NEW APPROACH FOR QSAR AND QSTR TREND ANALYSIS ON LARGE SAMPLE DATA SET BY THE KY-METHODS

○ Kohtaro Yuta In Silico Data, Ltd. (http://www.insilicodata.com)

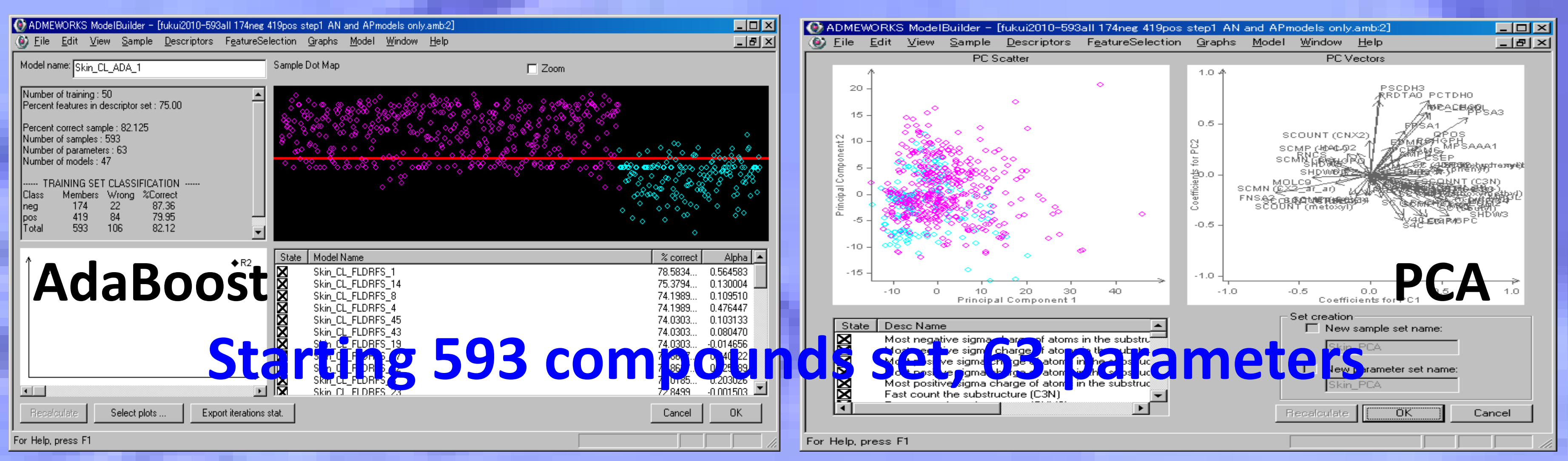
◆ Objection : Big data analysis by the KY-method on QSAR oriented research works

◆ Used samples : Skin sensitization data

Total ; 593, Positive 419, Negative; 174

◆ Classification Results by various methods (63params)

Methods	Total	Positive	Negative
N.N.	86.0%	86.2%	85.6%
SVM	91.7%	98.3%	75.9%
LDA	87.0%	95.2%	67.2%
KNN (K=5)	77.7%	86.9%	55.8%
AdaBoost	82.1%	80.0%	87.4%

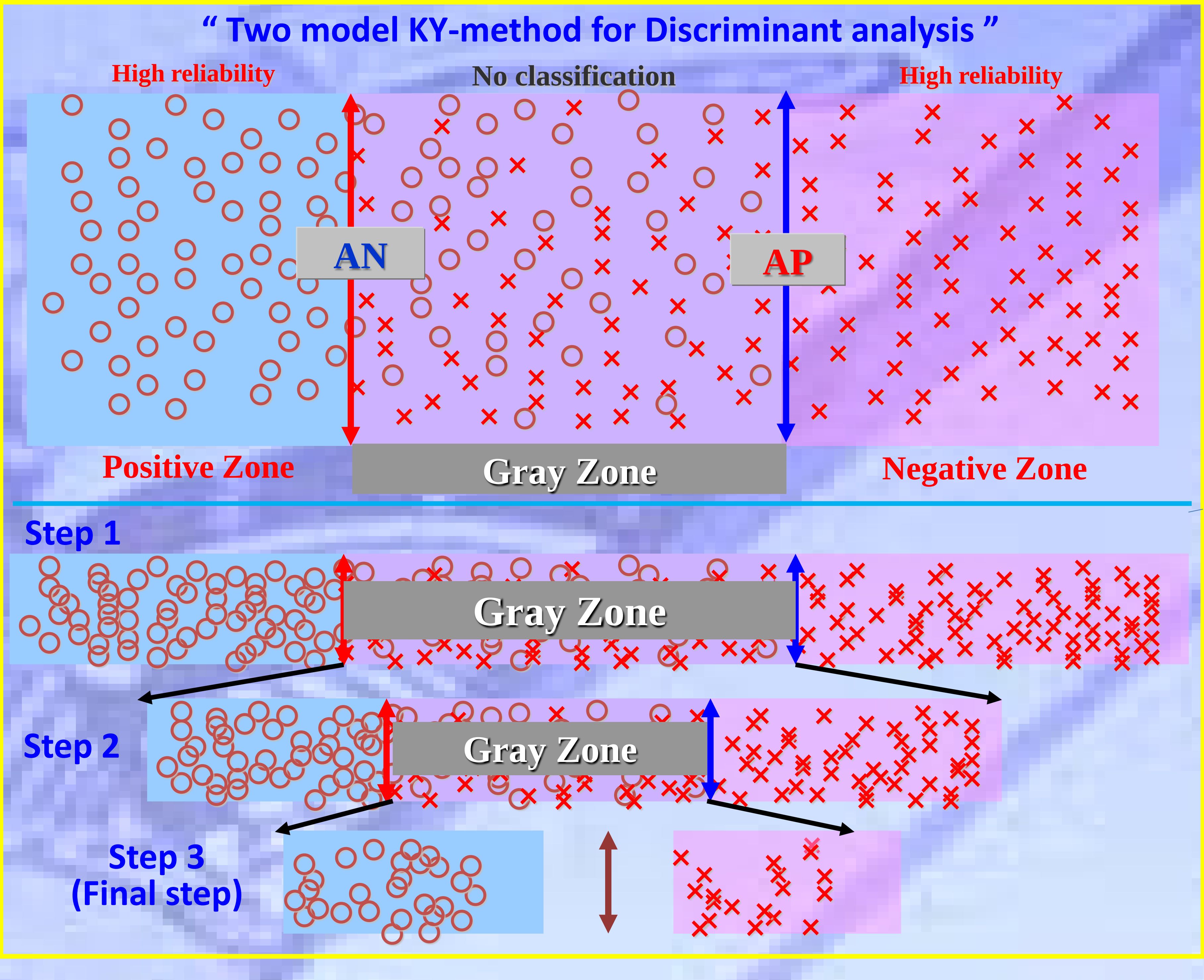


Starting 593 compounds, 63 parameters

◆ Features of the KY (K-step Yard sampling) methods

1. Always achieve perfect (100%) classification under any conditions
   - ･Highly overlapped class sample data set
   - ･Quite large number of sample data set (tens and several thousands of)
2. Starting sample set was divided into
   - ･ small and clean sample set
   - ･ small and hierarchical sample set
3. Applicable not only the discriminant but multi-regression analysis

" Two model KY-method for Discriminant analysis "



◆ Variation of the "KY (K-step Yard sampling) methods"

List of the available "KY methods" : Total 6 approaches

◇ Binary classification ; 3approaches

1. Two model KY− discriminant method
2. One model KY− discriminant method
3. Model free KY− discriminant method

◇ Fitting (multi regression); 3approaches

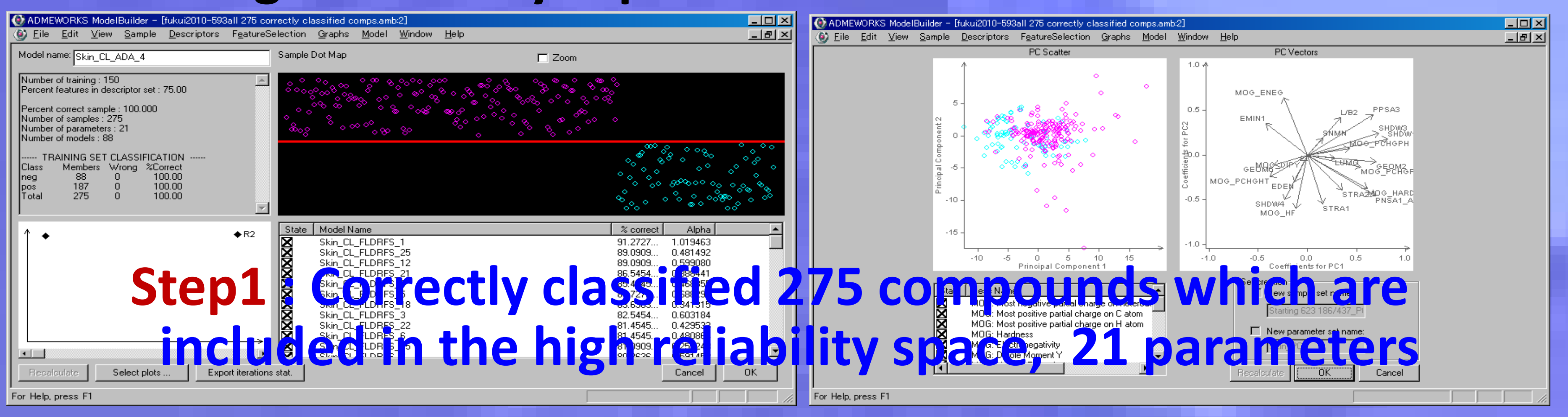
1. KY−fitting with discriminant method
2. Three zone KY−fitting method
3. Model free KY−fitting method

* All six KY-methods are patented and pending applications in JP, USA and EU

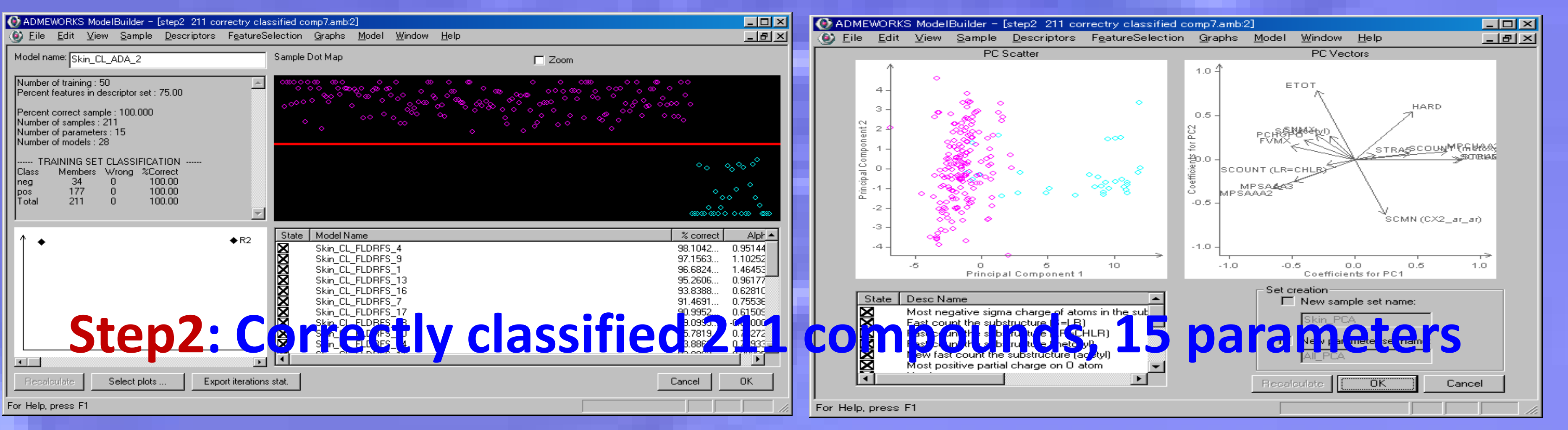
◆ Classification result by the KY-method (100% correct)

Step1;	Positive 187	Negative 88	Grey zone 318
Step2;	Positive 177	Negative 34	Grey zone 107
Step3;	Positive 55	Negative 52	Grey zone 0

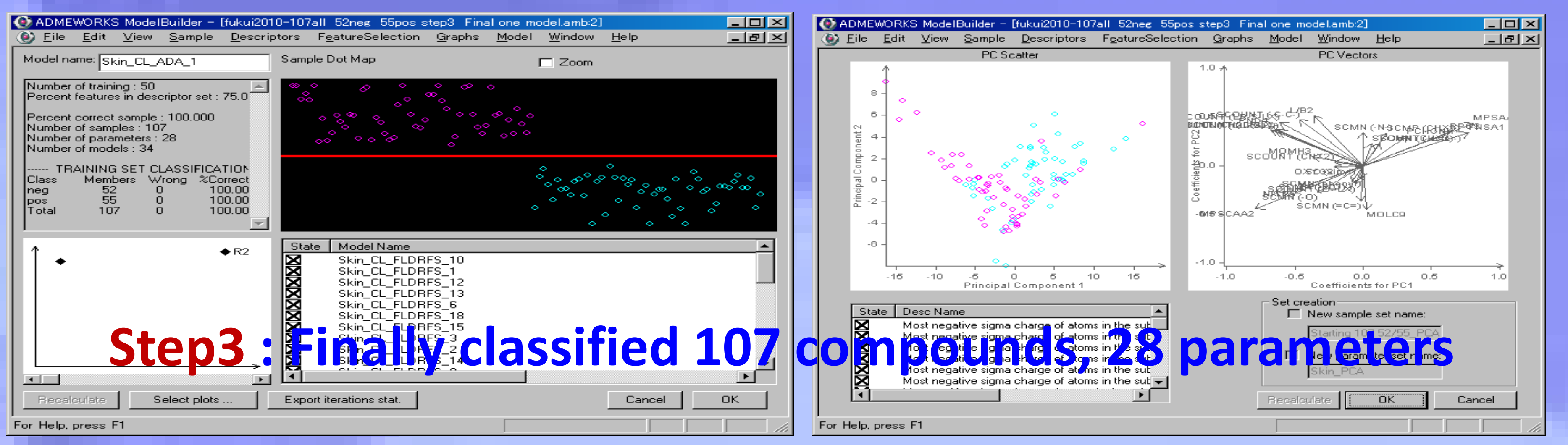
◆ Classification results of compounds which are included in the "High reliability" space of the "KY-methods"



Step1 : Correctly classified 275 compounds which are included in the high reliability space, 21 parameters

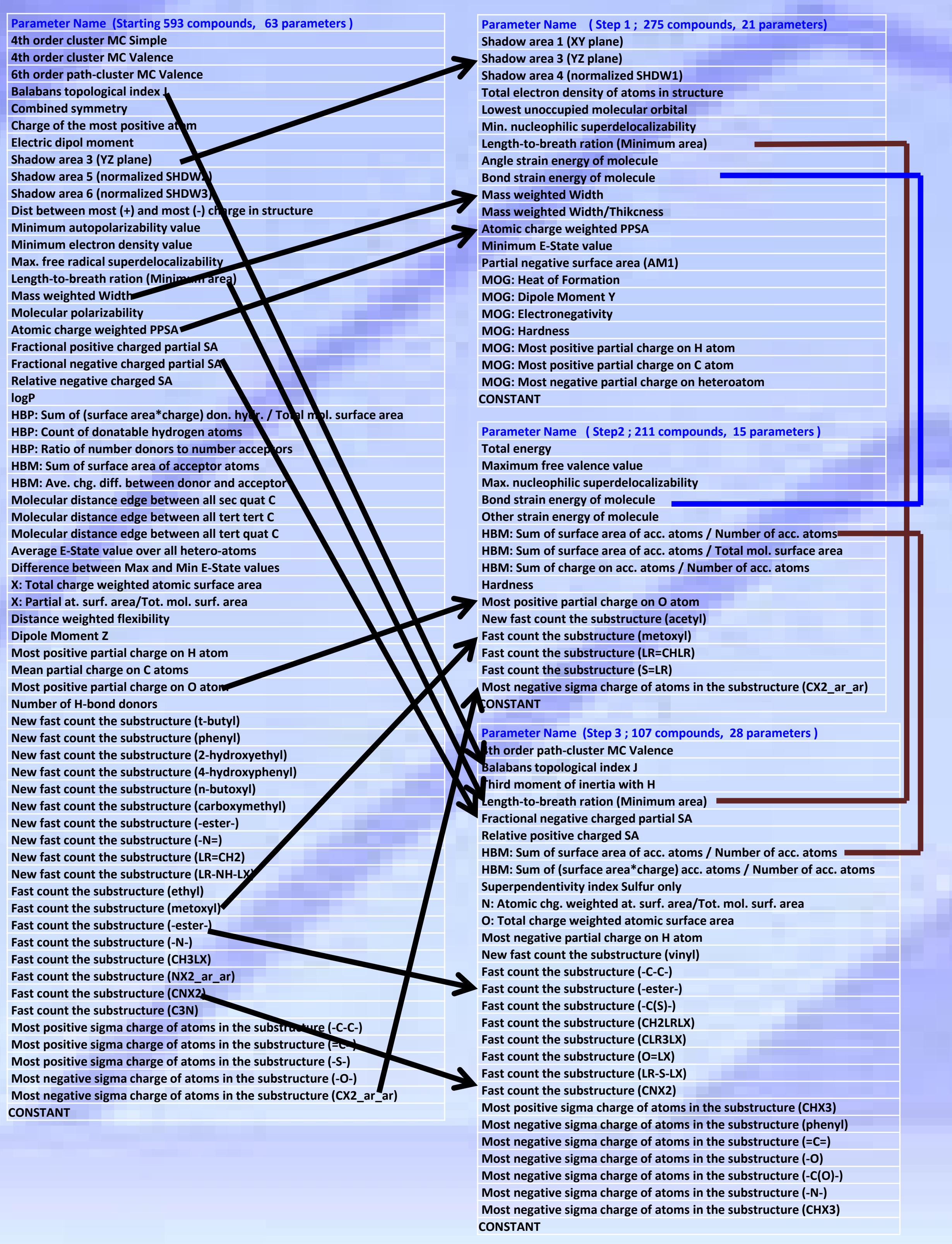


Step2: Correctly classified 211 compounds, 15 parameters



Step3 : Finally classified 107 compounds, 28 parameters

◆ Parameters used in the original and each steps in KY-method



Parameter Name (Starting 593 compounds, 63 parameters)
4th order cluster MC Simple
4th order cluster MC Valence
6th order path-cluster MC Valence
Balabans topological index J
Combined symmetry
Charge of the most positive atom
Electric dipol moment
Shadow area 3 (YZ plane)
Shadow area 5 (normalized SHDW2)
Shadow area 6 (normalized SHDW3)
Dist between most (+) and most (-) charge in structure
Minimum autopolarizability value
Minimum electron density value
Max. free radical superdelocalizability
Length-to-breath ration (Minimum area)
Mass weighted Width
Molecular polarizability
Atomic charge weighted PPSA
Fractional positive charged partial SA
Fractional negative charged partial SA
Relative negative charged SA
logP
HBP: Sum of (surface area*charge) don. hydr. / Total mol. surface area
HBP: Count of donatable hydrogen atoms
HBP: Ratio of number donors to number acceptors
HBM: Sum of surface area of acceptor atoms
HBM: Ave. chg. diff. between donor and acceptor
Molecular distance edge between all sec quat C
Molecular distance edge between all tert tert C
Molecular distance edge between all tert quat C
Average E-State value over all hetero-atoms
Difference between Max and Min E-State values
X: Total charge weighted atomic surface area
X: Partial at. surf. area/Tot. mol. surf. area
Distance weighted flexibility
Dipole Moment Z
Most positive partial charge on H atom
Mean partial charge on C atoms
Most positive partial charge on O atom
Number of H-bond donors
New fast count the substructure (t-butyl)
New fast count the substructure (phenyl)
New fast count the substructure (2-hydroxyethyl)
New fast count the substructure (4-hydroxyphenyl)
New fast count the substructure (n-butoxyl)
New fast count the substructure (carboxymethyl)
New fast count the substructure (-ester-)
New fast count the substructure (-N=)
New fast count the substructure (LR=CH2)
New fast count the substructure (LR-NH-LX)
Fast count the substructure (ethyl)
Fast count the substructure (metoxyl)
Fast count the substructure (-ester-)
Fast count the substructure (-N-)
Fast count the substructure (CH3LX)
Fast count the substructure (NX2_ar_ar)
Fast count the substructure (CNX2)
Fast count the substructure (C3N)
Most positive sigma charge of atoms in the substructure (-C-C-)
Most positive sigma charge of atoms in the substructure (=C-)
Most positive sigma charge of atoms in the substructure (-S-)
Most negative sigma charge of atoms in the substructure (-O-)
Most negative sigma charge of atoms in the substructure (CX2_ar_ar)
CONSTANT

Parameter Name (Step 1 ; 275 compounds, 21 parameters)
Shadow area 1 (XY plane)
Shadow area 3 (YZ plane)
Shadow area 4 (normalized SHDW1)
Total electron density of atoms in structure
Lowest unoccupied molecular orbital
Min. nucleophilic superdelocalizability
Length-to-breath ration (Minimum area)
Angle strain energy of molecule
Bond strain energy of molecule
Mass weighted Width
Mass weighted Width/Thikcness
Atomic charge weighted PPSA
Minimum E-State value
Partial negative surface area (AM1)
MOG: Heat of Formation
MOG: Dipole Moment Y
MOG: Electronegativity
MOG: Hardness
MOG: Most positive partial charge on H atom
MOG: Most positive partial charge on C atom
MOG: Most negative partial charge on heteroatom
CONSTANT

Parameter Name (Step2 ; 211 compounds, 15 parameters)
Total energy
Maximum free valence value
Max. nucleophilic superdelocalizability
Bond strain energy of molecule
Other strain energy of molecule
HBM: Sum of surface area of acc. atoms / Number of acc. atoms
HBM: Sum of surface area of acc. atoms / Total mol. surface area
HBM: Sum of charge on acc. atoms / Number of acc. atoms
Hardness
Most positive partial charge on O atom
New fast count the substructure (acetyl)
Fast count the substructure (metoxyl)
Fast count the substructure (LR=CHLR)
Fast count the substructure (S=LR)
Most negative sigma charge of atoms in the substructure (CX2_ar_ar)
CONSTANT

Parameter Name (Step 3 ; 107 compounds, 28 parameters)
4th order path-cluster MC Valence
Balabans topological index J
Third moment of inertia with H
Length-to-breath ration (Minimum area)
Fractional negative charged partial SA
Relative positive charged SA
HBM: Sum of surface area of acc. atoms / Number of acc. atoms
HBM: Sum of (surface area*charge) acc. atoms / Number of acc. atoms
Superpendentivity index Sulfur only
N: Atomic chg. weighted at. surf. area/Tot. mol. surf. area
O: Total charge weighted atomic surface area
Most negative partial charge on H atom
New fast count the substructure (vinyl)
Fast count the substructure (-C-C-)
Fast count the substructure (-ester-)
Fast count the substructure (-C(S)-)
Fast count the substructure (CH2LRLX)
Fast count the substructure (CLR3LX)
Fast count the substructure (O=LX)
Fast count the substructure (LR-S-LX)
Fast count the substructure (CNX2)
Most positive sigma charge of atoms in the substructure (CHX3)
Most negative sigma charge of atoms in the substructure (phenyl)
Most negative sigma charge of atoms in the substructure (=C=)
Most negative sigma charge of atoms in the substructure (-O)
Most negative sigma charge of atoms in the substructure (-C(O)-)
Most negative sigma charge of atoms in the substructure (-N-)
Most negative sigma charge of atoms in the substructure (CHX3)
CONSTANT

* Number of parameters used in common with other sample set

Starting set: 63 params
To step 1; 3 params
To step 2; 3 params
To step 3; 5 params

Step 1: 21 params
Starting ; 3 params
To step 2; 1 params
To step 3; 1 params

Step 2: 15 params
Starting ; 3 params
To step 1; 1 params
To step 3; 1 params

Step 3: 28 params
Starting ; 5 params
To step 1; 1 params
To step 2; 1 params

◆ Conclusions:

1. Sophisticated and detailed QSAR analysis has been done by the KY-method even if used sample number was large .
2. Perfect (100%) classification ratio was achieved by the KY-method
   - ･Other traditional methods could not be achieved perfect classification ratio
3. The KY-method applicable not only classification but QSAR works

All research works and screen displays were executed and generated by the ADMEWORKS : ModelBuilder program developed by FJQS (Fujitsu Kyushu Systems Ltd.)