of aromatic rings), nRingBlocks (total number of distinct ring blocks), nAromBlocks (total number of "aromatically connected components"), nRings3, 5, 6, 7 (individual breakdown of small rings), tpsaEfficiency (Polar surface area expressed as a ratio to molecular size), VABC (Atomic and Bond Contributions of van der Waals volume), HybRatio (the ratio of heavy atoms in the framework to the total number of heavy atoms in the molecule.), tpsaEfficiency.1 (Polar surface area expressed as a ratio to molecular size), TopoPSA.1 (Topological polar surface area), topoShape.1 (A measure of the anisotropy in a molecule), FMF (FMF descriptor characterizing complexity of a molecule)

*End of table*

### 2.2.3 Regression analysis of pesticides recovery rate prediction by machine learning

Regression algorithms can be classified into two categories, i.e. (1) ordinary learning approaches which construct one learner from training data and (2) ensemble methods which construct a set of learners and combine them. caret package for machine learning in R program is introduced as publicly accessible learning resources and tools related to machine learning [21]. In the present study, we examined regression models of 69 ordinary and 20 ensemble learning methods in caret (Table 2.3). Most of the ordinary learning methods correspond to regression models with kernel and simple linear models. In regression models with kernels, gaussprLinear (Gaussian Process), rvmLinear (Relevance Vector Machines with Linear Kernel), svmLinear (Support Vector Machines with Linear Kernel by using kernlab package with the kernel principal component analysis), svmLinear2 (Support Vector Machines with Linear Kernel by e1071 package) and svmLinear3 (L2 Regularized Support Vector Machine (dual) with Linear Kernel) implement linear kernel; gaussprRadial (Gaussian Process with Radial Basis Function Kernel), krlsRadial (Radial Basis Function Kernel Regularized Least Squares), svmRadial (Support Vector Machines with Class Weights), svmRadialSigma (Support Vector Machines with Radial Basis Function Kernel), rvmRadial (Relevance Vector Machines with Radial Basis Function Kernel), svmRadialCost (Support Vector Machines with Radial Basis Function Kernel) implement radial basis function kernels; and gaussprPoly (Gaussian Process with Polynomial Kernel), krlsPoly (Polynomial Kernel Regularized Least Squares), rvmPoly (Relevance Vector Machines with Polynomial Kernel), and svmPoly (Support Vector Machines with

Polynomial Kernel) implement polynomial kernels. Simple linear models have been developed on the basis of multilinear regression models (lm, leapSeq, leapForward, leapBackward, lmStepAIC), generalized linear models (glm, bayesglm, glmStepAIC, plsRglm), principal component regressions (icr, pcr, superpc) and partial least square models (nnls, simples, pls). Subsequently, diverged sparse models have been developed; nine methods lasso (The lasso), blassoAveraged (Bayesian Ridge Regression(Model Averaged)), blasso (Bayesian lasso), penalized (Penalized Linear Regression), relaxo (Relaxed Lasso), lars (Least Angle Regression with the parameter of fraction), lars2 (Least Angle Regression with the parameter of steps), glmnet, and enet (Elasticnet) are a type of Least Absolute Shrinkage and Selection Operation (LASSO, [22]); two methods foba (Ridge Regression with Variable Selection) and ridge (Ridge Regression) are a type of ridge regression, [23]. In addition, methods belonging to neural network, decision tree and y-value prediction based on neighboring samples in X variable space [24, 25], regression models using spline function and the others were implemented in caret package (Table 2.3). In contrast, 14 regression models using decision tree, i.e., random forest, methods using simple linear regression models, and models using spline were available in caret package. Thus, we could implement 89 methods and compared them in terms of performance of precision and execution time.

Table 2.3. Regression methods in caret evaluated in this study

Algorithm	Methods in caret
(a) Ordinary learning methods	
Kernel (17)	gaussprRadial, gaussprPoly, krlsPoly, gaussprLinear, krlsRadial, rvmLinear, rvmRadial, rvmPoly, svmRadial, svmRadialCost, svmRadialSigma, svmLinear, svmLinear2, svmPoly, svmLinear3, kernelpls(PLS), widekernelpls(PLS)
Simple Linear (16)	lm, leapSeq, leapForward, leapBackward, lmStepAIC, bridge, bayesglm(GLM), glmStepAIC(GLM), icr(ICA), pcr(PCA), superpc(PCA), superpc(PCA), nnls(PLS), simpls(PLS), pls(PLS), plsRglm(PLS, GLM), glm(GLM)
Sparse modeling (11)	penalized, blassoAveraged, foba, ridge, relaxo, lasso, Blasso, lars, lars2, glmnet, enet
Neural Network (9)	rbfDDA, dnn, neuralnet, brnn, mlpML, mlp, mlpWeightDecay, msaenet, monmlp
Decision Tree (8)	rpart2, rpart1SE, ctrees, ctrees2, evtree, M5Rules, M5, WM
Centroid,kNN (3)	knn, kknn, SBC
Spline (2)	gcvEarth, earth
Others (3)	ppr, spikeslab, xyf(LVQ)
(b) Ensemble learning methods	
Decision Tree (15)	cforest, ranger, qrf, rf, parRF, extraTrees, Rborist, RRFglobal, RRF, treebag, bstTree, gbm, xgbTree, nodeHarvest, xgbDART
Simple Linear (3)	BstLm, glmboost(GLM), xgbLinear
Spline (2)	bagEarthGCV, bagEarth

## 2.3 Results and Discussion

### 2.3.1 Pesticide recovery rate distribution

Understanding the data structure is important step before building the machine learning model. Before developing regression models for pesticide recovery prediction, we examined the correlations between recovery rates of 248 pesticides and each 178 molecular descriptor for seven crops [20] using the Pearson correlation coefficients [26]. Distributions of Pearson correlations between recovery rates for seven crops and all 178 molecular descriptors (Figure 2.2, left) indicate that all correlations are weak, i.e. Pearson correlation coefficients were in the range of -0.254 to 0.523. Thus it is difficult to construct regression models by using any single chemical descriptor. So, we performed multivariate regression models using the caret package for predicting the recovery rate by the molecular descriptors. Initially, regression models were built with the 10-fold cross validation(CV-10) to evaluate the performance of prediction [27]. Figure 2.2 right diagram shows distribution of Pearson correlations between experimental and predicted recovery rates in 89 regression models. Recovery rates for five crops(potato, brown rice, spinach, cabbage and soybean) are correlated with each other, but correlations of apple and orange to those five are very low, but obviously, we could obtain regression models with very high correlations for all seven crops though a few methods failed to make regression models to estimate recovery rates from the descriptors. Figure 2.3 shows heat map with the dendrogram of hierarchical cluster analysis [28] of recovery rates among crops. Figure 2.2 shows distribution of Pearson correlations between recovery rates and molecular descriptors(left) where X-axis represents Pearson correlation coefficients; while Y-axis represents the count of descriptors in the bin of histogram, and distribution of Pearson correlations between experimental and predicted recovery rates in 89 regression models(right) where X-axis represents Pearson correlation coefficients; while Y-axis represents the count of regression models.

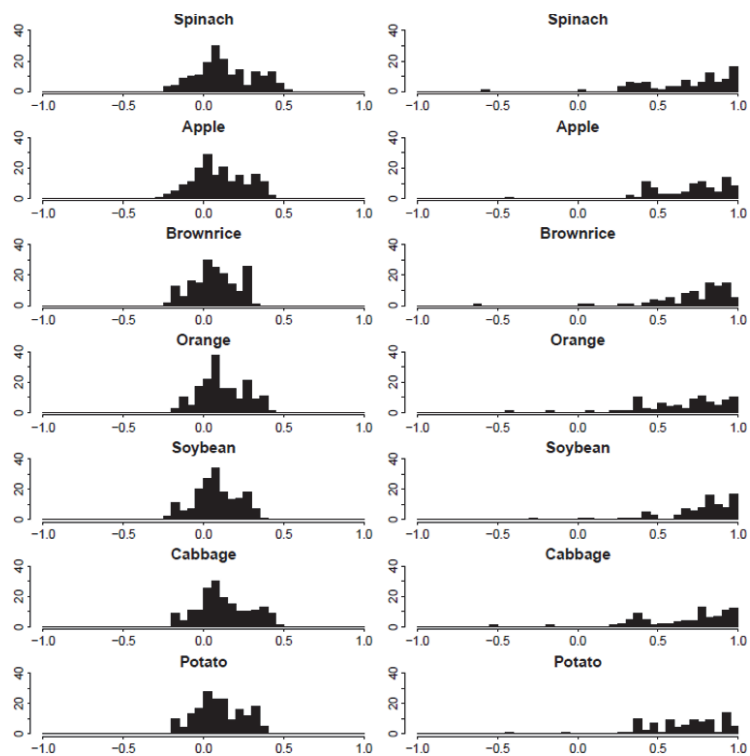


Figure 2.2. Distribution of Pearson correlations between recovery rates and molecular descriptors(left) where X-axis represents Pearson correlations; while Y-axis represents the count of descriptors, and Distribution of Pearson correlations between experimental and predicted recovery rates in 89 regression models(right) where X-axis represents Pearson correlations; while Y-axis represents the count of regression models.



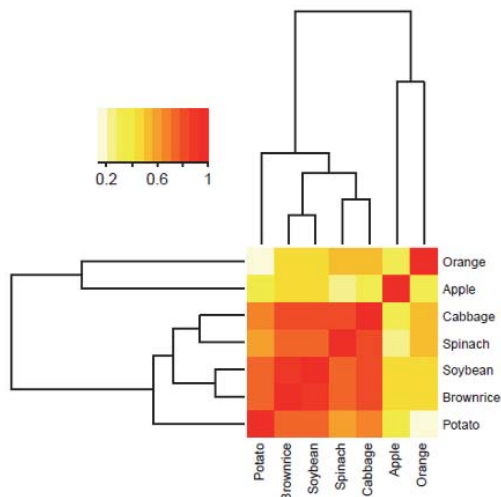


Figure 2.3. Dendrogram and heat map of hierarchical cluster analysis of pesticide recovery for seven crops

### 2.3.2 Machine Learning results for execution time and prediction error

Though Pearson correlation coefficients between observed and predicted recovery rates are indexes to represent validation of regression model, it is not enough to examine prediction performance. Execution time of individual algorithms should also be considered in practical prediction of recovery rates from chemical structures because some machine learning methods need several hours to complete building the prediction model. To estimate the model prediction performance for each regression model, we use prediction model error (PE) represented in Equation 2.3 and the Execution Time (ET) measured by the `system.time()` function of R program for building the machine learning method. The specifications of the personal computer used in this study are shown in the Table C.1 in the appendix section.

$$PE_j = \frac{\sum_{i=1}^N (y_{obs}^{(ij)} - y_{pred}^{(ij)})^2}{\sum_{i=1}^N (y_{obs}^{(ij)} - \bar{y})^2} \quad (2.3)$$

Here,  $\hat{y}_{ij}$  and  $\hat{y}_j$  represent the actual and predicted recovery ratio of  $i$ th pesticide in each crop, respectively, and  $\bar{y}_j$  represents the average recovery ratio in  $j$ th crop ( $j = 1, 2, \dots, 7$ ). Execution time (sec.) for constructing individual regression models was measured by “system.time( )” functions of R program.

Figure 2.4 shows the distribution for averages of log (PE) and log (ET) in seven crops. A few regression models were very large prediction errors except for two crops cabbage and soybeans, which correspond to the maximum values in log(PE), but most of regression models ranged approximately between -4 (0.0001 in PE) and 0 (1 in PE). Regression models with the highest PE are as follows: SBC(Subtractive Clustering and Fuzzy c-Means Rules) for 3 crops [brown rice(-3.72), cabbage(-3.13), orange(-3.13)], monmlp(Monotone Multi-Layer Perceptron Neural Network) for 2 crops [brown rice(-3.10), soybean(-3.31)], and xgbLinear(eXtreme Gradient Boosting) for 6 crops [brown rice(-3.97), spinach(-3.10), potato(-3.22), cabbage(-3.31), orange(-3.22)]. In contrast, those with the worst are rbfDDA(Radial Basis Function Network with the dynamic decay adjustment) for 3 crops [(spinach(0.07), cabbage(0.07), brownrice(0.12)), lasso for 4 crops [brown rice(9.07), spinach(12.5), potato(16.4), orange(24.6)], lars for 2 crops [soybean(5.81), brownrice(9.71)], and bagEarth (Bagged MARS) for orange(6.24). Those indicate that three regression methods, i.e., SBC, monmlp and xgbLinear are optimal for evaluation of the recovery rates for crops. In contrast, lasso and rbfDDA did not perform well for the present data. The former and latter belong to sparse and neural network modeling, respectively. In case of execution times for constructing regression models, the highest performances are obtained in knn(k-Nearest Neighbors) [brown rice(-0.05 sec in logarithm of execution time), Soybeans(-0.03), Potato(-0.02)] and in nnls(Non-Negative Least Squares) [brown rice(-0.27), Soybean(-0.12), Spinach(-0.05), potato(-0.08), Cabbage(-0.09), Apple(-0.08) and Orange(-0.08)]. In contrast, three methods krlsPoly, bridge and blassoAveraged needed long execution times, i.e., krlsPoly for seven crops [brown rice(3.85), soybean(3.84), spinach(3.81), potato(3.83), cabbage(3.89), apple(3.88), Orange(3.86)], bridge for soybean(3.76), blassoAveraged for cabbage(3.76), and blassoAveraged for apple(3.76). Consequently, as shown in Figure 2-4, we can observe intrinsic properties in log(PE) and log(ET) for seven crops.

Top 20 machine learning methods in seven crops average of  $\log(\text{PE})$  in Figure 2.5 indicate that xgbLinear and SBC carried out the highest prediction performance for all crops, in contrast, monmlp, ppr, extraTrees, and xgbDART also show the highest prediction performance for several crops. ET of the top 20 ranged from 1.24 sec. to 46 min.

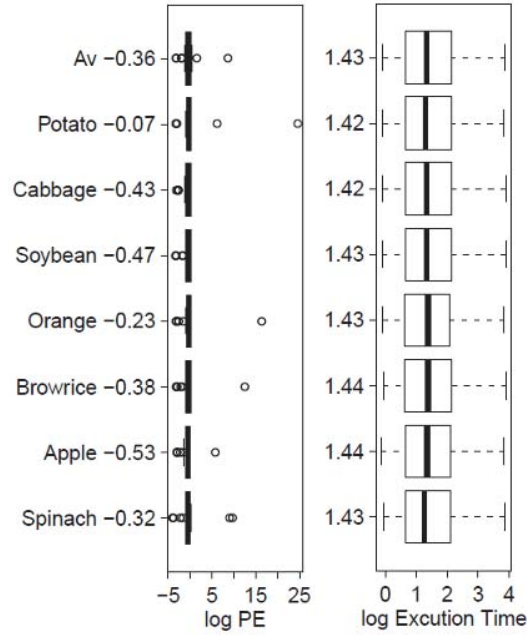


Figure 2.4. Distribution of log PE and log Execution Time for seven crops. Averages of log PE and log Execution Time corresponding to all regression models are shown along the crop names. “Av” in row represents averages of log PE- and Log Execution Time-averages for seven crops.

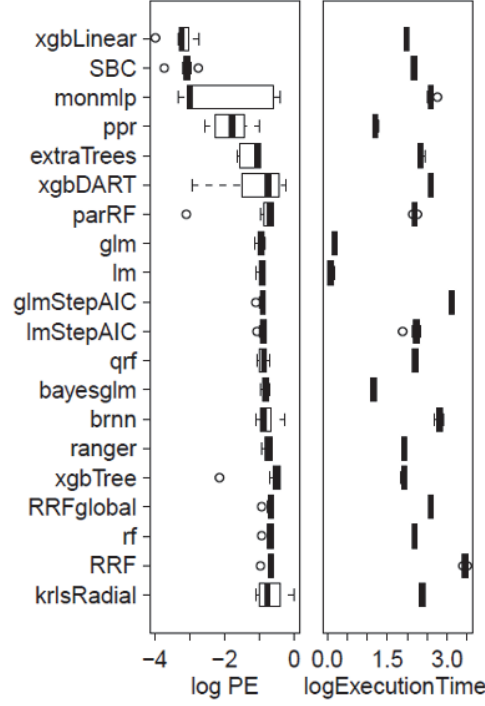


Figure 2.5. Averages of log PE for seven crops corresponding to a number of regression methods

### 2.3.3 Generalization performance for prediction

Generalized performance usually refers to the ability of regression methods to be effective across a range of input data and regression methods may always have strengths and weaknesses. It is generally considered that a single model is unlikely to fit every possible data and ensemble regression methods trend to achieve higher generalized performance than any of the ordinary models. Ensemble regression makes it possible to obtain better results when there is considerable diversity among the base classifiers, regardless of the measure of diversity [29]. In the present analysis, the generalized performance in different crops is the one of the topics for selecting regression methods. Average  $PE_k$  for seven crops ( $j=1, 2, \dots, M$ , and  $M=7$ ) in  $k$ th regression method represented in Eq. 2.4 is one of the indexes for generalized performance taking different crops into consideration.

$$PE_k = \frac{\sum_{j=1}^M PE_j^{(k)}}{M} \quad (2.4)$$

In addition to  $PE_k$ , average rank of  $k$ th method for  $PE_j^{(k)}, (k = 1, 2, \dots, 89)$  for seven crops becomes an index of generalized performance for different crops. Thus we compared between  $PE_k$  and the average rank for individual methods and obtained clear linear relationships between them (Figure 2.6). Top four regression methods, xgbLinear, SBC, ppr (Projection Pursuit Regression) and extraTrees (Random Forest by Randomization) are with the lowest PE with the best ranks. To compare trends of regression methods, we classified 89 regression models to 11 categories in Table 1. Figure 2.7 represents box plots for two performance rates, i.e.,  $\log(PE)$  and average of  $\log(ET)$  for 11 categories of machine learning methods. It can be observed that in terms of the median values of  $\log PE$ , the following three types of methods performed better: (i) ensemble decision trees, (ii) ordinal kernel type of regressions and (iii) ordinal centroid types of regressions using distance function for individual samples. In addition to those general trends, it should be worthy of notice that methods belonging to ensemble simple linear and ordinal centroid type also have very high performance in the context of generalized prediction error represented by Eq. 2.4.

On the execution time, ordinary neural network regressions and ensemble-type of regressions using decision trees and splines trend to need relatively long execution time but this trend are opposite in ensemble-type simple linear models. The performance rates in prediction errors and execution times for 89 methods (Table A.2) are visualized in Figure 2.8. In the present data, SBC and xgbLinear have the highest performance in prediction errors, ppr and monmlp show the second level performance; whereas gaussprLinear and widely used lm (Linear Regression) are good methods in terms of execution time.

Tuning parameters of xgbLinear are listed in the Table 2.4. The parameters were the same for all seven crops.

Table 2.4. Tuning parameters of xgbLinear for pesticide recovery prediction.

Parameter	Value
Number of rounds (Boosting Iterations)	50
Lambda (L2 Regularization)	0.0001
Alpha (L1 Regularization)	0.0001
Eta (Learning Rate)	0.3

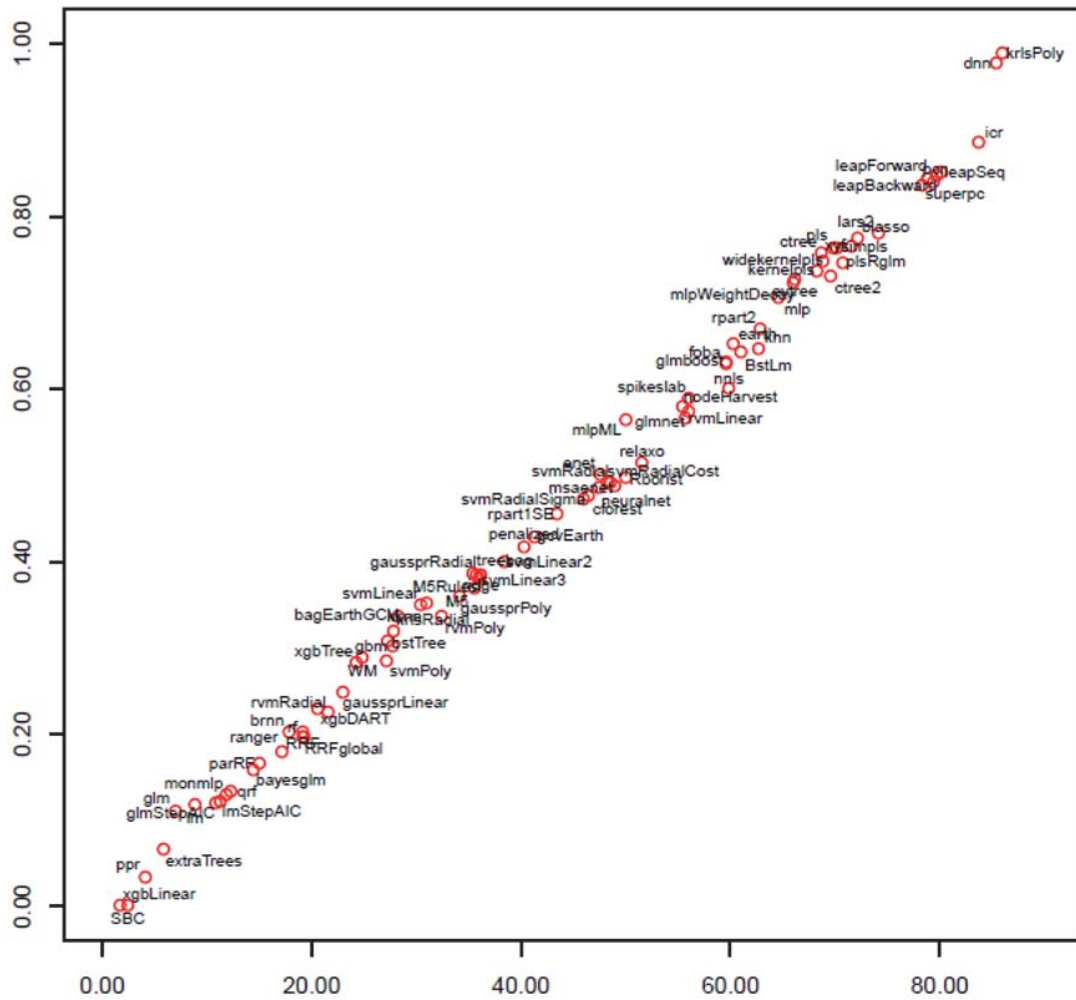


Figure 2.6. Relationships between average of rank(X-axis) and PE<sub>k</sub>(Y-axis) for seven crops. The range of PE<sub>k</sub> is set between 0 and 1 because *k*th regression method with *PE* larger than 1 means no prediction performance according to Eq. 2.3

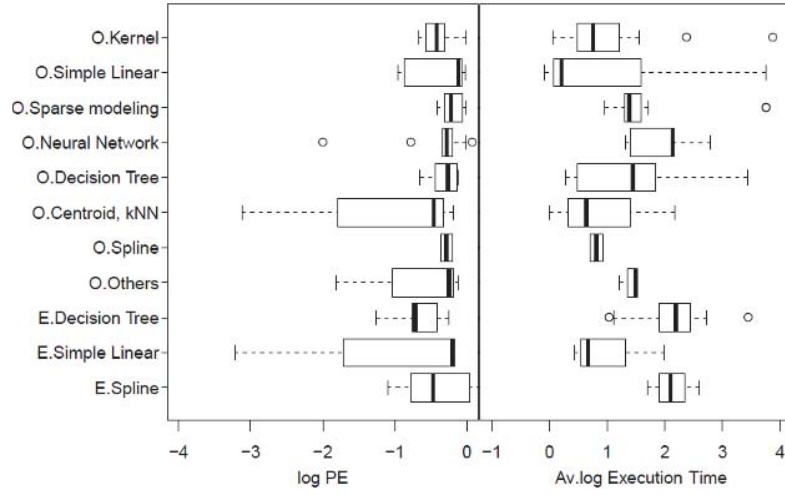


Figure 2.7. Box plots for log PE and average of log(Execution Time) for seven crops corresponding to 11 categories of regression methods.



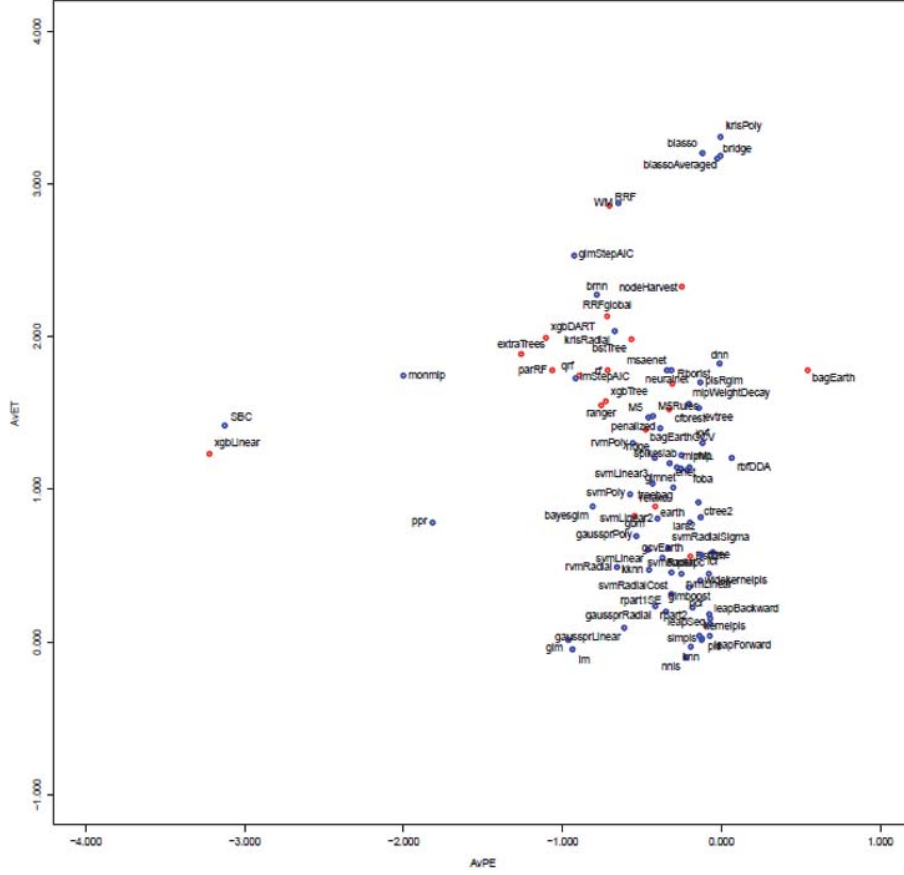


Figure 2.8. Relationships between average of  $\log(\text{PE})$  and average of  $\log(\text{Execution Time})$  for regression methods. Ordinary learning and Ensemble learning methods corresponds to blue and red circles, respectively.

## 2.4 Conclusion of Chapter 2

In Chapter 2, the prediction model of pesticide recovery rate was developed and 89 machine learning methods in ‘caret’ package. Most machine learning methods were not able to predict the pesticide recovery rate successfully, but two methods of xgbLinear (eXtreme Gradient Boosting Liner) and SBC (Subtractive and K-Means Fuzzy Clustering Method) performed well in prediction at shorter time for building the model, less than 2 minutes.

### **3. Selection of optimum molecular descriptors for recovery rate prediction using graph clustering tool**

#### **3.1 Introduction**

In Chapter 2, I proposed the procedure to estimate the pesticide recovery using the literature of JPL [20]. With rcdk, 178 molecular descriptors(MDs) were obtained by the canonical SMILES of each pesticide. All MDs were used as the explanatory variables for predicting the pesticide recovery rate. Some combinations among these MDs are correlated strongly that can influence on the performance of regression for pesticide recovery rate prediction such as the multicollinearity [30]. In this chapter, I propose the procedure to select the optimum MD for regression analysis using the correlation analysis and clustering method for optimum selection of MDs. Following two strategies were taken into consideration of MD selection.

##### **1. Reduction of highly correlated MDs**

Select unique MDs utilizing the correlation analysis, i.e. select the MD with less correlations with any other MDs.

##### **2. Minimize the loss of information**

Select as many MDs as possible in order for minimizing the loss of the information utilizing the clustering tool.

#### **3.2 Materials and Methods**

##### **3.2.1 Correlation analysis and cluster analysis**

In order to select the optimum MDs for machine learning based on the two considerations, I propose the process of the flow chart for MD selection shown in the Figure 3.1. Two clustering methods were used in selection of MDs in this thesis, DP Clus of graph clustering tool and Hierarchical Clustering Analysis. In this chapter, the MD selection using DP Clus is discussed. MD selection by hierarchical cluster analysis is discussed in the next chapter.

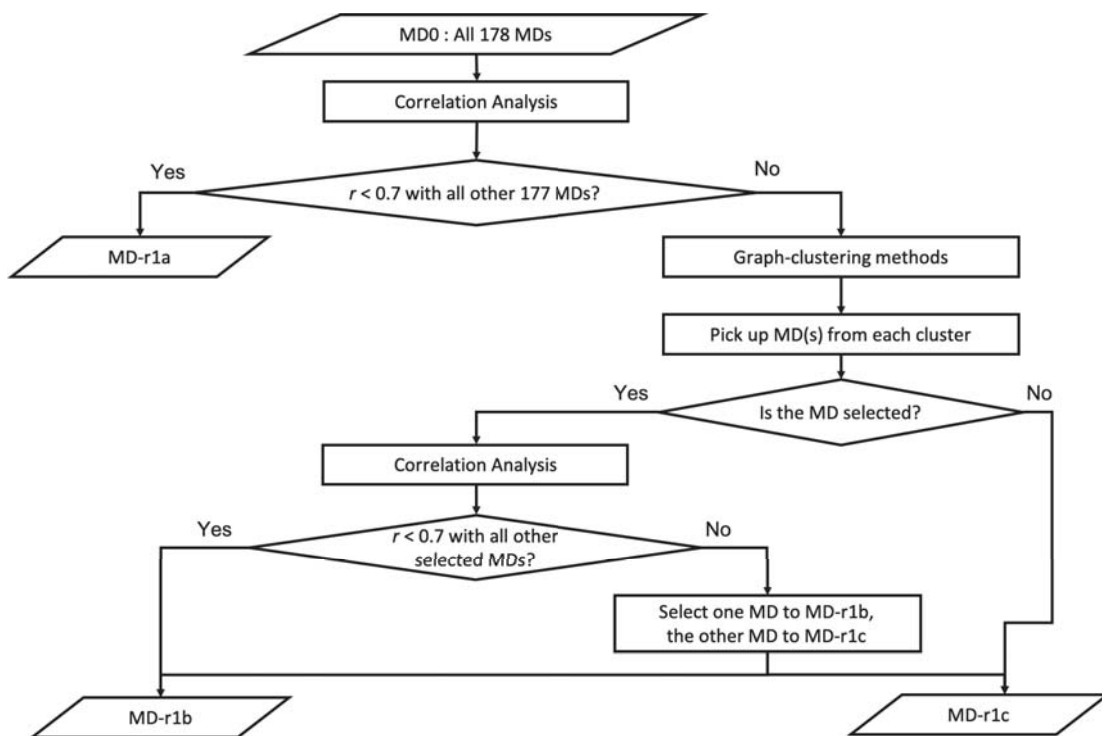


Figure 3.1. Process chart of selecting the optimum MDs

There are five steps for selection of MDs, 1) Input for correlation analysis, 2) Select the MDs of weak correlation with any other MDs, 3) cluster analysis to pick up representative MD(s) from each cluster and removal of other MDs, 4) correlation analysis for the MDs selected by Step 3 and 5) final selection of MDs from the cluster analysis.

**1st step: List the correlations of all possible combinations** The first step is to list the correlations of all possible combinations among 178 MDs. MD-MD correlations were calculated by the Pearson’s correlation coefficient( $r$ ) [26] using ‘corr’ package of R program and stretch function [31] for all 178 MDs. Based on the guidances of Pearson correlation coefficient [32], I set the threshold at  $r = 0.7$  for the “Highly correlated” of MDs on the present study. The 178 MDs were divided into two groups by this threshold  $r \geq 0.7$ . The MDs in the combinations of  $r \geq 0.7$  are classified as “Strongly correlated MD” and the other MDs are “Weakly correlated MD group” in the present study.

**2nd step: Pick up the MDs with weak correlations with other MDs** The second step is pick up the MDs of weak correlation(i.e.  $r < 0.7$ ) with any other MDs. These MDs are grouped as “MD-r1a” which are used for regression analysis of machine learning later.

**3rd step: Pick up the MDs by graph clustering tool** The third step is to visualize the correlations of strongly correlated MDs by the graph-clustering method called DP Clus [33] and pick up the representative MD(s) from each cluster. The parameters of DP Clus software is set as Table 3.1.

Table 3.1. DP Clus parameters

Parameter	Value
Cluster Property Value ( $cp_{nk}$ )	0.5
Density Value ( $d_k$ )	0.9
Minimum Cluster Value	2

The Density value is the threshold of the cluster density in the MD-MD connection network. In the MD-MD network of DP Clus, MD is represented as the

node and MD-MD connection is represented as the edge. The density value  $d_k$  of any cluster  $k$  is the ratio of the number of edges present in the cluster ( $|E_k|$ ) and the maximum possible number of edges in the cluster ( $|E_k|_{max}$ ), which is defined as Eq 3.1.

$$d_k = \frac{|E_k|}{|E_k|_{max}} = \frac{2 \times |E_k|}{|N_k| \times (|N_k| - 1)} \quad (3.1)$$

where  $|N_k|$  is the number of MDs in the cluster. If the number of the edges (connection of the MDs) are 90% of maximum combination of the edges, the MDs are grouped into the same cluster.

The cluster property value is the threshold of the connection around the cluster defined as Eq 3.2.

$$cp_{nk} = \frac{|E_{nk}|}{d_k \times |N_k|} \quad (3.2)$$

where  $|E_{nk}|$  is the total number of edges between the node  $n$  and each of the nodes of cluster  $k$ .

#### 4th step: Confirm the correlation of MDs picked up by graph clustering

The fourth step using DP Clus is to confirm the second correlation analysis among the representative MDs picked up from each cluster. Threshold of correlation is set at  $r \geq 0.7$ .

#### 5th step: Finalize the MDs for prediction

The fifth step is to select the MD(s) based on the step 4. The MDs of weak correlation with other MDs (i.e.  $r < 0.7$ ) in step 4 are grouped as “MD-r1b”, which are used for regression analysis of machine learning later. The MDs of the combination with the strong correlation in step 4 are divided into two groups according to the combination of  $r$ , i.e. “MD-r1b” and “MD-r1c”. MDs in MD-r1c are excluded from further regression analysis. Thus, 178 MDs are divided into three groups as listed in the table 3.2 according to the process in Figure 3.1.

Table 3.2. Group of MDs for optimum selection using DP Clus

MD group	MDs to be classified
MD-r1a	MD of $r < 0.7$ with any of other 177 MDs(Weakly correlated)
MD-r1b	MD of $r \geq 0.7$ with any of other 177 MDs(Strongly correlated) and selected by graph-clustering method
MD-r1c	MD of $r \geq 0.7$ with any of other 177 MDs(Strongly correlated) and excluded by graph-clustering method

### 3.2.2 Machine learning by ‘caret’ package

Prediction Error(PE) and Execution Time(ET) were measured the same procedure as Chapter 2 for 89 regression machine learning methods, then these performances were compared.

### 3.3 Results and Discussion

#### 3.3.1 Correlation analysis among molecular descriptors

Figure 3.2 is the histogram of the combination of MDs with the correlation analysis. The x-axis is the Pearson correlation coefficient and y-axis is the count(frequency) of MD combinations. There are 15,753 combinations of 178 MDs in total and 658 combinations(4.2%) consisted of 118 MDs were  $r \geq 0.7$ , which correlate strongly. Other 60 MDs were correlated with the other MDs at 0.7, which was classified as the MD group MD-r1a in the Table 3.1.

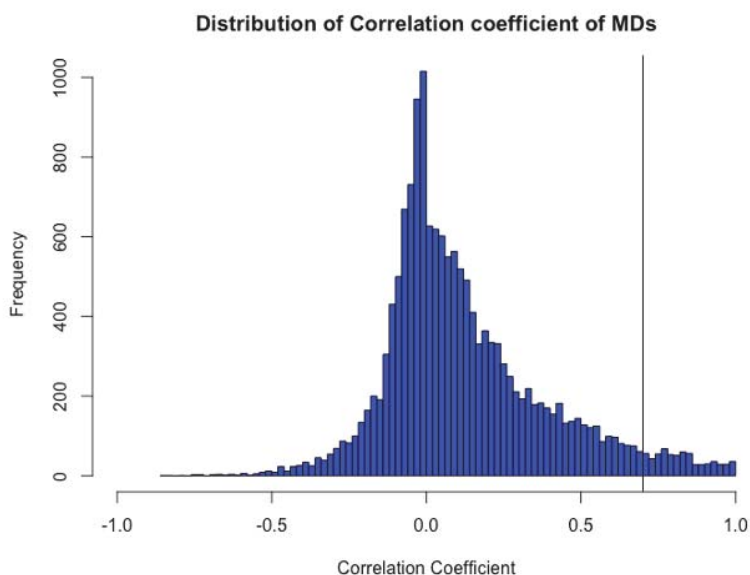
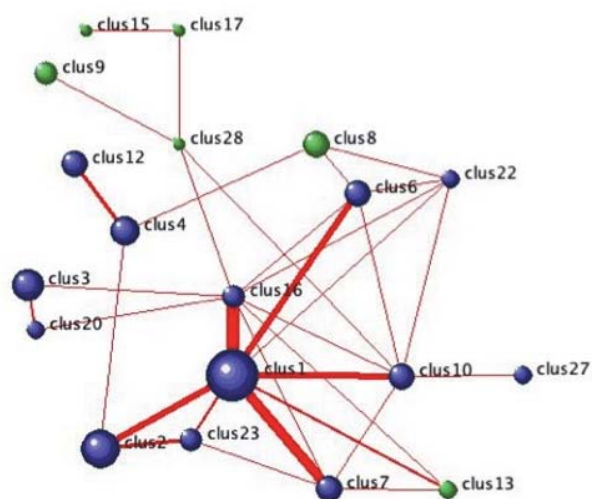
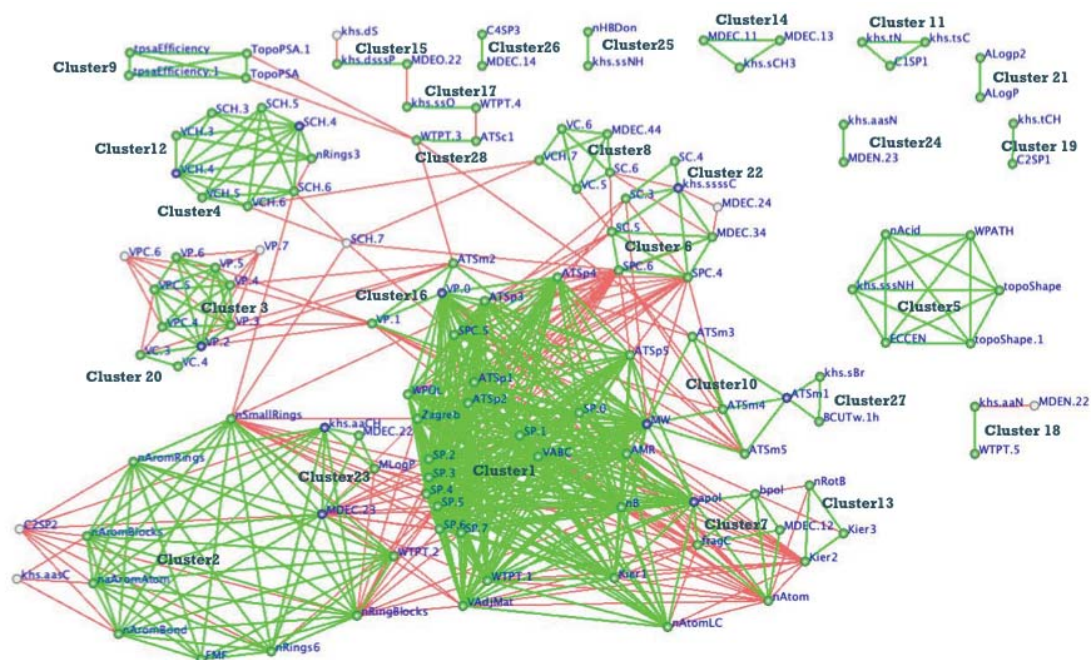


Figure 3.2. Distribution of correlation coefficient of MD

#### 3.3.2 Selection of molecular descriptors by the clustering tool - DP Clus

Relationships of 118 MDs of strongly correlated with any other MD(s) were classified into 28 clusters according to the cluster analysis by DP Clus, shown in Figure 3.3. 19 clusters are correlated each other dependently as shown in Figure 3.4. MDs in the other 9 clusters correlate independently.





**Criteria of selection from each cluster** I set the criteria as shown in the decision tree(Figure 3.5) for selection of the MD(s) from each cluster as the candidates of the machine learning for prediction of pesticide recovery. There are six decisions for selections of MD(s) from each cluster in order to avoid redundancy and loss of information. The first step is the criteria for number of MDs in the cluster. There were several combinations of MDs with weak correlation in the same cluster due to the presence of other many MDs with strong correlations. In order to reduce the loss of information, this should be the first check point for selection of MDs from each cluster. The second filter is applied only for the cluster with the number of MD  $\geq 5$ . If there is any combination of MDs with  $r < 0.5$ , both MDs is selected for from that cluster. Else, third filter was applied for selection of MDs. Before third filter, correlation analysis in the cluster was performed and four combinations of MDs with least  $r$  were picked up. The third filter is the frequency of MD in the combinations of least  $r$  within the cluster. There are several clusters that have several MDs with same frequency in the combination. The forth criteria is the type of MD, “not integer” is added for more information from that MD for better prediction of regression. If there is a single MD that meets these criteria in the four combinations of MDs, that MD is selected. The fifth filter is the number of connection with the other MDs in the cluster. This is the assumption that the node with less edge(s) should be less correlation. The final filter is the number of inter cluster connection in the cluster. This is also the assumption that less edges should be less correlation with the other MDs.



Figure 3.5. Decision tree to select the MD(s) from each cluster

**Selection of molecular descriptors by the clustering tool - Cluster 1** In DP Clus, the MDs are denoted as the node by small circle and the relation of MDs are indicated as edge, green line for intra cluster and red line for inter cluster. In Cluster 1, 26 molecular descriptors of polarizability and molecular shape (Topology descriptors) are included and correlated, i.e. AMR(Molar polarizability), apol (Sum of atomic polarizabilities), ATSp1-5 (Autocorrelation polarizability), Kier1 (First Kappa Shape Index), MW(Molecular Weight), nAtomLC (Number of atoms in the largest chain), nB (Number of bonds), SP.0-7 (Chi path descriptor, single path), SPC.5 (Chi path cluster descriptor, single path, order 5), VABC (Atomic and bond contributions of Van der Waals volume), VAdjMat (Vertex adjacency information), VP.0 (Chi path descriptor, valence path, order 0), WPOL (Weiner Polarity Number), WTPT.1 (Molecular ID) and Zagreb (Sum of the squares of atom degree over all heavy atoms). Cluster 1 connects to the other eight clusters (Cluster 2, 6, 7, 10, 13, 16, 22, 23). According to the decision tree

in Figure 3.5, SPC.5 and nAtomLC are selected. The correlation coefficient of these two MDs were at  $r = 0.329$ .

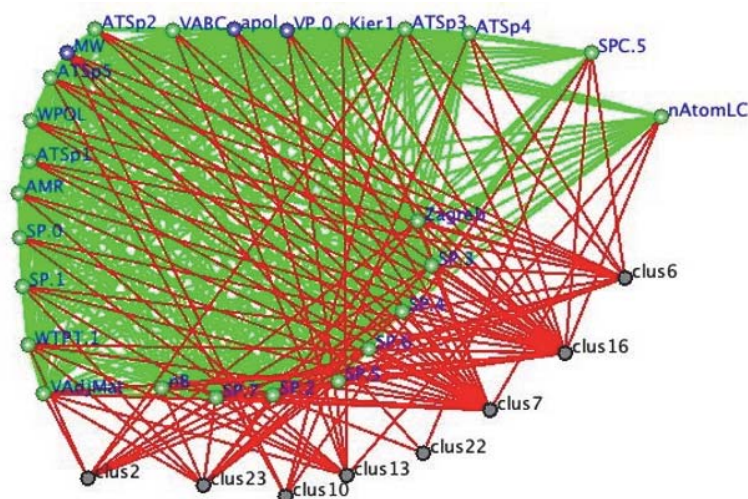


Figure 3.6. Cluster 1: 26 molecular descriptors correlates as shown in green lines. Eight other clusters are connected by the red lines.

**Selection of molecular descriptors by the clustering tool - Cluster 2** In Cluster 2, 11 molecular descriptors of ring count and carbon related descriptors are included and correlated, i.e. FMF(FMF descriptor characterizing complexity of a molecule), khs.aaCH(Count of atom type E-state CH and connection with two aromatic rings), MDEC.23(Between all secondary and tertiary carbons), naAromAtom(Number of aromatic atoms), nAromBlocks(Number of aromatically connected bonds), nAromBond(Number of aromatic bonds), nAromRings(Number of aromatic rings), nRingBlocks(Total number of distinct ring blocks), nRings6(individual breakdown of small rings), nSmallRings(Number of small rings from size 3 to 9), WTPT.2(Molecular ID / number of atoms) as shown in the Figure 3.7. Many MDs in cluster 2 are the aromatic and ring descriptors. Cluster 2 connects to the other three clusters(Cluster 1,4,23). Most of combinations of cluster 2 are over 0.6 of correlation coefficient. The MDs in least 4 combinations of  $r$  were nSmallRings, WTPT.2, FMF and khs.aaCH. According

to the decision tree in Figure 3.5, khs.aaCH is selected from the cluster 2.

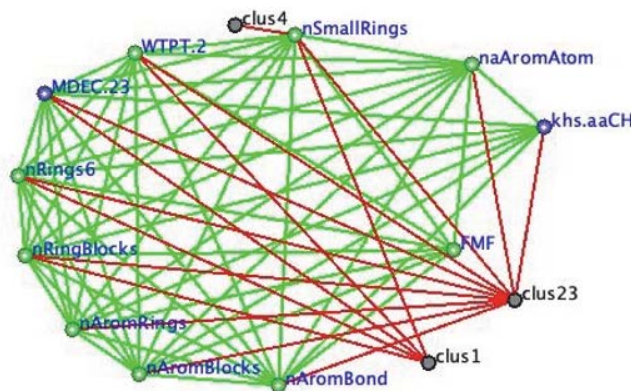


Figure 3.7. Cluster 2: 11 molecular descriptors correlates as shown in green lines. Three other clusters are connected by the red lines.

### Selection of molecular descriptors by the clustering tool - Cluster 3

In cluster 3, seven Chi path descriptors(topological descriptors) are correlated, i.e. VP.2-6(Chi path descriptor, valence path, order 2 to 6), VPC.4-5(Chi path cluster descriptor, valence path, order 4 and 5) as shown in Figure 3.8. Cluster 3 connects to the other two clusters(cluster 16, 20). According to the decision tree in Figure 3.5, VP.6 is selected from the cluster 3.

### Selection of molecular descriptors by the clustering tool - Cluster 4

In cluster 4, six MDs of Chi chain descriptors(topological descriptors) are connected , i.e. SCH.4-6(Chi chain descriptors, simple chain order 4 to 6) and VCH.4-6(Chi chain descriptors, valence chain order 4 to 6) as shown in Figure 3.9. Cluster 4 connects to the other three clusters(Cluster 2, 8, 12). According to the decision tree in Figure 3.5, VCH.4 is selected.

### Selection of molecular descriptors by the clustering tool - Cluster 5

Cluster 5 is the independent cluster, i.e. no connection to the other clusters. Six molecular descriptors in the cluster 5 are ECCEN(A topological descriptor combining distance and adjacency information), khs.sssNH(Count of atom type

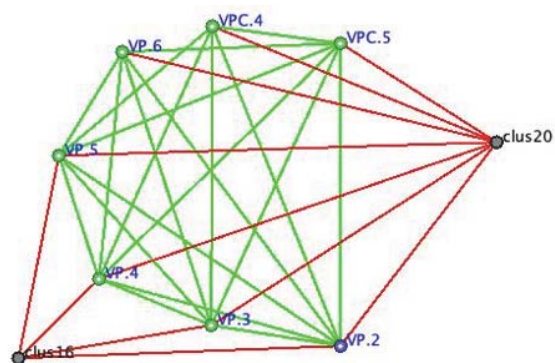


Figure 3.8. Cluster 3: Seven molecular descriptors correlate as shown in green lines. Two other clusters are connected by the red lines.

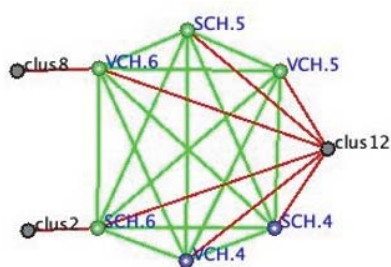


Figure 3.9. Cluster 4: Six molecular descriptors correlate as shown in green lines. Three other clusters are connected by the red lines.

Table 3.3. Molecular descriptors comparison between Nereistoxin and average of other 247 pesticides

Pesticides	nAcid	WPATH	topoShape	khs.sssNH	ECCEN	topoShape.1
Nereistoxin oxalate	2	48,000,000,043	999	1	230,196,198	999
Average of other 247 pesticides	0	870	1	0	315	1

E-state NH with the other three single bonds), nAcid(Acidic group count descriptor), topoShape(A measure of the anisotropy in a molecule), topoShape.1(A measure of the anisotropy in a molecule) and WPATH(Weiner path number) as shown in Figure 3.10. Five molecular descriptors in cluster 5 are the topological descriptors. Nereistoxin oxalate was the only one pesticide with the unique value of molecular descriptors as shown in the Table 3.3. This pesticide is consisted with two molecules as shown in Figure 3.11, that gives the unique value on the MDs of cluster 5, whereas all the other 247 pesticides are single molecule. According to the decision tree in Figure 3.5, TopoShape is selected.

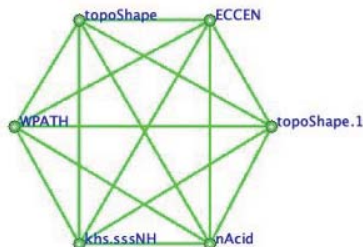


Figure 3.10. Cluster 5: Six molecular descriptors are correlated as the green line.

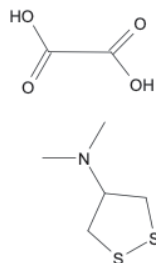


Figure 3.11. Chemical structure of Nereistoxin oxalate, the pesticide with two molecules

**Selection of molecular descriptors by the clustering tool - Cluster 6** In cluster 6, five MDs were correlated, i.e. khs.ssssC(Count of atom type E-state tertiary carbon), MDEC.34(between tertiary and quarterly carbons), SC.5(Chi cluster descriptor, simple path, order 5) and SPC.4 and 6(Chi path cluster descriptor, simple path, order 4 and 6) as shown in Figure 3.12. Cluster 6 connects to other five clusters(Cluster 1, 8, 10, 16, 22) According to the decision tree in Figure 3.5, MDEC.34 is selected.

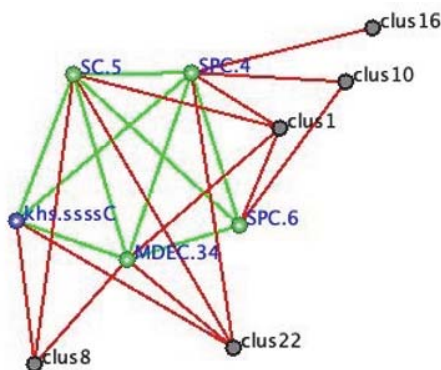


Figure 3.12. Cluster 6: Five MDs correlated as green lines. Other five clusters are connecting with red lines.

**Selection of molecular descriptors by the clustering tool - Cluster 7** In cluster 7, five MDs are correlated, apol (Sum of the atomic polarizabilities),

bpol (um of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule), fragC (Complexity of a system), MDEC.12 (between all primary and secondary carbons) and nAtom(NumberOf atoms) as shown in Figure 3.13. Cluster 7 connects to other four clusters(Cluster 1, 13, 16, 23). According to the decision tree in Figure 3.5, MDEC.12 is selected.

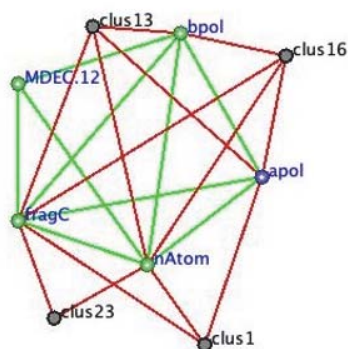


Figure 3.13. Cluster 7: Five MDs are correlated as green lines. Cluster 7 connects with the other four clusters as shown in red lines

**Selection of molecular descriptors by the clustering tool - Cluster 8** In cluster 8, five MDs are correlated, i.e. MDEC.44(between all quarterly carbons), SC.6(Chi cluster descriptor, simple cluster, order6), VC.5 and 6(Chi cluster descriptor, valence cluster, order 5 and 6) and VCH.7(Chi chain descriptor, valence chain, order 7) are included on the Cluster 8 as shown in Figure 3.14. Cluster 8 connects to other three clusters(Cluster 4, 6, 22). According to the correlation analysis within Cluster 8. Most of combinations are over 0.65 of correlation coefficient, with the result in the unique value as shown in Table 3.4 caused by the molecular structure of  $\alpha$ -Endosulfan as shown in Figure 3.15). According to the decision tree in Figure 3.5, VCH.7 is selected.

**Selection of molecular descriptors by the clustering tool - Cluster 9** In cluster 9, four MDs are correlated, TopoPSA(topological polar surface area), TopoPSA.1(topological polar surface area), tpsaEfficiency(Polar surface area ex-



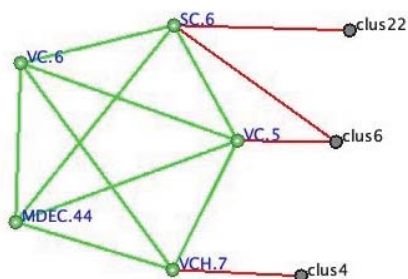


Figure 3.14. Cluster 8: Five MDs are correlated as green lines. Cluster 8 connects with the other four clusters as shown in red lines

Table 3.4. Molecular descriptors comparison between  $\alpha$ -Endosulfan and average of other 247 pesticides

Pesticides	MDEC.44	SC.5	VC.5	VC.6	VCH.7
$\alpha$ -Endosulfan	2.38	2.39	2.70	1.04	2.01
Average of other 247 pesticides	0.03	0.36	0.15	0.02	0.10

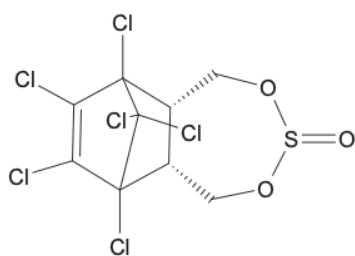


Figure 3.15. Chemical structure of  $\alpha$ -Endosulfan

pressed as a ratio to molecular size) and tpsaEfficiency.1(Polar surface area expressed as a ratio to molecular size) as shown in Figure 3.16, connecting to cluster 28. These clusters are the molecular descriptors of topological polar surface area. According to the decision tree in Figure 3.5, tpsaEfficiency is selected.

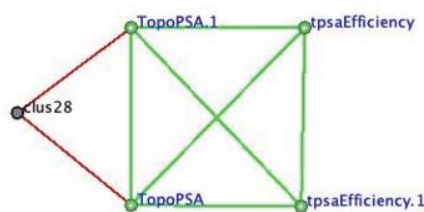


Figure 3.16. Cluster 9: Four MDs correlate as shown in green lines. Cluster 9 connects with cluster 28 in red lines.

### Selection of molecular descriptors by the clustering tool - Cluster 10

In cluster 10, five MDs are included, i.e. ATSm1,3-5(Autocorrelation descriptor, weighted by scaled atomic mass) and MW(Molecular weight) as shown in Figure 3.17. Cluster 10 connects to other seven clusters(Cluster 1, 6, 7, 16, 22, 27, 28). According to the decision tree in Figure 3.5, ATSm.1 is selected.

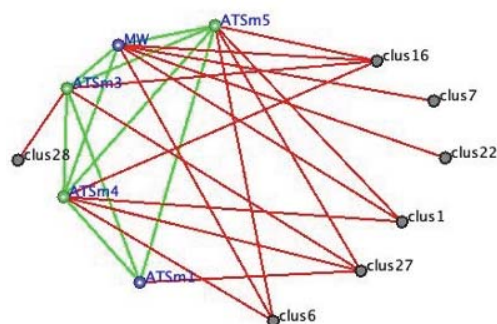


Figure 3.17. Cluster 10: Five MDs are correlated as shown in green lines. Cluster 10 connects with the other seven clusters as red lines.

### Selection of molecular descriptors by the clustering tool - Cluster 11

In cluster 11, three MDs are correlated, i.e. C1SP1(Triply bound carbon bound to one other carbon), khs.tN(Count of atom type E-state tertiary nitrogen) and khs.tsC(Count of atom type E-state tertiary carbon with single bond) as shown in the Figure 3.18, no connection with the other cluster. These MDs are unique to the pesticides that includes cyano group in the molecule. According to the decision tree in Figure 3.5, khs.tsC is selected.

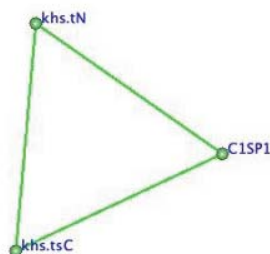


Figure 3.18. Cluster 11: Three MDs are correlated in green lines.

### Selection of molecular descriptors by the clustering tool - Cluster 12

In cluster 12, five MDs correlate, i.e. nRings3(individual breakdown of small rings with 3 members), SCH.3-4(Chi chain descriptor, simple chain, orders 3 and 4), VCH.3-4(Chi chain descriptor, valence chain, orders 3 and 4) as shown in Figure 3.19. Cluster 12 connects with cluster 4. According to the decision tree in Figure 3.5, SCH.3 and VCH.3 are highest frequency with the smallest number of nodes connected to the other MDs and clusters. SCH.3 is selected.

### Selection of molecular descriptors by the clustering tool - Cluster 13

In cluster 13, three MDs correlated, i.e. Kier2(Kappa shape indices descriptor, second kappa shape index), Kier3(Kappa shape indices descriptor, third kappa shape index) and nRotB(Number of rotatable bonds, excluding terminal bonds) as shown in Figure 3.20. Cluster 13 connects to other three clusters(Cluster 1, 7, 16). According to the decision tree in Figure 3.5, Kier3 is selected.

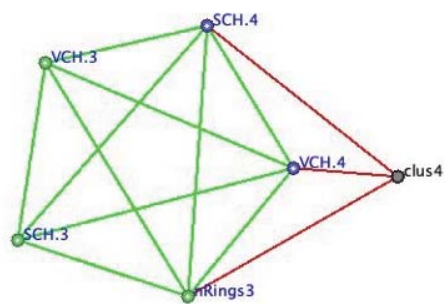


Figure 3.19. Cluster 12: Five MDs are correlated as shown in green lines. Cluster 12 connects with cluster 4 with red lines.

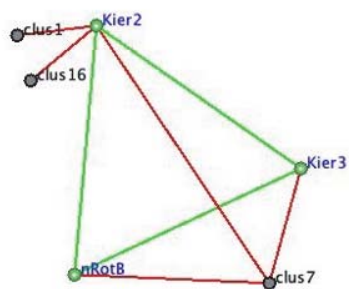


Figure 3.20. Cluster 13: Three MDs correlate as shown in green lines. Cluster 13 connects with three clusters in red lines.

#### Selection of molecular descriptors by the clustering tool - Cluster 14

In cluster 14, three MDs correlate, i.e. khs.sCH3(Count of atom type E-state of CH3 with single bond), MDEC.11(Molecular Distance Edge Descriptor, between all primary carbons) and MDEC.13(Molecular Distance Edge Descriptor, between all primary and tertiary carbons) as shown in Figure 3.21, no connection with the other cluster. According to the decision tree in Figure 3.5, MDEC.13 is selected.

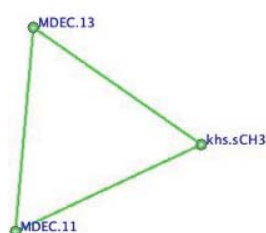


Figure 3.21. Cluster 14: Three MDs correlate as shown in green lines.

#### Selection of molecular descriptors by the clustering tool - Cluster 15

Cluster 15 includes two MDs, khs.dsssP(Count of atom type E-state of phosphorus with one double bond and three single bonds) and MDEO.22(Molecular distance edge descriptor, between all secondary oxygens). MDEO.22 is connecting to khs.ssO of the cluster 17 as shown in the Figure 3.22. According to the decision tree in Figure 3.5, khs.dsssP is selected



Figure 3.22. Cluster 15: Two MDs correlated with green line, and MDEO.22 is connecting with the khs.ssO of cluster 17

#### Selection of molecular descriptors by the clustering tool - Cluster 16

In cluster 16, four MDs are correlated strongly, i.e. ATSm2(Autocorrelation

descriptor, weighted by scaled atomic mass), MW(Molecular weight) and VP.0-1(Chi path descriptor, valence path, order 0 and 1) as shown in Figure 3.23. Cluster 16 connects with other nine clusters(Cluster 1, 3, 6, 7, 10, 13, 20, 22 28). According to the decision tree in Figure 3.5, VP.0 is selected.

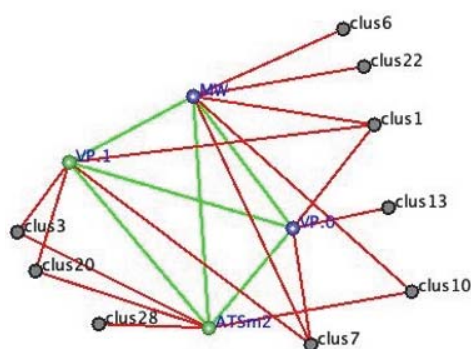


Figure 3.23. Cluster 16: Four MDs correlate as shown in green lines. Cluster 16 connects with nine clusters in red lines.

### Selection of molecular descriptors by the clustering tool - Cluster 17

Two MDs correlated, i.e. khs.ssO(Count of atom type E-state of oxygen with two single bonds) and WTPT.4(Weighted path descriptor, Sum of path lengths starting from oxygens) with the correlation coefficient 0.826. Both MDs are connecting to the other different clusters(Figure 3.24). WTPT.4 is selected based on the decision tree in Figure 3.5.



Figure 3.24. Cluster 17: Two MDs correlate as green line. Both MDs connect to the other clusters.

### Selection of molecular descriptors by the clustering tool - Cluster 18

In cluster 18, two MDs correlated, i.e. khs.aaN(Count of atom type E-state of nitrogen with two aromatic groups) and WTPT.5(Sum of path lengths starting from nitrogens) as shown in Figure 3.24, no connection with the other cluster. According to the flow chart of decision tree of Figure 3.5, WTPT.5 is selected.



Figure 3.25. Cluster 18: Two MDs correlated as in green line.

### Selection of molecular descriptors by the clustering tool - Cluster 19

In cluster 19, two MDs correlated, i.e. C2SP1(Triply bound carbon bound to two other carbons) and khs.tCH(Count of atom type E-state of triple bonds of carbon) are in the cluster 19 with completely correlate each other as shown in Figure 3.26. Two pesticides, Propyzamide and Flumioxazin(Figure 3.27) have unique value of these MDs as shown in Table 3.5, which includes triple bond of carbon( $\text{-C}\equiv\text{CH}$ ). khs.tCH is selected according to the decision tree of Figure 3.5.



Figure 3.26. Cluster 19: Two MDs correlate as shown in green line.

Table 3.5. Molecular descriptors comparison among Propyzamide, Flumioxazin and average of other 246 pesticides

Pesticides	khs.tCH	C2SP1
Propyzamide	1	1
Flumioxazin	1	1
Average of other 246 pesticides	0	0

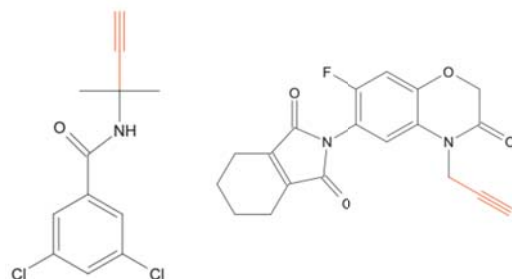


Figure 3.27. Chemical structure of Propyzamide (left) and Flumioxazin (right). (-C≡CH) is colored in red

### Selection of molecular descriptors by the clustering tool - Cluster 20

In cluster 20, three MDs of topology descriptors of VP.2-4 (Chi path descriptor, valence path, order 2 to 4) are in the cluster 20 with high correlation one another. VP.2 connected to the cluster 3 and cluster 16, and VP.3 to cluster 3 as shown in Figure 3.28. According to the decision tree of Figure 3.5, VP.4 is selected.

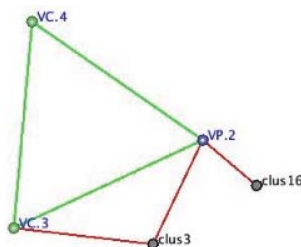


Figure 3.28. Cluster 20: Three MDs correlated as shown in green lines. Cluster 20 connected with the cluster 3 and cluster 16 as shown in red lines

### Selection of molecular descriptors by the clustering tool - Cluster 21

In cluster 21, two MDs of AlogP (Ghose-Crippen  $LogK_{ow}$ ) and ALogP2 (Square of ALogP) are included on Cluster 21 (Figure 3.29, no connection with the other cluster). AlogP2 is the square of AlogP by definition and confirmed from the 248 pesticides result as shown in the scatter plot Figure 3.30. According to the decision tree of Figure 3.5, AlogP2 is selected from cluster 21.





Figure 3.29. Cluster 21: Two MDs correlate by green line.

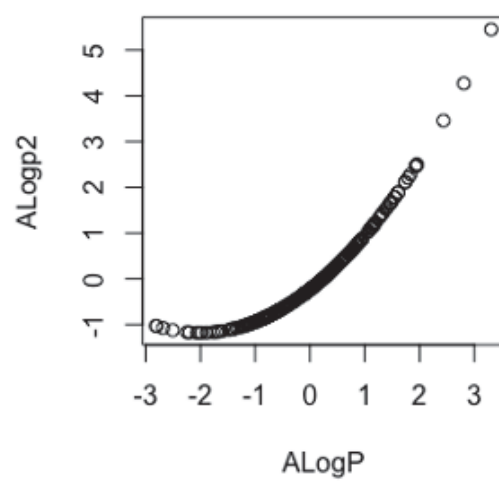


Figure 3.30. Scatter plot of 248 pesticides, AlogP vs. AlogP2

### Selection of molecular descriptors by the clustering tool - Cluster 22

In cluster 22, three MDs correlated, i.e. khs.ssssC(Count of atom type E-state of four single bond carbon) and SC.3-4(Chi cluster descriptor, simple cluster, orders 3 and 4) with high correlation one another(Figure 3.31). SC.3 was connected to other four clusters(Cluster 1, 6, 8, 16) and khs.ssssC connected to the cluster 6 and 8. According to the decision tree of Figure 3.5, SC.4 is selected.

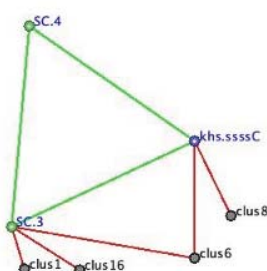


Figure 3.31. Cluster 22: Three MDs correlate in green line. Two MDs connected to the other clusters in red lines.

### Selection of molecular descriptors by the clustering tool - Cluster 23

In cluster 23, four MDs correlated, i.e. khs.aaCH(Count of atom type E-state of carbon connecting with two aromatic groups and hydrogen), MDEC.22(between all secondary carbons), MDEC.23(between all secondary and tertiary carbons) and MLogP(Mannhold LogP) as shown in Figure 3.32. Cluster 23 connects to the other 3 clusters(Cluster 1, 2, 7). MDEC.22 is selected according to the decision tree of Figure 3.5.

### Selection of molecular descriptors by the clustering tool - Cluster 24

In cluster 24, two MDs correlated, i.e. khs.aasN(Count of atom type E-state of nitrogen connecting with two aromatic groups and one single bond) and MDEN.23(Molecular distance edge descriptor, between all secondary and tertiary nitrogens) with the correlation coefficient 0.829(Figure 3.33), no connection with the other cluster. According to the decision tree of Figure 3.5, MDEN.23 is selected.

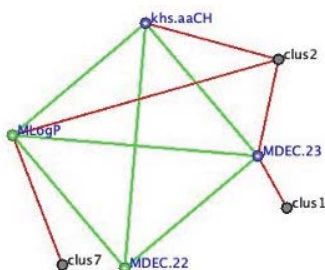


Figure 3.32. Cluster 23: Four MDs correlated as shown in green lines. Cluster 23 connects with the other clusters as shown in red lines.



Figure 3.33. Cluster 24: Two MDs correlate as shown in green line.

#### Selection of molecular descriptors by the clustering tool - Cluster 25

In cluster 25, two MDs correlate, i.e. khs.ssNH(Count of atom type E-state of nitrogen connecting with hydrogen and two other single bonds) and nHBDon(Number of hydrogen bond donors) with the correlation coefficient 0.807(Figure 3.34), no connection with the other cluster. According to the decision tree of Figure 3.5, nHBDon is selected.



Figure 3.34. Cluster 25: Two MDs correlate as shown in green line.

#### Selection of molecular descriptors by the clustering tool - Cluster 26

In cluster 26, two MDs correlated, i.e. C4SP3(Singly bound carbon bound to four other carbons) and MDEC.14 (Molecular distance edge descriptor, between all

primary and quaternary carbons) with the correlation coefficient 0.806 (Figure 3.35), no connection with the other cluster. According to the decision tree of Figure 3.5, MDEC.14 is selected.



Figure 3.35. Cluster 26: Two MDs correlated as shown in the green line.

### Selection of molecular descriptors by the clustering tool - Cluster 27

In cluster 27, three MDs correlated, i.e. ATSm1 (ATS autocorrelation descriptor, weighted by scaled atomic mass), BCUTw.1h (The number of lowest eigenvalue highest atom weighted BCUTS) and khs.sBr (Count of atom type E-state of bromine). ATSm1 is connected to the cluster 10(Figure 3.36). According to the decision tree of Figure 3.5, BCUTw.1h is selected.

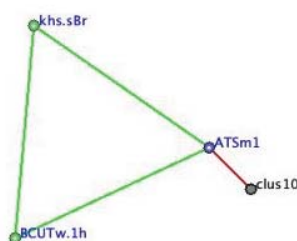


Figure 3.36. Cluster 27: Tree MDs correlate as shown in green lines. Cluster 27 connects to the cluster 10 as shown in red line.

### Selection of molecular descriptors by the clustering tool - Cluster 28

In cluster 28, two MDs correlated, i.e. ATSc1(ATS autocorrelation descriptor, weighted by charges) and WTPT.3(Sum of path lengths starting from heteroatoms) with the correlation coefficient 0.747(Figure 3.37). ATSc1 is less number of nodes connected to the other cluster than WTPT.3. According to the decision tree of Figure 3.5, ATSc1 is selected.

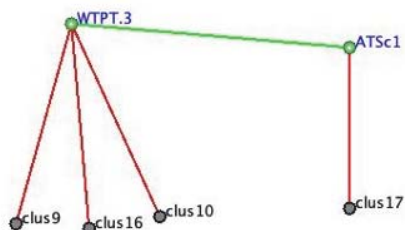


Figure 3.37. Cluster 28: Two MDs correlate as shown in green line. Both MDs connect to the other clusters by red lines.

### 3.3.3 Selection of molecular descriptors for machine learning after cluster analysis by DP Clus

**Correlation analysis of MDs selected by cluster analysis** Table 3.6 shows the candidates of 29 MDs selected by the clustering according to the decision tree in Figure 3.5 from 28 clusters.

Table 3.6. Candidates of MD2(29 MDs from 28 clusters)

Cluster	Molecular Descriptors
1	SPC.5, nAtomLC
2	khs.aaCH
3	VP.6
4	VCH.4
5	TopoShape
6	MDEC.34
7	MDEC.12
8	VCH.7
9	tpsaEfficiency
10	ATSm1
11	khs.tsC
12	SCH.3
13	Kier3
14	MDEC.13
15	khs.dsssP
16	VP.0
17	WTPT.4
18	WTPT.5
19	khs.tCH
20	VC.4
21	Alogp2
22	SC.4
23	MDEC.22
24	MDEN.23
25	nHBDOn
26	MDEC.14
27	BCUTw.1h
28	ATSc1

I performed correlation analysis for these MDs for final selection. The combinations of MDs with  $r \geq 0.7$  among them are listed in the Table 3.7. As shown in the Figure 3.38, these MDs were selected from the dependent clusters, which results in high correlation after the selection of cluster analysis.

Table 3.7. Combination of MDs with the  $r \geq 0.7$ 

MD-a	MD-b	$r$	Cluster of MD-a	Cluster of MD-b
khs.aaCH	MDEC.22	0.833	2	23
ATSm1	BCUTw.1h	0.777	10	27
VCH.4	SCH.3	0.769	4	12
nAtomLC	VP.0	0.751	1	16
WTPT.4	ATSc1	0.736	17	28
SPC.5	MDEC.34	0.731	1	6

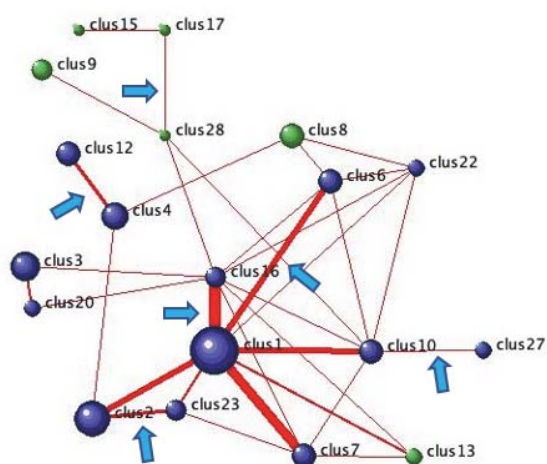


Figure 3.38. Cluster diagram for dependent clusters

**Final selection of MDs for machine learning** The MDs of the column MD-a in the Table 3.7 are selected and MD-b are deleted for machine learning. Table 3.8 shows the 23 MDs finally selected by the cluster analysis for machine learning of pesticide recovery rate prediction.

Table 3.8. Final MDs selected by cluster analysis

Cluster	Molecular Descriptors
3	VP.6
5	TopoShape
6	MDEC.34
7	MDEC.12
8	VCH.7
9	tpsaEfficiency
11	khs.tsC
12	SCH.3
13	Kier3
14	MDEC.13
15	khs.dsssP
16	VP.0
18	WTPT.5
19	khs.tCH
20	VC.4
21	Alogp2
22	SC.4
23	MDEC.22
24	MDEN.23
25	nHBDon
26	MDEC.14
27	BCUTw.1h
28	ATSc1

Table 3.9 is the summary of MDs selected by the decision tree (Figure 3.5).

Table 3.9. Summary of molecular descriptors selected by the correlation analysis and cluster analysis

MD group	Description of MDs	Number of MDs	Selected
MD-r1a	MD of $r < 0.7$ with any of other 177 MDs	60	Yes
MD-r1b	MD of $r \geq 0.7$ with any of other 177 MDs and selected by graph-clustering method	23	Yes
MD-r1c	MD of $r \geq 0.7$ with any of other 177 MDs and excluded by graph-clustering method	95	No

Thus, MD-r1a and MD-r1b (both MDs are combined as the MD group of “MD2”) will be used for the regression analysis by caret. Here, all 178 MDs(original MDs discussed in Chapter 2) are expressed as “MD0” for comparison of the performance against MD2.



### 3.3.4 Comparison of machine learning performance between with and without selection of molecular descriptors

By selecting the MDs, 57 machine learning methods gave better PE for regression analysis and 32 methods got worse as shown in the Figure 3.39.

Table 3.10 and Table 3.11 are the lists of best and worst top 20 machine learning methods of Prediction Error.

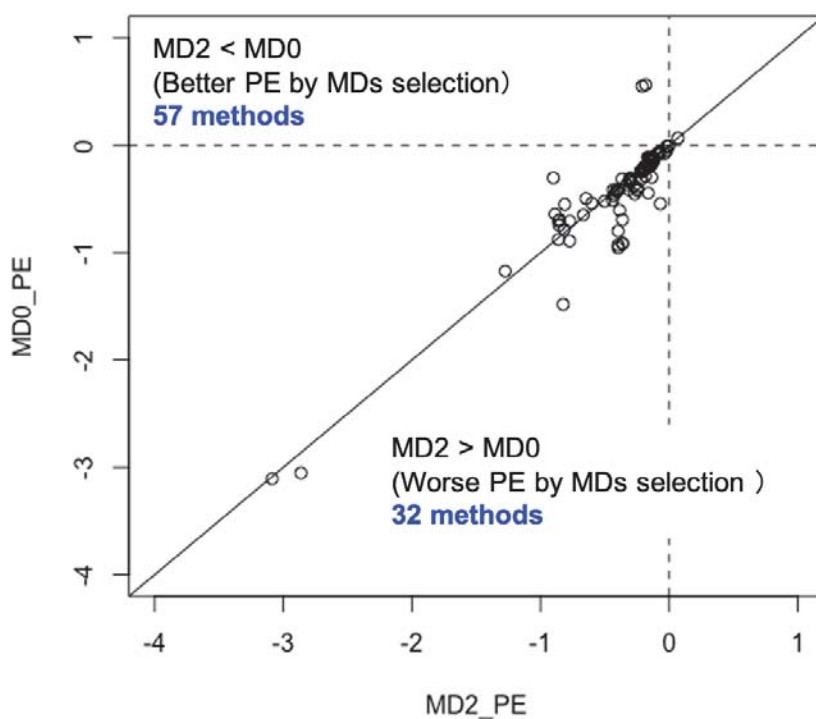


Figure 3.39. Comparison of Prediction Error by machine learning method on with and without selection of molecular descriptors by correlation analysis and cluster analysis. Plots on the rectangular line shows no difference by selection of molecular descriptors.

Table 3.10. Top 20 method for MD2 > MD0(Worse PE by selecting the molecular descriptors sorted by the Prediction Error difference)

Method	Category	MD2 PE	MD0 PE	PE Diff.
bagEarthGCV	E. Spline	3.501	-0.473	3.974
ppr	O. Others	-0.824	-1.482	0.657
glm	O. Simple Liner	-0.397	-0.956	0.559
glmStepAIC	O. Simple Liner	-0.363	-0.918	0.556
lmStepAIC	O. Simple Liner	-0.363	-0.912	0.550
lm	O. Simple Liner	-0.397	-0.930	0.532
svmPoly	O. Kernel	-0.069	-0.545	0.476
bayesglm	O. Simple Liner	-0.397	-0.798	0.401
brnn	O. Neural Network	-0.364	-0.693	0.329
gaussprPoly	O. Kernel	-0.163	-0.445	0.282
gaussprLinear	O. Kernel	-0.386	-0.605	0.220
SBC	O.Centroid,kNN	-2.862	-3.052	0.190
svmLinear	O. Kernel	-0.268	-0.455	0.187
svmLinear3	O. Kernel	-0.247	-0.421	0.174
enet	O. Sparse Modeling	-0.136	-0.300	0.164
svmLinear2	O. Kernel	-0.247	-0.398	0.151
monmlp	O. Neural Network	-0.774	-0.891	0.117
penalized	O. Sparse Modeling	-0.274	-0.380	0.107
relaxo	O. Sparse Modeling	-0.185	-0.289	0.103
ridge	O. Sparse Modeling	-0.313	-0.415	0.103

Table 3.11. Top 20 method for MD2 < MD0(Better PE by selecting the molecular descriptors sorted by the Prediction Error difference)

Method	Category	MD2 PE	MD0 PE	PE Diff.
lasso	O. Sparse Modeling	-0.232	23.723	-23.955
lars	O. Sparse Modeling	-0.161	8.871	-9.032
bagEarth	E. Spline	-0.283	5.398	-5.681
bridge	O. Simple Liner	-0.211	0.552	-0.762
blassoAveraged	O. Sparse Modeling	-0.183	0.564	-0.747
Rborist	E. Decision Tree	-0.901	-0.304	-0.597
xgbTree	E. Decision Tree	-0.812	-0.550	-0.262
xgbDART	E. Spline	-0.888	-0.640	-0.248
RRF	E. Decision Tree	-0.851	-0.694	-0.158
rf	E. Decision Tree	-0.860	-0.708	-0.152
krlsRadial	O. Kernel	-0.647	-0.496	-0.151
ranger	E. Decision Tree	-0.855	-0.744	-0.111
extraTrees	E. Decision Tree	-1.275	-1.173	-0.102
RRFglobal	E. Decision Tree	-0.773	-0.709	-0.064
WM	O. Decision Tree	-0.601	-0.541	-0.061
blasso	O. Sparse Modeling	-0.167	-0.111	-0.057
neuralnet	O. Neural Network	-0.367	-0.314	-0.052
simpls	O. Simple Liner	-0.157	-0.117	-0.040
pls	O. Simple Liner	-0.156	-0.117	-0.039
parRF	E. Decision Tree	-0.817	-0.782	-0.035
lars2	O. Sparse Modeling	-0.141	-0.107	-0.034

bagEarthGCV (ensemble spline method, LogPE -0.473 to 3.501), ppr (ordinary other method, -1.482 to -0.824) and four ordinary simple liner methods(glm -0.956 to -0.397, glmStepAIC -0.918 to -0.363, lmStepAIC -0.912 to -0.397 and lm -0.930 to -0.397) got worse in prediction error by the selection of MDs with cluster analysis, as shown in the Table 3.10. lasso(LogPE 23.723 to -0.232) and lars(8.871 to -0.161) of the ordinal Sparse methods and bagEarth(5.398 to -0.283) of the ensemble spline method were very high prediction error (poor performance in prediction) in MD0, which were improved in MD2 by selecting the MDs with cluster analysis as Table 3.11 for this data set. Figure 3.40 shows the distribution of the Prediction Error by the machine learning method category. The distribution of Execution Time (ET) by machine learning method category is shown in Figure 3.41. Both box plots compare the performances between MD0 (MDs before removal of highly correlated MDs) and MD2 (MDs after removal of highly correlated MDs). Many machine learning methods in the Ordinary Simple Linear category got worse in Prediction Error by the selection of molecular descriptors

due to the less explanatory variables (MDs) by reduction of highly correlated MDs. Ordinary Decision trees, Ordinary Centroid and Ensemble Simple Liner were small differences in prediction error by the selection of MDs. ET of most machine learning methods was decreased by selecting the molecular descriptors as shown in the Figure 3.41. Table 3.12 shows the top 20 machine learning methods with shorter ET and Table 3.13 shows the top 20 machine learning methods with longer ET by MD selection with clustering. Time consuming methods (krlsPoly of ordinary Kernel Log ET 3.870 to 2.509, blasso of ordinary sparse modeling 3.754 to 2.768 and bridge of ordinary simple liner 3.743 to 2.813) were improved by reducing the MDs. On the other hand, the methods with small ET methods (BstLm of ensemble simple liner 0.675 to 0.980, knn of ordinary centroid, kNN -0.002 to 0.273 and simpls of ordinary simple liner 0.030 to 0.304) showed slight increases by selecting the MDs with cluster analysis.

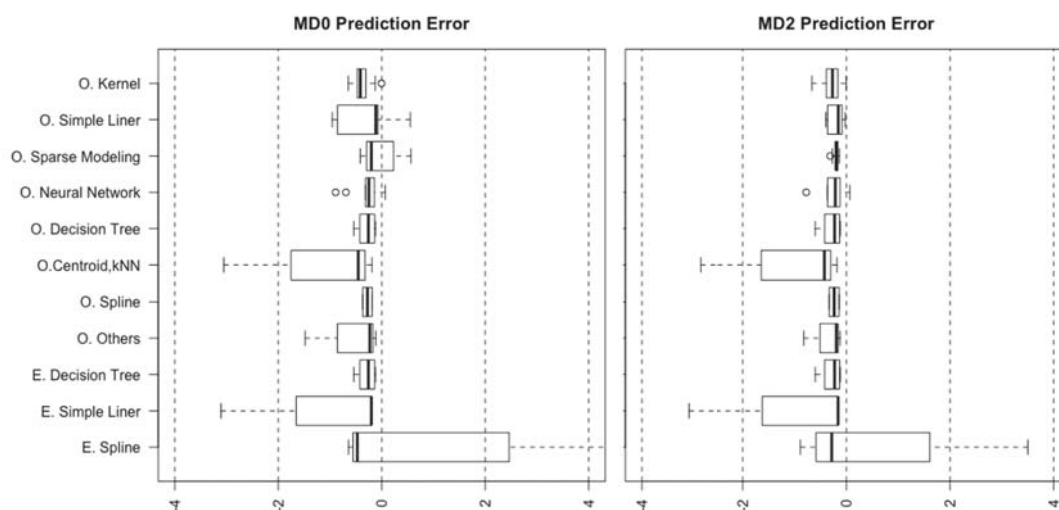


Figure 3.40. Boxplot of Prediction Error comparison of the molecular descriptors before selection(left) and after selection(right).

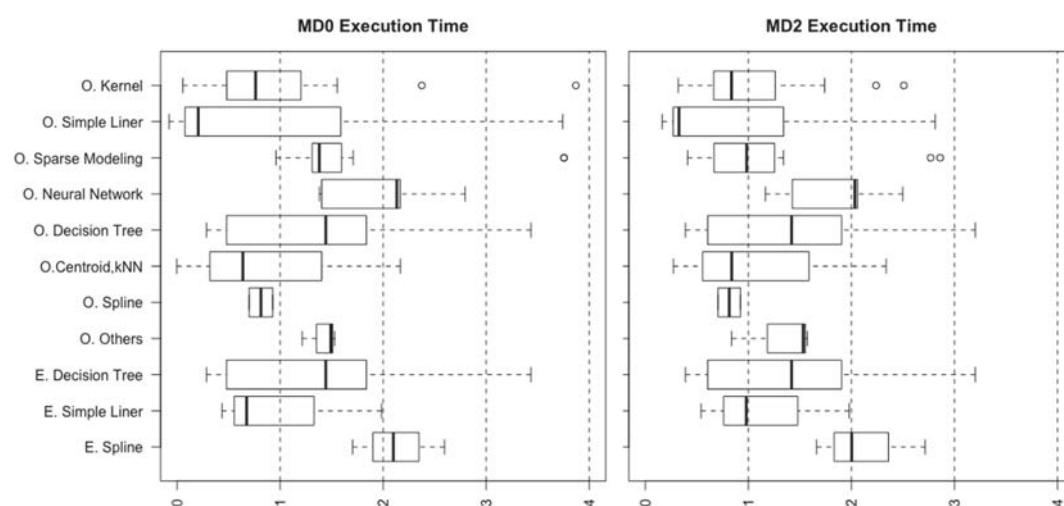


Figure 3.41. Boxplot of Execution Time comparison of the molecular descriptors before selection(left) and after selection(right).

Table 3.12. Top 20 method for MD2 < MD0 (Shorter ET in MD2 by selecting the molecular descriptors sorted by the Execution Time difference)

Method	Category	MD2 ET	MD0 ET	ET Diff.
krlsPoly	O. Kernel	2.509	3.870	-1.361
blasso	O. Sparse Modeling	2.768	3.754	-0.985
bridge	O. Simple Liner	2.813	3.743	-0.930
blassoAveraged	O. Sparse Modeling	2.861	3.755	-0.895
glmStepAIC	O. Simple Liner	2.242	3.116	-0.873
brnn	O. Neural Network	2.058	2.795	-0.738
lasso	O. Sparse Modeling	0.609	1.341	-0.732
glmnet	O. Sparse Modeling	0.723	1.380	-0.657
lars	O. Sparse Modeling	0.410	0.984	-0.574
enet	O. Sparse Modeling	0.935	1.438	-0.502
ridge	O. Sparse Modeling	0.983	1.482	-0.499
lars2	O. Sparse Modeling	0.462	0.960	-0.498
bayesglm	O. Simple Liner	0.731	1.166	-0.434
ppr	O. Others	0.838	1.213	-0.375
penalized	O. Sparse Modeling	1.340	1.709	-0.369
RRF	E. Decision Tree	3.130	3.442	-0.312
relaxo	O. Sparse Modeling	0.988	1.280	-0.292
parRF	E. Decision Tree	1.947	2.188	-0.241
WM	O. Decision Tree	3.203	3.434	-0.232
RRFglobal	E. Decision Tree	2.380	2.604	-0.224

Table 3.13. Top 20 method for MD2 > MD0 (Longer ET in MD2 by selecting the molecular descriptors sorted by the Execution Time difference)

Method	Category	MD2 ET	MD0 ET	ET Diff.
BstLm	E. Simple Liner	0.980	0.675	0.305
knn	O.Centroid,kNN	0.273	-0.002	0.276
simpls	O. Simple Liner	0.304	0.030	0.274
kernelps	O. Kernel	0.319	0.056	0.264
extraTrees	E. Decision Tree	2.613	2.362	0.251
nnls	O. Simple Liner	0.165	-0.078	0.243
rvmRadial	O. Kernel	0.908	0.682	0.226
widekernelps	O. Kernel	0.700	0.482	0.219
lm	O. Simple Liner	0.309	0.099	0.210
gaussprLinear	O. Kernel	0.417	0.208	0.209
pls	O. Simple Liner	0.248	0.039	0.209
leapForward	O. Simple Liner	0.258	0.056	0.202
kknn	O.Centroid,kNN	0.838	0.638	0.200
rpart2	O. Decision Tree	0.478	0.285	0.193
rvmPoly	O. Kernel	1.739	1.556	0.184
gaussprRadial	O. Kernel	0.533	0.352	0.181
rvmLinear	O. Kernel	0.738	0.559	0.178
SBC	O.Centroid,kNN	2.337	2.168	0.169
glm	O. Simple Liner	0.325	0.175	0.149
evtree	O. Decision Tree	1.962	1.812	0.149

Tuning parameters of xgbLinear for MD2 are listed in the Table 3.14. The parameters were the same for all seven crops.

Table 3.14. Tuning parameters of xgbLinear (eXtreme Gradient Boosting Linear) for pesticide recovery prediction (MD2).

Parameter	Value
Number of rounds (Boosting Iterations)	50
Lambda (L2 Regularization)	0.0001
Alpha (L1 Regularization)	0.0001
Eta (Learning Rate)	0.3

### 3.4 Conclusion of Chapter 3

In Chapter 3, graph clustering tool was used to select the molecular descriptors for prediction model of pesticide recovery. DP Clus successfully classify the strongly

correlated molecular descriptors into 28 clusters and molecular descriptors from each cluster were selected. Correlation analysis of the selected molecular descriptors was required because some clusters independently connect in the MD-MD network. By selecting the molecular descriptors, execution time for building prediction model was shorten for most machine learning methods. Prediction error of xgbLiner (eXtreme Gradient Boosting Linear) was the same with and without the selection of molecular descriptors. On the other hand, prediction error of SBC (Subtractive and K-Means Fuzzy Clustering Method) got worse by selecting molecular descriptors due to the reduction of explanatory parameter for machine learning.



## 4. Optimum molecular descriptors selection by hierarchical cluster analysis

### 4.1 Introduction

In Chapter 2, I proposed the procedure to estimate the pesticide recovery using the literature of JPL [20] and the selection of MDs for machine learning is discussed in the Chapter 3. With rcdk, 178 molecular descriptors(MDs) were obtained by the canonical SMILES of each pesticide. All MDs were used as the explanatory variables for predicting the pesticide recovery rate. In this chapter, I propose the procedure to select the optimum MD using the hierarchical cluster analysis, then compare the results of pesticide recovery prediction obtained in the Chapter 2 and Chapter3. Just the same as chapter 3, two strategies below for selecting MDs for pesticide recovery rate prediction.

#### 1. Reduction of highly correlated MDs

Select unique MDs utilizing the correlation analysis, i.e. select the MD with less correlations with any other MDs.

#### 2. Minimize the loss of information

Select as many MDs as possible in order for minimizing the loss of the information utilizing the clustering tool.

### 4.2 Materials and Methods

#### 4.2.1 Correlation analysis among molecular descriptors

In order to select the optimum MDs for machine learning based on the two considerations, I propose the process of the flow chart for MD selection shown in the Figure 4.1. In this chapter, the MD selection using the hierarchical cluster analysis is discussed.

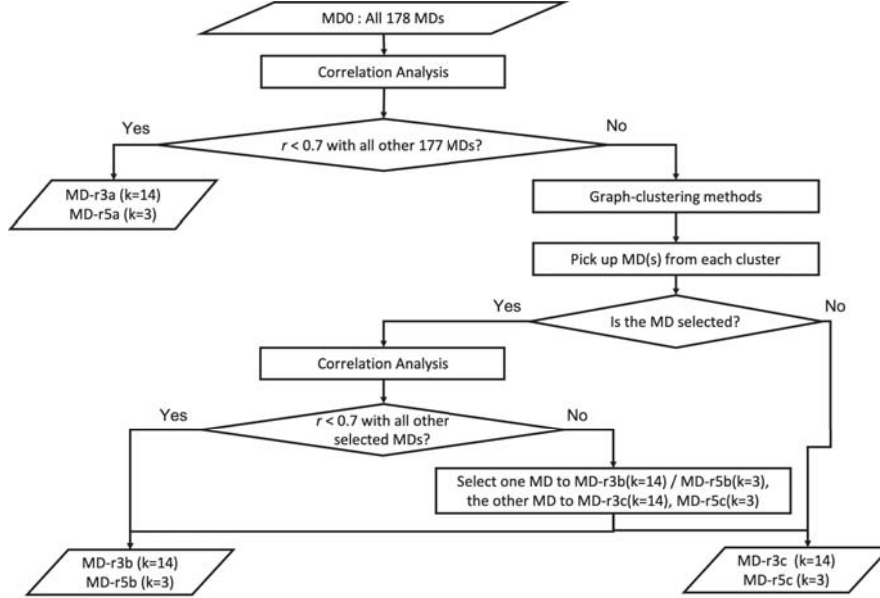


Figure 4.1. Selection of MD-MD pair from the dendrogram

There are five steps for selection of MDs, 1) Input for correlation analysis, 2) Select the MDs of weak correlation with any other MDs, 3) cluster analysis to pick up representative MD(s) from each cluster and removal of other MDs, 4) correlation analysis for the MDs selected by Step 3 and 5) final selection of MDs from the cluster analysis.

**1st step: List the correlations of all possible combinations** The first step is to list the correlations of all possible combinations among 178 MDs. MD-MD correlations were calculated by the Pearson's correlation coefficient( $r$ ) [26] using 'corr' package of R program and stretch function [31] for all 178 MDs. Based on the guidances of Pearson correlation coefficient [32], I set the threshold at  $r = 0.7$  for the "Highly correlated" of MDs on the present study. The 178 MDs were divided into two groups by this threshold  $r \geq 0.7$ . The MDs in the combinations of  $r \geq 0.7$  are classified as "Strongly correlated MD" and the other MDs are "Weakly correlated MD group" in the present study.

**2nd step: Pick up the MDs with weak correlations with other MDs**

The second step is pick up the MDs of weak correlation(i.e.  $r < 0.7$ ) with any other MDs. In this chapter, two conditions of hierarchical cluster analysis is discussed, number of clusters (k) at 14 and 3. These MDs are grouped as “MD-r3a” for k=14 and “MD-r5a” for k=3 which were used for regression analysis of machine learning later.

**3rd step: Pick up the MDs by hierarchical cluster analysis**

The third step is the hierarchical cluster analysis according to the similarity in the molecular descriptor profile among 248 pesticides. Dendrogram of molecular descriptors were obtained with the parameters shown in Table 4.1.

Table 4.1. Hierarchical Cluster Analysis parameters

Parameter	Value
Distance metric	Euclidian distance
Linkage criteria	Ward.D2

After obtaining the dendrogram of MDs, the optimum number of clusters (k) from the dendrogram were determined using the NbClust package [34]. The MDs from each cluster is picked up based on the following criteria.

**Pick up the MD-MD pair at the lowest distance in the cluster on the dendrogram**

Select the MD-MD pair of the most similar from each cluster according to the dendrogram as shown in the Figure 4.2.

**Pick up the MD that includes more information than the other**

Pick up the MD that is more range of z-score of 248 pesticides

Pick up the MD with real number rather than integer

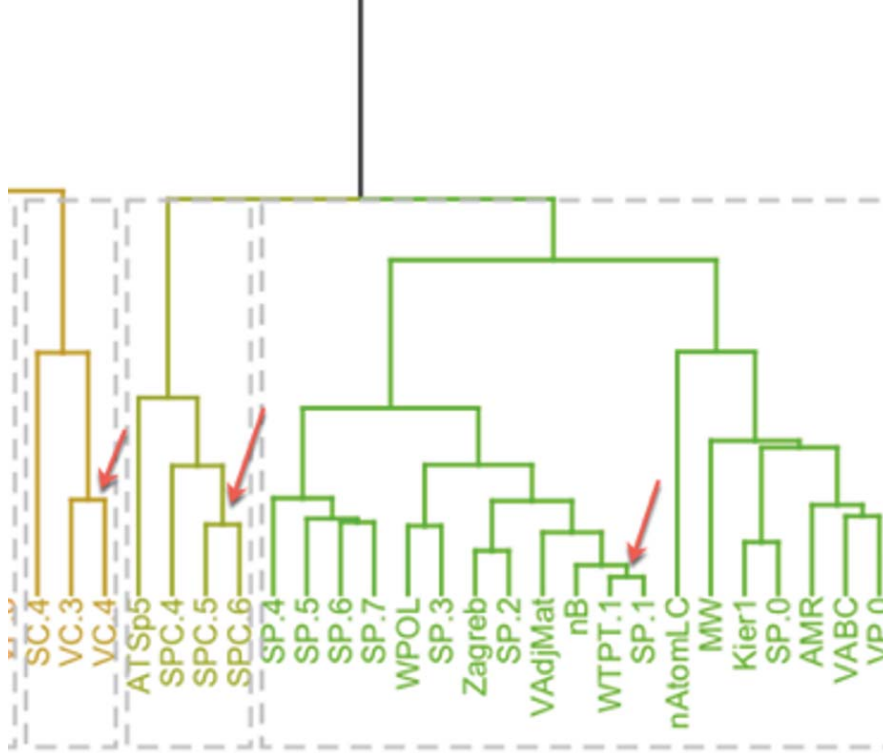


Figure 4.2. Selection of MD-MD pair from the dendrogram

**4th step: Confirm the correlation of MDs picked up by hierarchical cluster analysis** The fourth step using hierarchical cluster analysis is to confirm the correlation of MDs among picked up from each cluster. Threshold of correlation is set at  $r \geq 0.7$ .

**5th step: Finalize the MDs for prediction** The fifth step is the final selection the MD(s) based on the step 4. The MDs of weak correlation with other MDs(i.e.  $r < 0.7$ ) in step 4 are grouped as “MD-r3a” for  $k=14$  and “MD-r5a” for  $k=3$ , which are used for regression analysis of machine learning later. The MDs of the combination with the strong correlation in step 4 are divided into two groups, i.e. “MD-r3b” and “MD-r3c” for  $k=14$  and “MD-r5b” and “MD-r5c” for  $k=3$ . MDs in MD-r3c ( $k=14$ ) and MD-r5c ( $k=3$ ) were excluded from further

regression analysis. Thus, 178 MDs are divided into three groups as listed in the table according to the process in Figure 4.3.

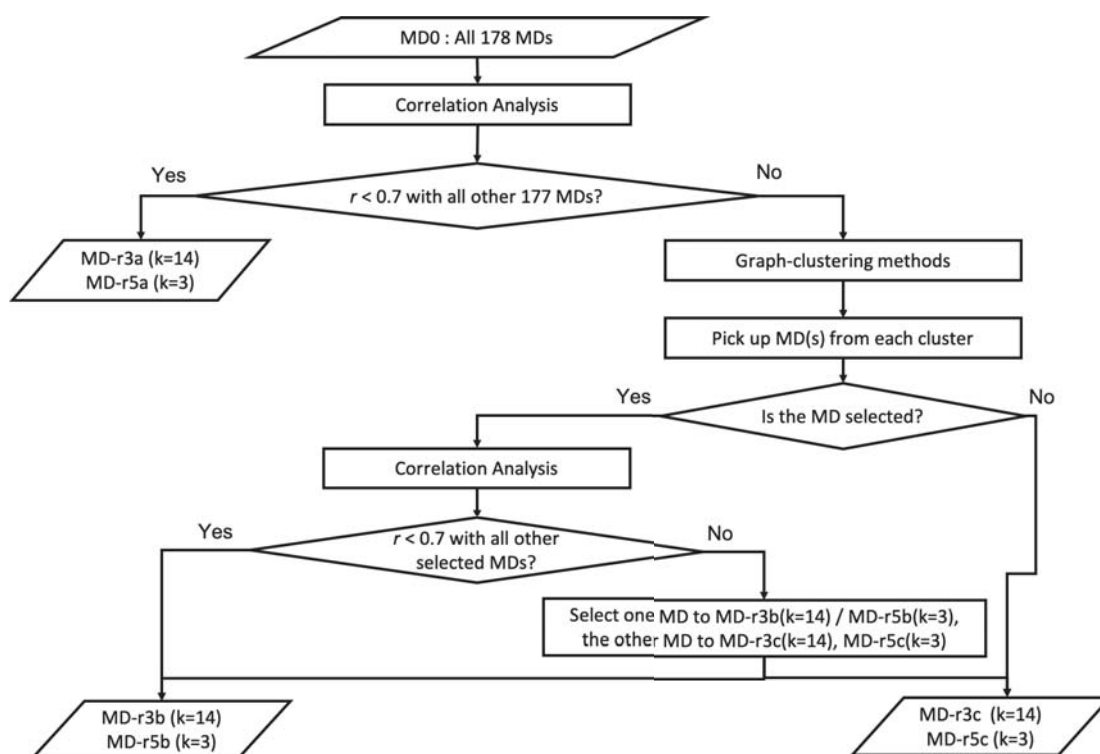


Figure 4.3. Process chart of selecting the optimum MDs

#### 4.2.2 Machine learning by ‘caret’ package

Prediction Error(PE) and Execution Time(ET) were measured the same procedure as Chapter 2 for 89 regression machine learning methods, then these performances were compared.

### 4.3 Results and Discussion

#### 4.3.1 Correlation analysis among molecular descriptors

Figure 4.4 is the histogram of the combination of MDs with the correlation analysis, which is the same as that of chapter 3. The x-axis is the Pearson correlation

coefficient and y-axis is the count(frequency) of MD combinations. There are 15,753 combinations of 178 MDs in total and 658 combinations(4.2%) consisted of 118 MDs were  $r \geq 0.7$ , which correlate strongly. Other 60 MDs were correlated with the other MDs at 0.7, which was classified as the MD group MD-r3a in the Table 4.2.

Table 4.2. Group of MDs for optimum selection using hierarchical cluster tree	
MD group	MDs to be classified
MD-r3a	MD of $r < 0.7$ with any of other 177 MDs(Weakly correlated)
MD-r3b	MD of $r \geq 0.7$ with any of other 177 MDs(Strongly correlated) and selected by hierarchical cluster analysis at $k=14$
MD-r3c	MD of $r \geq 0.7$ with any of other 177 MDs(Strongly correlated) and excluded by hierarchical cluster analysis at $k=14$
MD-r5a	MD of $r < 0.7$ with any of other 177 MDs(Weakly correlated)
MD-r5b	MD of $r \geq 0.7$ with any of other 177 MDs(Strongly correlated) and selected by hierarchical cluster analysis at $k=3$
MD-r5c	MD of $r \geq 0.7$ with any of other 177 MDs(Strongly correlated) and excluded by hierarchical cluster analysis at $k=3$

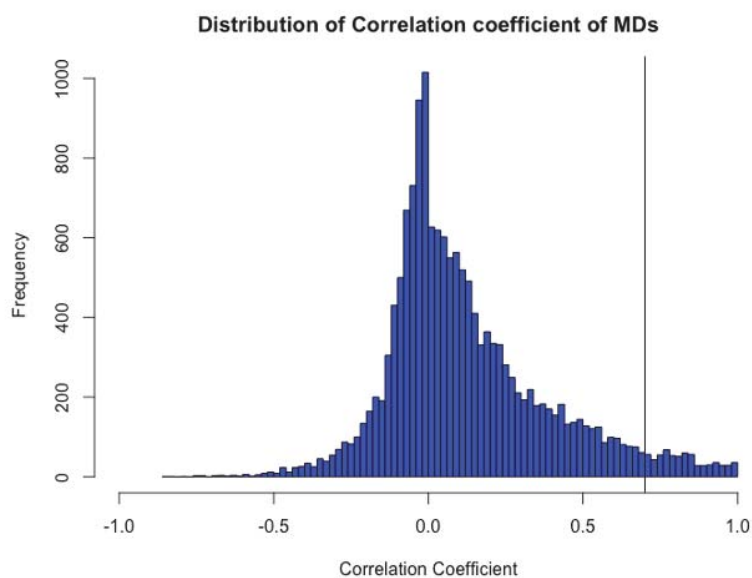


Figure 4.4. Distribution of correlation coefficient of MD

#### 4.3.2 Selection of molecular descriptors by the clustering tool - hierarchical cluster analysis

**Determination of number of clusters on the dendrogram by gap function** Dendrogram of 118 MDs according to the similarity of 248 pesticides was shown in the Figure 4.5

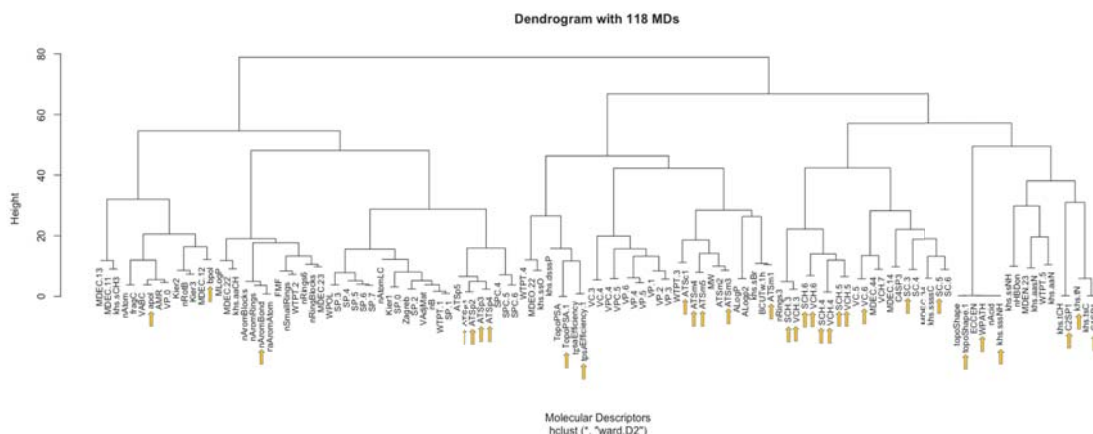


Figure 4.5. Dendrogram of 118 MDs according to the similarity of 248 pesticides

31 MDs out of 118 MDs were excluded on implementing the NbClust package of R program because they correlated with the other MDs strongly ( $r=1.0$ ) that caused error when executing NbClust package as highlighted with arrows in Figure 4.5. Therefore, 87 MDs were used for determination of number of cluster on hierarchical cluster analysis using NbClust, as shown in the dendrogram in the Figure 4.6.

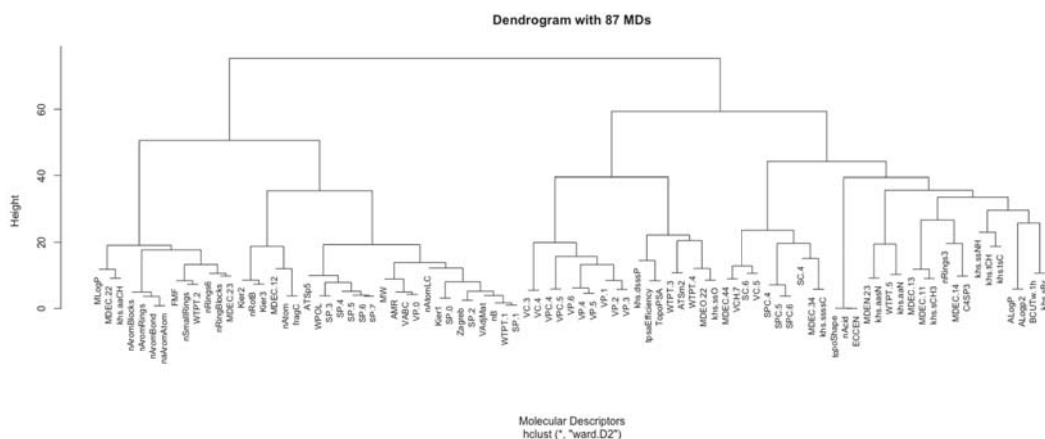


Figure 4.6. Dendrogram of 87 MDs for NbClust



With the NbClust package, the result of optimum numbers of cluster were suggested in the histogram (Figure 4.7), numbers of cluster (k) 2, 3 and 14 were suggested as the optimum number of clusters according to the gap function. Parameters of NbClust is listed in the Table 4.3.

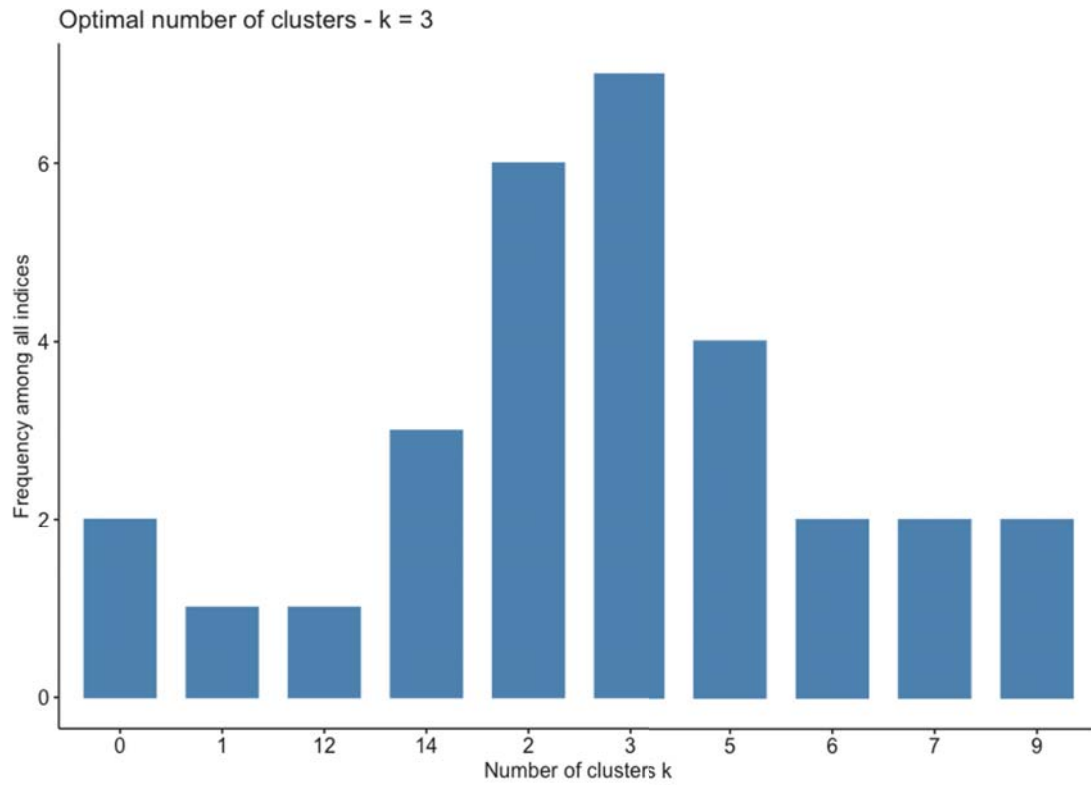


Figure 4.7. Histogram of the result of optimum number of cluster according to the gap function

Table 4.3. NbClust parameters

Parameter	Value
Distance	Euclidean
Minimal number of clusters	2
Maximal number of clusters	14
Cluster analysis method	ward.D2

According to the dendrograms in Figure 4.8, Figure 4.9 and Figure 4.10,  $k=2$  loses much information from the MDs in spite of optimum number according to gap function. Based on the strategy of MD selection for minimizing the loss of information by MD selection,  $k=2$  was excluded, i.e. two numbers of clusters at  $k=14$  and  $k=3$  were discussed for comparisons of Prediction Error (PE) and Execution Time (ET) of pesticide recovery rate prediction.

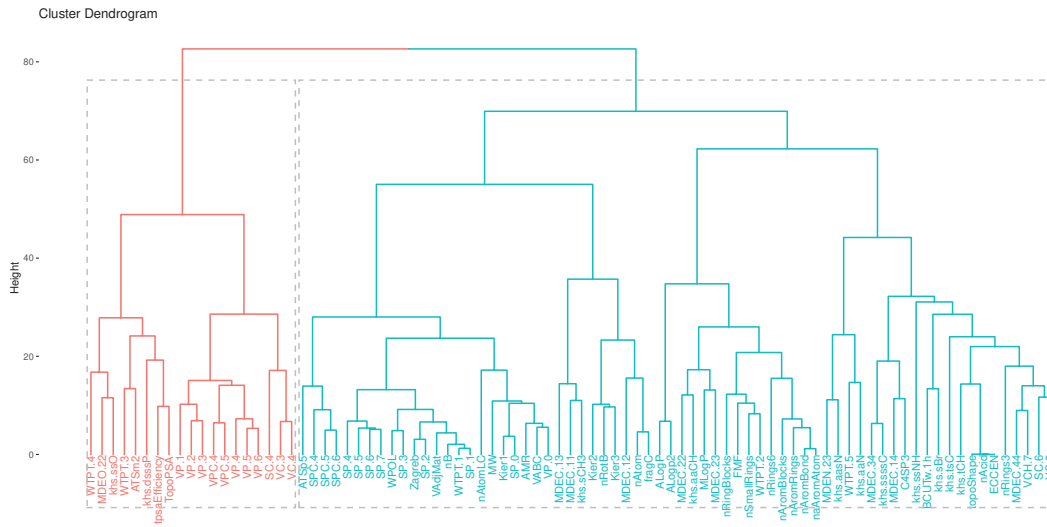


Figure 4.8. Dendrogram of 87 MDs at  $k=2$



based on the selection of MDs by clustering analysis, including the original MDs without the chapter of MDs. (MD0, MD2, MD4 and MD6)

Table 4.4. Molecular Descriptor group in this chapter

Molecular descriptor group	Molecular descriptors
MD0	All MDs obtained by rcdk package (Chapter 2)
MD2	MDs selected by DP Clus (Chapter 3)
MD4	MDs selected by hierarchical cluster analysis with k=14
MD6	MDs selected by hierarchical cluster analysis with k=3

**Selection of MDs from the dendrogram at k=14 (MD4)** 28 MDs from 87 MDs were selected as shown in the Table 4.7. 14 MDs were picked up as the candidate of machine learning for pesticide recovery prediction. The combination of SP.1 - fragC ( $r = 0.7577$ ) and SPC.6 - MDEC.34 (0.7352) were correlated strongly. SP.1 and SPC.6 were selected because their range of z-score is more than the other MD. 12 MDs was finally selected for machine learning of pesticide recovery prediction shown in the Table 4.6.

Table 4.5. Molecular Descriptors selected by the correlation analysis and cluster analysis (k=14)

MD	Cluster	Integer or Real number	Z-score range	Final candidates	Final Selection
BCUTw.1h	1	R	5.405	x	x
khs.sBr	1	I	6.687		
WTPT.1	2	R	6.048		
SP.1	2	R	6.073	x	x
nAtom	3	I	5.018		
fragC	3	R	5.11	x	
nAromBond	4	I	4.652	x	x
naAromAtom	4	I	4.497		
ALogP	5	R	6.117		
ALogp2	5	R	6.635	x	x
nAcid	6	I	15.748		
topoShape	6	R	15.758	x	x
ECCEN	6	R	15.748		
tpsaEfficiency	7	R	5.072		
TopoPSA	7	R	5.969	x	x
MDEN.23	8	R	5.726	x	x
khs.aasN	8	I	4.63		
MDEC.11	9	R	8.562	x	x
khs.sCH3	9	I	5.191		
MDEC.34	10	R	5.928	x	
khs.ssssC	10	I	4.529		
khs.ssNH	11	I	3.966	x	x
VP.5	12	R	8.884		
VP.6	12	R	10.767	x	x
SPC.5	13	R	6.388		
SPC.6	13	R	6.82	x	x
VC.3	14	R	5.746		
VC.4	14	R	6.061	x	x

Table 4.6. Summary of molecular descriptors selected by the correlation analysis and cluster analysis (k=14)

MD group	Description of MDs	Number of MDs	Selected
MD-r3a	MD of $r < 0.7$ with any of other 177 MDs	60	Yes
MD-r3b	MD of $r \geq 0.7$ with any of other 177 MDs and selected by graph-clustering method	12	Yes
MD-r3c	MD of $r \geq 0.7$ with any of other 177 MDs and excluded by graph-clustering method	106	No

Thus, MD-r3a and MD-r3b (both MDs are combined as the MD group of “MD4”) will be used for the regression analysis by caret. Here, all 178 MDs(original



Table 4.7. Molecular Descriptors selected by the correlation analysis and cluster analysis (k=14)

MD	Cluster	Integer or Real number	Z-score range	Final candidates	Final Selection
BCUTw.1h	1	R	5.405	x	x
BCUTw.1h	1	R	5.405	x	x
khs.sBr	1	I	6.687		
WTPT.1	2	R	6.048		
SP.1	2	R	6.073	x	x
nAtom	3	I	5.018		
fragC	3	R	5.11	x	
nAromBond	4	I	4.652	x	x
naAromAtom	4	I	4.497		
ALogP	5	R	6.117		
ALogp2	5	R	6.635	x	x
nAcid	6	I	15.748		
topoShape	6	R	15.758	x	x
ECCEN	6	R	15.748		
tpsaEfficiency	7	R	5.072		
TopoPSA	7	R	5.969	x	x
MDEN.23	8	R	5.726	x	x
khs.aasN	8	I	4.63		
MDEC.11	9	R	8.562	x	x
khs.sCH3	9	I	5.191		
MDEC.34	10	R	5.928	x	
khs.ssssC	10	I	4.529		
khs.ssNH	11	I	3.966	x	x
VP.5	12	R	8.884		
VP.6	12	R	10.767	x	x
SPC.5	13	R	6.388		
SPC.6	13	R	6.82	x	x
VC.3	14	R	5.746		
VC.4	14	R	6.061	x	x

Most other machine learning methods were below -1 in Log(PE), poor performance in prediction. Tuning parameters of xgbLinear for MD4 are listed in the Table 4.8. The parameters were the same for all seven crops.

Table 4.8. Tuning parameters of xgbLinear for pesticide recovery prediction (MD4).

Parameter	Value
Number of rounds (Boosting Iterations)	50
Lambda (L2 Regularization)	0.0001
Alpha (L1 Regularization)	0.0001
Eta (Learning Rate)	0.3

**Comparison of PE among MD4, MD2 and MD0** Figure 4.12 and Figure 4.13 show the comparison of Log(PE) among MD4, MD2 and MD0. There was no difference in Log(PE) between MD4 and MD2 for all 89 machine learning methods. SBC and ppr got worse in Log(PE) with MD4, which was caused by the reduction of MDs, i.e. explanatory of variables for prediction. xgbLinear performed pretty well in MD4, in addition to MD2 and MD0.



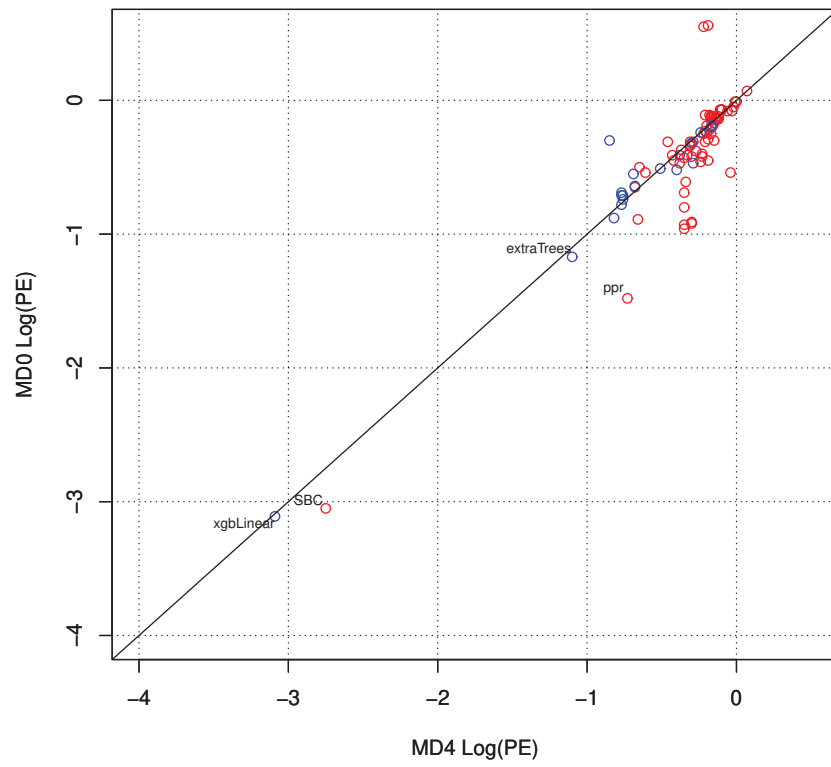


Figure 4.12. Log(PE) comparison between MD4 and MD0

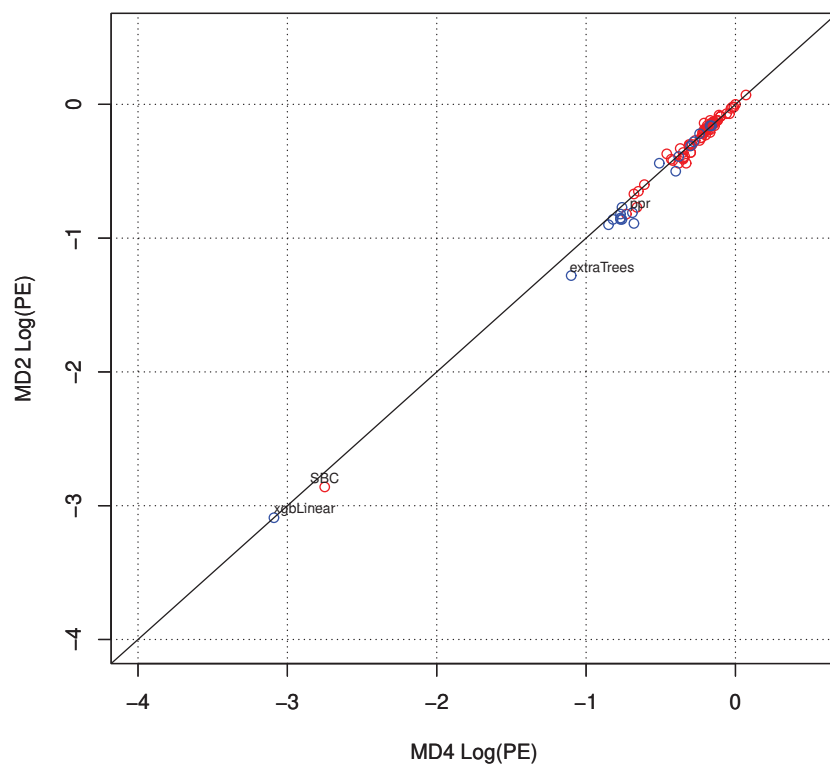


Figure 4.13. Log(PE) comparison between MD4 and MD2

**Selection of MDs from the dendrogram at k=3 (MD6)** 3 MDs from 87 MDs were selected as shown in the Table 4.9 for k=3. These MDs correlate weakly, so all of them were selected for machine learning of pesticide recovery prediction.

Table 4.9. Molecular Descriptors selected by the correlation analysis and cluster analysis (k=3)

MD	Cluster	Integer or Real number	Z-score range	Final candidates	Final Selection
VP.6	1	R	10.767	x	x
WTPT.1	2	R	6.048		
SP.1	2	R	6.073	x	x
nAcid	3	I	15.748		
topoShape	3	R	15.758	x	x
ECCEN	3	R	15.748		

Table 4.10 is the summary of MDs selected by the decision tree Figure 3.5. 60 MDs of MD-r5a and 3 MDs of MD-r5b were combined as the MD6 for regression to predict the pesticide recovery.

Table 4.10. Summary of molecular descriptors selected by the correlation analysis and cluster analysis (k=3)

MD group	Description of MDs	Number of MDs	Selected
MD-r5a	MD of $r < 0.7$ with any of other 177 MDs	60	Yes
MD-r5b	MD of $r \geq 0.7$ with any of other 177 MDs and selected by hierarchical cluster analysis	3	Yes
MD-r5c	MD of $r \geq 0.7$ with any of other 177 MDs and excluded by hierarchical cluster analysis	115	No

**PE and ET result in MD6** Figure 4.14 shows the plot of Log(PE) and Log(ET) of 89 machine learning methods with MD6. xgbLinear and SBC performed well in prediction of pesticide recovery rate as well as MD4. Most other machine learning methods were below -1 in Log(PE), poor performance in prediction.

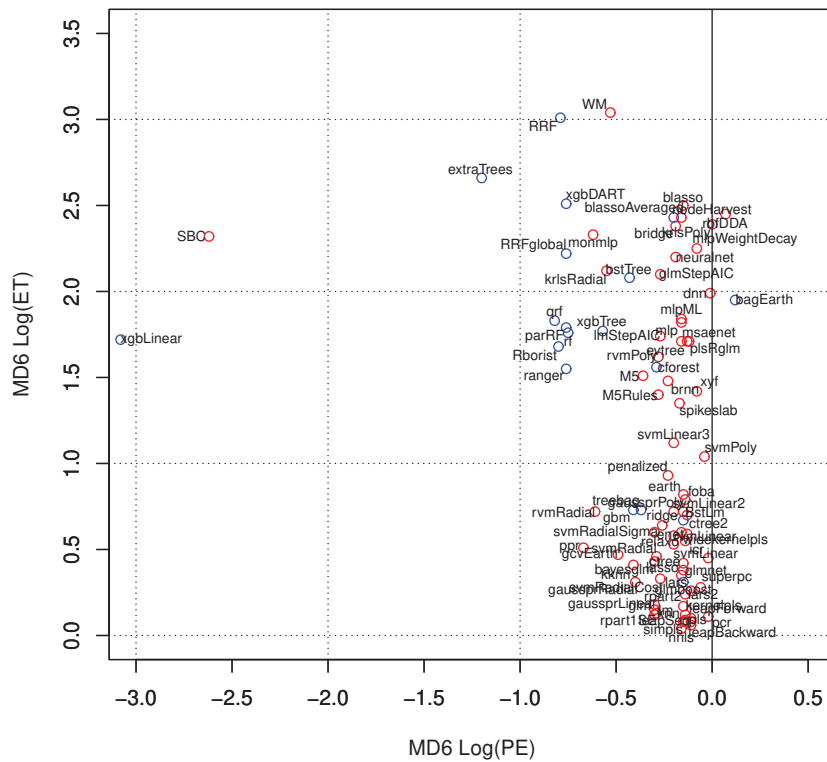


Figure 4.14. Log(PE) and Log(ET) for 89 methods using the MD6

Tuning parameters of xgbLinear for MD6 are listed in the Table 4.11. The parameters were the same for all seven crops.

Table 4.11. Tuning parameters of xgbLinear for pesticide recovery prediction (MD6).

Parameter	Value
Number of rounds (Boosting iterations)	50
Lambda (L2 Regularization)	0.1
Alpha (L1 Regularization)	0.1
Eta (Learning Rate)	0.3

**Comparison of machine learning performance among MD6, MD4, MD2 and MD0** Figure 4.15, 4.16 and Figure 4.17 shows the comparison of  $\text{Log(PE)}$  among MD6, MD4, MD2 and MD0. There was no difference in  $\text{Log(PE)}$  among MD6, MD4 and MD2 for all 89 machine learning methods. SBC and ppr (Projection pursuit regression) got worse in  $\text{Log(PE)}$  with MD6 as well as MD4, which was caused by the reduction of MDs, i.e. explanatory variables for prediction. xgbLinear performed pretty well in MD4, in addition to MD2 and MD0.

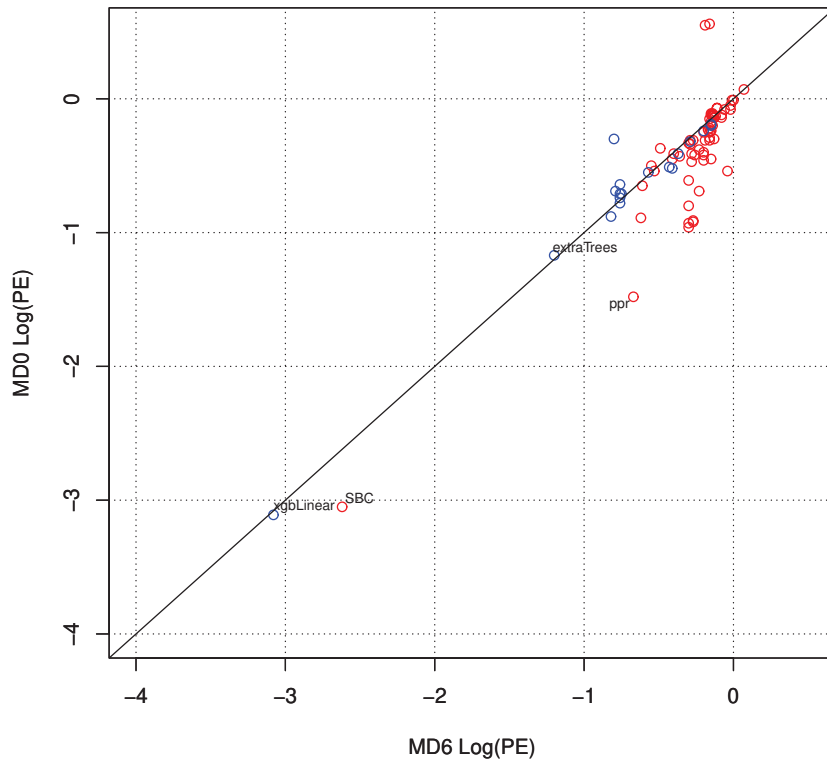


Figure 4.15.  $\text{Log(PE)}$  comparison between MD6 and MD0

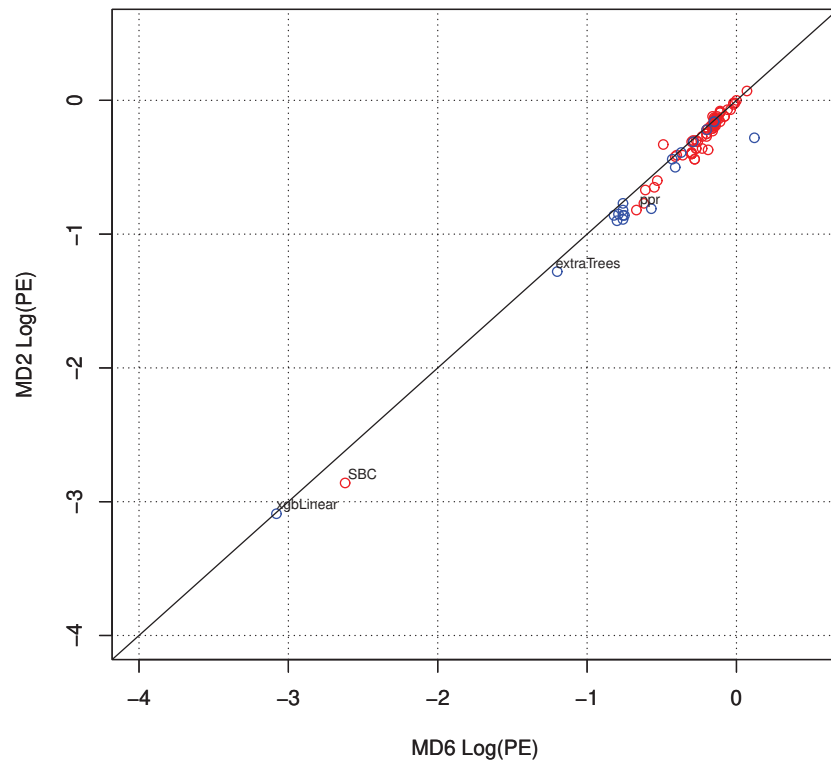


Figure 4.16. Log(PE) comparison between MD6 and MD2

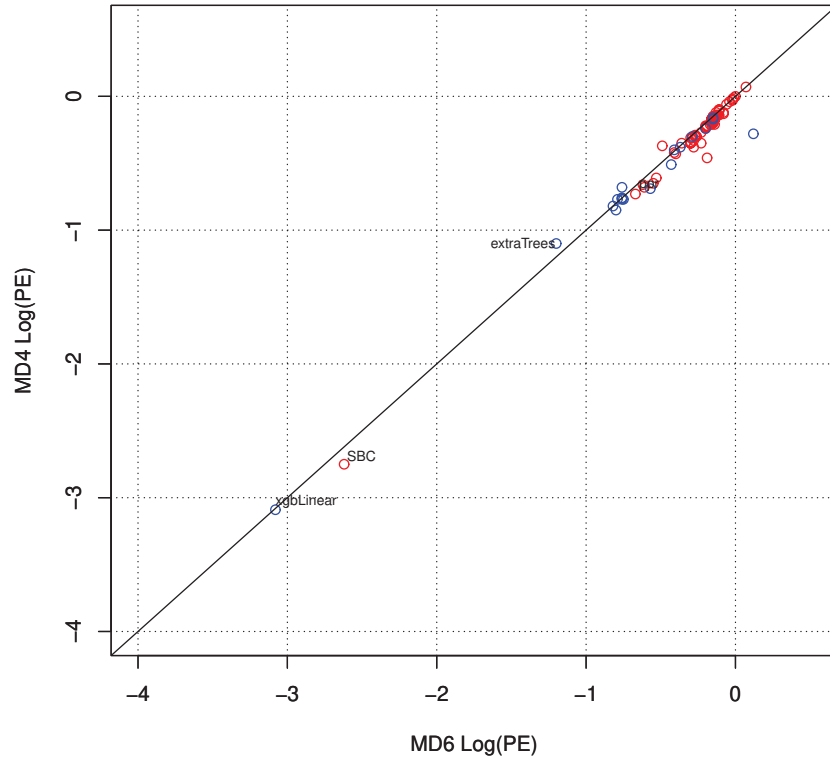


Figure 4.17. Log(PE) comparison between MD6 and MD4

#### 4.3.3 Comparison of PE among MD6, MD4, MD2 and MD0 for machine learning method category

**Comparison of PE** Figure 4.18 shows the box plots of Log(PE) for machine learning methods compared among MD6, MD4, MD2 and MD0. Results of PE comparison among MD6, MD4, MD2 and MD0 by the machine learning method category are summarized in the Table 4.12.

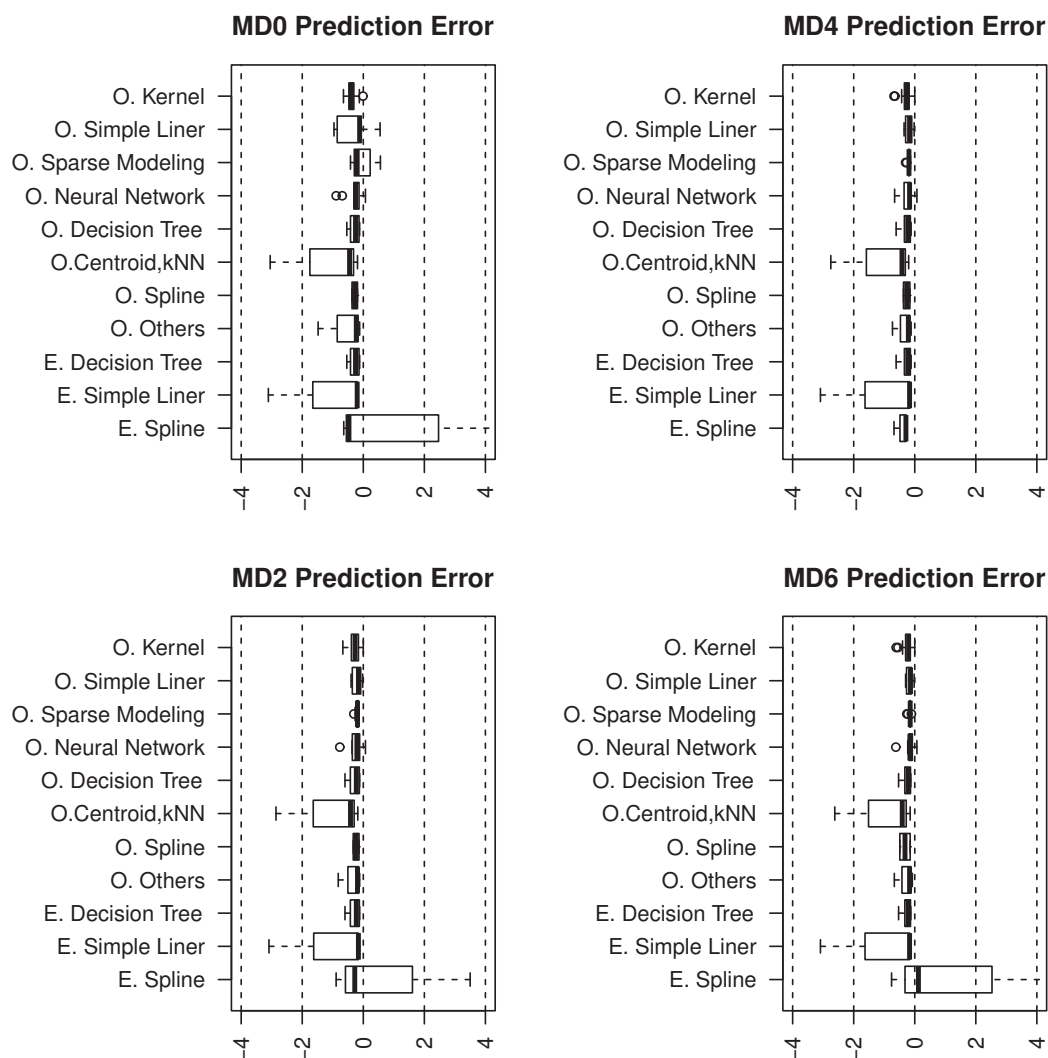


Figure 4.18. Comparison of Log(PE) comparison by learning method category among MD6, MD4, MD2 and MD0



Table 4.12. Results of machine learning method category among the MD groups

Results	Machine learning method category
Poor performance in PE on all MDs	O. Kernel, O. Simple Linear, O. Sparse Modeling, O. Neural Network, O. Decision Tree, O. Spline, O. Others, E. Decision Tree and E. Spline
Good performance in MD0, worse by reduction of MDs	O. Simple Linear
Some method(s) performed well but others were poor in all MDs	O. centroid, kNN and E. Simple Liner

**Comparison of ET among MD6, MD4, MD2 and MD0** Figure 4.19 shows the ET of SBC and xgbLinear for MD6, MD4, MD2 and MD0 with 5 replicates. ET of SBC is not changed by reduction of MDs but ET of xgbLinear reduces the ET by reduction of MDs.

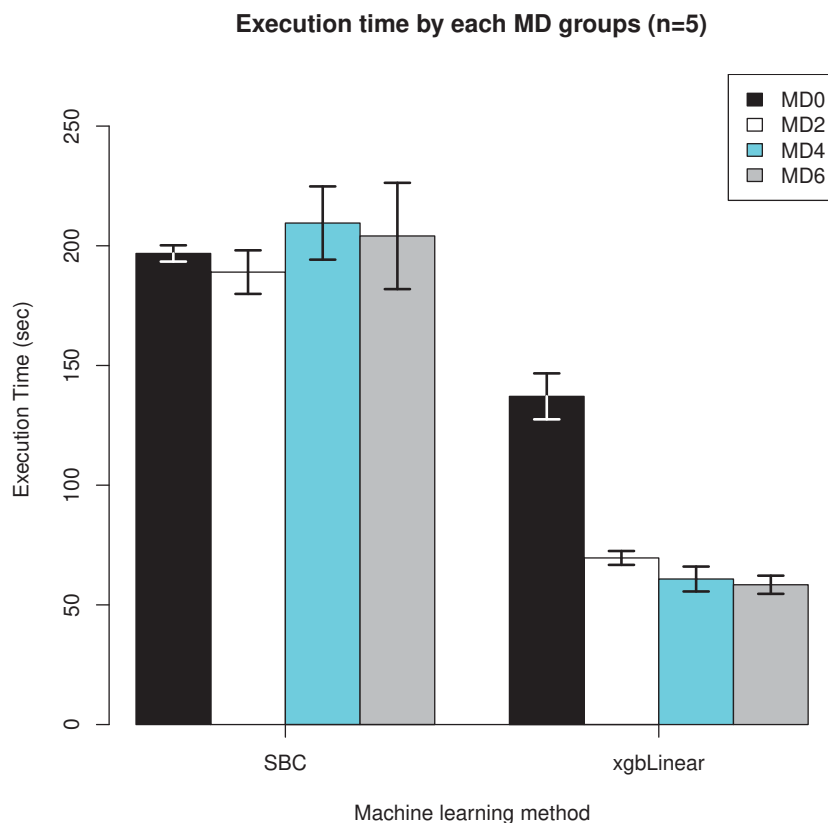


Figure 4.19. ET of SBC and xgbLinear for MD6, MD4, MD2 and MD0 with 5 replicates

#### 4.4 Conclusion of Chapter 4 and pesticide recovery prediction by regression model

Considering the results, xgbLinear (eXtremely Gradient Boosting Linear) performs excellently for prediction of pesticide recovery rate for this data set. Execution time of building the prediction model of xgbLinear is reduced when MDs are reduced while maintaining the performance of PE for prediction of pesticide recovery. Optimum condition will be using xgbLinear with the selection of MDs by hierarchical cluster analysis. In Chapter 4, hierarchical cluster analysis was

used to select the molecular descriptors for prediction model of pesticide recovery.

For prediction of pesticide recovery for this data set, xgbLinear with the hierarchical cluster analysis with  $k=3$  is the optimum condition.

## **5. Classification of pesticides amenability between LC or GC**

### **5.1 Introduction**

One of the frequently asked questions by food analysis chemists who are currently using GC-MS or LC-MS is whether that pesticide is “GC-amenable” or “LC-amenable”. This is because neither LC-MS nor GC-MS can analyze all pesticides by any single technology, comparisons of the pesticides has been researched [11, 12, 35]. There are several guidelines for the selection between LC-amenable and GC-amenable for pesticides based on the physical and chemical properties [36], and experienced chemists can predict the answer to this question based on the experiences for some degree. In this chapter, I build the prediction model for answering to this question of the pesticide amenability between LC- and GC-, by the classification model using the molecular descriptors(MDs) using validated pesticide literatures [37, 38].

### **5.2 Materials and Methods**

#### **5.2.1 Preparation of the pesticide list**

Pesticide information for classification model were obtained from two validation reports of residual pesticide analysis in foods [37, 38]. Details of the pesticides and technologies are listed in the Table B.1 in the Appendix.

##### **U.S. Food and Drug Administration(FDA) List [37]**

The validation report of 136 pesticides analysis in Avocado using both LC-MS and GC-MS.

##### **EU Reference Laboratories for Residues of Pesticides(EURL)[38]**

The validation report of 127 pesticides analysis in Olive Oil using both LC-MS and GC-MS.

202 pesticides in total were included on both literatures. For improving the classification capability of machine learning, eight pesticides were excluded from

the machine learning which were analyzed differently between both, i.e. by GC-MS in EURL while by LC-MS in FDA as shown in Figure 5.1. 194 pesticides were used for building the classification model by machine learning.

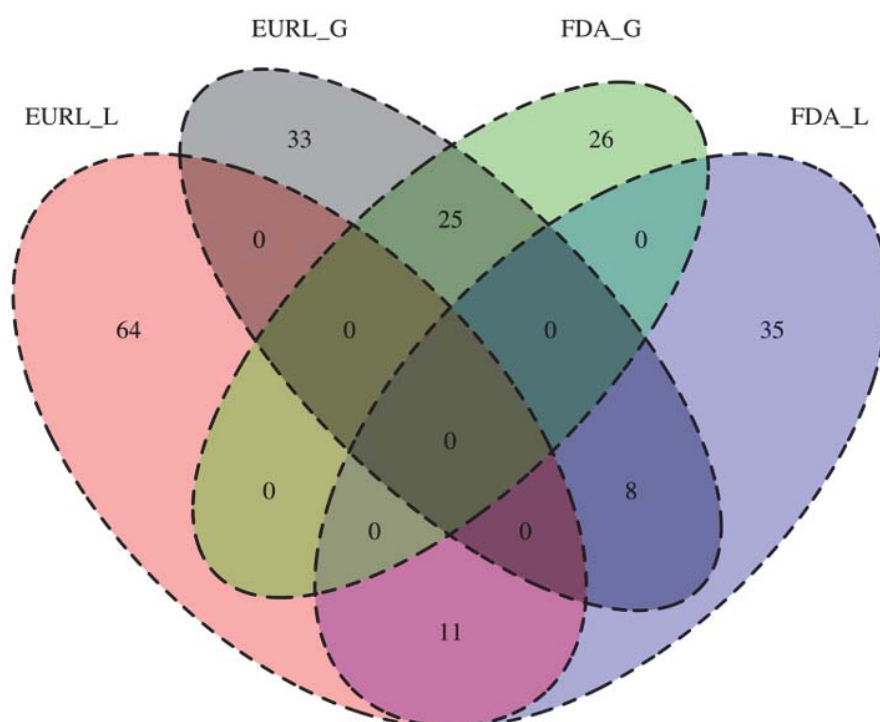


Figure 5.1. Venn diagram to describe the number of pesticides by the list(FDA or EURL) and the technology used for analysis(L:LC-MS and G:GC-MS).

### 5.2.2 Molecular descriptors of the pesticides

The canonical SMILES of 194 pesticides were obtained from the PubChem website as listed in the Table B.1. 224 molecular descriptors(MDs) of these pesticides were obtained by rcdk package of R program. The MDs with the zero variance

among 194 pesticides were removed in order to avoid the error in machine learning, 176 MDs were eventually obtained as Table 5.1. Each molecular descriptor was standardized for comparable as expressed by the Equation 5.1 (z-score), where  $z_i$  is the standardized value to be used for machine learning,  $x_i$  is the raw value from rcdk,  $\mu_i$  is the average of 194 pesticides and  $\sigma_i$  is the standard deviation of 194 pesticides for  $i$ th molecular descriptor.

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \quad (5.1)$$

Table 5.1: Summary of molecular descriptors obtained by SMILES for chemicals

Descriptor Class	Descriptor (Description)
ALOGP Descriptor (2)	ALogP (Ghose-Crippen LogKow), ALogP2 (Square of ALogP)
APol Descriptor (1)	Apol (Sum of the atomic polarizabilities (including implicit hydrogens))
Aromatic Atoms Count Descriptor (1)	naAromAtom (Number of aromatic atoms)
Aromatic Bonds Count Descriptor (1)	nAromBond (Number of aromatic bonds)
Atom Count Descriptor (2)	nAtom (Number of atoms), nB (Number of binding)
Autocorrelation Descriptor Charge (5)	ATSc1, ATSc2, ATSc3, ATSc4, ATSc5 (ATS autocorrelation descriptor, weighted by charges)
Autocorrelation Descriptor Mass (5)	ATSm1, ATSm2, ATSm3, ATSm4, ATSm5 (ATS autocorrelation descriptor, weighted by scaled atomic mass)
Autocorrelation Descriptor Polarizability (5)	ATSp1, ATSp2, ATSp3, ATSp4, ATSp5 (ATS autocorrelation descriptor, weighted by polarizability)
BCUT Descriptor (6)	BCUTw.1l (nhigh lowest atom weighted BCUTS), BCUTw.1h (nlow highest atom), BCUTc.1l (nhigh lowest partial charge), BCUTc.1h (nlow highest partial charge) BCUTp.1l (nhigh lowest polarizability), BCUTp.1h (nlow highest polarizability)
BPolDescriptor (1)	bpol (Sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule (including implicit hydrogens))
Carbon Types Descriptor (9)	C1SP1 (Triply bound carbon bound to one other carbon), C2SP1 (Triply bound carbon bound to two other carbons), C1SP2 (Doubly bound carbon bound to one other carbon), C2SP2 (Doubly bound carbon bound to two other carbons), C3SP2 (Doubly bound carbon bound to three other carbons), C1SP3 (Singly bound carbon bound to one other carbon), C2SP3 (Singly bound carbon bound to two other carbons), C3SP3 (Singly bound carbon bound to three other carbons), C4SP3 (Singly bound carbon bound to four other carbons)
Chi Chain Descriptor (10)	SCH.3-7 (Simple chain, orders 3-7), VCH.3-7 (Valence chain, orders 3-7)
Chi Cluster Descriptor (8)	SC.3-6 (Simple cluster, orders 3-6), VC.3-6 (Valence cluster, orders 3-6)
Chi Path Cluster Descriptor (6)	SPC.4-6 (Simple path cluster, orders 4 to 6), VPC.4-6 (Valence path cluster, orders 4-6)
Chi Path Descriptor (16)	SP.0-7 (Simple path, orders 0-7), VP.0-7Valence path, orders 0-7)
Eccentric Connectivity Index Descriptor (37)	ECCEN (A topological descriptor combining distance and adjacency information), khs.sCH3 (Count of atom-type E-State: -CH3), khs.dCH2 (=CH2), khs.ssCH2 (-CH2-), khs.tCH (#CH), khs.dsCH (=CH-), khs.aaCH (:CH:), khs.sssCH (>CH-), khs.tsC (#C-), khs.dssC (=C<), khs.aasC (:C:-), khs.aaaC (::C:), khs.ssssC (>C<), khs.sNH2 (-NH2), khs.ssNH (-NH2+), khs.aaNH (:NH:), khs.tN (#N), khs.dsN (=N-), khs.aaN (:N:), khs.sssN (>N-), khs.ddN (-N<<), khs.aasN (:N:-), khs.sOH (-OH), khs.dO (=O), khs.ssO (-O-), khs.aaO (:O:), khs.sF (-F), khs.ssssSi (>Si<), khs.dsssP (->P=), khs.dS (=S), khs.ssS (-S-), khs.aas (aSa), khs.dssS (>S=), khs.ddssS (>S==), khs.sCl (-Cl), khs.sBr (-Br)

*Continue to next page*

*Continue from previous page*

Descriptor Class	Descriptor (Description)
Fragment Complexity Descriptor (1)	fragC (Complexity of a system)
Ghose Crippen Molecular Refractivity Descriptor (1)	AMR (Molar refractivity)
H Bond Acceptor Count Descriptor (1)	nHBAcc (Number of hydrogen bond acceptors)
H Bond Donor Count Descriptor (1)	nHBDon (Number of hydrogen bond donors)
KappaShape Indices Descriptor (3)	Kier1-3 (First, Second, Third kappa ( $\kappa$ ) shape indexes)
Largest Chain Descriptor (1)	nAtomLC (Number of atoms in the largest chain)
Longest Aliphatic Chain Descriptor (1)	nAtomLAC (Number of atoms in the longest aliphatic chain)
Mannhold LogP Descriptor (1)	MLogP (Mannhold LogP)
MDEDescriptor (19)	MDEC.11 (Molecular distance edge between all primary carbons), MDEC.12 (between all primary and secondary carbons), MDEC.13 (between all primary and tertiary carbons), MDEC.14 (between all primary and quaternary carbons), MDEC.22 (between all secondary carbons), MDEC.23 (between all secondary and tertiary carbons), MDEC.24 (between all secondary and quaternary carbons), MDEC.33 (between all tertiary carbons), MDEC.34 (between all tertiary and quaternary carbons), MDEC.44 (between all quaternary carbons), MDEO.11 (between all primary oxygens), MDEO.12 (between all primary and secondary oxygens), MDEO.22 (between all secondary oxygens), MDEN.11 (between all primary nitrogens), MDEN.12 (between all primary and secondary nitrogens), MDEN.13 (between all primary and tertiary nitrogens), MDEN.22 (between all secondary nitrogens), MDEN.23 (between all secondary and tertiary nitrogens), MDEN.33 (between all tertiary nitrogens)
Petitjean Number Descriptor (1)	PetitjeanNumber (Petitjean number)
Rotatable Bonds Count Descriptor (1)	nRotB (Number of rotatable bonds, excluding terminal bonds)
Rule Of Five Descriptor (1)	LipinskiFailures (Number failures of the Lipinski's Rule Of 5)
TPSA Descriptor (19)	TopoPSA (Topological polar surface area)
VAdjMat Descriptor (1)	VAdjMat (Vertex adjacency information (magnitude))
Weight Descriptor (1)	MW (Molecular weight)
Weighted Path Descriptor (5)	WTPT.1 (Molecular ID), WTPT.2 (Molecular ID / number of atoms), WTPT.3 (Sum of path lengths starting from heteroatoms), WTPT.4 (Sum of path lengths starting from oxygens), WTPT.5 (Sum of path lengths starting from nitrogens)
Wiener Numbers Descriptor (2)	WPATH (Weiner path number), WPOL (Weiner polarity number)
XLogP Descriptor (1)	XLogP (XLogP)
Zagreb Index Descriptor (1)	Zagreb (Sum of the squares of atom degree over all heavy atoms i)
Petitjean Shape Index Descriptor (1)	topoShape (Petitjean topological shape index)
Others (16)	nBase (Basic group count descriptor), nSmallRings (the number of small rings from size 3 to 9), nAromRings (the number of aromatic rings), nRingBlocks (total number of distinct ring blocks), nAromBlocks (total number of "aromatically connected components"), nRings3, 5, 6, 7 (individual breakdown of small rings), tpsaEfficiency (Polar surface area expressed as a ratio to molecular size), VABC (Atomic and Bond Contributions of van der Waals volume), HybRatio (the ratio of heavy atoms in the framework to the total number of heavy atoms in the molecule.), tpsaEfficiency.1 (Polar surface area expressed as a ratio to molecular size), TopoPSA.1 (Topological polar surface area), topoShape.1 (A measure of the anisotropy in a molecule), FMF (FMF descriptor characterizing complexity of a molecule)

*End of table*

### 5.2.3 Classification of pesticides by the machine learning

Either G(GC-MS) or L(LC-MS) of technology flag is assigned on each 194 pesticides based on the literatures as Table B.1. 119 machine learning methods of the classification in caret package listed in the Table 5.2 [19] are evaluated in the present study. These machine learning methods are expected to classify the 194 pesticides between G(GC-MS) or L(LC-MS) using the 176 molecular descriptors. Classification algorithms can be categorized into two categories, i.e. (1) ordinary learning approaches which construct one learner from training data and (2) ensemble methods which construct a set of learners and combine them. Machine learning is implemented using the caret package of R program. In the present study, 84 ordinary and 35 ensemble methods. Most of ordinary learning methods correspond to classification models with kernel and decision tree models. In classification models with kernels, `dwdPoly` (Distance Weighted Discrimination with Polynomial Kernel), `dwdRadial` (Distance Weighted Discrimination with Radial Basis Function Kernel), `gaussprRadial` (Gaussian Process with Radial Basis Function Kernel), `kernelpls` (Partial Least Squares), `lssvmRadial` (Least Squares Support Vector Machine with Radial Basis Function Kernel), `stepQDA` (Quadratic Discriminant Analysis with Stepwise Feature Selection), `svmLinear` (Least Squares Support Vector Machine), `svmLinear2` (Support Vector Machines with Linear Kernel), `svmLinear3` (L2 Regularized Support Vector Machine (dual) with Linear Kernel), `svmLinearWeights` (Linear Support Vector Machines with Class Weights), `svmLinearWeights2` (L2 Regularized Linear Support Vector Machines with Class Weights), `svmPoly` (Support Vector Machines with Polynomial Kernel), `svmRadial`, `svmRadialCost`, `svmRadialSigma` (Support Vector Machines with Radial Basis Function Kernel), `svmRadialWeights` (Support Vector Machines with Class Weights) and `widekernelpls` (Partial Least Squares) implements polynomial kernels. Ordinary decision trees have been developed on the basis of C5.0 classification models (C5.0, C5.0Cost, C5.0Rules and C5.0Tree), recursive partitioning using `rpart` package (`rpart`, `rpart1SE`, `rpart2`, `rpartCost`, `rpartScore`), conditional inference trees by recursively making binary splittings on the variables with the highest association to the class (`ctree` and `ctree2`), `evtree` (Tree Models from Genetic Algorithms), `J48` (C4.5-like Trees), `JRip` (Rule-Based Classifier), `LMT` (Logistic Model



Trees), OneR (Single Rule Classification) and PART(Rule-Based Classifier) for classification. Ordinary simple linear models, neural network, spline and other models are also available for classification of the pesticides which were available in caret package. In contrast, various ensemble decision trees (ada(Boosted Classification Trees), AdaBag(Bagged AdaBoost), adaboost(AdaBoost Classification Trees), AdaBoost.M1, blackboost(Boosted Tree), bstTree(Boosted Tree), cforest(Conditional Inference Random Forest), deepboost, extraTrees (Random Forest by Randomization), gbm (Stochastic Gradient Boosting), nodeHarvest (Tree-Based Ensembles), ORFpls (Oblique Random Forest with pls), ORFridge (Oblique Random Forest with ridge), ORFsvm (Oblique Random Forest with Support Vector Machine), parRF (Parallel Random Forest), ranger (Random Forest), Rborist(Random Forest), rf (Random Forest), rFerns (Random Ferns), rfRules (Random Forest Rule-Based Model), rotationForest (Rotation Forest), rotationForestCp (Rotation Forest), RRF (Regularized Random Forest), RRF-global (Regularized Random Forest), treebag (Bagged CART), wsrf (Weighted Subspace Random Forest), xgbDART (eXtreme Gradient Boosting with Dropout and Additive Regression Tree), xgbTree(eXtreme Gradient Boosting with Tree)) are also available in caret package for classification of pesticides on this data set. We could compare the performance of the classification on these machine learning methods. Prediction performance of classification is measured by the accuracy of resamples from the 10-fold cross-validation(CV10) iterations [39] and execution time. Execution time is obtained by the “system.time()” function of R program.

Table 5.2. Classification methods in caret evaluated in this study

Algorithm	Methods in caret
(a) Ordinary learning methods	
Kernel (17)	dwdPoly, dwdRadial, gaussprRadial, kernelps, lssvmRadial, stepQDA, svmLinear, svmLinear2, svmLinear3, svmLinearWeights, svmLinearWeights2, svmPoly, svmRadial, svmRadialCost, svmRadialSigma, svmRadialWeights, widekernelps
Simple Linear (12)	bayesglm, CSimca, glm, glmStepAIC, multinom, ordinalNet, plr, pls, regLogistic, rrla, RSimca, simpls
Sparse modeling (2)	glmnet, sdwd
Neural Network (11)	avNNNet, dnn, mlp, mlpML, mlpWeightDecay, mlpWeightDecayML, mon-mlp, msaenet, nnet, pcaNNNet, rbfDDA
Decision Tree (17)	C5.0, C5.0Cost, C5.0Rules, C5.0Tree, ctree, ctree2, evtree, J48, JRip, LMT, OneR, PART, rpart, rpart1SE, rpart2, rpartCost, rpartScore
Centroid, kNN (6)	kknn, knn, lvq, ownn, pam, snn
Spline (4)	earth, gamLoess, gamSpline, gcvEarth
Naive Bayes (2)	naive bayes, nb
Others (13)	dwdLinear, fda, hdda, null, pda, pda2, rda, rocc, sda, slda, sparseLDA, stepLDA, xyf
(b) Ensemble learning methods	
Decision Tree (28)	ada, AdaBag, adaboost, AdaBoost.M1, blackboost, bstTree, cforest, deepboost, extraTrees, gbm, nodeHarvest, ORFpls, ORFridge, ORFsvm, parRF, ranger, Rborist, rf, rFerns, rfRules, rotationForest, rotationForestCp, RRF, RRFglobal, treebag, wsrf, xgbDART, xgbTree
Simple Linear (4)	BstLm, glmboost, LogitBoost, xgbLinear
Spline (3)	bagEarth, bagEarthGCV, bagFDA

### 5.3 Results and Discussion

Results of CV10 accuracy and Execution Time for 119 machine learning methods are listed in the Table B.2.

#### 5.3.1 Classification Performance (Accuracy of CV10 resample)

The box plot in the Figure 5.2 shows the distribution of accuracy for each machine learning method category. The overall accuracy of CV10 is calculated by the Eq. 5.2.

$$CV10\ Accuracy(\%) = \frac{\sum_{i=1}^{10} (Accuracy\ of\ ith\ test)}{Number\ of\ test(10)} \times 100 \quad (5.2)$$

Overall accuracy across the 119 methods was 77%, ranged between 60% and 85% as shown in Figure 4.5. Machine learning methods in the ensemble spline method category show larger variability in accuracy than the others. Three machine learning methods, bagEarth(Bagging Earth, 27%), bagEarthGCV(Bagging Earth generalized cross validation, 16%) and bagFDA(Bagging flexible discriminant analysis, 81%) were included on this category. According to this result, two methods of bagging earth were not suitable in classifications for this data set.

#### 5.3.2 Execution time (ET)

The result of ET of each the machine learning method is shown in the Figure 5.3. Methods of ordinary neural network(ranged LogET 1.08 to 2.36) and ensemble spline categories(LogET 1.63 to 2.71) require more execution time than the other categories. The machine learning method with the maximum ET is glmStepAIC (Generalized Linear Model with Stepwise Feature Selection) with the LogET 4.12, i.e. 3 hours and 41 minutes.

#### 5.3.3 Total performance - both Accuracy and Execution Time

The results of Accuracy and ET by the machine learning method are shown in the Figure 5.4 and Figure 5.5.

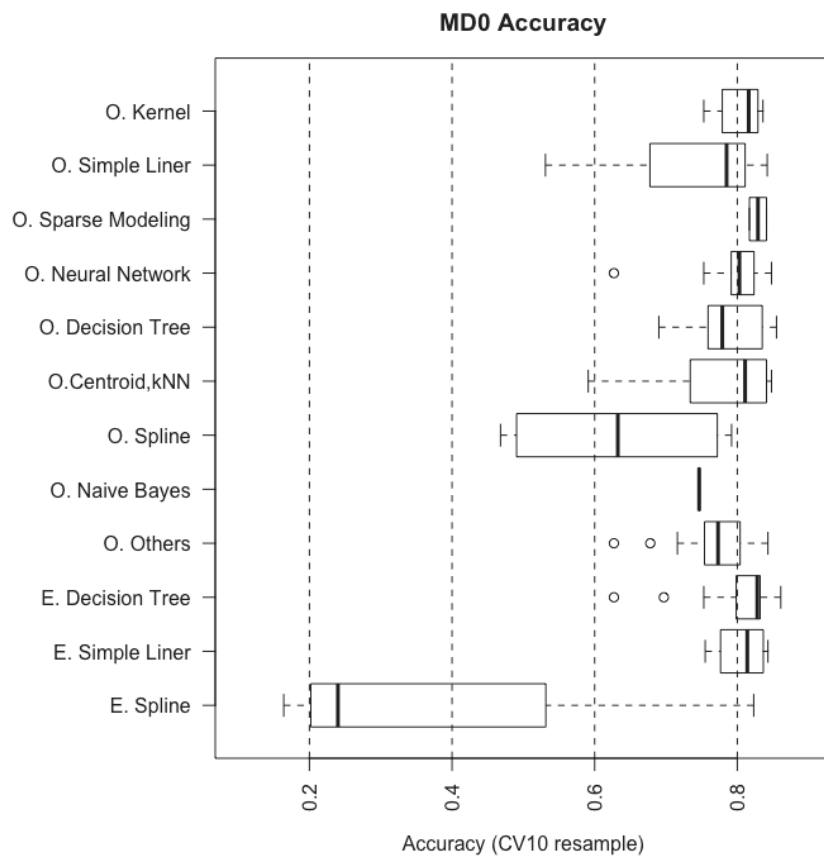


Figure 5.2. Accuracy of classification(CV10 resample) for 119 machine learning methods

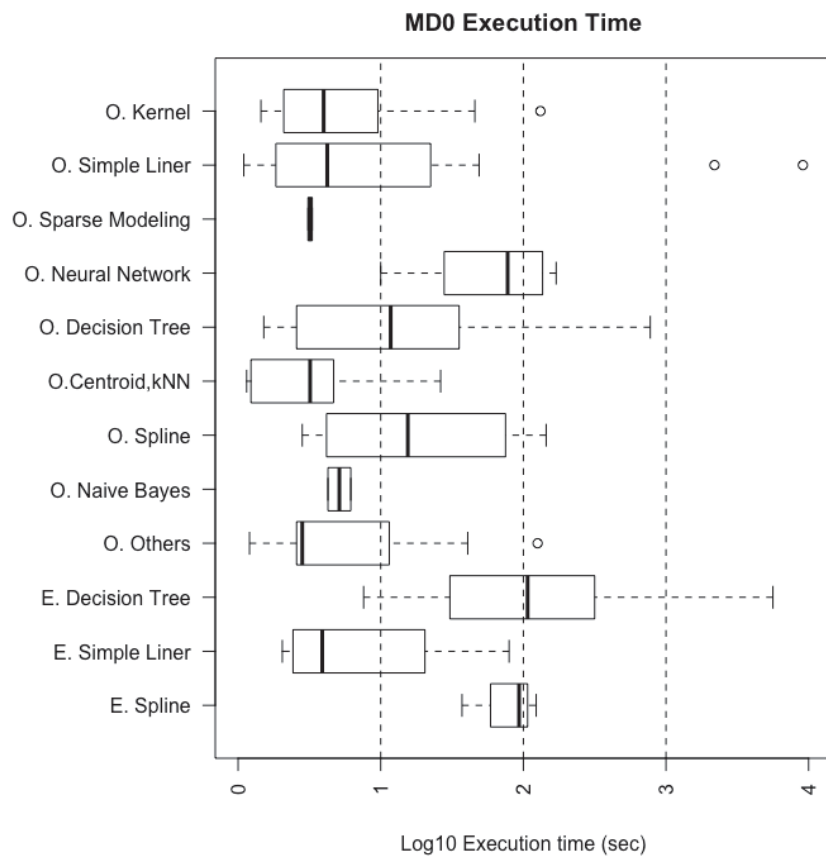


Figure 5.3. Execution time for 119 machine learning methods

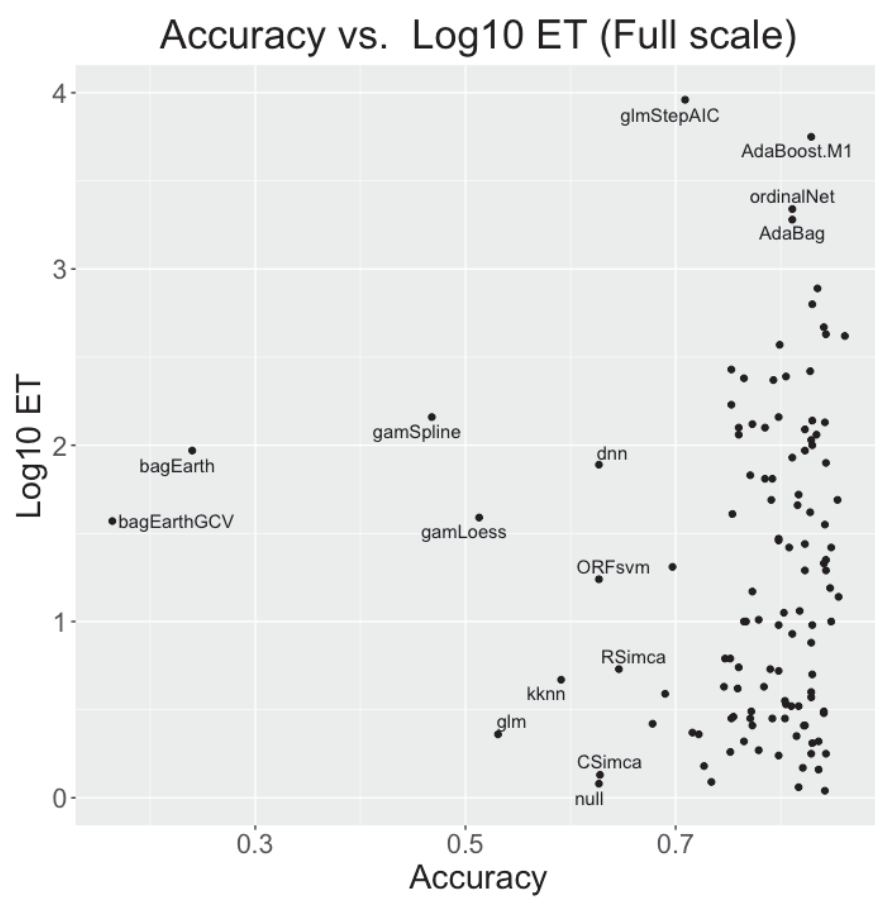


Figure 5.4. Accuracy and Execution time for 119 machine learning methods in full scale

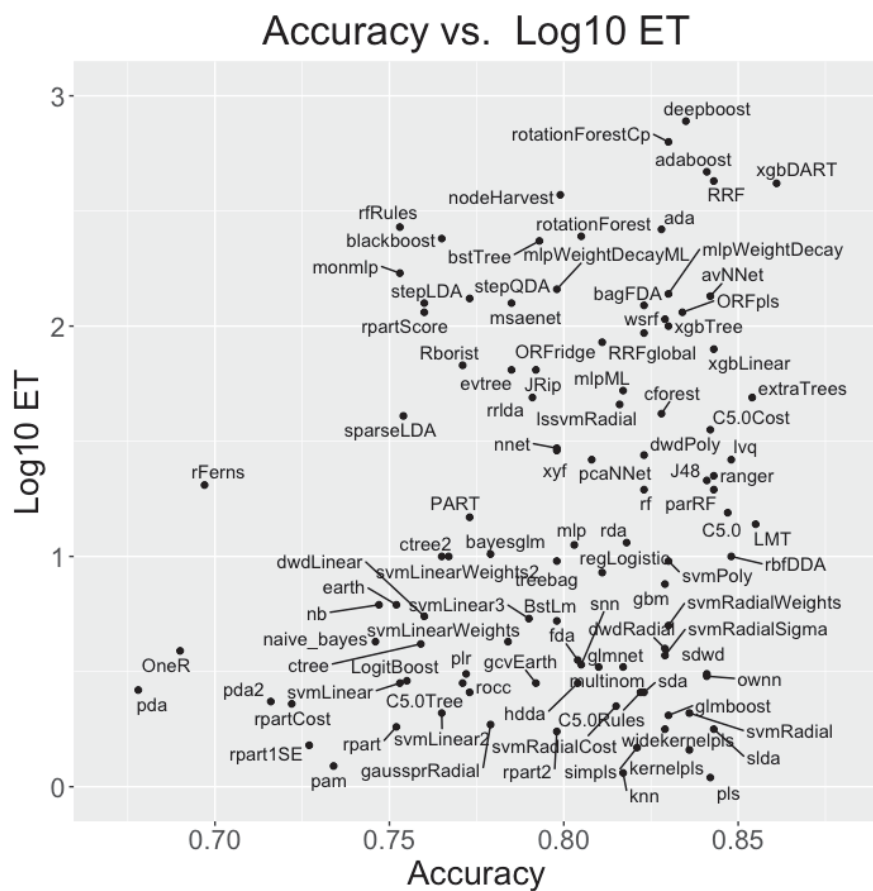


Figure 5.5. Accuracy and Execution time in expanded view

Table 5.3 shows top 20 methods of accuracy. Accuracy of the classification of the amenability between LC-MS and GC-MS ranged from 85.0%(xgbDART) to 83.0%(svmPoly) in the Table 5.3. Six methods of Ensemble Decision Tree showed higher accuracy for the present data set of GC-MS and LC-MS amenability. The best machine learning method of accuracy is xgbDART of Ensemble Decision Tree (85.0% accuracy with 10 minutes 46 seconds), xgbTree of Ensemble Decision Tree(85.0% of accuracy and 2 minutes of ET) and extraTrees of Ensemble Decision Tree(84.6% in accuracy with 1 minutes and 10 seconds of ET) were higher accuracy with the shorter Execution time. I suggest using xgbDART with higher accuracy and reasonable execution time for classification and if longer time is not acceptable, xgbTree and extraTrees will be the second choice for this data set when using all 176 molecular descriptors.

Table 5.3. Top 20 method for classification of pesticides sorted by accuracy

Method in caret	Category	Accuracy	LogET
xgbDART	E. Decision Tree	0.850	2.81
xgbTree	E. Decision Tree	0.850	2.07
extraTrees	E. Decision Tree	0.846	1.83
AdaBoost.M1	E. Decision Tree	0.845	3.84
C5.0	O. Decision Tree	0.841	1.58
C5.0Cost	O. Decision Tree	0.841	1.94
msaenet	O. Neural Network	0.840	2.25
ORFsvm	E. Decision Tree	0.840	2.95
gbm	E. Decision Tree	0.840	1.06
adaboost	E. Decision Tree	0.839	2.93
wsrf	E. Decision Tree	0.839	2.28
glmboost	E. Simple Liner	0.835	0.38
ordinalNet	O. Simple Liner	0.835	3.34
regLogistic	O. Simple Liner	0.835	1.06
dwdRadial	O. Kernel	0.835	0.78
svmRadialSigma	O. Kernel	0.835	0.87
xyf	O. Others	0.830	1.59
glmnet	O. Sparse Modeling	0.830	0.53
ownn	O. Centroid, kNN	0.830	0.54
svmPoly	O. Kernel	0.830	1.06

Tuning parameters of xgbDART are listed in the Table 5.4.



Table 5.4. Tuning parameters of xgbDART for pesticide classification prediction.

Parameter	Value
Boosting Iterations	50
Maximum tree depth	2
Eta (Shrinkage)	0.4
Gamma (Minimum loss Reduction)	0
Subsample (Subsample percentage)	1
Colsample by tree (Subsample ratio of columns)	0.6
Rate drop (Fraction of trees dropped)	0.5
Skip Drop (Probability of skipping drop-out)	0.95
Minimum child weight (Minimum sum of instance weight)	1

#### 5.3.4 Pesticides which were not accurately predicted by the xgbDART

35 pesticides out of 194 pesticides were not accurately predicted by the best machine learning method (xgbLinear) with 10-fold cross validation using 176 MDs. Table 5.5 is the list of 35 pesticides with the characteristics of chemical features. 21 pesticides include the nitrogen-heterocyclic structure and 12 pesticides include the amide bonds, both include nitrogen that can cause the polarization in the molecules, as shown in the Figure 5.6 to Figure 5.10. Other pesticides also includes the heteroatoms such as sulfur, phosphorous and halogens that can also cause the polarization in the molecule as well as nitrogens. For simple visualization by decision tree to understand the characteristics at MD level, ‘rpart’ package is used for confirmation of the molecular descriptors that classify the pesticides by the decision tree with the minimization of gini impurity for classification with the parameter in Table 5.6.

Table 5.5. Chemical characteristics of wrongly classified 35 pesticides by xgb-DART with 176 MDs

Pesticide	Tech	Nitrogen-heterocyclic compound	Amide bond	Sulfur	Phosphorous	Halogen
Flufenacet	L	x	x	x		x
iprodione	G	x	x			x
Diiflufenican	G	x	x			x
Fluopicolide	L	x	x			x
oxadixyl	G	x	x			
azinphos-methyl	L	x		x	x	
phosmet	L	x		x	x	
Pyrazophos	G	x		x	x	
Etridiazole	G	x		x		x
Pyridaben	G	x		x		x
Bupirimate	G	x		x		
vinclozolin	G	x				x
Tebuconazole	G	x				x
Tetraconazole	G	x				x
fludioxinil	L	x				x
penconazole	L	x				x
Diniconazole	L	x				x
pyriproxifen	G	x				
Mepanipyrim	G	x				
fenpropimorph	L	x				
Fenpyroximate	L	x				
Methidathion	G		x	x	x	
Zoxamide	L		x			x
pronamide	G		x			x
propanil	G		x			x
Benalaxyl	G		x			
Metalaxyl	G		x			
napropamide	G		x			
Ethoprophos	G			x	x	
Isocarbophos	G			x	x	
coumaphos	L			x	x	x
dichlorfluanid	L			x		x
Fenamiphos	L				x	x
Dicloran	G					x
bifenthrin	L					x

Table 5.6. Tuning parameters of ‘rpart’ for decision tree to classify 194 pesticides by 176 MDs

Parameter	value
Method	Class
Minimum Split	4
Complexity parameter	0.001
Split	gini

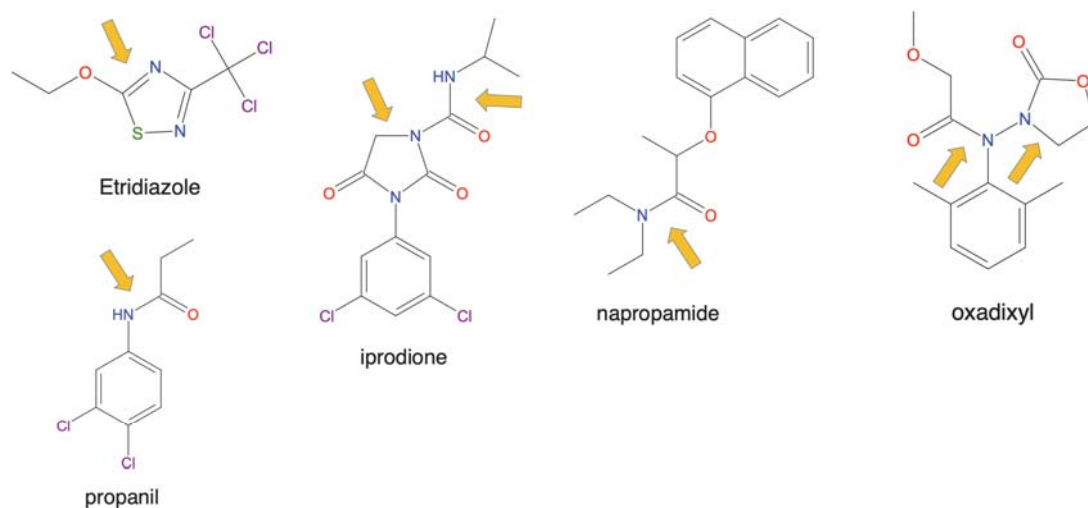


Figure 5.6. Chemical structure of pesticides wrongly classified by xgbDART (1 of 5)

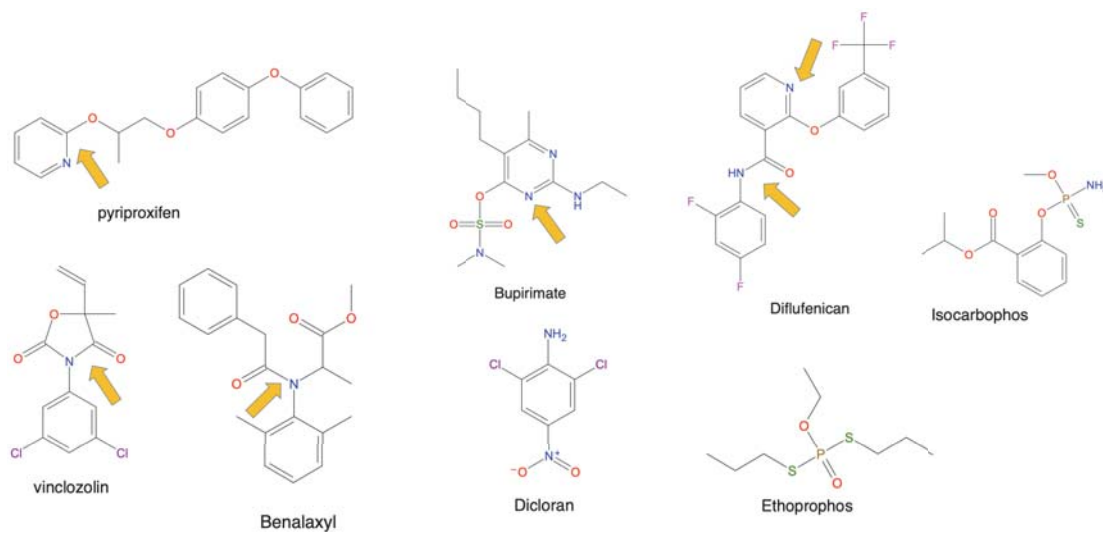


Figure 5.7. Chemical structure of pesticides wrongly classified by xgbDART (2 of 5)

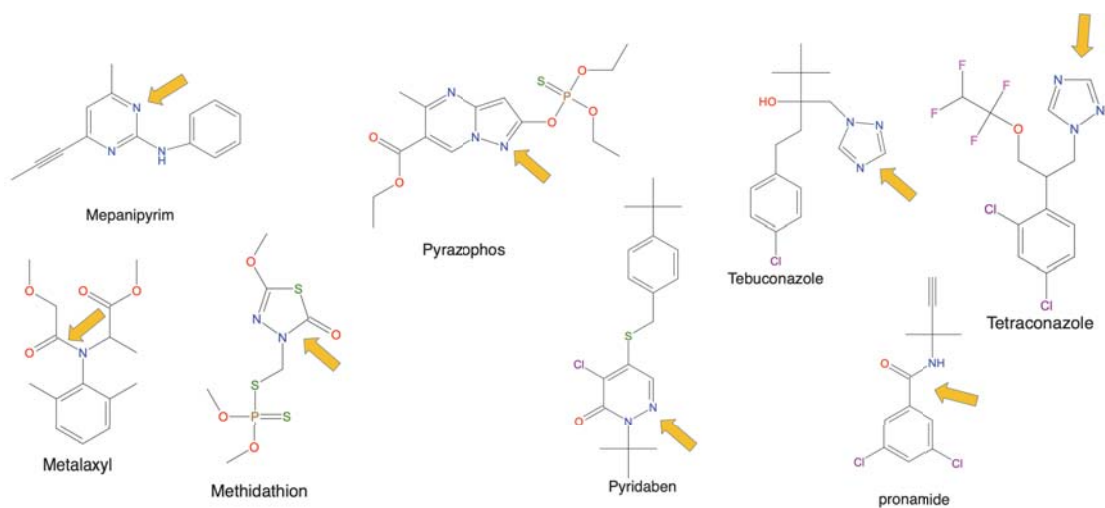


Figure 5.8. Chemical structure of pesticides wrongly classified by xgbDART (3 of 5)

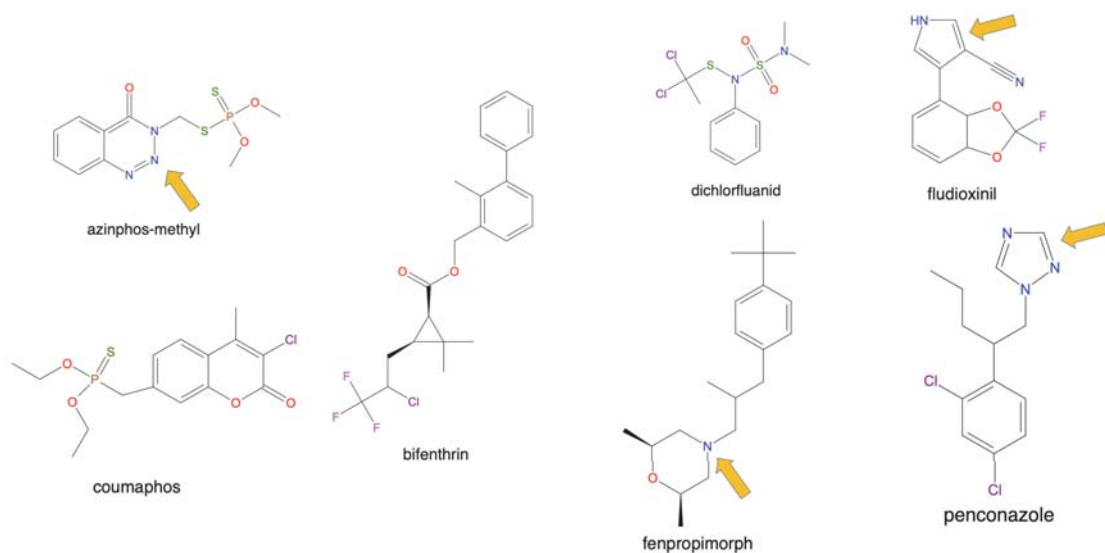


Figure 5.9. Chemical structure of pesticides wrongly classified by xgbDART (4 of 5)

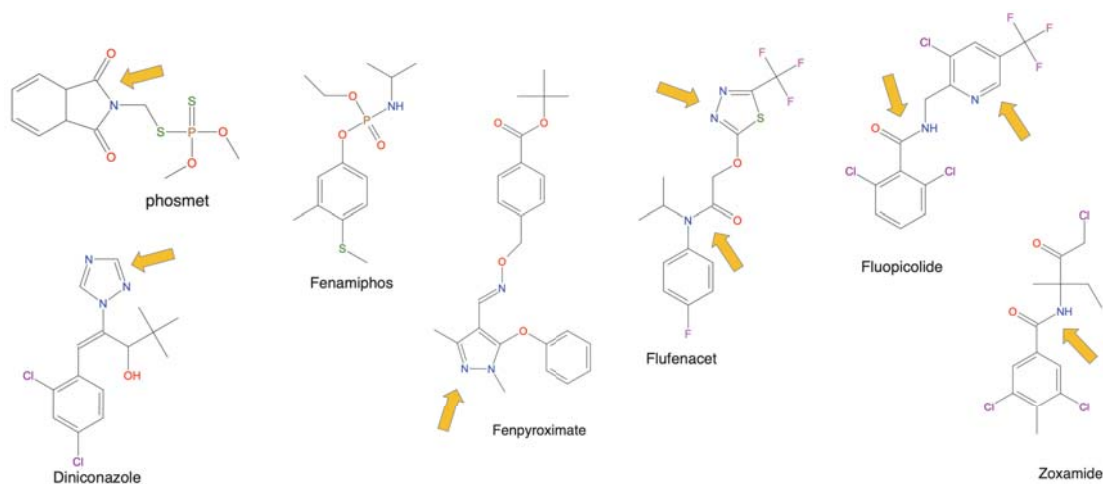


Figure 5.10. Chemical structure of pesticides wrongly classified by xgbDART (5 of 5)

XLogP, ECCEN, Kier3, WTPT.4, ALogP and BCUTp.11 were the major MDs that classify the pesticides between LC-MS and GC-MS. Box plots of Figure 5.12 shows the comparisons of XLogP, one of the characteristic MDs that shows opposite distribution between all 194 pesticides and the 35 pesticides wrongly classified. As shown in the Table 5.7 and Table 5.8, the average of XLogP of all pesticides of LC-MS was 2.33 and the average of all GC-MS was 3.72. On the other hand, the average of XLogP of wrongly classified LC-MS was 3.72 and GC-MS was 2.98.

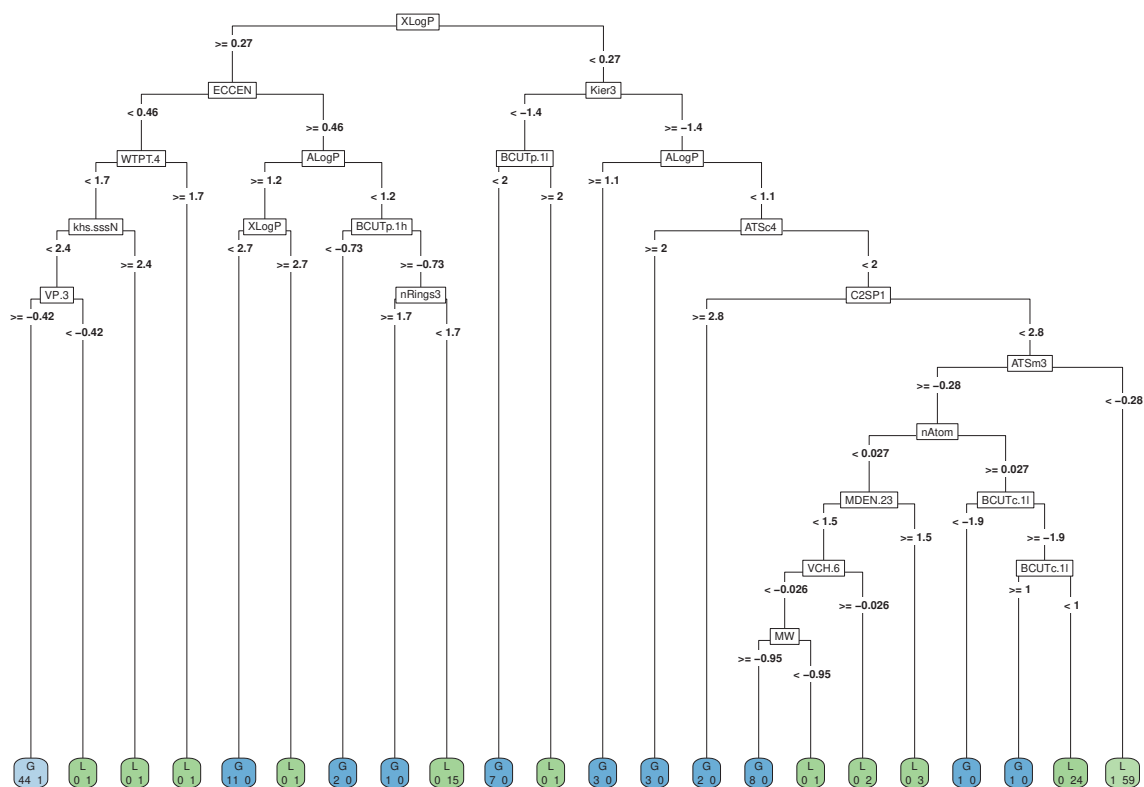


Figure 5.11. Decision tree of 'rpart' to classify 194 pesticides using 176 MDs

Thus, the presence of nitrogen and other heteroatoms gives the influence on the XLogP, that makes the machine learning prediction difficult to predict accurately. Generally, LogP is one of the well-known prime indicators of chemical property of the molecules, so removal of XLogP will be losing the prediction capability of the other pesticides. Other possible solution to improve the accuracy on this data set will be categorize the pesticides in sub groups according to the chemical class for reducing the influence of the other pesticides' effect.

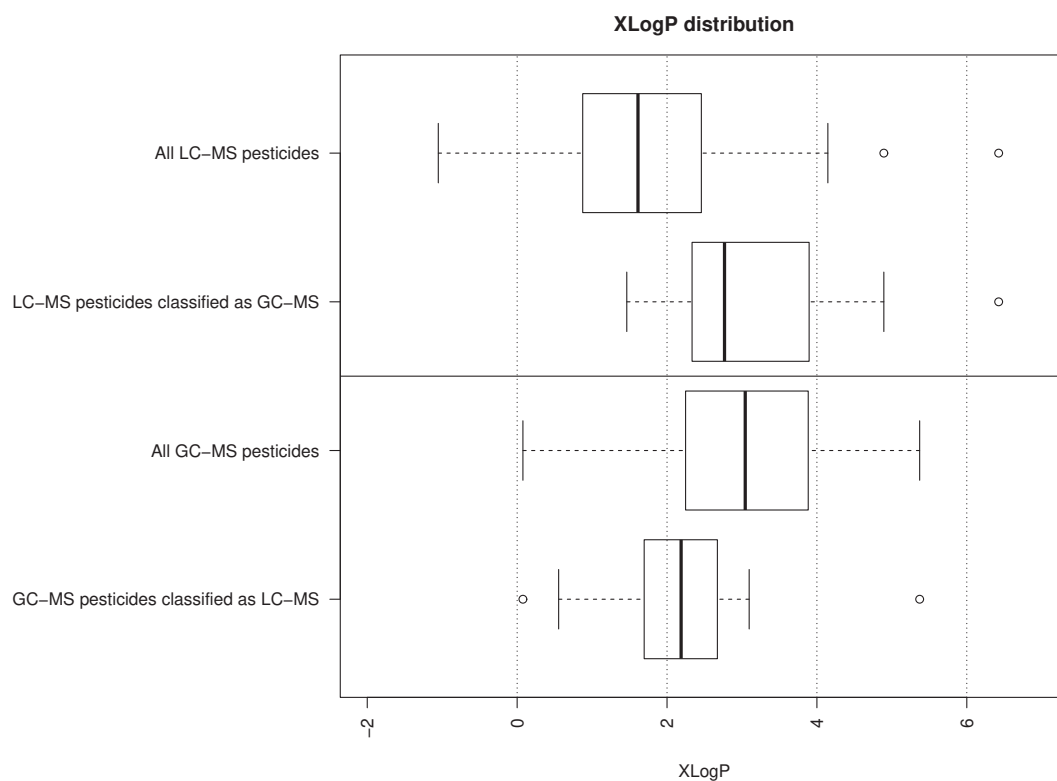


Figure 5.12. Comparison of XLogP between all 194 pesticides and 35 pesticides wrongly classified by xgbDART

Table 5.7. XLogP of LC-MS pesticides (All 110 pesticides and 14 pesticides wrongly classified)

LC-MS pesticides	All 110 pesticides	14 pesticides wrongly classified
Average	2.33	3.72
Standard deviation	1.41	1.14

Table 5.8. XLogP of GC-MS pesticides (All 84 pesticides and 21 pesticides wrongly classified)

GC-MS pesticides	All 84 pesticides	21 pesticides wrongly classified
Average	3.72	2.98
Standard deviation	1.14	1.34

## 5.4 Conclusion of Chapter 5

In Chapter 5, machine learning method to classify the pesticides between LC-amenable and GC-amenable by the molecular descriptors from two validated pesticide reports was developed. Overall accuracy of the classification was up to 85% and execution time to building the prediction model ranged from less than 10 seconds to over 3 hours. xgbDART (eXtreme Gradient Boosting and Additive Regression Tree) was the best machine learning method to classify the 194 pesticides. The common chemical structure of wrongly classified pesticides by xgbDART was the presence of heteroatoms such as nitrogens, oxygens and halogens. Decision tree by ‘rpart’ package of R program showed XLogP is the major molecular descriptor to classify the pesticides for this data set. The distribution of XLogP showed that all LC-MS pesticides and all GC-MS pesticides were opposite to that of wrongly classified pesticides. This can be interpreted by the fact that heteroatoms can influence on the polarizabilities of the pesticide molecule, that can influence on the XLogP descriptor. Considering the importance of LogP (Water - Octanol coefficient) among molecular descriptors, removal of XLogP from machine learning will not be the right solution which gives other wrong classification of pesticides. For improving the accuracy of the classification, pre-classification of pesticides according to the chemical class is required instead of classifying the pesticides by the prediction model with all pesticides.



## 6. Classification of pesticides amenability between LC or GC with graph clustering tool

### 6.1 Introduction

In Chapter 5, I proposed the procedure to classify the amenability between LC and GC by machine learning using the molecular descriptors(MDs). Just the same as Chapter 3, there is the opportunity in optimizing the selection of MDs as the explanatory variables of machine learning by correlation analysis and cluster analysis according to the correlation coefficient among molecular descriptors. In this chapter, similar approach as Chapter 3 is applied on the selection of MDs for machine learning for classification of pesticide amenability between LC and GC. I propose the method to optimize the selection of MDs utilizing the correlation analysis and graph-clustering tool, i.e. DP Clus Software [33]. I propose two considerations below for selection of the optimum MD.

#### 1. Reduction of highly correlated MDs

Select unique MDs utilizing the correlation analysis, i.e. select the MD with less correlations with any other MDs

#### 2. Minimize the loss of information

As many MDs as possible in order to avoid losing the information by Selection utilizing the graph-clustering tool

### 6.2 Materials and Methods

#### 6.2.1 Correlation analysis and cluster analysis

In order to select the optimum MDs for machine learning which reduces the MDs of high correlation, I propose the process shown in the Figure 6.1.

There are five steps for selection of MDs same as Chapter 3, 1) Input for correlation analysis, 2) Select the MDs of weak correlation with any other MDs, 3) cluster analysis to pick up representative MD(s) from each cluster and removal of other MDs, 4) correlation analysis for the MDs selected by Step 3 and 5) final selection of MDs from the cluster analysis.

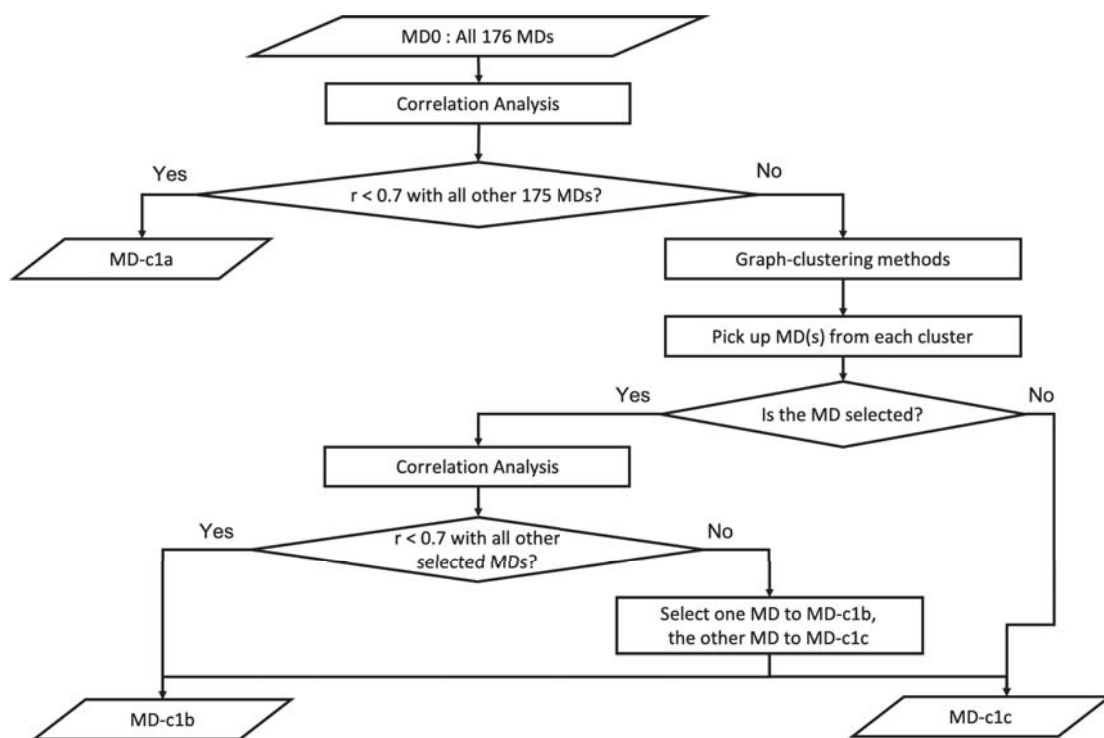


Figure 6.1. Process chart of selecting the optimum MDs

**1st step: List the correlations of all possible combinations** The first step is to list the correlations of all possible combinations among 176 MDs. MD-MD correlations were calculated by the Pearson’s correlation coefficient( $r$ )[26] using corrr package of R program and stretch function [31] for all 176 MDs. Based on the guidances of Pearson correlation coefficient [32], I set the threshold at  $r=0.7$  for the “Highly correlated” of MDs on the present study. The 176 MDs were divided into two groups by this threshold  $r \geq 0.7$ . The MDs in the combinations of  $r \geq 0.7$  are classified as “Strongly correlated MD” and the other MDs are “Weakly correlated MD group” in the present study.

**2nd step: Pick up the MDs with weak correlations with other MDs** The second step is to pick up the MDs of weak correlation(i.e.  $r < 0.7$ ) with any other MDs. These MDs are grouped as “MD-c1a” which are used for regression analysis of machine learning later.

**3rd step: Pick up the MDs by graph clustering tool** The third step is to visualize the correlations of strongly correlated MDs by the method of graph-clustering method called DP Clus [33] and pick up the representative MD(s) from each cluster. The parameters of DP Clus software is set as Table 6.1.

Table 6.1. DP Clus parameters	
Parameter	Value
Cluster Property Value ( $cp_{nk}$ )	0.5
Density Value ( $d_k$ )	0.9
Minimum Cluster Value	2

The Density value is the threshold of the cluster density in the MD-MD connection network. In the MD-MD network, MD is represented as the node and MD-MD connection is represented as the edge. The density value  $d_k$  of any cluster  $k$  is the ratio of the number of edges present in the cluster ( $|E_k|$ ) and the maximum possible number of edges in the cluster ( $|E_k|_{max}$ ), which is defined as Eq 6.1.

$$d_k = \frac{|E_k|}{|E_k|_{max}} = \frac{2 \times |E_k|}{|N_k| \times (|N_k| - 1)} \quad (6.1)$$

where  $|N_k|$  is the number of MDs in the cluster. If the number of the edges (connection of the MDs) are 90% of maximum combination of the edges, the MDs are grouped into the same cluster.

The cluster property value is the threshold of the connection around the cluster defined as Eq 6.2.

$$cp_{nk} = \frac{|E_{nk}|}{d_k \times |N_k|} \quad (6.2)$$

where  $|E_{nk}|$  is the total number of edges between the node  $n$  and each of the nodes of cluster  $k$ .

**4th step: Confirm the correlation of MDs picked up by graph clustering** The fourth step is to confirm the second correlation analysis among the representative MDs picked up from each cluster. Threshold of correlation is set at  $r \geq 0.7$ .

**5th step: Finalize the MDs for prediction** The fifth step is to select the MD(s) based on the step 4. The MDs of weak correlation with other MDs (i.e.  $r < 0.7$ ) in step 4 are grouped as “MD-c1b”, which are used for regression analysis of machine learning later. The MDs of the combination with the strong correlation in step 4 are divided into two groups according to the combination of  $r$ , i.e. “MD-c1b” and “MD-c1c”. MDs in MD-c1c are excluded from further regression analysis.

Thus, 176 MDs are divided into three groups as listed in the table 6.2 according to the process in Figure 6.1.

### 6.2.2 Machine learning by ‘caret’ package

Accuracy and Execution Time(ET) were measured the same procedure as Chapter 5 for 119 regression machine learning methods, then these performances were compared.

Table 6.2. Group of MDs for optimum selection

MD group	MDs to be classified
MD-c1a	MD of $r < 0.7$ with any of other 175 MDs
MD-c1b	MD of $r \geq 0.7$ with any of other 175 MDs and selected by graph-clustering method
MD-c1c	MD of $r \geq 0.7$ with any of other 175 MDs and excluded by graph-clustering method

## 6.3 Results and Discussion

### 6.3.1 Correlation analysis results among molecular descriptors

Figure 6.2 shows the histogram of correlation distribution among the 176 MDs. The x-axis is the Pearson correlation coefficient of the combination and y-axis is the count(frequency) in each bin.

### 6.3.2 Selection of molecular descriptors by the clustering tool - DP Clus

There are 3365 MD-MD combinations of  $r \geq 0.7$  which are consisted of 130 MDs. 46 MDs out of 176 MDs were classified as MD-c1a. 130 MDs were classified into 25 clusters by DP Clus software as shown in Figure 6.3. 16 clusters are correlated dependently as shown in Figure 6.4 and MDs in nine clusters are not connected to the other clusters.

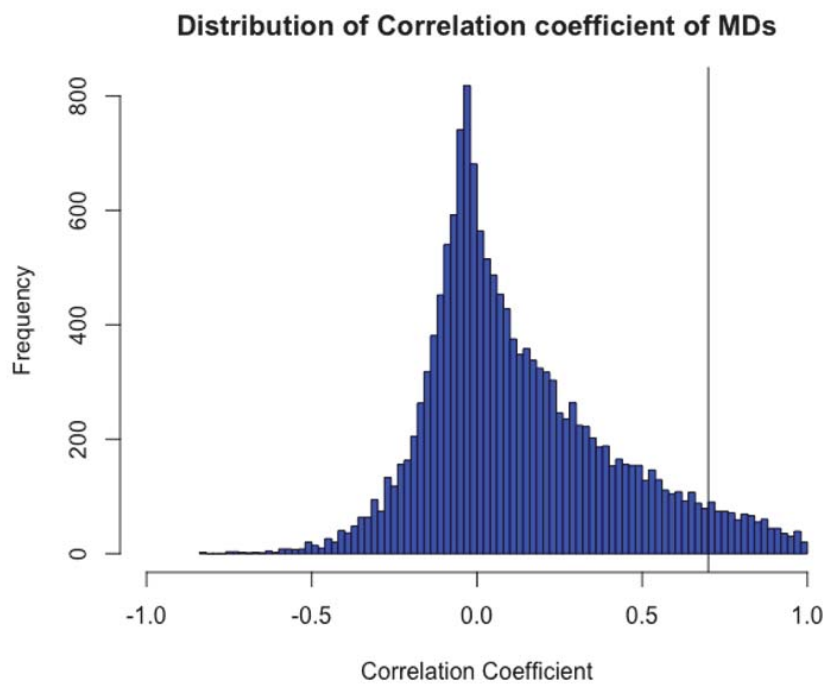


Figure 6.2. Distribution of correlation coefficient of MD

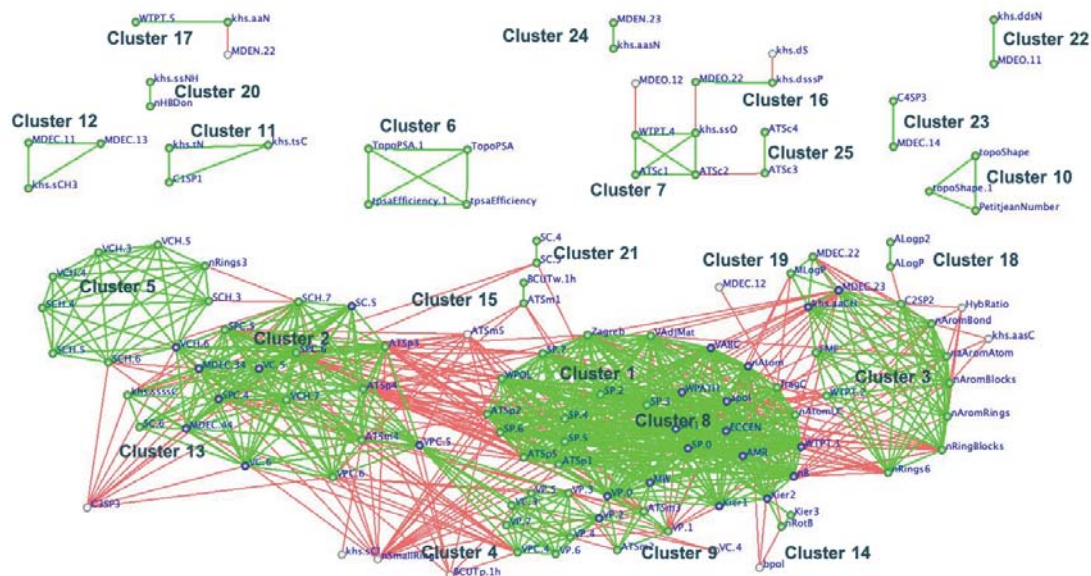


Figure 6.3. Cluster analysis result by DP Clus

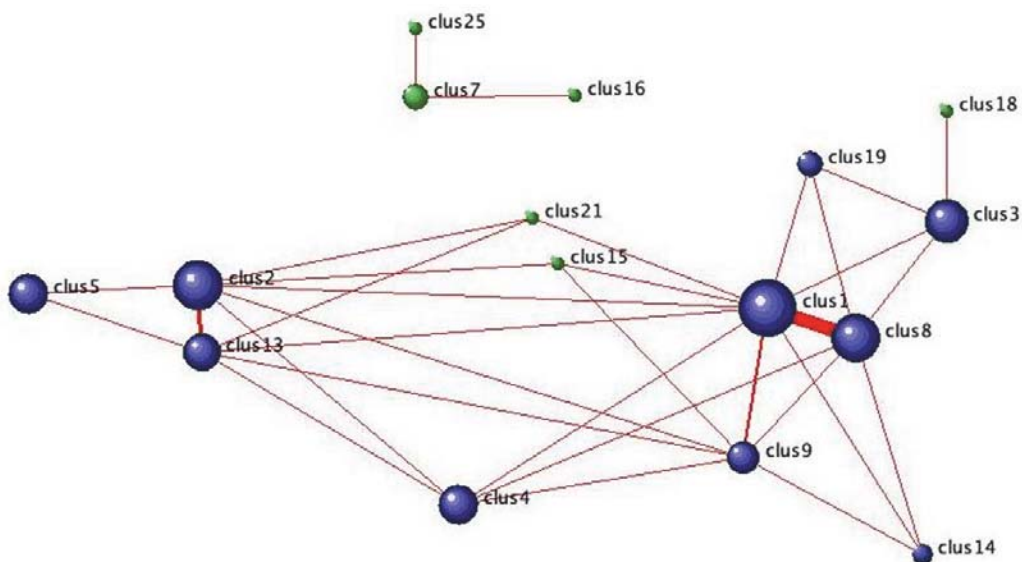


Figure 6.4. Relationship dependent clusters by DP Clus

**Criteria of selection from each cluster** I set the criteria as shown in the decision tree(Figure 6.5) to select the MDs from each cluster as the candidate of machine learning for classification. There are six decisions for selections of MD(s) from each cluster in order to avoid redundancy and loss of information. The first step is the criteria for number of MDs in the cluster. According to the correlation analysis of each cluster(discussed later), there are several combinations of MDs with weak correlation in the same cluster due to the presence of other many MDs with strong correlations. In order to reduce the loss of information, this should be the first check point for selection of MDs from each cluster. The second filter is applied only for the cluster with the number of MD  $\geq 5$ . If there is any combination of MDs with  $r < 0.5$ , both MDs is selected for from that cluster. Else, third filter was applied for selection of MDs. Before third filter, correlation analysis in the cluster was performed and four combinations of MDs with least  $r$  were picked up. The third filter is the frequency of MD in the combinations of least  $r$  within the cluster. There are several clusters that have several MDs with same frequency in the combination. The forth criteria is the type of MD, “not integer” is added for more information from that MD for better prediction of regression. If there is a single MD that meets these criteria in the four combinations of MDs, that MD is selected. The fifth filter is the number of connection with the other MDs in the cluster. This is the assumption that the node with less edge(s) should be less correlation. The final filter is the number of inter cluster connection in the cluster. This is also the assumption that less edges should be less correlation with the other MDs.



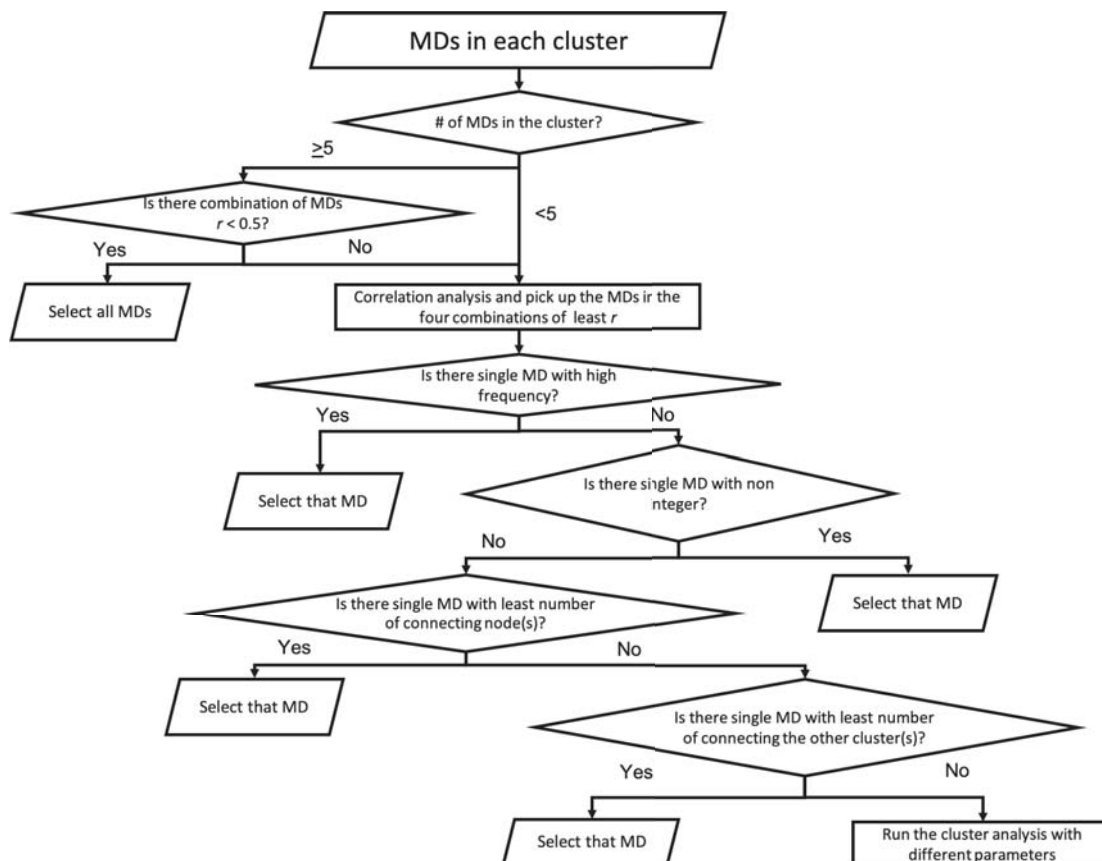


Figure 6.5. Decision tree to select the optimum MDs from each cluster

### Selection of molecular descriptors by the clustering tool - Cluster 1

In DP Clus, the MDs are denoted as the node by small circle and the relation of MDs are indicated as edge, green line for intra cluster and red line for inter cluster. In Cluster 1, 25 MDs are correlated, i.e. AMR(Molar polarizability), apol(Sum of atomic polarizabilities), ATSp1,2,5(Autocorrelation polarizability), ECCEN(A topological descriptor combining distance and adjacency information), Kier1(First Kappa Shape Index), MW(Molecular Weight), nAtom(Number of atoms), nB(Number of bonds), SP.0-7(Chi path descriptor, single path), VABC(Atomic and bond contributions of Van der Waals volume), VAdjMat(Vertex adjacency information), VP.0(Chi path descriptor, valence path, order 0), WPOL

(Weiner Polarity Number), WPATH (Weiner path number), WTPT.1 (Molecular ID) and Zagreb (Sum of the squares of atom degree over all heavy atoms) are included as shown in the Figure 6.6. Cluster 1 connects with the other nine clusters(Cluster 2, 3, 8, 9, 13, 14, 15, 19, 21). According to the decision tree(Figure 6.5), ATSp2 and nAtom are selected from cluster 1. They are correlated at  $r = 0.443$ .

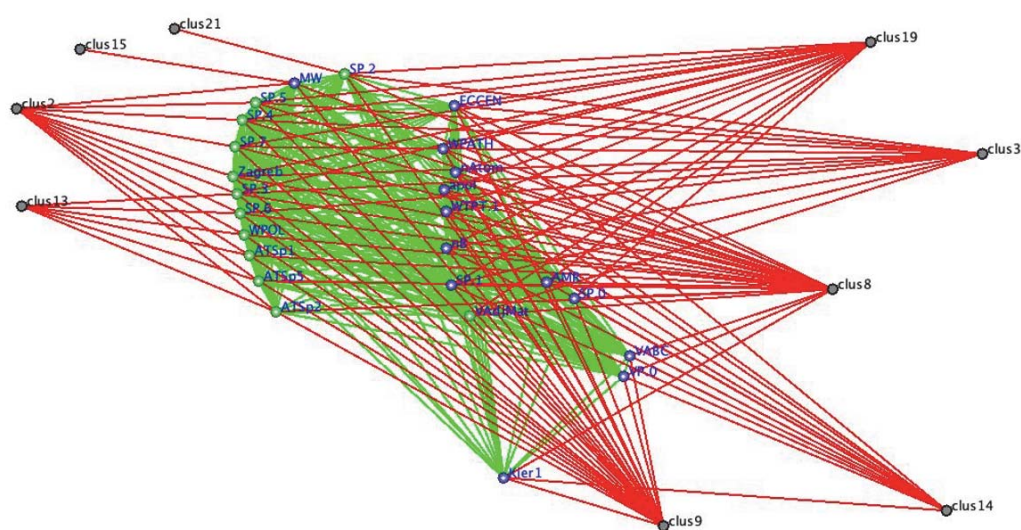


Figure 6.6. Cluster 1: 25 molecular descriptors are correlated as shown in green lines. Cluster 1 connects with the other nine clusters as in red lines.

### Selection of molecular descriptors by the clustering tool - Cluster 2

In Cluster 2, 16 MDs are correlated, i.e. ATSm4(ATS autocorrelation descriptor, weighted by scaled atomic mass), ATSp3-4(ATS autocorrelation descriptor, weighted by polarizability), MDEC.34(Molecular distance edge descriptor, between all tertiary and quaternary carbons), MDEC.44(Molecular distance edge descriptor, between all quaternary carbons), SC.5(Chi cluster descriptor, simple path, order 5), SCH.7(Chi chain descriptors, simple chain order 7), SPC.4-6(Chi path cluster descriptor, simple path, order 4-6), VC.5 and 6(Chi cluster descriptor, valence cluster, order 5 and 6), VCH.6 and 7(Chi chain descriptors, valence chain order 6 and 7) and VPC.5-6(Chi path cluster descriptor, valence path, or-

der 5 and 6) as shown in the Figure 6.7. Many of these MDs are topological descriptors. Cluster 2 connects with the other seven clusters(Cluster 1, 4, 5, 9, 15, 21). According to the decision tree(Figure 6.5), VCH.6 is selected.

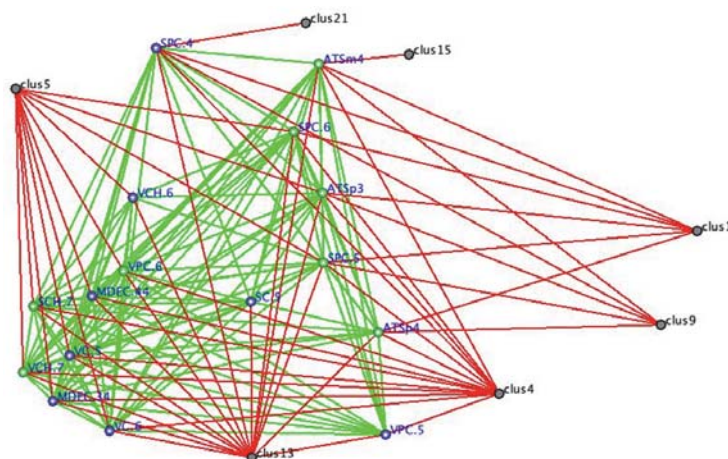


Figure 6.7. Cluster 2: 16 MDs correlated as shown in the green lines. Cluster 2 connects with other seven clusters as shown in red lines.

**Selection of molecular descriptors by the clustering tool - Cluster 3** In Cluster 3, 11 MDs correlated, i.e. C2SP2(Doubly bound carbon bound to two other carbons), FMF(FMF descriptor characterizing complexity of a molecule), khs.aaCH(Count of atom type E-state of carbon connecting with two aromatic groups and hydrogen), MDEC.23(Molecular distance edge descriptor, between all secondary and tertiary carbons), naAromAtom(Number of aromatic atoms), nAromBlocks(Number of aromatically connected bonds), nAromBond(Number of aromatic bonds), nAromRings(Number of aromatic rings), nRingBlocks(Total number of distinct ring blocks), nRings6(individual breakdown of small rings), and WTPT.2(Molecular ID / number of atoms) as shown in Figure 6.8. Cluster 3 connects to other four clusters(Cluster 1, 8, 18, 19). These MDs are related to the aromatic compounds or aromatic bonds. According to the decision tree(Figure 6.5), WTPT.2 is selected.

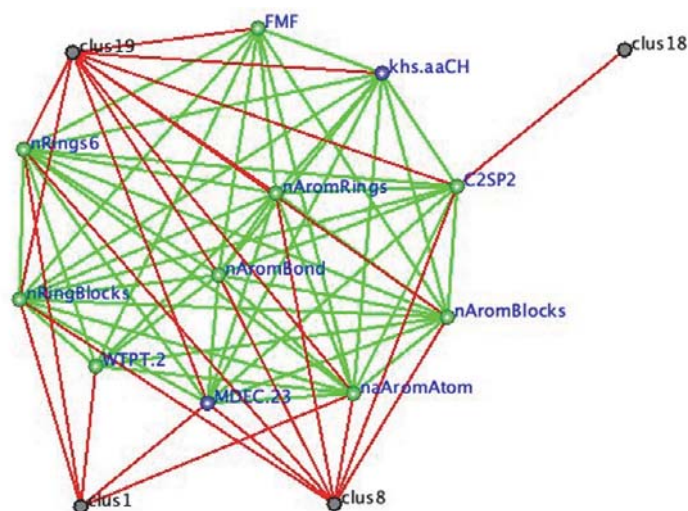


Figure 6.8. Cluster 3: 11 MDs correlated as shown in green lines. Cluster 3 connects to other four clusters as shown in red lines.

#### Selection of molecular descriptors by the clustering tool - Cluster 4

In Cluster 4, nine MDs are correlated, i.e. VC.3 (Chi cluster descriptor, valence cluster, order 3), VP.2-7 (Chi path descriptor, valence path, order 2 to 7) and VPC.4-5 (Chi path cluster descriptor, valence path, order 4 and 5) as shown in Figure 6.9. Cluster 4 connects to the other three clusters (Cluster 2, 9, 13). These MDs are topological descriptors. According to the decision tree (Figure 6.5), VC.3 is selected from cluster 4.

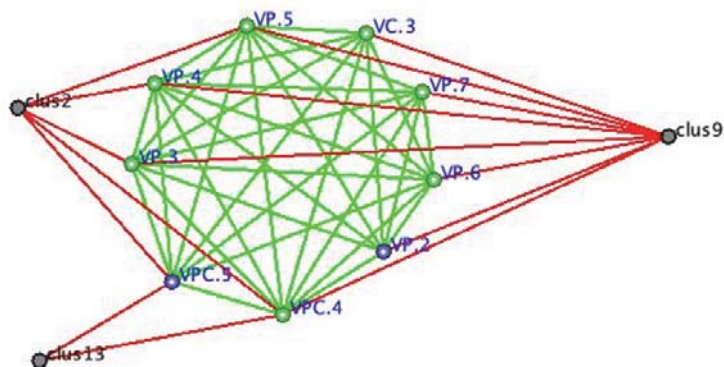


Figure 6.9. Cluster 4: Nine MDs correlate as shown in green lines. Cluster 4 connects with the other three clusters as shown in red lines.

**Selection of molecular descriptors by the clustering tool - Cluster 5** In Cluster 5, nine MDs are correlated, i.e. nRings3(individual breakdown of small rings with 3 members), SCH.3-6(Chi chain descriptor, simple chain, orders 3 to 6) and VCH.3-6(Chi chain descriptor, valence chain, orders 3 to 6) as shown in Figure 6.10. Cluster 5 connects to the other two clusters(Cluster 2, 13). VCH.6 is selected according to the decision tree Figure 6.5.

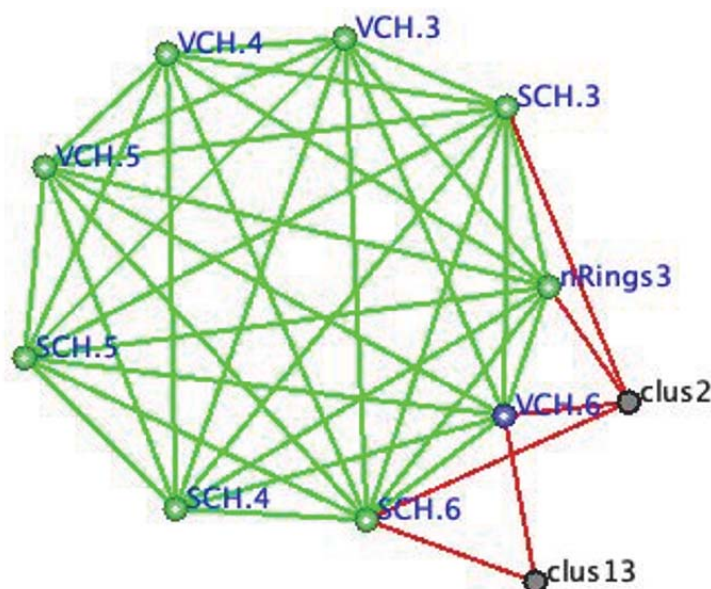


Figure 6.10. Cluster 5: Nine MDs correlate as shown in green lines. Cluster 5 connects with other two clusters as shown in red lines.

### Selection of molecular descriptors by the clustering tool - Cluster 6

In Cluster 6, four MDs of topology correlate, i.e. TopoPSA(topological polar surface area), TopoPSA.1(topological polar surface area), tpsaEfficiency(Polar surface area expressed as a ratio to molecular size) and tpsaEfficiency.1(Polar surface area expressed as a ratio to molecular size) are included as shown in Figure 6.11. Cluster 6 doesn't connect to any other clusters. These are the MDs of topological surface area of the molecules. TopoPSA selected according to the decision tree Figure 6.5.



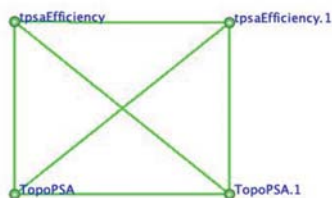


Figure 6.11. Cluster 6: Four MDs correlated as shown in green lines.

### Selection of molecular descriptors by the clustering tool - Cluster 7

In Cluster 7, four MDs correlated, i.e. ATSc1-2(ATSc autocorrelation descriptor, weighted by charges), khs.ssO(Count of atom type E-state of oxygen with two single bonds) and WTPT.4(Weighted path descriptor, Sum of path lengths starting from oxygens) as shown in Figure 6.12. Cluster 7 connects with other two clusters(Cluster 16, 25). WTPT.4 selected according to the decision tree Figure 6.5.

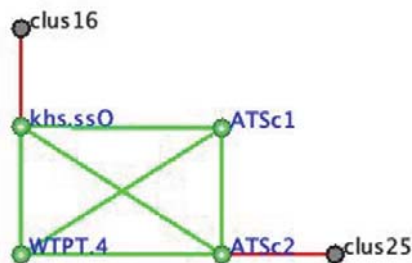


Figure 6.12. Cluster 7: Four MDs correlate as shown in green lines. Cluster 7 connects with other two clusters as shown in red lines.

### Selection of molecular descriptors by the clustering tool - Cluster 8

In Cluster 8, 15 MDs correlate, i.e. AMR(Molar polarizability), apol(Sum of atomic polarizabilities), ECCEN(A topological descriptor combining distance and adjacency information), fragC(Complexity of a system), Kier1(First Kappa Shape Index), Kier2(Second Kappa Shape Index), nAtom, nAtomLC(Number of atoms in the largest chain), nB(Number of bonds), SP.0-1(Chi path descriptor,

single path order 0 and 1), VABC(Atomic and bond contributions of Van der Waals volume), VP.0(Chi path descriptor, valence path, order 0), WPATH and WTPT.1(Molecular ID) as shown in Figure 6.13. Cluster 8 connects with the other five clusters(Cluster 1, 3, 9, 14, 19). Kier2 is selected according to the decision tree Figure 6.5.

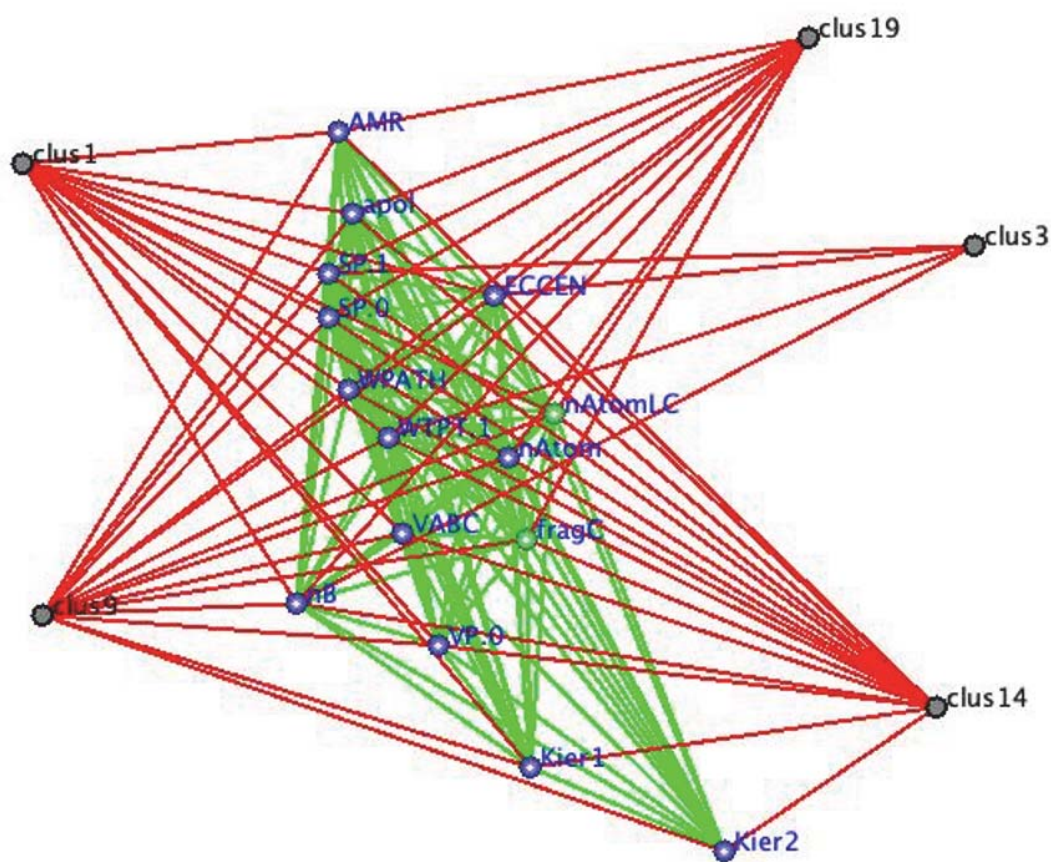


Figure 6.13. Cluster 8: 15 MDs correlate as shown in green lines. Cluster 8 connects with other five clusters as shown in red lines.

**Selection of molecular descriptors by the clustering tool - Cluster 9** In Cluster 9, six MDs correlate, i.e. ATSm2-3(Autocorrelation descriptor, weighted by scaled atomic mass), MW(Molecular weight), VP.0-2(Chi path descriptor,



valence path, orders 0-2) as shown in Figure 6.14. Cluster 9 connects with other seven clusters(Cluster 1, 2, 4, 8, 13, 14, 15). VP.2 is selected according to the decision tree Figure 6.5.

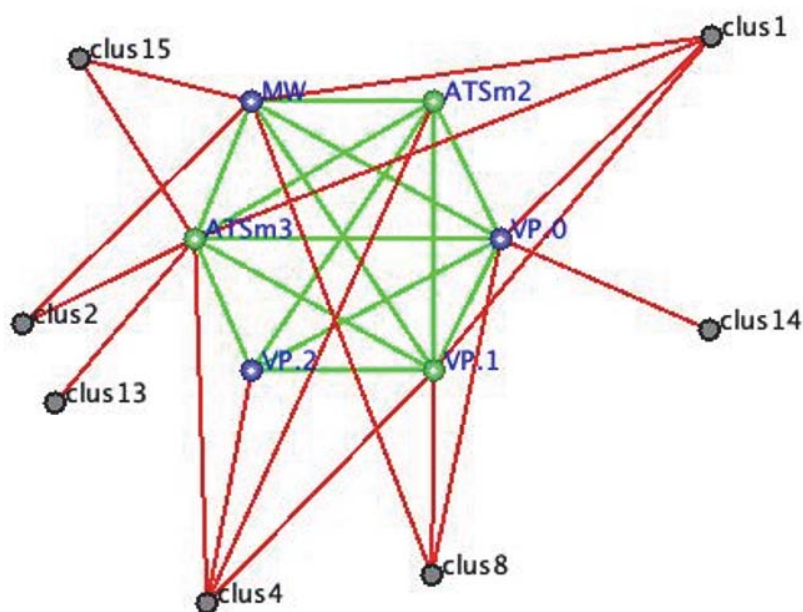


Figure 6.14. Cluster 9: Six MDs correlate as shown in green lines. Cluster 9 connects with other seven clusters as shown in red lines.

### Selection of molecular descriptors by the clustering tool - Cluster 10

In Cluster 10, three MDs correlate, i.e. PetitjeanNumber(Petitjean number, graph-theoretical shape coefficient), topoShape(A measure of the anisotropy in a molecule) and topoShape.1(A measure of the anisotropy in a molecule) shown in Figure ??, no connection with the other cluster. topoShape is selected.

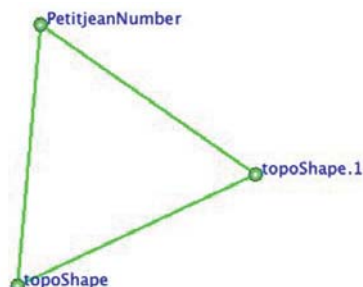


Figure 6.15. Cluster 10: Three MDs correlate as shown in green lines.

### Selection of molecular descriptors by the clustering tool - Cluster 11

In Cluster 11, three MDs correlate, i.e. C1SP1(Triply bound carbon bound to one other carbon), khs.tN(Count of atom type E-state tertiary nitrogen) and khs.tsC(Count of atom type E-state tertiary carbon with single bond) as shown in Figure 6.16, no connection with the other cluster. Azoxystrobin, Chlorfenapyr, Cyfluthrin, Cyhalothrin Lambda, Cymoxanil, Cypermethrin, Deltamethrin, Fenvalerate, Fenpropathrin, Fluvalinate and Myclobutanil are the pesticides of which value of these MDs' are "1" and they include the cyano group( $\text{-C}\equiv\text{N}$ ) in their chemical structure as shown in Figure 6.17. C1SP1 is selected according to the decision tree Figure 6.5.

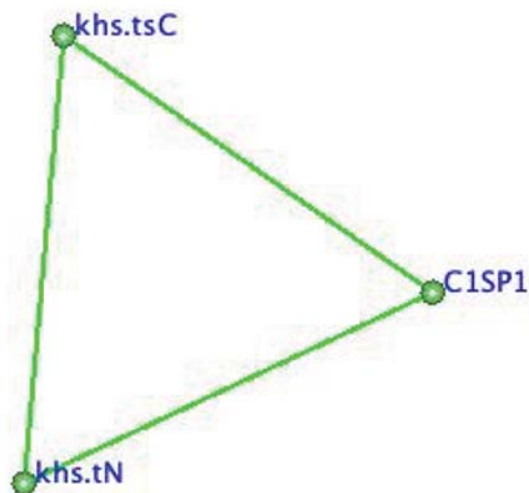


Figure 6.16. Cluster 11: Three MDs correlate as shown in green lines.

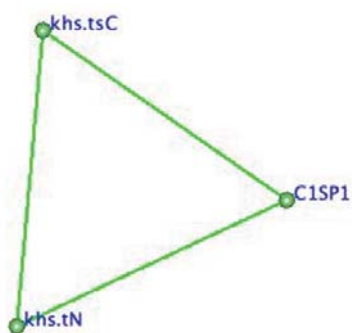


Figure 6.17. Chemical structure of pesticides with the unique value of MDs for cluster 11.

### Selection of molecular descriptors by the clustering tool - Cluster 12

In Cluster 12, three MDs correlate, i.e. khs.sCH3(Count of atom type E-state of CH3 with single bond), MDEC.11(Molecular Distance Edge Descriptor, between all primary carbons) and MDEC.13(Molecular Distance Edge Descriptor, between all primary and tertiary carbons) as shown in Figure 6.18, no connection with the other cluster. MDEC.13 is selected according to the decision tree Figure 6.5.

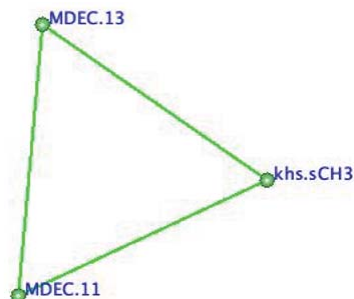


Figure 6.18. Cluster 12: Three MDs correlate as shown in green lines.

### Selection of molecular descriptors by the clustering tool - Cluster

**13** In Cluster 13, eight MDs correlate, i.e. khs.ssssC(Count of atom type E-state tertiary carbon), MDEC.34(between tertiary and quarterly carbons), MDEC.44(between all quaternary carbons), SC.5-6(Chi cluster descriptor, simple path, order 5 and 6), SPC.4(Chi path cluster descriptor, simple path, order 4) and VC.5-6(Chi cluster descriptor, valence path, order 5 and 6) as shown in Figure 6.19. Cluster 13 connects with other six clusters(Cluster 1, 2, 4, 5, 9, 21). khs.ssssC is selected according to the decision tree Figure 6.5.

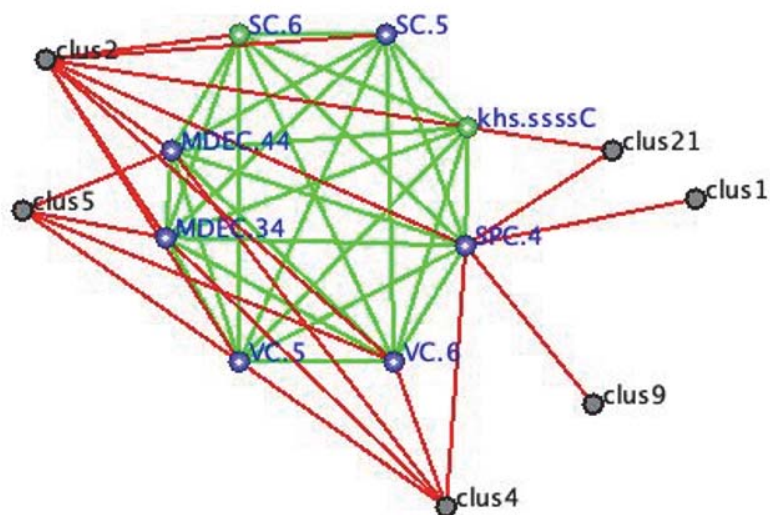


Figure 6.19. Cluster 13: Eight MDs correlate as shown in green lines. Cluster 13 connects with other six clusters as shown in red lines.

#### Selection of molecular descriptors by the clustering tool - Cluster 14

In Cluster 14, three MDs correlate, i.e. i.e. Kier2(Kappa shape indices descriptor, second kappa shape index), Kier3(Kappa shape indices descriptor, third kappa shape index) and nRotB(Number of rotatable bonds, excluding terminal bonds) as shown in Figure 6.20. Cluster 14 connects with other two clusters(Cluster 8, 9). Kier3 is selected according to the decision tree Figure 6.5.

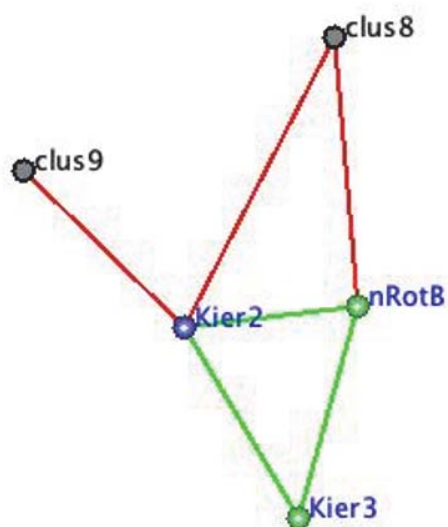


Figure 6.20. Cluster 14: Three MDs correlate as shown in green lines. Cluster 14 connects with other two clusters as shown in red lines.

#### Selection of molecular descriptors by the clustering tool - Cluster 15

In Cluster 15, two MDs correlate, i.e. ATSm1(ATS autocorrelation descriptor, weighted by scaled atomic mass) and BCUTw.1h(The number of lowest eigenvalue highest atom weighted BCUTS) as shown in Figure 6.21. Cluster 15 connects with the other 2 clusters(Cluster 2, 9). BCUTw.1h is selected according to the decision tree Figure 6.5.

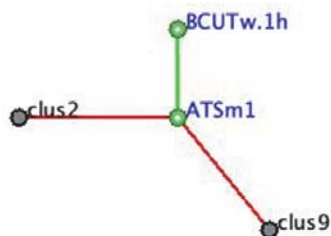


Figure 6.21. Cluster 15: Two MDs correlate as shown in green lines. Cluster 15 connects with other two clusters as shown in red lines.

### Selection of molecular descriptors by the clustering tool - Cluster 16

In Cluster 16, two MDs correlate, i.e. `khs.dsssP`(Count of atom type E-state of phosphorus with one double bond and three single bonds) and `MDEO.22`(Molecular distance edge descriptor, between all secondary oxygens) as shown in Figure 6.22. Cluster 16 connects with cluster 7. `khs.dsssP` is selected according to the decision tree Figure 6.5.



Figure 6.22. Cluster 16: Two MDs correlate as shown in green lines. Cluster 16 connects with cluster 7 as shown in red lines.

### Selection of molecular descriptors by the clustering tool - Cluster 17

In Cluster 17, two MDs correlate, i.e. `khs.aaN`(Count of atom type E-state of nitrogen with two aromatic groups) and `WTPT.5`(Sum of path lengths starting from nitrogens) as shown in Figure 6.23. `WTPT.5` is selected according to the decision tree Figure 6.5.



Figure 6.23. Cluster 17: Two MDs correlate as shown in green lines.

### Selection of molecular descriptors by the clustering tool - Cluster 18

In Cluster 18, two MDs of `AlogP`(Ghose-Crippen  $\text{Log}K_{ow}$ ) and `ALogP2`(Square of `AlogP`) are included as shown in Figure 6.24. Cluster 18 connects with cluster 3. `ALogP2` is the square of `AlogP`, as discussed in Chapter 3. `ALogP2` is selected according to the decision tree Figure 6.5.



Figure 6.24. Cluster 18: Two MDs correlate as shown in green lines. Cluster 18 connects with cluster 3 as shown in red lines.

### Selection of molecular descriptors by the clustering tool - Cluster 19

In Cluster 19, four MDs correlate, i.e. khs.aaCH(Count of atom type E-state of carbon connecting with two aromatic groups and hydrogen), MDEC.22(between all secondary carbons), MDEC.23(Molecular distance edge descriptor, between all secondary and tertiary carbons)and MLogP(Mannhold LogP) as shown in Figure 6.25. Cluster 19 connects with other three clusters(Cluster 1, 3, 8). MLogP is selected according to the decision tree Figure 6.5.

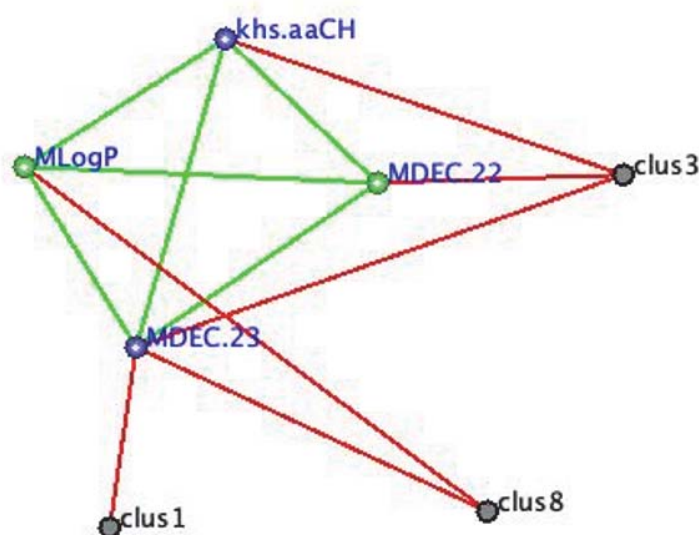


Figure 6.25. Cluster 19: Four MDs correlate as shown in green lines. Cluster 19 connects with three clusters as shown in red lines.



### Selection of molecular descriptors by the clustering tool - Cluster 20

In Cluster 20, two MDs correlate, i.e. khs.ssNH(Count of atom type E-state of nitrogen connecting with hydrogen and two other single bonds) and nHBDon(Number of hydrogen bond donors) as shown in Figure 6.26. nHBDon is selected according to the decision tree Figure 6.5.



Figure 6.26. Cluster 20: Two MDs correlate as shown in green lines.

### Selection of molecular descriptors by the clustering tool - Cluster 21

In Cluster 21, two MDs correlate, i.e. SC.3-4(Chi cluster descriptor, simple cluster, orders 3 and 4) are included as shown in Figure 6.27. Cluster 21 connects with the other 2 clusters(Cluster 1, 13). SC.4 is selected according to the decision tree Figure 6.5.

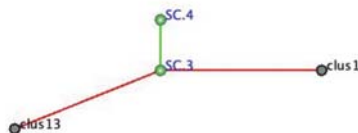


Figure 6.27. Cluster 21: Two MDs correlate as shown in green lines. Cluster 21 connects with two clusters as shown in red lines.

### Selection of molecular descriptors by the clustering tool - Cluster 22

In Cluster 22, two MDs correlate, i.e. khs.ddsN(Count of atom type E-state of nitrogen connecting with two double bonds and one single bond) and MDEO.11(Molecular distance edge descriptor, between all primary oxygens) as shown in Figure 6.28. MDEO.11 is selected according to the decision tree Figure 6.5.



Figure 6.28. Cluster 22: Two MDs correlate as shown in green lines.

#### Selection of molecular descriptors by the clustering tool - Cluster 23

In Cluster 23, TWO MDs correlate, i.e. C4SP3(Singly bound carbon bound to four other carbons) and MDEC.14(Molecular distance edge descriptor, between all primary and quaternary carbons) as shown in Figure 6.29. MDEC.14 is selected according to the decision tree Figure 6.5.



Figure 6.29. Cluster 23: Two MDs correlate as shown in green lines.

#### Selection of molecular descriptors by the clustering tool - Cluster 24

In Cluster 24, two MDs correlate, i.e. khs.aasN(Count of atom type E-state of nitrogen connecting with two aromatic groups and one single bond) and MDEN.23(Molecular distance edge descriptor, between all secondary and tertiary nitrogens) as shown in Figure 6.30. MDEN.23 is selected according to the decision tree Figure 6.5.



Figure 6.30. Cluster 24: Two MDs correlate as shown in green lines.

#### Selection of molecular descriptors by the clustering tool - Cluster 25

In Cluster 25, 2 MDs correlate, i.e. ATSc3-4(ATSc autocorrelation descriptor, weighted by charges) as shown in Figure 6.31. Cluster 25 connects with cluster 7. ATSc4 is selected according to the decision tree Figure 6.5.



Figure 6.31. Cluster 25: Two MDs correlate as shown in green lines. Cluster 25 connects with cluster 7 as shown in red line.

### 6.3.3 Molecular descriptors for machine learning

**Correlation analysis of MDs selected by cluster analysis** Table 6.3 shows the candidates of 26 MDs from 25 clusters selected by the clustering according to the decision tree in Figure 6.5.

Table 6.3. Candidates of MD2(26 MDs from 25 clusters)

Cluster	Molecular Descriptors
1	ATSp2, nAtom
2	VCH.6
3	WTP.T.2
4	VC.3
5	VCH.6
6	TopoPSA
7	WTP.T.4
8	Kier2
9	VP.2
10	topoShape
11	C1SP1
12	MDEC.13
13	khs.ssssC
14	Kier3
15	BCUTw.1h
16	khs.dsssP
17	WTP.T.5
18	ALogp2
19	MLogP
20	nHBDOn
21	SC.4
22	MDEO.11
23	MDEC.14
24	MDEN.23
25	ATSc4

Correlation analysis was performed on these 26 MDs. The combinations of MDs with  $r \geq 0.7$  are listed in the Table 6.4. As shown in Figure 6.32, these MDs were selected from the dependent clusters which results in high correlation after the selection of cluster analysis.

Table 6.4. Combination of MDs with the  $r \geq 0.7$ 

MD-a	MD-b	Corr	Cluster of MD-a	Cluster of MD-b
VC.3	VP.2	0.872	4	9
Kier3	Kier2	0.849	8	14
nAtom	Kier2	0.836	1	14
nAtom	MLogP	0.736	1	19

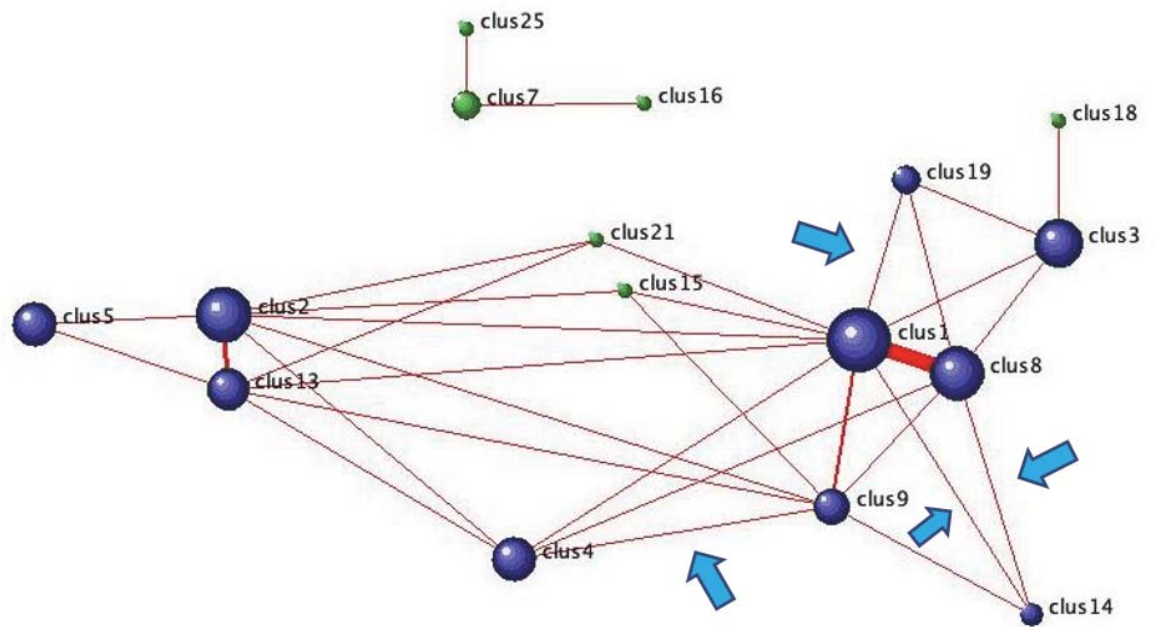


Figure 6.32. Cluster diagram for dependent clusters

**Final selection of MDs for machine learning** The MDs of the column MD-a and MDs of Table 6.5 are finally selected.

Table 6.5. Final MDs selected by cluster analysis

Cluster	Molecular Descriptors
1	ATSp2, nAtom
2	VCH.6
3	WTPT.2
4	VC.3
6	TopoPSA
7	WTPT.4
10	topoShape
11	C1SP1
12	MDEC.13
13	khs.ssssC
14	Kier3
15	BCUTw.1h
16	khs.dsssP
17	WTPT.5
18	ALogp2
20	nHBDon
21	SC.4
22	MDEO.11
23	MDEC.14
24	MDEN.23
25	ATSc4

Table 6.6 is the summary of MDs selected by the decision tree(Figure 6.5).

Table 6.6. Summary of molecular descriptors selected by the correlation analysis and cluster analysis

MD group	MDs to be classified	Number of MDs	Selected
MD-c1a	MD of $r < 0.7$ with any of other 175 MDs	59	Yes
MD-c1b	MD of $r \geq 0.7$ with any of other 175 MDs and selected by graph-clustering method	22	Yes
MD-c1c	MD of $r \geq 0.7$ with any of other 175 MDs and excluded by graph-clustering method	95	No

Thus, MD-c1a and MD-c1b(both MDs are combined as the MD2) were used for the classification analysis.

#### **6.3.4 Comparison of machine learning performance between with and without selection of molecular descriptors**

By selecting the MDs, 30 machine learning methods gave better PE for regression analysis and 89 methods got worse as shown in the Figure 6.33. ORFsvm of Ensemble Decision Tree and kkn of O.Centroid,kNN showed more than 30% improvement in the accuracy of classification when the highly correlated molecular descriptors were removed. Table 6.7 and Table 6.8 are the lists of top 20 machine learning methods of classification performance, sorted by the change ratio of accuracy by selecting the MDs.

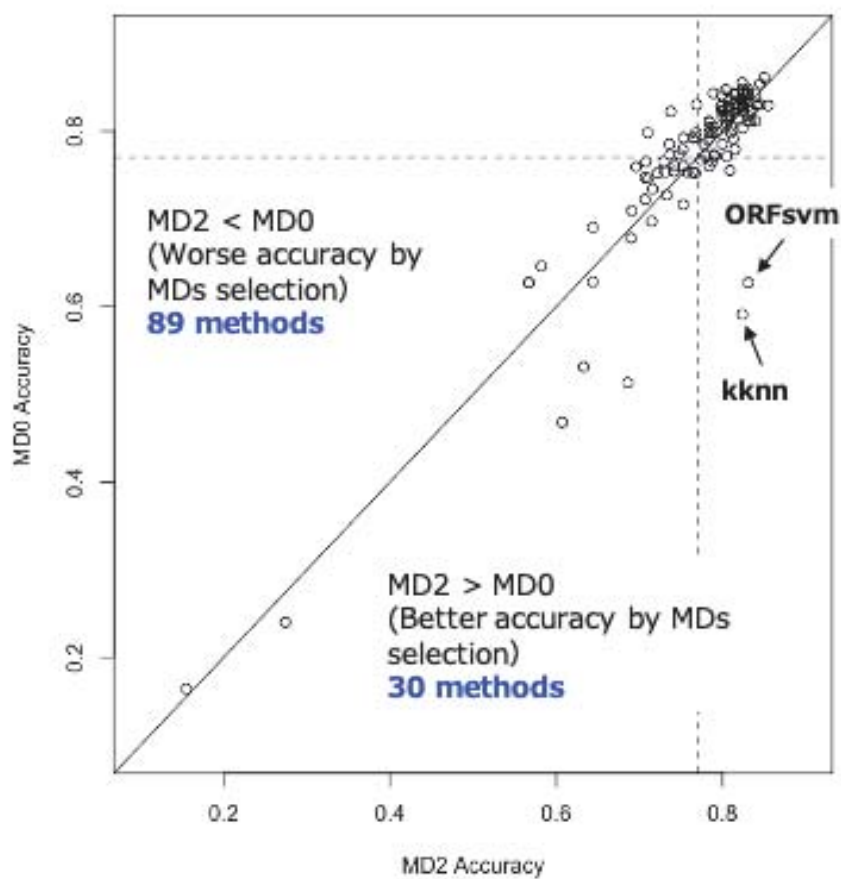


Figure 6.33. Comparison of accuracy by machine learning method on with and without selection of molecular descriptors by correlation analysis and cluster analysis. Plots on the rectangular line shows no difference by selection of molecular descriptors.



Table 6.7. Top 20 method for MD2 < MD0(Worse accuracy by selecting the molecular descriptors)

Method	Category	MD2 Accuracy	MD0 Accuracy	Accuracy ratio
bagEarth	E. Spline	0.254	0.274	0.93
nb	O. Naive Bayes	0.675	0.721	0.94
slda	O. Others	0.763	0.814	0.94
rbfDDA	O. Neural Network	0.778	0.824	0.94
kkn	O. Centroid, kNN	0.784	0.829	0.95
earth	O. Spline	0.773	0.814	0.95
naive bayes	O. Naive Bayes	0.690	0.726	0.95
lssvmRadial	O. Kernel	0.787	0.824	0.95
regLogistic	O. Simple Liner	0.799	0.835	0.96
msaenet	O. Neural Network	0.804	0.840	0.96
fda	O. Others	0.768	0.799	0.96
J48	O. Decision Tree	0.778	0.809	0.96
xyf	O. Others	0.799	0.830	0.96
rfRules	E. Decision Tree	0.741	0.766	0.97
rotationForestCp	E. Decision Tree	0.787	0.814	0.97
rotationForest	E. Decision Tree	0.793	0.819	0.97
RRFglobal	E. Decision Tree	0.792	0.813	0.97
RRF	E. Decision Tree	0.793	0.813	0.97
ordinalNet	O. Simple Liner	0.814	0.835	0.98
pcaNNet	O. Neural Network	0.804	0.824	0.98

Table 6.8. Top 20 method for MD2 > MD0(Better accuracy by selecting the molecular descriptors)

Method	Category	MD2 Accuracy	MD0 Accuracy	Accuracy ratio
glm	O. Simple Liner	0.738	0.541	1.36
gamLoess	O. Spline	0.754	0.624	1.21
gamSpline	O. Spline	0.748	0.635	1.18
bagEarthGCV	E. Spline	0.175	0.149	1.17
pda	O. Others	0.784	0.681	1.15
glmStepAIC	O. Simple Liner	0.779	0.680	1.15
pam	O. Centroid, kNN	0.783	0.716	1.09
sparseLDA	O. Others	0.809	0.743	1.09
rpartCost	O. Decision Tree	0.710	0.658	1.08
rocc	O. Others	0.772	0.726	1.06
rFerns	E. Decision Tree	0.777	0.731	1.06
C5.0Tree	O. Decision Tree	0.768	0.726	1.06
multinom	O. Simple Liner	0.790	0.747	1.06
OneR	O. Decision Tree	0.710	0.675	1.05
svmLinear	O. Kernel	0.763	0.727	1.05
svmLinear3	O. Kernel	0.799	0.763	1.05
treebag	E. Decision Tree	0.834	0.798	1.05
CSimca	O. Simple Liner	0.684	0.659	1.04
nodeHarvest	E. Decision Tree	0.819	0.788	1.04
ORFridge	E. Decision Tree	0.851	0.820	1.04

bagEarth(Bagged MARS, 27.4% to 25.4%) of ensemble Spline, nb(Naive Bayes, 72.1% to 67.5%) of ordinary Naive Bayes and slda(Stabilized Linear Discriminant Analysis, 81.4% to 76.3%) were top 3 methods that got worse in classification by removing the strongly correlated MDs. On the other hand, glm(Generalized Linear Model, 54.1% to 73.8%) of ordinary ordinary Simple Liner, gamLoess(Generalized Additive Model using LOESS, 62.4% to 75.4%) of ordinary spline and gamSpline (Generalized Additive Model using Splines, 63.5% to 74.8%) of ordinary Spline were top 3 methods that got better in classification.

The comparison of accuracy and execution time by the machine learning categories are shown in the Figure 6.34 and Figure 6.35 respectively. According to the Figure 6.34, ordinary sparse modeling category and ordinary Naive Bayes got worse totally. Methods in other categories showed small difference in accuracy. The execution time are reduced for most machine learning methods by selecting the MDs, especially time consuming methods in the category of ordinary simple linear. In Chapter 4, the optimum machine learning method for classification for

this data set was adaboost (AdaBoost Classification Trees of ensemble decision tree model), which was better accuracy (83.9% to 86.0%) by selecting the MDs.

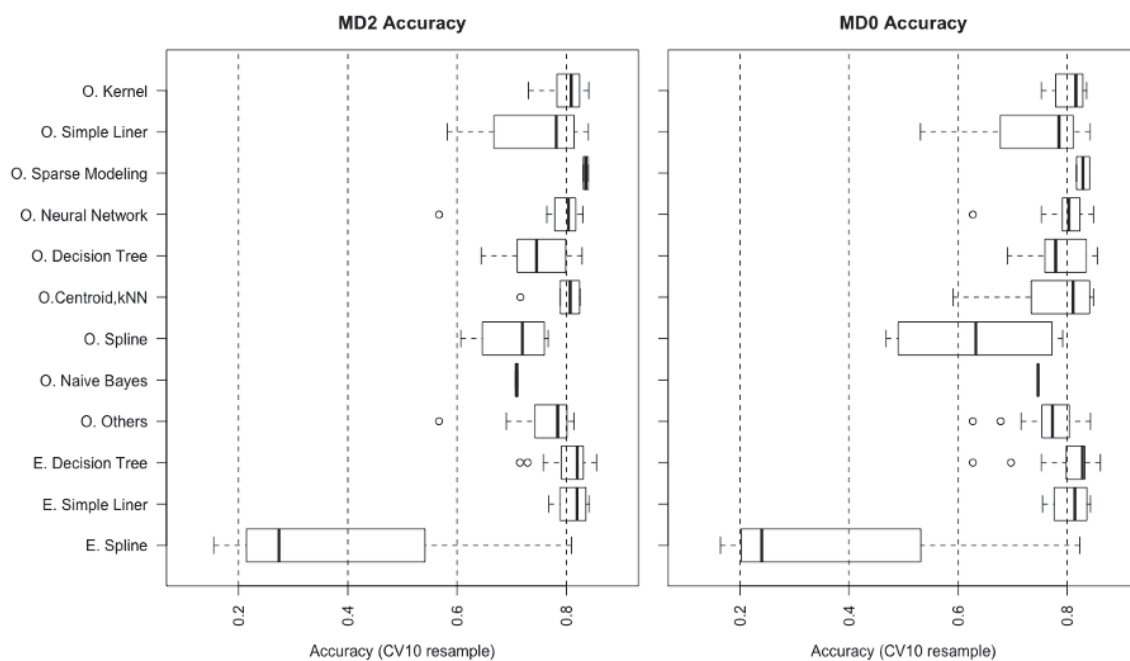


Figure 6.34. Boxplot of accuracy comparison of the molecular descriptors before selection(left) and after selection(right).

Tuning parameters of xgbDART are listed in the Table 6.9.

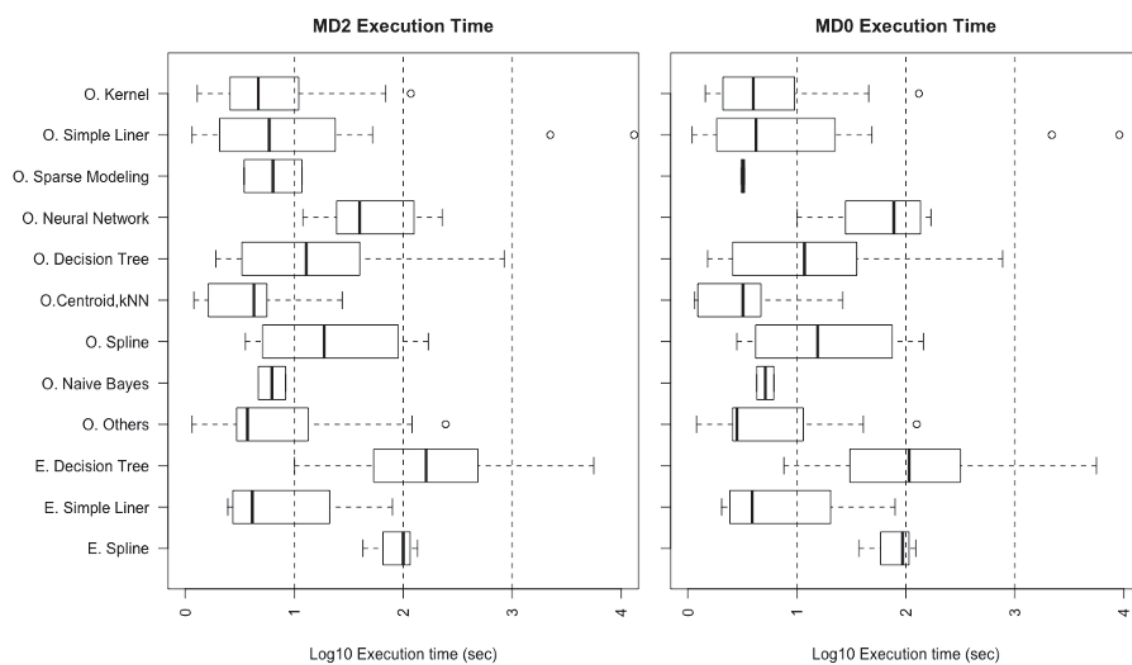


Figure 6.35. Boxplot of Execution Time comparison of the molecular descriptors before(left) and after selection(right).

Table 6.9. Tuning parameters of xgbDART for pesticide classification prediction. (MD2)

Parameter	Value
Boosting Iterations	150
Maximum tree depth	1
Eta (Shrinkage)	0.4
Gamma (Minimum loss Reduction)	0
Subsample (Subsample percentage)	1
Colsample by tree (Subsample ratio of columns)	0.8
Rate drop (Fraction of trees dropped)	0.5
Skip Drop (Probability of skipping drop-out)	0.95
Minimum child weight (Minimum sum of instance weight)	1

## 6.4 Conclusion of Chapter 6

In Chapter 6, graph clustering tool was used to select the molecular descriptors for prediction model of pesticide recovery. DP Clus successfully classify the strongly correlated molecular descriptors into 25 clusters and molecular descriptors from each cluster were selected. Correlation analysis of the selected molecular descriptors was required because some clusters independently connect in the MD-MD network. By selecting the molecular descriptors, execution time for building prediction model was shorten for most machine learning methods. Accuracy of adaboost was improved (83.9% to 86.0%) by selecting the MDs with DP Clus.

## **7. Classification of pesticides amenability between LC or GC with hierarchical cluster analysis**

### **7.1 Introduction**

In Chapter 5, I proposed the procedure to classify the amenability between LC and GC by machine learning using the molecular descriptors(MDs). Just the same as Chapter 3 and Chapter 4, there is the opportunity in optimizing the selection of MDs as the explanatory variables of machine learning by correlation analysis and cluster analysis according to the correlation coefficient among molecular descriptors. In this chapter, similar approach as Chapter 5 is applied on the selection of MDs for machine learning for classification of pesticide amenability between LC and GC, correlation analysis and hierarchical cluster analysis.

Just the same as chapter 6, two strategies below for selecting the MDs for classification of pesticides.

#### **1. Reduction of highly correlated MDs**

Select unique MDs utilizing the correlation analysis, i.e. select the MD with less correlations with any other MDs.

#### **2. Minimize the loss of information**

Select as many MDs as possible in order for minimizing the loss of the information utilizing the clustering tool.

### **7.2 Materials and Methods**

#### **7.2.1 Correlation analysis among molecular descriptors**

In order to select the optimum MDs for machine learning based on the two considerations, I propose the process of the flow chart for MD selection shown in the Figure 7.3. In this chapter, the MD selection using the hierarchical cluster analysis is discussed.

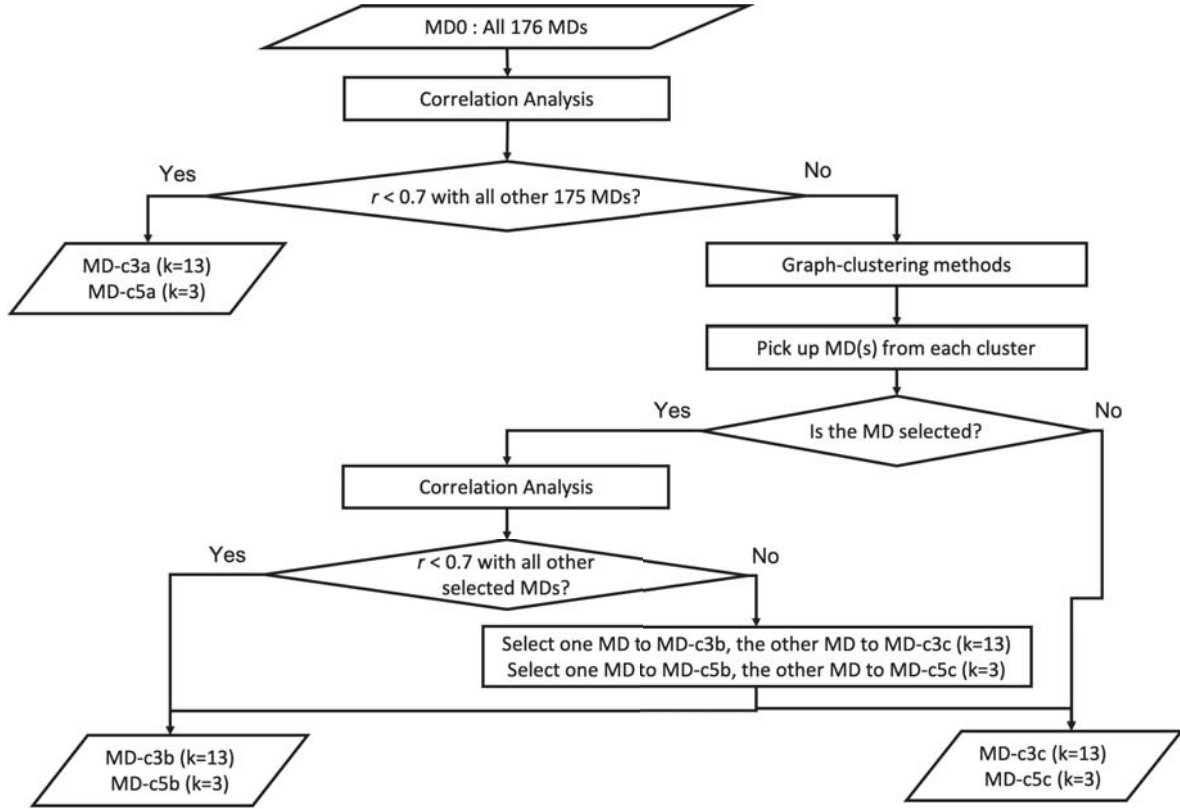


Figure 7.1. Process chart of selecting the optimum MDs

There are five steps for selection of MDs, 1) Input for correlation analysis, 2) Select the MDs of weak correlation with any other MDs, 3) cluster analysis to pick up representative MD(s) from each cluster and removal of other MDs, 4) correlation analysis for the MDs selected by Step 3 and 5) final selection of MDs from the cluster analysis.

**1st step: List the correlations of all possible combinations** The first step is to list the correlations of all possible combinations among 176 MDs. MD-MD correlations were calculated by the Pearson's correlation coefficient( $r$ ) [26] using 'corr' package of R program and stretch function [31] for all 176 MDs. Based on the guidances of Pearson correlation coefficient [32], I set the threshold at  $r = 0.7$  for the "Highly correlated" of MDs on the present study. The 176 MDs were

divided into two groups by this threshold  $r \geq 0.7$ . The MDs in the combinations of  $r \geq 0.7$  are classified as “Strongly correlated MD” and the other MDs are “Weakly correlated MD group” in the present study.

**2nd step: Pick up the MDs with weak correlations with other MDs**

The second step is pick up the MDs of weak correlation(i.e.  $r < 0.7$ ) with any other MDs. In this chapter, two conditions of hierarchical cluster analysis is discussed, number of clusters (k) at 13 and 3. These MDs are grouped as “MD-c3a” for k=14 and “MD-c5a” for k=3 which were used for regression analysis of machine learning later.

**3rd step: Pick up the MDs by hierarchical cluster analysis** The third step is the hierarchical cluster analysis according to the similarity in the molecular descriptor profile among 194 pesticides. Dendrogram of molecular descriptors were obtained with the parameters shown in Table 7.1.

Table 7.1. Hierarchical Cluster Analysis parameters

Parameter	Value
Distance metric	Euclidian distance
Linkage criteria	Ward.D2

After obtaining the dendrogram of MDs, the optimum number of clusters (k) from the dendrogram were determined using the NbClust package [34]. The MDs from each cluster is picked up based on the following criteria.

**1. Pick up the MD-MD pair at the lowest distance in the cluster on the dendrogram**

Select the MD-MD pair of the most similar from each cluster according to the dendrogram as shown in the Figure 7.2.

**2. Pick up the MD that includes more information than the other**

Pick up the MD that is more range of z-score of 194 pesticides



Pick up the MD with real number rather than integer

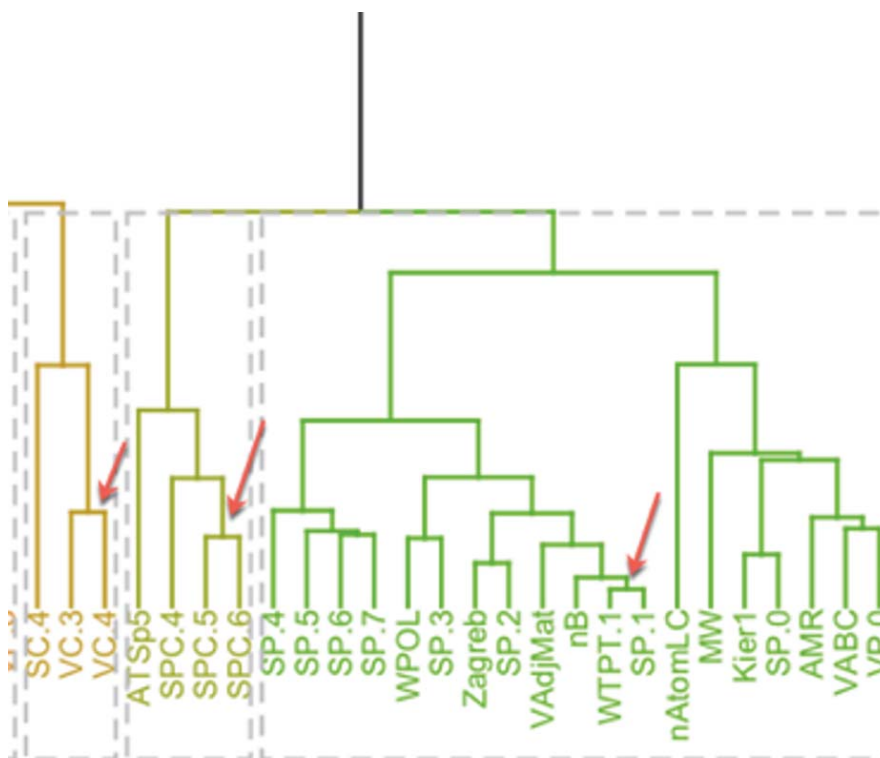


Figure 7.2. Selection of MD-MD pair from the dendrogram

**4th step: Confirm the correlation of MDs picked up by hierarchical cluster analysis** The fourth step using hierarchical cluster analysis is to confirm the correlation of MDs among picked up from each cluster. Threshold of correlation is set at  $r \geq 0.7$ .

**5th step: Finalize the MDs for prediction** The fifth step is the final selection the MD(s) based on the step 4. The MDs of weak correlation with other MDs(i.e.  $r < 0.7$ ) in step 4 are grouped as “MD-c3a” for  $k=13$  and “MD-c5a” for  $k=3$ , which are used for regression analysis of machine learning later. The

MDs of the combination with the strong correlation in step 4 are divided into two groups, i.e. “MD-c3b” and “MD-c3c” for  $k=13$  and “MD-c5b” and “MD-c5c” for  $k=3$ . MDs in MD-c3c ( $k=13$ ) and MD-c5c ( $k=3$ ) were excluded from further regression analysis. Thus, 176 MDs are divided into three groups as listed in the table according to the process in Figure 7.3.

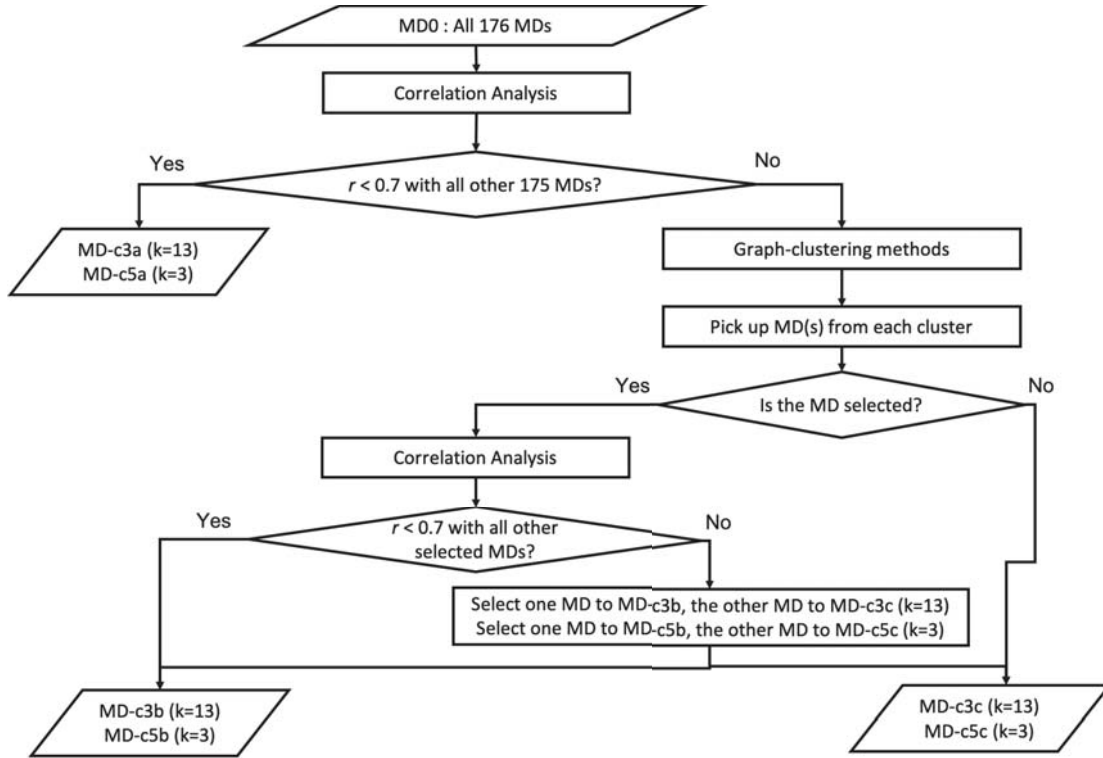


Figure 7.3. Process chart of selecting the optimum MDs

### 7.2.2 Machine learning by ‘caret’ package

Accuracy and Execution Time(ET) were measured the same procedure as Chapter 2 for 119 classification machine learning methods, then these performances were compared.

## 7.3 Results and Discussion

### 7.3.1 Correlation analysis results among molecular descriptors

Figure 7.4 shows the histogram of correlation distribution among the 176 MDs. The x-axis is the Pearson correlation coefficient of the combination and y-axis is the count(frequency) in each bin.

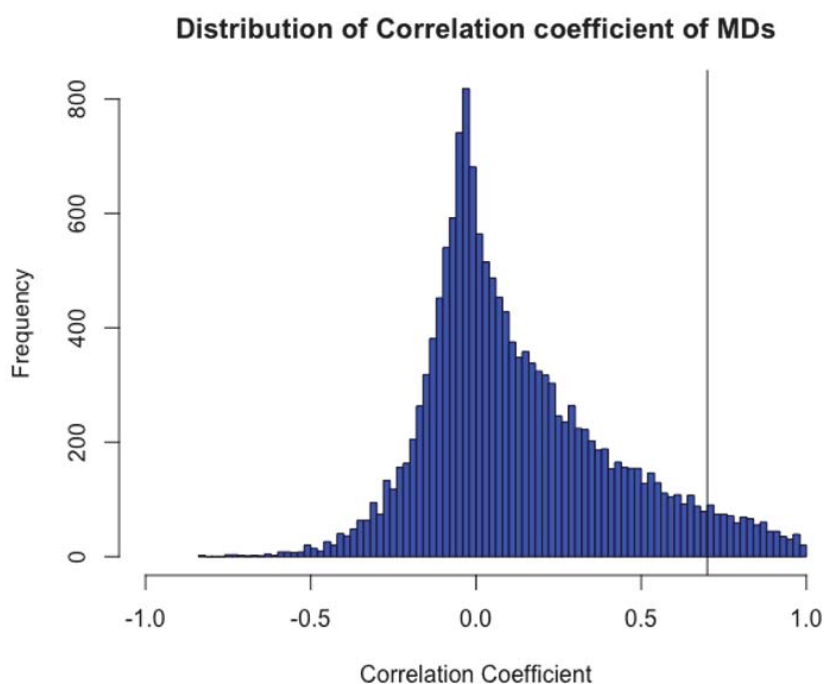


Figure 7.4. Distribution of correlation coefficient of MD

### 7.3.2 Selection of molecular descriptors by the clustering tool - hierarchical cluster analysis

There are 3365 MD-MD combinations of  $r \geq 0.7$  which are consisted of 130 MDs. 46 MDs out of 176 MDs were classified as MD-c1a in the Table 7.3.

Table 7.2. Group of MDs for optimum selection using hierarchical cluster tree

MD group	MDs to be classified
MD-c3a	MD of $r < 0.7$ with any of other 175 MDs(Weakly correlated)
MD-c3b	MD of $r \geq 0.7$ with any of other 175 MDs(Strongly correlated) and selected by hierarchical cluster analysis at $k=13$
MD-c3c	MD of $r \geq 0.7$ with any of other 175 MDs(Strongly correlated) and excluded by hierarchical cluster analysis at $k=13$
MD-c5a	MD of $r < 0.7$ with any of other 175 MDs(Weakly correlated)
MD-c5b	MD of $r \geq 0.7$ with any of other 175 MDs(Strongly correlated) and selected by hierarchical cluster analysis at $k=3$
MD-c5c	MD of $r \geq 0.7$ with any of other 175 MDs(Strongly correlated) and excluded by hierarchical cluster analysis at $k=3$

**Determination of number of clusters on the dendrogram by gap function** Dendrogram of 130 MDs according to the similarity of 194 pesticides was shown in the Figure 7.5.

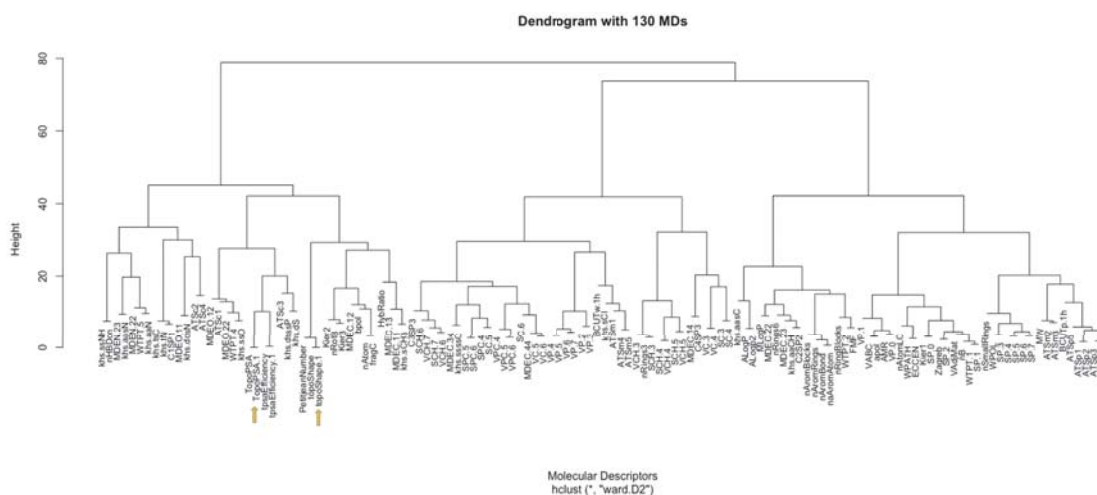


Figure 7.5. Dendrogram of 130 MDs according to the similarity of 194 pesticides

2 MDs out of 130 MDs were excluded when implementing the NbClust package of R program due to very strong correlations with the other MDs. Therefore, 128

MDs were used for determination of number of cluster on hierarchical cluster analysis using NbClust. Dendrogram of 128 MDs was shown in the Figure 7.6.

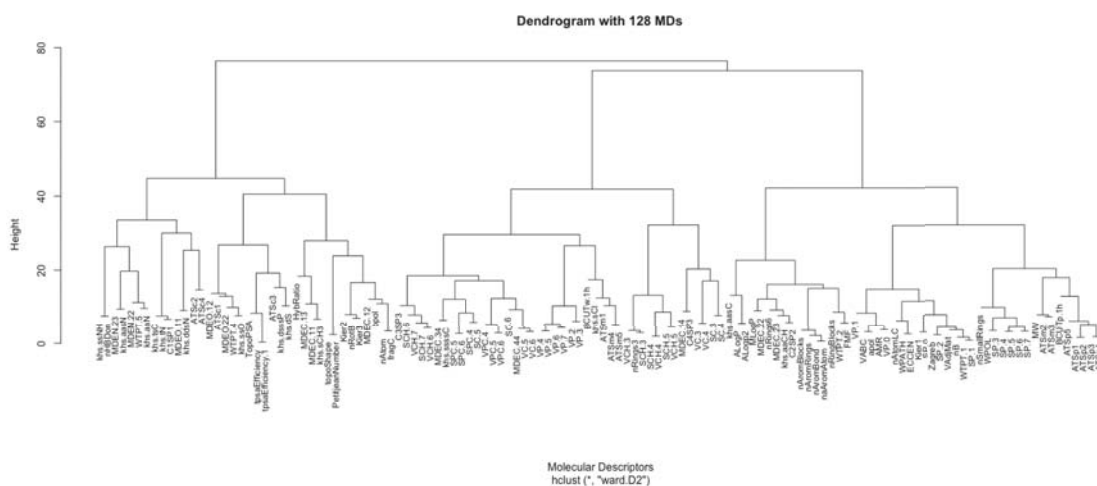


Figure 7.6. Dendrogram of 128 MDs

With the NbClust package, the result of optimum number of cluster was suggested in the histogram (Figure 7.7), numbers of cluster (k) 3 and 13 were optimum according to the gap function. Parameters of NbClust is listed in the Table 7.3.

Table 7.3. NbClust parameters

Parameter	Value
Distance	Euclidean
Minimal number of clusters	2
Maximal number of clusters	13
Cluster analysis method	ward.D2

Based on the strategy of MD selection for minimizing the loss of information by MD selection, two numbers of clusters at k=13 and k=3 were discussed for comparisons of Accuracy and Execution Time (ET) of pesticide classification prediction.

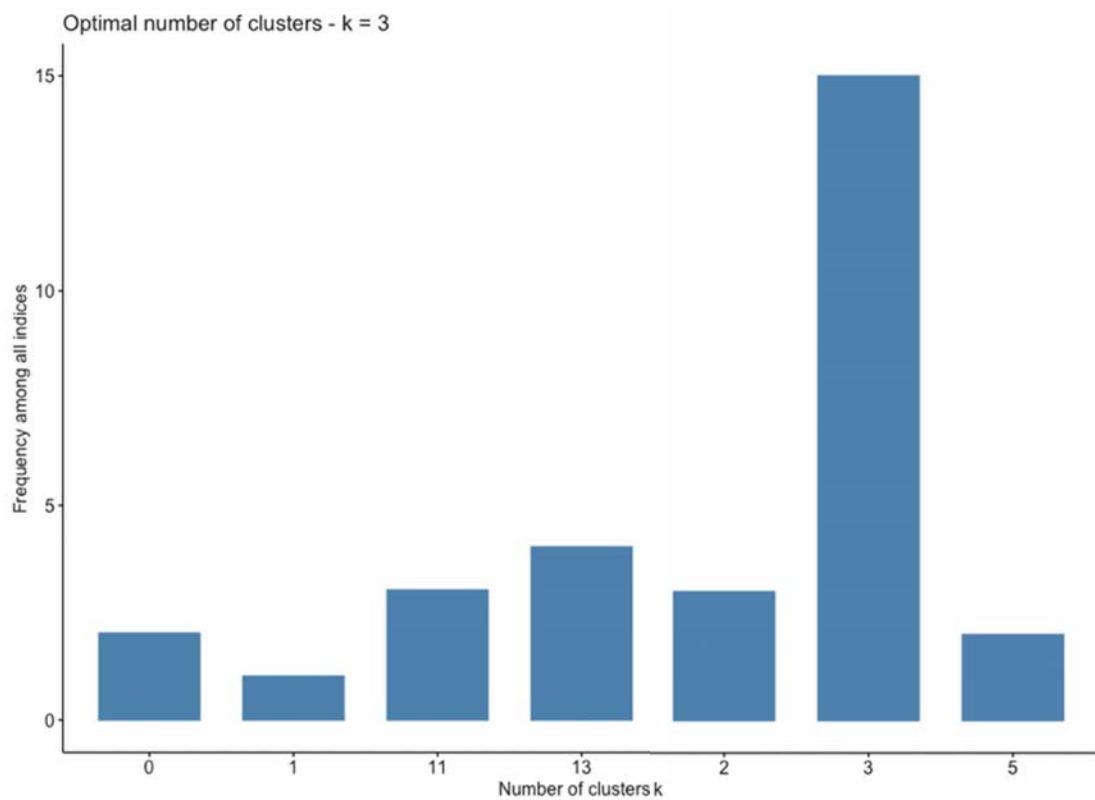


Figure 7.7. Histogram of the result of optimum number of cluster according to the gap function

**Molecular descriptor group in this chapter** Table 7.4 describes the name of MD on this chapter for comparison of machine learning prediction performance, based on the selection of MDs by clustering analysis, including the original MDs without the selection of MDs.

Table 7.4. Molecular Descriptor group in this chapter

Molecular descriptor group	Molecular descriptors
MD0	All MDs obtained by rcdk package (Chapter 5)
MD2	MDs selected by DP Clus (Chapter 6)
MD4	MDs selected by hierarchical cluster analysis with k=13
MD6	MDs selected by hierarchical cluster analysis with k=3

**Selection of MDs from the dendrogram at k=13 (MD4)** Figure 7.8 shows the dendrogram of the MDs with number of cluster at 13.

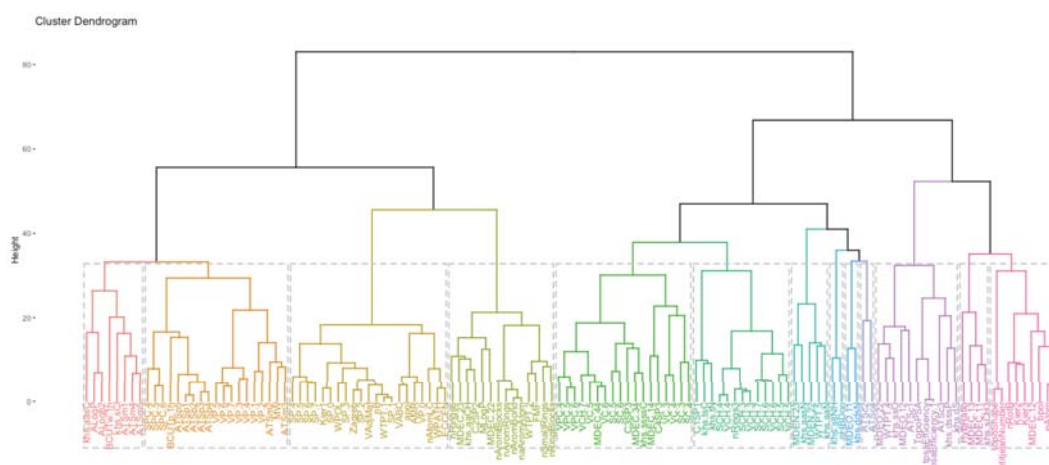


Figure 7.8. Dendrogram of MDs at k=13

26 MDs from 128 MDs were selected as shown in the Table 7.5. 13 MDs were picked up as the candidate of machine learning for pesticide recovery prediction.

Table 7.5. Molecular Descriptors selected by the correlation analysis and cluster analysis (k=13)

MD	Cluster	Integer or Real number	Z-score range	Final candidates	Final Selection
ATSm4	1	R	6.077	x	
ATSm5	1	R	5.866		
ATSp1	2	R	5.225		
ATSp2	2	R	5.602	x	x
topoShape	3	R	5.117		
PetitjeanNumber	3	R	5.98	x	x
nAromBond	4	I	4.405	x	
naAromAtom	4	I	4.242		
WTPT.1	5	R	5.616		
SP.1	5	R	5.656	x	x
SCH.4	6	R	5.275		
VCH.4	6	R	5.679	x	x
tpsaEfficiency	7	R	5.229		
tpsaEfficiency.1	7	R	5.249	x	x
WTPT.5	8	R	4.037	x	x
khs.aaN	8	I	3.364		
MDEC.11	9	R	4.075	x	x
khs.sCH3	9	I	3.831		
VPC.5	10	R	6.158		
VPC.6	10	R	6.486	x	x
MDEO.11	11	R	5.266	x	x
khs.ddsN	11	I	5.542		
khs.ssNH	12	I	3.426	x	x
nHBDon	12	I	3.263		
ATSc2	13	R	4.871		
ATSc4	13	R	5.982	x	x

The combination of VPC.6 - ATSm4 ( $r = 0.8617$ ), ATSp2 - ATSm4 ( $r = 0.7904$ ) and SP.1 - nAromBond ( $r = 0.7176$ ) were correlated strongly. ATSm4 and nAromBond were selected from these combinations because their range of z-score is less than the other MDs.

11 MDs were finally selected for machine learning of pesticide recovery prediction.

Table 7.6 is the summary of MDs selected by the decision tree Figure xx.



Table 7.6. Summary of molecular descriptors selected by the correlation analysis and cluster analysis (k=13)

MD group	Description of MDs	Number of MDs	Selected
MD-c3a	MD of $r < 0.7$ with any of other 175 MDs	46	Yes
MD-c3b	MD of $r \geq 0.7$ with any of other 175 MDs and selected by graph-clustering method	11	Yes
MD-c3c	MD of $r \geq 0.7$ with any of other 175 MDs and excluded by graph-clustering method	119	No

Thus, MD-c3a and MD-c3b(both MDs are combined as the MD group of “MD4”) will be used for the classification analysis by caret. Here, all 176 MDs(original MDs discussed in Chapter 5) are expressed as “MD0” for comparison of the performance against MD4.

**Accuracy and ET result in MD4** Figure 7.9 and Figure 7.10 shows the plot of Accuracy and Log(ET) of 119 machine learning methods with MD4.

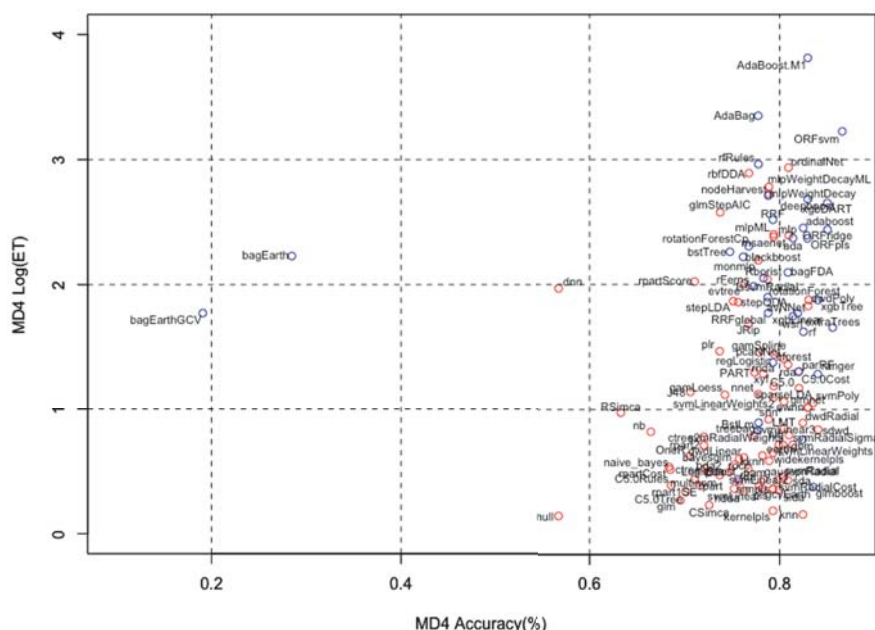


Figure 7.9. Accuracy and Log(ET) of 119 machine learning methods with MD4. Plot in red is ordinary method and blue is ensemble method.

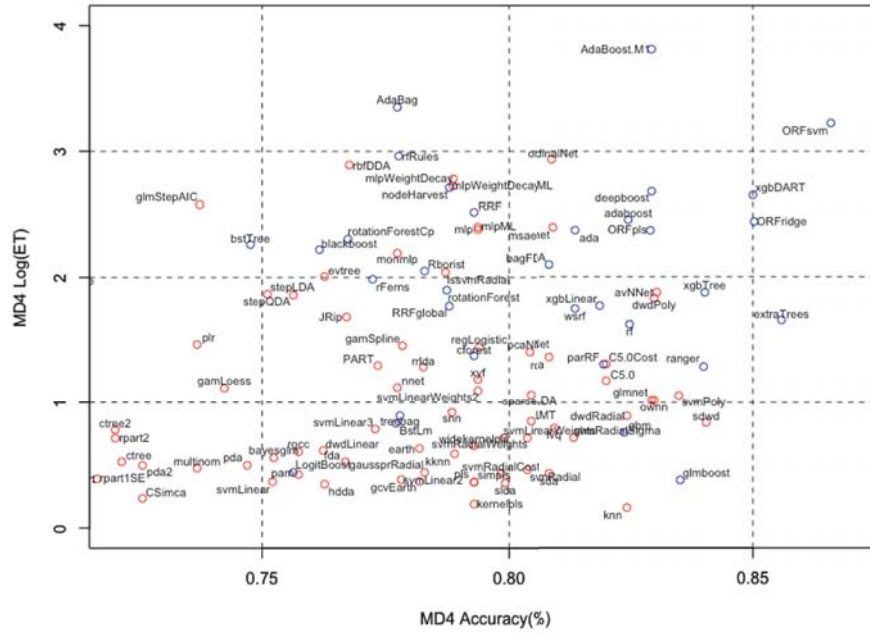


Figure 7.10. Accuracy and Log(ET) of 119 machine learning methods with MD4 in expanded view. Plot in red is ordinary method and blue is ensemble method.

**Comparison of Accuracy among MD4, MD2 and MD0** Figure 7.11 and Figure 7.12 show the comparison of Accuracy among MD4, MD2 and MD0. There was no difference in Accuracy between MD4 and MD2 for all 119 machine learning methods.

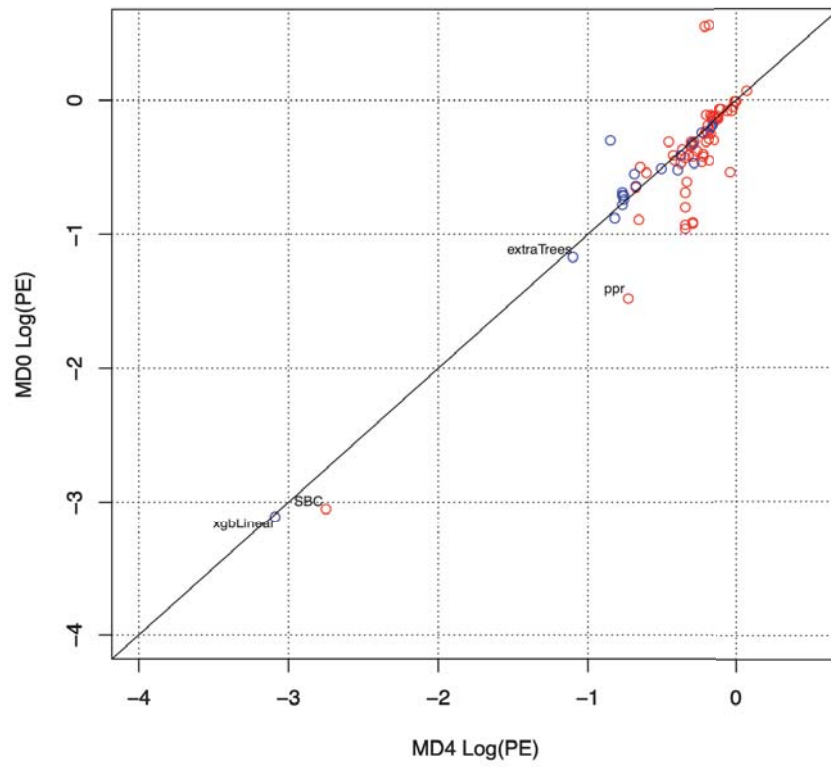


Figure 7.11. Accuracy comparison between MD4 and MD0. Plot in red is ordinary method and blue is ensemble method.

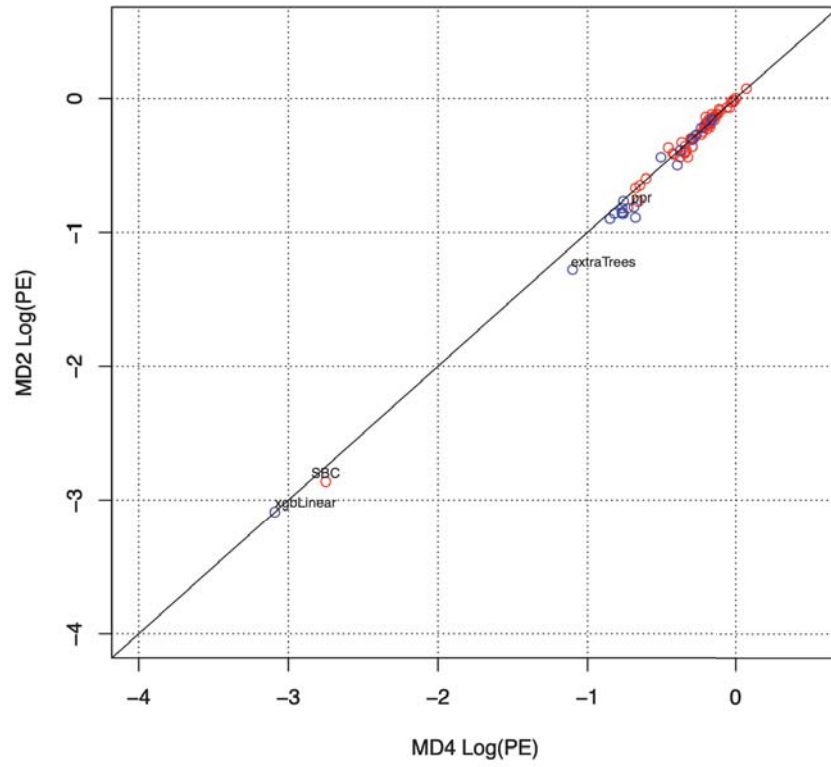


Figure 7.12. Accuracy comparison between MD4 and MD2. Plot in red is ordinary method and blue is ensemble method.

Tuning parameters of xgbDART are listed in the Table 7.7.

Table 7.7. Tuning parameters of xgbDART for pesticide classification prediction. (MD4)

Parameter	Value
Boosting Iterations	50
Maximum tree depth	2
Eta (Shrinkage)	0.4
Gamma (Minimum loss Reduction)	0
Subsample (Subsample percentage)	0.5
Colsample by tree (Subsample ratio of columns)	0.8
Rate drop (Fraction of trees dropped)	0.01
Skip Drop (Probability of skipping drop-out)	0.95
Minimum child weight (Minimum sum of instance weight)	1

**Selection of MDs from the dendrogram at k=3 (MD6)** Figure 7.13 shows the dendrogram of the MDs with number of cluster at 13.

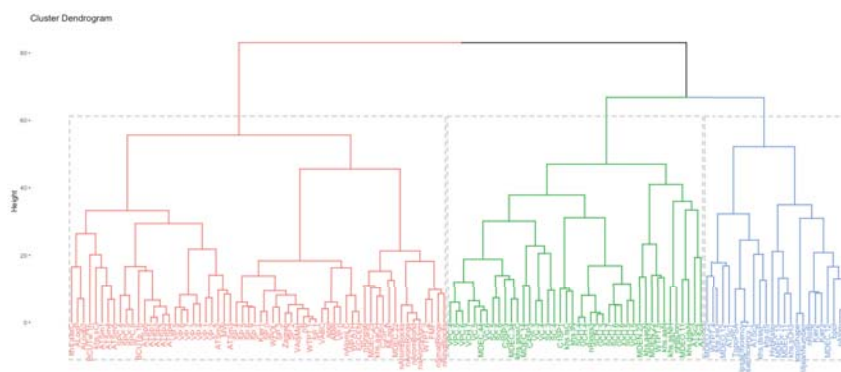


Figure 7.13. Dendrogram of MDs at k=3

3 MDs from 128 MDs were selected as shown in the table xx. These MDs correlate weakly, one another so all of them were selected for machine learning of pesticide recovery prediction.

Table 7.8 is the summary of MDs.

Table 7.8. Summary of molecular descriptors selected by the correlation analysis and cluster analysis (k=3)

MD group	Description of MDs	Number of MDs	Selected
MD-c5a	MD of $r < 0.7$ with any of other 175 MDs	46	Yes
MD-c5b	MD of $r \geq 0.7$ with any of other 175 MDs and selected by graph-clustering method	3	Yes
MD-c5c	MD of $r \geq 0.7$ with any of other 175 MDs and excluded by graph-clustering method	127	No

**Accuracy and ET result in MD6** Figure 7.14 and Figure 7.15 shows the plot of Log(PE) and Log(ET) of 119 machine learning methods with MD6.

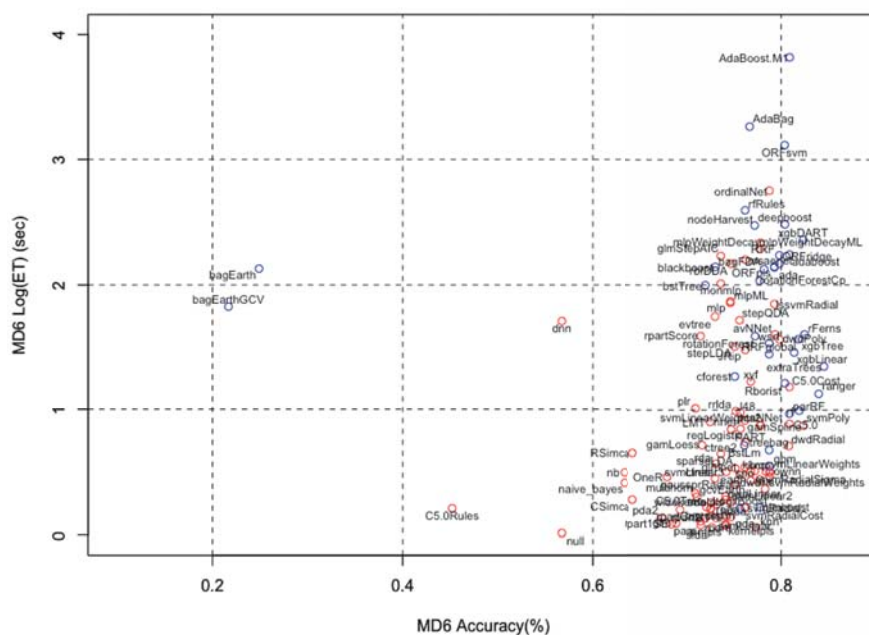


Figure 7.14. Accuracy and Log(ET) for 119 methods using the MD6. Plot in red is ordinary method and blue is ensemble method.

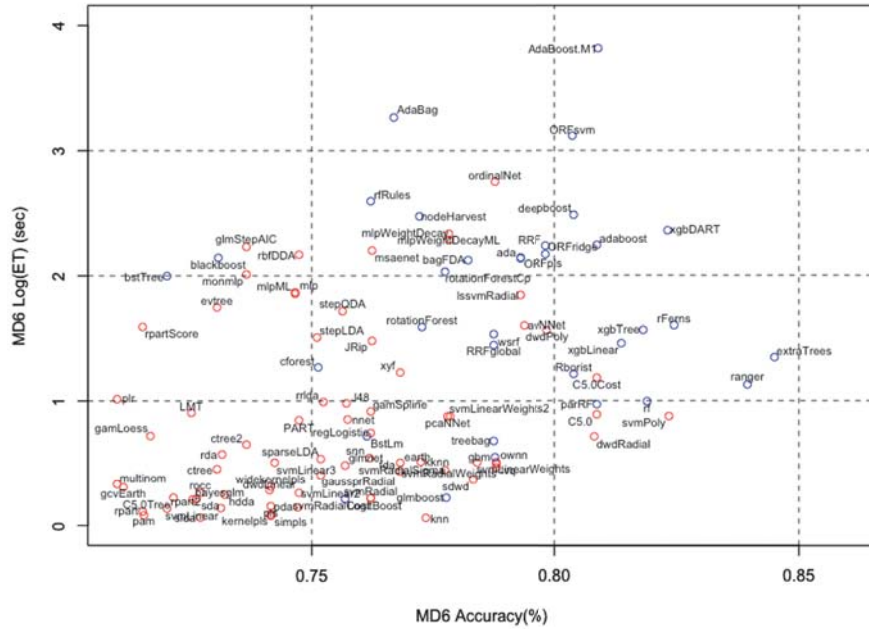


Figure 7.15. Accuracy and Log(ET) for 119 methods using the MD6 (expanded view). Plot in red is ordinary method and blue is ensemble method.

**Comparison of machine learning performance among MD6, MD4, MD2 and MD0** Figure 7.16, Figure 7.17 and Figure 7.18 shows the comparison of Accuracy among MD6, MD4, MD2 and MD0. There was no difference in Accuracy among MD6, MD4 and MD2 for all 119 machine learning methods.

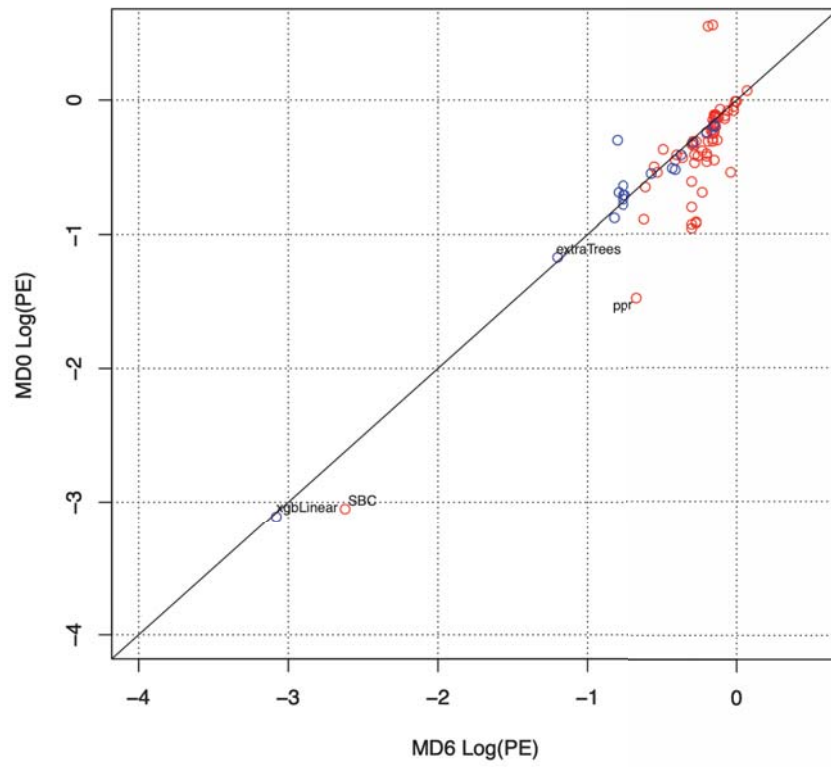


Figure 7.16. Accuracy comparison between MD6 and MD0. Plot in red is ordinary method and blue is ensemble method.



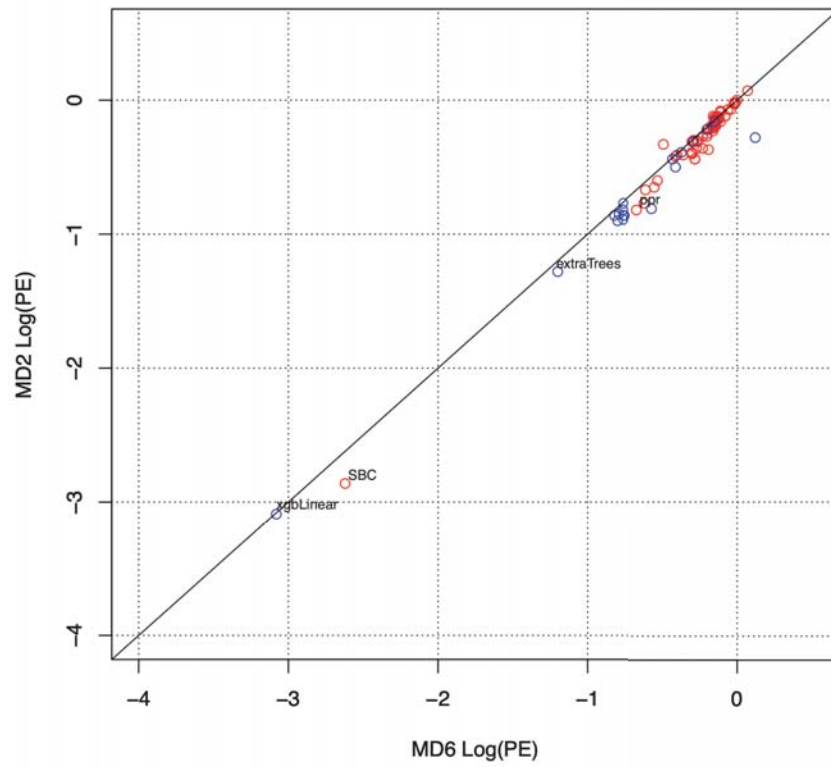


Figure 7.17. Accuracy comparison between MD6 and MD2. Plot in red is ordinary method and blue is ensemble method.

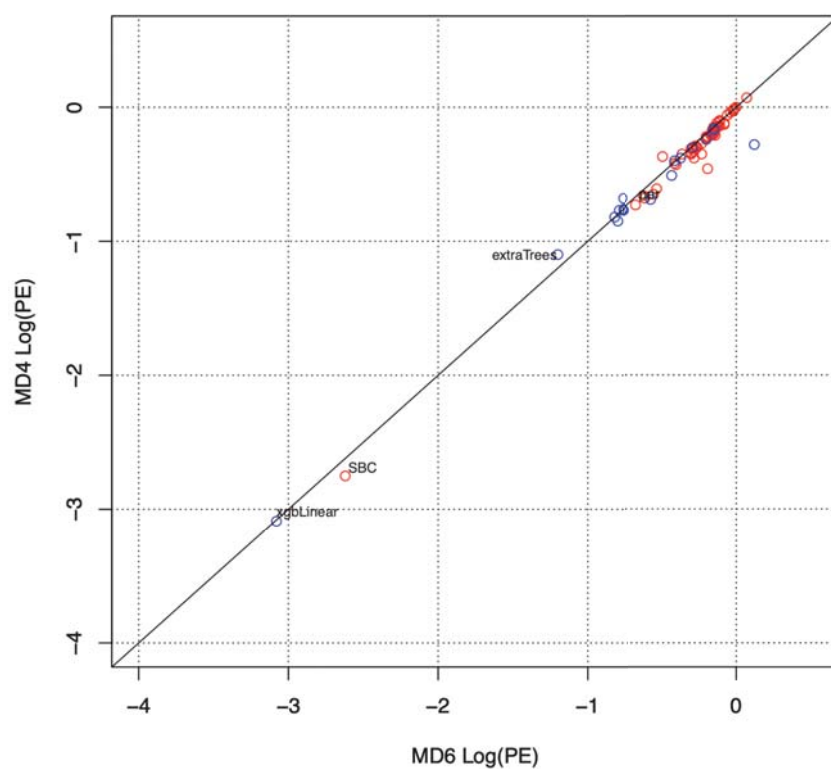


Figure 7.18. Accuracy comparison between MD6 and MD4. Plot in red is ordinary method and blue is ensemble method.

Tuning parameters of xgbDART are listed in the Table 7.9.

Table 7.9. Tuning parameters of xgbDART for pesticide classification prediction.(MD6)

Parameter	Value
Boosting Iterations	50
Maximum tree depth	3
Eta (Shrinkage)	0.4
Gamma (Minimum loss Reduction)	0
Subsample (Subsample percentage)	0.75
Colsample by tree (Subsample ratio of columns)	0.6
Rate drop (Fraction of trees dropped)	0.01
Skip Drop (Probability of skipping drop-out)	0.95
Minimum child weight (Minimum sum of instance weight)	1

### 7.3.3 Comparison of Accuracy among MD6, MD4, MD2 and MD0 for machine learning method category

**Comparison of PE** Figure 7.19 shows the box plots of Accuracy for machine learning methods compared among MD6, MD4, MD2 and MD0. sorted by the machine learning category Results of the Accuracy by machine learning categories among MD6, MD4, MD2 and MD0 are listed in the Table 7.10.

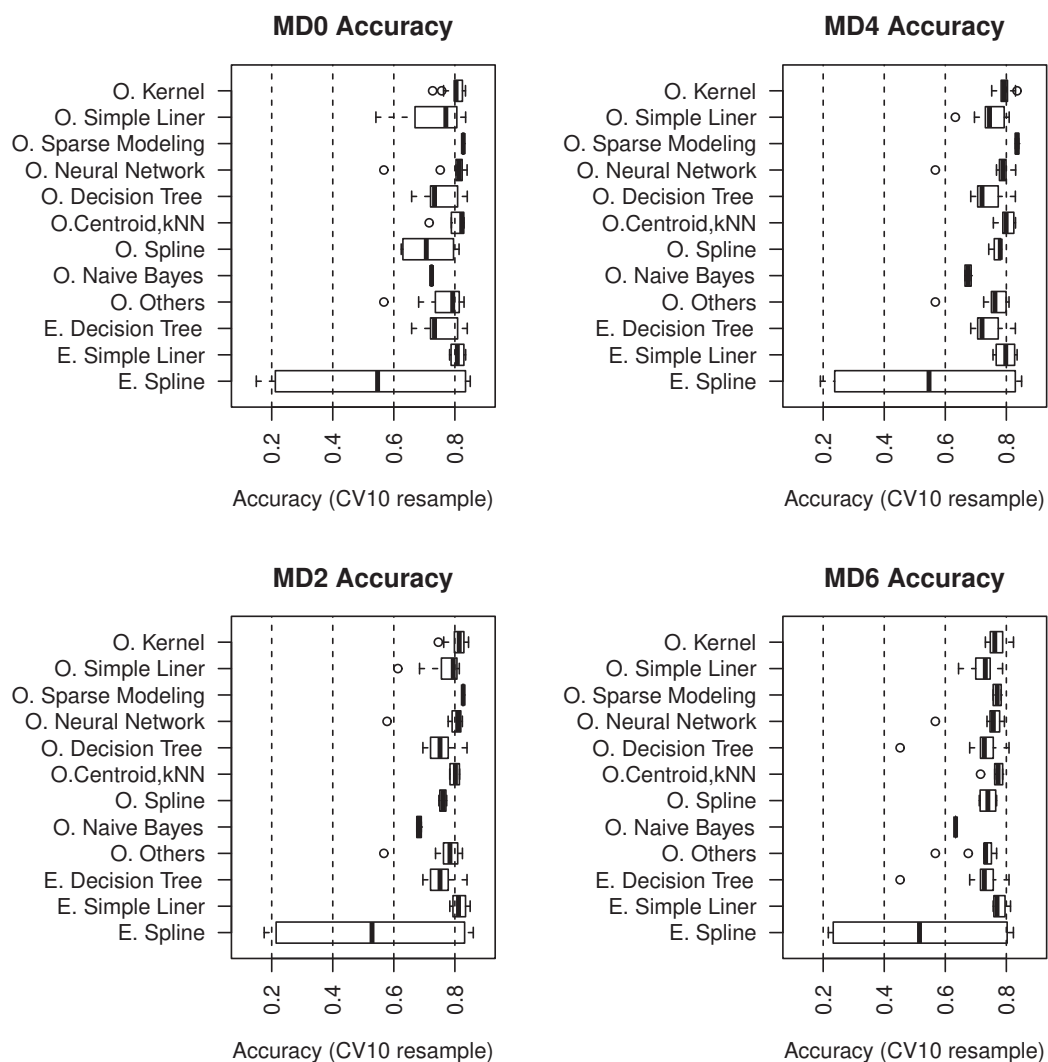


Figure 7.19. Comparison of Accuracy by learning method category among MD6, MD4, MD2 and MD0

Table 7.10. Results of machine learning method accuracy category among the MD groups

Results	Machine learning method category
Poor performance in Accuracy on all MDs	O. Kernel, O. Simple Linear, O. Sparse Modeling, O. Neural Network, O. Decision Tree, O. Spline, O. Others, E. Decision Tree and E. Spline
O. Simple Linear	Good performance in MD0, worse by reduction of MDs
Some method(s) performed well but others were poor in all MDs	O. centroid, kNN and E. Simple Liner

**Comparison of accuracy** Figure 7.20 shows shows the ET of extraTree (MD6), ORFsvm (MD4), Adaboost (MD2) and xgbDART (MD0) with 5 replicates.

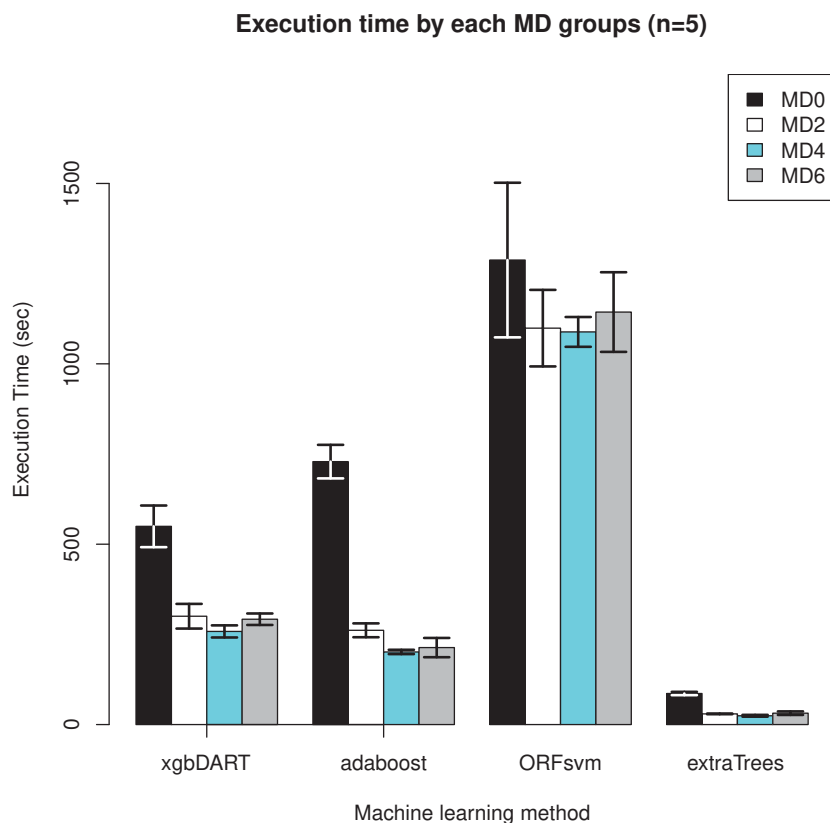


Figure 7.20. ET of best machine learning method of MD6, MD4, MD2 and MD0 with 5 replicates

## 7.4 Conclusion of Chapter 7 and pesticide classification prediction models in caret package

In Chapter 7, hierarchical cluster analysis was used to select the molecular descriptors for prediction model of pesticide recovery. Considering the results, ORFsvm (Oblique Random Forest with Support Vector Machine) was best accuracy (86.6%) by MD4 (k=13) and extraTrees (Random Forest by Randomization, 84.5%). Execution time was reduced for most machine learning methods by reducing the molecular descriptors. All of these machine learning methods were

the ensemble decision tree category. The best accuracy among the MD0, MD2, MD4 and MD6 for this data set was the ORFsvm with MD4, 86.6 % accuracy, but it requires 1000 seconds (about 16 minutes) for execution time to build the machine learning model. If 16 minutes for execution time is acceptable, ORFsvm with MD4 (hierarchical cluster analysis with  $k=13$  is the optimum condition. As discussed in the Chapter 5, heteroatoms in the pesticide molecule also gives wrong classification. Classification of the pesticides such as the common structure will be required for more accurate prediction.

## 8. Concluding remarks

To ensure the food safety of global trade, analyzing the multiresidue pesticides in crops are needed. LC-MS and GC-MS are widely used as the tool to detect the multiresidues in crops in the world, but there are technical challenges such as altering the recovery rate due to the called “Matrix Effect” , a chemical reaction between the residual pesticides and sample matrix(compounds in the sample other than pesticides) even the clean up treatment before injection to GC-MS and LC-MS and limitation of separation by liquid chromatography and gas chromatography. Prediction of pesticides are needed because not all Food Labs own both instruments, i.e. using either LC-MS or GC-MS. Managers in Food Lab need to consider the return of investments for owning both instruments due to the initial cost (100k USD for conventional GC-MS and 250k USD for conventional LC-MS), therefore there are common cases for chemists of Food Labs to analyze the additional pesticides newly with current owned instrument. Researches of residual pesticides in food are mainly method validation, the sample preparation and type of crops for example. This study utilizes the results of such validation results for prediction of pesticides detection by LC-MS and GC-MS, using the molecular descriptors and machine learning.

In Chapter 2, I developed the the prediction of residual pesticide recovery rate of GC-MS by the 89 regression methods for seven crops with 178 molecular descriptors. Prediction error and execution time were used for performance measure of machine learning method. Prediction error of machine learning methods in the ordinary sparse model showed divergence value. The prediction error showed variances among seven crops for some machine learning methods, thus the generalization performance error( $PE_k$ ) using each rank of prediction error in the crop was used for reasonable comparison of machine learning methods in this study.  $PE_k$  showed the good correlation with the average of prediction error while  $PE_k$  can easily interpret the order of machine learning method prediction capability across the seven crops. By using the machine learning method with the molecular descriptors, SBC (Subtractive and K-means Fuzzy Clustering Method) and xgbLinear (eXtreme Gradient Boosting Linear) predicted the recovery rate of pesticides within 2 minutes for seven crops.

In Chapter 3, the removal of highly correlated molecular descriptors on the



data set of Chapter 2 was discussed using the graph clustering tool. In order to reduce the molecular descriptors with high correlation and minimize the loss of information by selecting the molecular descriptors, I created the process of molecular descriptors selection. Correlation analysis was applied on all 178 molecular descriptors. 658 combinations of 118 molecular descriptors showed higher correlation with the Pearson's correlation coefficient  $\geq 0.7$ . 83 molecular descriptors (60 molecular descriptors of low correlation with other molecular descriptors and 23 molecular descriptors selected by the cluster analysis) were used in the regression analysis of 89 machine learning methods for seven crops. Selection of molecular descriptors reduces the execution times on most machine learning methods. The effect on the prediction error varies on the machine learning methods. 57 methods of regression get better by the molecular descriptor selection, and 32 methods got worse. Prediction error of xgbLiner (eXtreme Gradient Boosting Linear) was the same with and without the selection of molecular descriptors. On the other hand, prediction error of SBC (Subtractive and K-Means Fuzzy Clustering Method) got worse by selecting molecular descriptors due to the reduction of explanatory parameter for machine learning.

In Chapter 4, Hierarchical cluster analysis was used to select the molecular descriptors for prediction model of pesticide recovery. Considering the results, xgbLiner (eXtremely Gradient Boosting Linear) performs excellently for prediction of pesticide recovery rate for this data set. Execution time of building the prediction model of xgbLiner is reduced when MDs are reduced while maintaining the performance of PE for prediction of pesticide recovery. Optimum condition is the xgbLiner with the selection of MDs by hierarchical cluster analysis. For prediction of pesticide recovery for this data set, xgbLiner with the hierarchical cluster analysis with  $k=3$  is the optimum condition for this data set.

In Chapter 5, I developed the classification method of pesticides amenability between LC and GC, based on the conversations with food analysis chemists. 119 machine learning methods for classification was investigated using 176 molecular descriptors obtained by the 194 pesticides of two validation reports. Prediction accuracy and execution time are the measure of the machine learning method performance. The accuracy of prediction ranged from 15% to 85%. The recommended machine learning method in the Chapter 4 is xgbDART (eXtreme Gradi-

ent Boosting and Additive Regression Tree) with 85% accuracy that requires less than 2 minutes for execution. The common chemical structure of wrongly classified pesticides by xgbDART was the presence of heteroatoms such as nitrogens, oxygens and halogens. Decision tree by 'rpart' package of R program showed XLogP is the major molecular descriptor to classify the pesticides for this data set. The distributions of XLogP showed that all LC-MS pesticides and all GC-MS pesticides were opposite to that of wrongly classified pesticides. Considering the importance of LogP (Water - Octanol coefficient) among molecular descriptors, removal of XLogP from machine learning will not be the right solution which gives other wrong classification of pesticides. For improving the accuracy of the classification, pre-classification of pesticides according to the chemical class is required instead of classifying the pesticides by the prediction model with all pesticides.

In Chapter 6, graph clustering tool was used to select the molecular descriptors for prediction model of pesticide recovery. DP Clus successfully classify the strongly correlated molecular descriptors into 25 clusters and molecular descriptors from each cluster were selected. Correlation analysis of the selected molecular descriptors was required because some clusters independently connect in the MD-MD network. By selecting the molecular descriptors, execution time for building prediction model was shorten for most machine learning methods. Accuracy of adaboost (AdaBoost Classification Trees) was improved (83.9% to 86.0%) by selecting the MDs with DP Clus.

In Chapter 7, hierarchical cluster analysis was used to select the molecular descriptors for prediction model of pesticide recovery. Considering the results, ORFsvm (Oblique Random Forest with Support Vector Machine) was best accuracy (86.6%) by MD4 (k=13) and extraTrees (Random Forest by Randomization, 84.5%). Execution time was reduced for most machine learning methods by reducing the molecular descriptors. All of these machine learning methods were the ensemble decision tree category. The best accuracy among the MD0, MD2, MD4 and MD6 for this data set was the ORFsvm with MD4, 86.6 % accuracy, but it requires 1000 seconds (about 16 minutes) for execution time to build the machine learning model. If 16 minutes for execution time is acceptable, ORFsvm with MD4 (hierarchical cluster analysis with k=13 is the optimum condition for this data set.

The knowledge of residual pesticide analysis in foods is increasing in the era of big data. By using the machine learning tool like the approaches in this study, such pile of data on the pesticides can be utilized for regression analysis of pesticides recovery and classifications. The classification of the pesticides can be enhanced by increasing the number of pesticides from other validation reports and further enhancement of machine learning algorithms.

## URLs

### **PubChem Website**

<https://pubchem.ncbi.nlm.nih.gov>

### **Online SMILES Translator and Structure File Generator**

<https://cactus.nci.nih.gov/translate/>

### **The Comprehensive R Archive Network**

<https://cran.r-project.org>

### **The caret Package on github**

<http://topepo.github.io/caret/index.html>

### **Daylight Theory Manual**

<https://www.daylight.com/dayhtml/doc/theory/>

### **Molecular Descriptors on OCHEM**

<http://wiki.qspr-thesaurus.eu/w/CDK.html>

### **IUPAC agrochemicals**

<http://agrochemicals.iupac.org>

## References

- [1] L.G.Costa, C.L.Galli, S.D.Murphy. 1987. Toxicology of Pesticides: A Brief History. *Toxicology of Pesticides* 1–10.
- [2] H. OHTA. 2013. Historical Development of Pesticides in Japan. *Survey Reports on the Systemization of Technologies* **18**: 1–6.
- [3] IUPAC Agrochemicals <http://agrochemicals.iupac.org/index.php>(accessed on May 1st, 2020).
- [4] Alimentarius Commission, Codex general standard for contaminants and toxins in food and feed(CODEX STAN 193-1995).
- [5] SANCO. 2009. METHOD VALIDATION AND QUALITY CONTROL PROCEDURES FOR PESTICIDE RESIDUES ANALYSIS IN FOOD AND FEED. Document No. SANCO/10684/2009.
- [6] Ministry of Health, Labour and Welfare in Japan. 2006. Introduction of the Positive List System for Agricultural Chemical Residues in Foods.
- [7] B. K. Matuszewski, M. L. Constanzer and C. M. Chavez. 2003. Strategies for the Assessment of Matrix Effect in Quantitative Bioanalytical Methods Based on HPLC-MS/MS. *Eng, Anal. Chem.* **75**: 3019–3030.
- [8] A. Krueve, A. Künnäpas, K. Herodes and I. Leito. 2008. Matrix effects in pesticide multi-residue analysis by liquid chromatography mass spectrometry. *Journal of Chromatography A* **1187**: 58–66.
- [9] José Fenoll, Pilar Hellín, Carmen M. Martínez and Pilar Flores. 2004. Multiresidue Analysis of Pesticides in Vegetables and Citrus Fruits by LC-MS-MS. *Chromatographia*, **72**: 857–866.
- [10] J.J. Kirkland. 2004. Development of some stationary phases for reversed-phase high-performance liquid chromatography. *Journal of Chromatography A* **1060**: 9–21.

- [11] Z. Barganska, P. Konieczka and J. Namiesnik. 2018. Comparison of Two Methods for the Determination of Selected Pesticides in Honey and Honeybee Samples. *Molecules* **23**: 2582.
- [12] P. He and D. S. Aga. 2019. Comparison of GC-MS/MS and LC-MS/MS for the analysis of hormones and pesticides in surface waters: advantages and pitfalls. *Anal. Methods*, **11**: 1436.
- [13] Y. C., Lo, S. E. Rensi, W. Torng and R. B. Altman. 2018. Machine learning in chemoinformatics and drug discovery. *Drug Recovery Today* **23**: 1538–1546.
- [14] A. Pyka, M. Babuška, and M. Zachariasz. 2006. A COMPARISON OF THEORETICAL METHODS OF CALCULATION OF PARTITION COEFFICIENTS FOR SELECTED DRUGS. *Acta Poloniae Pharmaceutica - Drug Research* **63**: 159–167.
- [15] J. C. Dearden and T. Ghafourian. 1999. Hydrogen Bonding Parameters for QSAR: Comparison of Indicator Variables, Hydrogen Bond Counts, Molecular Orbital and Other Parameters. *J. Chem. Inf. Comput. Sci.* **39**: 231–235.
- [16] F. R. Burden. 1989. Molecular Identification Number for Substructure Searches. *J. Chem. Comput. Sci.* **29**: 225–227.
- [17] R. Guha. 2007. Chemical Informatics Functionality in R. *J. Statistical software* **18**: 1–18.
- [18] Thomas Engel and Johann Gasteiger. 2018. Chemoinformatics: Basic Concepts and Methods, First Edition 44–47.
- [19] M. Kuhn. 2008. Building Predictive Models in R Using the caret Package. *J. Statistical Software* **28**: 1–26.
- [20] S. Nakamura, T. Yamagami, Y. Ono, K. Toubou and S. Daishima. 2013. Multi-residue Analysis of Pesticides in Agricultural Products by GC/MS Using Synchronous SIM/Scan Acquisition *BUNSEKI KAGAKU* **62**: 229–241.

- [21] A. Butler, P. Hoffman, P. Smibert, P. Efthymia and R. Satija. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, **36**: 411–420.
- [22] R. Tibshirani. 1996. Regression Shrinkage and Selection Via the Lasso. *Royal Statistical Soc. Ser. B.* **58**: 267–288.
- [23] D. W. Marquardt and R. Snee. 1975. Ridge regression in practice. *Am. Stat.* **29**: 3–20.
- [24] N. S. Altman, 1992. An introduction to kernel and nearest-neighbor non-parametric regression. *Am. Stat.* **46**:, 175–185.
- [25] R.R. Yager and D. P. Filev. 1994. Generation of Fuzzy Rules by Mountain Clustering. *J. Intelligent Fuzzy Sys.* **2**: 209–219.
- [26] A. G. Asuero, A. Sayago and A. G. Gonzalez. 2006. The Correlation Coefficient: An Overview. *Critical Rev. Anal. Chem.* **46**: 41–59.
- [27] R. Bro, K. Kjeldahl, A. K. Smilde and H. A. L. Kiers. 2008. Cross-validation of component models: A critical look at current methods. *Anal. Bioanal. Chem.* **390**: 1241–1251.
- [28] K. Drab and M. Daszykowski. 2014. Clustering in Analytical Chemistry *J. AOAC Int.* **97**: 29–38.
- [29] L. I. Kuncheva and C. J. Whitaker. 2003. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning* **51**: 181–207.
- [30] A. Garg and K. Tai. 2013. Comparison of statistical and machine learning methods in modeling of data with multicollinearity. *International Journal of Modelling, Identification and Control(IJMIC)* **18**: 295–312.
- [31] M. Kuhn, S. Jackson and J. Cimentada. 2020. Package ‘corr’ version 0.4.2. *Documentation on CRAN*.
- [32] M.M. Mukaka. 2012. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J.* **24**: 69–71.

- [33] M. Altaf-Ul-Amin, Y. Shibo, K. Mihara, K. Kurokawa and S. Kanaya. 2006. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics* **7**: 207.
- [34] M. Charrad, N. Ghazzali, V. Boiteau and A. Niknafs. 2012. Package ‘NbClust’ version 1.0. *Documentation on CRAN*.
- [35] C. Anagnostopoulos and G.E.Miliadis. 2013. Development and validation of an easy multiresidue method for the determination of multiclass pesticide residues using GC-MS/MS and LC-MS/MS in olive oil and olives. *Talanta* **1121**: 1–10.
- [36] Pesticide Analytical Manual Vol. I, Appendix II, Food and Drug Administration. 1999.
- [37] N. Chamkasem, L. W. Ollis, T. Harmon, S. Lee and Greg Mercer. 2013. Analysis of 136 Pesticides in Avocado Using a Modified QuEChERS Method with LC-MS/MS and GC-MS/MS. *J. Agric. Food Chem.* **61**: 2315 –-2329.
- [38] EURL-FV(2012-M6) Validation Data of 127 Pesticides Using a Multiresidue Method by LC-MS/MS and GC-MS/MS in Olive Oil, EU Reference Laboratories for residues of pesticides. 2012.
- [39] Raschka S. 2018. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv :1811.12808*.



## Achievements

### List of publications and manuscripts in Preparation

Takeshi Serino, Yoshizumi Takigawa, Sadao Nakamura, Ming Huang, Naoaki Ono, Altaf-Ul-Amin, Shigehiko Kanaya. Chemoinformatics Approach for Estimating Recovery Rates of Pesticides in Fruits and Vegetables. 2019. *Journal of Computer Aided Chemistry* **20**: 92-103.

### International Conferences

1. Takeshi Serino, Sadao Nakamura, Yoshizumi Takigawa, Norton Kitagawa, Shigehiko Kanaya. 67th American Society for Mass Spectrometry. *Comprehensive Machine Learning Prediction of GC/MS Pesticide Recovery Based on the Molecular Fingerprinting for Food QA/QC*. Atlanta, GA. June, 2019.(in Chapter 2)
2. Takeshi Serino, Sadao Nakamura, Yoshizumi Takigawa, Tarun Anumol, Md. Altaf-UI-Amin, Shigehiko Kanaya. American Society for Mass Spectrometry 2020 Reboot. *Optimum molecular descriptors based on 89 machine learning methods for predicting the recovery rate of pesticides in crops by GC-MS*. United States. June, 2020.(in Chapter 3)
3. Sadao Nakamura, Takeshi Serino, Takeshi Otsuka, Yoshizumi Takigawa, Tarun Anumol, Shigehiko Kanaya. American Society for Mass Spectrometry 2020 Reboot. *Classifying the pesticides in foods between GC-amenable and LC-amenable using the prediction model with molecular descriptors*. United States. June, 2020.(in Chapter 4)

### Local Conferences

1. Takeshi Serino, Sadao Nakamura, Yoshizumi Takigawa, Shigehiko Kanaya. 第79回分析化学討論会. *Challenge of regression analysis of food data with machine learning*. Kokura. May, 2019

2. **Takeshi Serino**, Sadao Nakamura, Yoshizumi Takigawa, Shigehiko Kanaya. The 68th annual conference on Mass Spectrometry, Japan. *Data science approach on the Multi-compounds simultaneous analysis of mass spectrum data*. Tsukuba. May, 2019
3. **Takeshi Serino**, Sadao Nakamura, Yoshizumi Takigawa, Shigehiko Kanaya. The 68th annual conference on Mass Spectrometry, Japan. *Classification of pesticides amenability between LC or GC by 119 machine learning methods using the dataset of residual pesticides in foods of Japan Positive List*. May(2020)
4. **Takeshi Serino**, Kuniyo Sugitate, Seiya Tanaka, Hirokazu Sawada, Sadao Nakamura, Shigehiko Kanaya. 第 80 回分析化学討論会. *A study on the prediction method of mass spectrometer for pesticide residue analysis in food(LC/MS or GC/MS)*. May, 2020.
5. Kazuyuki Yamashita, Takashi Kasamatsu, Sadao Nakamura, **Takeshi Serino**, Shigehiko Kanaya. 日本農薬学会第 45 回大会. *Study on the variation of pesticide recovery rates in tomatoes and soybeans due to different pretreatment methods*. March, 2020.

## Acknowledgements

First, I'd like to acknowledge Professor Shigehiko Kanaya, who has been the advisor of the research in this dissertation. I decided to start this research by his advice at the conversation in Osaka. Professor Kanaya always supported me to complete this dissertation and I was always encouraged. I'm also thankful to Professor Keiichi Yasumoto, Associate Professor Naoaki Ono and Assistant Professor Ming Huang for reviewing this dissertation and advices. Their advices enhanced the contents of this thesis greatly. I'd also thank to Associate Professor Md. Altaf-Ul-Amin for advices and providing the DP Clus software. I'm impressed at the practical function and attractive user interface. Without DP Clus, I cannot achieve the optimum selection of molecular descriptors for machine learning.

I appreciate Dr. Sadao Nakamura of Agilent Technologies for providing the pesticide data and advices on the behavior of the pesticides on Mass Spectrometer. I'd also like to thank to Yoshizumi Takigawa, Norton Kitagawa, Tarun Anumol, Shweta Shukradas and other colleagues in Agilent for the advices based on their various backgrounds and proof readings of the articles.

Finally, I'd like to thank my family to support me. My wife, Hitoko always encourages me and brings me the different viewpoint through the daily conversations, which prevent me on sticking on the single aspect or idea.

# Appendix

## A. Additional information for regression analysis

### A.1 Pesticides and Canonical SMILES for regression analysis

Table A.1: Pesticides and Canonical SMILES for regression analysis

Pesticide name	Canonical SMILES
cyanofenphos	<chem>CCOP(=S)(C1=CC=CC=C1)OC2=CC=C(C=C2)C#N</chem>
Chinomethionat	<chem>CC1=CC2=C(C=C1)N=C3C(=N2)SC(=O)S3</chem>
Thiocyclam	<chem>CN(C)C1CSSSC1</chem>
Captan	<chem>C1C=CCC2C1C(=O)N(C2=O)SC(Cl)(Cl)Cl</chem>
Chlormephos	<chem>CCOP(=S)(OCC)SCCl</chem>
Butylate	<chem>CCSC(=O)N(CC(C)C)CC(C)C</chem>
Nereistoxin oxalate	<chem>C[NH+](C)C1CSSC1.C(=O)(C(=O)[O-])O</chem>
EPTC	<chem>CCCN(CCC)C(=O)SCC</chem>
Etridiazole	<chem>CCOC1=NC(=NS1)C(Cl)(Cl)Cl</chem>
Molinate	<chem>CCSC(=O)N1CCCCC1</chem>
bifenazate	<chem>CC(C)OC(=O)NNC1=C(C=CC(=C1)C2=CC=CC=C2)OC</chem>
Methacrifos	<chem>CC(=COP(=S)(OC)OC)C(=O)OC</chem>
Dichlorvos	<chem>COP(=O)(OC)OC=C(Cl)Cl</chem>
Folpet	<chem>C1=CC=C2C(=C1)C(=O)N(C2=O)SC(Cl)(Cl)Cl</chem>
Dichlobenil	<chem>C1=CC(=C(C(=C1)Cl)C#N)Cl</chem>
Biphenyl	<chem>C1=CC=C(C=C1)C2=CC=CC=C2</chem>
etobenzanid	<chem>CCOCOC1=CC=C(C=C1)C(=O)NC2=C(C(=CC=C2)Cl)Cl</chem>
Hymexazol	<chem>CC1=CC(=O)NO1</chem>
Trichlorfon	<chem>COP(=O)(C(C(Cl)(Cl)Cl)O)OC</chem>
Dibrom	<chem>COP(=O)(OC)OC(C(Cl)(Cl)Br)Br</chem>
Ditalimfos	<chem>CCOP(=S)(N1C(=O)C2=CC=CC=C2C1=O)OCC</chem>
spirodiclofen	<chem>CCC(C)(C)C(=O)OC1=C(C(=O)OC12CCCCC2)C3=C(C=C(C=C3)Cl)Cl</chem>
Crimidine	<chem>CC1=CC(=NC(=N1)Cl)N(C)C</chem>
Tecnazene	<chem>C1=C(C(C(=C(C(=C1Cl)Cl)[N+](=O)[O-])Cl)Cl</chem>
Formothion	<chem>CN(C=O)C(=O)CSP(=S)(OC)OC</chem>
Monocrotophos	<chem>CC(=CC(=O)NC)OP(=O)(OC)OC</chem>
pyrimidifen	<chem>CCC1=C(C(=NC=N1)NCCOC2=C(C(=C(C=C2)CCOCC)C)C)Cl</chem>
a-BHC	<chem>C1(C(C(C(C1Cl)Cl)Cl)Cl)Cl</chem>
Omethoate	<chem>CNC(=O)CSP(=O)(OC)OC</chem>
azamethiphos	<chem>COP(=O)(OC)SCN1C2=NC=C(C=C2OC1=O)Cl</chem>
Mevinphos	<chem>CC(=CC(=O)OC)OP(=O)(OC)OC</chem>
Diphenylamine	<chem>C1=CC=C(C=C1)NC2=CC=CC=C2</chem>
Quintozene	<chem>C1(=C(C(=C(C(=C1Cl)Cl)Cl)Cl)Cl)[N+](=O)[O-]</chem>
OPP	<chem>C1=CC=C(C=C1)C2=CC=CC=C2O</chem>
Oxabetrinil	<chem>C1COC(O1)CON=C(C#N)C2=CC=CC=C2</chem>
Parathion	<chem>CCOP(=S)(OCC)OC1=CC=C(C=C1)[N+](=O)[O-]</chem>
Fenitrothion	<chem>CC1=C(C=CC(=C1)OP(=S)(OC)OC)[N+](=O)[O-]</chem>
myclobutanil	<chem>CCCC(CN1C=NC=N1)(C#N)C2=CC=C(C=C2)Cl</chem>
Isofenphos	<chem>CCOP(=S)(NC(C)C)OC1=CC=CC=C1C(=O)OC(C)C</chem>
Trichlamide	<chem>CCCCOC(C(Cl)(Cl)Cl)NC(=O)C1=CC=CC=C1O</chem>
Etrimfos	<chem>CCC1=NC(=CC(=N1)OP(=S)(OC)OC)OCC</chem>
Iprobenfos	<chem>CC(C)OP(=O)(OC(C)C)SCC1=CC=CC=C1</chem>
Dimethametryn	<chem>CCNC1=NC(=NC(=N1)SC)NC(C)C(C)C</chem>
Flutolanil	<chem>CC(C)OC1=CC=CC(=C1)NC(=O)C2=CC=CC=C2C(F)(F)F</chem>

Continue to next page

*Continue from previous page*

Pesticide name	Canonical SMILES
Tefluthrin	<chem>CC1=C(C(=C(C(=C1F)F)COC(=O)C2C(C2(C)C)C=C(C(F)(F)F)Cl)F)F</chem>
Benfuresate	<chem>CCS(=O)(=O)OC1=CC2=C(C=C1)OCC2(C)C</chem>
uniconazole	<chem>CC(C)(C)C(C(=CC1=CC=C(C=C1)Cl)N2C=NC=N2)O</chem>
Prometryn	<chem>CC(C)NC1=NC(=NC(=N1)SC)NC(C)C</chem>
mepronil	<chem>CC1=CC=CC=C1C(=O)NC2=CC(=CC=C2)OC(C)C</chem>
Diethofencarb	<chem>CCOC1=C(C=C(C=C1)NC(=O)OC(C)C)OCC</chem>
Propanil	<chem>CCC(=O)NC1=CC(=C(C=C1)Cl)Cl</chem>
Fenamiphos	<chem>CCOP(=O)(NC(C)C)OC1=CC(=C(C=C1)SC)C</chem>
Profenofos	<chem>CCCSP(=O)(OCC)OC1=C(C=C(C=C1)Br)Cl</chem>
Propaphos	<chem>CCCOP(=O)(OCCC)OC1=CC=C(C=C1)SC</chem>
Dicloran	<chem>C1=C(C=C(C(=C1Cl)N)Cl)[N+](=O)[O-]</chem>
Benthiocarb	<chem>CCN(CC)C(=O)SCC1=CC=C(C=C1)Cl</chem>
Metalaxyl	<chem>CC1=C(C(=CC=C1)C)N(C(C)C(=O)OC)C(=O)COC</chem>
Pendimethalin	<chem>CCC(CC)NC1=C(C=C(C(=C1[N+](=O)[O-])C)C)[N+](=O)[O-]</chem>
Methyl parathion	<chem>COP(=S)(OC)OC1=CC=C(C=C1)[N+](=O)[O-]</chem>
Phorate	<chem>CCOP(=S)(OCC)SCSCC</chem>
Chlorpyrifos	<chem>CCOP(=S)(OCC)OC1=NC(=C(C=C1)Cl)Cl</chem>
Cadusafos	<chem>CCC(C)SP(=O)(OCC)SC(C)CC</chem>
pyridaphenthion	<chem>CCOP(=S)(OCC)OC1=NN(C(=O)C=C1)C2=CC=CC=C2</chem>
fluacrypyrim	<chem>CC(C)OC1=NC(=CC(=N1)OCC2=CC=CC=C2C(=COC)C(=O)OC)C(F)(F)F</chem>
lenacil	<chem>C1CCC(CC1)N2C(=O)C3=C(CCC3)NC2=O</chem>
nitralin	<chem>CCCN(CCC)C1=C(C=C(C=C1[N+](=O)[O-])S(=O)(=O)C)[N+](=O)[O-]</chem>
furametpyr	<chem>CC1C2=C(C=CC=C2NC(=O)C3=C(N(N=C3C)Cl)C(O1)(C)C</chem>
Terbacil	<chem>CC1=C(C(=O)N(C(=O)N1)C(C)(C)Cl</chem>
Fipronil	<chem>C1=C(C=C(C(=C1Cl)N2C(=C(C(=N2)C#N)S(=O)C(F)(F)F)N)Cl)C(F)(F)F</chem>
Propachlor	<chem>CC(C)N(C1=CC=CC=C1)C(=O)CCl</chem>
bifenox	<chem>COC(=O)C1=C(C=CC(=C1)OC2=C(C=C(C=C2)Cl)Cl)[N+](=O)[O-]</chem>
etoxazole	<chem>CCOC1=C(C=CC(=C1)C(C)(C)C2COC(=N2)C3=C(C=CC=C3F)F</chem>
Tolyfluamid	<chem>CC1=CC=C(C=C1)N(SC(F)(Cl)Cl)S(=O)(=O)N(C)C</chem>
ACN	<chem>C1=CC=C2C(=C1)C(=O)C(=C(C2=O)Cl)N</chem>
Dithiopyr	<chem>CC(C)CC1=C(C(=NC(=C1C(=O)SC)C(F)(F)F)C(F)F)C(=O)SC</chem>
Dimepiperate	<chem>CC(C)(C1=CC=CC=C1)SC(=O)N2CCCCC2</chem>
Dimethenamid	<chem>CC1=CSC(=C1N(C(C)COC)C(=O)CCl)C</chem>
Xyllycarb	<chem>CC1=C(C=C(C=C1)OC(=O)NC)C</chem>
Tolclofos-methyl	<chem>CC1=CC(=C(C(=C1)Cl)OP(=S)(OC)OC)Cl</chem>
Pyrimethanil	<chem>CC1=CC(=NC(=N1)NC2=CC=CC=C2)C</chem>
Isoprothiolane	<chem>CC(C)OC(=O)C(=C1SCCS1)C(=O)OC(C)C</chem>
2,6-Dichlorobenzamide	<chem>C1=CC(=C(C(=C1)Cl)C(=O)N)Cl</chem>
MCPB ethyl ester	<chem>CCOC(=O)CCOC1=C(C=C(C=C1)Cl)C</chem>
carbophenothion	<chem>CCOP(=S)(OCC)SCSC1=CC=C(C=C1)Cl</chem>
Sulfotep	<chem>CCOP(=S)(OCC)OP(=S)(OCC)OCC</chem>
Fenothiocarb	<chem>CN(C)C(=O)SCCCCOC1=CC=CC=C1</chem>
Cyprodinyl	<chem>CC1=NC(=NC(=C1)C2CC2)NC3=CC=CC=C3</chem>
triazophos	<chem>CCOP(=S)(OCC)OC1=NN(C(=N1)C2=CC=CC=C2</chem>
isoxathion	<chem>CCOP(=S)(OCC)OC1=NOCC(=C1)C2=CC=CC=C2</chem>
cyproconazole	<chem>CC(C1CC1)C(CN2C=NC=N2)(C3=CC=C(C=C3)Cl)O</chem>
Cyanophos	<chem>COP(=S)(OC)OC1=CC=C(C=C1)C#N</chem>
chlorthiophos	<chem>CCOP(=S)(OCC)OC1=CC(=C(C=C1)SC)Cl</chem>
Fenchlorphos	<chem>COP(=S)(OC)OC1=CC(=C(C=C1)Cl)Cl</chem>
Butamifos	<chem>CCC(C)NP(=S)(OCC)OC1=C(C=CC(=C1)C)[N+](=O)[O-]</chem>
Simeconazole	<chem>C[Si](C)(C)CC(CN1C=NC=N1)(C2=CC=C(C=C2)F)O</chem>
Isazophos	<chem>CCOP(=S)(OCC)OC1=NN(C(=N1)Cl)C(C)C</chem>
oxadiazon	<chem>CC(C)OC1=C(C=C(C(=C1)N2C(=O)OC(=N2)C(C)(C)C)Cl)Cl</chem>
Tetrachlorvinphos	<chem>COP(=O)(OC)OC(=CCl)C1=CC(=C(C=C1)Cl)Cl</chem>
pyriminobac-methyl(Z)	<chem>CC(=NOC)C1=C(C(=CC=C1)OC2=NC(=CC(=N2)OC)OC)C(=O)OC</chem>
Penconazole	<chem>CCCC(CN1C=NC=N1)C2=C(C=C(C=C2)Cl)Cl</chem>
Thiometon	<chem>CCSCCSP(=S)(OC)OC</chem>

*Continue to next page*

*Continue from previous page*

Pesticide name	Canonical SMILES
Acetochlor	<chem>CCC1=CC=CC(=C1N(COCC)C(=O)CCl)C</chem>
Desmedipham	<chem>CCOC(=O)NC1=CC(=CC=C1)OC(=O)NC2=CC=CC=C2</chem>
flusilazole	<chem>C[Si](CN1C=NC=N1)(C2=CC=C(C=C2)F)C3=CC=C(C=C3)F</chem>
leptophos	<chem>COP(=S)(C1=CC=CC=C1)OC2=CC(=C(C=C2Cl)Br)Cl</chem>
Triadimefon	<chem>CC(C)(C)C(=O)C(N1C=NC=N1)OC2=CC=C(C=C2)Cl</chem>
Terbucarb	<chem>CC1=CC(=C(C=C1)C(C)(C)OC(=O)NC)C(C)(C)C</chem>
Cyanazine	<chem>CCNC1=NC(=NC(=N1)Cl)NC(C)(C)C#N</chem>
Chlorothalonil	<chem>C(#N)C1=C(C(=C(C=C1Cl)Cl)Cl)C#N)Cl</chem>
Malathion	<chem>CCOC(=O)CC(C(=O)OCC)SP(=S)(OC)OC</chem>
Propyzamide	<chem>CC(C)(C#C)NC(=O)C1=CC(=CC(=C1)Cl)Cl</chem>
diniconazole	<chem>CC(C)(C)C(C(=CC1=C(C=C(C=C1)Cl)Cl)N2C=NC=N2)O</chem>
ethofenprox	<chem>CCOC1=CC=C(C=C1)C(C)(C)COCC2=CC(=CC=C2)OC3=CC=CC=C3</chem>
Phenothiol	<chem>CCSC(=O)COC1=C(C=C(C=C1)Cl)C</chem>
pyrazophos	<chem>CCOC(=O)C1=CN2C(=CC(=N2)OP(=S)(OCC)OCC)N=C1C</chem>
tebufenpyrad	<chem>CCC1=NN(C(=C1Cl)C(=O)NCC2=CC=C(C=C2)C(C)(C)C)C</chem>
edifenphos	<chem>CCOP(=O)(SC1=CC=CC=C1)SC2=CC=CC=C2</chem>
ethion	<chem>CCOP(=S)(OCC)SCSP(=S)(OCC)OCC</chem>
Atrazine	<chem>CCNC1=NC(=NC(=N1)Cl)NC(C)C</chem>
Fenpropimorph	<chem>CC1CN(CC(O1)C)CC(C)CC2=CC=C(C=C2)C(C)(C)C</chem>
Dimethipin	<chem>CC1=C(S(=O)(=O)CCS1(=O)=O)C</chem>
Thifluzamide	<chem>CC1=NC(=C(S1)C(=O)NC2=C(C=C(C=C2Br)OC(F)(F)F)Br)C(F)(F)F</chem>
anilofos	<chem>CC(C)N(C1=CC=C(C=C1)Cl)C(=O)CSP(=S)(OC)OC</chem>
azaconazole	<chem>C1COC(O1)(CN2C=NC=N2)C3=C(C=C(C=C3)Cl)Cl</chem>
amitraz	<chem>CC1=CC(=C(C=C1)N=CN(C)C=NC2=C(C=C(C=C2)C)C)C</chem>
Metominostrobin E	<chem>CNC(=O)C(=NOC)C1=CC=CC=C1OC2=CC=CC=C2</chem>
Demeton-S-methyl	<chem>CCSCCSP(=O)(OC)OC</chem>
Bromobutide	<chem>CC(C)(C)C(C(=O)NC(C)(C)C1=CC=CC=C1)Br</chem>
Prothiofos	<chem>CCCSP(=S)(OCC)OC1=C(C=C(C=C1)Cl)Cl</chem>
Espirocarb	<chem>CCN(C(C)C(C)C)C(=O)SCC1=CC=CC=C1</chem>
Pretilachlor	<chem>CCCOCCN(C1=C(C=CC=C1CC)CC)C(=O)CCl</chem>
phosmet	<chem>COP(=S)(OC)SCN1C(=O)C2=CC=CC=C2C1=O</chem>
dioxathion	<chem>CCOP(=S)(OCC)SC1C(OCCO1)SP(=S)(OCC)OCC</chem>
tebuconazole	<chem>CC(C)(C)C(CCC1=CC=C(C=C1)Cl)(CN2C=NC=N2)O</chem>
trifloxystrobin	<chem>CC(=NOC)C1=CC=CC=C1C(=NOC)C(=O)OC)C2=CC(=CC=C2)C(F)(F)F</chem>
Disulfoton	<chem>CCOP(=S)(OCC)SCCSCC</chem>
bifenthrin	<chem>CC1=C(C=CC=C1COC(=O)C2C(C2(C)C)C=C(C(F)(F)F)Cl)C3=CC=CC=C3</chem>
Dicrotophos	<chem>CC(=CC(=O)N(C)C)OP(=O)(OC)OC</chem>
pyriminobac-methyl(E)	<chem>CC(=NOC)C1=C(C(=CC=C1)OC2=NC(=CC(=N2)OC)OC)C(=O)OC</chem>
Diphenamid	<chem>CN(C)C(=O)C(C1=CC=CC=C1)C2=CC=CC=C2</chem>
benalaxyl	<chem>CC1=C(C(=CC=C1)C)N(C(C)C(=O)OC)C(=O)CC2=CC=CC=C2</chem>
Chlorpyrifos-methyl	<chem>COP(=S)(OC)OC1=NC(=C(C=C1Cl)Cl)Cl</chem>
chlorfenapyr	<chem>CCOCCN1C(=C(C(=C1C(F)(F)F)Br)C#N)C2=CC=C(C=C2)Cl</chem>
diclobutrazol	<chem>CC(C)(C)C(C(CCC1=C(C=C(C=C1)Cl)Cl)N2C=NC=N2)O</chem>
Dichlofuanid	<chem>CN(C)S(=O)(=O)N(C1=CC=CC=C1)SC(F)(Cl)Cl</chem>
Lindane	<chem>C1(C(C(C(C1Cl)Cl)Cl)Cl)Cl)Cl</chem>
Sweep	<chem>COC(=O)NC1=CC(=C(C=C1)Cl)Cl</chem>
azinphos-ethyl	<chem>CCOP(=S)(OCC)SCN1C(=O)C2=CC=CC=C2N=N1</chem>
Paclobutrazol	<chem>CC(C)(C)C(C(CCC1=CC=C(C=C1)Cl)N2C=NC=N2)O</chem>
Methidathion	<chem>COC1=NN(C(=O)S1)CSP(=S)(OC)OC</chem>
Benfluralin	<chem>CCCCN(CC)C1=C(C=C(C=C1[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]</chem>
Pyroquilon	<chem>C1CC(=O)N2CCC3=CC=CC1=C32</chem>
Prohydrojasmon	<chem>CCCCCC1C(CCC1=O)CC(=O)OCCC</chem>
cyhalofop-butyl	<chem>CCCCOC(=O)C(C)OC1=CC=C(C=C1)OC2=C(C=C(C=C2)C#N)F</chem>
phosalone	<chem>CCOP(=S)(OCC)SCN1C2=C(C=C(C=C2)Cl)OC1=O</chem>
chlomethoxynil	<chem>COC1=C(C=C(C=C1)OC2=C(C=C(C=C2)Cl)Cl)[N+](=O)[O-]</chem>
a-Endosulfan	<chem>C1C2C(COS(=O)O1)C3(C(=C(C2(C3(Cl)Cl)Cl)Cl)Cl)Cl</chem>
Simazine	<chem>CCNC1=NC(=NC(=N1)Cl)NCC</chem>

*Continue to next page*

*Continue from previous page*

Pesticide name	Canonical SMILES
clomeprop	<chem>CC1=C(C=CC(=C1Cl)OC(C)C(=O)NC2=CC=CC=C2)Cl</chem>
iprodione	<chem>CC(C)NC(=O)N1CC(=O)N(C1=O)C2=CC(=CC(=C2)Cl)Cl</chem>
Metolachlor	<chem>CCC1=CC=CC(=C1N(C(C)COC)C(=O)CCl)C</chem>
fensulfthion	<chem>CCOP(=S)(OCC)OC1=CC=C(C=C1)S(=O)C</chem>
diflufenican	<chem>C1=CC(=CC(=C1)OC2=C(C=CC=N2)C(=O)NC3=C(C=C(C=C3)F)F)C(F)(F)F</chem>
sulprofos	<chem>CCCSP(=S)(OCC)OC1=CC=C(C=C1)SC</chem>
kresoxim-methyl	<chem>CC1=CC=CC=C1OCC2=CC=CC=C2C(=NOC)C(=O)OC</chem>
b-endosulfan	<chem>C1C2C(COS(=O)O1)C3(C(=C(C2(C3(Cl)Cl)Cl)Cl)Cl)Cl</chem>
Phthalide	<chem>C1C2=CC=CC=C2C(=O)O1</chem>
piperophos	<chem>CCCOP(=S)(OCCC)SCC(=O)N1CCCCC1C</chem>
Carbetamide	<chem>CCNC(=O)C(C)OC(=O)NC1=CC=CC=C1</chem>
azinphos-methyl	<chem>COP(=S)(OC)SCN1C(=O)C2=CC=CC=C2N=N1</chem>
Isocarbophos	<chem>CC(C)OC(=O)C1=CC=CC=C1OP(=S)(N)OC</chem>
Cinmethylin	<chem>CC1=CC=CC=C1COC2CC3(CCC2(O3)C)C(C)C</chem>
Chlorfenson	<chem>C1=CC(=CC=C1OS(=O)(=O)C2=CC=C(C=C2)Cl)Cl</chem>
deltamethrin	<chem>CC1(C(C1C(=O)OC(C#N)C2=CC(=CC=C2)OC3=CC=CC=C3)C=C(Br)Br)C</chem>
piperonyl butoxide	<chem>CCCCOCCOCCOCC1=CC2=C(C=C1CCC)OCO2</chem>
pyributicarb	<chem>CC(C)(C)C1=CC(=CC=C1)OC(=S)N(C)C2=NC(=CC=C2)OC</chem>
Fenthion	<chem>CC1=C(C=CC(=C1)OP(=S)(OC)OC)SC</chem>
cafenstrole	<chem>CCN(CC)C(=O)N1C=NC(=N1)S(=O)(=O)C2=C(C=C(C=C2C)C)C</chem>
Bromophos	<chem>COP(=S)(OC)OC1=CC(=C(C=C1Cl)Br)Cl</chem>
Napropamide	<chem>CCN(CC)C(=O)C(C)OC1=CC=CC2=CC=CC=C21</chem>
Chlorpropham	<chem>CC(C)OC(=O)NC1=CC(=CC=C1)Cl</chem>
indanofan	<chem>CCC1(C(=O)C2=CC=CC=C2C1=O)CC3(CO3)C4=CC(=CC=C4)Cl</chem>
silafiuofen	<chem>CCOC1=CC=C(C=C1)[Si](C)(C)CCCC2=CC(=C(C=C2)F)OC3=CC=CC=C3</chem>
Dioxabenzofos	<chem>COP1(=S)OCC2=CC=CC=C2O1</chem>
Mecarbam	<chem>CCOC(=O)N(C)C(=O)CSP(=S)(OCC)OCC</chem>
pyridaben	<chem>CC(C)(C)C1=CC=C(C=C1)CSC2=C(C(=O)N(N=C2)C(C)(C)C)Cl</chem>
4,4-dichlorobenzophenone	<chem>C1=CC(=CC=C1C(=O)C2=CC=C(C=C2)Cl)Cl</chem>
Procymidone	<chem>CC12CC1(C(=O)N(C2=O)C3=CC(=CC(=C3)Cl)Cl)C</chem>
Alachlor	<chem>CCC1=C(C(=CC=C1)CC)N(COC)C(=O)CCl</chem>
Diazinon	<chem>CCOP(=S)(OCC)OC1=NC(=NC(=C1)C)C(C)C</chem>
pyraflufen-ethyl	<chem>CCOC(=O)COC1=C(C=C(C=C1)C2=NN(C(=C2Cl)OC(F)F)C)F)Cl</chem>
fluquinconazole	<chem>C1=CC2=C(C(C=C1F)C(=O)N(C(=N2)N3C=NC=N3)C4=C(C=C(C=C4)Cl)Cl</chem>
Pirimiphos-methyl	<chem>CCN(CC)C1=NC(=CC(=N1)OP(=S)(OC)OC)C</chem>
Terbufos	<chem>CCOP(=S)(OCC)SCSC(C)(C)C</chem>
EPN	<chem>CCOP(=S)(C1=CC=CC=C1)OC2=CC=C(C=C2)[N+](=O)[O-]</chem>
Ferimzone	<chem>CC1=CC=CC=C1C(=NNC2=NC(=CC(=N2)C)C)C</chem>
mefenacet	<chem>CN(C1=CC=CC=C1)C(=O)COC2=NC3=CC=CC=C3S2</chem>
Butachlor	<chem>CCCCOCCN(C1=C(C=CC=C1CC)CC)C(=O)CCl</chem>
Phenthoate	<chem>CCOC(=O)C(C1=CC=CC=C1)SP(=S)(OC)OC</chem>
fenoxaprop-ethyl	<chem>CCOC(=O)C(C)OC1=CC=C(C=C1)OC2=NC3=C(O2)C=C(C=C3)Cl</chem>
Simetryn	<chem>CCNC1=NC(=NC(=N1)SC)NCC</chem>
cyflufenamide	<chem>C1CC1CONC(=NC(=O)CC2=CC=CC=C2)C3=C(C=CC(=C3F)F)C(F)(F)F</chem>
carfentrazone-ethyl	<chem>CCOC(=O)C(CC1=CC(=C(C=C1Cl)F)N2C(=O)N(C(=N2)C)C(F)F)Cl</chem>
chlornitrofen	<chem>C1=CC(=CC=C1[N+](=O)[O-])OC2=C(C=C(C=C2Cl)Cl)Cl</chem>
fenarimol	<chem>C1=CC=C(C(=C1)C(C2=CC=C(C=C2)Cl)(C3=CN=CN=C3)O)Cl</chem>
chlorobenzilate	<chem>CCOC(=O)C(C1=CC=C(C=C1)Cl)(C2=CC=C(C=C2)Cl)O</chem>
nitrofen	<chem>C1=CC(=CC=C1[N+](=O)[O-])OC2=C(C=C(C=C2)Cl)Cl</chem>
DEF	<chem>CCCCSP(=O)(SCCCC)SCCCC</chem>
flumioxazin	<chem>C#CCN1C(=O)COC2=CC(=C(C=C21)N3C(=O)C4=C(C3=O)CCCC4)F</chem>
indoxacarb	<chem>COC(=O)C12CC3=C(C1=NN(CO2)C(=O)N(C4=CC=C(C=C4)OC(F)(F)F)C(=O)OC)C=CC(=C3)Cl</chem>
chloropropylate	<chem>CC(C)OC(=O)C(C1=CC=C(C=C1)Cl)(C2=CC=C(C=C2)Cl)O</chem>
fluthiacet-methyl	<chem>COC(=O)CSC1=C(C=C(C=C1)N=C2N3CCCCN3C(=O)S2)F)Cl</chem>
acrinathrin	<chem>CC1(C(C1C(=O)OC(C#N)C2=CC(=CC=C2)OC3=CC=CC=C3)C=CC(=O)OC(C(F)(F)F)C(F)(F)F)C</chem>
halfenprox	<chem>CC(C)(COCC1=CC(=CC=C1)OC2=CC=CC=C2)C3=CC=C(C=C3)OC(F)(F)Br</chem>
Hexaconazole	<chem>CCCC(CN1C=NC=N1)(C2=C(C=C(C=C2)Cl)Cl)O</chem>

*Continue to next page*

*Continue from previous page*

Pesticide name	Canonical SMILES
Ethoprophos	<chem>CCCSP(=O)(OCC)SCCC</chem>
dialifos	<chem>CCOP(=S)(OCC)SC(CCl)N1C(=O)C2=CC=CC=C2C1=O</chem>
butafenacil	<chem>CC(C)(C(=O)OCC=C)OC(=O)C1=C(C=CC(=C1)N2C(=O)C=C(N(C2=O)C)C(F)(F)F)Cl</chem>
Trifluralin	<chem>CCCN(CCC)C1=C(C=C(C=C1[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]</chem>
Dichlofenthion	<chem>CCOP(=S)(OCC)OC1=C(C=C(C=C1)Cl)Cl</chem>
Ametryn	<chem>CCNC1=NC(=NC(=N1)SC)NC(C)C</chem>
quizalofop-ethyl	<chem>CCOC(=O)C(C)OC1=CC=C(C(=C1)OC2=CN=C3C=C(C=CC3=N2)Cl</chem>
fenoxycarb	<chem>CCOC(=O)NCCOC1=CC=C(C(=C1)OC2=CC=CC=C2</chem>
pyraclofos	<chem>CCCSP(=O)(OCC)OC1=CN(N=C1)C2=CC=C(C=C2)Cl</chem>
fenoxanil	<chem>CC(C)C(C)(C#N)NC(=O)C(C)OC1=C(C=C(C=C1)Cl)Cl</chem>
fenbuconazole	<chem>C1=CC=C(C=C1)C(CCC2=CC=C(C=C2)Cl)(CN3C=NC=N3)C#N</chem>
oxadixyl	<chem>CC1=C(C(=CC=C1)C)N(C(=O)COC)N2CCOC2=O</chem>
Quinalphos	<chem>CCOP(=S)(OCC)OC1=NC2=CC=CC=C2N=C1</chem>
pyriproxyfen	<chem>CC(COC1=CC=C(C=C1)OC2=CC=CC=C2)OC3=CC=CC=N3</chem>
pyrazoxyfen	<chem>CC1=NN(C(=C1C(=O)C2=C(C=C(C=C2)Cl)Cl)OCC(=O)C3=CC=CC=C3)C</chem>
Metribuzin	<chem>CC(C)(C)C1=NN=C(N(C1=O)N)SC</chem>
azoxystrobin	<chem>COC=C(C1=CC=CC=C1OC2=NC=NC(=C2)OC3=CC=CC=C3C#N)C(=O)OC</chem>
Fonofos	<chem>CCOP(=S)(CC)SC1=CC=CC=C1</chem>
Vinclozolin	<chem>CC1(C(=O)N(C(=O)O1)C2=CC(=CC(=C2)Cl)Cl)C=C</chem>
tetradifon	<chem>C1=CC(=CC=C1S(=O)(=O)C2=CC(=C(C=C2Cl)Cl)Cl)Cl</chem>
fenpropathrin	<chem>CC1(C(C1(C)C)C(=O)OC(C#N)C2=CC(=CC=C2)OC3=CC=CC=C3)C</chem>
bromopropylate	<chem>CC(C)OC(=O)C(C1=CC=C(C=C1)Br)(C2=CC=C(C=C2)Br)O</chem>
Ethychlozate	<chem>CCOC(=O)CC1=C2C=C(C(=CC2=NN1)Cl</chem>
tolfenpyrad	<chem>CCC1=NN(C(=C1Cl)C(=O)NCC2=CC=C(C=C2)OC3=CC=C(C=C3)C)C</chem>
Bromacil	<chem>CCC(C)N1C(=O)C(=C(NC1=O)C)Br</chem>
Dimethoate	<chem>CNC(=O)CSP(=S)(OC)OC</chem>
buprofezin	<chem>CC(C)N1C(=NC(C)(C)C)SCN(C1=O)C2=CC=CC=C2</chem>
thenylchlor	<chem>CC1=C(C(=CC=C1)C)N(CC2=C(C=CS2)OC)C(=O)CCl</chem>
pentoxazone	<chem>CC(=C1C(=O)N(C(=O)O1)C2=CC(=C(C=C2F)Cl)OC3CCCC3)C</chem>
chloridazon	<chem>C1=CC=C(C=C1)N2C(=O)C(=C(C=N2)N)Cl</chem>
Fludioxonil	<chem>C1=CC(=C2C(=C1)OC(O2)(F)F)C3=CNC=C3C#N</chem>
bioresmethrin	<chem>CC(=CC1C(C1(C)C)C(=O)OCC2=COC(=C2)CC3=CC=CC=C3)C</chem>
captafol	<chem>C1C=CCC2C1C(=O)N(C2=O)SC(C(Cl)Cl)(Cl)Cl</chem>

*End of table*



## A.2 Average of log PE(Av log PE) and Average of log Execution Time(Av log ET). Regression methods(Method) are ordered according to Av log PE.

Table A.2. Average of log PE(Av log PE) and Average of log Execution Time(Av log ET). Regression methods(Method) are ordered according to Av log PE.

Method	Av logPE	Av log ET	Method	Av logPE	Av log ET	Method	Av logPE	Av log ET
xgbLinear	-3.221	1.233	M5	-0.459	1.476	earth	-0.199	0.782
SBC	-3.124	1.416	kknn	-0.457	0.473	BstLm	-0.195	0.560
monmlp	-2.002	1.743	svmLinear3	-0.434	1.040	knn	-0.192	-0.034
ppr	-1.817	0.780	M5Rules	-0.430	1.484	rpart2	-0.181	0.229
extraTrees	-1.260	1.889	ridge	-0.419	1.208	ctree2	-0.144	0.913
xgbDART	-1.103	1.995	gaussprRadial	-0.417	0.232	evtree	-0.141	1.533
parRF	-1.064	1.781	treebag	-0.417	0.892	kernelpls	-0.140	0.038
glm	-0.963	0.015	svmLinear2	-0.403	0.813	plsRglm	-0.133	1.702
lm	-0.937	-0.050	penalized	-0.386	1.405	widekernelpls	-0.132	0.401
glmStepAIC	-0.926	2.537	gcvEarth	-0.371	0.549	lars2	-0.130	0.819
lmStepAIC	-0.918	1.730	rpart1SE	-0.347	0.204	ctree	-0.127	0.573
qrf	-0.894	1.745	neuralnet	-0.343	1.784	simpls	-0.124	0.015
bayesglm	-0.809	0.887	svmRadialSigma	-0.333	0.616	pls	-0.122	0.024
brnn	-0.783	2.272	cforest	-0.328	1.528	blasso	-0.119	3.208
ranger	-0.757	1.552	enet	-0.326	1.171	xyf	-0.119	1.303
xgbTree	-0.728	1.576	msaenet	-0.315	1.777	superpc	-0.078	0.447
RRFglobal	-0.718	2.136	svmRadial	-0.314	0.457	pcr	-0.077	0.182
rf	-0.717	1.778	svmRadialCost	-0.314	0.312	leapForward	-0.074	0.037
RRF	-0.706	2.856	Rborist	-0.310	1.691	leapBackward	-0.072	0.155
krlsRadial	-0.672	2.041	relaxo	-0.304	1.009	leapSeq	-0.070	0.121
rvmRadial	-0.658	0.494	mlpML	-0.281	1.147	icr	-0.053	0.591
WM	-0.649	2.877	glmnet	-0.255	1.134	bridge	-0.025	3.166
gaussprLinear	-0.611	0.092	rvmLinear	-0.253	0.443	dnn	-0.011	1.822
svmPoly	-0.576	0.971	spikeslab	-0.252	1.225	blassoAveraged	-0.007	3.183
bstTree	-0.567	1.982	nodeHarvest	-0.249	2.326	krlsPoly	-0.005	3.312
rvmPoly	-0.558	1.304	nnls	-0.223	-0.098	rbfDDA	0.065	1.210
gbm	-0.546	0.828	foba	-0.215	1.129	bagEarth	0.542	1.785
gaussprPoly	-0.534	0.698	mlpWeightDecay	-0.204	1.562	lars	1.572	0.712
bagEarthGCV	-0.475	1.397	glmboost	-0.203	0.357	lasso	8.646	1.021
svmLinear	-0.461	0.607	mlp	-0.201	1.148			

## B. Additional information for classification analysis

### B.1 Pesticide and Canonical SMILES for classification analysis

Table B.1: Summary of molecular descriptors obtained by SMILES for classification. Technology L is analyzed by LC-MS, G is GC-MS. List of E is EURL list, F is FDA list and Both is both EURL and FDA list.

Pesticides	Technology	List	Canonical SMILES
$\alpha$ -BHC	G	Both	<chem>C1(C(C(C(C1Cl)Cl)Cl)Cl)Cl</chem>
$\alpha$ -endosulfan	G	Both	<chem>C1C2C(COS(=O)O1)C3(C(=C(C2(C3(Cl)Cl)Cl)Cl)Cl)Cl</chem>
acetamiprid	L	Both	<chem>CC(=NC#N)N(C)CC1=CN=C(C=C1)Cl</chem>
Aldicarb	L	E	<chem>CC(C)(C=NOC(=O)NC)SC</chem>
Aldicarb Sulfone	L	E	<chem>CC(C)(C=NOC(=O)NC)S(=O)(=O)C</chem>
Aldicarb Sulfoxide	L	E	<chem>CC(C)(C=NOC(=O)NC)S(=O)C</chem>
ametryn	L	F	<chem>CCNC1=NC(=NC(=N1)SC)NC(C)C</chem>
aminocarb	L	F	<chem>CC1=C(C=CC(=C1)OC(=O)NC)N(C)C</chem>
amitraz	G	F	<chem>CC1=CC(=C(C=C1)N=CN(C)C=NC2=C(C=C(C=C2)C)C)C</chem>
azinphos-methyl	L	F	<chem>COP(=S)(OC)SCN1C(=O)C2=CC=CC=C2N=N1</chem>
Azoxystrobin	L	E	<chem>COC=C(C1=CC=CC=C1OC2=NC=NC(=C2)OC3=CC=CC=C3C#N)C(=O)OC</chem>
$\beta$ -endosulfan	G	Both	<chem>C1C2C(COS(=O)O1)C3(C(=C(C2(C3(Cl)Cl)Cl)Cl)Cl)Cl</chem>
Benalaxyl	G	E	<chem>CC1=C(C(=CC=C1)C)N(C(C)C(=O)OC)C(=O)CC2=CC=CC=C2</chem>
bendiocarb	L	F	<chem>CC1(OC2=C(O1)C(=CC=C2)OC(=O)NC)C</chem>
Benfluralin	G	Both	<chem>CCCCN(CC)C1=C(C=C(C=C1[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]</chem>
Bifenox	G	E	<chem>COC(=O)C1=C(C=CC(=C1)OC2=C(C=C(C=C2)Cl)Cl)[N+](=O)[O-]</chem>
bifenthrin	L	F	<chem>CC1=C(C(=CC=C1COC(=O)C2C(C2(C)C)C=C(C(F)(F)F)Cl)C3=CC=CC=C3</chem>
boscalid	L	F	<chem>C1=CC=C(C(=C1)C2=CC=C(C=C2)Cl)NC(=O)C3=C(N=CC=C3)Cl</chem>
bromopropylate	G	Both	<chem>CC(C)OC(=O)C(C1=CC=C(C=C1)Br)(C2=CC=C(C=C2)Br)O</chem>
Bupirimate	G	E	<chem>CCCCC1=C(N=C(N=C1OS(=O)(=O)N(C)C)NCC)C</chem>
cadusafos	G	F	<chem>CCC(C)SP(=O)(OCC)SC(C)CC</chem>
Carbendazim	L	E	<chem>COC(=O)NC1=NC2=CC=CC=C2N1</chem>
Carbofuran	L	E	<chem>CC1(CC2=C(O1)C(=CC=C2)OC(=O)NC)C</chem>
Carbofuran 3-Oh	L	E	<chem>CC1(C(C2=C(O1)C(=CC=C2)OC(=O)NC)O)C</chem>
Carfentrazone Ethyl	L	E	<chem>CCOC(=O)C(CC1=CC(=C(C=C1Cl)F)N2C(=O)N(C(=N2)C)C(F)F)Cl</chem>
Chlofenvinphos	G	E	<chem>CCOP(=O)(OCC)OC(=CCl)C1=C(C=C(C=C1)Cl)Cl</chem>
chlordimeform	L	F	<chem>CC1=C(C=CC(=C1)Cl)N=CN(C)C</chem>
Chlorfenapyr	G	E	<chem>CCOCN1C(=C(C(=C1C(F)(F)F)Br)C#N)C2=CC=C(C=C2)Cl</chem>
Chloridazon	L	E	<chem>C1=CC=C(C=C1)N2C(=O)C(=C(C=N2)N)Cl</chem>
chlorothalonil	G	F	<chem>C(#N)C1=C(C(=C(C(=C1Cl)Cl)Cl)C#N)Cl</chem>
Chloroxuron	L	E	<chem>CN(C)C(=O)NC1=CC=C(C=C1)OC2=CC=C(C=C2)Cl</chem>
chlorpyrifos-methyl	G	Both	<chem>COP(=S)(OC)OC1=NC(=C(C=C1Cl)Cl)Cl</chem>
Chlorthiophos	G	E	<chem>CCOP(=S)(OCC)OC1=CC(=C(C=C1Cl)SC)Cl</chem>
Chlozolinate	G	E	<chem>CCOC(=O)C1(C(=O)N(C(=O)O1)C2=CC(=CC(=C2)Cl)Cl)C</chem>
Clortoluron	L	E	<chem>CC1=C(C=C(C=C1)NC(=O)N(C)C)Cl</chem>
Clothianidin	L	E	<chem>CNC(=N[N+](=O)[O-])NCC1=CN=C(S1)Cl</chem>
coumaphos	L	F	<chem>CCOP(=S)(OCC)OC1=CC2=C(C=C1)C(=C(C(=O)O2)Cl)C</chem>
cyanazine	L	F	<chem>CCNC1=NC(=NC(=N1)Cl)NC(C)(C)C#N</chem>
cycluron	L	F	<chem>CN(C)C(=O)NC1CCCCCCC1</chem>
Cyfluthrin	G	E	<chem>CC1(C(C1C(=O)OC(C#N)C2=CC(=C(C=C2)F)OC3=CC=CC=C3)C=C(Cl)Cl)C</chem>
cyhalothrin	G	Both	<chem>CC1(C(C1C(=O)OC(C#N)C2=CC(=CC=C2)OC3=CC=CC=C3)C=C(C(F)(F)F)Cl)C</chem>
Cymoxanil	L	E	<chem>CCNC(=O)NC(=O)C(=NOC)C#N</chem>
cypermethrin	G	Both	<chem>CC1(C(C1C(=O)OC(C#N)C2=CC(=CC=C2)OC3=CC=CC=C3)C=C(Cl)Cl)C</chem>
cyproconazole	L	F	<chem>CC(C1CC1)C(CN2C=NC=N2)(C3=CC=C(C=C3)Cl)O</chem>
dacthal	G	Both	<chem>COC(=O)C1=C(C(=C(C(=C1Cl)Cl)C(=O)OC)Cl)Cl</chem>
DDE(4,4')	G	F	<chem>C1=CC(=CC=C1C(=C(Cl)Cl)C2=CC=C(C=C2)Cl)Cl</chem>
DDT(2,4')	G	F	<chem>C1=CC=C(C(=C1)C(C2=CC=C(C=C2)Cl)C(Cl)(Cl)Cl)Cl</chem>

*Continue to next page*

*Continue from previous page*

Pesticides	Technology	List	Canonical SMILES
DDT(4,4')	G	F	<chem>C1=CC(=CC=C1C(C2=CC=C(C=C2)Cl)C(Cl)(Cl)Cl)Cl</chem>
DEF	G	F	<chem>CCCCSP(=O)(SCCCC)SCCCC</chem>
Deltamethrin	G	E	<chem>CC1(C(C1C(=O)OC(C#N)C2=CC(=CC=C2)OC3=CC=CC=C3)C=C(Br)Br)C</chem>
desmedipham	L	F	<chem>CCOC(=O)NC1=CC(=CC=C1)OC(=O)NC2=CC=CC=C2</chem>
Desmethyl Pirimicarb	L	E	<chem>CC1=C(N=C(N=C1OC(=O)N(C)C)NC)C</chem>
dichlorfluamid	L	F	<chem>CN(C)S(=O)(=O)N(C1=CC=CC=C1)SC(F)(Cl)Cl</chem>
dichlorvos	L	F	<chem>COP(=O)(OC)OC=C(Cl)Cl</chem>
Dicloran	G	E	<chem>C1=C(C=C(C(=C1Cl)N)Cl)[N+](=O)[O-]</chem>
dicrotophos	L	Both	<chem>CC(=CC(=O)N(C)C)OP(=O)(OC)OC</chem>
dieldrin	G	F	<chem>C1C2C3C(C1C4C2O4)C5(C(=C(C3(C5(Cl)Cl)Cl)Cl)Cl)Cl</chem>
difenoconazole	L	Both	<chem>CC1COC(O1)(CN2C=NC=N2)C3=C(C=C(C=C3)OC4=CC=C(C=C4)Cl)Cl</chem>
Diflufenican	G	E	<chem>C1=CC(=CC(=C1)OC2=C(C=CC=N2)C(=O)NC3=C(C=C(C=C3)F)F)C(F)(F)F</chem>
Dimefuron	L	E	<chem>CC(C)(C)C1=NN(C(=O)O1)C2=C(C=C(C=C2)NC(=O)N(C)C)Cl</chem>
Dimethachlor	L	E	<chem>CC1=C(C(=CC=C1)C)N(CCOC)C(=O)CC1</chem>
Dimethenamid	L	E	<chem>CC1=CSC(=C1N(C(C)COC)C(=O)CC1)C</chem>
dimethoate	L	Both	<chem>CNC(=O)CSP(=S)(OC)OC</chem>
dimethomorph	L	Both	<chem>COC1=C(C=C(C=C1)C(=CC(=O)N2CCOCC2)C3=CC=C(C=C3)Cl)OC</chem>
Dimoxystrobin	L	E	<chem>CC1=CC(=C(C=C1)C)OCC2=CC=CC=C2C(=NOC)C(=O)NC</chem>
Diniconazole	L	E	<chem>CC(C)(C)C(C(=CC1=C(C=C(C=C1)Cl)Cl)N2C=NC=N2)O</chem>
dinitramine	G	F	<chem>CCN(CC)C1=C(C=C(C(=C1[N+](=O)[O-])N)C(F)(F)F)[N+](=O)[O-]</chem>
dioxacarb	L	F	<chem>CNC(=O)OC1=CC=CC=C1C2OCCO2</chem>
Dmst	L	E	<chem>CC1=CC=C(C=C1)NS(=O)(=O)N(C)C</chem>
endosulfan sulphate	G	Both	<chem>C1C2C(COS(=O)(=O)O1)C3(C(=C(C2(C3(Cl)Cl)Cl)Cl)Cl)Cl</chem>
endrin	G	F	<chem>C1C2C3C(C1C4C2O4)C5(C(=C(C3(C5(Cl)Cl)Cl)Cl)Cl)Cl</chem>
EPN	G	F	<chem>CCOP(=S)(C1=CC=CC=C1)OC2=CC=C(C=C2)[N+](=O)[O-]</chem>
epoxiconazole	L	Both	<chem>C1=CC=C(C(=C1)C2C(O2)(CN3C=NC=N3)C4=CC=C(C=C4)F)Cl</chem>
Etaconazol	L	E	<chem>CCC1COC(O1)(CN2C=NC=N2)C3=C(C=C(C=C3)Cl)Cl</chem>
ethiolate	L	F	<chem>CCN(CC)C(=O)SCC</chem>
ethofumesate	L	F	<chem>CCOC1C(C2=C(O1)C=CC(=C2)OS(=O)(=O)C)(C)C</chem>
Ethoprophos	G	E	<chem>CCCCSP(=O)(OCC)SCCC</chem>
Etridiazole	G	F	<chem>CCOC1=NC(=NS1)C(Cl)(Cl)Cl</chem>
Fenamiphos	L	E	<chem>CCOP(=O)(NC(C)C)OC1=CC(=C(C=C1)SC)C</chem>
Fenamiphos Sulfone	L	E	<chem>CCOP(=O)(NC(C)C)OC1=CC(=C(C=C1)S(=O)(=O)C)C</chem>
Fenamiphos Sulfoxide	L	E	<chem>CCOP(=O)(NC(C)C)OC1=CC(=C(C=C1)S(=O)C)C</chem>
fenarimol	G	Both	<chem>C1=CC=C(C(=C1)C(C2=CC=C(C=C2)Cl)(C3=CN=CN=C3)O)Cl</chem>
fenbuconazole	L	F	<chem>C1=CC=C(C(=C1)C(CCC2=CC=C(C=C2)Cl)(CN3C=NC=N3)C#N</chem>
Penitrothion	G	E	<chem>CC1=C(C=CC(=C1)OP(=S)(OC)OC)[N+](=O)[O-]</chem>
Fenobucarb	L	E	<chem>CCC(C)C1=CC=CC=C1OC(=O)NC</chem>
fenoxycarb	L	F	<chem>CCOC(=O)NCCOC1=CC=C(C=C1)OC2=CC=CC=C2</chem>
Fenpropathrin	G	E	<chem>CC1(C(C1(C)C)C(=O)OC(C#N)C2=CC(=CC=C2)OC3=CC=CC=C3)C</chem>
fenpropimorph	L	F	<chem>CC1CN(CC(O1)C)CC(C)CC2=CC=C(C=C2)C(C)(C)C</chem>
Fenpyroximate	L	E	<chem>CC1=NN(C(=C1C=NOCC2=CC=C(C=C2)C(=O)OC(C)(C)C)OC3=CC=CC=C3)C</chem>
Fenuron	L	E	<chem>CN(C)C(=O)NC1=CC=CC=C1</chem>
fenvalerate	G	Both	<chem>CC(C)C(C1=CC=C(C=C1)Cl)C(=O)OC(C#N)C2=CC(=CC=C2)OC3=CC=CC=C3</chem>
Flazasulfuron	L	E	<chem>COC1=CC(=NC(=N1)NC(=O)NS(=O)(=O)C2=C(C=CC=N2)C(F)(F)F)OC</chem>
fludioxinil	L	F	<chem>C1=CC(=C2C(=C1)OC(O2)(F)F)C3=CC=CC=C3C#N</chem>
Flufenacet	L	E	<chem>CC(C)N(C1=CC=C(C=C1)F)C(=O)COC2=NN=C(S2)C(F)(F)F</chem>
Fluopicolide	L	E	<chem>C1=CC(=C(C(=C1)Cl)C(=O)NCC2=C(C=C(C=C2)C(F)(F)F)Cl)Cl</chem>
Fluoxastrobin	L	E	<chem>CON=C(C1=CC=CC=C1OC2=C(C(=NC=N2)OC3=CC=CC=C3Cl)F)C4=NOCCO4</chem>
fluquinconazole	L	Both	<chem>C1=CC2=C(C=C1F)C(=O)N(C(=N2)N3C=NC=N3)C4=C(C=C(C=C4)Cl)Cl</chem>
Flurtamone	L	E	<chem>CNC1=C(C(=O)C(O1)C2=CC=CC=C2)C3=CC(=CC=C3)C(F)(F)F</chem>
Flusilazole	G	E	<chem>C[Si](CN1C=NC=N1)(C2=CC=C(C=C2)F)C3=CC=C(C=C3)F</chem>
flutolanil	L	F	<chem>CC(C)OC1=CC=CC(=C1)NC(=O)C2=CC=CC=C2C(F)(F)F</chem>
Flutriafol	L	E	<chem>C1=CC=C(C(=C1)C(CN2C=NC=N2)(C3=CC=C(C=C3)F)O)F</chem>
Fluvalinate	G	Both	<chem>CC(C)C(C(=O)OC(C#N)C1=CC(=CC=C1)OC2=CC=CC=C2)NC3=C(C=C(C=C3)C(F)(F)F)Cl</chem>
Forchlorfenuron	L	E	<chem>C1=CC=C(C=C1)NC(=O)NC2=CC(=NC=C2)Cl</chem>
Furalaxyl	G	E	<chem>CC1=C(C(=CC=C1)C)N(C(C)C(=O)OC)C(=O)C2=CC=CO2</chem>
heptachlor epoxide	G	F	<chem>C12C(C(C3C1O3)Cl)C4(C(=C(C2(C4(Cl)Cl)Cl)Cl)Cl)Cl</chem>
hexachlorobenzene	G	F	<chem>C1(=C(C(=C(C(=C1Cl)Cl)Cl)Cl)Cl)Cl</chem>
hexaconazole	L	F	<chem>CCCC(CN1C=NC=N1)(C2=C(C=C(C=C2)Cl)Cl)O</chem>

*Continue to next page*

*Continue from previous page*

Pesticides	Technology	List	Canonical SMILES
Hexythiazox	L	E	<chem>CC1C(SC(=O)N1C(=O)NC2CCCCC2)C3=CC=C(C=C3)Cl</chem>
imazalil	L	F	<chem>C=CCOC(CN1C=CN=C1)C2=C(C=C(C=C2)Cl)Cl</chem>
Imidacloprid	L	E	<chem>C1CN(C(=N1)N[N+](=O)[O-])CC2=CN=C(C=C2)Cl</chem>
iprodione	G	Both	<chem>CC(C)NC(=O)N1CC(=O)N(C1=O)C2=CC(=CC(=C2)Cl)Cl</chem>
Iprovalicarb	L	E	<chem>CC1=CC=C(C=C1)C(C)NC(=O)C(C(C)C)NC(=O)OC(C)C</chem>
Isocarbophos	G	E	<chem>CC(C)OC(=O)C1=CC=CC=C1OP(=S)(N)OC</chem>
Isufenphos-Methyl	G	E	<chem>CC(C)NP(=S)(OC)OC1=CC=CC=C1C(=O)OC(C)C</chem>
linuron	L	Both	<chem>CN(C(=O)NC1=CC(=C(C=C1)Cl)Cl)OC</chem>
Malaoxon	L	E	<chem>CCOC(=O)CC(C(=O)OCC)SP(=O)(OC)OC</chem>
Mepanipyrim	G	E	<chem>CC#CC1=NC(=NC(=C1)C)NC2=CC=CC=C2</chem>
Metalaxyl	G	E	<chem>CC1=C(C(=CC=C1)C)N(C(C)C(=O)OC)C(=O)COC</chem>
Metamitron	L	E	<chem>CC1=NN=C(C(=O)N1N)C2=CC=CC=C2</chem>
Metconazole	L	E	<chem>CC1(CCC(C1(CN2C=NC=N2)O)CC3=CC=C(C=C3)Cl)C</chem>
methamidophos	L	F	<chem>COP(=O)(N)SC</chem>
Methidathion	G	E	<chem>COC1=NN(C(=O)S1)CSP(=S)(OC)OC</chem>
Methiocarb	L	E	<chem>CC1=CC(=CC(=C1SC)C)OC(=O)NC</chem>
Methiocarb Sulfone	L	E	<chem>CC1=CC(=CC(=C1S(=O)(=O)C)C)OC(=O)NC</chem>
Methiocarb Sulfoxide	L	E	<chem>CC1=CC(=CC(=C1S(=O)C)C)OC(=O)NC</chem>
Methomyl	L	E	<chem>CC(=NOC(=O)NC)SC</chem>
methyl parathion	G	Both	<chem>COP(=S)(OC)OC1=CC=C(C=C1)[N+](=O)[O-]</chem>
metolachlor	L	F	<chem>CCC1=CC=CC(=C1N(C(C)COC)C(=O)CCl)C</chem>
metolcarb	L	F	<chem>CC1=CC(=CC=C1)OC(=O)NC</chem>
Metosulam	L	E	<chem>CC1=C(C(=C(C=C1)Cl)NS(=O)(=O)C2=NN3C(=CC(=NC3=N2)OC)OC)Cl</chem>
mevinphos	L	F	<chem>CC(=CC(=O)OC)OP(=O)(OC)OC</chem>
MGK-264	G	F	<chem>CCCCC(CC)CN1C(=O)C2C3CC(C2C1=O)C=C3</chem>
monocrotophos	L	Both	<chem>CC(=CC(=O)NC)OP(=O)(OC)OC</chem>
monolinuron	L	F	<chem>CN(C(=O)NC1=CC=C(C=C1)Cl)OC</chem>
napropamide	G	F	<chem>CCN(CC)C(=O)C(C)OC1=CC=CC2=CC=CC=C21</chem>
Neburon	L	E	<chem>CCCCN(C)C(=O)NC1=CC(=C(C=C1)Cl)Cl</chem>
o-phenylphenol	G	Both	<chem>C1=CC=C(C=C1)C2=CC=CC=C2O</chem>
o,p'-methoxychlor	G	F	<chem>COC1=CC=C(C=C1)C(C2=CC=CC=C2OC)C(Cl)(Cl)Cl</chem>
omethoate	L	Both	<chem>CNC(=O)CSP(=O)(OC)OC</chem>
oxadixyl	G	F	<chem>CC1=C(C(=CC=C1)C)N(C(=O)COC)N2CCOC2=O</chem>
Oxamyl	L	E	<chem>CNC(=O)ON=C(C(=O)N(C)C)SC</chem>
Oxyfluorfen	G	E	<chem>CCOC1=C(C(=CC(=C1)OC2=C(C=C(C=C2)C(F)(F)F)Cl)[N+](=O)[O-])</chem>
Paclobutrazole	L	E	<chem>CC(C)(C)C(C(C1=CC=C(C=C1)Cl)N2C=NC=N2)O</chem>
Paraoxon Methyl	L	E	<chem>COP(=O)(OC)OC1=CC=C(C=C1)[N+](=O)[O-]</chem>
parathion	G	F	<chem>CCOP(=S)(OCC)OC1=CC=C(C=C1)[N+](=O)[O-]</chem>
penconazole	L	F	<chem>CCCC(CN1C=NC=N1)C2=C(C=C(C=C2)Cl)Cl</chem>
Pencycuron	L	E	<chem>C1CCC(C1)N(CC2=CC=C(C=C2)Cl)C(=O)NC3=CC=CC=C3</chem>
Pendimethalin	G	E	<chem>CCC(CC)NC1=C(C=C(C(=C1[N+](=O)[O-])C)C)[N+](=O)[O-]</chem>
pentachloroaniline	G	F	<chem>C1(=C(C(=C(C(=C1Cl)Cl)Cl)Cl)Cl)N</chem>
pentachlorobenzene	G	F	<chem>C1=C(C(=C(C(=C1Cl)Cl)Cl)Cl)Cl</chem>
permethrin	G	F	<chem>CC1(C(C1C(=O)OCC2=CC(=CC=C2)OC3=CC=CC=C3)C=C(C1)Cl)C</chem>
Pethoxamid	L	E	<chem>CCOCCN(C(=O)CCl)C(=C(C)C)C1=CC=CC=C1</chem>
Phenthoate	G	E	<chem>CCOC(=O)C(C1=CC=CC=C1)SP(=S)(OC)OC</chem>
Phosalone	G	Both	<chem>CCOP(=S)(OCC)SCN1C2=C(C=C(C=C2)Cl)OC1=O</chem>
phosmet	L	Both	<chem>COP(=S)(OC)SCN1C(=O)C2=CC=CC=C2C1=O</chem>
Picolinafen	G	E	<chem>C1=CC(=CC(=C1)OC2=CC=CC(=N2)C(=O) \ NC3=CC=C(C=C3)F)C(F)(F)F</chem>
Picoxystrobin	L	E	<chem>COC=C(C1=CC=CC=C1COC2=CC=CC(=N2)C(F)(F)F)C(=O)OC</chem>
Piperonyl butoxide	L	F	<chem>CCCCOCCOCCOCC1=CC2=C(C=C1CCC)OCO2</chem>
Piridafenthion	G	E	<chem>CCOP(=S)(OCC)OC1=NN(C(=O)C=C1)C2=CC=CC=C2</chem>
pirimiphos-methyl	G	F	<chem>CCN(CC)C1=NC(=CC(=N1)OP(=S)(OC)OC)C</chem>
prochloraz	L	F	<chem>CCCN(CCOC1=C(C=C(C=C1Cl)Cl)Cl)C(=O)N2C=NC=C2</chem>
procymidone	G	Both	<chem>CC12CC1(C(=O)N(C2=O)C3=CC(=CC(=C3)Cl)Cl)C</chem>
profenofos	G	Both	<chem>CCCS(=O)(OCC)OC1=C(C=C(C=C1)Br)Cl</chem>
prometryn	L	F	<chem>CC(C)NC1=NC(=NC(=N1)SC)NC(C)C</chem>
pronamide	G	Both	<chem>CC(C)(C#C)NC(=O)C1=CC(=CC(=C1) \ Cl)Cl</chem>
propachlor	L	F	<chem>CC(C)N(C1=CC=CC=C1)C(=O)CCl</chem>
propanil	G	F	<chem>CCC(=O)NC1=CC(=C(C=C1)Cl)Cl</chem>
propargite	L	F	<chem>CC(C)(C)C1=CC=C(C=C1)OC2CCCCC2OS(=O)OCC#C</chem>

*Continue to next page*

*Continue from previous page*

Pesticides	Technology	List	Canonical SMILES
Pymetrozine	L	E	<chem>CC1=NNC(=O)N(C1)N=CC2=CN=CC=C2</chem>
Pyraclostrobin	L	E	<chem>COC(=O)N(C1=CC=CC=C1COC2=NN(C=C2)C3=CC=C(C=C3)Cl)OC</chem>
Pyrazophos	G	E	<chem>CCOC(=O)C1=CN2C(=CC(=N2)OP(=S)(OCC)OCC)N=C1C</chem>
Pyridaben	G	E	<chem>CC(C)(C)C1=CC=C(C(=C1)CSC2=C(C(=O)N(N=C2)C(C)(C)C)Cl</chem>
pyriproxifen	G	Both	<chem>CC(COC1=CC=C(C(=C1)OC2=CC=CC=C2)OC3=CC=CC=N3</chem>
quinalphos	G	Both	<chem>CCOP(=S)(OCC)OC1=NC2=CC=CC=C2N=C1</chem>
Tebuconazole	G	E	<chem>CC(C)(C)C(CCC1=CC=C(C(=C1)Cl)(CN2C=NC=N2)O</chem>
Tefluthrin	G	E	<chem>CC1=C(C(=C(C(=C1F)F)COC(=O)C2C(C2(C)C)C=C(C(F)(F)F)Cl)F)F</chem>
Terbufos	G	E	<chem>CCOP(=S)(OCC)SCSC(C)(C)C</chem>
Terbutryn	L	E	<chem>CCNC1=NC(=NC(=N1)SC)NC(C)(C)C</chem>
Tetraconazole	G	E	<chem>C1=CC(=C(C=C1Cl)Cl)C(CN2C=NC=N2)COC(C(F)F)(F)F</chem>
tetradifon	G	F	<chem>C1=CC(=CC=C1S(=O)(=O)C2=CC(=C(C=C2Cl)Cl)Cl)Cl</chem>
Thiabendazole	L	E	<chem>C1=CC=C2C(=C1)NC(=N2)C3=CSC=N3</chem>
Thiacloprid	L	E	<chem>C1CSC(=NC # N)N1CC2=CN=C(C=C2)Cl</chem>
Thiamethoxam	L	E	<chem>CN1COCN(C1=N[N+](=O)[O-])CC2=CN=C(S2)Cl</chem>
tolclofos-methyl	G	Both	<chem>CC1=CC(=C(C(=C1)Cl)OP(=S)(OC)OC)Cl</chem>
Triadimefon	L	E	<chem>CC(C)(C)C(=O)C(N1C=NC=N1)OC2=CC=C(C=C2)Cl</chem>
Triadimenol	L	E	<chem>CC(C)(C)C(C(N1C=NC=N1)OC2=CC=C(C=C2)Cl)O</chem>
triallate	G	F	<chem>CC(C)N(C(C)C)C(=O)SCC(=C(Cl)Cl)Cl</chem>
Triazophos	G	E	<chem>CCOP(=S)(OCC)OC1=NN(C=N1)C2=CC=CC=C2</chem>
Trifloxystrobin	L	E	<chem>CC(=NOCC1=CC=CC=C1C(=NOC)C(=O)OC)C2=CC(=CC=C2)C(F)(F)F</chem>
Triflumizole	L	E	<chem>CCCOCC(=NC1=C(C=C(C=C1)Cl)C(F)(F)F)N2C=CN=C2</chem>
Trifluralin	G	Both	<chem>CCCN(CCC)C1=C(C=C(C=C1[N+](=O)[O-])C(F)(F)F)[N+](=O)[O-]</chem>
Triticonazole	L	E	<chem>CC1(CCC(=CC2=CC=C(C=C2)Cl)C1(CN3C=NC=N3)O)C</chem>
vinclozolin	G	Both	<chem>CC1(C(=O)N(C(=O)O1)C2=CC(=CC(=C2)Cl)Cl)C=C</chem>
Zoxamide	L	E	<chem>CCC(C)(C(=O)CC1NC(=O)C1=CC(=C(C(=C1)Cl)C)Cl</chem>

*End of table*

## B.2 Accuracy and log Execution Time(Av log ET). Classification methods (Method) are ordered according to accuracy.

Table B.2. Accuracy and log Execution Time (Av log ET). Classification methods (Method) are ordered according to the accuracy.

Method	Accuracy	Av log ET	Method	Accuracy	Av log ET	Method	Accuracy	Av log ET
xgbDART	0.850	2.81	rotationForest	0.819	2.54	svmLinear2	0.768	0.40
xgbTree	0.850	2.07	LMT	0.819	1.19	rfRules	0.766	2.62
extraTrees	0.846	1.83	avNNet	0.819	2.40	svmLinear3	0.763	0.71
AdaBoost.M1	0.845	3.84	rrlda	0.814	1.79	evtree	0.757	2.07
C5.0	0.841	1.58	dwdPoly	0.814	1.70	stepQDA	0.757	2.19
C5.0Cost	0.841	1.94	mlpWeightDecay	0.814	2.95	monmlp	0.752	2.26
msaenet	0.840	2.25	mlpWeightDecayML	0.814	2.98	multinom	0.747	0.59
ORFsvm	0.840	2.95	rotationForestCp	0.814	2.81	pda2	0.747	0.54
gbm	0.840	1.06	sda	0.814	0.49	plr	0.747	0.61
adaboost	0.839	2.93	slda	0.814	0.40	blackboost	0.747	2.65
wsrf	0.839	2.28	earth	0.814	0.94	sparseLDA	0.743	2.83
glmboost	0.835	0.38	RRF	0.813	2.77	stepLDA	0.736	2.24
ordinalNet	0.835	3.34	RRFglobal	0.813	2.04	C5.0Rules	0.736	0.82
regLogistic	0.835	1.06	ada	0.809	2.69	rFerns	0.731	1.71
dwdRadial	0.835	0.78	J48	0.809	1.53	rpartScore	0.731	2.18
svmRadialSigma	0.835	0.87	Rborist	0.804	1.96	svmLinear	0.727	0.24
glmnet	0.830	0.53	gaussprRadial	0.804	0.49	naive_bayes	0.726	0.74
ownnn	0.830	0.54	mlp	0.804	2.58	C5.0Tree	0.726	0.79
svmPoly	0.830	1.06	mlpML	0.804	2.56	rocc	0.726	0.58
xyf	0.830	1.59	dwdLinear	0.799	1.01	nb	0.721	0.92
parRF	0.829	1.46	svmLinearWeights	0.799	0.72	ctree	0.720	0.80
svmRadialCost	0.829	0.49	svmLinearWeights2	0.799	1.12	ctree2	0.720	1.24
kknn	0.829	0.71	fda	0.799	0.58	rpart	0.720	0.30
sdwd	0.825	1.15	kernelpls	0.798	0.19	rpart2	0.720	0.32
lvq	0.825	1.48	pls	0.798	0.16	pam	0.716	0.17
ORFpls	0.825	2.20	simpls	0.798	0.24	rpart1SE	0.700	0.30
svmRadial	0.825	0.47	widekernelpls	0.798	0.35	pda	0.681	0.57
svmRadialWeights	0.825	0.67	cforest	0.798	1.77	glmStepAIC	0.680	4.25
deepboost	0.824	3.14	treebag	0.798	1.03	OneR	0.675	0.56
lssvmRadial	0.824	1.89	bayesglm	0.794	0.99	CSimca	0.659	0.43
rf	0.824	1.44	hdda	0.793	0.51	rpartCost	0.658	0.47
rbfDDA	0.824	2.71	LogitBoost	0.792	0.54	gamSpline	0.635	2.36
pcaNNet	0.824	1.45	nodeHarvest	0.788	2.68	gamLoess	0.624	1.75
nnet	0.824	1.69	snn	0.788	0.82	RSimca	0.592	0.83
xgbLinear	0.824	1.97	AdaBag	0.783	3.35	dnn	0.567	2.09
knn	0.820	0.22	BstLm	0.783	0.69	null	0.567	0.00
ORFridge	0.820	2.11	bstTree	0.782	2.55	glm	0.541	0.44
ranger	0.820	1.50	gcvEarth	0.777	0.56	bagEarth	0.274	2.27
bagFDA	0.819	2.33	JRip	0.774	1.97	bagEarthGCV	0.149	1.72
rda	0.819	1.07	PART	0.774	1.30			

## C. Specification of PC and Software programs

Table C.1. Specification of the PC and Software programs

Component	Specification
Personal Computer	Apple Macbook Air 13-inch, early 2015
Processor	1.6GHz Intel Core i5
Memory	8GB 1600MHz DDR3
Operation System	macOS Mojave version 10.14.6
R program	version 3.6.1
R Studio	1.2.5019