Poster Reprint

ASMS 2020
ThP 177

Optimum molecular descriptors based on 89 machine learning methods for predicting the recovery rate of pesticides in crops by GC-MS

Takeshi Serino [1, 2], Sadao Nakamura [1], Yoshizumi Takigawa [1], Tarun Anumol [3], Md. Altaf-UI-Amin [2], Shigehiko Kanaya [2]

[1] Agilent Technologies, Hachioji, Tokyo, Japan

[2] Nara Institute of Science and Technology, Ikoma, Nara, Japan

[3] Agilent Technologies, Wilmington, DE, United States

Introduction

GC-MS is widely used for analysis of residual pesticides in fruits and vegetables (crops). Pesticide recovery ratio by GC-MS can vary by the residual matrix of the crops. We previously developed the prediction models with 89 machine learning methods (Table 1) for pesticide recovery rate using molecular descriptors (MDs)[1]. With rcdk package of R program, 178 MDs were obtained by the canonical SMILES of each pesticide (Table 2). All MDs were used as the explanatory variables for predicting the pesticide recovery rate. Correlation coefficient of some MDs obtained by rcdk were over 0.7, i.e. highly correlated. Some combinations among these MDs are correlated strongly that can influence the performance of regression for pesticide recovery rate prediction such as the multicollinearity[2]. The procedure to select the optimum MD for regression analysis using the correlation analysis and graph clustering tool[3] is developed.

Experimental

There are two considerations below on selection of MDs for machine learning.

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 to minimize the loss of the information utilizing the graph clustering tool.

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 1. There are 5 steps for selection of MDs as below.

The 1st 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 using “corrr” package of R program and “strech” function for all 178 MDs. Based on the guidance of pearson correlation coefficient, the threshold at r = 0.7 is set for the “Highly correlated” of MD combination. The MDs in the combinations of r > 0.7 are classified as “Strongly correlated MD” and the other MDs are “Weakly correlated MD group”.

The 2nd 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.

Experimental

Table 1. Machine Learning methods for regression analysis used in present 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, ctree, ctree2, 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 (14)	cforest, ranger, qrf, rf, parRF, extraTrees, Rborist, RRFglobal, RRF, treebag, bstTree, gbm, xgbTree, nodeHarvest
Simple Linear (3)	BstLm, glmboost (GLM), xgbLinear
Spline (3)	bagEarthGCV, bagEarth, xgbDART

Table 2. 178 molecular descriptors in present study

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 boron atoms)
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 hound 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 (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.sssSi (>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)
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 (κ) 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 nitroqens), MDEN.22 (between all secondary nitroqens), MDEN.23 (between all secondary and tertiary nitrogens), MDEN.33 (between all tertiary nitrogens)
PetitjeanNumberDescriptor (1)	PetitjeanNumber (Petitjean number)
RotatableBondsCountDescriptor (1)	nRotB (Number of rotatable bonds, excluding terminal bonds)
RuleOfFiveDescriptor (1)	LipinskiFailures (Number failures of the Lipinski's Rule Of 5)
TPSADescriptor (19)	TopoPSA (Topological polar surface area)
VAdjMaDescriptor (1)	VAdjMat (Vertex adjacency information (magnitude))
WeightDescriptor (1)	MW (Molecular weight)
WeightedPathDescriptor (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)
WienerNumbersDescriptor (2)	WPATH (Weiner path number), WPOL (Weiner polarity number)
XLogPDescriptor (1)	XLogP (XLogP)
ZagrebIndexDescriptor (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 (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)

Table 3. DPClus parameters

Parameter	Value
CP Value	0.5
Density Value	0.9
Minimum Cluster Value	2

Experimental

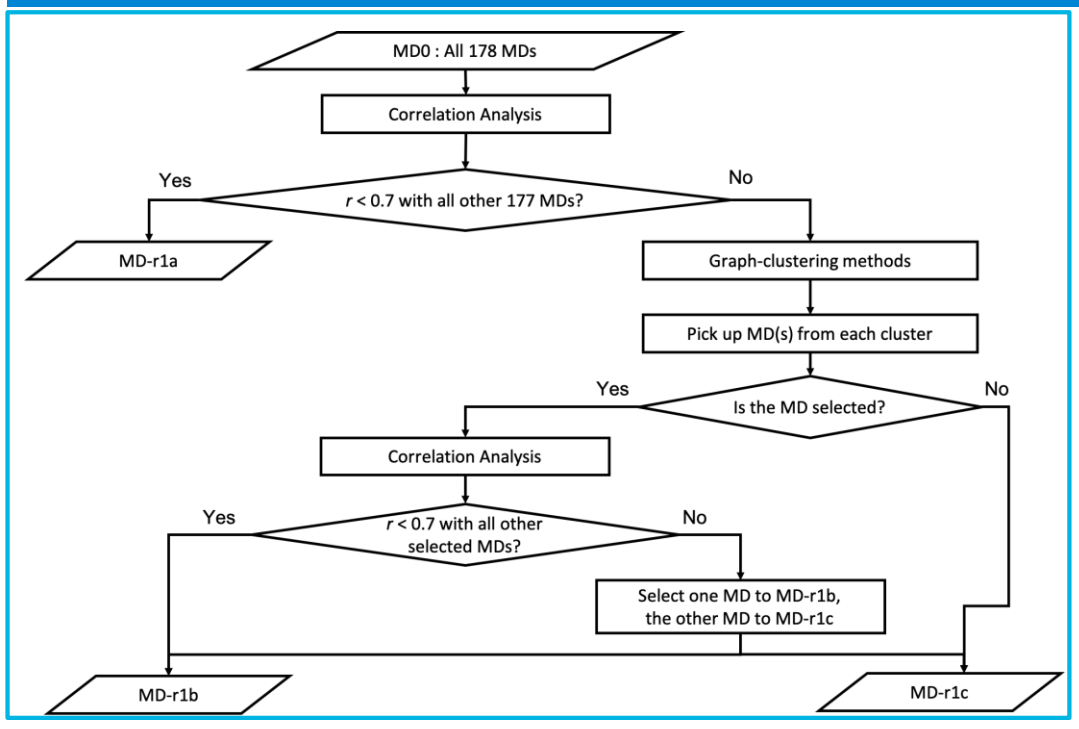

Figure 1. Process chart of selecting the optimum MDs

The 3rd step is to visualize the correlations of strongly correlated MDs by the method of graph clustering method called DPClus[3] and pick up the representative MD(s) from each cluster according to the flow chart in Figure 2, based on the considerations of removal of highly correlated MDs while minimizing the loss of information. The parameters of DPClus software is set as in Table 3.

The 4th step is the second correlation analysis among the representative MDs picked up from each cluster. Threshold of correlation is $r > 0.7$.

The final step is to select the MD(s) based on the step 4. The MDs of weak correlation with other MDs in step 4 are grouped as “MD-r1b”, which are used for regression analysis. Other MDs in step 4 are divided into two groups, ”MD-r1b” and “MD-r1c”. MDs in group MD-r1c are excluded from regression analysis.

Thus, 178 MDs are divided into three groups as listed in the Table 4 according to the process in Figure 1.

Prediction Error (PE)

$$PE_j = \frac{\sum_{i=1}^{N}\left(y_{obs}^{(ij)} - y_{pred}^{(ij)}\right)^2}{\sum_{i=1}^{N}\left(y_{obs}^{(ij)} - \bar{y}^{(j)}\right)^2} \qquad (1)$$

Table 4. Molecular Descriptors selected by the correlation analysis and cluster analysis

MD group	Description of MDs	# 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

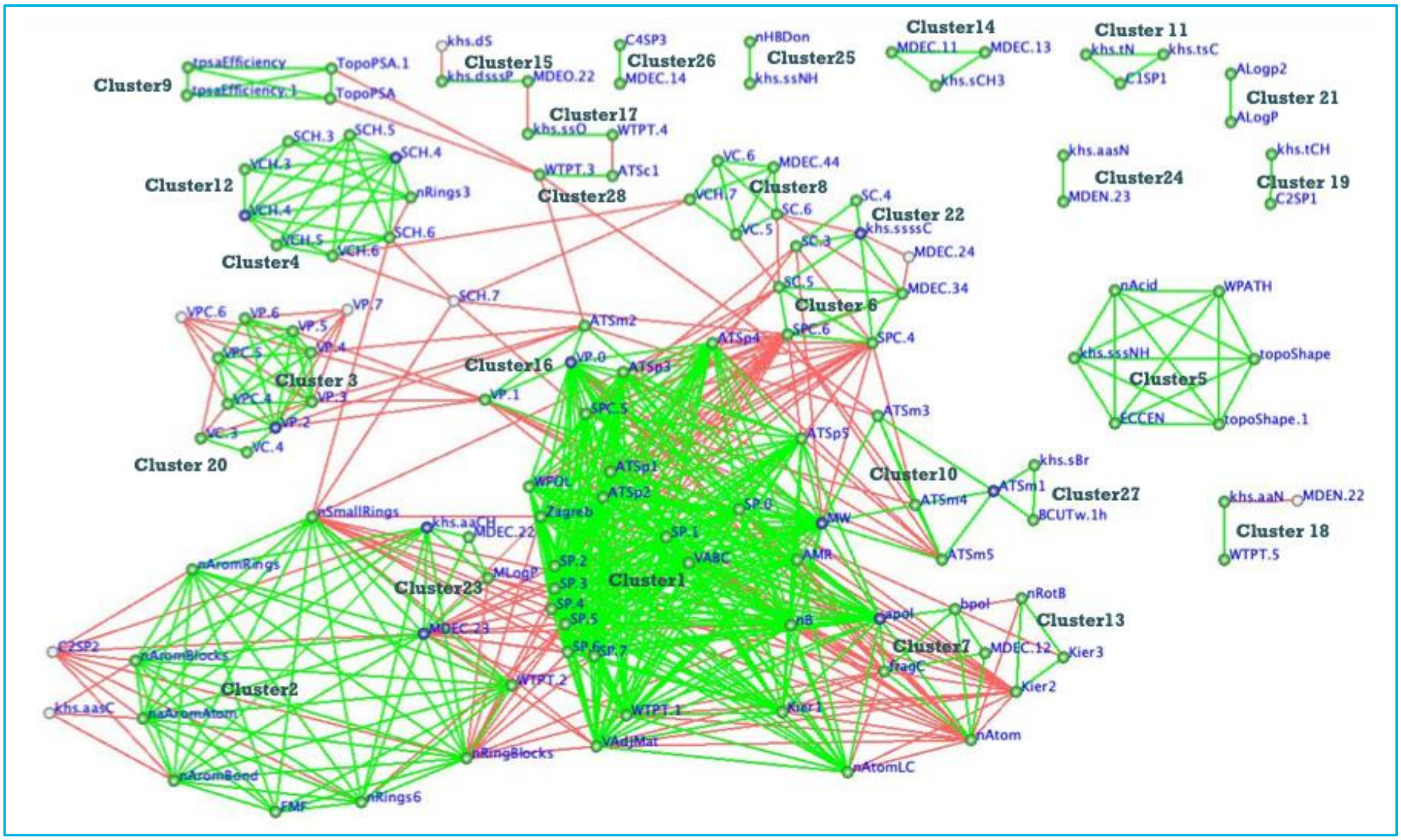

Figure 3. Result of cluster analysis result by DPClus

Experimental

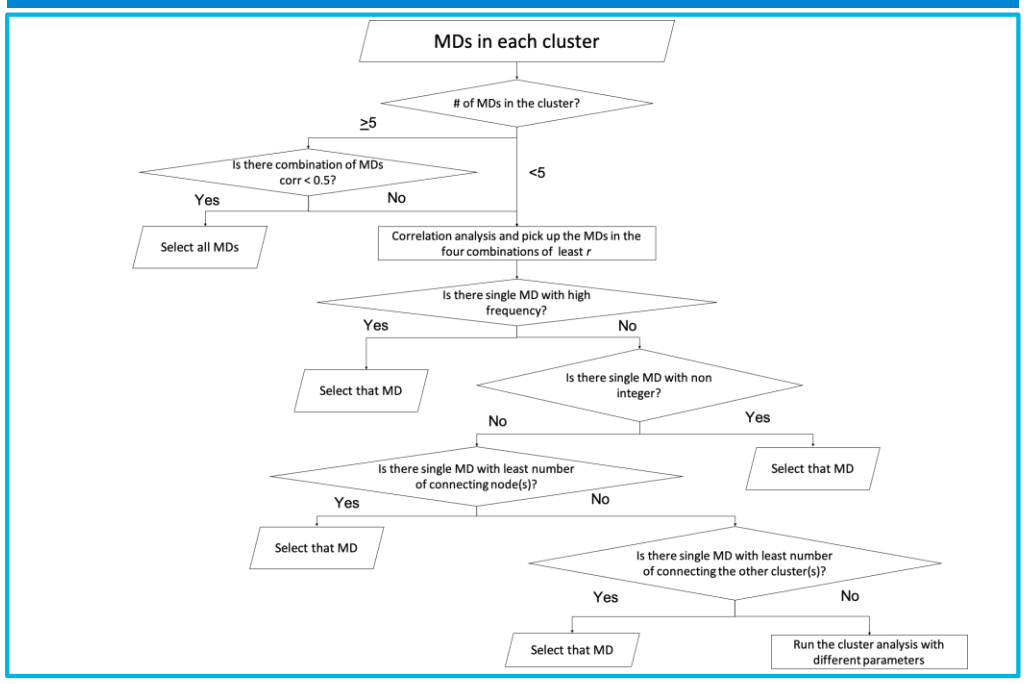

Figure 2. Process chart of selecting the optimum MDs from each cluster

Results and discussion

Correlation analysis among molecular descriptors

658 combinations consisted of 118 MDs were $r \geq 0.7$, which correlate strongly. Other 60 MDs were correlated with the other MDs at $r < 0.7$, which was classified as the MD group MD-r1a in the Table 4.

Selection of molecular descriptors by the clustering tool

Relationships of 118 MDs of strongly correlated with any other MD(s) were classified into 28 clusters according to the cluster analysis by DPClus, shown in Figure 3. 19 clusters are connected by red lines each other dependently and MDs in the other 9 clusters has no connection with the other clusters.

Selection of molecular descriptors for machine learning after cluster analysis

The combinations of MDs with $r > 0.7$ among them are listed in the Table 5. The MDs of the column MD-a in the Table 5 are classified to MD-1rb and MD-b are classified to MD-r1c of Table 4.

Table 5. Combination of MDs with the r > 0.7 after after selection of cluster analysis

MD-a	MD-b	Correlation coefficient	MD-a	MD-b	Correlation coefficient
khs.aaCH	MDEC.22	0.833	nAtomLC	VP.0	0.751
ATSm1	BCUTw.1h	0.777	WTPT.4	ATSc1	0.736
VCH.4	SCH.3	0.769	SPC.5	MDEC.34	0.731

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. Top and bottom 10 machine learning methods in prediction error by selecting MDs are listed in Table 6.

bagEarthGCV(LogPE -0.473 to 3.501), ppr(-1.482 to -0.824) and 4 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. lasso(LogPE 23.723 to -0.232) and lars(8.871 to -0.161) bagEarth(5.398 to -0.283) were improved by selecting the MDs with cluster analysis. Ordinary Decision trees, Ordinary Centroid and Ensemble Simple Liner show-small differences in prediction error by the selection of MDs.

Results and discussion

Table 6. Top 10 methods of best (right) and worst (left) in differences of prediction error by selecting the molecular descriptors, sorted by the Prediction Error difference

Met	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. Newral Network	-0.364	-0.693	0.329
gaussprPoly	O. Kernel	-0.163	-0.445	0.282

Met	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

Conclusions

The procedure to remove the highly correlated explanatory variables of molecular descriptors with correlation analysis and graph clustering tool are developed for the data set of residual pesticide recovery. Correlation analysis is applied on all 178 molecular descriptors and finally 83 molecular descriptors were selected for regression analysis of 89 machine learning methods. Prediction error of machine learning methods in the ordinary sparse model improved by removal of highly correlated molecular descriptors. On the other hand, prediction error of the methods of ordinary simple liner got worse by the selection of molecular descriptors.

References

[1]Takeshi Serino, et al. 67th American Society for Mass Spectrometry TP-298. *Comprehensive Machine Learning Prediction of GC/MS Pesticide Recovery Based on the Molecular Fingerprinting for Food QA/QC*. Atlanta, GA. June, 2019.

[2] A. Garg and K. Tai. 2013. Comparison of statistical and machine learning methods in modelling of data with multicollinearity. *International Journal of Modelling, Identification and Control(IJMIC)* **18:** 295-312.

[3] 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

Agilent
Trusted Answers